Session 10. Survival analysis

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Santiago Perez-Hoyos

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca 29/10/2020

Outline

- Introduction to Survival analysis
- Data structure
- Survival & hazard functions
- Kaplan-Meier Estimation
- Plotting survival curves
- Comparing survival curves

What do we need in survival analysis?

- Group of individuals followed from a time point(origin) up to an event in time T

Origin	Event
Birth or entry time	Death
Remission	Recurrence
HIV Infection	Aids
Disease diagnosis	Healing

Objectives of survival analysis

- **Calculate probability of event free at time T**

(ESTIMATION)

- **Compare survival experiences among groups**

(HYPOTHESIS TEST)

- **Analyse risk factors related to survival**

(REGRESSION)

Required Data

Origin

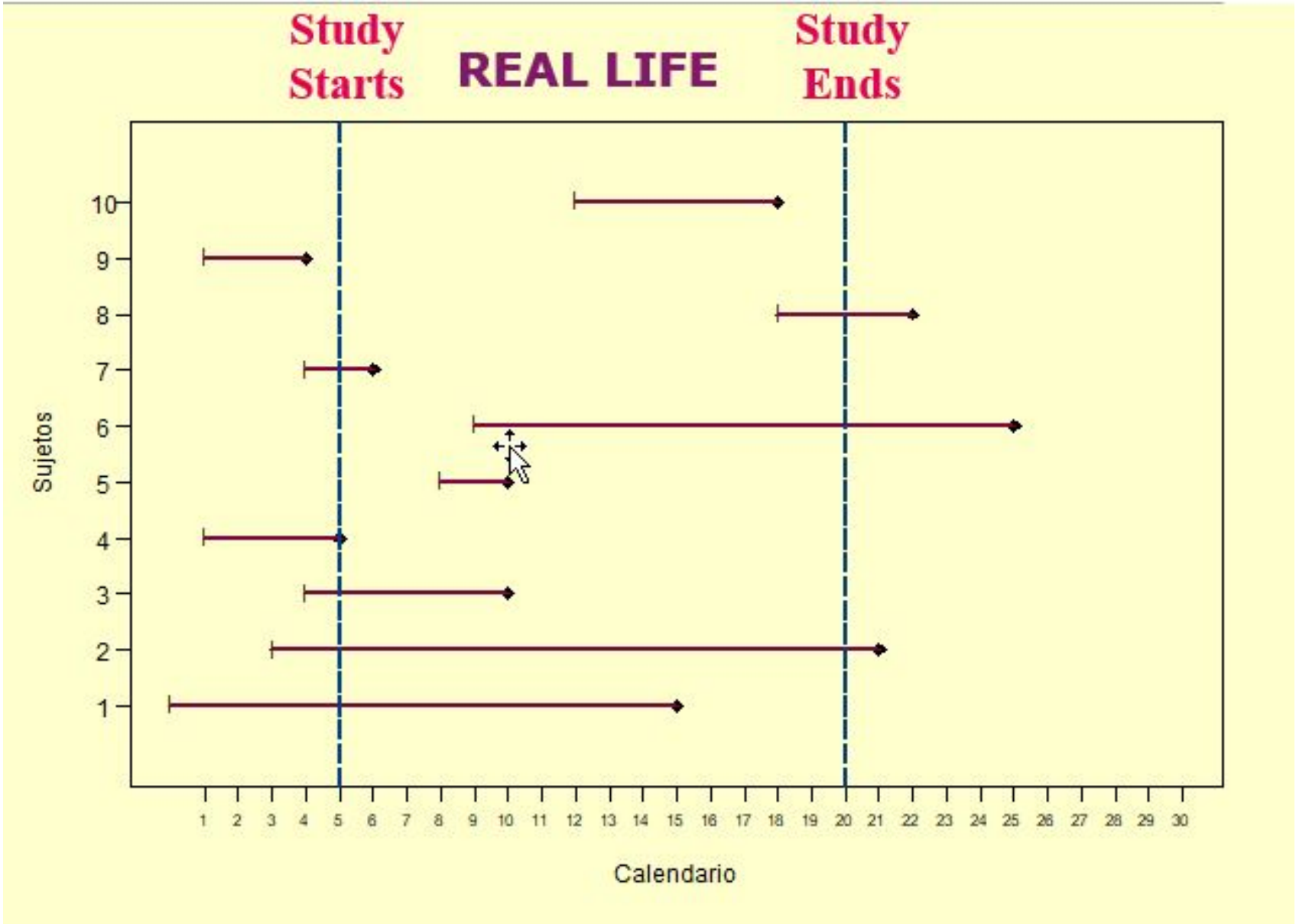
Infection, Hospital Admission, Randomized time, Diagnosis, Surgery etc.

