Survival Analysis Project

1 Introduction

The goal of this project is to use Survival Analysis techniques to determine how long it takes for a DSTI student to get an internship. The below questions should be answered.

How long does it take for a DSTI student to start an internship?

1) How long does it take in general?
2) Is the situation changing over cohorts?
3) Does the educational background have an impact?
4) Build a predictive model to identify students at high risk of a long search.

2 Survival Analysis techniques

A brief description follows of each of the survival analysis techniques that will be employed to do the analysis.

2.1 Non-Parametric Survival Analysis

Non-parametric tests do not assume anything about the underlying distribution. They are best used on right-skewed data with censored events. As we have many censored events in our dataset these tests are entirely suitable.

1) Kaplan-Meier

The most popular estimation method for non-parametric survival analysis is Kaplan-Meier. The Kaplan-Meier curve is an estimator that is used to estimate the survival function. The curve is a visual representation of the survival function that shows the probability of an event over a specific time interval. The curve is a step plot.

2) Nelson-Aalen

The Nelson-Aalen estimator is an estimator of the cumulative hazard rate function. It is used to estimate the cumulative number of expected events within a certain period of time.

3) Logrank

The Logrank test allows one to compare the survival distribution of two groups or more. For example, in a clinical trial this test could answer the question: Which treatment has a better survival probability? This is also a non-parametric test.

The Logrank test compares the hazard functions of the two groups at each observed event time.

1. It computes the observed and expected number of events in a group at each observed event time.
2. These results are then summed to get a total across all time points where there was an event.

The null hypothesis is that the hazard rate in each group is similar. It is rejected if the test statistic of Z is not normally distributed.

2.2 Semi-Parametric Survival Analysis

1) Cox Proportional Hazards Model

This model allows one to evaluate simultaneously the effect of several covariates on survival. It enables one to analyse how certain covariates influence the rate of a particular event occurring at a particular point in time. The rate is more commonly known as the hazard rate.

The formula is the following:

$$h_i(t) = h_0(t)\exp(x_i'\beta)$$

$h_i(t)$: hazard for subject i at time t

h_0: baseline hazard function

x_i: vector of covariates for subject i

β: vector of effects of each covariate on risk

Given an observed dataset one can estimate beta without having to specify the baseline hazard h_0. It is therefore a semi-parametric model. Due to censoring the classical likelihood cannot be calculated so a partial likelihood is calculated instead which takes censoring times into account similar to Kaplan-Meier.

3 Data Clean-up

The data for the project was provided in a csv file named DSTI_survey.csv and contained 82 rows and 13 columns. However, the data was not clean.

Firstly, there were 18 rows where no internship start or end dates were specified. With both of these missing it is not possible to determine how long it would take for these students to get an internship. Therefore, these rows were removed from the dataset. → 18 rows removed

Secondly, there was a row where the start date was in 1980. This was assumed to be a rogue row as the school did not even exist in 1980, therefore this row was also removed. → 1 row removed

Then there were 8 rows where the internship start date was greater than the survey date. These rows were definitely wrong and therefore removed. → 8 rows removed

Finally, there were 3 rows where an internship start and end date were specified but the internship found column was 'No'. This did not make any sense to me as getting an internship is an integral part of the course so it didn't make sense that a student specified an end date and then set the internship found column to No. So, these rows were also removed. → 3 rows removed

Total number of rows removed → 30 rows

No. of rows remaining → 52 rows

In addition to the above the following changes were also made:

- Some renaming was done to make the column headers names and column entries shorter and more readable e.g. shortening educational background as the names were too long (using method demonstrated in class).
- The timestamps were changed to POSIX ct timestamps.

- A new column for time to internship, tti was calculated: in the following way: *calculated in days using the difference between the int_start and int_stop dates when an internship was found or the survey date minus int_start date if no internship was found.*
- The 'internship found' column was renamed to int_found and changed to a boolean, 0 for censored events and 1 for the event occurring.
- A new column 'age' was calculated as: survey date – year of birth (demonstrated in class)
- The smoking columns were dropped as it was deemed that these columns were used to teach survival analysis in classes rather than to be a contributing factor for getting an internship.

The dataset after the clean-up is as below:

age <dbl>	sex <fctr>	education <fctr>	cohort <fctr>	int_start <S3: POSIXct>	int_stop <S3: POSIXct>	int_found <dbl>	edu_years <dbl>	children <chr>	tti <dbl>
28	Male	math	A20	2020-11-02	<NA>	0	20	No	1
27	Female	math	A20	2020-10-19	<NA>	0	17	No	15
34	Male	bio	S20	2020-09-01	2020-10-31	1	22	Yes	60
27	Male	math	A18	2018-11-01	2018-12-31	1	16	No	60
28	Female	lit	A19	2020-03-01	2020-07-01	1	18	No	122
25	Male	fin	A19	2019-10-01	2020-02-01	1	16	No	123
31	Male	math	S20	2020-11-02	<NA>	0	18	No	1
38	Female	math	A20	2020-10-29	<NA>	0	20	No	5
23	Female	math	A20	2020-11-02	<NA>	0	20	No	1
50	Male	math	S20	2020-08-20	2020-11-01	1	20	Yes	73

1-10 of 52 rows Previous 1 2 3 4 5 6 Next

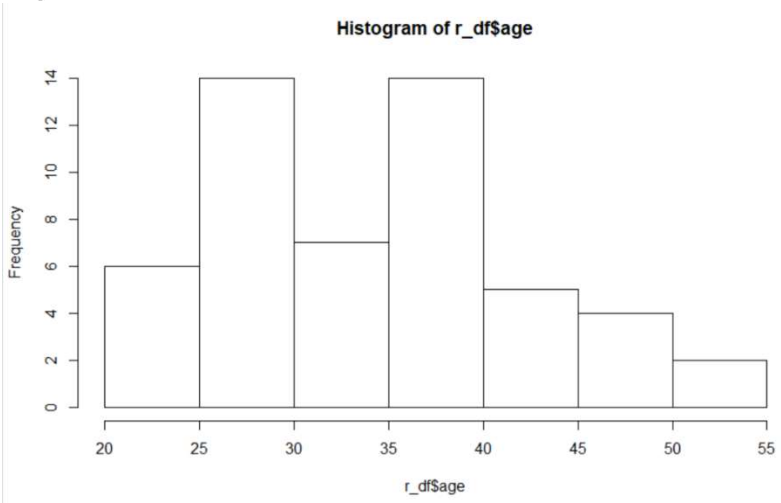
4 Exploratory Data Analysis

Let's first do some exploratory data analysis of the variables to understand the data better.

4.1 Age

From the below summary of the age column, we can see that the median age is 34.5 and 50% of the students have an age between 27 and 39.

```
> summary(r_df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  23.00   27.00   34.50   34.67   39.00   54.00
```



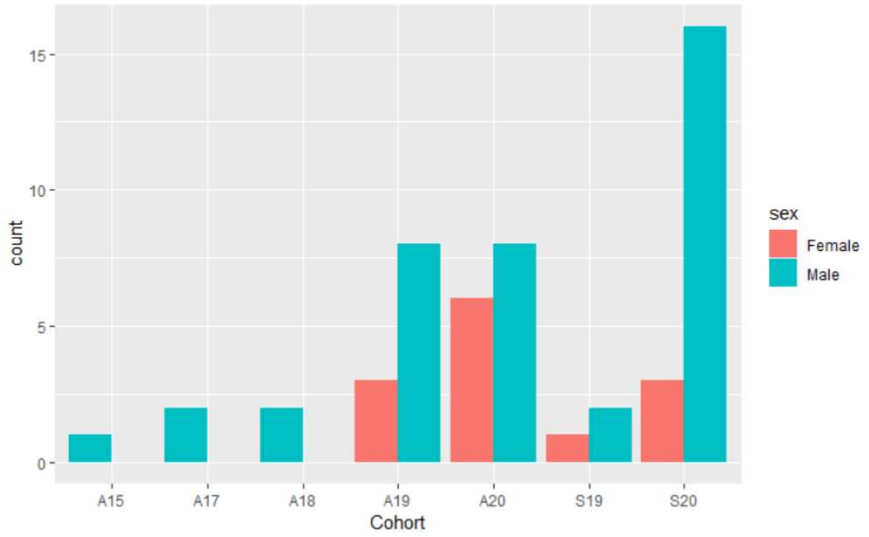
4.2 Sex

Below we can see the distribution of males and females over all cohorts. There are 3 times as many men than women that have responded to the survey.

```
Female   Male   <NA>
    13     39      0
```

Let's now look at the male/female distribution per cohort.

```
ggplot(d_sex, aes(x = Cohort, fill = sex)) +
  geom_bar(position = position_dodge(preserve = "single"))
```



In the earlier cohorts there is no female representation at all. This does not necessarily mean that there were no females in these cohorts only that no females responded to the survey from these cohorts. There are only female respondents to the survey from the S19 cohort. It's only in the more recent cohorts that we see female numbers increasing but there are still fewer females compared to males.

