Data Analysis

Juste Simanauskaite & Patricia Rivera

Contents


```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3, fig.width=5, fig.align="center")
library(tidyverse)
library(broom)
library(plyr)
library(survival)
library(survminer)
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")
dim(aids)
```

```
## [1] 851  16
```

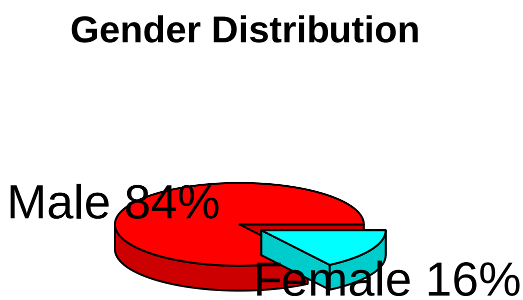
```
summary(aids)
```

```
##        id              time           censor            time_d
##  Min.   :   1.0   Min.   :  1.0   Min.   :0.00000   Min.   :  1.0
##  1st Qu.: 287.5   1st Qu.:179.5   1st Qu.:0.00000   1st Qu.:199.5
##  Median : 581.0   Median :257.0   Median :0.00000   Median :266.0
##  Mean   : 579.5   Mean   :231.8   Mean   :0.08108   Mean   :243.4
##  3rd Qu.: 873.0   3rd Qu.:300.0   3rd Qu.:0.00000   3rd Qu.:306.0
##  Max.   :1156.0   Max.   :362.0   Max.   :1.00000   Max.   :362.0
##     censor_d            tx             txgrp            strat2
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :2.000   Median :1.0000
##  Mean   :0.0235   Mean   :0.5041   Mean   :1.504   Mean   :0.6157
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :2.000   Max.   :1.0000
##       sex            raceth          ivdrug          hemophil
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.00000
##  Median :1.000   Median :1.000   Median :1.000   Median :0.00000
##  Mean   :1.157   Mean   :1.706   Mean   :1.317   Mean   :0.03408
```

```
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.00000
##  Max.   :2.000   Max.   :5.000   Max.   :3.000   Max.   :1.00000
##      karnof            cd4            priorzdv           age
##  Min.   : 70.00   Min.   :  0.00   Min.   :  3.00   Min.   :15.00
##  1st Qu.: 90.00   1st Qu.: 22.25   1st Qu.: 11.00   1st Qu.:33.00
##  Median : 90.00   Median : 75.00   Median : 21.00   Median :38.00
##  Mean   : 91.34   Mean   : 86.45   Mean   : 30.63   Mean   :38.81
##  3rd Qu.:100.00   3rd Qu.:135.75   3rd Qu.: 44.00   3rd Qu.:44.00
##  Max.   :100.00   Max.   :348.00   Max.   :288.00   Max.   :73.00
```

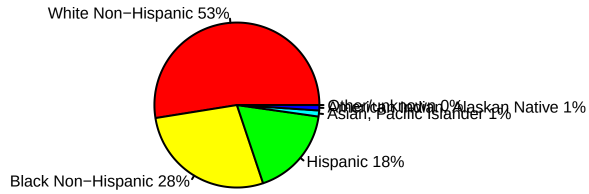
The data set contains a sample size equal to 851 participants and 16 different variables.

```
library(plotrix)
male<-sum(aids$sex==1)
female<-sum(aids$sex==2)
slices <- c(male, female)
lbls <- c("Male", "Female")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie3D(slices,labels=lbls,explode=0.1,
   main="Gender Distribution ", cex.lab=0.1)
```



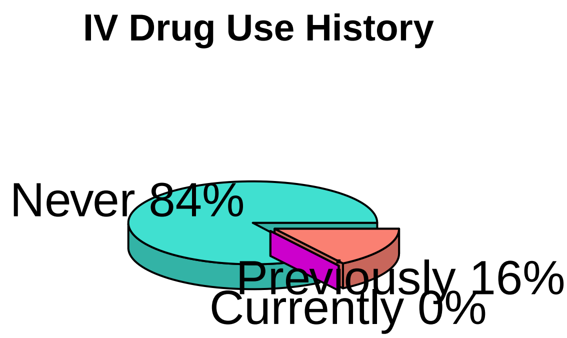
The Pie Chart represents the gender distribution in the sample, with 84% male and 16% female. This shows the potential for the data to not be able to correctly represent the difference of the data variance by gender, if there were to be one. Therefore, gender is something to look into in future data analysis.

```
wnh<-sum(aids$raceth==1)
bnh<-sum(aids$raceth==2)
h<-sum(aids$raceth==3)
api<-sum(aids$raceth==4)
aian<-sum(aids$raceth==5)
oth<-sum(aids$raceth==6)
slices <- c(wnh,bnh,h,api,aian,oth)
lbls <- c("White Non-Hispanic", "Black Non-Hispanic", "Hispanic","Asian, Pacific Islander", "American I
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie(slices,lbls,col = rainbow(length(lbls)), cex=0.5 )
```



