12/16/2019

Inferential Statistics

Liver Cancer

Kenny Barza
USJ FS

Contents

I. Introduction

The human body, from the outside, is fairly simple. Each human have two arms, two legs, two eyes, a mouth, a tongue, a brain … you name it! These are what we call organs. Each organ has the responsibility to assure the functionality of the human being. Some of them make our live easier. A great example is the eye, which grants us the vision sense. In that case, they are important, but not essentials for a person to survive. Other organs are absolutely necessary to the point that humans cannot survive without them. We are talking about organs such as the heart, the liver and so on.

Now, if we dive deeper into the subject, it is without hesitation that we can confirm that the human body is extremely complex. People spend 7 to 10 years studying it, yet they cannot fully understand it. That is because our body – and so each organ- is formed by millions of microscopic organisms. These organisms is what we call the "cells", and they are literally everywhere in our body. In fact, a large group of cells form a tissue, and a large group of tissue forms an organ. From here we can confirm how important these microorganisms are. They are responsible for the well-functioning of the organs and as a result, for the well-functioning of a living being.

A logical question to ask is the following: How are these cells created? Well, we will not go into details about how the cell reproduction mechanism work. All that I can say is that 1 cell gets divided into 2 cells. The 2 newly formed cells will be divided into 4, and so on… This mechanism happen in a very periodic way, and need 24 hour to be completed. In addition to that, the reproduction mechanism is divided into multiple steps. The transition between one step to the other happens very smoothly. You see, problem happen when these steps are not followed properly. I Stage, called the mitosis (stage where the cells start to divide) is very crucial, and if it happens, that due to some external factors the mitosis does not work properly, that is where we will start getting abnormal cells. And this is the beginning of something dangerous. In fact, these abnormal cells do not multiply normally. They proliferate in an unusual way, forming what we call a tumour. I take back my words. This is not the beginning of something dangerous. This is the beginning of cancer. People die from cancer because the tumour formed in an organ become so big, that it sort of invades the normal cells. This will result into the death of an organ, and so the death of a living being.

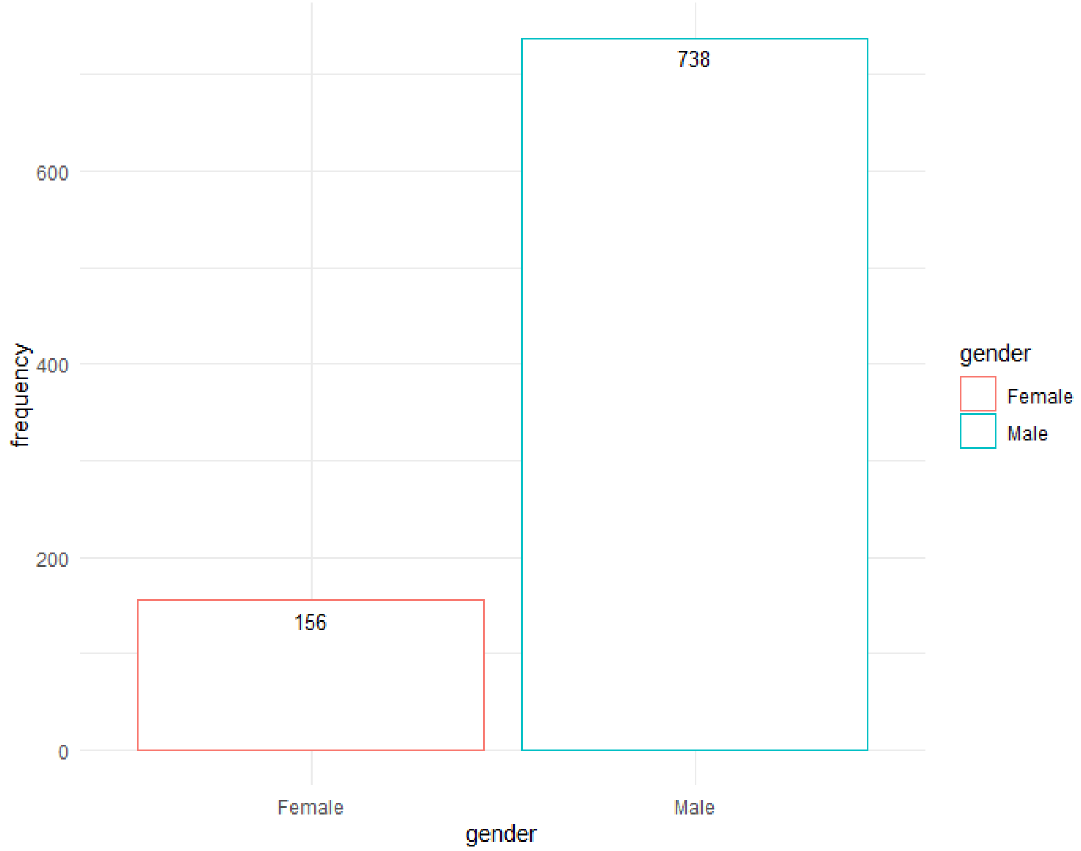
We are interested specifically in the liver cancer, and we have a dataset that we will be analysing in order to answer multiple questions surrounding this disease

II. Descriptive Statistic

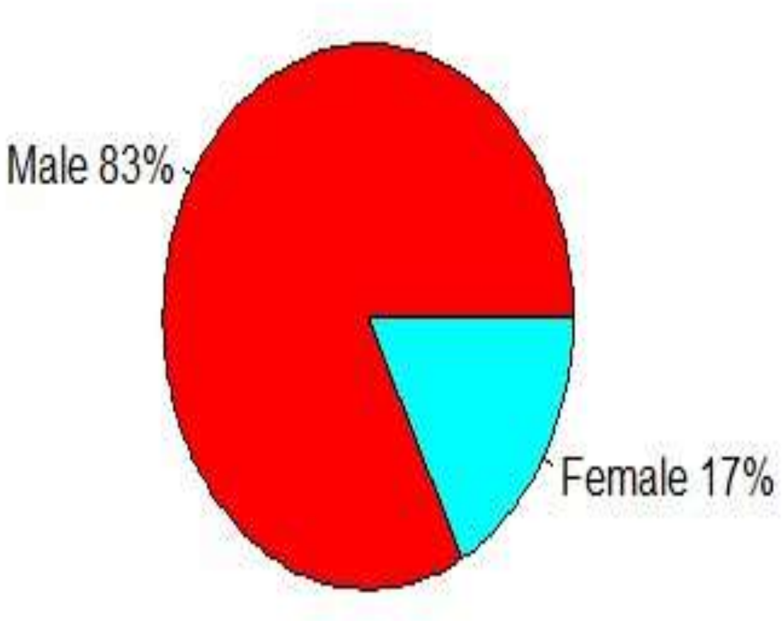
Before we start and answer our problematic questions, it is essential to start with some basic descriptive statistic just to have an idea about our dataset. We are working with relatively large dataset, with 28 features (Gender, Age, HCV, Tumour size, Diabetes…) and 894 individuals. And it would be interesting to analyse briefly each feature.

A. Gender

Despite being relatively big, our sample is mostly formed of Males. In fact 83% of the individuals in question are Male. The other 17% represent the Female (Doc 1). If we want to take at this proportion in a more concrete way, we can take a look at this histogram, which tells us that we have 738 Males and 156 Females (Doc 2).



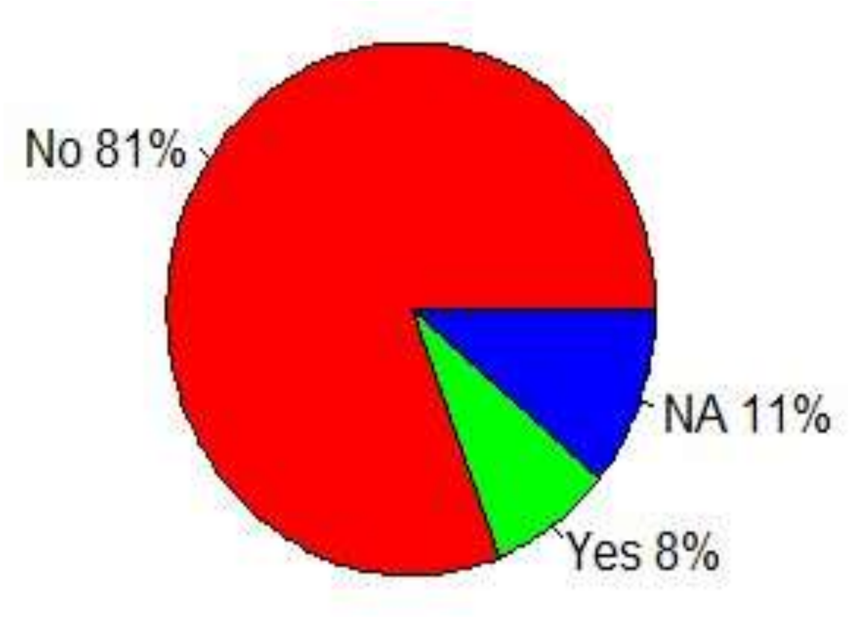
Doc 1: Barplot showing the frequency of male and female



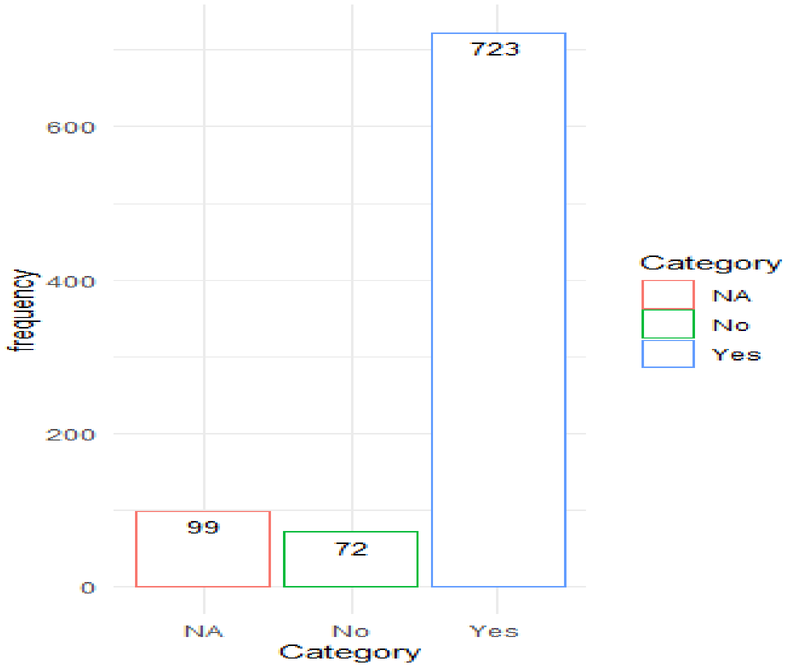
Doc 2: Pie chart representing the percentage of male and female

B. HBV

HBV, is what known as hepatitis B. It is in fact a serious liver infection caused by the hepatitis B virus. It spreads from person to person not through coughing or sneezing. In fact, the most common ways of spreading this virus are sexual contact, sharing needles (drug and so on), from mother to child and so on. In other way the virus is passed from person to person through blood, semen or other body fluids. It would be interesting to see if such disease increases the risk of liver cancer. More on that later on. Now we want to see how many person in our sample have HBV. Only 8% of our sample contain (72 individuals) have HBV. 81% do not have such disease (723 individuals). Finally, we have missing information on 11% of or sample concerning the HBV.(Docs 3,4)



Doc 3: Pie chart representing the percentage of HBV.

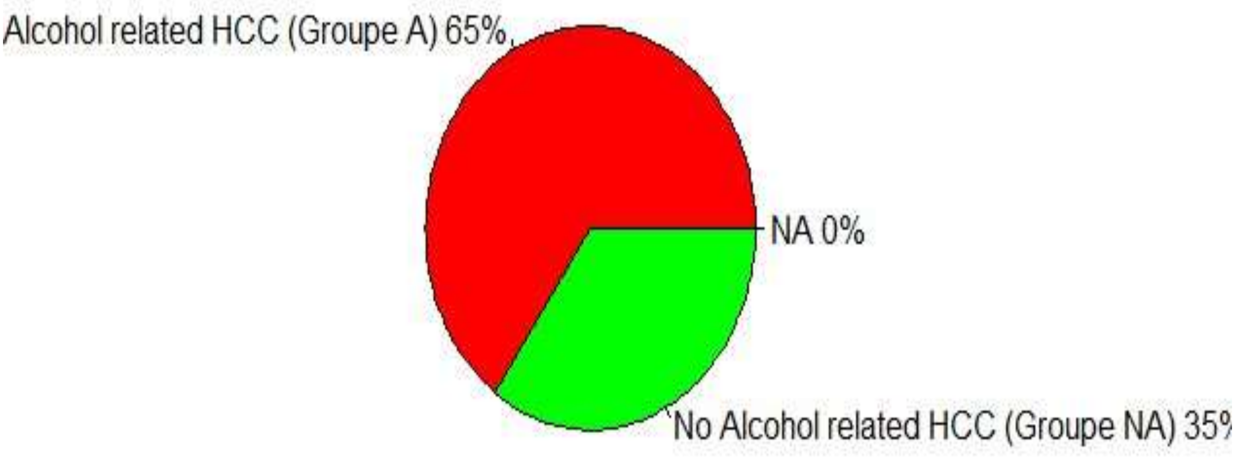


Doc 4: Barplot showing the frequency of HBV

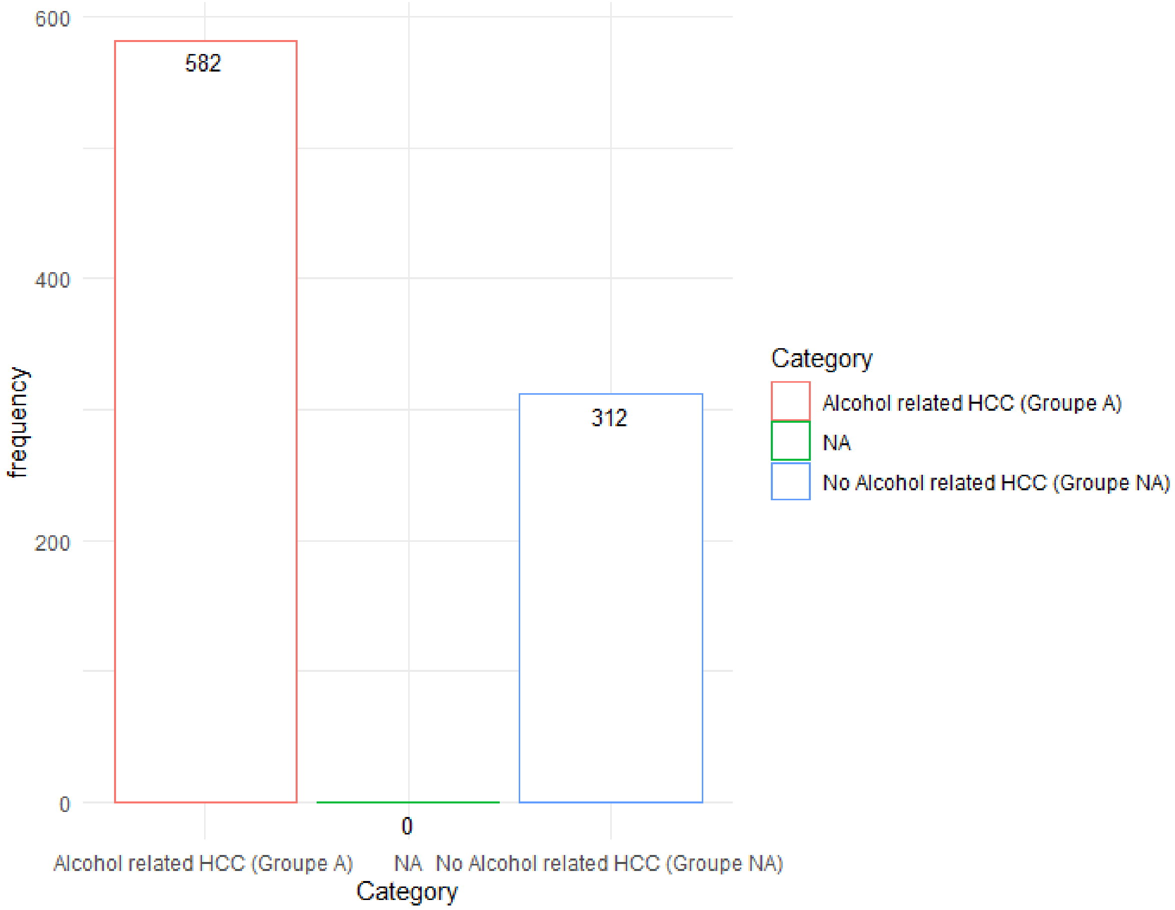
C. Alcohol

Whether or not people drink alcohol could increase the risk of developing Liver Cancer. Here, we are interested in the proportion. Group A, which refers to the people that drinks alcohol is fairly high with 65% (582).

Group NA refers to the people who do not drink alcohol related to HCC (hepatocellular carcinoma), and represents only 35% (312). (Docs 5,6)



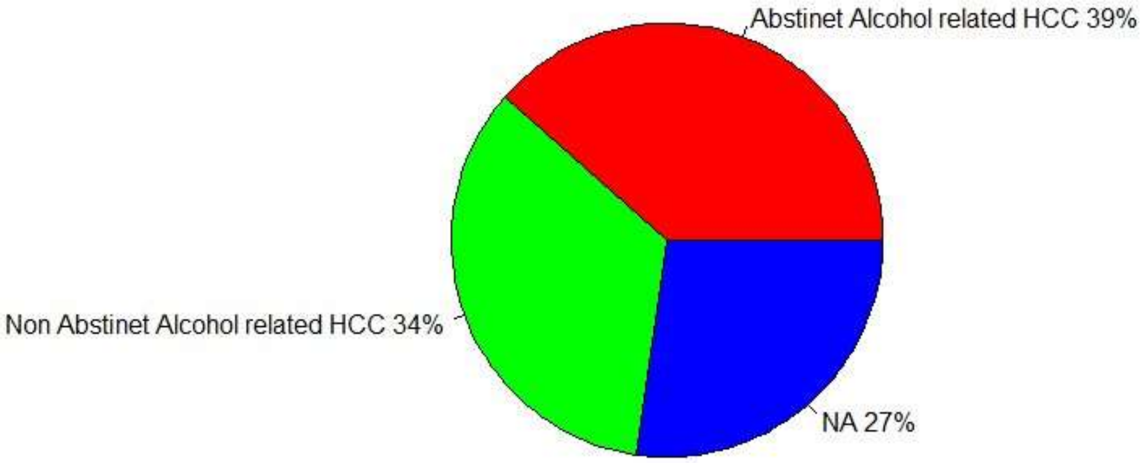
Doc 5: Pie chart representing the percentage of Alcoholic patient.



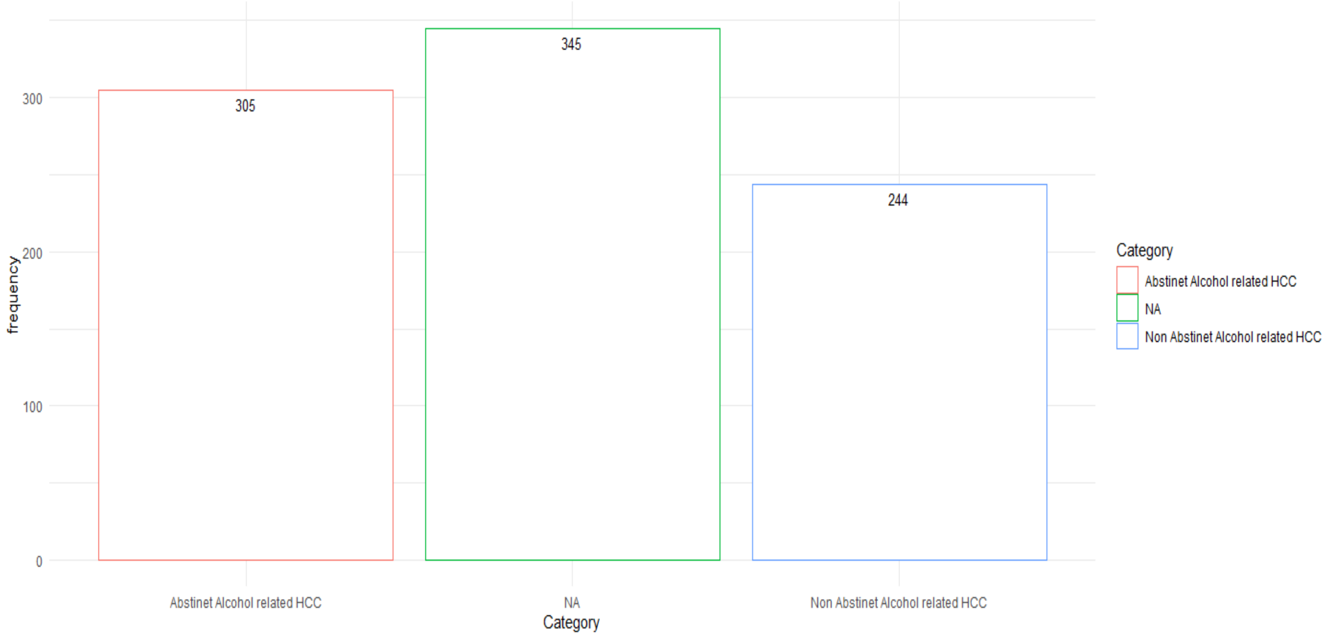
Doc 6: Barplot showing the frequency of people who drink alcohol

D. Abstinence Alcohol

Here we are checking if individuals have stopped drinking alcohol or not. And so we have two categories. The first one is called non abstinent from alcohol related to HCC, which represents 34% (). The other one is called abstinent from alcohol related to HCC and form 39% (). Finally we have 27% of missing values in this section. (Docs 7,8)



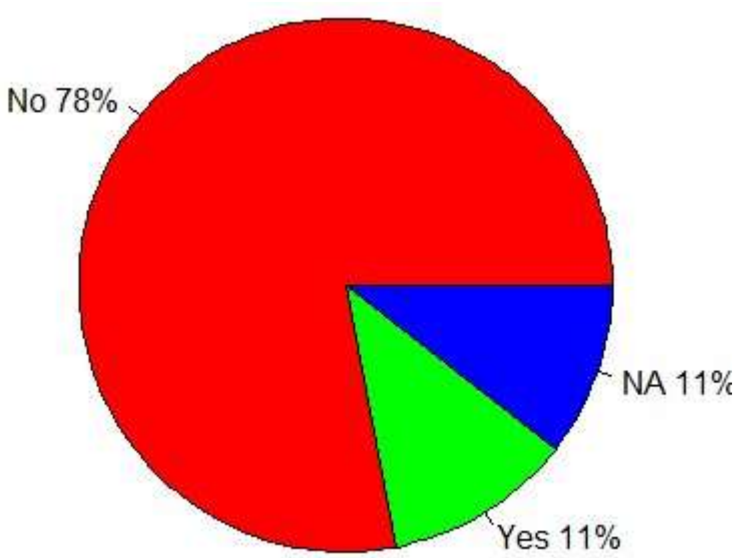
Doc 7: Pie chart representing the percentage of Alcohol abstinence.



