Final Exam

Louis Bensard

May 14, 2018

Problem 1:

The data we have is pretty straightforward, we have 18 patient suffering from arthritis and they are experimenting a new anti-arthritic supplement (GNO44). They rank their pain level on a scale (0 to 10) at the start of the trial and then every 3 months for a year. First, when I look quickly at the data, I can clearly notice a drop in the arthritic pain ranked by the patients over time. Now let's analyse this data more closely.

This type of data is called "Ordinal Data". This type a data is really subjective and very patient dependent, therefore the data is not suitable for a parametric test. This is why we are going to use a Kruskal-Wallis test.

This test is more flexible and does not need the typical assumptions we would need for a parametric model. We basically need independence of observations within each patient's data and across patients, plus each patient's data need to be identicaly distributed. Those assumptions are easily assumed to be met and thus we can proceed with our test. Notice that we don't need the normality assumption, which makes this test very versatile and useful for ordinal data.

From the data, along with the gender, we basically have 5 column of values for each time of the year mentioned above, we will say that each of these column of value belong to a "subpopulation". A Kruskal-Wallis test is basically a hypothesis test with null hypothesis being that the cumulative distributions of those subpopulations are all equal (versus at least two of them being different). A Kruskal-Wallis test will tell us if the differences between the groups are so large that they are unlikely to have occurred by chance.

This might seem a little technical but basiclly, this Kruskal-Wallis test is going to test whether the new treatment significantly decreases the arthritic pain of the patients over time.

Here is the result of this test:

```
print(p_val)
```

```
## [1] 2.4934e-05
```

This p-value is very low and below 0.05, therefore the test found enough evidence to reject the null hypothesis and thus we can conlcude that the new treatment does significantly decreases the pain the patients go through over time.

Now let's find out whether this effect is similar if we focus on one gender at a time. If we analyse only the data collected from males, we get:

```
print(p_valM)
```

```
## [1] 0.10495
```

Hmm, this p-value is more than 0.05, it looks like males are having a hard time feeling the benefits of this new treatment on their arthritic pain.

Let's now focus on data collected from women:

```
print(p_valF)
```

```
## [1] 5.0308e-05
```

That p-value is less that 0.05, thus it seems that the women in that study are significantly feeling the benefit of the new treatment over time.

But how do we know whether the new treatment works better on women or if women just hanle the pain better than men?

A modification of the study might help, let's jump to part (b) and see how we can improve it.

(b)

The first problem I noticed is that there is no control. How do we know if the pain is not just going away as the time passes by? We need to record the pain over time of males and females who do not receive the new treatment so we can compare the results of the treatment group to something. Also, a control group would be useful to assess whether women feel relatively less pain than men in identical scenarios. This is in my opinion critical to interpret the results of the Kruskal-Wallis test.

Secondly, I think more data would help stengthen the significance of this study, 18 people is not enough and is hardly representative of a whole population. I would gather more data using two different ways: First get more people into the study and second gather more data about the patient themselves. I think on top of gender, we should at least know the gge, the medical background of the patient and the gravity of their arthritis, so we can isolate what part of the population the drug is the most efficient on and what part it is useless on.

Lastly, it would be good (if possible) to have a objective scientific way of mesuring the evolution of the arthritis of each patients. Scientific measurment are easier to analyse and more reliable than the pain felt by a specific person at a specific moment.

This last improvement might be hard to obtain but I am sure than the first two improvements can be relatively easy to implement.

Problem 2:

We interested in reporting on the effects of zalcitabine (ddC) vs. didanosine (ddI) in terms of prolonging life for a patient diagnosed with AIDS. First, out of the data, we can compare the mortality ratios of people in the ddC group as opposed to the ddI group.

```
print(prop_ddc_dead); print(prop_ddi_dead)
```

```
## [1] 0.37131
```

```
## [1] 0.43478
```

Thus, we found that 37% of people in group ddC died against 43% of death for group ddI. The ddC drug seems slightly slightly better but we cannot tell whether it significantly extends the life of patients yet.