4.3 Educational Background

Below we see the educational breakdown of the students.
We can see that the maths group has the highest number of students.

```
> table(d_edu$education)

 bio  fin  lit math mgmt  oth
   6    5    1   33    6    1
```

This is probably because the maths group has so many subjects together:
Mathematics, Physics, Chemistry, Computer Science, Statistics

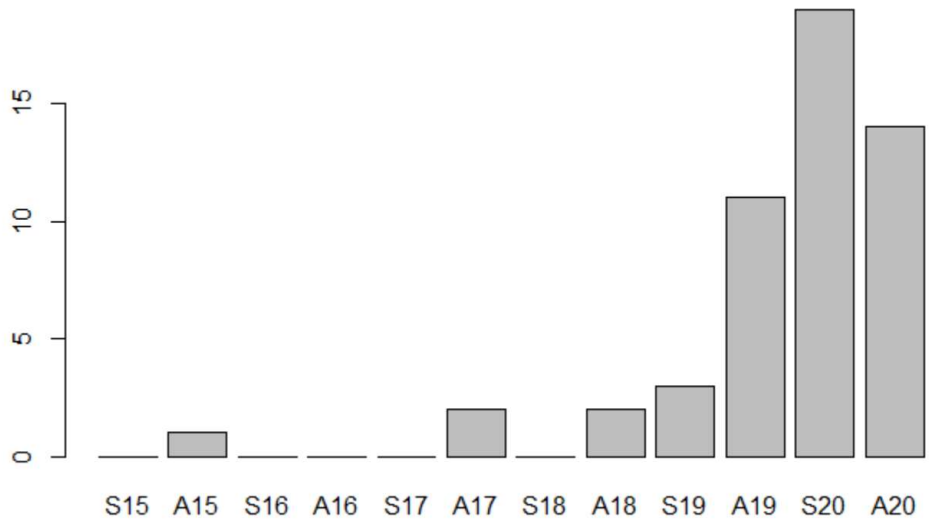
It might be worth breaking this group up by grouping Chemistry with Biology and Medicine (bio group) and putting Computer Science in its own group. For instance, I have a background in Computer Science but not Maths. So, I would not feel I belong in the maths group but that's what I would have had to select if I did the survey.

4.4 Cohorts

There are very few respondents from the earlier cohorts. The numbers only start increasing from cohort A19 with cohorts A19, S20 and A20 containing the most respondents to the survey,

```
> table(d_cohort$cohort, useNA = "always")
```

```
S15  A15  S16  A16  S17  A17  S18  A18  S19  A19  S20  A20 <NA>
  0    1    0    0    0    2    0    2    3   11   19   14    0
```



4.5 Children

Around 30% of students have children and 70% do not.

```
> table(d_fin$children, useNA = "always")
```

```
  No  Yes <NA>
  36   16    0
```

4.6 Years of Education

The count, n, per number of years of education can be seen below. Most students have either 18 or 20 years of education.

```
                n
edu_years=14  2
edu_years=15  1
edu_years=16  6
edu_years=17  3
edu_years=18 12
edu_years=19  4
edu_years=20 10
edu_years=21  5
edu_years=22  2
edu_years=23  1
edu_years=24  2
edu_years=25  2
```

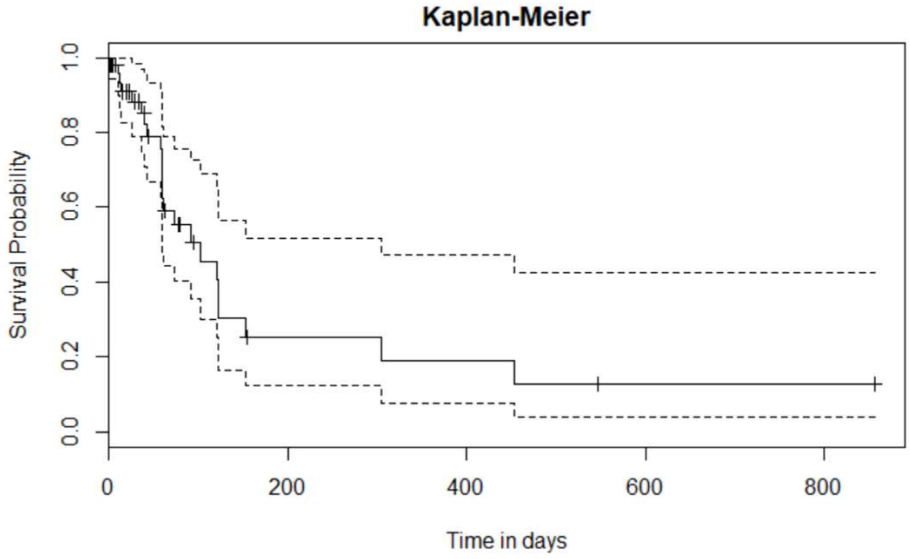
5 General Survival Analysis using Kaplan-Meier & Nelson-Aalen

5.1 Kaplan-Meier

Here we use the **survfit** function in the survival package in R to obtain a survival plot for time to internship within a specific time interval.

```
kmfit <- survfit(tti ~ 1, data = d_fin)
plot(kmfit, mark.time = TRUE,
    main = "Kaplan-Meier",
    xlab = "Time in days",
    ylab = "Survival Probability")
```

Below we have the Kaplan-Meier survival curve. The survival probability is on the y axis. It's the % students that have not experienced the event at the point in time on the x axis.

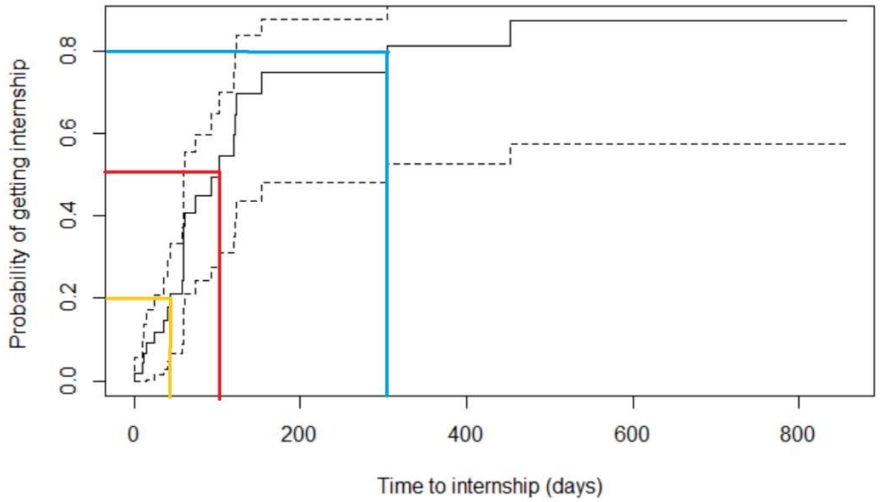


Each step down in the plot represents the event of a student getting an internship at that point in time.

Censored events are marked with a cross on the above plot. These represent students that did not find an internship by the survey date. As we can see from the plot there are many censored events within the first 100 days. It is a mistake to exclude these censored events as this could lead to bias.

As we are trying the find the probability of students getting an internship, the plot above is not very intuitive as it shows the probability decreasing with time and we want to see the opposite. We can use the fun = "F" to plot 1 – *S (survival function)* or incidence. The y axis on the below plot has been labelled 'Probability of students getting an internship' as this is what it now represents. The censored events are not displayed in the below plot as they eclipse the steps of the plot making it difficult to see when events happen.

```
#Plot incidence not displaying censored values
plot(kmfit, fun= "F",
    xlab = "Time to internship (days)",
    ylab = "Probability of getting internship")
```

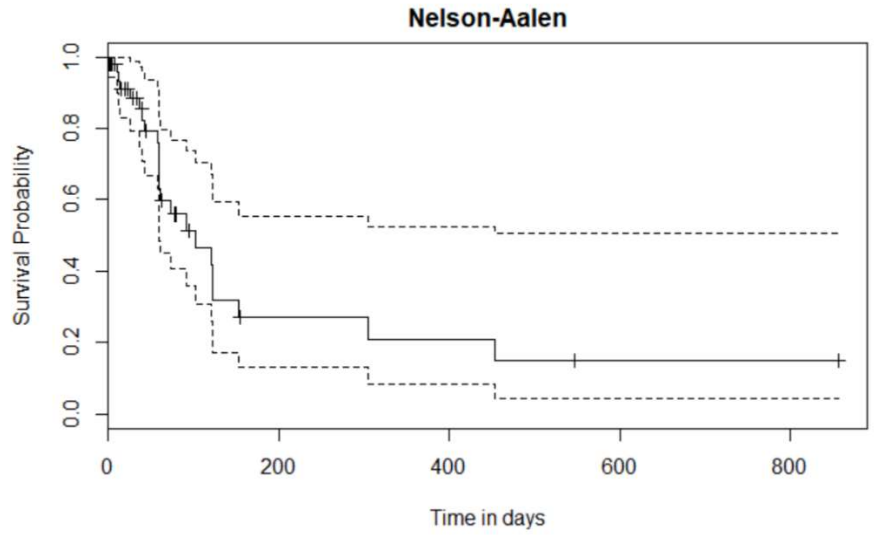


The above plot shows that 50% of students find an internship within 100 days of starting to look for one (red line). This is the median survival time.
80% of students have experienced the event that is they find an internship within approximately 300 days (blue line). 20% of students have found an internship within approximately 40 days (yellow line).