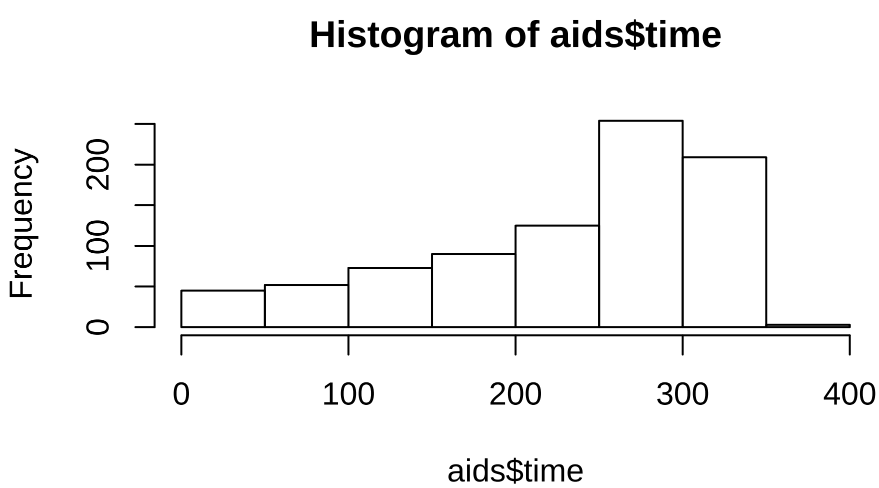
The distribution od race/ethnicity shows that the greatest number of participants consists of white non-hispanic identifying indicidiuals, with black non-hispanic following and hispanic as the 3rd largest represented group.

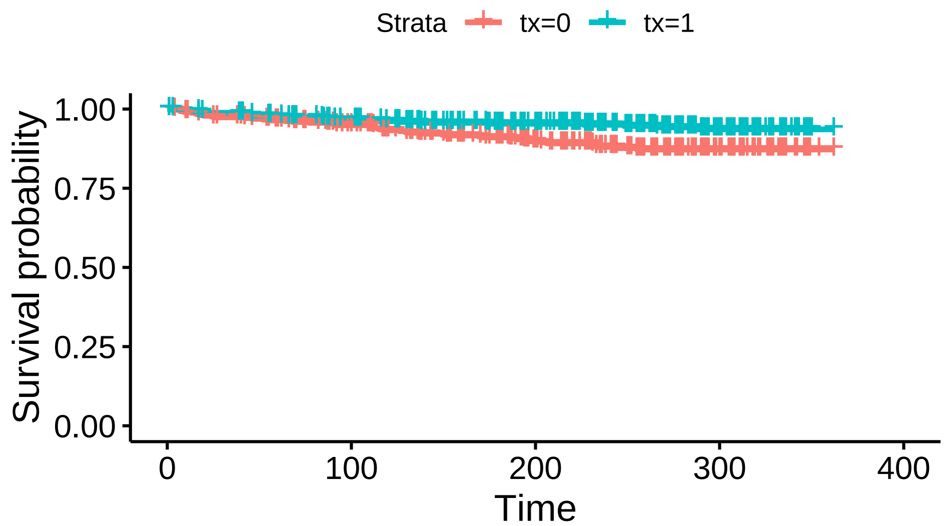
```
never<-sum(aids$ivdrug==1)
cur<-sum(aids$ivdrug==2)
prev<-sum(aids$ivdrug==3)
slices <- c(never,cur,prev)
lbls <- c("Never", "Currently", "Previously")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie3D(slices,labels=lbls,explode=0.1,col=c("turquoise","magenta","salmon"),cex.sub=0.5,
   main="IV Drug Use History ")
```



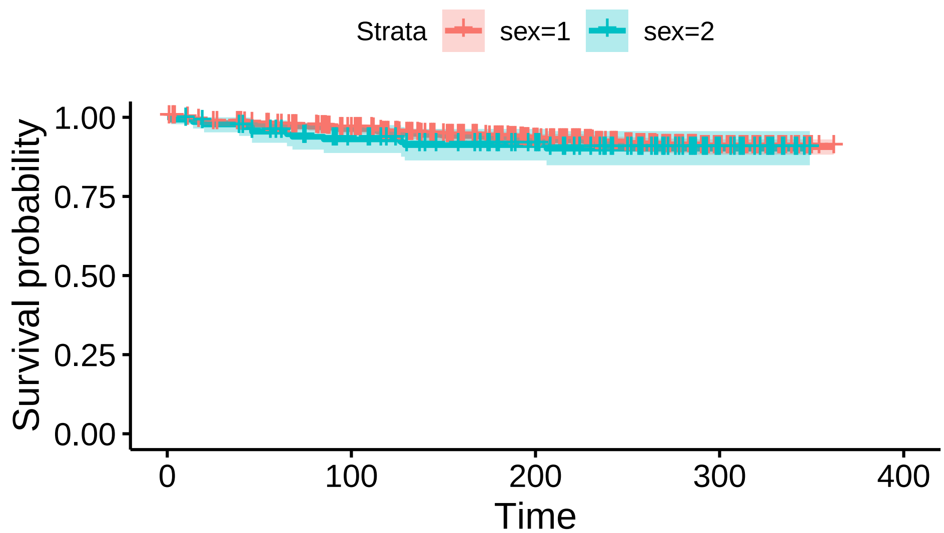
From this chart we see that most of the participants (84%) have never used IV drugs, whereas 16% of participants have some type of history of usage and none of the participants reported to be currently using the drugs.

```
hist(aids$time)

###Data Plots

fit <- survfit(Surv(time,censor)~tx, data = aids)
ggsurvplot(fit,data = aids,conf.int = FALSE)
```