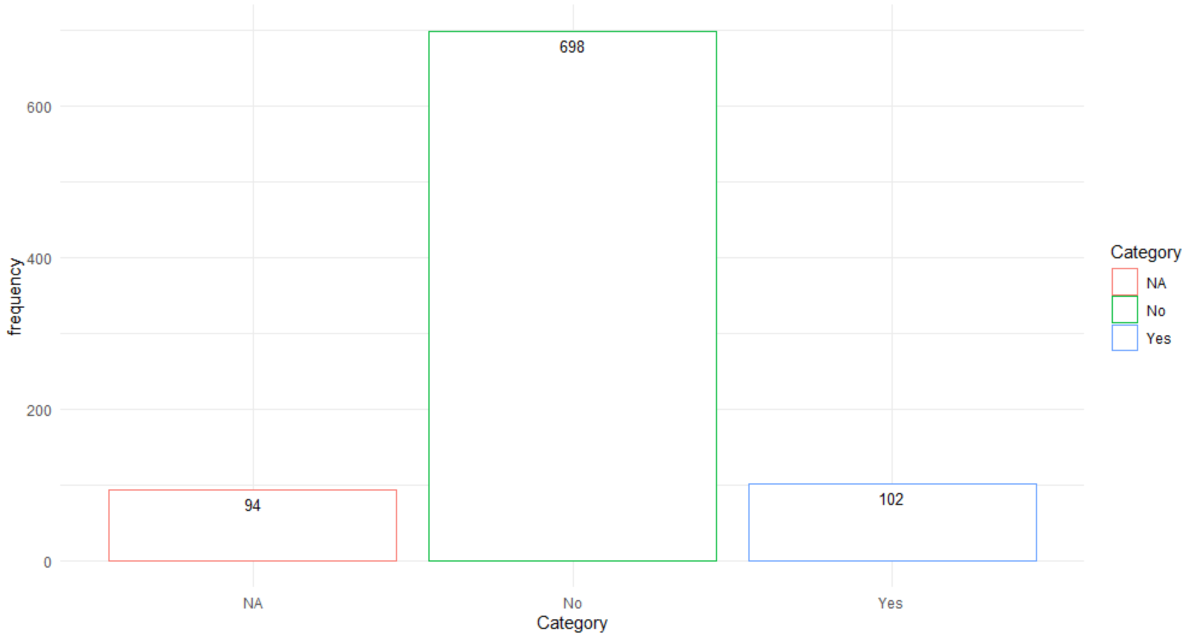
Doc 8: Barplot showing the frequency of the abstinent alcohol variable.

E. HCV

HCV, is known as hepatitis C. It is in fact a serious liver infection caused by the hepatitis C virus which can be responsible of both acute and chronic hepatitis, ranging in severity from a mild illness lasting a few weeks to a serious, lifelong illness. It spreads just like HCV does. Now we want to see how many person in our sample have HCV. Only 11% of our sample contain (102 individuals) have HBV. 78% do not have such disease (698 individuals). Finally, we have missing information on 11% of or sample concerning the HBV. (Docs 9,10)



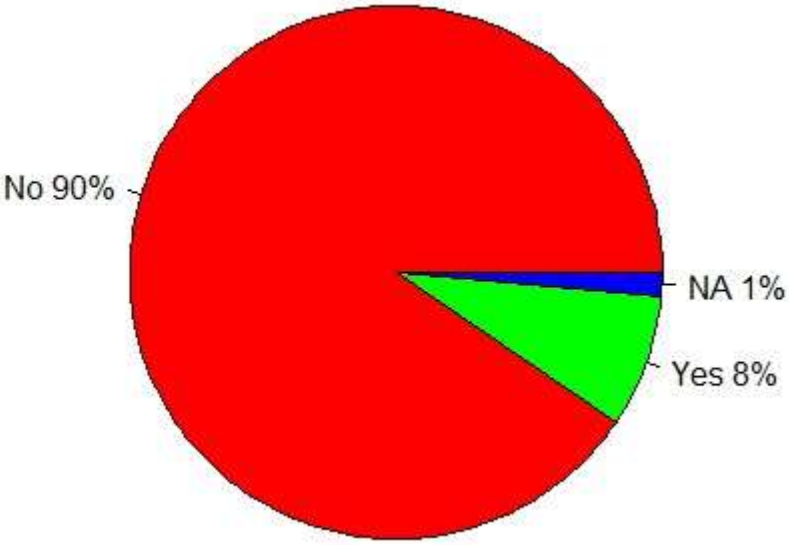
Doc 9: Pie chart representing the percentage of HCV.



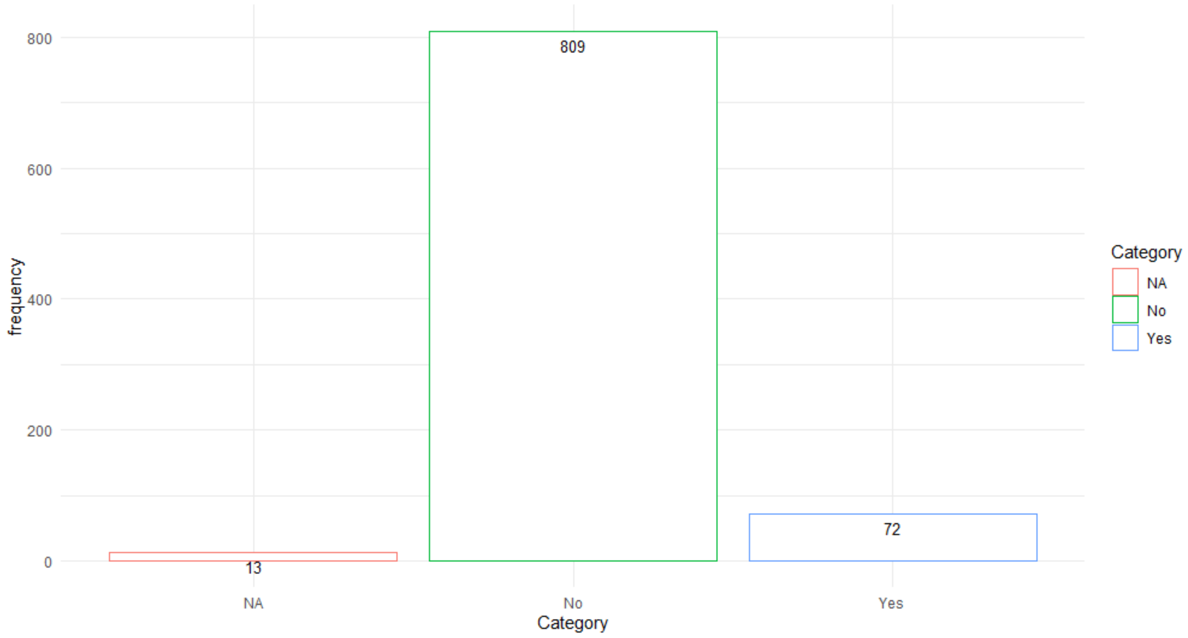
Doc 10: Barplot representing the frequency of HCV.

F. Other

In this section, we are checking diseases that affect the liver other than HBV and HCV. An example of such disease would HAV (hepatitis A). We found out that 90% do not have such diseases (). 8% do have other diseases related to the liver. Finally we have very few missing data (1%). (Doc11,12)



Doc 11: Pie chart representing the percentage of other disease

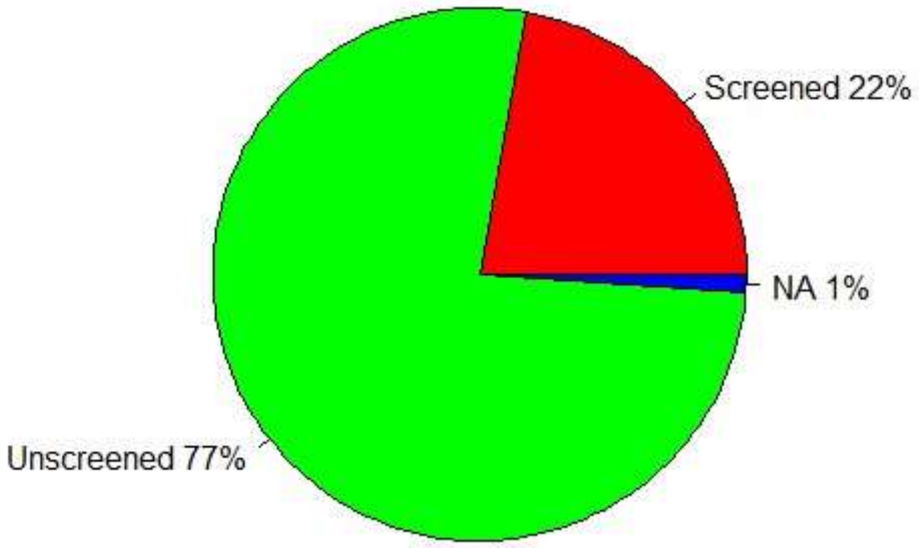


Doc 12: Barplot representing the frequency of other disease

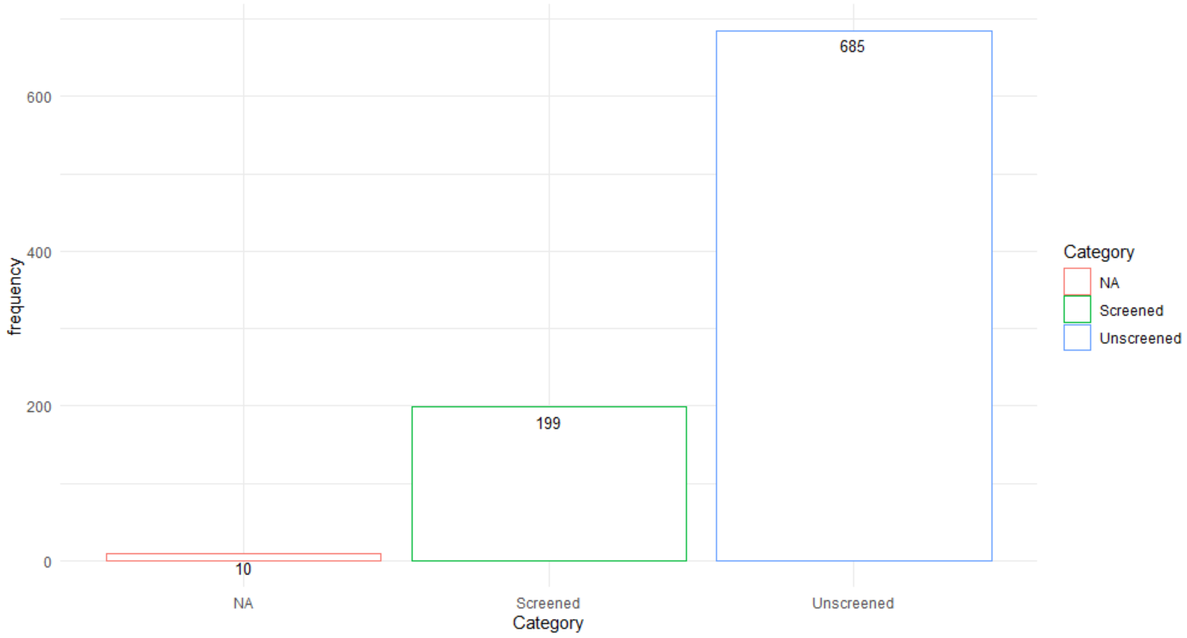
G. Screening

Screening, in medicine, is a strategy used to look for as-yet-unrecognised conditions or risk markers in individuals without signs or symptoms. The defining features of screening programmes are that the people tested do not have signs or symptoms and the implied promise is future risk reduction from an undesirable disease outcome. As such, screening tests are somewhat unusual in that they are offered to and performed on persons apparently in good health. This kind of test is used very often to detect cancer before it starts developing and entering the later stages. In fact, patient do not show any signs or symptoms immediately thus the importance of screening. In our sample, only 22% (199) have done a screening test. 77% are unscreened (685).

Finally we are dealing with only 1% of missing data.(Docs 13,14)



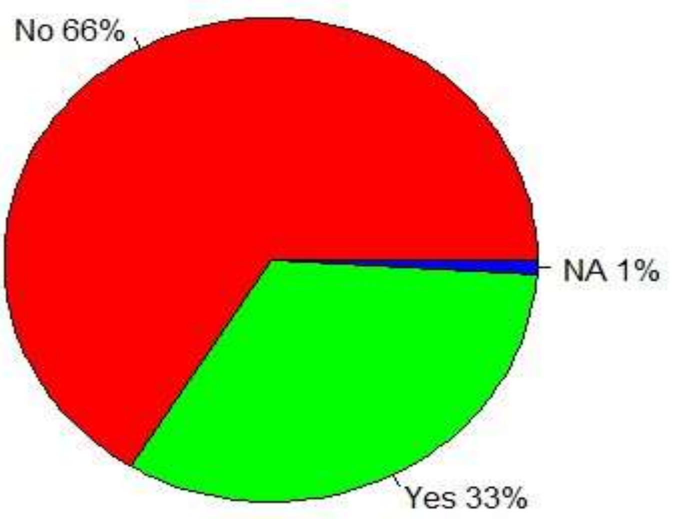
Doc 13: Pie chart representing the percentage of Screened patient



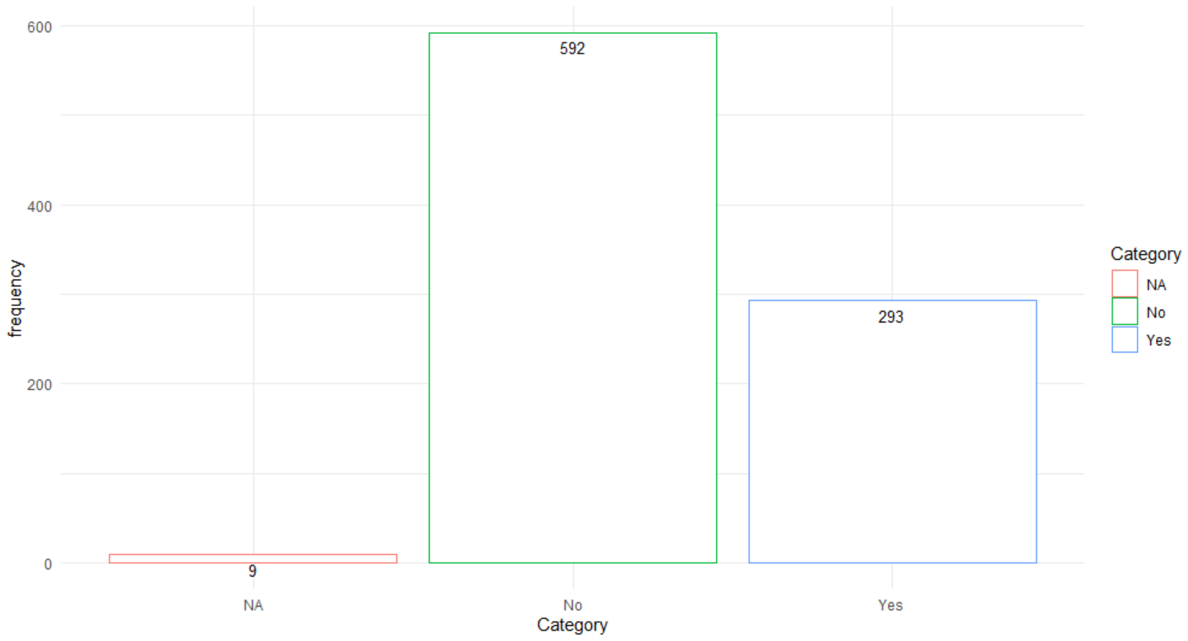
Doc 14: Barplot representing the percentage of Screened patient

H. Diabetes

Diabetes is a metabolic disorder that causes higher than normal blood sugar levels. Diabetes occurs when your body cannot make or effectively use its own insulin, a hormone made by special cells in the pancreas called islets. Insulin serves as a “key” to open your cells, to allow the sugar (glucose) from the food you eat to enter. Then, your body uses that glucose for energy. Let’s check what the percentage of diabetes in our sample is. 33% of our sample suffer from diabetes (293). 66% do not have diabetes whatsoever.(592) (Docs 15,6)



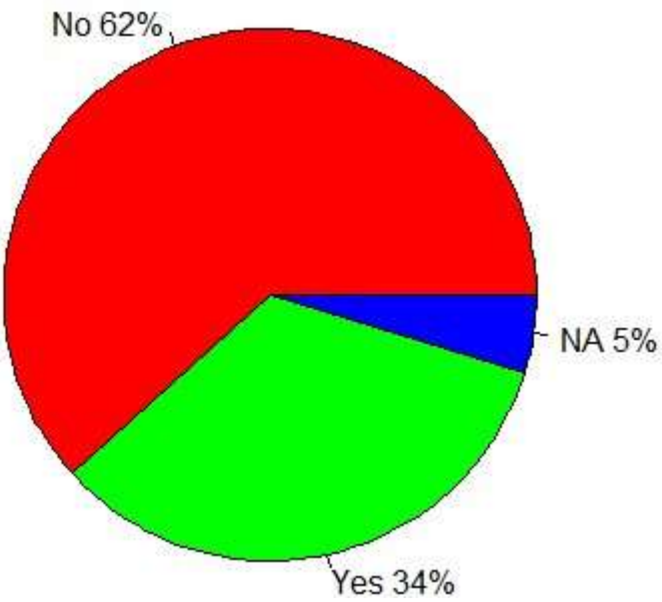
Doc 15: Pie chart representing the percentage of patient with Diabetes



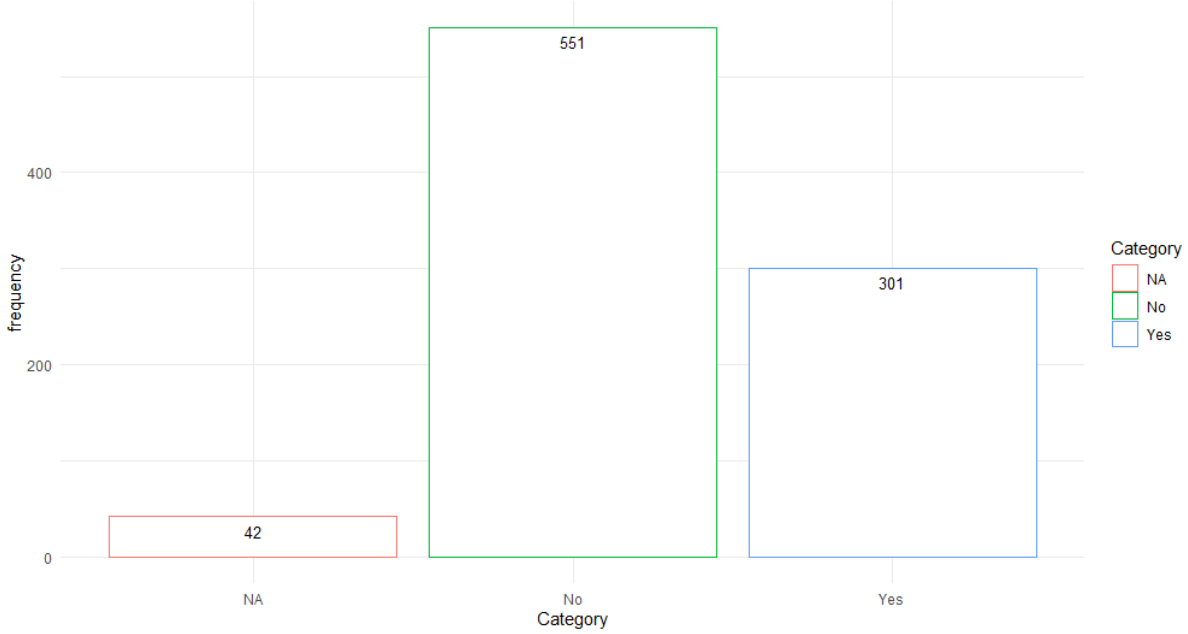
Doc 16: Histogram representing the percentage of Screened patient

I. Smokers

In our sample, people do not seem to smoke that much. In fact, 34% are smokers (), and 62% do not smoke (). It is also important to mention that we are dealing with 5% missing data. (Docs 17,18)



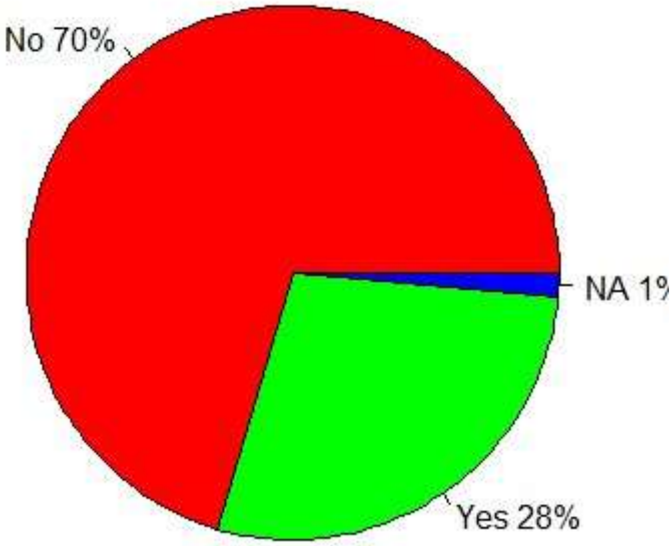
Doc 17: Pie Chart representing the percentage of smokers