5.2 Nelson-Aalen

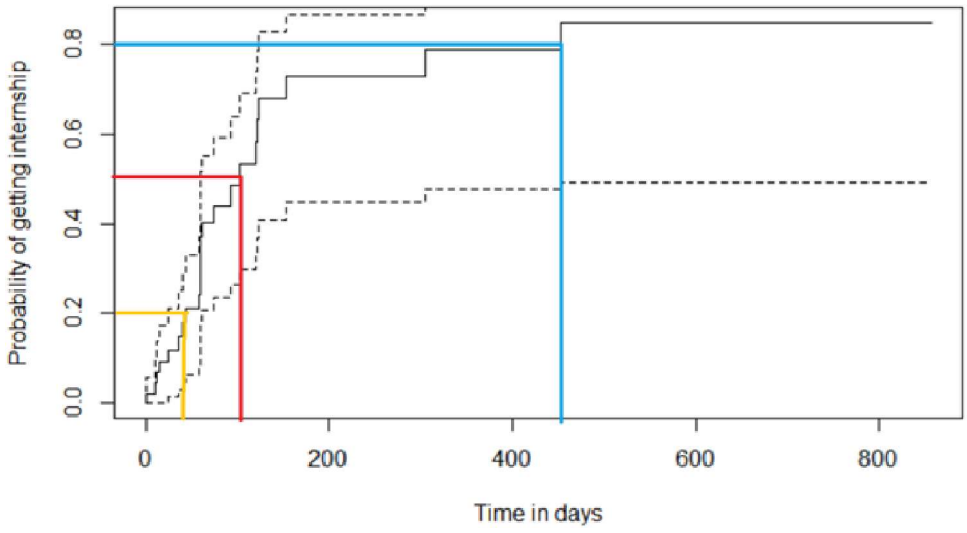
Another method of estimation of survival is Nelson-Aalen (Fleming-Harrington). This estimates the cumulative hazard. It is not as popular as Kaplan-Meier but nevertheless can be used alongside the Kaplan-Meier estimator. In R it can be obtained using the '**survfit**' function passing a **type** parameter equal to "**fh**".

```
#Nelson Aalen
nafit <- survfit(tti ~ 1, data = d_fin, type = "fh")
plot(nafit, mark.time = TRUE,
    main = "Nelson-Aalen",
    xlab = "Time in days",
    ylab = "Survival Probability")
```



Let's plot the incidence so that we can see the increasing probability for students getting an internship.

```
#Plot incidence Nelson Aalen
plot(nafit, fun= "F",
    main = "Nelson-Aalen",
    xlab = "Time in days",
    ylab = "Probability of getting internship")
```



As we can see the curve is similar to the Kaplan-Meier. Only the 80% probability is very different in that it takes students around 450 days to find an internship compared with 300 days in the Kaplan-Meier curve.

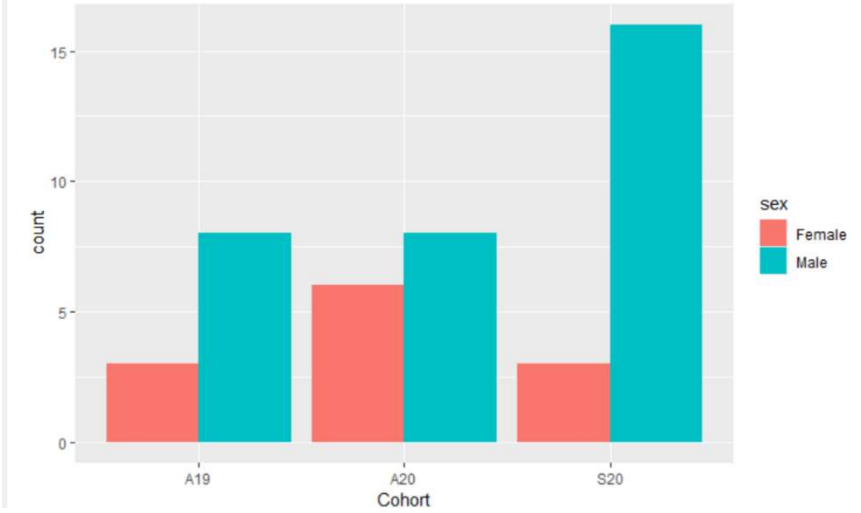
6 Analysis of Groups/Covariates using Kaplan-Meier, Logrank & Cox Proportional Hazards Model

In this section, time to internship will be measured against each of the variables in the dataset: sex, cohort, educational background, years of education, children and age.

6.1 Sex

Let's look at the gender distribution over the three most recent cohorts.

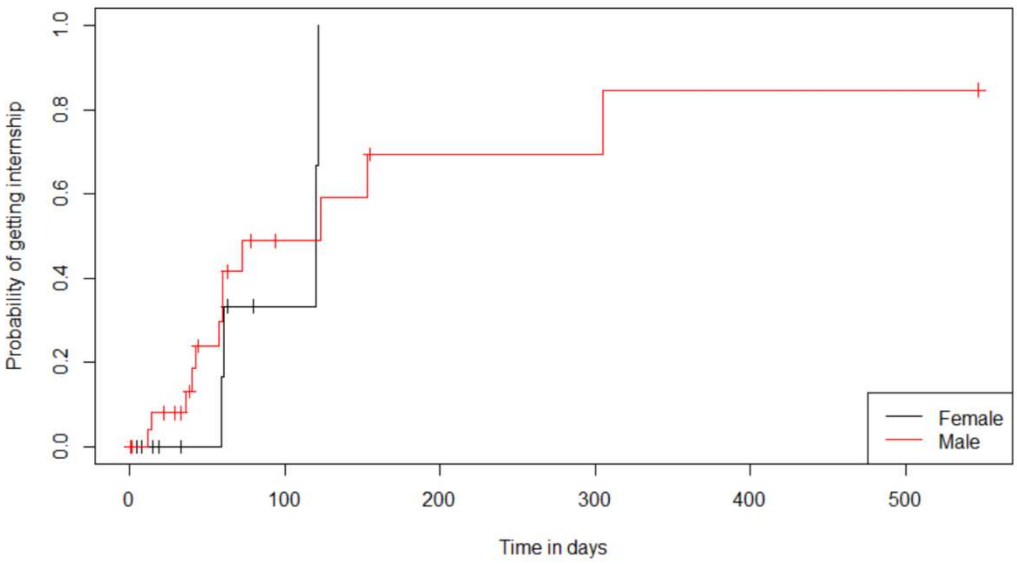
```
ggplot(d_recent, aes(x = Cohort, fill = sex)) +
  geom_bar(position = position_dodge(preserve = "single"))
```



6.1.1 Kaplan-Meier

Let's plot the Kaplan-Meier for Males/Females time to internship over the last three cohorts:

```
#Sex
sfit_tti_sex <- survfit(tti ~ sex, data = d_recent)

plot(sfit_tti_sex, col = 1:2, fun="F", mark.time=TRUE,
  xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:2, legend=c("Female","Male"), lty=1)
```



The curve for males is initially higher but the female curve converges to it at around 120 days. Also, the female curve reaches almost 100% probability while the male curve does not. Are these differences statistically significant? Let's carry out the Logrank test to find out.