To do so, we are going to perform a Survival Analysis by converting the data into a Kaplan-Meier table. The particularity of this data structure is that every time someone dies or leaves the study, a new data point is created. Here is how it works, below is the first 6 rows of that table:

```
##   time  n1  n2 d1 d2   prop.d1   prop.d2  s.hat1  s.hat2
## 1 0.47 237 230  0  1 0.0000000 0.0043478 1.00000 0.99565
## 2 0.77 237 229  1  0 0.0042194 0.0000000 0.99578 0.99565
## 3 0.90 236 229  0  1 0.0000000 0.0043668 0.99578 0.99130
## 4 1.03 236 228  1  0 0.0042373 0.0000000 0.99156 0.99130
## 5 1.07 235 228  2  0 0.0085106 0.0000000 0.98312 0.99130
## 6 1.13 233 228  0  1 0.0000000 0.0043860 0.98312 0.98696
```

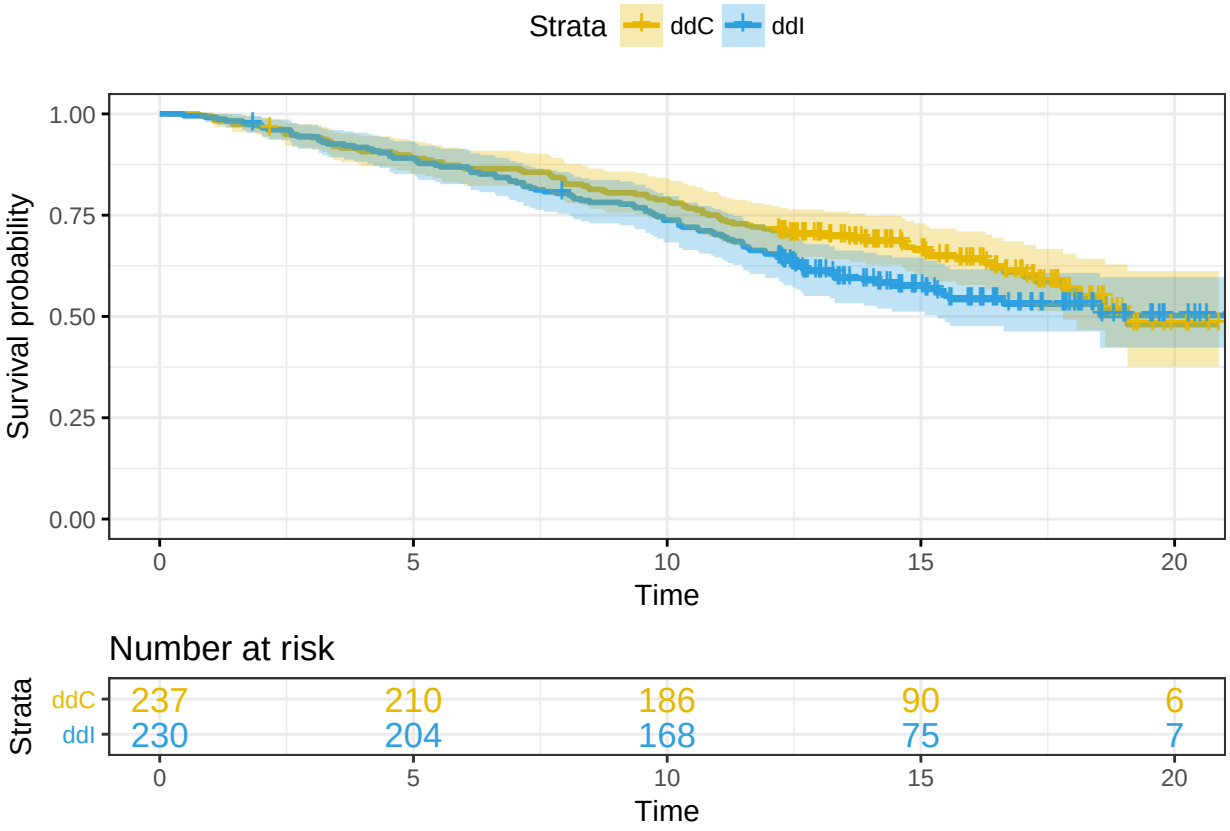
The time is assumed to be in months, d1 and d2 are respectiveley the number of death for group ddC and ddI at each given time, and n1 and n2 the number of patient left in the study (who died or left) for each group.

We can see that out of these 5 first column, we can compute some kind of "instantaneous probability of dying". For instance, 1 patient out of 237 using the ddC drug died at time 0.77, thus the proportion of death in group ddC are about 0.4% (in that case the "instantaneous probability of living" would be 1-0.004 = 99.6%). You can easily convince yourself that this proportion at this exact moment cannot be interpreted nor used to make predictions. If so, someone using drug ddC would have 0.4% of dying at time 0.77 but 0% chance of dying at time 0.90, this is non-sensical.