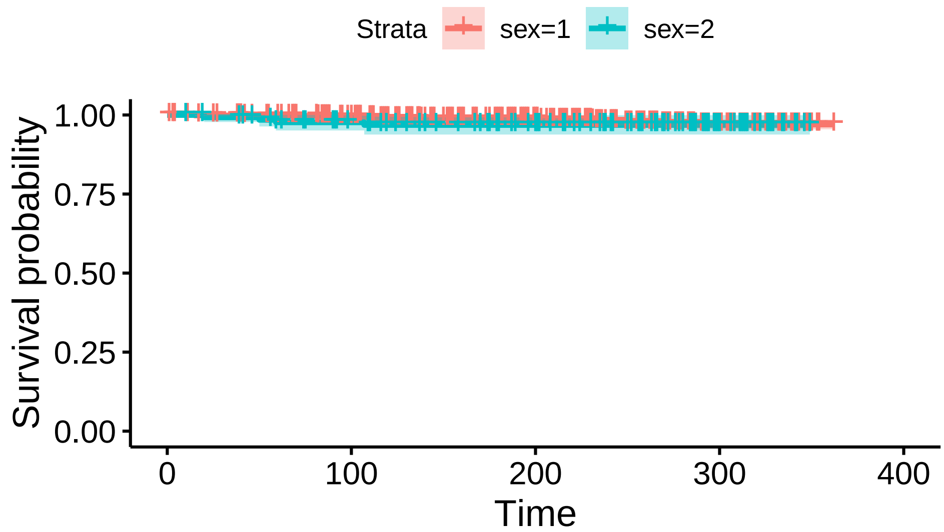





```
aids_fit_time <- survfit(Surv(time, censor) ~ sex, data=aids)
ggsurvplot(aids_fit_time, data=aids,  conf.int = TRUE)
```




```
aids_fit_time.d <- survfit(Surv(time_d, censor_d) ~ sex, data=aids)
ggsurvplot(aids_fit_time.d, data=aids,  conf.int = TRUE)
```



Survival Analysis

```
#mutation of age
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")
aids <- aids %>%
  mutate(age = ifelse(age <= 20, "under20",
                             ifelse(age <=30, "20-30",
                                    ifelse(age <= 40, "30-40",
                                           ifelse(age <=50, "40-50",
                                                  ifelse(age <=60, "50-60",
                                                         ifelse(age <=70, "60-70", "over70"))))))) %>%
  mutate(age = factor(age,
                       levels = c("under20", "20-30", "30-40","40-50", "50-60","60-70","over70")),  sex
```

```
library(survival)
library (survminer)
library(ggplot2)
library(broom)

coxph(Surv(time_d,censor_d) ~ sex , data=aids)  %>% tidy()
```

```
## # A tibble: 1 x 7
##   term    estimate std.error statistic p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 sexmale    0.390     0.559     0.697   0.486   -0.706      1.49
```

```
coxph(Surv(time,censor) ~ sex, data=aids) %>% tidy()
```

```
## # A tibble: 1 x 7
##   term    estimate std.error statistic p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 sexmale    0.199     0.318     0.625   0.532   -0.424     0.821
```

```
coxph(Surv(time,censor) ~ age+ txgrp+ karnof, data=aids) %>% tidy()
```

```
## # A tibble: 8 x 7
```

```
##   term      estimate std.error statistic       p.value conf.low conf.high
##   <chr>        <dbl>     <dbl>     <dbl>         <dbl>    <dbl>     <dbl>
## 1 age20-30   -0.438     1.07    -0.409   0.682           -2.53     1.66
## 2 age30-40   -0.442     1.02    -0.434   0.665           -2.44     1.55
## 3 age40-50   -0.361     1.03    -0.352   0.725           -2.37     1.65
## 4 age50-60    0.460     1.04     0.442   0.659           -1.58     2.50
## 5 age60-70   -0.780     1.42    -0.551   0.582           -3.55     2.00
## 6 ageover70 -14.1    2688.      -0.00525 0.996          -Inf      Inf
## 7 txgrp      -0.844     0.257   -3.28    0.00103         -1.35    -0.340
## 8 karnof     -0.0814    0.0138  -5.89    0.00000000385   -0.109   -0.0543
```

```
cox.zph(coxph(Surv(time,censor) ~ age + txgrp+karnof, data=aids))
```

```
##                rho    chisq     p
## age20-30   0.09054 5.70e-01 0.450
## age30-40   0.19294 2.53e+00 0.112
## age40-50   0.14871 1.50e+00 0.220
## age50-60   0.19861 2.69e+00 0.101
## age60-70   0.16251 1.81e+00 0.179
## ageover70  0.16355 2.57e-07 1.000
## txgrp     -0.10779 8.34e-01 0.361
## karnof     0.00121 1.03e-04 0.992
## GLOBAL          NA 7.98e+00 0.435
```

```
coxph(Surv(time,censor) ~ age *txgrp*karnof, data=aids) %>% tidy()
```