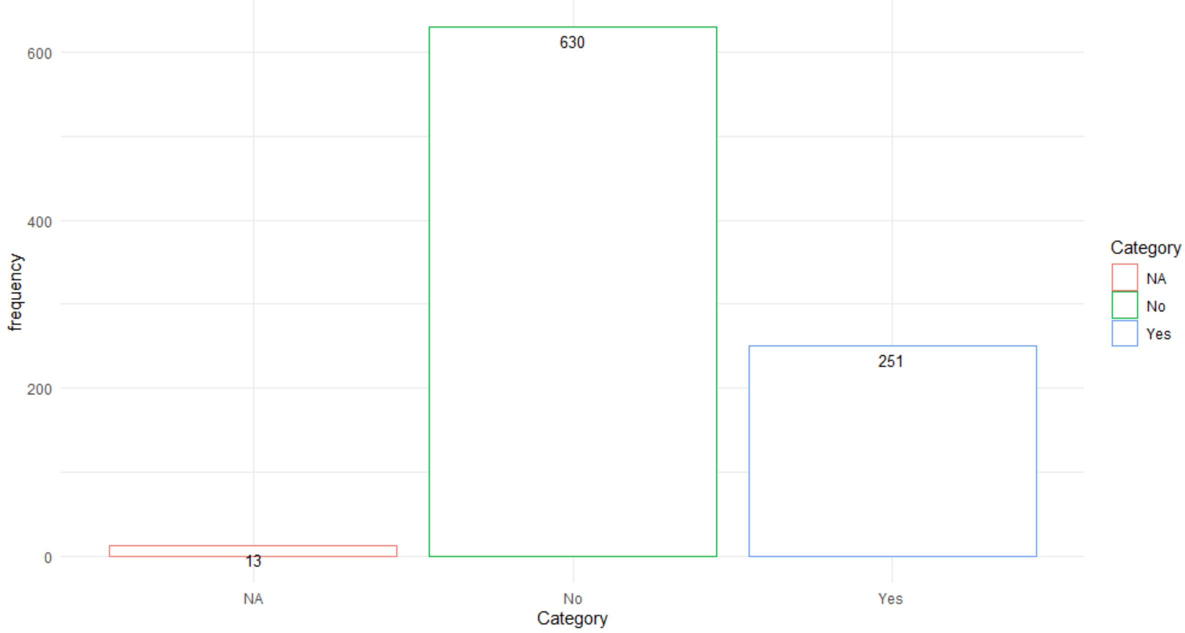
Doc 18: Barplot representing the percentage of Screened patient

J. Thrombosis

Thrombosis is the formation of a blood clot, known as the thrombus, within a blood vessel. It prevents blood from flowing normally through the circulatory system. Only 28% of our sample have thrombosis (251). 70% do not present this issue (630). It is also worth mentioning that we are dealing with 1% missing data. (Docs 19 20)



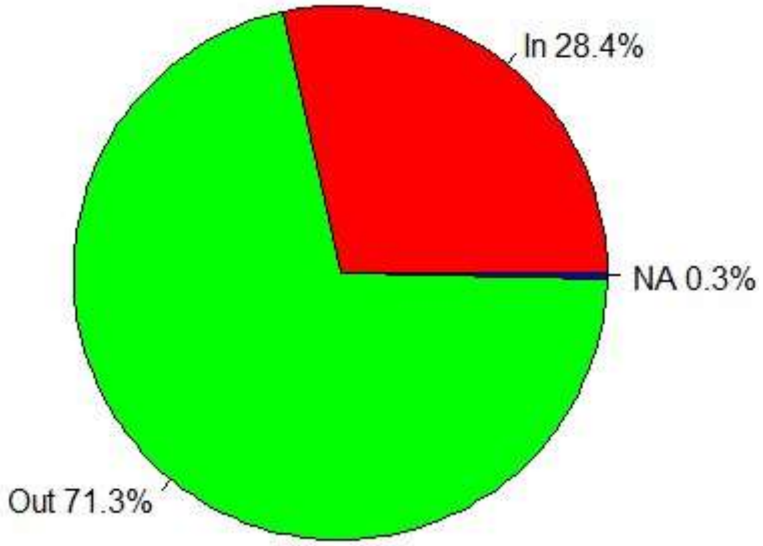
Doc 19: Pie Chart representing the percentage of patient with thrombosis



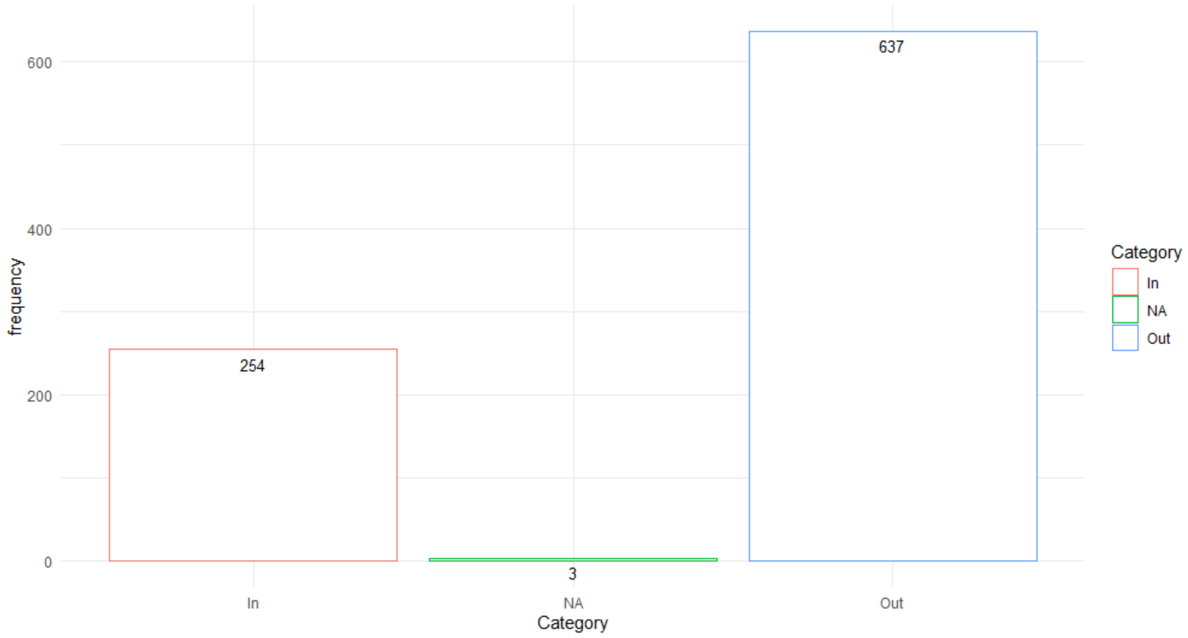
Doc 20: Barplot representing the frequency of patient with thrombosis

K. Criteria

Criteria is just a categorical variable. We are saying that if the tumour size of a patient is greater than 5cm, the criteria is “in”. If it is less than 5cm the criteria is “out”. As we can see, 28.4% (254) have a tumour size greater than 5 and 71.3% (637) have a criteria less than 5. Just for the sake of mentioning it, we have only 0.3% of data concerning this feature. (Docs 21,22)



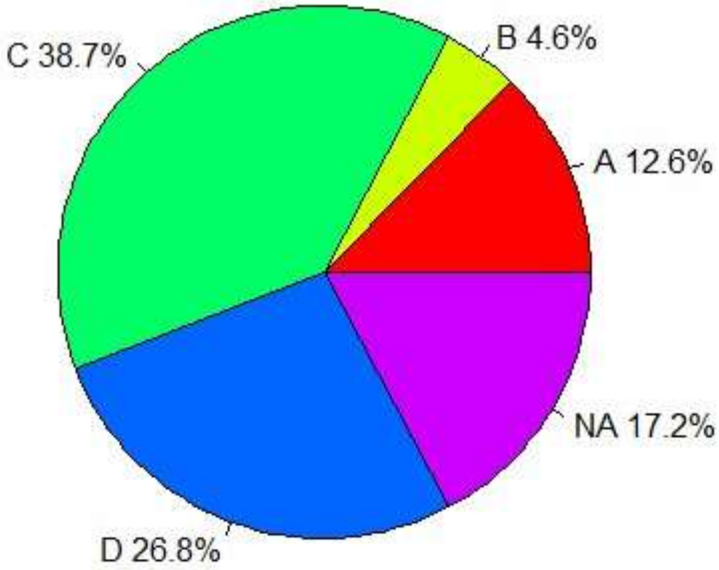
Doc 21: Pie Chart representing the percentage of patient Criteria variable



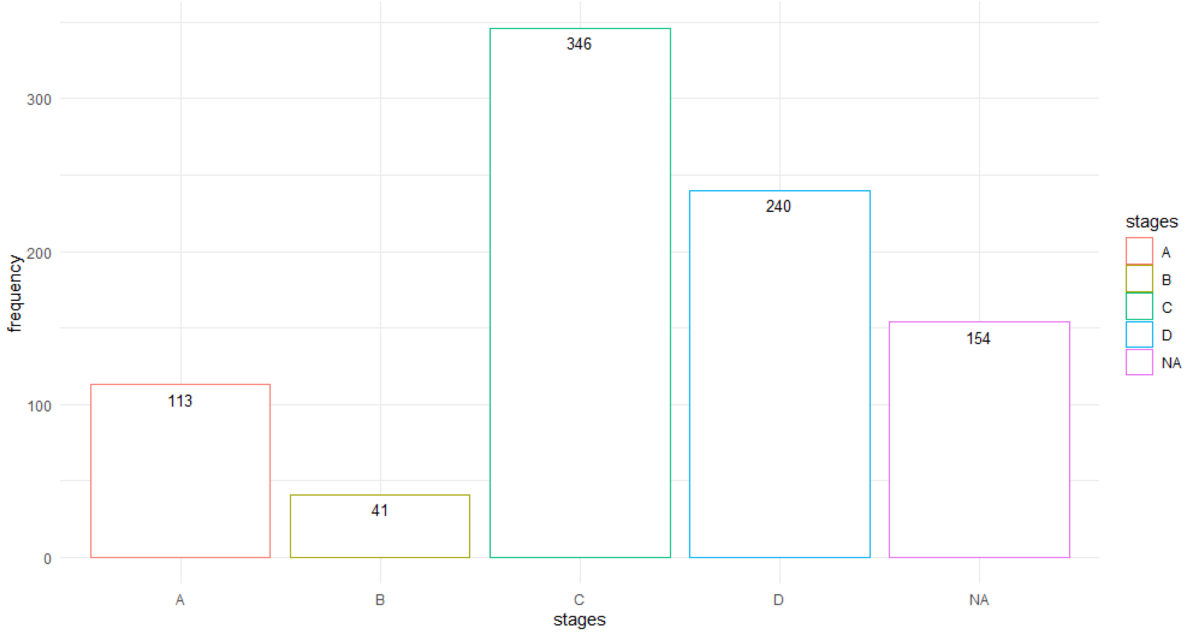
Doc 22: Barplot representing the frequency of the criteria level.

L. Cancer stages

In order to have a better understanding of Cancer, the science community broke it down into several stages. Stage A means the cancer is small and only in one area. This is also called early-stage cancer. Stage B and C mean the cancer is larger and has grown into nearby tissues or lymph nodes. Stage D means the cancer has spread to other parts of your body. In our sample, we can see that 12.6% (113) have cancer stage A, 4.6% (41) have cancer stage B, 38.7% (346) have cancer stage C, and 26.8% (240) have cancer stage D. Also, we do have a significant amount of missing values (17.2%) (Doc 23,24)



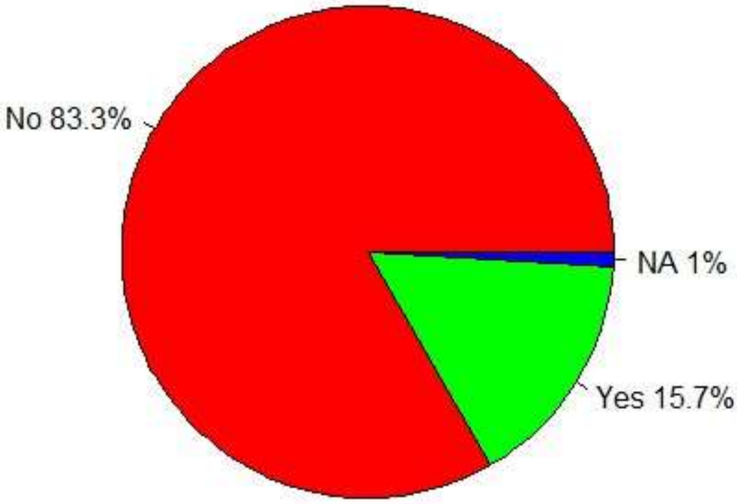
Doc 24: Pie Chart representing the percentage of the different cancer stages



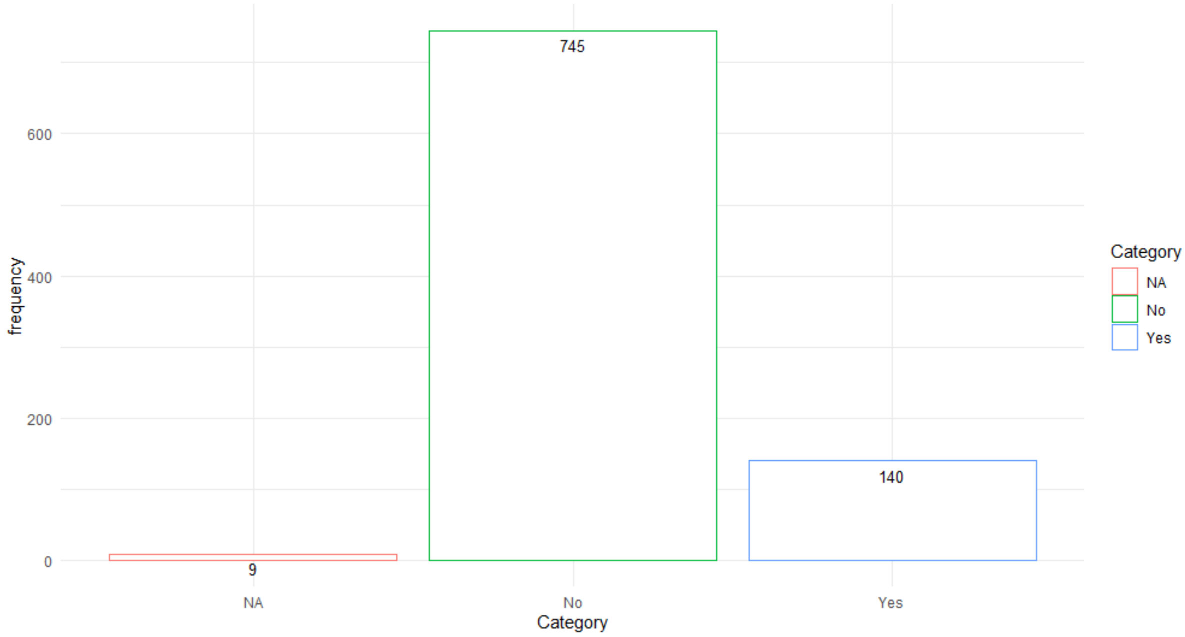
Doc 24: Barplot representing the frequency of the different cancer stages

M. Diffuse cancer

Some patient may undergo an operation in which they try and completely remove the tumour. We want to look at the proportion of patients that have diffused cancer. We can see that only 15.7% (140) have diffused the tumour. 83.3% (745) did not do such operation. Finally, missing values represent only 1%. (Docs 25,26)



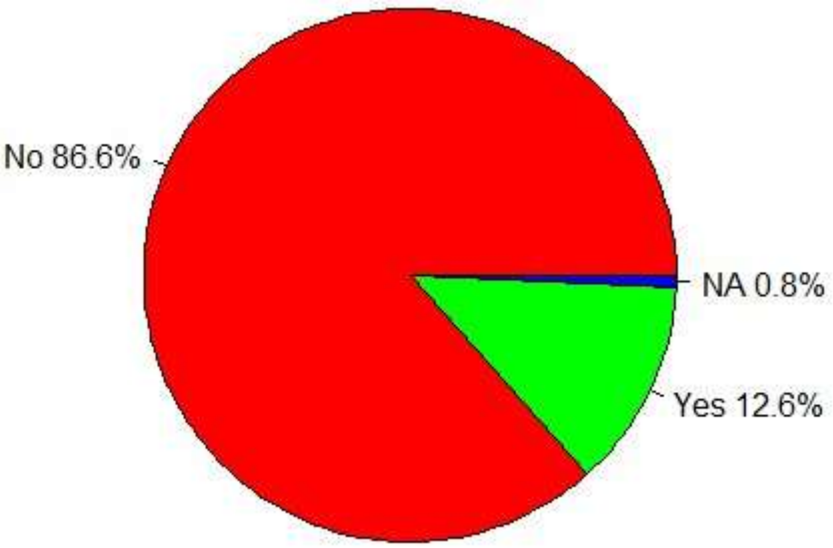
Doc 25: Pie Chart representing the percentage of patient who diffuse their cancer



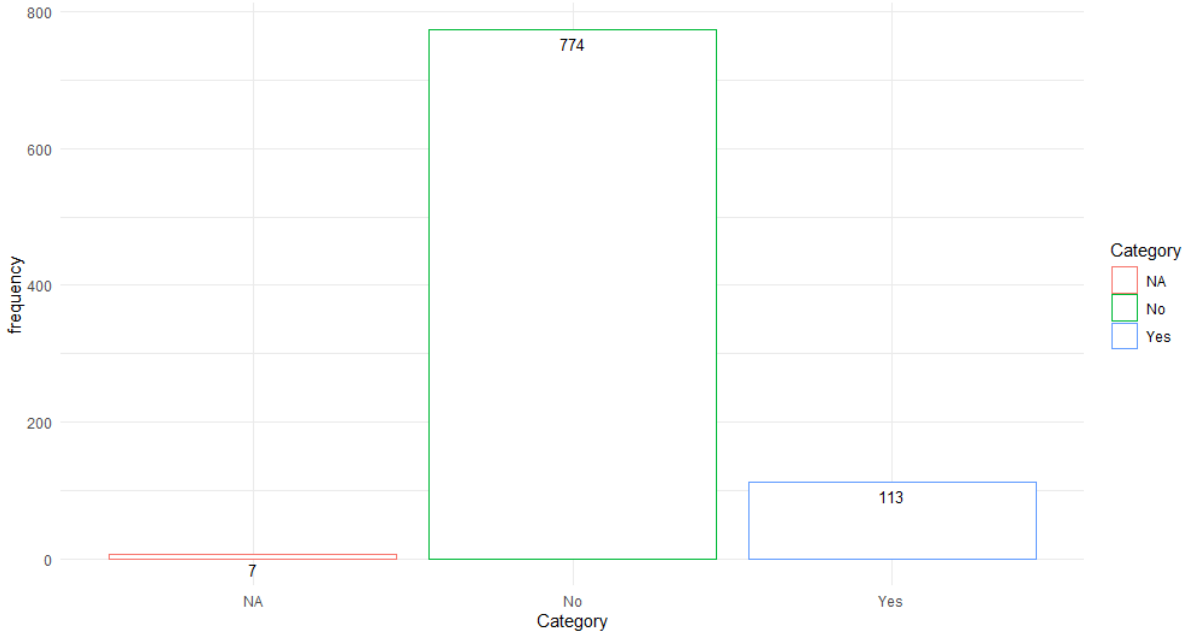
Doc 26: Barplot representing the frequency of patient who diffused their cancer

N. Metastatic Cancer

The main reason why cancer is serious is because of its ability to spread to other part of the body. When Cancer start spreading, that is when we call it metastatic cancer. Once again. We are interested in how many patient in our sample has a metastatic cancer. It turned out that 12.6% have it (113) and 86.6% do not have it (774). (Docs 27,28)



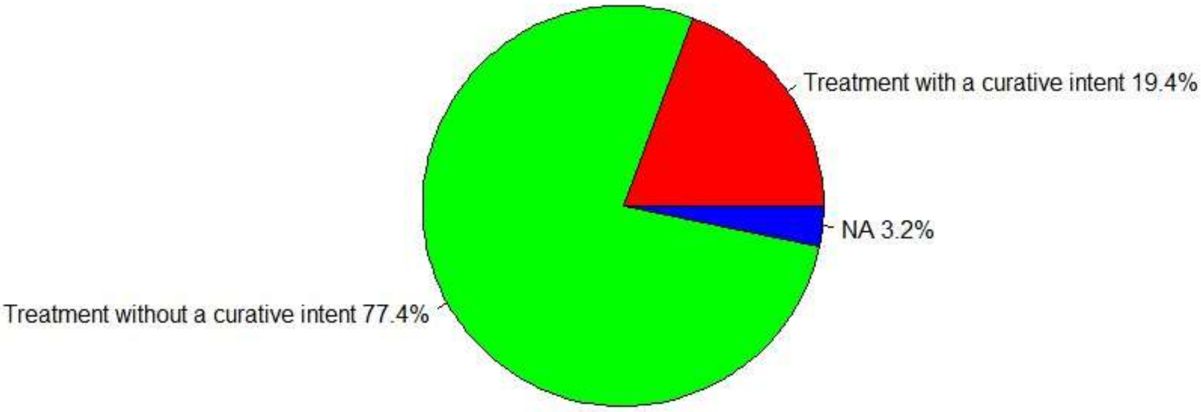
Doc 27: Pie Chart representing the percentage of patient with metastatic cancer