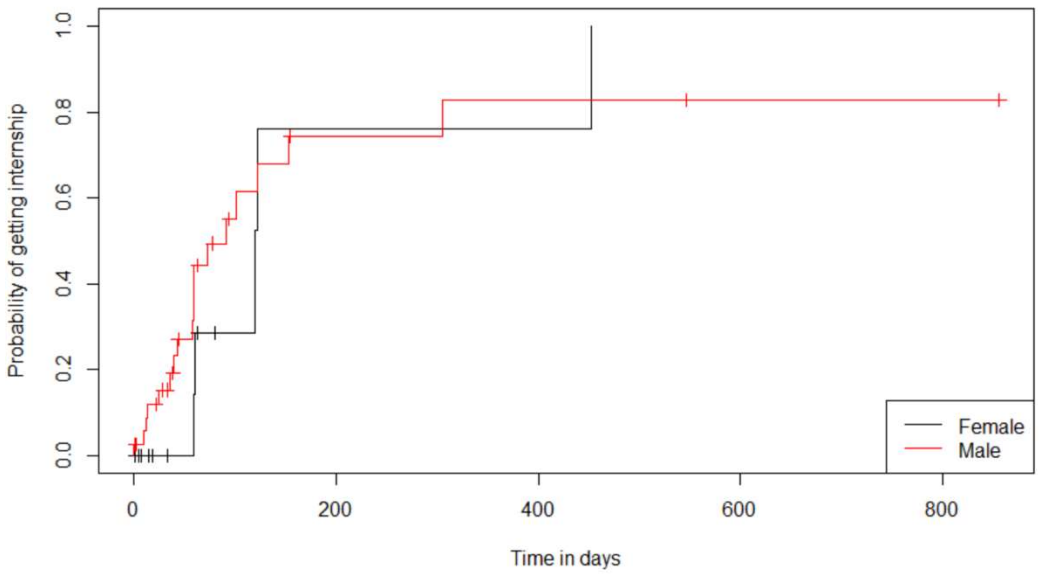
6.1.2 Logrank Test

```
Call:
survdiff(formula = tti ~ sex, data = d_recent)

             N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Female 12        4     3.64    0.0353    0.0499
sex=Male   32       12    12.36    0.0104    0.0499

 Chisq= 0  on 1 degrees of freedom, p= 0.8
```

The p-value is very large therefore the difference is not statistically significant.
The difference could be due to the fact that there are three times as many events observed for males compared to females, rather than a difference in gender.

Let's look at men and women over all the cohorts:

```
#Sex
sfit_tti_sex <- survfit(tti ~ sex, data = d_fin)

plot(sfit_tti_sex, col = 1:2, fun="F", mark.time=TRUE,
  xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:2, legend=c("Female","Male"), lty=1)
```



The difference is less apparent compared with the previous three cohort plot. In this plot the female curve converges with the male one after around 150 days.

Let's do the Logrank test to find out if there is any statistical significance.

```
> survdiff(tti ~ sex, data = d_fin)
Call:
survdiff(formula = tti ~ sex, data = d_fin)

            N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Female 13        5     6.02    0.1733      0.24
sex=Male   39       18    16.98    0.0615      0.24

 Chisq= 0.2  on 1 degrees of freedom, p= 0.6
```

The p-value is a little lower than previously but still not statistically significant.

6.1.3 Cox Proportional Hazards Model

Let's compare sex using the Cox regression model. Here we are using all the cohorts.

```
#Sex
coxph_fit <- coxph(tti ~ sex, data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ sex, data = d_fin)

  n= 52, number of events= 23

          coef exp(coef) se(coef)     z Pr(>|z|)
sexMale 0.2558    1.2915   0.5077 0.504    0.614

        exp(coef) exp(-coef) lower .95 upper .95
sexMale     1.291     0.7743    0.4774     3.494

Concordance= 0.572  (se = 0.037 )
Rsquare= 0.005   (max possible= 0.918 )
Likelihood ratio test= 0.27  on 1 df,   p=0.6
Wald test            = 0.25  on 1 df,   p=0.6
Score (logrank) test = 0.26  on 1 df,   p=0.6
```

The beta coefficient for sex = 0.26 indicates that males have a higher risk (lower survival) than females, in these data.
The Hazard Ratio is 1.29. This means that being male increases the hazard by 29%. As the hazard is higher for males it means the time to internship is faster by 29% for males compared to females.

The p-value for the Wald test is large at 0.6 this means that the difference is not statistically significant.

The point estimate is 1.3. The 95% confidence is interval is also large between 0.48 and 3.5.

Concordance is the probability of agreement between two randomly chosen observations. It tells us the chance of being correct in selecting the one observation with the higher risk of an event from two randomly chosen ones. We need the concordance to be as close to 1 as possible. Anything lower than 0.5 is a bad model. Here the concordance is greater than 0.5 but not by much.

The results also reflect our findings in the Kaplan-Meier plot where the curve for males was initially higher than that for females. However, as there are only one third as many females as males in the dataset this may also account for the difference.

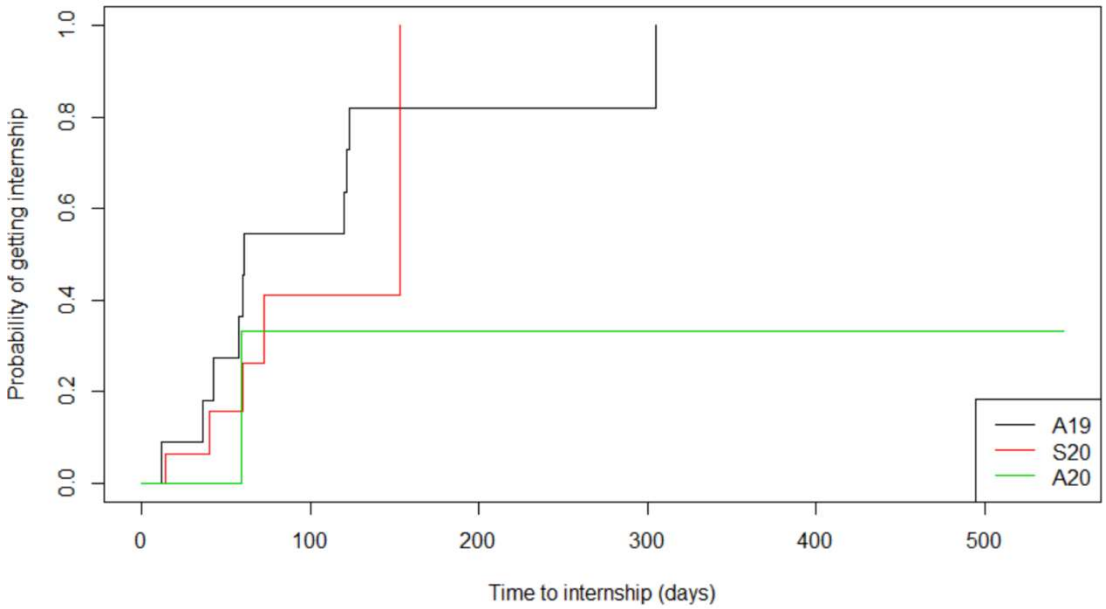
6.2 Cohorts

From the exploratory data analysis, we saw that most of the previous cohorts do not have much data. Let's just take a look at the time to internship of the three most recent cohorts: A19, S20 and A20.

6.2.1 Kaplan-Meier

First of all, let's plot the Kaplan-Meier to get a visualisation of the data.

```
#KM incidence curve for recent cohorts
d_recent$tti <- with(d_recent, Surv(tti, int_found))
sfit_tti_cohort <- survfit(tti ~ cohort, data = d_recent)

plot(sfit_tti_cohort, col = 1:3, fun="F",
  xlab = "Time to internship (days)",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:3, legend=c("A19","S20","A20"), lty=1)
```



We see from the above plot that there are differences in the curves for the different cohorts. A19 students seem to initially be faster to internship with S20 converging to the A19 curve after 150 days. A20 appears to have only one event so its presence does not add to the power of the test.

Are these differences statistically significant? Let's do the Logrank test to find out.

6.2.2 Logrank Test

```
survdiff(tti ~ cohort, data = d_recent)
```

```
Call:
survdiff(formula = tti ~ cohort, data = d_recent)

             N Observed Expected (O-E)^2/E (O-E)^2/V
cohort=A19 11       10     7.35     0.956     1.873
cohort=S20 19        5     5.78     0.106     0.186
cohort=A20 14        1     2.87     1.215     1.580

 Chisq= 2.4  on 2 degrees of freedom, p= 0.3
```

The p-value, 0.3 is quite high which means the difference in time to internship between the different cohorts is not statistically significant.

6.2.3 Stratified Logrank Test

We saw earlier that the number of women between the cohorts is different and that the time to internship is also different between men and women. Does the sex stratification have an impact on our conclusions? Let's compare the two cohorts, A19 and S20 while controlling for the potential confounder of sex.

We can do a stratified Logrank test.

```
Call:
survdiff(formula = tti ~ cohort + strata(sex), data = d_recent)

            N Observed Expected (O-E)^2/E (O-E)^2/V
cohort=A19 11       10     8.71     0.191     0.618
cohort=S20 19        5     6.29     0.264     0.618

 Chisq= 0.6  on 1 degrees of freedom, p= 0.4
```

The p-value is 0.4, therefore still non-significant even after we stratify for sex.