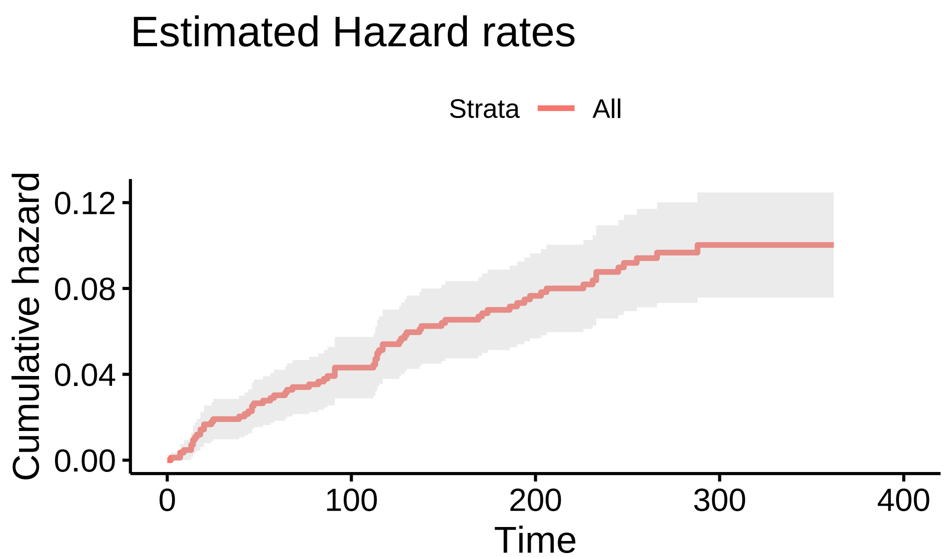
```
## # A tibble: 27 x 7
##    term           estimate std.error  statistic p.value conf.low conf.high
##    <chr>             <dbl>     <dbl>      <dbl>   <dbl>    <dbl>     <dbl>
##  1 age20-30          307.     138277.  0.00222     0.998     -Inf       Inf
##  2 age30-40          319.     138277.  0.00231     0.998     -Inf       Inf
##  3 age40-50          327.     138277.  0.00237     0.998     -Inf       Inf
##  4 age50-60          343.     138277.  0.00248     0.998     -Inf       Inf
##  5 age60-70          287.     176491.  0.00163     0.999     -Inf       Inf
##  6 ageover70          -1.66    29414. -0.0000565   1.000     -Inf       Inf
##  7 txgrp             150.      92392.  0.00163     0.999     -Inf       Inf
##  8 karnof              3.36     1424.  0.00236     0.998     -Inf       Inf
##  9 age20-30:txgrp  -144.       92392. -0.00156     0.999     -Inf       Inf
## 10 age30-40:txgrp  -146.       92392. -0.00158     0.999     -Inf       Inf
## # ... with 17 more rows
```

```
cox.zph(coxph(Surv(time,censor) ~ age *txgrp*karnof, data=aids))
```

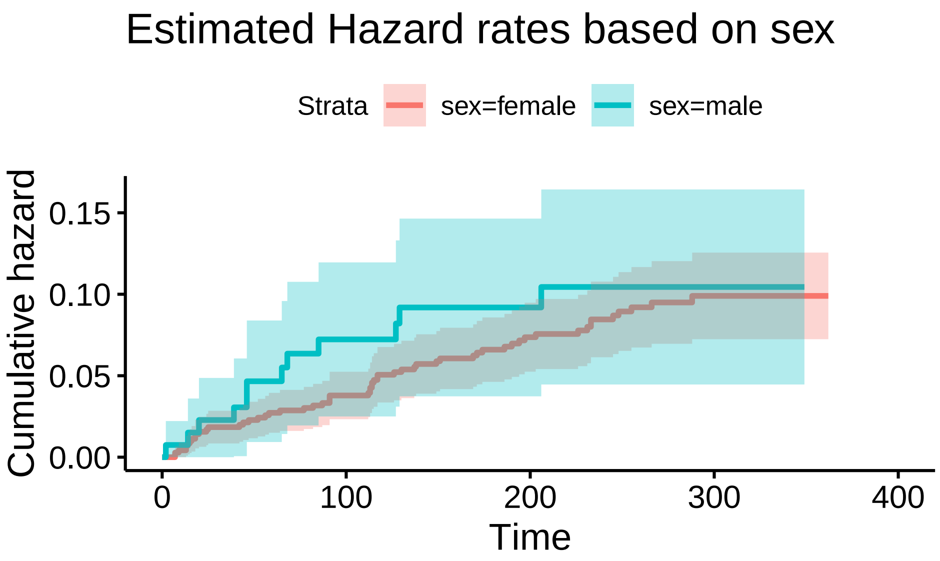
```
##                             rho    chisq     p
## age20-30                -0.1008 4.31e-08 1.000
## age30-40                -0.1583 3.15e-08 1.000
## age40-50                -0.0965 1.25e-08 1.000
## age50-60                -0.2071 6.53e-08 1.000
## age60-70                -0.2062 3.04e-08 1.000
## ageover70               -0.2493 7.81e-11 1.000
## txgrp                   -0.2032 2.68e-08 1.000
## karnof                  -0.1974 5.24e-08 1.000
## age20-30:txgrp           0.0921 2.14e-08 1.000
## age30-40:txgrp           0.1142 1.08e-08 1.000
## age40-50:txgrp           0.0826 5.64e-09 1.000
## age50-60:txgrp           0.1851 3.47e-08 1.000
```

```
## age60-70:txgrp          0.2102 2.15e-08 1.000
## ageover70:txgrp         0.1967 3.96e-11 1.000
## age20-30:karnof         0.0984 4.53e-08 1.000
## age30-40:karnof         0.1524 3.44e-08 1.000
## age40-50:karnof         0.0938 1.40e-08 1.000
## age50-60:karnof         0.2053 7.78e-08 1.000
## age60-70:karnof         0.1978 3.00e-08 1.000
## ageover70:karnof            NA      NaN   NaN
## txgrp:karnof            0.1996 2.81e-08 1.000
## age20-30:txgrp:karnof  -0.0910 2.15e-08 1.000
## age30-40:txgrp:karnof  -0.1020 9.71e-09 1.000
## age40-50:txgrp:karnof  -0.0823 6.23e-09 1.000
## age50-60:txgrp:karnof  -0.1796 3.72e-08 1.000
## age60-70:txgrp:karnof  -0.1981 1.98e-08 1.000
## ageover70:txgrp:karnof      NA      NaN   NaN
## GLOBAL                      NA 1.84e+01 0.891
```