However, we can compute what we call "cumulative survival probabilities" for each group. Basically, we start we $\hat{s} = 1$ and for each time, we multiply the previous $\hat{s}$ value by the "instantaneous probability of living" we talked about earlier. Let's follow what happens for s.hat2 in the table.

At time 0 we have s.hat2 = 1, then at time 0.47 someone dies, which gives us a proportion of surviving for that specific time of 1 - 0.0043 = 99.57%. Multiply this number by the previous s.hat2 (which is 1) and you get s.hat2 for time 0.47. Apply the same process over and over multiplying by the very last s.hat2 each time.

This cumulative survival probability can be easily interpreted as the probability for a patient using that drug to survive t months. Logically, the higher the $\hat{s}$, at a given time, the better. Let's plot the curves of s.hat1 and s.hat2 against time to see if one dominates the other.



Hmm, we can't see a real difference between those two curves, hard to tell whether one drug is significantly extenting life of patients.

(Side note: All those "+" on the graph are people who left the study, we can see that a lot of people left after a year in the study, this is a problem, the company in charge of those studies should consider giving a more benefits (ie $$$) to people who stay in the study).

Now let's go back to our analysis and compute a Log-Rank hypothesis test and see if the resulting p-value helps us differenciate those drugs. This hypothesis test has for null hypothesis s.hat1 = s.hat2. To perform this test, we cleaverly use the elements of the Kaplan-Meier table to set up a Maentel Haenzel χ^2 test.

```
print(p_value)
```