Time Scale

Years, Months, Days, Seconds

Event

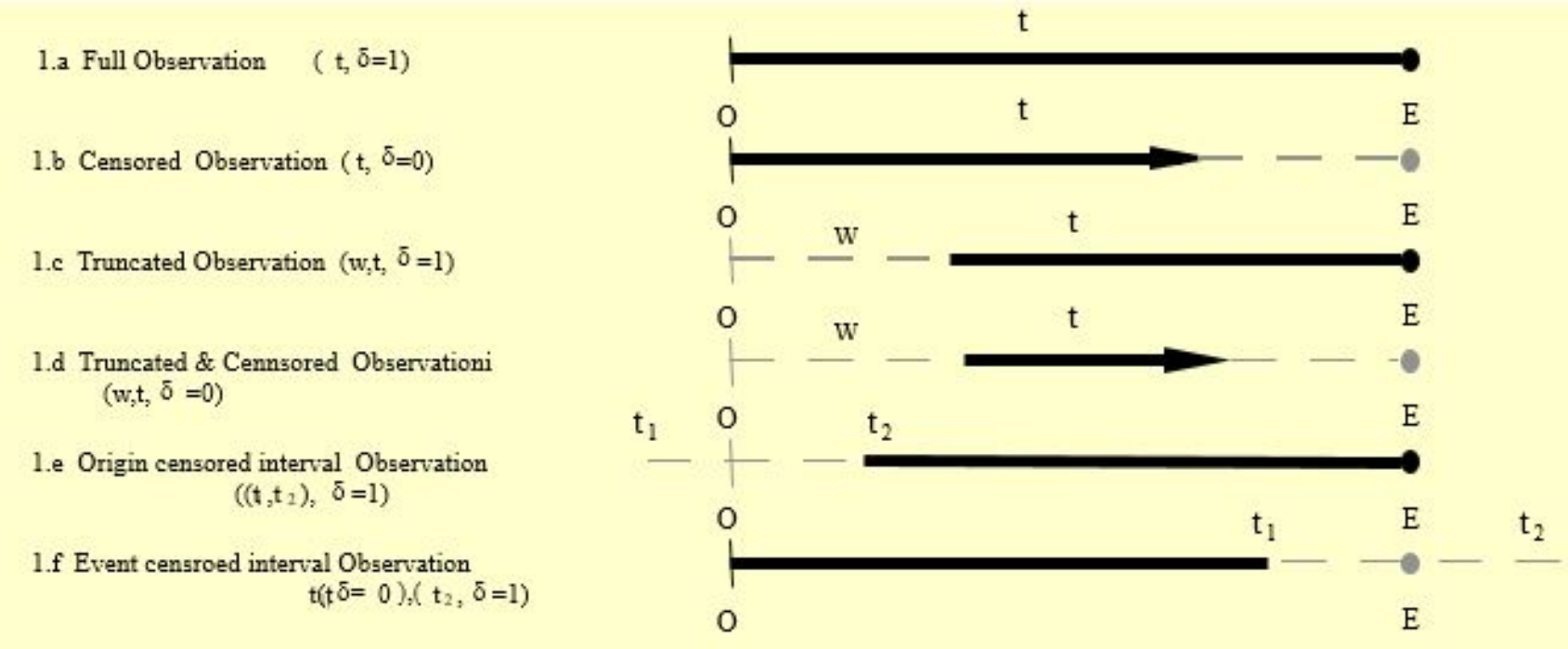
AIDS, Cancer onset, Cure, Recurrence, Death


Study
Starts
REAL LIFE
Study
Ends
Sujetos
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Calendario