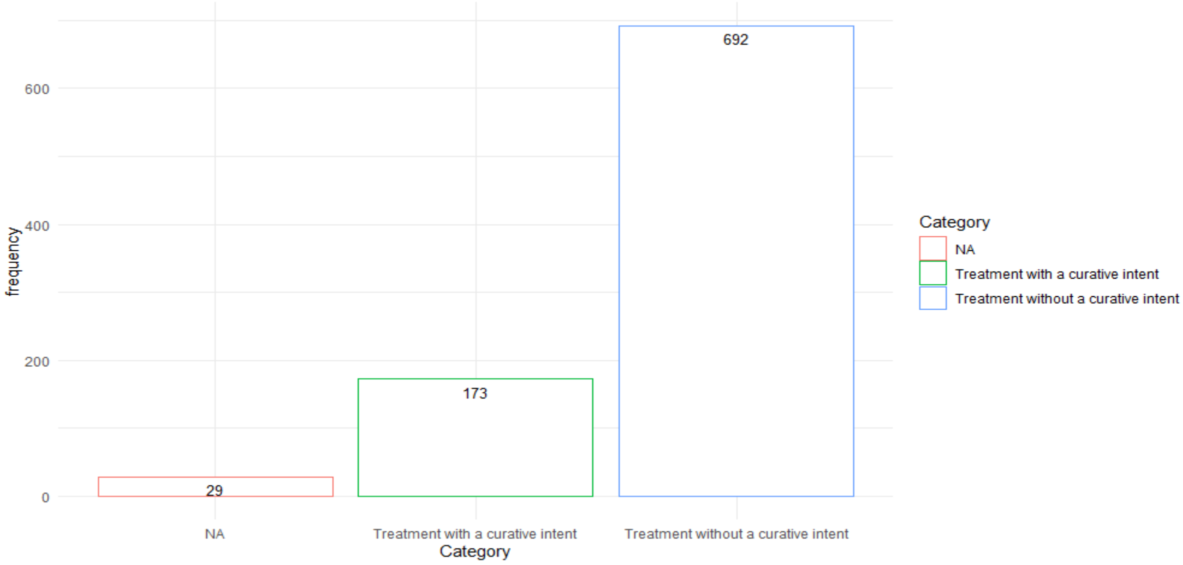
Doc 28: Barplot representing the frequency of patient with metastatic cancer

O. Curative treatment

Curative care refers to health care practices that treat patients with the intent of curing them, not just reducing their pain or stress. An example is chemotherapy, which seeks to cure cancer patients. We can see that 77.4% (692) have done a treatment without curative intent, and only 19.4% (173) have done a treatment with a curative intent. Note that 3.2% are missing values. (Doc 29,30)



Doc 29: Pie Chart representing the percentage of patient with curative treatment

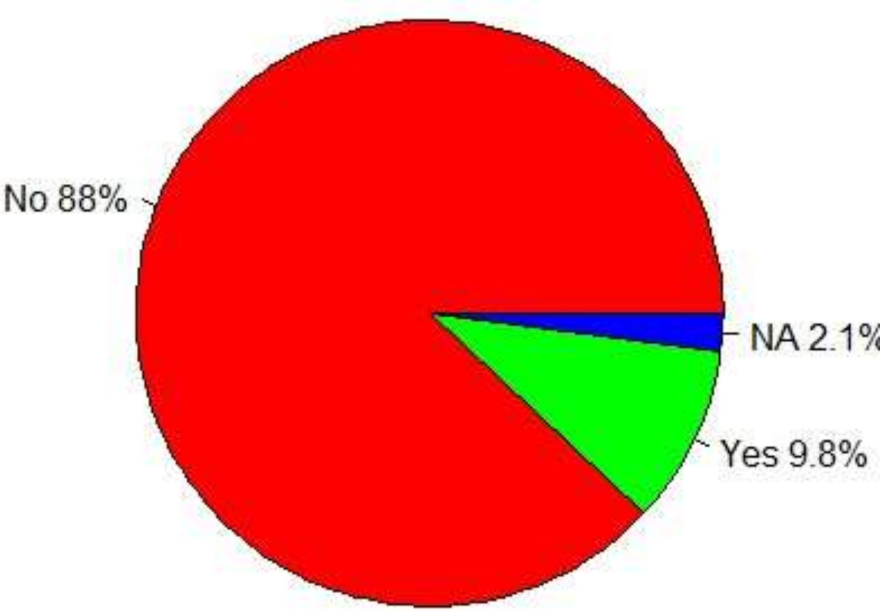


Doc 30: Barplot representing the frequency of patient with curative treatment

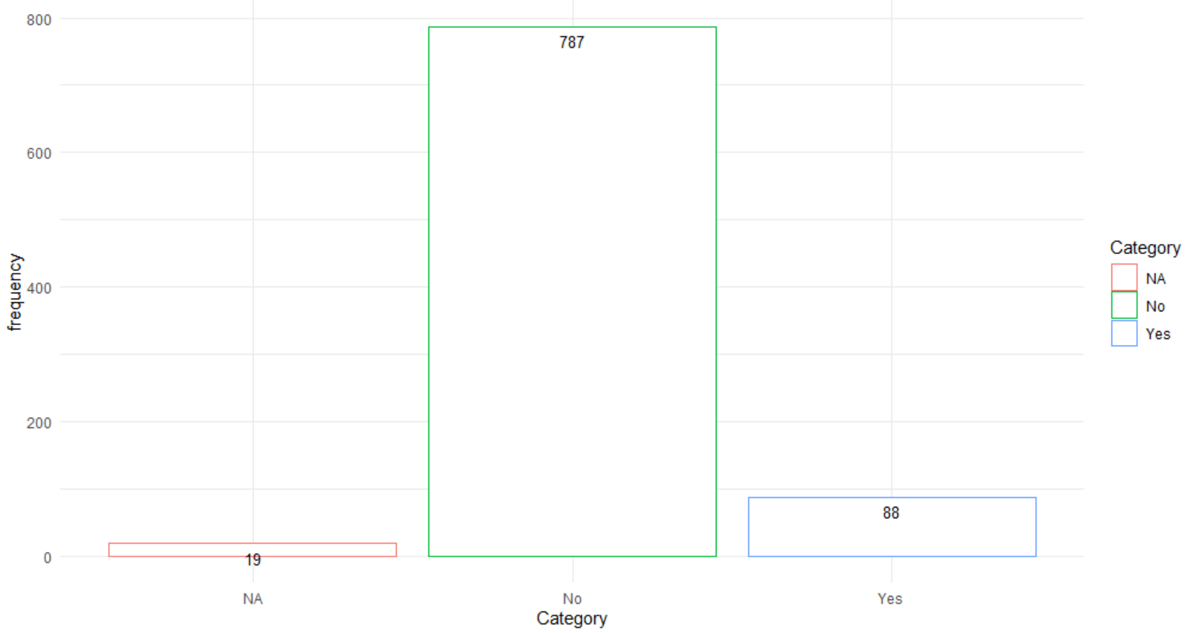
Next, we are going to take a brief overview at each treatment, from 1 to 6.

P. Treatment 1

In the case of treatment 1, only 9.8% (88) have done it. We can also see that 88% have not been treated with this treatment (787) . Finally 2.1% represent the missing values (Docs 31,32)



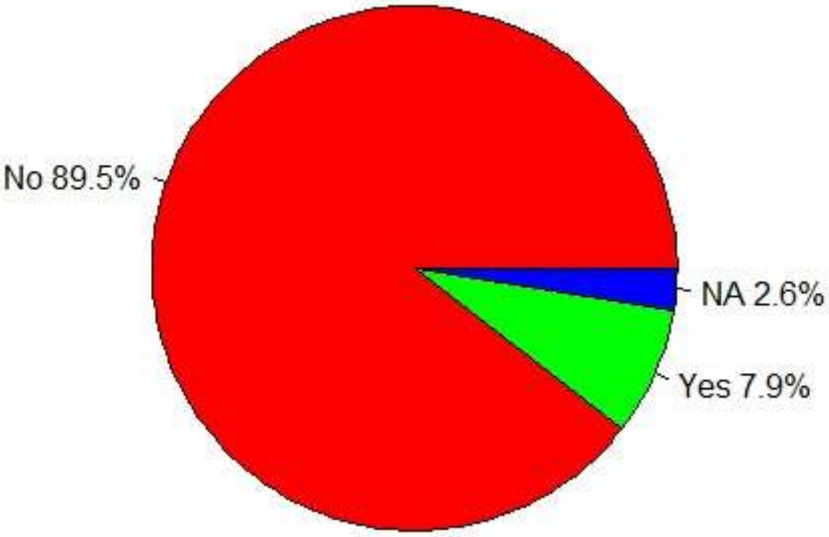
Doc 31: Pie Chart representing the percentage of treatment 1 done



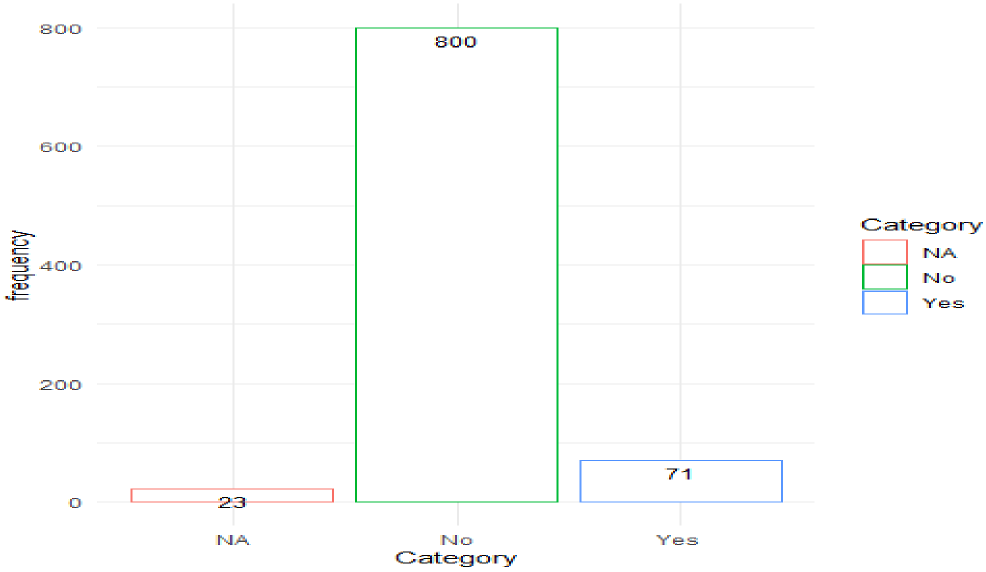
Doc 32: Barplot representing the frequency of patient with treatment 1

Q. Treatment 2

In the case of treatment 1, only 7.9% (71) have done it. We can also see that 89.5% have not been treated with this treatment (800). Finally 2.6% represent the missing values. (Docs 33,34)



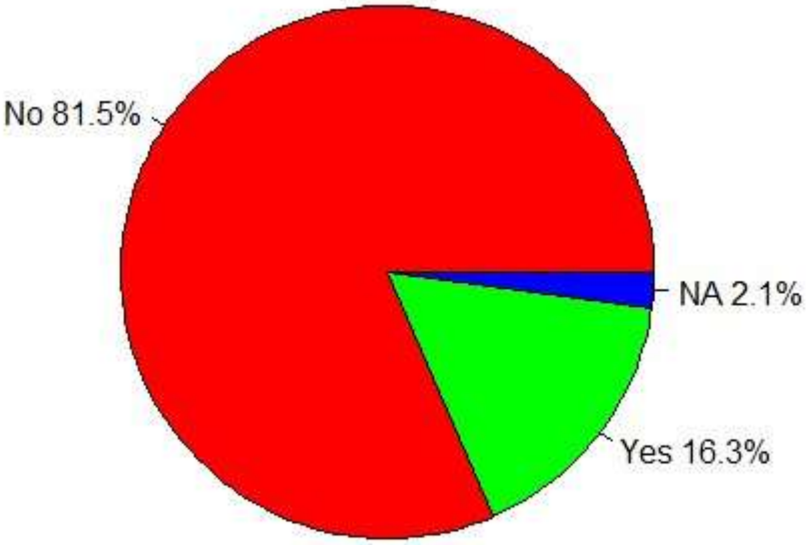
Doc 33: Pie Chart representing the percentage of treatment 2 done



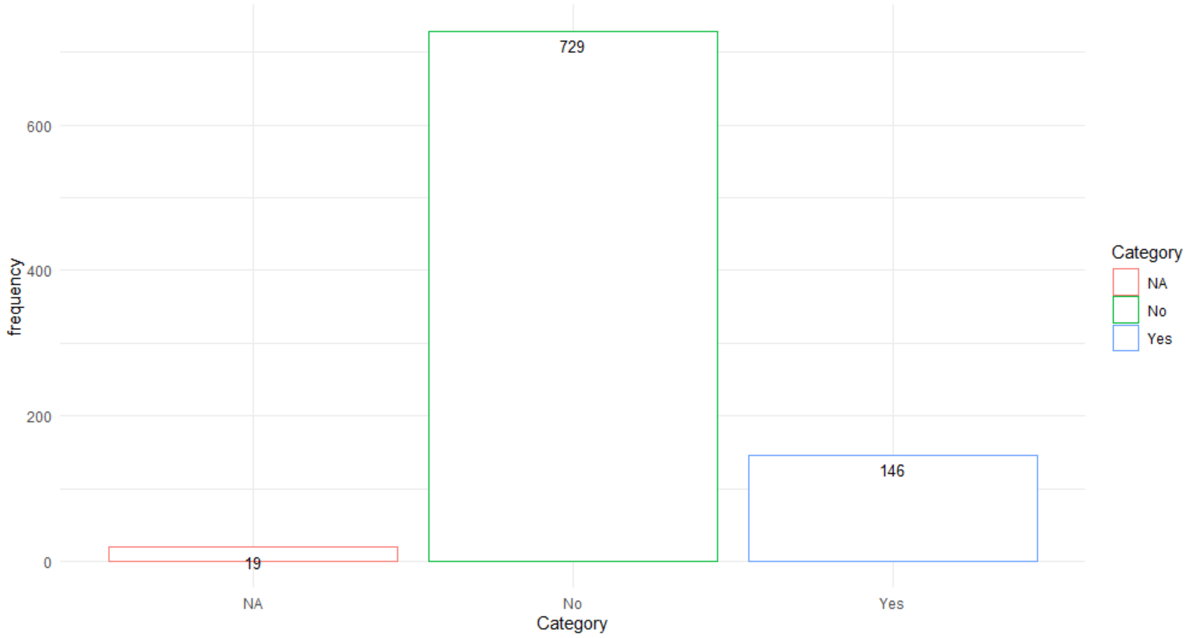
Doc 34: Barplot representing the frequency of patient with treatment 2

R. Treatment 3

In the case of treatment 1, only 16.3% (146) have done it. We can also see that 81.5% have not been treated with this treatment (729) . Finally 2.1% represent the missing values. (Docs 35,36



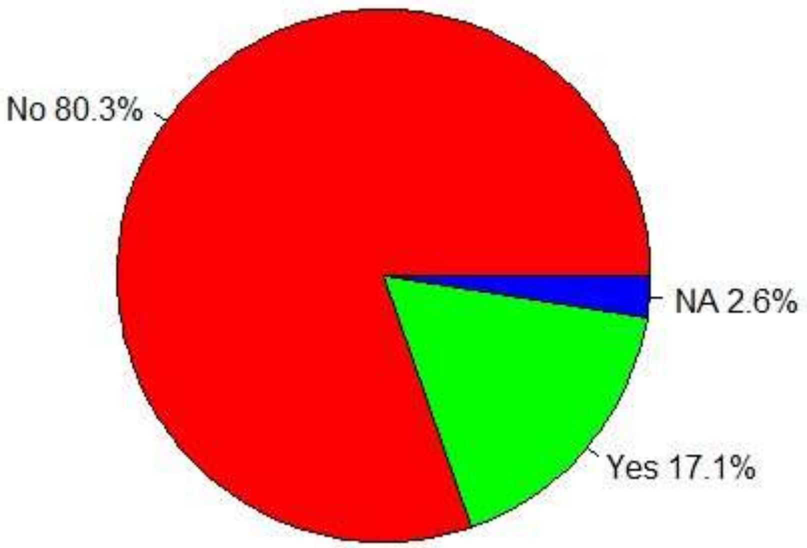
Doc 35: pie chart representing the percentage of patient with treatment 3



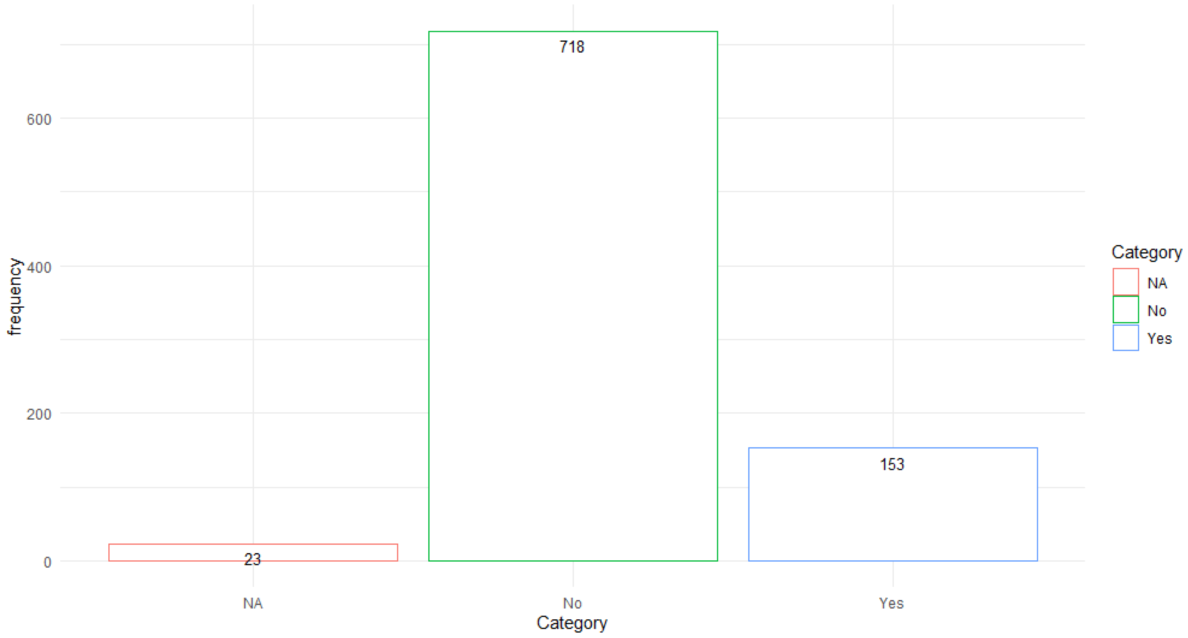
Doc 36: Barplot representing the frequency of patient with treatment 3

S. Treatment 4

In the case of treatment 1, only 17.1% (153) have done it. We can also see that 80.3% have not been treated with this treatment (718) . Finally 2.6% represent the missing values. (Docs 37,38)



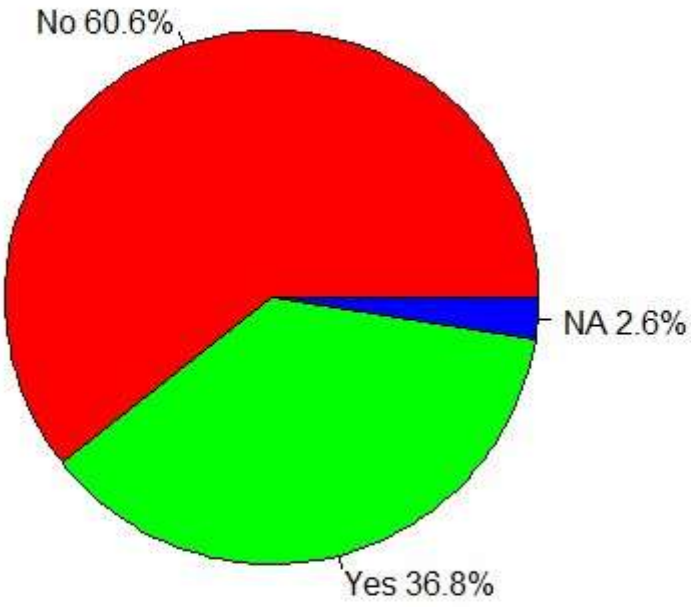
Doc 37: Pie chart representing the percentage of patient with treatment 4



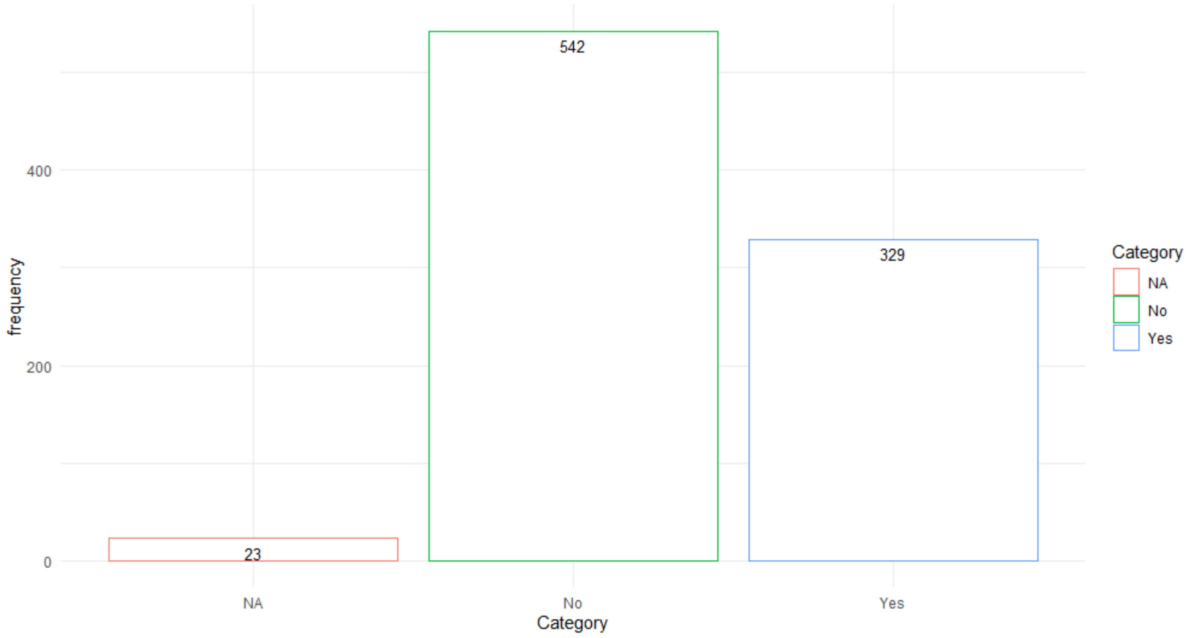
Doc 38: Barplot representing the frequency of patient with treatment 4