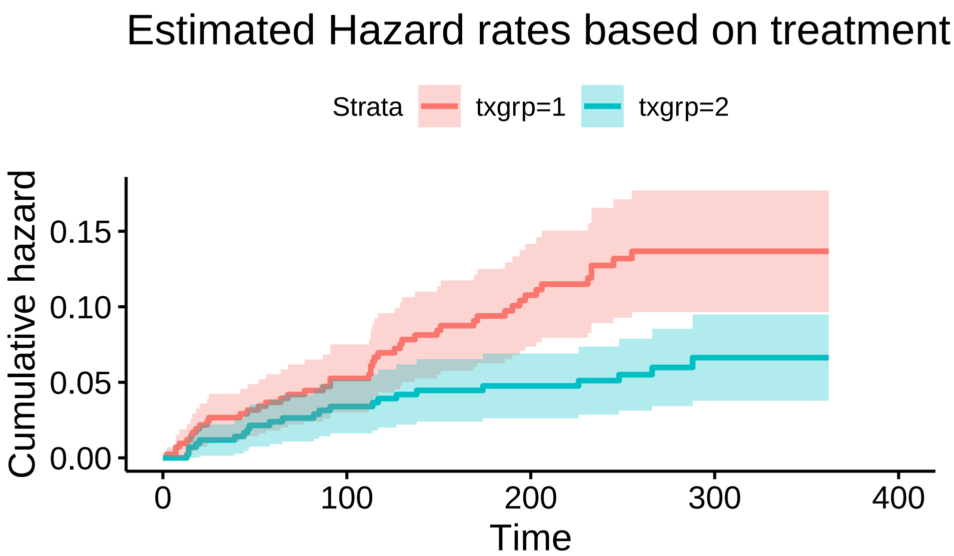
```
ggsurvplot(survfit(Surv(time,censor) ~ 1, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates")
```



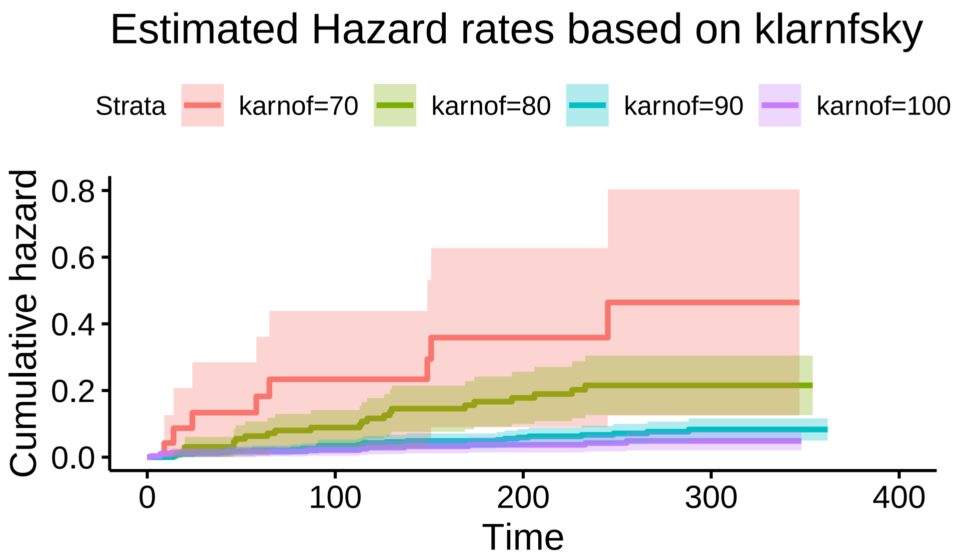

```
ggsurvplot(survfit(Surv(time,censor) ~ sex, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on sex")
```