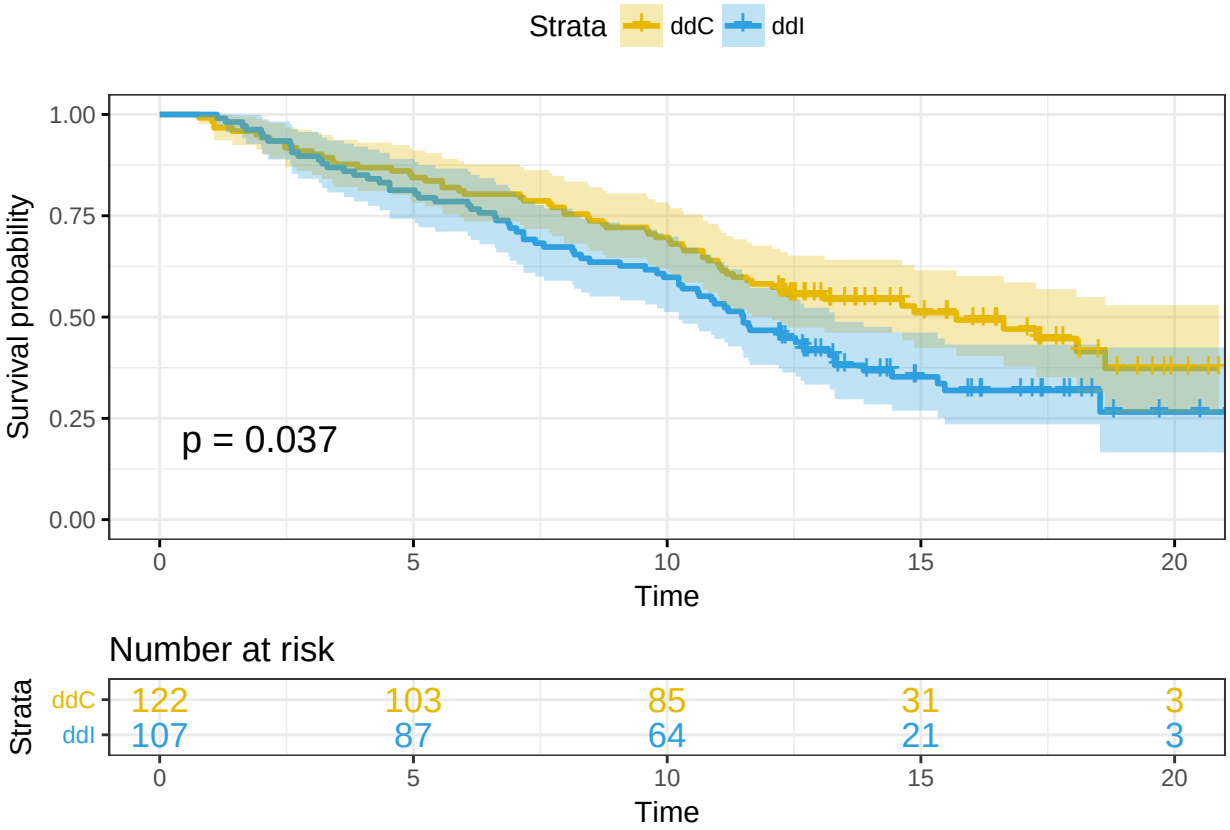
```
## [1] 0.17091
```

This p_value is really high, there is not enough evidence to reject the null hypothesis, one drug does not have any significant impact on prolonging the life of patients diagnosed with AIDS relative to the other drug. The p-value has the last word this time.... Or does it?!

Let's do our last stand and pull out the last piece of data that we have: the White Blood Cell count at initial observation. What we are going to do is very simple, we are going to split the inital data into 2 group. One group with relatively low WBC count (<6) and the other with relatively high WBC count (≥ 6). Then we are going to perform the same test as before with each new data set.

What we did with the Log Rank test was really fun but there is actually a "p-value" agrument in the plot I used previously (from the package surviminer).

Here is the new plot of the cumulative survival probabilities for the low WBC count data set:



That is an interesting result, this p-value of 0.037 is less than 0.05 therefore we can reject the null hypothesis and concluse that the ddC drug does extend the life of patients diagnosed with AIDS, relative to the ddI drug. (Of course I double checked this p-value performing another Log Rank test in the background).

On the other hand, for the high WBC count data set, we have a very high p-value of 0.67, thus none of these two drugs have better performance than the other for this specific subgroup of patients.

Conclusion:

At first sight, the data did not show any significant difference between the effect of the ddC and ddI drugs. A plot of the cumulative survival probability curves against time did not help us either and a Log-Rank test resulted in a non-significance of any drug relative to the other for extending the life of those patients. However, after focusing on patients with low White Blood Cell count, we found that the ddC is actually significantly extending the life of patients diagnosed with AIDS, relative to the ddI drug.

Problem 3:

(a)

First, let's try to use the data very simply to have an idea and an overview of the impact that Mono/HIV can have on the death prior ro 5 years of Hodgkins diagnosis.

The variable of interest (alive5yr and HIV.mono) are both binary variable, so a plot would not help a lot to assess the situation. Instead, only for patients who have not had Mono or HIV in the past, I am going to compute the proportion of death over total number of those patients. Then, only for patients who have had Mono or HIV in the past, I am going to compute the proportion of death over total number of those patients. And then compare those two proportions.

```
print(prop_no_mono); print(prop_mono)
```

```
## [1] 0.18553
```

```
## [1] 0.54167
```

Therefore, only 19% of patients who have not had Mono of HIV in the past died within 5 years of a Hodgkin's diagnosis, against 54% for patient who have had Mono or HIV.

Those are very dry numbers but now we have an idea of the kind of impact Mono/HIV can have on the death of patients.

The data provided has a categorical response and a mix of categorical and continuous predictors, this is why I want to use Logistic Regression to fit my data. I started with a linear model inside the glm() function containing all the predictors and with response variable being my variable *Alive5yr*. Then I used the *car* package to analyse the residual plots of my model. All the errors were nicely spread around 0, therefore I conclude that I have the right model:

$$\text{model} = \text{glm}(\text{Alive5yr} \sim \text{Stage} + \text{Age} + \text{RBC} + \text{Gender} + \text{HIV.mono} + \text{WBC}, \text{family="binomial"})$$

Now before doing anything with my model, I want to make sure that I am not overfitting by using unsignificant predictors. To search for possible unsignificant predictors, I used the Wald test. This test perform for each regression coefficient, an hypothesis test with the null hypothesis being that this coefficient is equal to 0. I used the regTermTest() function in the *survey* package. Here is a summary of the results of this Wald test (inspired by Anna's midterm):

Only the Gender has a p-value less than 0.05, therefore, as expected, the Wald test did not find enough evidence to reject the null hypothesis and thus Gender is unsignificant in my model.