6.2.4 Cox Proportional Hazards Model

As the A20 cohort contains only one observed event let's remove it from the dataframe as it does not increase the power of the test. So now, we will only compare A19 and S20 cohorts.

```
#Compare only cohorts A19 and S20
d_recent <-filter(d_cohort, cohort %in% c("A19", "S20")) %>% droplevels()

coxph_fit <- coxph(tti ~ cohort, data = d_recent)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ cohort, data = d_recent)

  n= 30, number of events= 15

             coef exp(coef) se(coef)      z Pr(>|z|)
cohortS20 -0.4235    0.6548   0.5777 -0.733    0.464

          exp(coef) exp(-coef) lower .95 upper .95
cohortS20    0.6548      1.527     0.211     2.032

Concordance= 0.58  (se = 0.08 )
Rsquare= 0.018   (max possible= 0.898 )
Likelihood ratio test= 0.55  on 1 df,   p=0.5
Wald test            = 0.54  on 1 df,   p=0.5
Score (logrank) test = 0.54  on 1 df,   p=0.5
```

The R summary for the Cox PH model gives the hazard ratio (HR) for the second group relative to the first group, that is, S20 versus A19. The beta coefficient for cohortS20 = -0.42 which indicates that S20 has a lower risk than A19, in these data.
The Hazard Ratio is 0.65. This means that being in S20 reduces the hazard by a factor of 0.65 or 35%. In our scenario for getting an internship we want the risk to be higher as this means a shorter time to internship. As the hazard is lower it means the time to internship is longer by 35% for students in cohort S20.

The p-value for the Wald test is large at 0.464 this means that the difference is not statistically significant.

The 95% confidence is interval is also large between 0.2 and 2.

The concordance is greater than 0.5 but not by much so the quality of the model is adequate.

The results also reflect our findings in the Kaplan-Meier plot where the A19 curve was initially higher than that for S20.

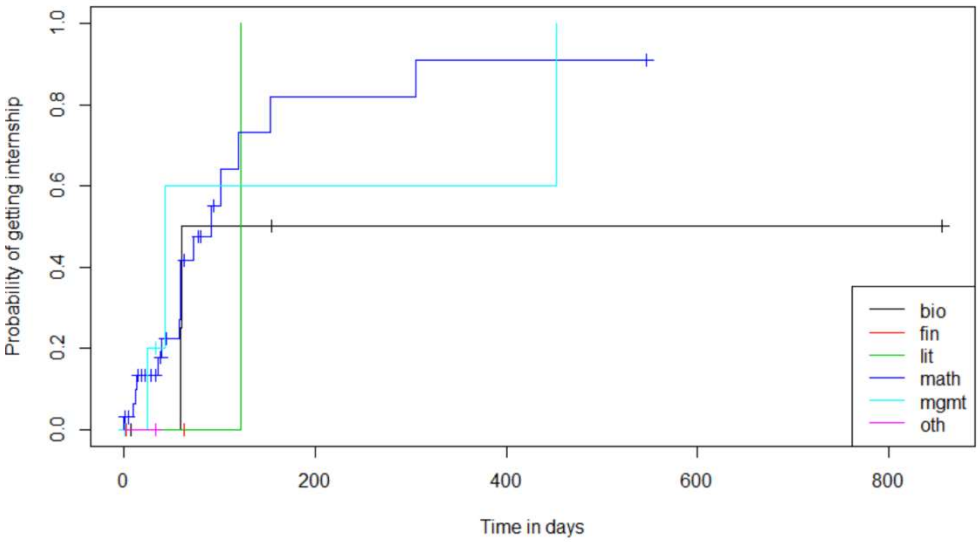
The difference could be explained by the fact that the survey was taken while the S20 course was still ongoing and when many students had not yet found internships. There are only 5 observed events for S20 compared with 10 for A19.

6.3 Educational Background

Let's see if a student's educational background has any bearing on their time to internship.

6.3.1 Kaplan-Meier

```
#Educational Background
sfit_tti_edubg <- survfit(tti ~ education, data = d_fin)

plot(sfit_tti_edubg, col = 1:6, mark.time = TRUE, fun="F",
 xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:6, legend=c("bio","fin", "lit", "math", "mgmt", "oth"), lty=1)
```



The curves are certainly different for the groups but the plot is not so easy to interpret as there are many events for the maths group but not so many for the other groups. Initially, in the first 100 days it seems all groups climb at approximately the same rate except for literature. However, there is only one event for literature so perhaps the results are not really comparable.

Let's see if there is any statistical significance using Logrank.

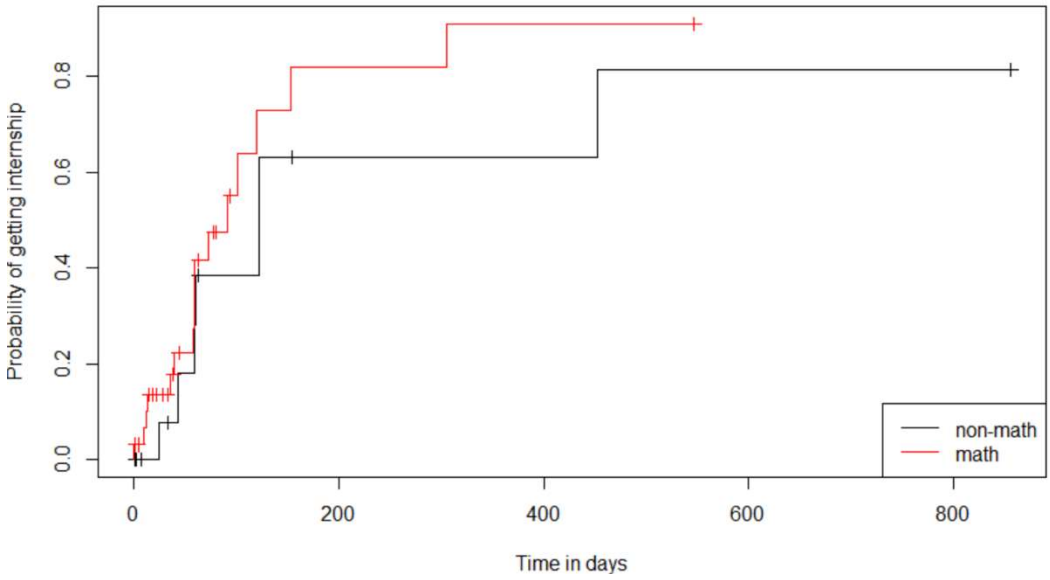
6.3.2 Logrank Test

```
Call:
survdiff(formula = tti ~ education, data = d_fin)

                N Observed Expected (O-E)^2/E (O-E)^2/V
education=bio   6        2    4.188  1.14e+00  1.46e+00
education=fin   5        1    1.701  2.89e-01  3.23e-01
education=lit   1        1    0.993  4.84e-05  5.27e-05
education=math 33       16   13.489  4.68e-01  1.20e+00
education=mgmt  6        3    2.506  9.72e-02  1.15e-01
education=oth   1        0    0.123  1.23e-01  1.26e-01

 Chisq= 2.2  on 5 degrees of freedom, p= 0.8
```

The p-value is high at 0.8. The difference is not significant.

Let's compare the maths students against all the other students.

```
#Maths vs everyone else
sfit_tti_edubg_math <- survfit(tti ~ I(education == "math"), data = d_fin)

plot(sfit_tti_edubg_math, col = 1:2, fun="F",
 xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:2, legend=c("non-math","math"), lty=1)
```



The curves this time are more comparable. It does seem that those students with a maths background have a faster time to internship compared with everyone else with 80% achieving an internship before 200 days. But is there any statistical significance? Let's do the Logrank test to find out.

```
Call:
survdiff(formula = tti ~ I(education == "math"), data = d_cohort)

                              N Observed Expected (O-E)^2/E (O-E)^2/V
I(education == "math")=FALSE 19        7     9.51     0.663       1.2
I(education == "math")=TRUE  33       16    13.49     0.468       1.2

 Chisq= 1.2  on 1 degrees of freedom, p= 0.3
```

The p-value is high at 0.3, therefore there is no statistical significance.