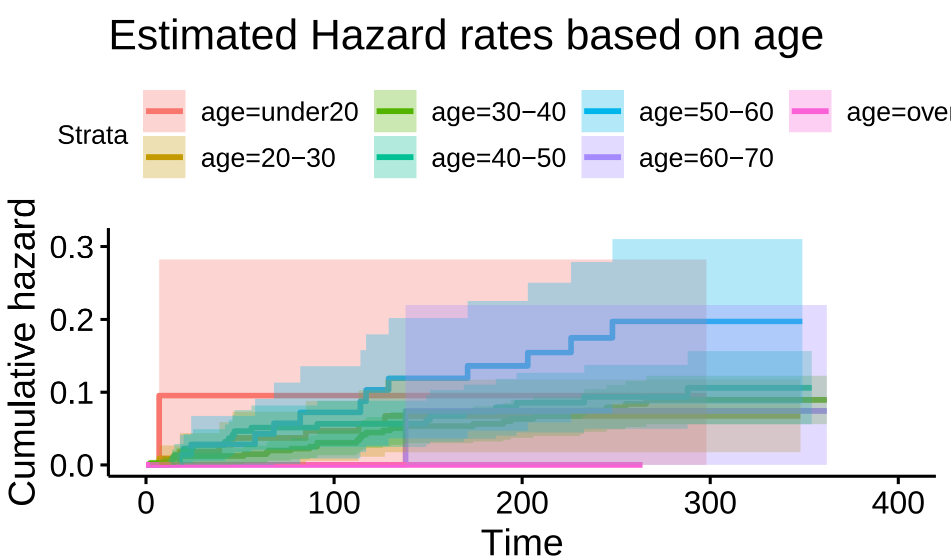

```
ggsurvplot(survfit(Surv(time,censor) ~ txgrp, data=aids),
          censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on treatment grou
```



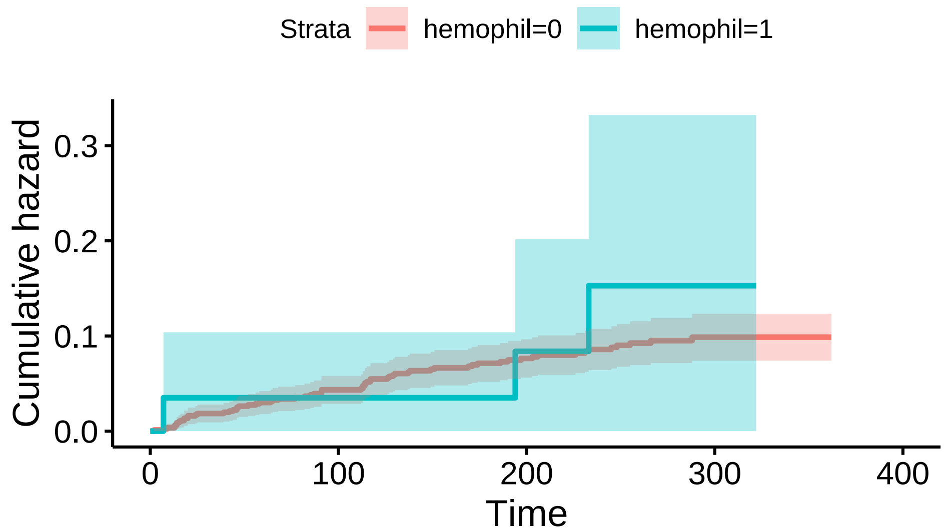

```
ggsurvplot(survfit(Surv(time,censor) ~ karnof, data=aids),
          censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on klarnfsky")
```



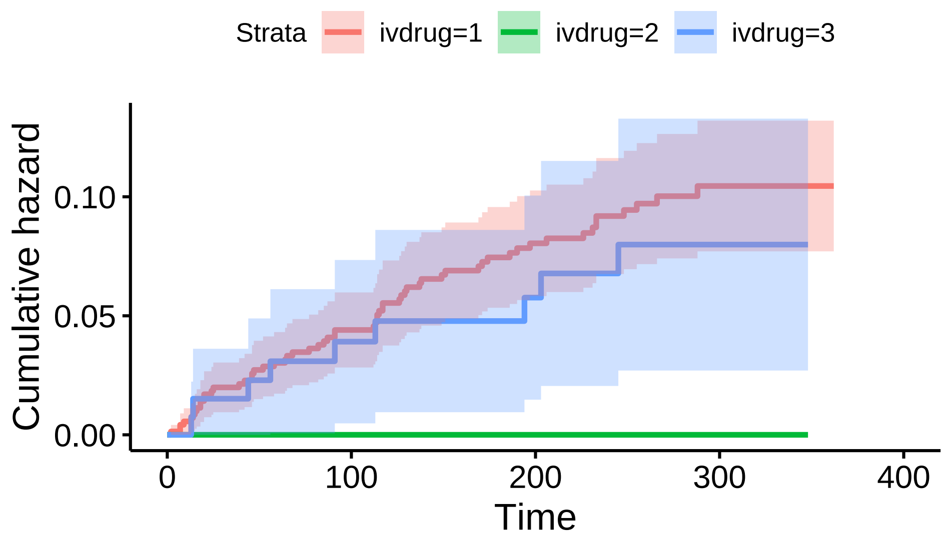

```
ggsurvplot(survfit(Surv(time,censor) ~ age, data=aids),
          censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on age")
```




```
ggsurvplot(survfit(Surv(time, censor)~hemophil, data = aids),
          censor=F, conf.int = T, fun = "cumhaz")
```




```r
ggsurvplot(survfit(Surv(time, censor)~ivdrug, data = aids),
           censor=F, conf.int = T, fun = "cumhaz")
```




```r
coxph(Surv(time,censor) ~ ivdrug, data=aids) %>% tidy()
```

```
## # A tibble: 1 x 7
##   term   estimate std.error statistic p.value conf.low conf.high
##   <chr>     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 ivdrug   -0.130     0.179    -0.723   0.470   -0.481     0.222
```

```r
coxph(Surv(time,censor) ~ ivdrug*karnof, data=aids) %>% tidy()
```

```
## # A tibble: 3 x 7
##   term          estimate std.error statistic p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 ivdrug        -0.711      1.71      -0.416 0.678    -4.07      2.64
## 2 karnof        -0.0903     0.0294    -3.07  0.00214  -0.148    -0.0326
## 3 ivdrug:karnof  0.00573    0.0201     0.285 0.775    -0.0336    0.0451
```

```r
coxph(Surv(time,censor)~sex+tx+age+txgrp, data = aids) %>% tidy()
```

```
## # A tibble: 9 x 7
```

```
##   term      estimate std.error statistic  p.value conf.low conf.high
##   <chr>        <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 sexmale     0.302      0.324   0.931    0.352    -0.333     0.937
## 2 tx         -0.790      0.256  -3.08     0.00205   -1.29     -0.288
## 3 age20-30   -0.424      1.07   -0.396    0.692     -2.52      1.67
## 4 age30-40   -0.214      1.02   -0.209    0.834     -2.21      1.79
## 5 age40-50   -0.0490     1.03   -0.0475   0.962     -2.07      1.97
## 6 age50-60    0.639      1.05    0.611    0.541     -1.41      2.69
## 7 age60-70   -0.328      1.42   -0.231    0.817     -3.11      2.46
## 8 ageover70 -14.1     2672.     -0.00528  0.996   -Inf       Inf
## 9 txgrp      NA          0      NA       NA         NA        NA
```