Our model is now :

$$\text{model} = \text{glm}(\text{Alive5yr} \sim \text{Stage} + \text{Age} + \text{RBC} + \text{HIV.mono} + \text{WBC}, \text{family="binomial"})$$

Predictor	P-value
Stage	0.0039
Age	0.0013
RBC	0.00036
Gender	0.43
HIV.mono	0.019
WBC	5.13e-10

Table 1: Summary of Wald test results

Our previous model had an AIC of 200.7 and this new model has and AIC of 199.3, the lower the better, thus we improved our model by dropping the Gender predictor.

This model is going to help up determine how more likely a patient who has had Mono/HIV in the past is to die prior to 5 years of a Hodgkin's diagnosis.

Using the regression coefficient of the predictor HIV.mono and its standard error, I am able to find a 95% confidence interval of the Odds Ratio of dying withing 5 years of patients who have vs patients who have not had HIV or Mono, with the other predictors being held constant.

This interval is [1.27771, 12.85362]. This is a wide range but since our confidence interval stays above one, we can say that having had Mono or HIV increases the risk of death prior to 5 years of a hodgkin's diagnosis.

More precisely, we are 95% confident that a patient that has had Mono or has test positive for HIV is between 28% and 1185% more likely to die within 5 years of a Hodgkin's diagnosis. I've had Mono before, so I am crossing fingers to never get diagnosed with Hodgkin's disease!

(b)

We are going to use the same method we used previously to find out how much more at risk an individual diagnosed with stage IV Hodgkins is, as opposed to someone diagnosed with stage I Hodgkins, with the other predictors being held constant.

The corresponding 95% confidence interval is : [2.04572, 17.11378]. Again, the interval is above 1 so we are confident to say that a patient diagnosed with stage IV are more likely to die within 5 years of Hodgkin's diagnosis than patient on stage I.

More precisely, we are 95% confident that a patient diagnosed with stage IV is between 104% and 1611% more likely to die within 5 years of a Hodgkin's diagnosis.

This result makes sense as Stage IV is the most progressed stage of the disease.

(c)

The function predict() applied to our model with the paramter values provided is going to help us help this gentleman. Here is the probability predicted by my model that this man will die within 5 years.