T. Treatment 5

In the case of treatment 1, only 36.8% (329) have done it. We can also see that 60.6% have not been treated with this treatment (542) . Finally 2.6% represent the missing values. (Docs 39,40)



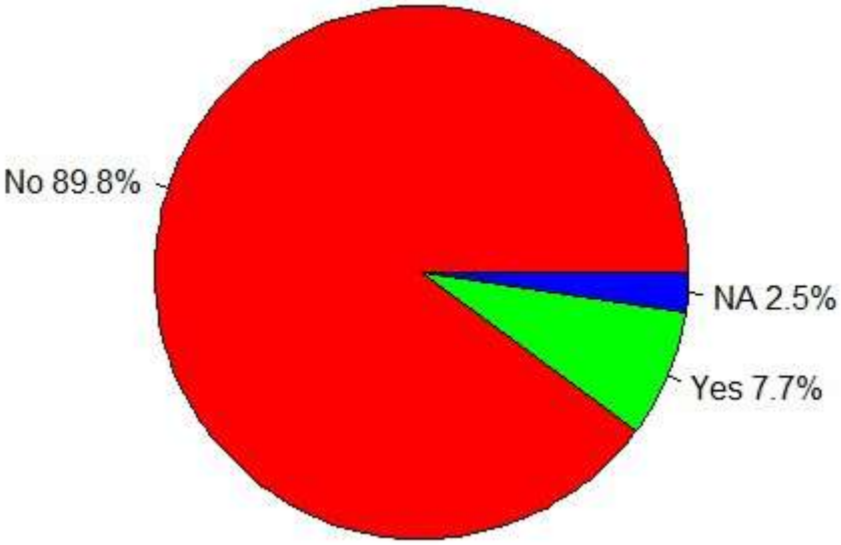
Doc 39: Pie chart representing the percentage of patient with treatment 5



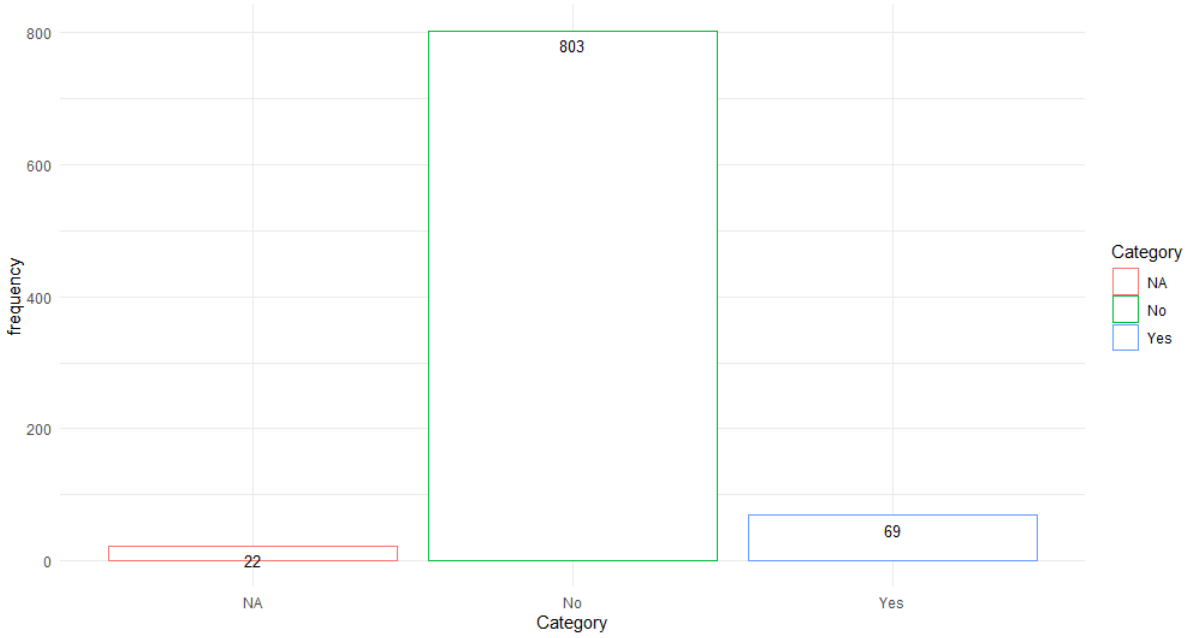
Doc 40: Barplot representing the frequency of patient with treatment 5

U. Treatment 6

In the case of treatment 1, only 7.7% (69) have done it. We can also see that 80.3% have not been treated with this treatment (803) . Finally 2.5% represent the missing values. (Docs 41,42)



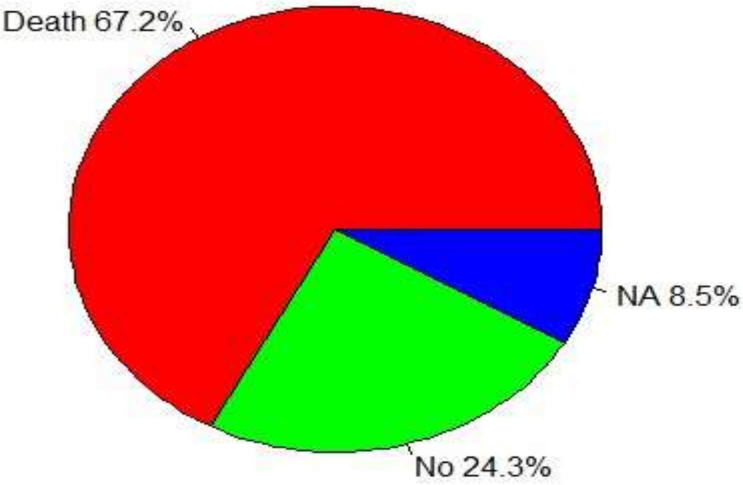
Doc 41: Pie chart representing the percentages of patient with treatment 6



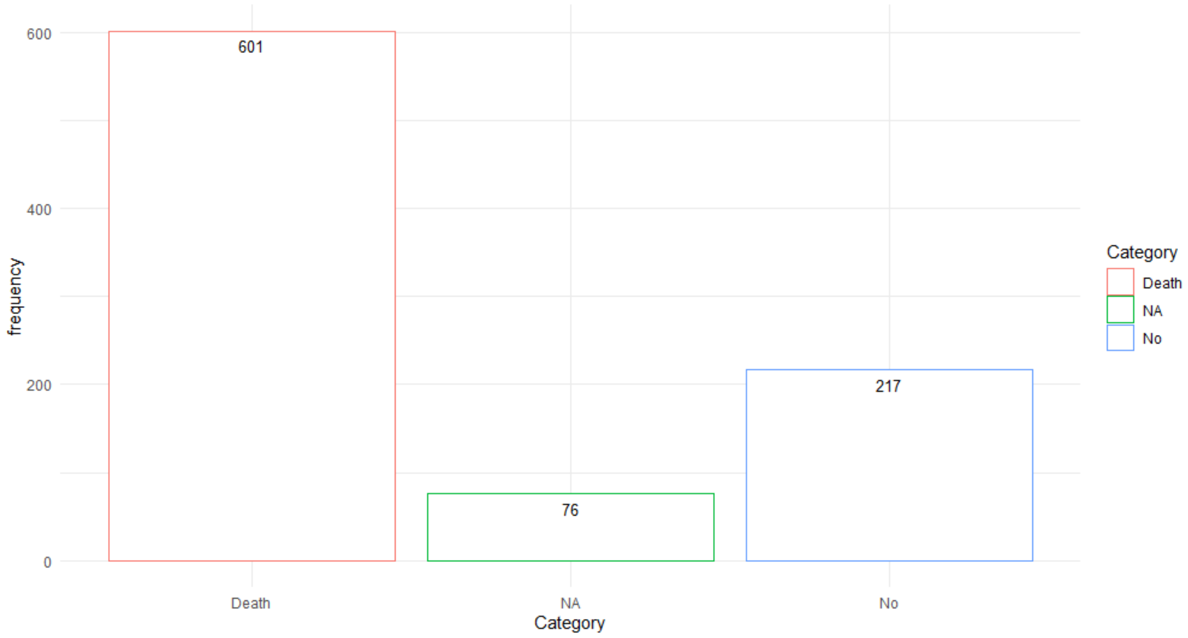
Doc 42: Barplot representing the frequency of patient with treatment 6

V. Death Status

Finally, it would be interesting to see how many people survived, especially in our sample. We found out that 24.3% have survived, whereas 67.2% passed away. Unfortunately, we have a relatively high percentage of missing value, people we do not know what their fate was. (Docs 42,43)



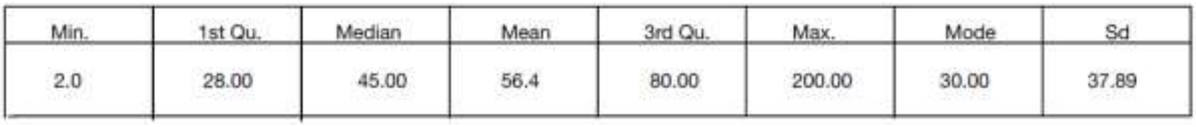
Doc 43: Pie chart representing the percentage of dead patient.



Doc 44: Barplot representing the frequency of dead patient

W. Tumour size

It would be interesting to analyse the tumour size of the liver cancer in this particular sample. Here, we can see a brief summary regarding this variable. (Doc 45)

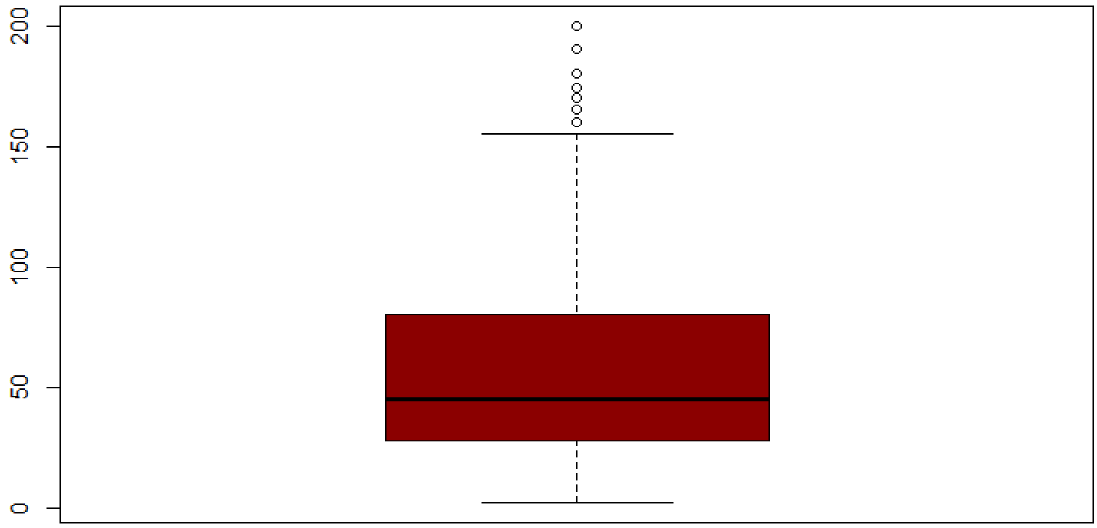


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mode	Sd
2.0	28.00	45.00	56.4	80.00	200.00	30.00	37.89

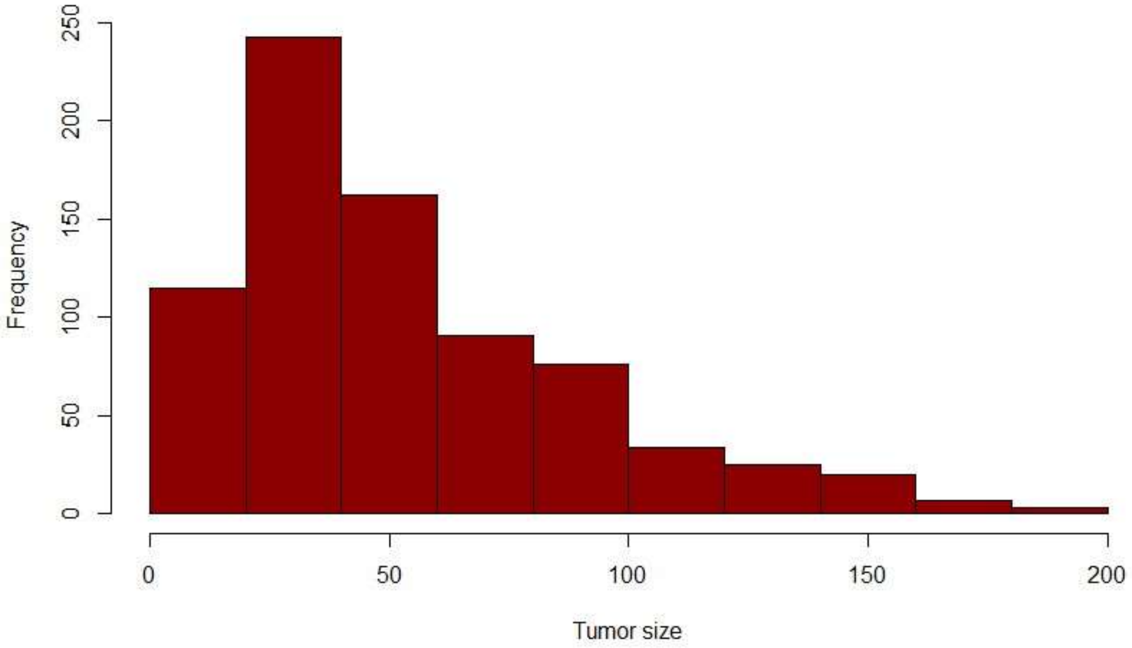
Doc 45: Summary of the Tumour Size variale

The minimum tumour size 2.00 mm. This probably refers to the first stage of cancer (stage A). The maximum tumours size in our sample is 200 mm. This is probably a cancer in its latest stages. The 1st Quantile is 28.00 mm, meaning that 25% of our sample has a tumour size below 28.00 mm and 75% have a tumour size greater than 28.00 mm. The median is 56.40 meaning that half of our sample have a tumour size below 56.40 mm. The other being greater than 50 mm. The 3rd Quantile is 80.00 mm. This means that 75% have a tumour size below 80.00 mm and 25% have a tumour size greater than 80.00 mm. The mean of our sample is 56.40 mm and the standard deviation is just 37.89 mm. Finally the most repeated value is 30.00 which is none other than the mode. Keep in mind, for visualization purposes, you can take a look at this boxplot. It is a good way to check the 1st Quantile, 3rd Quantile and the median. The distribution of tumour size in our sample can also be see in down below.

(Docs 46,47)



Doc 46: Tumour Size Boxplot

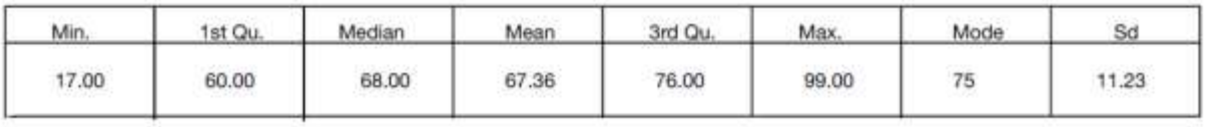


Doc 47: Tumour Histogram

One last thing before we move on to the next variable. It is essential to estimate the real mean of the tumour size in the whole population. For that, we need to compute the confidence interval which is I = [53.73 ; 59.06]. Please note that this is done with alpha=0.05 and so with that in mind, we can conclude that the real mean of liver tumour size is between 53.73 and 59.06, and this is for a 95% confidence level.

X. Age of our samples.

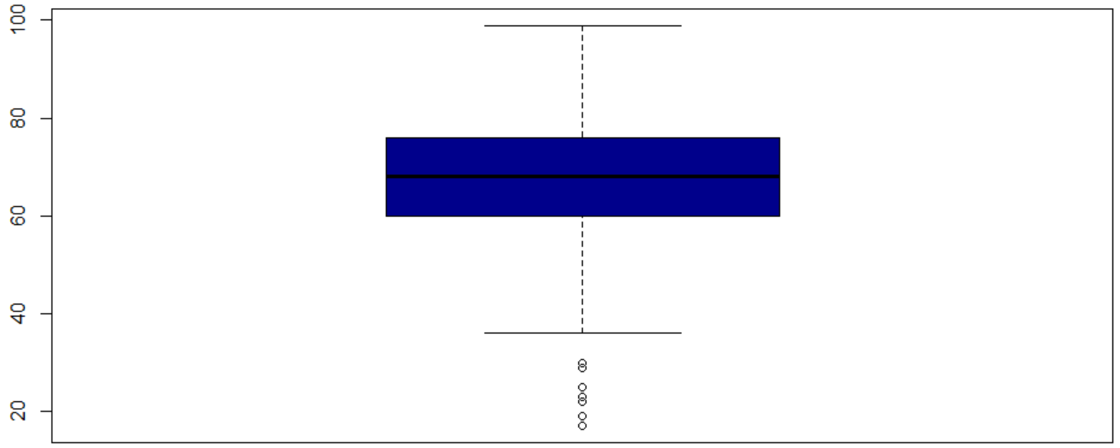
Let's take a look at the age of our samples. Once again, this is a brief summary regarding this variable. (Doc 48)



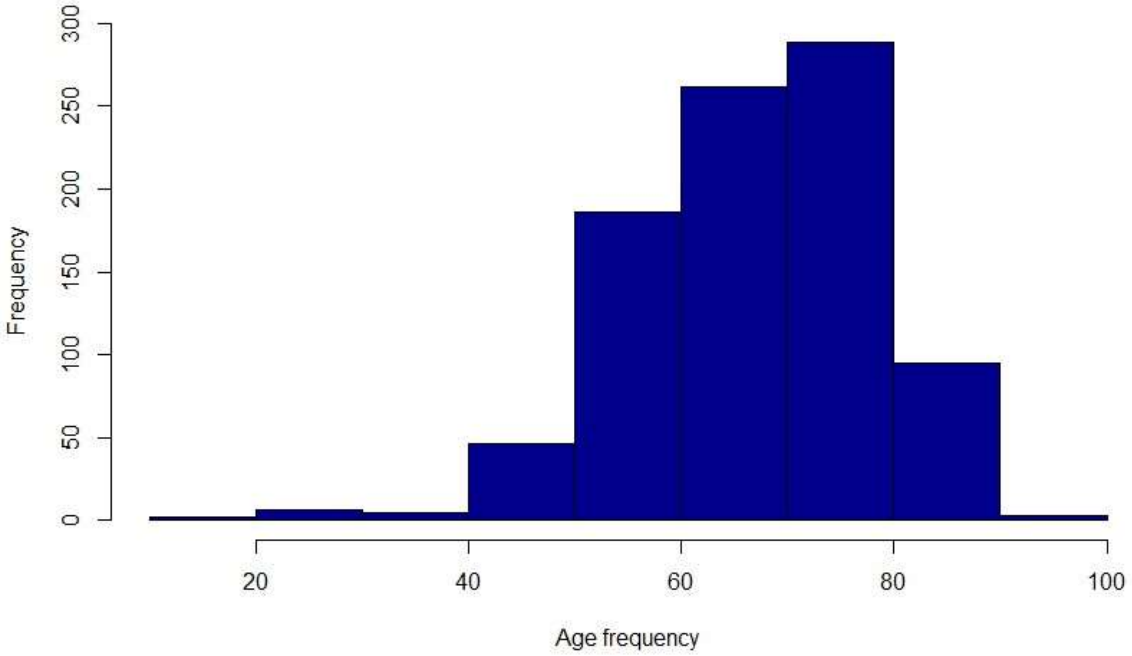
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mode	Sd
17.00	60.00	68.00	67.36	76.00	99.00	75	11.23

Doc 48: Summary of the Age variable

The minimum age of the patients is 17. The maximum age of the patient is 99. The 1st Quantile is 60, meaning that 25% of our sample are younger than 60 and 75% are older than 60. The median is 68.00 meaning that half of our sample is younger than 68. The other half being older than 68 mm. The 3rd Quantile is 76.0. This means that 75% are younger than 76.00 and 25% older than 76. The average age of our sample is 67.36 and the standard deviation is just 11.23. Finally the most repeated value is 75.00 which is none other than the mode. Keep in mind, for visualization purposes, you can take a look at this boxplot. It is a good way to check the 1st Quantile, 3rd Quantile and the median. The distribution o the age variable also figure out here. (Doc 49, 50)



Doc 49: Age Boxplot



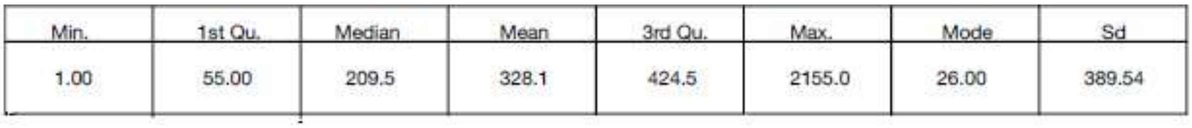
Doc 50: Histogram of the age variable

We can really see that we are dealing mostly with people who did pass their fifties. This is interesting and can lead us to think that older people have higher risk of getting liver cancer.

It is essential to estimate the real average age patient with liver cancer in the population. For that, we need to compute the confidence interval which is I = [66.62 ; 68.09]. Please note that this is done with alpha=0.05 and so with that in mind, we can conclude that the real average age of patient having is between 66.62 and 68.09, and this is for a 95% confidence level. The average is pretty high and sort of confirm our latest suggestion.