Patricia's "Something New"

I will be doing a power analysis by simulating survival analysis curves

1. What is the topic?

The topic is using sim.survdata in R to simulate survival data. Using that simulated data, we will make that the alternative and control for the coefficient beta by setting it equal to some value. Then using power analysis, we will see how many times we reject H_0.

2. How it is relevant? How it relates to survival analysis/analysis at hand?

Power analysis relates to survival analysis because if power is large after comparing our data to the simulated survival data, this tells us that there is a high chance that we would reject the null in favor of the alternative (control versus treatment?)

3. Resources to learn about the topic.

Below are some of the resources I have begun to use to learn about creating simulations of survival curves and performing power analysis:

a). https://cran.r-project.org/web/packages/coxed/vignettes/simulating_survival_data.html b). http://www.icssc.org/documents/advbiosgoa/tab%2026.00_survss.pdf

4. What will be challenging about learning something new?

Learning something new will be challenging because in this case, the concept of power analysis is something I just recently learned in Intro to Statistics. So learning to apply this concept in the context of survival analysis curves will be a challenge for me to learn. Learning how to simulate survival curves will also be challenging because I will have to learn how to use and interpret new functions in R.

Juste's "Something New"

I will be analyzing the Shoenfeld residuals for the Cox PH model.

1. What is goign on? What is the topic? 2. How it is relevant? How it relates to survival analysis/analysis at hand?

Cox proportional hazards (PH) model is considered a great way to identify combined effects of several covariates on the relative risk (hazard). This model assumes that the hazards of the different strata formed by the levels of the covariates are proportional. This proportional hazards assumption is particularly important and can be tested via three different clases of tests. The first class is focused on the piecewise estimation of models for subsets of data defined by stratification of time. The second one considers the interactions between covariates and some function of time. Final, third one is based on examinations of regression residuals. The Schoenfeld Residuals are a part of the third class of proportional hazard assumption testing and I will be exploring it in order to be able to eradicate a method for testing for the PH assumption in the current and future data set analyses. This topic is particularly important in relation to survival analysis since it provides an idea of whether the model is appropriate for the data set at hand and whether some covariates should be considered as variants of time in order to supply the best model for prediction of proportional hazards.

3. Resources to learn about the topic.

I have been researching articles and scientific journals that provide insights into tthe Schoenfeld residuals and their use in the Cox PH model. Sources include:

1. https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12176
2. https://rstudio-pubs-static.s3.amazonaws.com/39354_34153ff19e624116bd2fbdec7d2534aa.html

4. What will be challenging about learning something new?

Taking a completely new model of analyzing survival data is particulalrly difficult since the mathematical derivations and notations are also very varied from what we have seen in class. Although, I do remember some of the ideas behind parametric functions, their applications to statistical models are much more challenging than I have expected. Therefore, it will require me a lot of time and extensive research to be able to understand and learn how to apply this model to our data and other instances of survival analysis.

Explanation of the Theory Behind Schoenfeld Residuals

Let $z_{ij}(t)$ be the j^{th} covariate of the i^{th} unit, where $i = 1, 2, ..., n$ and $j = 1, 2..., p$

This notation indicates that z_{ij} is allowed to vary as a function of the time scale.

1) As we know from lecture, the Cox PH model assumes that $h(t)$ of the i^{th} individual satisfies:

- $h_i(t) = h_0(t)e^{z_i(t)\beta}$ where:
- h_0 -> baseline hazard
- $z_i(t)$ -> 1 x p vector of covariates for unit i each of which can be time fixed or time-varying.