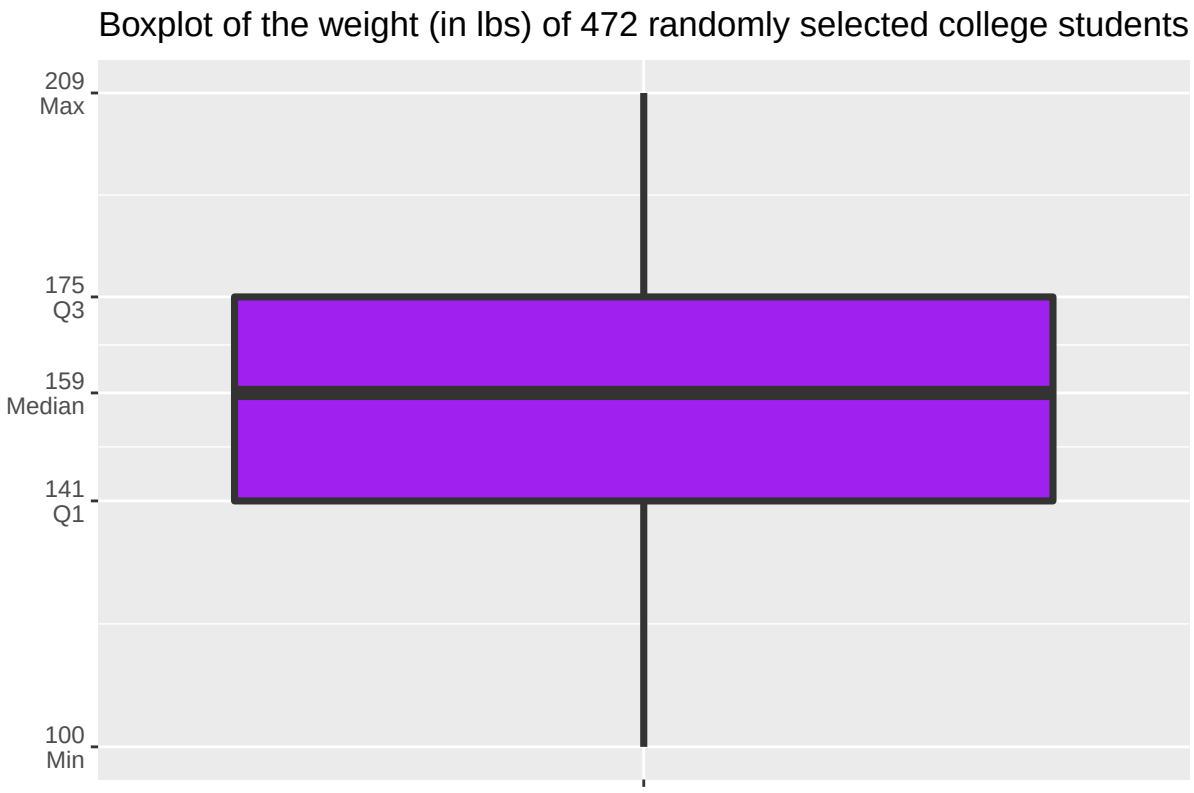
```
##       1
## 0.46314
```

Not very promising.. According to my model, this man has 54% chance of being alive 5 years from now. Having the result of part (a) in mind , I would advise this man to not catch Mono or HIV and wish him the best of luck for his coin flip!

Problem 4:

(a)

Here is a boxplot that summarizes the weight (in lbs) of 472 randomly selected College students:



The minimum weight recorded is 100 lbs, the max is 209 lbs, 25% of college students are below 159 lbs, 50% are below 158.7 lbs and 75% of college students are below 175 lb.

(b)

We are interested in creating a 95% confidence interval for the boxplot we created. A box plot is basically defined by 5 parameters: Min, 1st Quartile (Q1), Median, 3rd Quartile (Q3) and Max. What we can do is come up with a 95% confidence interval for all those parameters and build a 95% confidence interval boxplot from there.

Since we do not have any information about the population this data is coming from (no mean, standard deviation or distribution shape), and we cannot really go out and get thousands of sample of 400 college students to estimate those parameters, we are going to perform a boostrapping precedure in order to come up with a 95% confidence interval for the 5 parameter mentioned above.

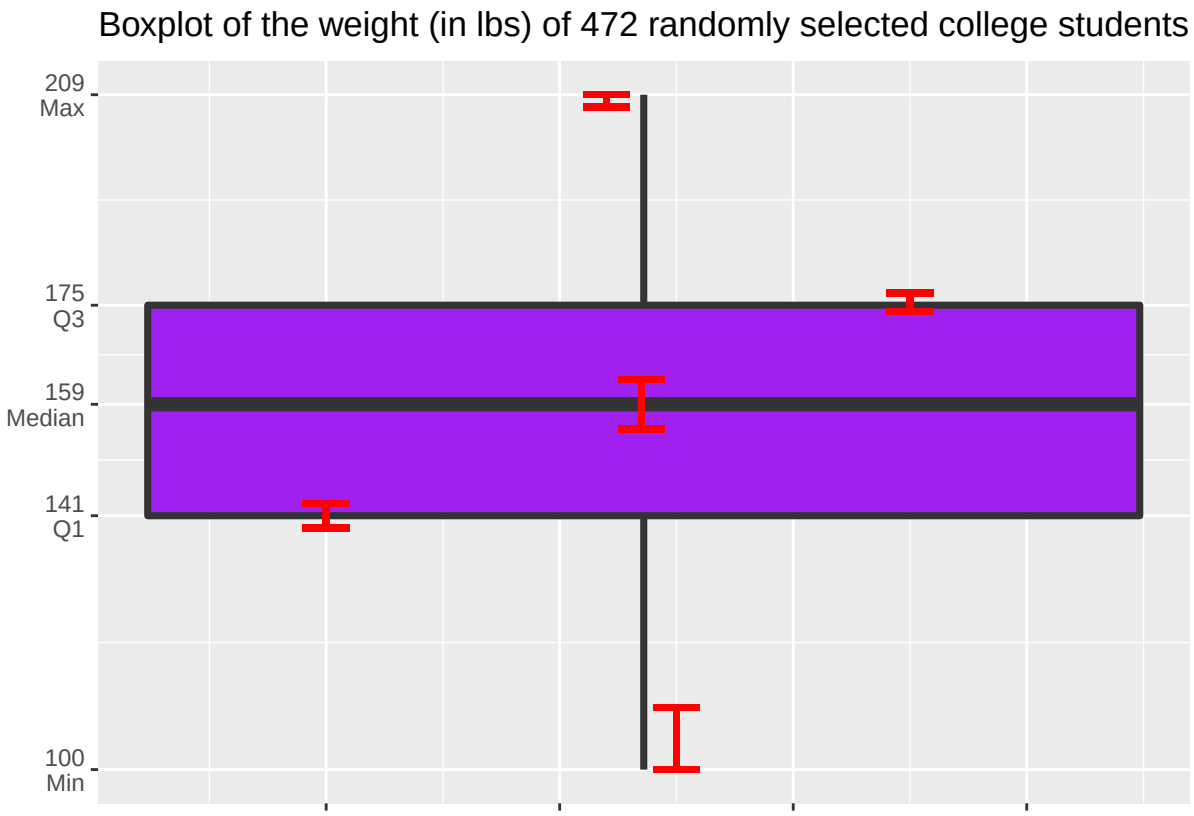
This means that instead of taking N sample from our population of college students, we are going to N samples of size $n = 472$ from our original sample, with replacement. We will make sure to store each parameter of each of those samples inside five distinct vectors at each iteration. Finally, our confidence intervals will be [2.5 %tile, 97.5 %tile] of the sorted resulting vectors of parameters.