6.3.3 Cox Proportional Hazards Model

Here we compare maths against all other subjects.

```
#Educational background
coxph_fit <- coxph(tti ~ I(education == "math"), data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ I(education == "math"), data = d_fin)

  n= 52, number of events= 23

                               coef exp(coef) se(coef)    z Pr(>|z|)
I(education == "math")TRUE 0.5005    1.6496   0.4634 1.08     0.28

                           exp(coef) exp(-coef) lower .95 upper .95
I(education == "math")TRUE      1.65     0.6062    0.6652     4.091

Concordance= 0.565  (se = 0.052 )
Rsquare= 0.023   (max possible= 0.918 )
Likelihood ratio test= 1.23  on 1 df,   p=0.3
Wald test            = 1.17  on 1 df,   p=0.3
Score (logrank) test = 1.19  on 1 df,   p=0.3
```

The beta coefficient for educational background = 0.5 which indicates that those with a maths background have a higher risk than those with another background, in these data.
The Hazard Ratio is 1.65. This means that having a maths background increases the hazard by 65%. As the hazard is higher for those with a maths background it means the time to internship is quicker by 65% for them compared to another background.

The p-value for the Wald test is large at 0.28 this means that the difference is not statistically significant.

The 95% confidence is interval is also large between 0.66 and 4.1.

The concordance is greater than 0.5 but not by much implying the quality of the model is not that great.

The results also reflect our findings in the Kaplan-Meier plot where the curve for students with a maths background was higher than the one for other backgrounds.

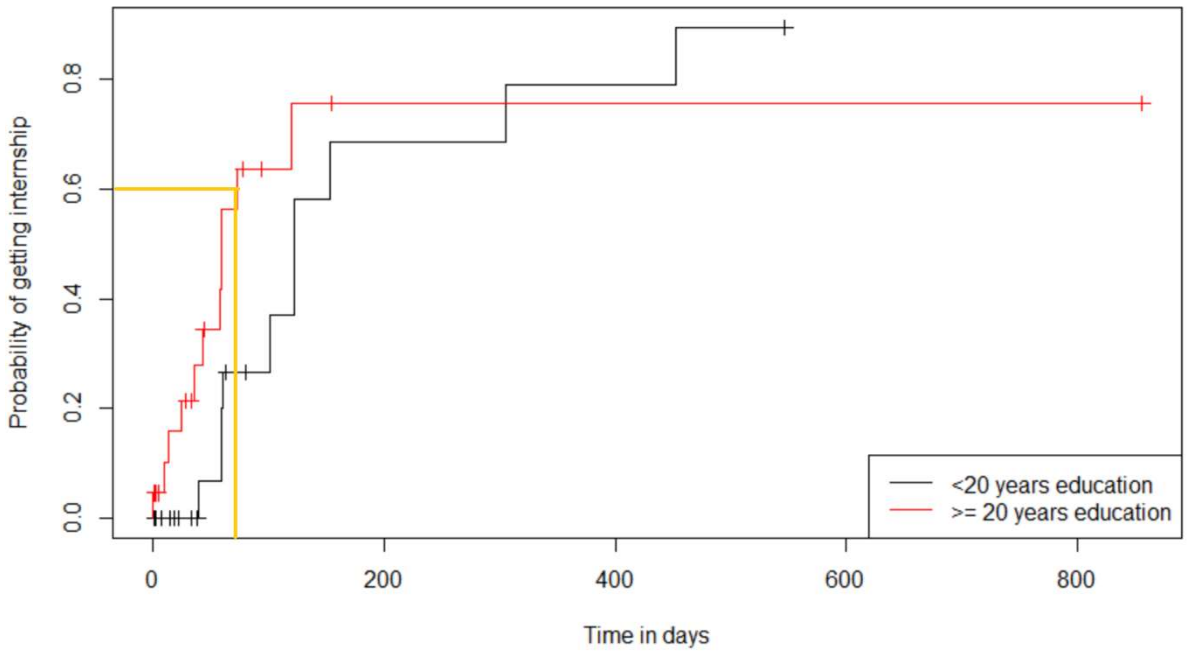
6.4 Years of education

The dataset contained some entries where the number of years of education were 5 or 6 years. These were deemed to be a mistake as in most countries it is compulsory to attend school until the age of 16. As you would need a school certification to enrol at DSTI, these rows with less than 10 years education, were removed from the dataset leaving 50 observations.

6.4.1 Kaplan-Meier

The years of education were split into two groups. Those with less than 20 years education and those with greater than or equal to 20 years education.

```
d_all$tti <- with(d_all, Surv(tti, int_found))

sfit_tti_edu_years <- survfit(tti ~ floor(I(edu_years/10)), data = d_all)

plot(sfit_tti_edu_years, col = 1:2, fun="F", mark.time=TRUE,
 xlab = "Time in days",
 ylab = "Probability of getting internship")
legend("bottomright", col = 1:2, legend=c("<20 years education",">= 20 years education"), lty=1)
```



The curves show that initially it is quicker for those with more than or equal to 20 years of education to find an internship. There is a probability that 60% of students find an internship before 100 days if they have more than 20 years of education (orange line). However, the two curves converge after 300 days or so. Let's see if there is any statistical significance in this difference.

6.4.2 Logrank Test

```
> survdiff(tti ~ floor(I(edu_years/10)), data = d_all)
Call:
survdiff(formula = tti ~ floor(I(edu_years/10)), data = d_all)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
floor(I(edu_years/10))=1 28       10    13.37     0.848      2.41
floor(I(edu_years/10))=2 22       11     7.63     1.486      2.41

 Chisq= 2.4  on 1 degrees of freedom, p= 0.1
```

The p-value is 0.1 therefore not statistically significant.
We do not reject the null hypothesis.

6.4.3 Cox Proportional Hazard Test

Here we compare DSTI students number of years of education in decades using floor to get a round number. This gives us two groups: <20 years education and >= 20 years education.

```
fit.cph <- coxph(tti ~ floor(I(edu_years/10)), data = d_all)
summary(fit.cph)
```

```
Call:
coxph(formula = tti ~ floor(I(edu_years/10)), data = d_all)

  n= 50, number of events= 21

                          coef exp(coef) se(coef)     z Pr(>|z|)
floor(I(edu_years/10)) 0.6677    1.9498   0.4410 1.514     0.13

                       exp(coef) exp(-coef) lower .95 upper .95
floor(I(edu_years/10))      1.95     0.5129    0.8214     4.628

Concordance= 0.669  (se = 0.052 )
Rsquare= 0.045   (max possible= 0.902 )
Likelihood ratio test= 2.28  on 1 df,   p=0.1
Wald test            = 2.29  on 1 df,   p=0.1
Score (logrank) test = 2.37  on 1 df,   p=0.1
```

The beta coefficient for years of education (in decades) = 0.67 indicates that the risk is increased with each decade, in these data. Note that there are only two possibilities for decades of education here, 1 or 2.

The Hazard Ratio is 1.95. This means that each increasing decade of education increases the hazard by 95%. As the hazard is higher it means the time to internship is 95% times quicker for each decade more of education.

The p-value for the Wald test is 0.13. This means that the difference is not statistically significant.

The 95% confidence is interval is big between 0.82 and 4.63.

The concordance is 0.67 which is not so bad.

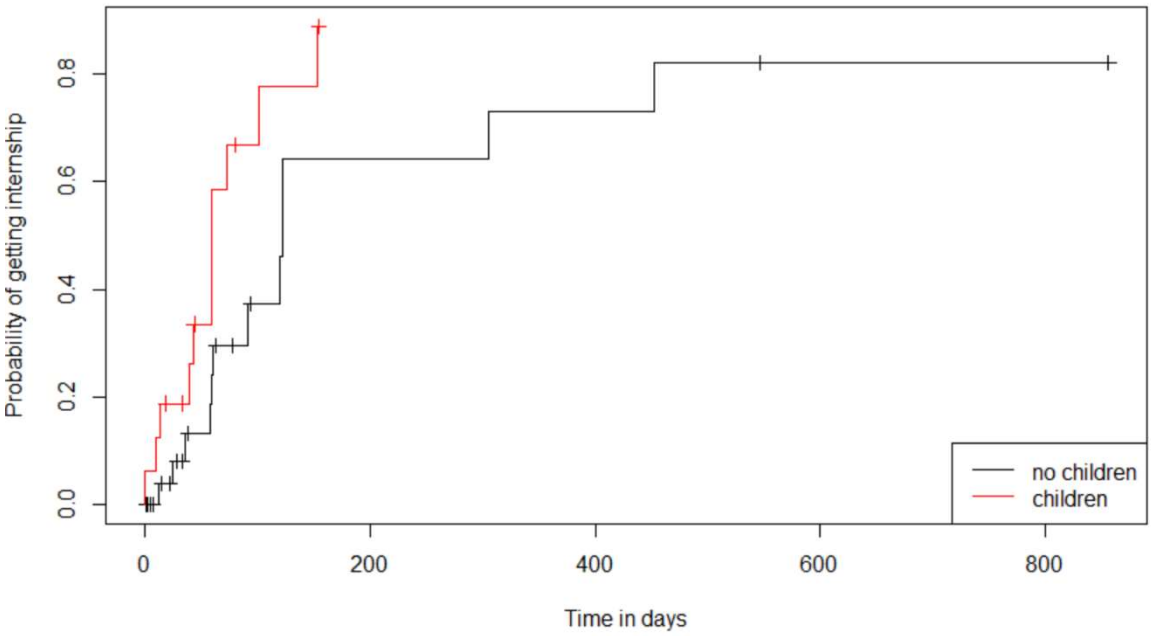
6.5 Children

Let's now see whether having children effects the time to internship of students.