Y. Survival in days.

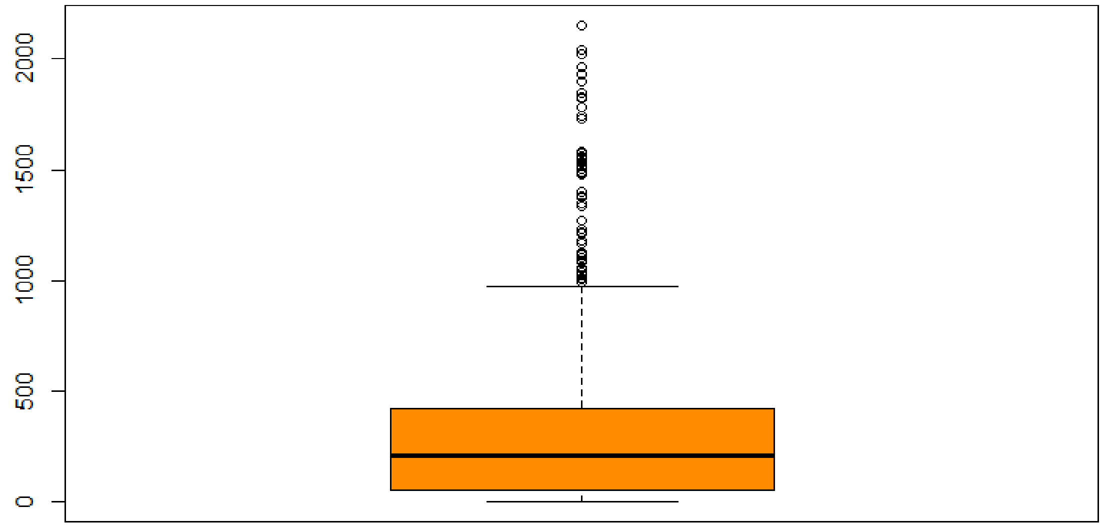
Let’s analyse how many days a patient can expect to live. Here, we can see a brief summary regarding this variable. (Doc 51)



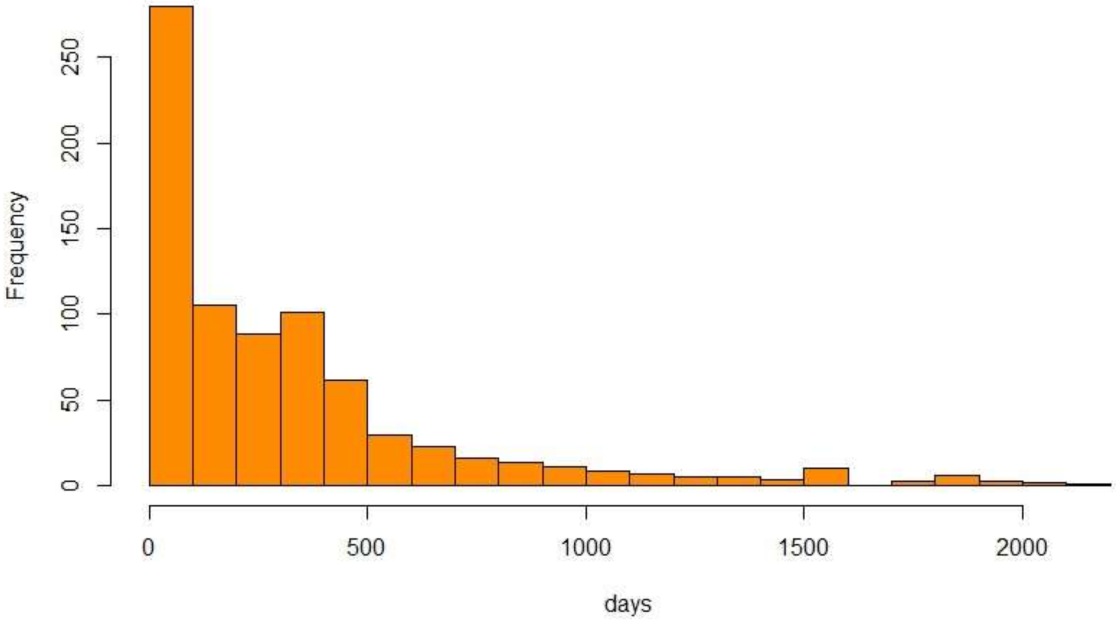
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mode	Sd
1.00	55.00	209.5	328.1	424.5	2155.0	26.00	389.54

Doc 51: Summary of the Survival In days variable

The minimum days of survival for the patients with liver cancer is 1. The maximum days of survival for the patients is 2155.0. The 1st Quantile is 55.00, meaning that 25% of our sample survived less than 55.00 days and 75% survived more than 55.00 days. The median is 209.5 meaning that half of our sample survived less than 209.5. The other half survived more than 209.5. The 3rd Quantile is 424.5. This means that 75% survived more than 424.5days and 25% survived less than 424.5days. The average survival day of our sample is 209.5 and the standard deviation is just 389.54. Finally the most repeated value is 26.00 which is none other than the mode. Keep in mind, for visualization purposes, you can take a look at this boxplot. It is a good way to check the 1st Quantile, 3rd Quantile and the median. The distribution of this variable in our sample is also present. (Doc 52,53)



Doc 52: Survival in days Boxplot

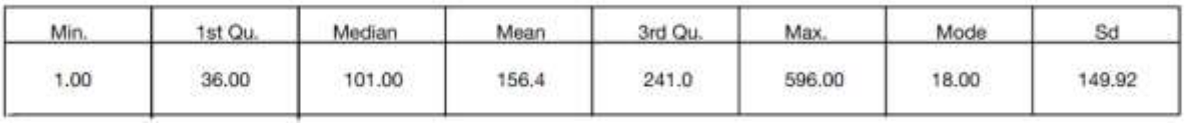


Doc 53: Survival in days histogram

We can see patient surviving fairly well. You see, here we are also working on people who actually were cured from the disease. This why we have high value. It would be interesting to see how many days patients, who were unfortunately dead, would actually survive. This is what we are going to look on later on. But for now, onto the confidence interval I = [300.82; 355.29]. Please note that this is done with alpha=0.05 and so with that in mind, we can conclude that the real average of day patient having the liver cancer should expect to survive is between 300.82 and 353.29, and this is for a 95% confidence level. The mean is pretty high and sort of confirm our latest suggestion.

Z. Survival in days for dead patient.

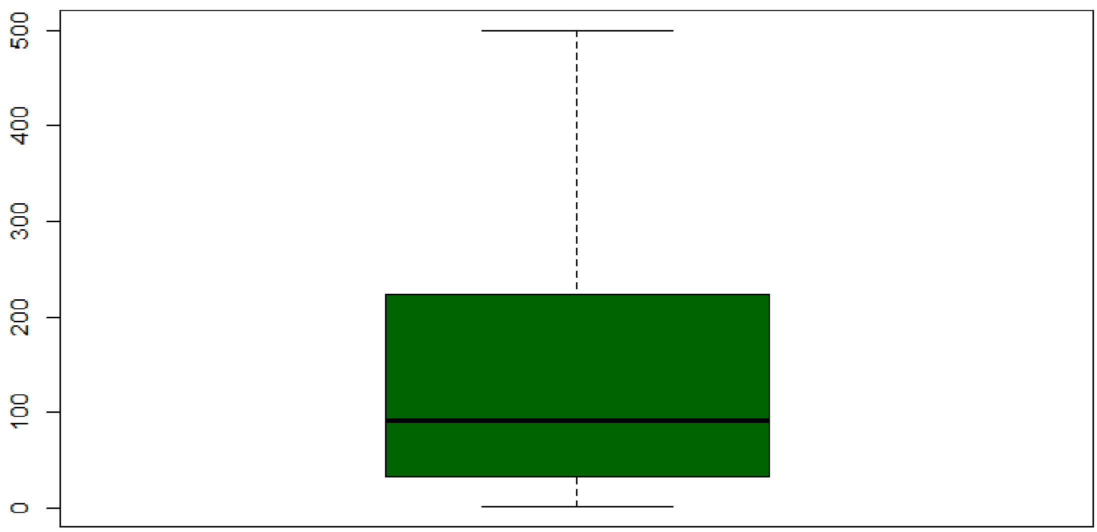
Let’s analyse how many days a patient who is dead lived. Here, we can see a brief summary regarding this variable. (Doc 54)



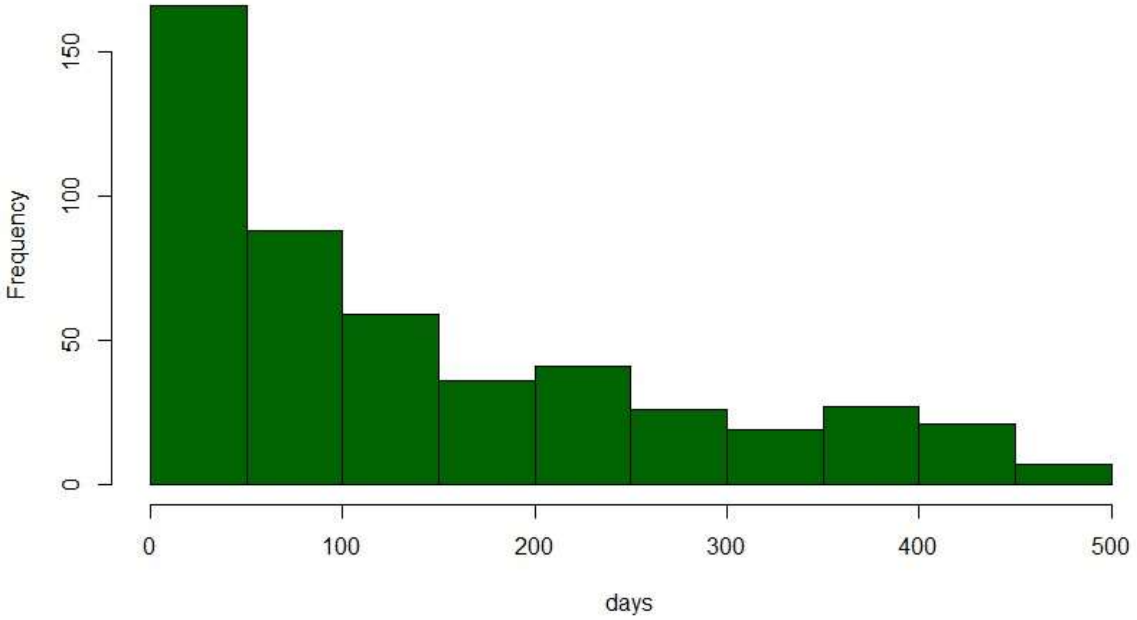
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mode	Sd
1.00	36.00	101.00	156.4	241.0	596.00	18.00	149.92

Doc 54: Summary of the Survival In days variable for dead people

The minimum days of survival for the patients with liver cancer who died is 1. The maximum days of survival for these patients is 596. The 1st Quantile is 36.00, meaning that 25% of our sample survived less than 36.00 and 75% survived more than 36.00. The median is 101.00 meaning that half of our sample survived less than 101.00. The other half survived more than 101.00. The 3rd Quantile is 241.0. This means that 75% survived more than 241.0 days and 25% survived less than 241.0 days. The average survival day of our sample is 156.4 and the standard deviation is just 149.92. Finally the most repeated value is 276.61 which is none other than the mode. Keep in mind, for visualization purposes, you can take a look at this boxplot. It is a good way to check the 1st Quantile, 3rd Quantile and the median. The distribution of this variable in our sample is also present. (Doc 55,56)



Doc 55: Survival in days Boxplot for dead patient

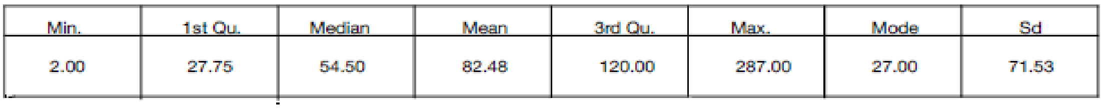


Doc 56: histogram of the survival in days for dead patient

You see? Even if we have almost have the same shape as before, we can notice that every statistical value have decreased giving us a better overview about patient who have no hope of winning the battle against cancer. Now, we could see for instance, for each cancer stage, especially the last 2, how many days do they survive. This will help us estimate how many days these helpless patient would expect to live, before ultimately losing the battle. More on that later on. For now, let us compute the confidence interval. I fact, I=[129.39 ; 152.38]. Please note that this is done with alpha=0.05 and so with that in mind, we can conclude that the real average of days patient who were dead due to the liver cancer survived between 129.39 and 152.38, and this is for a 95% confidence level.

AA. Survival in days for dead patient with cancer stage C or D

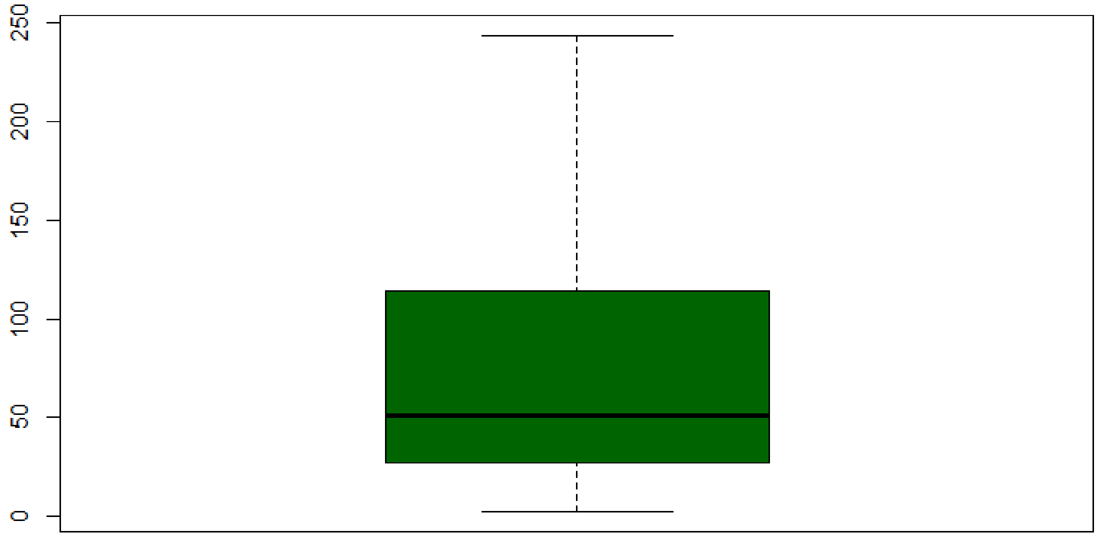
Let's analyse how many days a patient who is dead and have cancer stage C or D lived. Here, we can see a brief summary regarding this variable. (Doc 57)



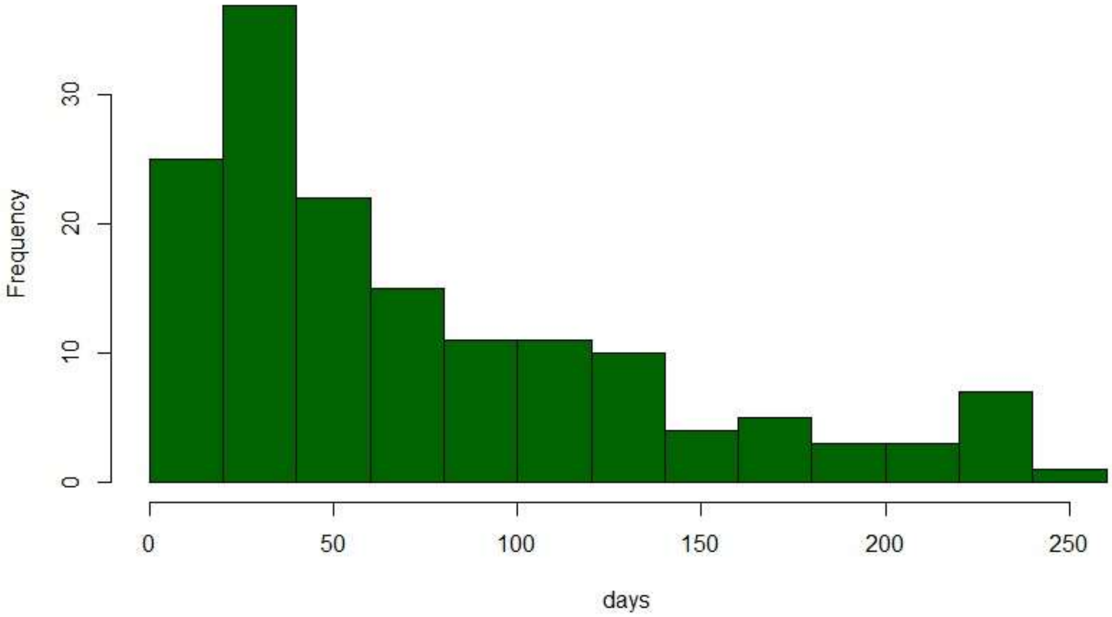
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Mode	Sd
2.00	27.75	54.50	82.48	120.00	287.00	27.00	71.53

Doc 57: Summary of the Survival In days for patient with cance C or D

The minimum days of survival for these patients is 2. The maximum days of survival for the patients is 287. The 1st Quantile is 27.75, meaning that 25% of our sample survived less than 27.75 and 75% survived more. The median is 87.48 meaning that half of our sample survived less than 87.48. The other half survived more than that. The 3rd Quantile is 120.00. This means that 75% survived more than 120.00 days and 25% survived less than that. The average survival day of our sample is 82.58 and the standard deviation is just 71.53. Finally the most repeated value is 27.00 which is none other than the mode. Keep in mind, for visualization purposes, you can take a look at this boxplot. It is a good way to check the 1st Quantile, 3rd Quantile and the median. The distribution of this variable in our sample is also present. (Doc 58,59)



Doc 58: Survival in days Boxplot for Cancer C and D



Doc 59: Histogram of the survival days for cancer C and D

Let’s compute the confidence interval in order to see what is the expected survival days for these patient. I=[65.39 ; 85.16]. This means that the real expected value for these people with cancer stage C and D, if they are expected to die is between 65.39 and 85.16 days and this is for a 95% confidence level.

III. Inferential Statistics

Our goal in this section is to tackle some specific question regarding our Data. Since we have taken a look at our variables in our descriptive statistics, now it is time for the real analysis which will be done in a scientific manner. The questions we will be answering are:

- What are the factors that influence the survival in days?
- What are the factors that influence the death status of a patient?
- The tumour size follows which distribution?
- The survival in days follows which distribution?

A. What are the factors that influence the survival in days?

Below, we listed only the factors that influenced survival in days variable. Note that the criterion used is α=0.05

1. Screening

This was an interesting result I must say. Before diving into it, we must start with our hypothesis

$$\begin{cases} Ho : \mu_S = \mu_{uns} \text{ (No difference between the screened and unscreened)} \\ H1 : \mu_S \neq \mu_{uns} \text{ (there is a difference between the screened and unscreened)} \end{cases}$$

The test that we have made, gave us a p-value <2.2e-16 which is lower than α=0.05 . This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 genders is different. The screening process seems to be important in determining how many days you survive, and perhaps winning the battle against cancer

2. Thrombosis

Here, the thrombosis variable have two modalities. The first one is for the people who have thrombosis and the second one is for the people who do not have. Let’s start with our hypothesis

$$\begin{cases} \text{Ho : } \mu_{thr} = \mu_{nothr} \text{ (No difference between the two groups)} \\ \text{H1 : } \mu_{thr} \neq \mu_{nothr} \text{ (there is a difference between the two groups)} \end{cases}$$

The test that we have made, gave us a p-value<2.2e-16 which is lower than $\alpha=0.05$. This means that we should reject Ho and this is for a 95% threshold. As a result, we can confirm on a 95% threshold that the means of survival days for the 2 groups is different. It seems that people suffering from thrombosis survive less.

This is another hypothesis test. The interesting part here is that we are checking if thrombosis decrease the day of survivorship,.

$$\begin{cases} \text{Ho : } \mu_{thrr} >= \mu_{nthr} \text{ (survival days thromb cancer is more than no thromb)} \\ \text{H1 : } \mu_{thr} < \mu_{nothr} \text{ (survival days of thromb is less than no thromb)} \end{cases}$$

We computed p-value and this what we got. P-value= 2.2e-16 <0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can say that having a thrombosis will decrease the survival in days.

3. Diffuse Cancer

Once again, the thrombosis variable have two modalities. The first one is for the people who diffused the cancer completely. The second one is for the patients that did not undergo such operation.

Ho : $\mu_{dif} = \mu_{nodif}$ (No difference between the two groups)

H1 : $\mu_{dif} \neq \mu_{nodif}$ (there is a difference between the two groups)

The test that we have made, gave us a p-value=2.852e-14< α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 groups is different. Perhaps diffusing cancer result in a better survivability. The word perhaps is interesting because that not the case… you will see what I am saying.

This is another hypothesis test. The interesting part here is that we are checking if diffusing cancer actually help people survive more

Ho : $\mu_{dif} >= \mu_{nodif}$ (survival days diffuse cancer is more than no diffuse cancer)

H1 : $\mu_{dif} < \mu_{nodif}$ (survival days of diffuse is less than no diffuse)

We computed p-value and this what we got. P-value= 1.42e-14 <0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can say that diffusing cancer actually decrease the survivability of the patient.

4. Metastatic Cancer

Once again, the Metastatic variable have two modalities. The first one is for the people who have a metastatic cancer. The second one is for the patients that did not have a metastatic variable.

Ho : $\mu_{met} = \mu_{nomet}$ (No difference between the two groups)

H1 : $\mu_{met} \neq \mu_{nodmet}$ (There is a difference between the two groups)

The test that we have made, gave us a p-value=1.621 -14< α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 groups is different.

5. Curative Treatment

Once again, the curative treatment variable have two modalities. The first one is for the people who have a curative treatment. The second one is for the patients that did not have a curative treatment.