Here are the 95% confidence interval of interest:

$$CI_{min} = [100, 115]$$
$$CI_{Q1} = [138.75, 143]$$
$$CI_{med} = [155, 163]$$
$$CI_{Q3} = [174, 177]$$
$$CI_{max} = [207, 209]$$

Therefore, we are 95% confident that each **population** parameter mentionned above lies within its corresponding confidence interval.

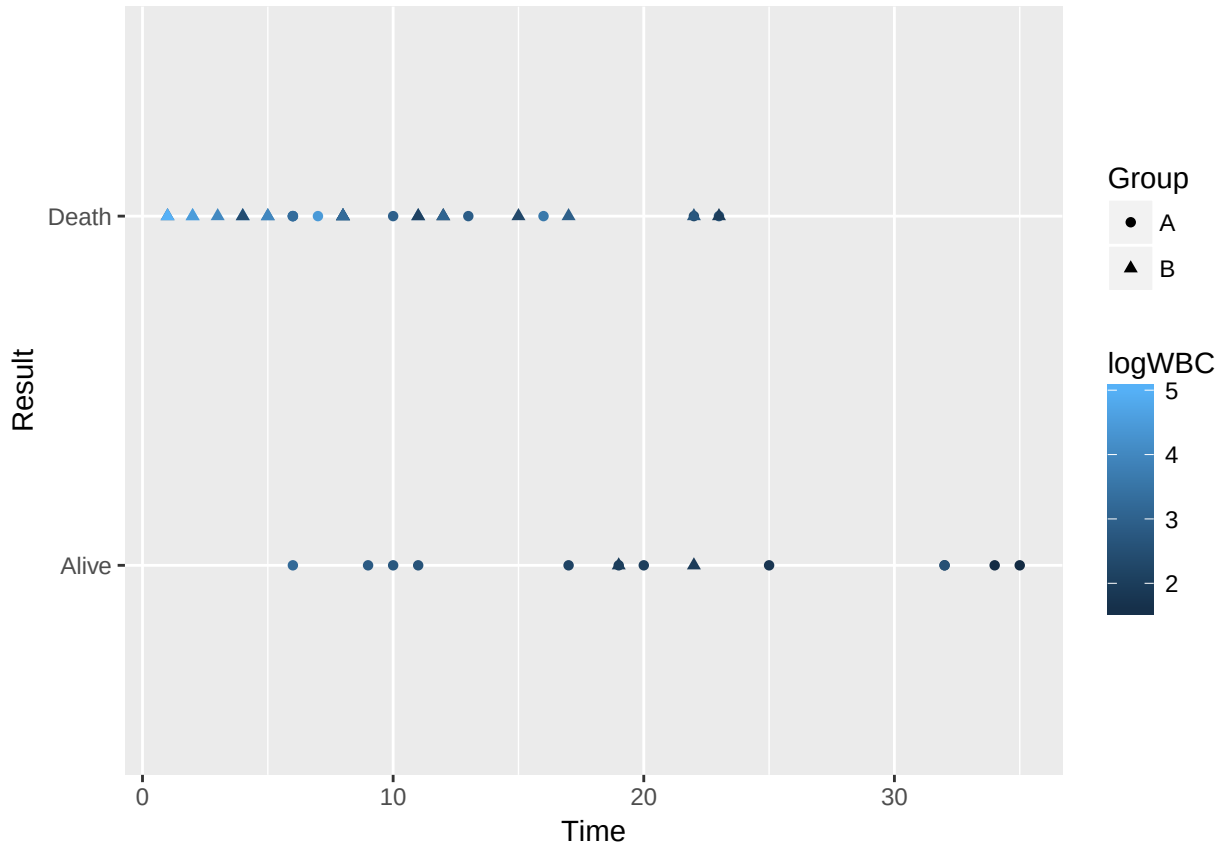
Here is the same box plot as before but we added the confidence intervals of each parameter in red. In other words, we are 95% confident that if we knew the Min, Q1, Median, Q3 and Max weight of the whole population of college students, and we drew the corresponding population boxplot, the resulting boxplot would fall within those red segments.