6.5.1 Kaplan-Meier

Let's first plot the Kaplan -Meier specifying children as a group:

```
#Children?
sfit_tti_children <- survfit(tti ~ children, data = d_fin)

plot(sfit_tti_children, col = 1:2, fun="F",
 xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:2, legend=c("no children","children"), lty=1)
```



The curve shows that students with children find an internship faster than those without children. Time to internship is less than 200 days for those with children. Let's see if this difference is statistically significant using the Logrank test.

6.5.2 Logrank Test

```
Call:
survdiff(formula = tti ~ children, data = d_fin)

              N Observed Expected (O-E)^2/E (O-E)^2/V
children=No  36       12    16.23      1.10      3.99
children=Yes 16       11     6.77      2.64      3.99

 Chisq= 4  on 1 degrees of freedom, p= 0.05
```

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. As the p-value is equal to 0.05 here this indicates strong evidence against the null hypothesis, as there is less than a

5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

So, we accept that there is a difference in time to internship with students that have children. In my eyes, it makes perfect sense that having small mouths to feed would push someone to try harder to find a job.

6.5.3 Cox Proportional Hazards Model

Here we compare students with and without children

```
#Children
coxph_fit <- coxph(tti ~ children, data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ children, data = d_fin)

  n= 52, number of events= 23

              coef exp(coef) se(coef)     z Pr(>|z|)
childrenYes 0.8594    2.3618   0.4397 1.955   0.0506 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
childrenYes     2.362     0.4234    0.9977     5.591

Concordance= 0.628  (se = 0.06 )
Rsquare= 0.07   (max possible= 0.918 )
Likelihood ratio test= 3.76  on 1 df,   p=0.05
Wald test            = 3.82  on 1 df,   p=0.05
Score (logrank) test = 4.05  on 1 df,   p=0.04
```

The beta coefficient for having children = 0.86 indicates that those with children have a higher risk than those without, in these data.

The Hazard Ratio is 2.36. This means that having children increases the hazard by a factor of 2.36. As the hazard is higher for those with children it means the time to internship is 2.36 times quicker for them compared to those without children.

The p-value for the Wald test is small at 0.05. This means that the difference is statistically significant and we can reject the null hypothesis.

The 95% confidence is interval is quite large between 0.998 and 5.6.

The concordance is 0.63 which is greater than 0.5 so we have an acceptable model.

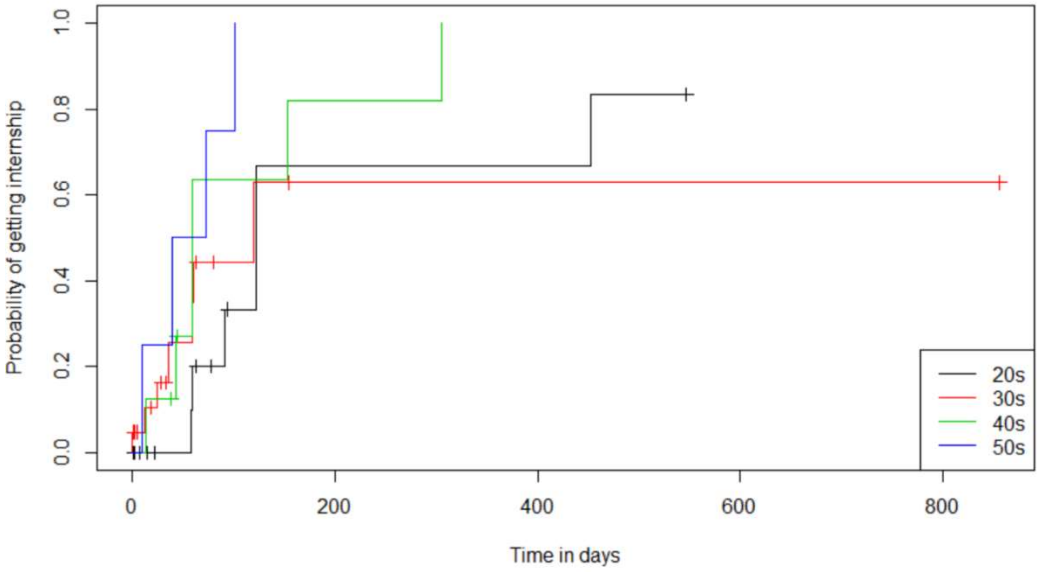
The p-values of the other tests such as the likelihood ratio test is also 0.05 suggesting that the null hypothesis can be rejected.

6.6 Age

Let's now see if age has an effect on the time to internship. We will use the decade of the student's age e.g. 20s, 30s etc rather than the students actual age.

6.6.1 Kaplan-Meier

```
#Age
sfit_tti_age <- survfit(tti ~ floor(I(age/10)), data = d_fin)

plot(sfit_tti_age, col = 1:4, fun="F", mark.time = TRUE,
 xlab = "Time in days",
  ylab = "Probability of getting internship")
legend("bottomright", col = 1:4, legend=c("20s","30s","40s","50s"), lty=1)
```



From the above plot it looks like students in their 50s have the fastest time to internship, followed by students in their 40s.

Could it be that the previous work experience of these two age groups could help them in finding an internship quicker? Is this difference statistically significant?

6.6.2 Logrank Test

```
Call:
survdiff(formula = tti ~ floor(I(age/10)), data = d_fin)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
floor(I(age/10))=2 19        6     9.35    1.1987     2.107
floor(I(age/10))=3 21        7     7.74    0.0707     0.109
floor(I(age/10))=4  8        6     4.34    0.6307     0.809
floor(I(age/10))=5  4        4     1.57    3.7690     4.233

 Chisq= 5.9  on 3 degrees of freedom, p= 0.1
```

A p-value of 0.1 is not particularly large but it is not <0.05 therefore not statistically significant either.

6.6.3 Cox Proportional Hazards Model

Here we see how students ages impacts their time to internship. The Cox regression model is said to be better than the Logrank test at dealing with continuous variables such as age and weight. We use the decade of the age rather than the actual age.

```
#Age in decades
coxph_fit <- coxph(tti ~ floor(I(age/10)), data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ I(age/10), data = d_fin)

  n= 52, number of events= 23

            coef exp(coef) se(coef)     z Pr(>|z|)
I(age/10) 0.5349    1.7073   0.2460 2.174   0.0297 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
I(age/10)     1.707     0.5857     1.054     2.765

Concordance= 0.689  (se = 0.055 )
Rsquare= 0.085   (max possible= 0.918 )
Likelihood ratio test= 4.64  on 1 df,   p=0.03
Wald test            = 4.73  on 1 df,   p=0.03
Score (logrank) test = 4.91  on 1 df,   p=0.03
```

The beta coefficient for age (in decades) = 0.53 indicates a positive relationship between age and risk. The higher the age, the higher the risk. This means that the risk is increased with each decade, in these data.

The Hazard Ratio is 1.7. This means that each increasing decade increases the risk by 70%. Therefore, the time to internship is 70% shorter for each decade older a student is.

The p-value for the Wald test is small at 0.03. This means that the difference is statistically significant and we can reject the null hypothesis. The p-values of the other tests such as the likelihood ratio test are also 0.03 suggesting that the null hypothesis can be rejected.

The 95% confidence is interval is not so big between 1.05 and 2.76.
The concordance is 0.69 which is a good concordance.