Ho : $\mu_{cur} = \mu_{nocur}$ (No difference between the two groups)

H1 : $\mu_{cur} \neq \mu_{nocur}$ (There is a difference between the two groups)

The test that we have made, gave us a p-value=2.2e-16< α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 groups is different. So let's check if the curative help with the survivability

This is another hypothesis test. The interesting part here is that we are checking if the curative treatment actually helps people survive more

Ho : μ_{cur} =< μ_{nocur} (survival days diffuse cancer is more than no diffuse cancer)

H1 : μ_{cur} > μ_{nocur} (survival days of diffuse is less than no diffuse)

We computed p-value and this what we got. P-value <0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can say that curative treatment actually increases the survivability of the patient.

6. Alcohol

Once again, the Alcohol variable have two modalities. The first one is for the people who drinks alcohol. The second one is for the patients that do not drink alcohol.

Ho : $\mu_{cur} = \mu_{nocur}$ (No difference between the two groups)

H1 : $\mu_{cur} \neq \mu_{nocur}$ (There is a difference between the two groups)

The test that we have made, gave us a p-value=0.0036< α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 groups is different. So let's check if drinking alcohol decreases the survivability days

This is another hypothesis test. The interesting part here is that we are checking if the alcohol is actually bad

Ho : μ_{alc} =< μ_{noalc} (survival days diffuse cancer is more than no diffuse cancer)

H1 : μ_{noalc} > μ_{alc} (survival days of diffuse is less than no diffuse)

We computed p-value and this what we got. P-value =0.0018<0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can confirm that Alcohol decreases the survivability of the patient.

7. HCV

Once again, the HCV variable have two modalities. The first one is for the people who have a HCV. The second one is for the patients that did not have HCV.

Ho : $\mu_{HCV} = \mu_{nohcv}$ (No difference between the two groups)

H1 : $\mu_{met} \neq \mu_{nodmet}$ (There is a difference between the two groups)

The test that we have made, gave us a p-value=0.005< α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that the means of survival days for the 2 groups is different. Having HCV influence the days of survivorship.

8. Cancer Stages

Cancer Stages have four modalities. These modalities refers to the different cancer stages as we saw earlier. Since we are dealing with 4 modalities, the hypothesis test is a little bit different.

Ho : $\mu_A = \mu_B = \mu_C = \mu_D$ (No difference between the two groups)

H1 : $\mu_A \neq \mu_{Bt} \neq \mu_C \neq \mu_D$ (There is a difference between at least one group)

The test that we have made, gave us a p-value<2.2e-16 < α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can confirm on a 95% threshold that at least one mean of a group is different from the other. My hypothesis now is, having a more developed cancer stage limit my days of living?

Here we are doing a series of one tailed tests. These tests will help us answer our questions

Ho : $\mu_A =< \mu_B$ (survival days of cancer A is less than Cancer B)

H1 : $\mu_A > \mu_B$ (survival days of cancer A is more than CancerB)

We computed p-value and this what we got. P-value =0.0058<0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can confirm that a patient with cancer stage A will survive longer than a patient with cancer stage B.

Ho : μ_B =< μ_C (survival days of cancer B is less than Cancer C)

H1 : $\mu_B > \mu_C$ (survival days of cancer B is more than Cancer C)

We computed p-value and this what we got. P-value =0.00079<0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can confirm that a patient with cancer stage B will survive longer than a patient with cancer stage C.

Ho : μ_C =< μ_D (survival days of cancer C is less than Cancer D)

H1 : $\mu_C > \mu_D$ (survival days of cancer C is more than Cancer D)

We computed p-value and this what we got. P-value <2.2e-16 <0.05 and so we actually reject Ho. As result, on a threshold of 95%, we can confirm that a patient with cancer stage C will survive longer than a patient with cancer stage D.

As a final mini conclusion, the more the cancer is developed the less days you should expect to live.

9. Criteria

Criteria variable have two modalities. The first one refers to the patients with tumour size below 5cm the second one refers to the tumour size above 5cm.

Ho : $\mu_{in} = \mu_{out}$ (No difference between the two groups)

H1 : $\mu_{in} \neq \mu_{out}$ (There is a difference between the two groups)

The test that we have made, gave us a p-value=2.2e-16 < α=0.05. This means that we should reject Ho and this is for a 95% threshold. As a result, We can conclude that on a 95% confidence level, the two variable, the varable criteria impact the survival in days

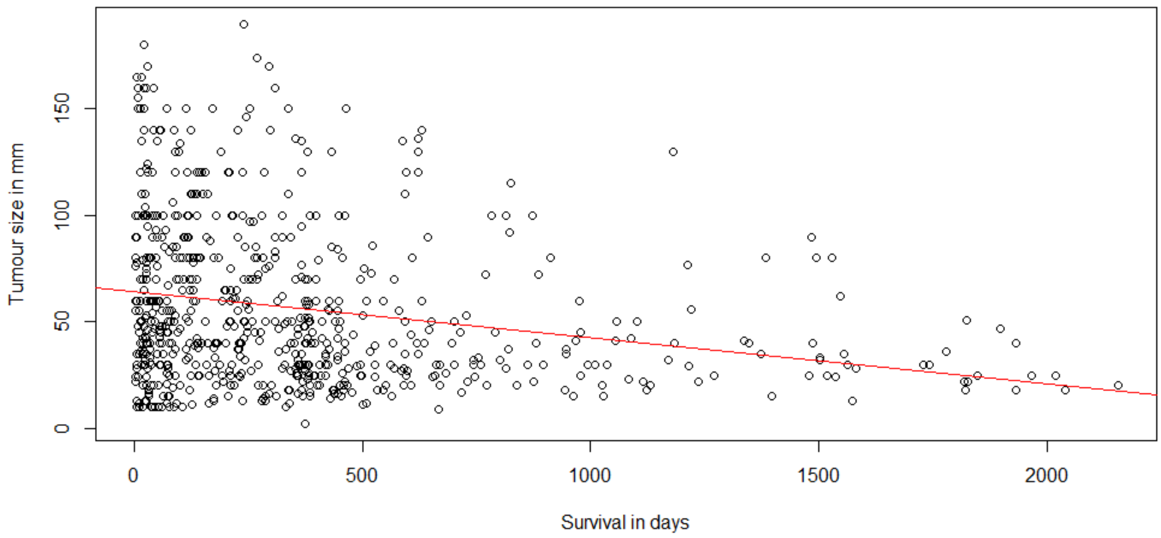
10. Tumour size

For the tumour size, we are going to perform a correlation test (Persean).

R is equal to 0

R is not equal to 0

The estimated coefficient of correlation is equal to -0.23. And the test tells us that the real coefficient of correlation is in this confidence interval I=[-0.30;-0.16]. Of course, we are working on a 95% confidence level. We can see that 0 does not figure out in this confidence interval, and so we are 95% confident that R is not equal to 0. So there is a linear relation between tumour size and survival in days. In fact it is a negative one. The bigger the tumour size the less time you are expected to live. Here is a regression model (Doc 60)



Doc 60: Linear regression and relation between tumour size and survival in days

For curiosity and prediction purpose here is the equation:

$$Y=-0.0218X+64.4275$$

Where Y is the tumour size in cm and X is the survival in days.

B. What are the factors that influence the death variable?

Let's take a look at the variable that influence whether a patient will die or no

1. Gender

Interesting result right here. It seem that the gender of the patient will influence the death variable. Let's perform ha hypothesis test

$$\begin{cases} Ho : p_M = p_F \text{ (Proportion of the two groups is the same)} \\ H1 : p_M \neq p_D \text{ (Proportion of the two groups is different)} \end{cases}$$

The p-value is equal to 0.02 which is less than 0.05. In that sense, we can say confirm, with a threshold of 95%, that the gender variable can influence whether or not a patient survive

2. Age group

Age group is a variable with two modalities as we have mentioned.

Ho : $p_{lower70} = p_{higher70}$ (Proportion of the two groups is the same)

H1 : $p_{lower70} \neq p_{higher}$ (Proportion of the two groups is different)

Pvalue ≈ 0.05. In that sense, we can say confirm, with a threshold of 95%, that the age group variable can influence whether or not a patient survive. It seems that the older you are, the more difficult it is the win the battle against cancer

3. Screening

Screening is a variable with two modalities as we have mentioned. There are patient who have been screened, and other who have not.

Ho : $p_{screening} = p_{noscr}$ (Proportion of the two groups is the same)

H1 : $p_{screening} \neq p_{noscr}$ (Proportion of the two groups is different)

Pvalue = 3.3e-5<0.05. In that sense, we can say confirm, with a threshold of 95% that the screening test can influence whether or not a patient survive. It seems that patient should undergo a screening test, for a better chance to be cured from this disease.

4. Thrombosis

Thrombosis is a variable with two modalities as we have mentioned. There are patient who have Thrombosis, and other who do not.

Ho : $p_{thr} = p_{nothr}$ (Proportion of the two groups is the same)

H1 : $p_{thr} \neq p_{nothr}$ (Proportion of the two groups is different)

Pvalue = 1.249e-8<0.05. In that sense, we can say confirm, with a threshold of 95% that the thrombosis disease can influence whether or not a patient survive. It seems that patient with thrombosis disease might present a higher risk of death.

5. Diffuse Cancer

Diffuse Cancer is a variable with two modalities as we have mentioned. There are patient who have diffused their tumour, and other who did not undergo such operation.

$$\left\{\begin{array}{l} \text{Ho : } p_{dif} = p_{nodif} \text{ (Proportion of the two groups is the same)} \\ \\ \text{H1 : } p_{dif} \neq p_{nodif} \text{ (Proportion of the two groups is different)} \end{array}\right.$$

Pvalue = 3.989e-06<0.05. In that sense, we can say confirm, with a threshold of 95% that this operation can play a role on the destiny of a patient.

6. Metastatic Cancer

Metastatic Cancer is a variable with two modalities as we have mentioned. There are patient who have diffused their tumour, and other who did not undergo such operation.

$$\left\{\begin{array}{l} \text{Ho : } p_{met} = p_{nomet} \text{ (Proportion of the two groups is the same)} \\ \\ \text{H1 : } p_{met} \neq p_{nomet} \text{ (Proportion of the two groups is different)} \end{array}\right.$$

Pvalue = 0.004<0.05. In that sense, we can say confirm, with a threshold of 95% that this having a metastatic cancer can influence on the death variable. As a logic response unrelated to the test, we can say having a metastatic cancer could increase the death of someone

7. Curative treatment

Curative treatment is a variable with two modalities as we have mentioned. There are patient who have undergone a curative treatment, and other who did not undergo such treatment.

Ho : $p_{cur} = p_{nocur}$ (Proportion of the two groups is the same)

H1 : $p_{cur} \neq p_{nocur}$ (Proportion of the two groups is the same)

Pvalue < 2.2e-16 <0.05. In that sense, we can say confirm, with a threshold of 95% that this having done a curative treatment can influence on the death variable. Maybe curative treatment help us with our battle against cancer?

8. Alcohol

Alcohol is a variable with two modalities as we have mentioned. There are patient who usually drink alcohol, and other who do not drink.

Ho : $p_{alc} = p_{noalc}$ (Proportion of the two groups is the same)

H1 : $p_{alc} \neq p_{noalc}$ (Proportion of the two groups is the same)

Pvalue = 4.77e-5 <0.05. In that sense, we can confirm, with a threshold of 95% that this having drinking alcohol is statistically significantly associated with the death of a patient.

9. HCV

HCV is a variable with two modalities as we have mentioned. There are patient who suffered from the HCV diseases and other who did not

Ho : p_{HCV} = p_{noHCV} (Proportion of the two groups is the same)

H1 : $p_{HCV} \neq p_{noHCV}$ (survival days of cancer C is more than Cancer D)

Pvalue = 0.00608 <0.05. In that sense, we can confirm, with a threshold of 95% that the HCV disease is stastically significantly associated with the death of a person

10. Criteria

Criteria is a variable with two modalities as we have mentioned. There are patient with a tumour size bigger than 5cm and others with a tumour size smaller than 5cm

Ho : p_{in} = p_{out} (Proportion of the two groups is the same)

H1 : $p_{in} \neq p_{out}$ (Proportion of the two groups is the same)

Pvalue < 2.2e-16 <0.05. In that sense, we can confirm, with a threshold of 95% that the the size of a tumours is stastically significantly associated with the death of a person

11. Cancer stage

Cancer stages is a variable with 4 modalities as we have mentioned.

$$\begin{cases} Ho : p_A = p_B = p_C = p_D \text{ (Proportion of the two groups is the same)} \\ H1 : p_A \neq p_B \neq p_C \neq p_D \text{ (Proportion of the two groups is the same)} \end{cases}$$

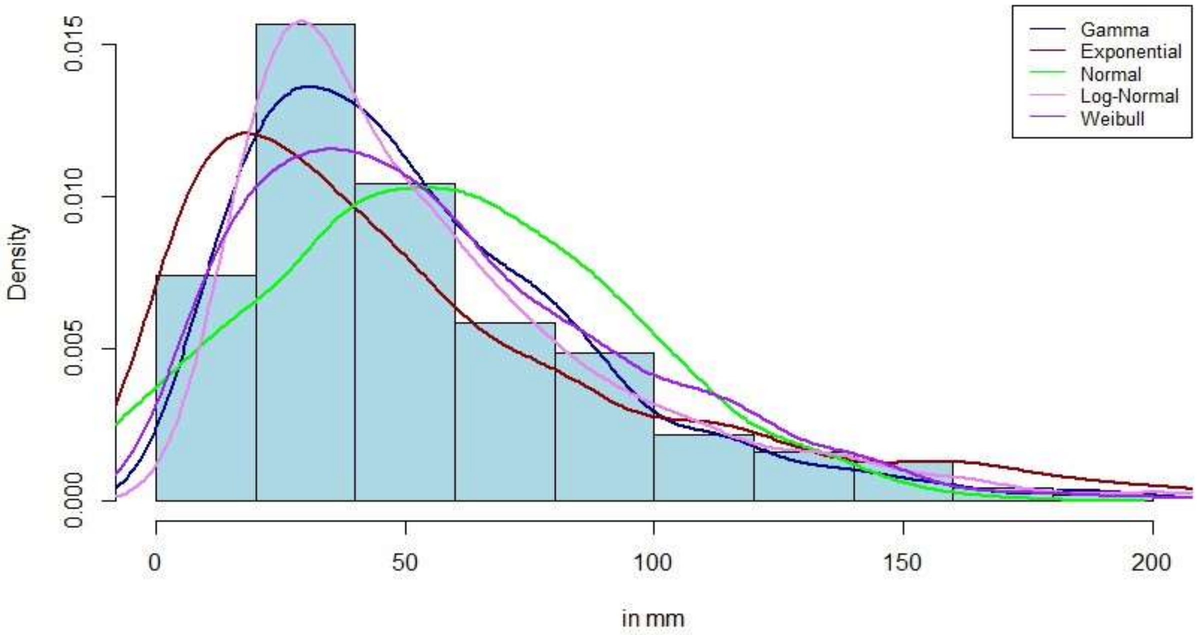
Pvalue < 1.03e-13 <0.05. In that case, we can confirm that at least one proportion above is different from the other, and this is for a threshold of 95%.

C. What is the distribution of the tumour size?

We have tried and approximate 5 distribution to the tumour size distribution. After calculating the parameter of each distribution, here is what we have got:

- Gamma(α=2.399,β=0.042)
- Exponential(λ=0.017)
- Normal distribution : N(56.398,37.869)
- Log Normal distribution : LN(3.809,0.686)
- Weibull distribution: Weibull(k=1.59 ,λ=63.28)

For visualization purposes, we have plotted all the 5 distributions, on top of the histogram of the tumour size. This is a good way to visually perceive which distribution does the tumour size approximately follows. (Doc 61)



We can see, at first glance, that the Log-Normal distribution fit our histogram perfectly. The normal, exponential does not fit the histogram fairly well. But in the end, we cannot conclude anything by just looking visually to the graph. Statistical test should be done before doing any sort of conclusion.

1. Gamma(α=2.399,β=0.042)

We will do some hypothesis testing once again

Ho: The tumour size follows a Gamma(α=2.399,β=0.042)

H1: The tumour does not follow a Gamma(α=2.399,β=0.042)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue=0.0006<0.05. We have done also another test called Anderson-Darling test, and we got pvalue=0.006<0.05. In both cases, we reject Ho 0n a 95% threshold. This means that tumour size does not follow a Gamma distribution.

2. Exponential(λ=0.017)

We will do some hypothesis testing once again

Ho: The tumour size follows an exponential(λ=0.017)

H1: The tumour does not follow an exponential(λ=0.017)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue<2.2e-16<0.05. We have done also another test called Anderson-Darling test, and we got pvalue=7.732e-07 <0.05. In both cases, we reject Ho 0n a 95% threshold. This means that tumour size does not follow an exponential(λ=0.017) distribution

3. Normal distribution : N(56.398,37.869)

We will do some hypothesis testing once again

Ho: The tumour size follows a Normal distribution N(56,398,37.869)

H1: The tumour does not follow a Normal distribution N(56,398,37.869)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue<2.998e-15<0.05. We have done also another test called Anderson-Darling test, and we got pvalue<7.732e-7<0.05. In both cases, we reject Ho 0n a 95% threshold. This means that tumour size does not follow a Normal distribution N(56,398,37.869)

4. Log Normal distribution : LN(3.809,0.686)

We will do some hypothesis testing once again

Ho: The tumour size follows a Log Normal distribution : LN(3.809,0.686)

H1: The tumour does not follow a Log Normal distribution : LN(3.809,0.686)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue≈ 0.05=0.05. We have done also another test called Anderson-Darling test, and we got pvalue=0.1062 >0.05. In both cases, we do not reject Ho on a 95% threshold. This means that tumour size does follow a LN(3.809,0.686).

5. Weibull distribution: Weibull(k=1.59 ,λ=63.28)

We will do some hypothesis testing once again

Ho: The tumour size follows a Weibull distribution: Weibull(k=1.59 ,λ=63.28)

H1: The tumour does not follow a Weibull distribution: Weibull(k=1.59 ,λ=63.28)

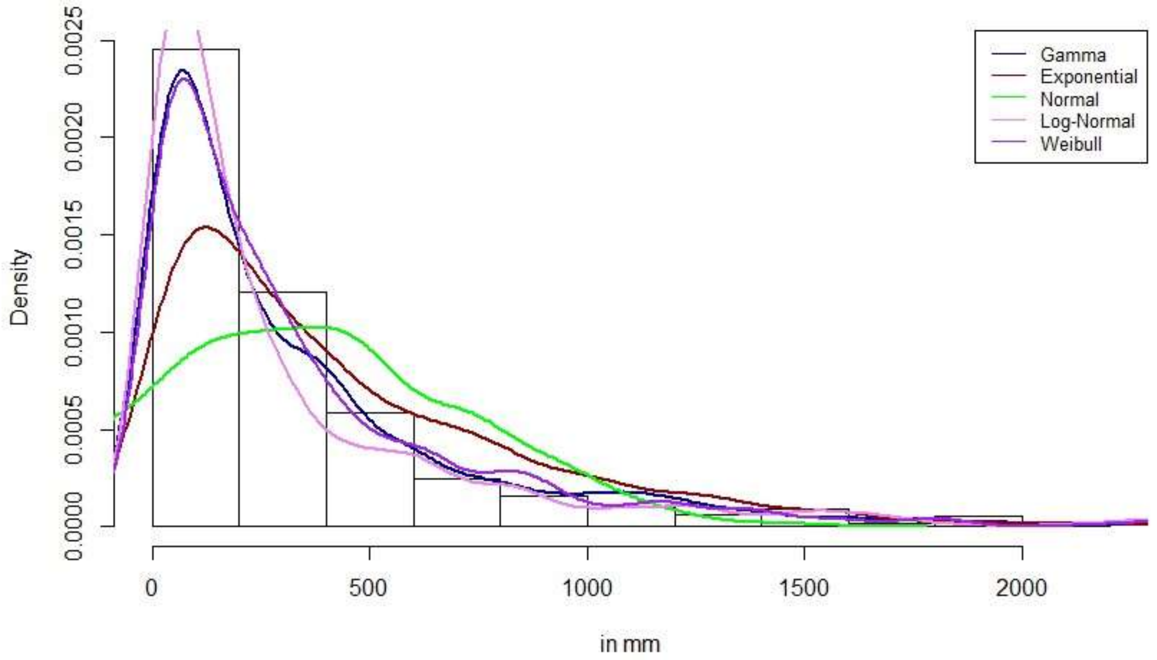
We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue= 3.348e-05<0.05. We have done also another test called Anderson-Darling test, and we got pvalue=0.0005 <0.05. In both cases, we do reject Ho on a 95% threshold. This means that tumour size does not follow a Weibull distribution: Weibull(k=1.59 ,λ=63.28)

D. What is the distribution of the survival in days variable?

We have tried and approximate 5 distribution to the survival in days distribution. After calculating the parameter of each distribution, here is what we have got:

- Gamma(α=0.762,β=0.0023)
- Exponential(λ=0.0021)
- Normal distribution : N(328.05,389.29)
- Log Normal distribution : LN(5.01,1.46)
- Weibull distribution: Weibull(k=0.833,λ=298.67)

For visualization purposes, we have plotted all the 5 distributions, on top of the histogram of the tumour size. This is a good way to visually perceive which distribution does the tumour size approximately follows. (Doc 62)



Doc 62: Survival in days Boxplot for Cancer C and D

If we want to conclude visually, we can say that the normal distribution is out of contest. Gamma and Weibull both have a great chance of representing our variable. But, as we said before further statistical test must be done in order tp confirm what we just have said.

1. Gamma(α=0.762,β=0.0023)

We will do some hypothesis testing once again

Ho: The tumour size follows a Gamma(α=2.399,β=0.042)

H1: The tumour does not follow a Gamma(α=2.399,β=0.042)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue=0.1201 >0.05. We have done also another test called Anderson-Darling test, and we got pvalue=0.052>0.05. In both cases, we do not reject Ho 0n a 95% threshold. This means that tumour size does follow a Gamma distribution.