Problem 5:

We are provided time to event data for 44 patients who were diagnosed with Leukemia. 21 of them received a new mediacal treatment (Group A: Treatment) and 23 of them received a traditional medical treatment (Group B: Control). We want to determine weather there is a benefit to being placed in the Treatment group as opposed to the Control group.

First, we plot the data in a relevant way to have an overview of the possible effects of the group placement:



We notice clearly that the Treatment group contains very few deaths while most of the patient in the Control group died. This plot also shows than the higher your white blood cells count at time of diagnosis, the sooner you will die.

Also, we can note that even though we assume that patient are assigned to groups at random, the number of patients having high white blood cell count seems higher for group B, this is only a small difference but more data would strengthen the following analysis.

Even though it is useful to get a big picture of the data, this plot is not enough to determine whether there is a real benefit to being placed in one group or another.

I am going to compute a Survival Analysis on the mortality rates. Let ID_A and ID_B be respectively the mortality rate of group A and group B. Let's perform the following one-sided hypothesis test and determine wether there is a significant improvement between the Treatment mortality rate and the Control mortality rate:

$$\begin{aligned} H_0: &\quad ID_A = ID_B \\ H_A: &\quad ID_A < ID_B \end{aligned}$$

Here is the p-value obtained:

```
print(pval)
```

```
## [1] 0.00036
```

$p - value = 3.60002 \times 10^{-4} < 0.05$ thus we conclue that the results are significants and thus we reject the null hypothesis. As a result, there is evidence that the mortality rate of the Treatment group is lower than the mortality rate of the Control group.

However, the data provides another ususeful information: the natural log of the white blood cell counts of individuals at time of initial diagnosis. What if the new Treatment if effective on patients having a relatively low white blood cell count (the lower the better) but does not work that well on patients having a high white blood cell count?

Let's arbitrarily split each group into two subgroups: High WBC count (patients with $logWBC > 3.5$) and Low WBC count (patients with $logWBC \leq 3.5$) . The analysis goes the same way as described above but we focus on one subgroup at a time.

High WBC count:

$$\begin{aligned} H_0: &\quad ID_{Ahigh} = ID_{Bhigh} \\ H_A: &\quad ID_{Ahigh} < ID_{Bhigh} \end{aligned}$$

Here is the p-value obtained:

```
print(pvalh)
```

```
## [1] 0.11777
```

The data is not that significant anymore, this p-value is high, we fail to reject the null hypothesis. If we focus on high white blood cell count patients, there is no evidence that the new treatment is better than the traditional one. However, considering there is only 10 high level patients in both groups combined, I would not be surprised to see this result change as we gather more data.

Low WBC count:

$$\begin{aligned} H_0: &\quad ID_{Alow} = ID_{Blow} \\ H_A: &\quad ID_{Alow} < ID_{Blow} \end{aligned}$$

Here is the p-value obtained:

```
print(pvall)
```

```
## [1] 0.0027918
```

Since this p-vale is less than 0.05, we reject the null hypothesis. Therefore, for low white blood cell count patients there is evidence that the new treatment is better than the traditional one.

Conclusion:

From the start when we plotted the data, we saw that the new treatment might decrease the mortality rate for patient diagnosed with Leukemia. The survival analysis performed between the Treatment group and the Control group confirmed this first thought and we proved that there is strong evidence the new treatment decreased the mortality rate. However, a further analysis showed that this evidence is only valid for patient having relatively low white blood cell count when they are diagnosed with Leukemia. I would suggest the company who ordered those test to gather more data to see more clearly whether this new treatment can be efficient on high white blodd cell count patients. But for now, I will conclude that this new treatment only benefits patients having a relative low white blood cell count when they are diagnosed with Leukemia.