2) However, another possibility has been presented by Therneau and Granbsh in 2000, where they proposed an idea that there ciuyld be an alternative to the current Cox model, where the coefficient of the estimate could also be varying as a function of time.

   The new hazard function would look like this: $h_i(t) = h_0(t)e^{z_i(t)\beta(t)}$

Therefore, in order to examine thee two models in a case when $\beta = \beta(t)$ requires a residual analysis that could indicate whether a model should consider a covariate as a variable with time.

Due to the fact that that some observations might be censored and in particular, regarding the Cox PH model, the baseline hazard is not estimated, in oprder to analyse the residuals a particular score process. The risk score for unit i at time t is thought to be $r_i(t) = e^{z_i(t)\beta}$, where $Y_i(t)$ is the indicator function and $Y_i(t) = 1$ indicates a point in which i is under risk and thus observation and it is equal to 0 in other occasions.

The Schoenfeld residuals are given by the equations:

1. $s_k = Z_{(k)} - \frac{\sum_i Y_i(t_k) r_i(t_k) Z_i(t_k)}{\sum_i Y_i(t_k) r_i(t_k)}$
2. $s_k = Z_{(k)} - \bar{z}(\hat{\beta}, t_k)$

In this case, the $Z(k)$ is the covariate vector of the particular unit that is experiencing the evnt at time k; $\hat{\beta}$ is the estimate of β and $\bar{z}(\hat{\beta}, t_k)$ is the wighted mean of covariate values.

Furthermore, the weighted variance can be represented by the derived equation at the k^{th} time as

$V(\beta, t_k) = \sum_i Y_i(t_k) r_i(t_k) Z_i(t_k) - \bar{z}(\hat{\beta}, t_k)' Z_i(t_k) - \frac{\bar{z}(\hat{\beta}, t_k)}{\sum_i Y_i(t_k) r_i(t_k)}$

From this, we can scale the Schoenfeld residuals by $V(\beta, t_k)$ of X at t_k via the equation:

$s_k^* = V^{-1}(\hat{\beta}, t_k) s_k$

The scaled Schoenfeld residuals can also be defined as follows:

$s_k^* = m \sum_{k=1}^{d} V(\hat{\beta}, t_k) s_k$

here, m is the total number of deaths in the data set.

Following the calculations, the residuals are plotted against time in order to test the prportional hazards assumption. If the assumption is correct, the residuals should be fiting around the line centered at zero (y=0). The further away this predicted line is form the horizontal of (y=0) the more likely one is to call the PH assumption to question and determine whether it is met through the model.

> To go a little deeper into the analysis of the resiaul calculation, one can look at the calculations of the test statistic for this residual mdoel.

By producing a least squares slope of regression and assuming a relationship between s_{kj}^* and t_{kj} or some function $g(t_k)$ allows to derive a test statistic for the proportional hazards assumption in regards to the j^{th} covariate, which is given by:

$T_j = \frac{[\sum_{k=1}^{d} (g(t_k) - \hat{g}) s_{kj}^{8}]^2}{d I^{jj} \sum_{k=1}^{d} (g(t_k) - \hat{g})^2}$

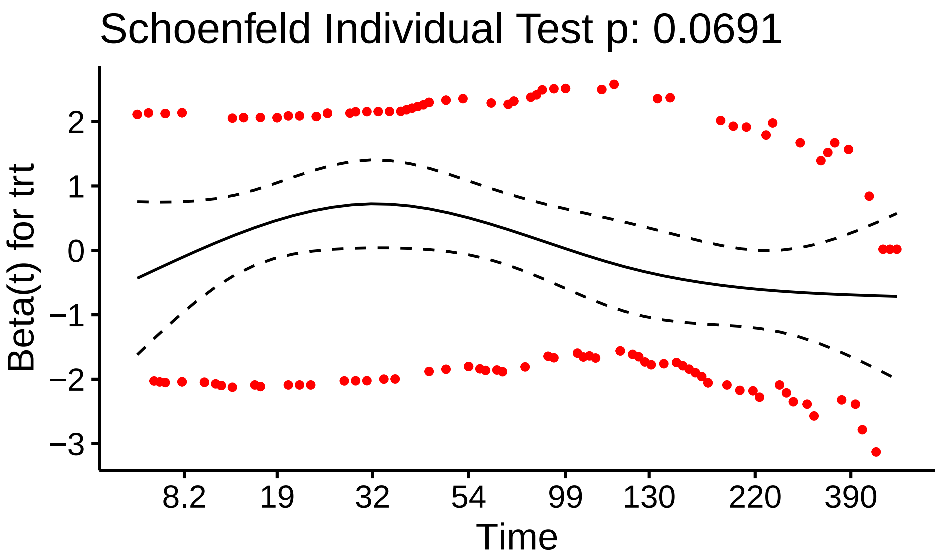
Here, the distribution is asymptotical as $X^2(1)$ stating the null hypothesis that the relationship between the covariate, in this case j and the evnt time follows the assumption of PH.

Schoenfeld:

```
veteran.ph <- coxph(Surv(time,status) ~ trt, data=veteran)
cox.veteran <- cox.zph(veteran.ph)
cox.veteran
```

```
##        rho chisq      p
## trt -0.16   3.3 0.0691
```

```
ggcoxzph(cox.veteran)
```




```r
ggcoxdiagnostics(veteran.ph, type="schoenfeld")
```