2. Exponential(λ=0.017)

We will do some hypothesis testing once again

Ho: The tumour size follows an exponential(λ=0.0021)

H1: The tumour does not follow an exponential(λ=0.0021)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue<2.2e-16<0.05. We have done also another test called Anderson-Darling test, and we got pvalue=7.634e-07 <0.05. In both cases, we reject Ho 0n a 95% threshold. This means that tumour size does not follow an exponential(λ=0.0021) distribution.

3. Normal distribution : N(328.05,389.29)

We will do some hypothesis testing once again

Ho: The tumour size follows a Normal distribution N(328.05 ,389.29)

H1: The tumour does not follow a Normal distribution N(328.05 ,390.29)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue<2.2e-16<0.05. We have done also another test called Anderson-Darling test, and we got pvalue<7.734e-7<0.05. In both cases, we reject Ho 0n a 95% threshold. This means that tumour size does not follow a Normal distribution N(328.05,390.29)

4. Log Normal distribution : LN(5.01,1.46)

We will do some hypothesis testing once again

Ho: The tumour size follows a Log Normal distribution : LN(5.01,1.46)

H1: The tumour does not follow a Log Normal distribution : LN(5.01,1.46)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue=2.286e-8<0.05. We have done also another test called Anderson-Darling test, and we got pvalue=5.065 <0.05. In both cases, we do not reject Ho on a 95% threshold. This means that tumour size does follow a LN(5.01,1.46).

5. Weibull distribution: Weibull(k=0.833,λ=298.67)

We will do some hypothesis testing once again

Ho: The tumour size follows a Weibull distribution: Weibull(k=1.59 ,λ=63.28)

H1: The tumour does not follow a Weibull distribution: Weibull(k=1.59 ,λ=63.28)

We have done two tests. The first one is called Kolmogorov-Smirnov, and it gave us pvalue= 0.1813>0.05. We have done also another test called Anderson-Darling test, and we got pvalue=0.07 >0.05. In both cases, we do not reject Ho on a 95% threshold. This means that tumour size does follow a Weibull distribution: Weibull(k=0.833,λ=298.67)

6. Weibull or Gamma?

We found out that the survival in days distribution follows two distribution and so we have to decide which in better. We can say that the with smaller maximum loglikely hood will be chosen.

|Loglik(“Weibull”)|<|loglik(“gamma”|

And so we decided that survival in days follows a Weibull distribution.

E. Treatment 1 to 6.

Before going any further, I wanted to stop and talk about these treatment. None of them affected the death or even the survival in days. Statistically speaking, these treatments do not prolonge the expected days of the patient to live. Even worse, they do not affect the variable death meaning they do not cure a person. Therefore they seem to be useless. I would opt for a Diffuse cancer operation or for a curative treatment.

IV. Estimation of Mean Sojourn time

The next, and final big thing about our article is to estimate the mean sojourn time. For decades, researcher have been estimating the mean sojourn time of breast cancer, liver cancer etc. This is primarily for identifying a suitable round length in a new let's say, in our case liver screening programme. In fact, the sojourn important because it will determine the effectiveness of a screening program. Furthermore, we call MST, the mean sojourn time, as the average time from becoming screen detectable, to becoming clinically apparent in the absence of screening. This is very interesting because with MST, we can know how long a patient, who already have been screen, should wait in order to undergo another screening. So a program could be made for each patient who are at risk of developing the liver cancer. Let's say, a person X is an alcoholic person, and had suffered HCV during his time on this planet. This guy have high risk of developing cancer and so, he should undergo a screening program. So at first, he may do a screening test, and he was quite fortunate that he did not have a liver cancer. But how do we know how many time should we wait for another screening to be done? This is where the MST will come in handy.

A. What is the formula of the MST?

The MST, is given by this formula:

$$MST = 3DT\frac{\ln\left(\frac{dt1}{dto}\right)}{\ln(2)}$$

- DT: doubling time

- dt1: Tumour size of unscreened patients
- dto: Minimum Tumour size detected

The doubling time is the time that it takes for a quantity to double in size or value. In our case, we are referring to the time needed for a liver tumour to double its size.

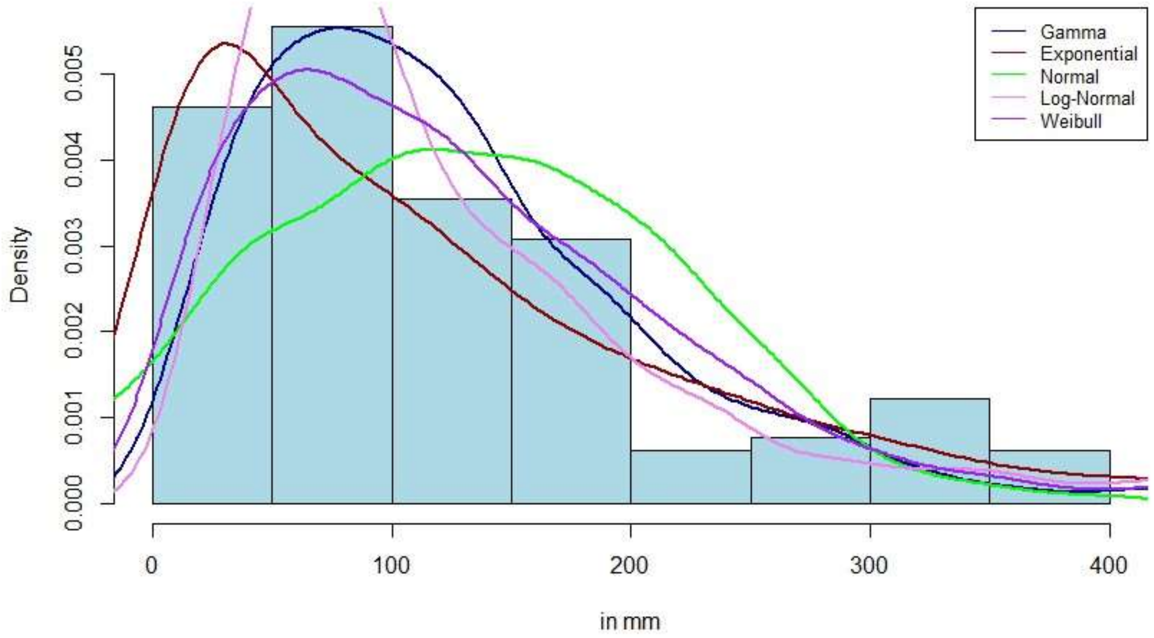
In order for us to determine the mean sojourn time, we need to know which distribution it follows. But, before we can do such thing. It is essential to know what distribution does the doubling time, the tumour size of unscreened patients and minimum tumour size follows. One thing to know is that we already the distribution of the minimum tumour size detected. In fact, it follows a triangular distribution (1; 1.2; 1.4).

B. Doubling Time

We have tried and approximate 5 distributions to the doubling time distribution. After calculating the parameter of each distribution, here is what we have got:

- Gamma(α=1.962,β= 0.0154)
- Exponential(λ=0.0078)
- Normal distribution : N(126.69,93.22)
- Log Normal distribution : LN(4.565,0.771)
- Weibull distribution: Weibull(k= 1.438,λ= 140.351)

For visualization purposes, we have plotted all the 5 distributions, on top of the histogram of the tumour size. This is a good way to visually perceive which distribution does the tumour size approximately follows. (Doc 63)



Doc 63: Survival in days Boxplot for Cancer C and D

Again, a lot of distribution are close to our histogram. We cannot conclude. We have to opt for some statistical testing.

1. Gamma(α=01.962,β=0.015)

We will do some hypothesis testing once again

Ho: The tumour size follows a Gamma(α=1.962,β= 0.015)

H1: The tumour does not follow a Gamma(α=1.962,β= 0.015)

We have done a test called Kolmogorov-Smirnov. pvalue =7.777e-16<0.05. With that in mind, we can reject Ho on a 95% confidence level. And so the doubling time does not follow a gamma distribution

2. Weibull(α= 1.438,β=140.35)

We will do some hypothesis testing once again

Ho: The tumour size follows a Weibull (k=1.438,λ= 140.35)

H1: The tumour does not follow a Weibull (α=140.35,β= 140.35)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.63>0.05. Another test will be done called Anderson-Darling. Pvalue=0.36.>0.05 With that in mind, we do not reject Ho on a 95% confidence level. And so the doubling time follow a Weibull distribution.

3. Exponential(λ= 0.0078)

We will do some hypothesis testing once again

Ho: The tumour size follows a Exponential (λ=0.0078)

H1: The tumour does not follow a Exponential (λ=0.0078)

We have done a test called kolmogorov-smirnov. pvalue =0.0059<0.05. With that in mind, we can reject Ho on a 95% confidence level. And so the doubling time does not follow an exponential distribution

4. Lognormal: LN(4.56, 0.77)

We will do some hypothesis testing once again

Ho: The tumour size follows a Lognormal: LN(4.56, 0.77)

H1: The tumour does not follow a Lognormal: LN(4.56, 0.77)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.47>0.05. Another test will be done called Anderson-Darling. Pvalue=0. 62.>0.05 With that in mind, we do not reject Ho on a 95% confidence level. And so the doubling time follow a lognormal distribution

5. Normal: N(126.69, 93.22)

We will do some hypothesis testing once again

Ho: The tumour size follows a Normal: N(126.69, 93.22)

H1: The tumour does not follow a Normal: N(126.69, 93.22)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.47>0.05. With that in mind, we do reject Ho on a 95% confidence level. And so the doubling time does not follow a normal distribution

6. Weibull or Lognormal?

We found out that the survival in days distribution follows two distribution and so we have to decide which one is better. We can say that the distribution with smaller maximum loglikely hood will be chosen.

|Loglik("lognormal")|<|loglik("Weibul"|

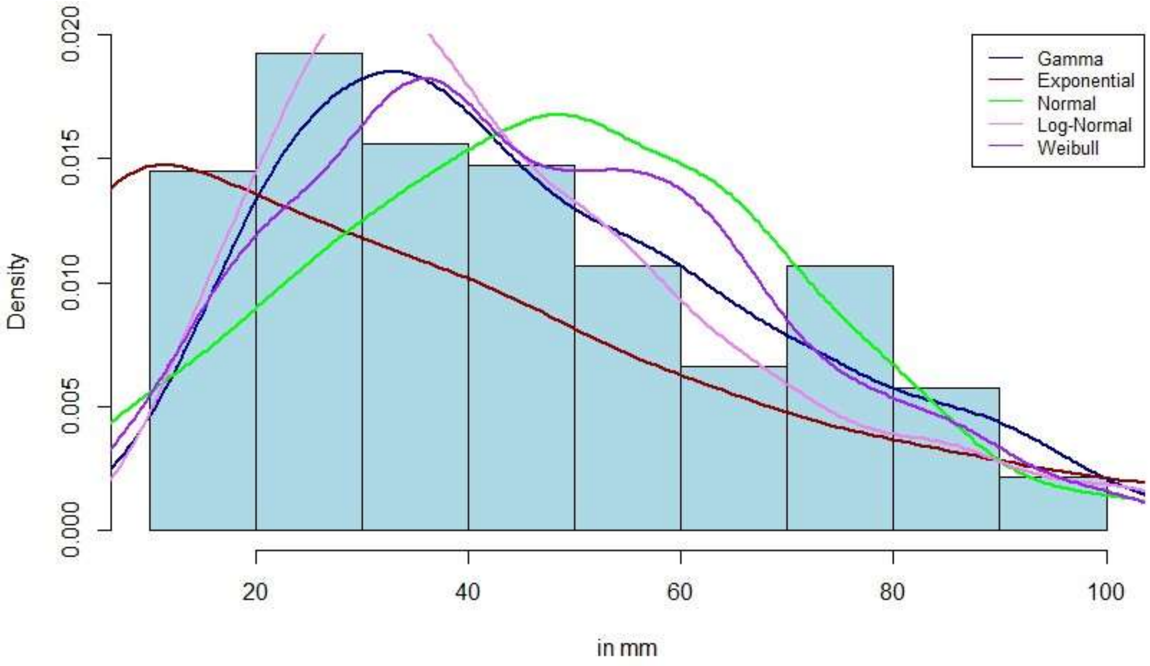
And so we decided that doubling time follows a lognormal distribution

C. Tumour size of unscreened patients

We have tried and approximate 5 distributions to the doubling time distribution. After calculating the parameter of each distribution, here is what we have got:

- Gamma(α=3.64,β=0.078)
- Exponential(λ=0.021)
- Normal distribution : N(46.316, 23.023)
- Log Normal distribution : LN(3.69, 0.56)
- Weibull distribution: Weibull(k= 2.149,,λ= 52.45)

For visualization purposes, we have plotted all the 5 distributions, on top of the histogram of the tumour size of unscreened patients. This is a good way to visually perceive which distribution does the tumour size of unscreened patients approximately follow. (Doc 64)



Doc 64: Which distribution does the tumour size of unscreened patient follows?

1. Gamma(α=0 3.64,β=0.078)

We will do some hypothesis testing once again

Ho: The tumour size follows a Gamma(α=3.64,β= 0.078)

H1: The tumour does not follow a Gamma(α=3.64,β= 0.078)

We have done a test called kolmogorov-smirnov. pvalue =7.772e-16<0.05. With that in mind, we can reject Ho on a 95% confidence level. And so the tumour size of unscreened patients does not follow a gamma distribution

2. Weibull(k= 2.14, λ = 52.45)

We will do some hypothesis testing once again

Ho: The tumour size follows a Weibul (k=2.14, λ = 52.45)

H1: The tumour does not follow a Weibul (k=2.14, λ = 52.45)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.01≈0.05. Another test will be done called Anderson-Darling. Pvalue=0.036≈0.05 With that in mind, we do not reject Ho on a 95% confidence level. And so the tumour size of unscreened patient does follow a Weibull distribution.

3. Exponential(λ= 0.021)

We will do some hypothesis testing once again

Ho: The tumour size follows a Exponential (λ= 0.021)

H1: The tumour does not follow a Exponential (λ= 0.021)

We have done a test called kolmogorov-smirnov. pvalue <2.2e-16<0.05. With that in mind, we can reject Ho on a 95% confidence level. And so the tumour size of unscreened patients does not follow an exponential distribution.

4. Lognormal: LN(3.69, 0.56)

We will do some hypothesis testing once again

Ho: The tumour size follows a Lognormal: LN(3.69, 0.56)

H1: The tumour does not follow a Lognormal: LN(3.69, 0.56)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.003 <0.05. With that in mind, we do reject Ho on a 95% confidence level. And so the doubling time follow a lognormal distribution

5. Normal: N(46.316, 23.02)

We will do some hypothesis testing once again

Ho: The tumour size follows a Normal: N(46.316,23.02)

H1: The tumour does not follow a Normal: N(46.316,23.02)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.00012 <0.05. With that in mind, we do reject Ho on a 95% confidence level. And so the tumour size of unscreened people does not follow a normal distribution.

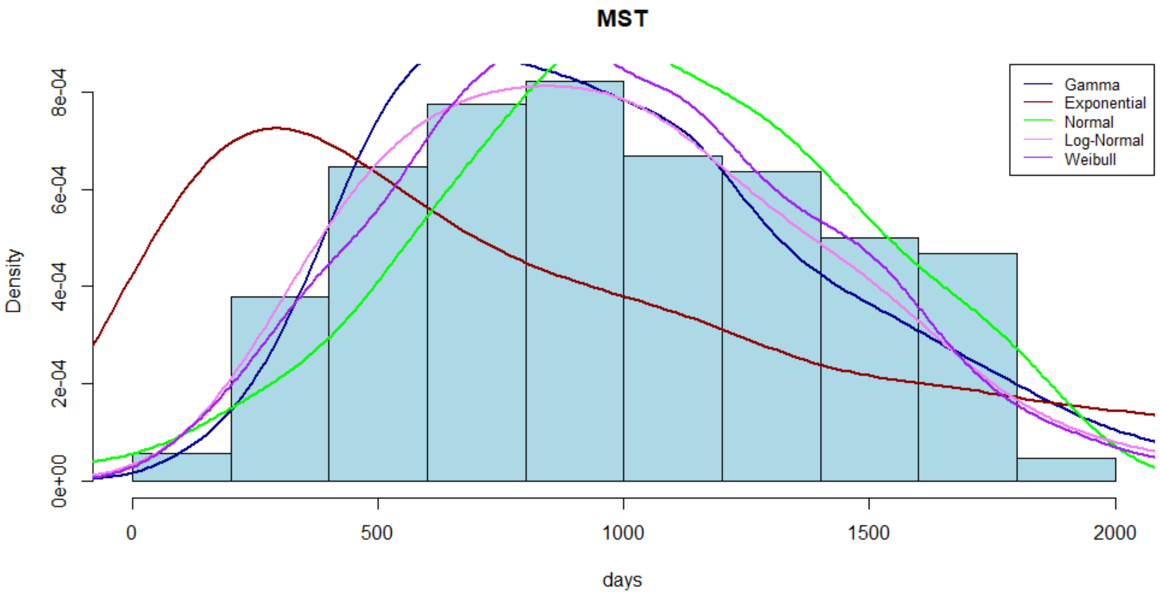
D. MST distribution

Now, it is time to check for the distribution of MST. In fact, thanks to the formula above, and since we have the distribution of every element in our equation, the project is now feasible.

We have tried and approximate 5 distributions to the MST distibution. After calculating the parameter of each distribution, here is what we have got:

- Gamma(α= 4.40,β=0.004)
- Exponential(λ= 1.011e-03)
- Normal distribution : N(988.41, 429.647)
- Log Normal distribution : LN(6.77, 0.52391550)
- Weibull distribution: Weibull(k= 2.49,λ= 1.117e+03)

For visualization purposes, we have plotted all the 5 distributions, on top of the histogram of the tumour size of unscreened patients. This is a good way to visually perceive which distribution does the MST follows.



1. Gamma(α=4.40,β=0.004)

We will do some hypothesis testing once again

Ho: The tumour size follows a Gamma(α=4.4,β= 0.004)

H1: The tumour does not follow a Gamma(α=4.4,β= 0.004)

We have done a test called kolmogorov-smirnov. pvalue = 0.02≈0.05. We have done another test and gave us pvalue=0.006 <0.05. With that in mind, we do not reject Ho on a 95% confidence level. And so the MST does not follow a gamma distribution.

2. Weibull(k=2.39 ,λ= 1.117e+3)

We will do some hypothesis testing once again

Ho: The tumour size follows a Weibull(k=2.49, λ = 1.117e+3)

H1: The tumour does not follow a Weibull(α=2.39, λ = 1.117e+3)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.33>0.05. Another test will be done called Anderson-Darling. Pvalue=0.07>0.05 With that in mind, we do not reject Ho on a 95% confidence level. And so the MST does follow a Weibull distribution.

3. Exponential(λ= 1.101e-3)

We will do some hypothesis testing once again

Ho: The tumour size follows a Exponential (λ=1.101e-3)

H1: The tumour does not follow a Exponential (λ= 1.101e-3)

We have done a test called Kolmogorov-Smirnov. Pvalue<2.2e-16<0.05. With that in mind, we can reject Ho on a 95% confidence level. And so the MST does not follow an exponential distribution.

4. Lognormal: LN(6.77, 0.52)

We will do some hypothesis testing once again

Ho: The tumour size follows a Lognormal: LN(6.77, 0.52)

H1: The tumour does not follow a Lognormal: LN(6.77, 0.52)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.5>0.05. Another test will be done called Anderson-Darling. Pvalue=0.3>0.05. With that in mind, we do not reject Ho on a 95% confidence level. And so the MST does follow a log normal distribution.

5. Normal: N(988.4123.02)

We will do some hypothesis testing once again

Ho: The tumour size follows a Normal: N(988.41, 429.647)

H1: The tumour does not follow a Normal: N(988.41, 429.647)

We have done a test called Kolmogorov-Smirnov. pvalue = 0.15 <0.05. Another test will be done called Anderson-Darling. Pvalue=0.01<0.05 With that in mind, we do reject Ho on a 95% confidence level. And so the MST does follow a log normal distribution.

E. Calculating the mean sojourn time

Since the MST follows a Weibull distribution, all we need to do is to calculate the mean of the Weibull distribution that we have got. What we have done is computed 1000 random values of our Weibull distribution. The value gotten is 986 days. So we can say that each 986, a patient should be screened. I would say, for safety issues, we choose 900 days. We also computed the interval confidence. The real mean would be somewhat between [953.49;1006.5511].

V. Conclusion

There is so much to discuss in the data that we have. We discovered a lot of new things just by analysing a relatively large data set. We now know what factors influence the death and survival in days. In fact, we found out that the age group, Gender, diffuse cancer, Metastatic cancer, curative treatment, HCV, Alcohol, Criteria and cancer stages impact the variable death. Also, the screening, Thrombosis, Diffuse Cancer, Metastatic, Curative, HCV, alcohol, Cancer stages, tumour size and criteria impact the variable survival in days. In fact, we saw how certain factors such as drinking alcohol or having

HCV or Thrombosis, negatively impact our chance of surviving the liver cancer or even, surviving more days. What is even surprising is that the operation of diffusing cancer, as we saw earlier, negatively impacted the survival in days of a person. Can we really say that is a bad operation to do? Maye if it is you only hope for winning the cancer battle, then of course, go for it.

Also, I wanted to mention how smoking did not impact the survival in days or the death variable. Smokers right around the globe, there is no reason to quit … Just kidding!

Furthermore we saw how important screening is. But also the curative treatment had a positive impact on the survival in days and death variable.

Something to note, is that the treatment 1 to 6 are really useless. They do not affect both variable (survival in days and death), and it really makes me think about why they are even there. Is it true that the hospitals, and the capitalism system is benefiting from us? Have cancer really become a profit for the government? We cannot really confirm. But it would be an interesting subject to dive into.

We also found which distribution the survival in days follows. It turned out that it follows a Weibull distribution. We found the distribution of the tumour size, which happens to be a log normal distribution. Now we can really see how both variables behave… Of course in the case of patients in the liver cancer.

Finally, I could not stress more about how important screening is. That is why, we must know how long we should wait in order to do a screening test. That's what we have done in the last part, where we identified the distribution of the MST. Having known which distribution it follows, we calculated the mean sojourn time, which will help us engineer a screening program for the patient at high risk of having the liver cancer.