6.7 Cox PH Summary Table

Covariate	Beta coef.	Hazard Rate	p-value	LB 95% CI	UB 95% CI
cohortS20	-0.42	0.65	0.464	0.211	2.032
sexMale	0.26	1.29	0.61	0.48	3.5
eduMathTrue	0.5	1.65	0.28	0.66	4.1
eduYearsDec	0.67	1.95	0.13	0.82	4.63
children	0.86	2.36	0.05*	0.998	5.6
age/10	0.53	1.7	0.03*	1.05	2.77

*significant

6.8 Multivariate Cox Proportional Hazards Model

6.8.1. Cox PH Model with Two Covariates

How can we see the effect of multiple covariates? For example, age and sex together. It is not possible to take the individual results above and combine them. We need to have both covariates in the model simultaneously. It is possible to carry out multivariate analysis with the Cox Proportional Hazard model. The model will give us two parameters. These parameters add up in the log scale. In the hazard scale they multiply.
Let's try to create a model with age in decades and sex together.

```
#Multivariate CPH
coxph_fit <- coxph(tti  ~ I(age/10) + sex, data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ I(age/10) + sex, data = d_fin)

  n= 52, number of events= 23

             coef exp(coef) se(coef)      z Pr(>|z|)
I(age/10)  0.5661    1.7614   0.2719  2.082   0.0373 *
sexMale   -0.1593    0.8527   0.5621 -0.283   0.7769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
I(age/10)    1.7614     0.5677    1.0338     3.001
sexMale      0.8527     1.1727    0.2833     2.566

Concordance= 0.672  (se = 0.057 )
Rsquare= 0.087   (max possible= 0.918 )
Likelihood ratio test= 4.72  on 2 df,   p=0.09
Wald test            = 4.77  on 2 df,   p=0.09
Score (logrank) test = 4.96  on 2 df,   p=0.08
```

The summary tells us that the beta coefficient for being male now has a negative effect = -0.16 (it was positive previously) which indicates that males have a lower risk than females, in these data.
The hazard ratio for being male is also lower, 0.85 compared to the model where only sex was present (HR: 1.29).
The p-value for sex is still not significant.

The beta coefficient for age in decades has increased slightly to 0.56 from 0.53 in the univariate model.
The hazard ratio for age in decades has also increased slightly to 1.76 compared to 1.7 in the univariate model.

Age in decades is still significant as the p-value is still small at 0.03. The confidence intervals are still quite wide for both covariates.

We can now use this model to make predictions. To calculate the risk for a male student based on his age we would need to multiply the hazard ratios for age in decades and sex.

6.8.2. Cox PH Model with Three Covariates

Let's now try and fit a model with three covariates.
We can estimate the effect of age while keeping fixed sex and whether the student has children. So, we do a Cox multiple regression with three covariates.

```
#Multivariate CPH
coxph_fit <- coxph(tti  ~ I(age/10) + sex + children, data = d_fin)
summary(coxph_fit)
```

```
Call:
coxph(formula = tti ~ I(age/10) + sex + children, data = d_fin)

  n= 52, number of events= 23

               coef exp(coef) se(coef)      z Pr(>|z|)
I(age/10)    0.4133    1.5118   0.3362  1.229    0.219
sexMale     -0.1456    0.8645   0.5630 -0.259    0.796
childrenYes  0.4319    1.5402   0.5610  0.770    0.441

            exp(coef) exp(-coef) lower .95 upper .95
I(age/10)      1.5118     0.6615    0.7822     2.922
sexMale        0.8645     1.1567    0.2868     2.606
childrenYes    1.5402     0.6493    0.5129     4.624

Concordance= 0.677  (se = 0.059 )
Rsquare= 0.097   (max possible= 0.918 )
Likelihood ratio test= 5.31  on 3 df,   p=0.2
Wald test            = 5.37  on 3 df,   p=0.1
Score (logrank) test = 5.67  on 3 df,   p=0.1
```

The summary tells us that the beta coefficient for being male again has a negative effect = -0.145 which indicates that males have a lower risk than females, in these data.
The hazard ratio for being male is also lower, 0.86 compared to the model where only a single covariate was present (HR: 1.29) but slightly higher than the previous model with two covariates.
The p-value for sex is still not significant.

The beta coefficient for age in decades has decreased slightly to 0.41 from 0.53 in the univariate model.
The hazard ratio for age in decades has also decreased to 1.51 compared to 1.7 in the univariate model.
Unfortunately, the summary is also telling us that the age in decades in no longer significant as the p-value is now 0.2 which is quite high.

The beta coefficient for children is lower compared to the single covariate model 0.43 compared to 0.86. The hazard ratio is lower compared to the univariate model 1.54 compared to 2.36 and the p-value is now large and therefore having children is now insignificant whereas previously in the single covariate model it was significant.

Unfortunately, we have lost the significance of all covariates in this model.

Having a model with more covariates is more versatile but it does mean that we have more variability in the estimates. There is however less bias as we are taking more covariates into account.

We could also investigate the effect of covariates with more than two levels such as educational background and cohorts but as the number of events for many of these covariate levels are very low (1 or 2) the power of the test would also be low and hence the decision has been made not to pursue this line of investigation.

6.9 Predictions

We can make predictions based on the models we have created. Let's use the Cox PH two covariate model created in section 6.8.1 with age and sex.

β_A=0.56 β_s=-0.16

Let's use these betas in a specific example with two students:

$h(t \mid M, AGE = 30) = h_0(t) \times \exp(-0.16 + 0.56 \times 3.0^*)$

$h(t \mid F, AGE = 40) = h_0(t) \times \exp(0.56 \times 4.0^*)$

*age in decades

(based on example in class)

We don't have a direct prediction of the actual risk as a number we only know it up to a proportionality constant. We can however use the exp part (highlighted in red above) to stratify a population by risk i.e. we can say which of the two students has the higher risk and by how much.

$\beta_s \times M + \beta_A + AGE\_DEC$ -> Linear Predictor (enables one to have a relative evaluation of risk i.e. compare two different people)

We can also use Cox's regression to give us an absolute probability of an event by plugging some empirical estimates for the baseline hazard into the model to get predictions of survival. The R function is: survival::survfit.coxph.

7 Conclusion

Over the course of the analysis, we have been able to answer the questions asked at the beginning.

The answers are summarised below.

1) How long does it take in general?

The answer to this question was seen in section 5.1 using a general Kaplan-Meier plot but is further elaborated here.

The KM plot in section 5.1 showed that 50% of students found an internship within 100 days of starting to look for one. This can be further substantiated by the following command in R:

```
> summary(kmfit, time=100)
Call: survfit(formula = tti ~ 1, data = d_fin)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  100     10      16    0.507  0.0928        0.354        0.725
```

Below is a summary table showing the time to internship based on days:

No. of days	% DSTI students to find internship
30	12
60	37.5
90	45
100	50
120	60
150	70
180	75
365	80

We can see from the above table that 75% of DSTI students are able to find an internship within 6 months of starting to look for one and 80% within a year.

2) *Is the situation changing over cohorts?*

This question was answered in section 6.2. We saw that the earlier cohorts did not have much data. There were not many respondents to the survey from the earlier cohorts. This left us with cohorts A19, S20 and A20 to use for our tests. After some analysis it was discovered that A20 only had one event so the power would have been very low. Therefore, A20 was removed from the dataset and this only left the A19 and S20 observations.

A difference was seen between the two cohorts in the Kaplan-Meier curve. However, the difference was not found to be statistically significant when using Logrank, perhaps due to the small sample size/number of events.

The Cox Proportional Hazard Model hazard ratio showed that the S20 time to internship was 35% longer compared to A19, but the p-value was large and therefore not significant.

Another reason for the difference could be that the survey was taken during the S20 course which means that many students were probably still looking for an internship hence more censored events whereas the A19 cohort was already complete so most would have found their internships already.

3) *Does the educational background have an impact?*

This question was answered in section 6.3. Firstly, all the different backgrounds were compared but the Kaplan-Meier plot was not so easy to interpret as the maths group had a high number of events but the other groups not so many. Despite this it looked like in the first three months most of the curves for the groups climbed at approximately the same rate.

Then the maths group was compared with all the other groups together and this time it looked like the maths group definitely had a faster time to internship from the Kaplan-Meier plot. However, the Logrank test showed that this difference was not significant.

The results of the Cox Proportional Hazard model showed the maths group had a hazard ratio of 1.65 which means there was a 65% higher risk and therefore 65% faster time to internship. However, the p-value was large and therefore the difference was found to be insignificant.

4) *Build a predictive model to identify students at high risk of a long search.*

This question was answered in section 6.8 where multivariate Cox Proportional Hazard models were created. One with two covariates and another with three covariates.

The model with two covariates: age in decades and sex, still showed age to be a significant factor in the model. However, when adding one more covariate to the model: children, the covariates that were shown to be significant when modelled on their own, lost their significance in the multivariate model with three covariates. This shows that while adding covariates may reduce the bias of a model, one must be careful to find a balance as otherwise we have too much variability in the covariates especially when the sample size is small as is the case here.

In section 6.9 we saw how these models could be used to make predictions.

All in all, we have used many survival analysis techniques to obtain our results. However, I feel the analysis would have yielded better results if the sample size was larger. Over a third of the dataset was lost after cleaning the dataset which was of course a necessary evil. Having a low sample size decreased the power of the tests but even with the sample size that was used, it was possible to identify some indicators/risk factors that could be used to identify students that may be at high risk of a long search for an internship.