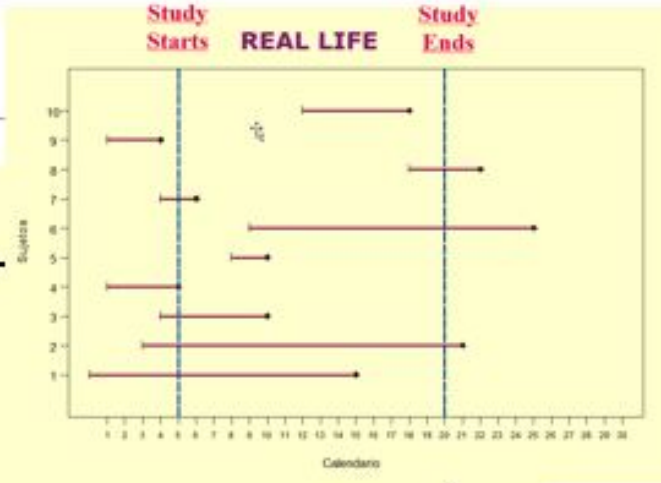
Problem with Data

Censoring

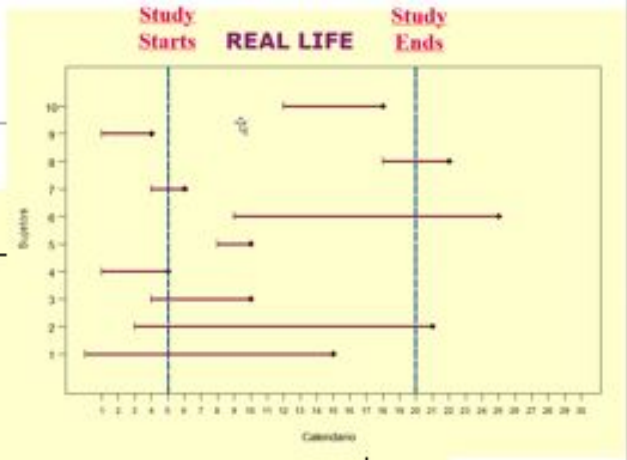
Truncating

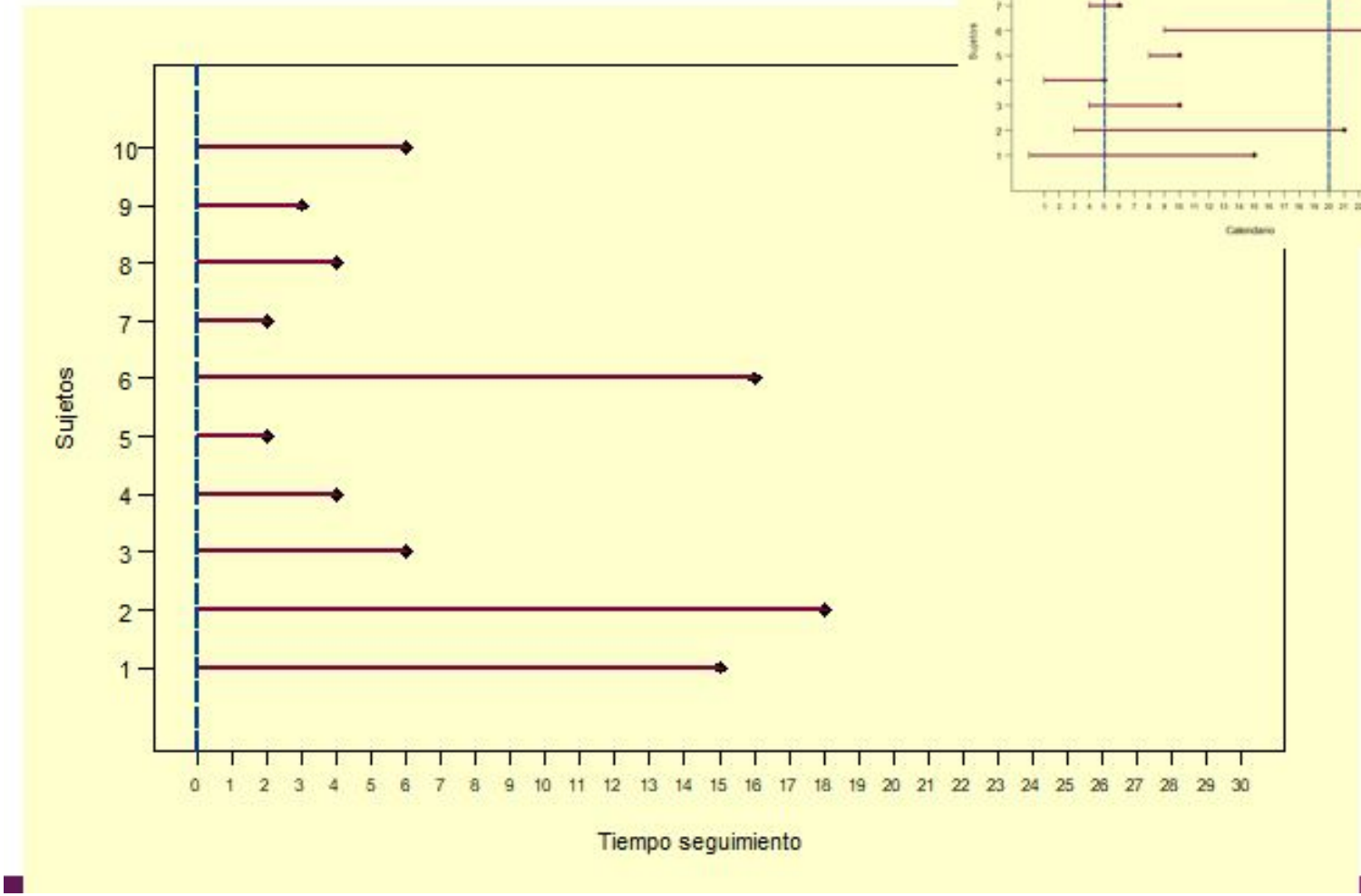


Follow up times observed



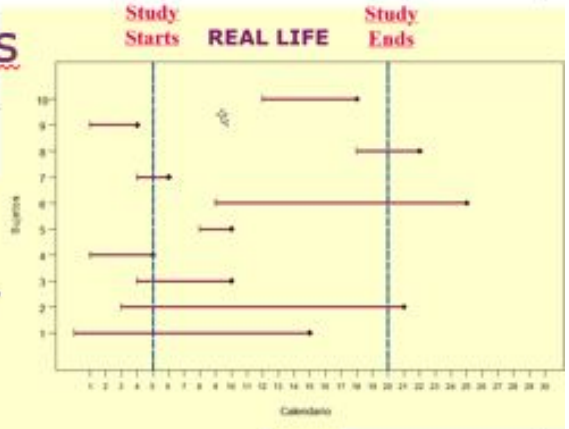
Real Follow up times not observed

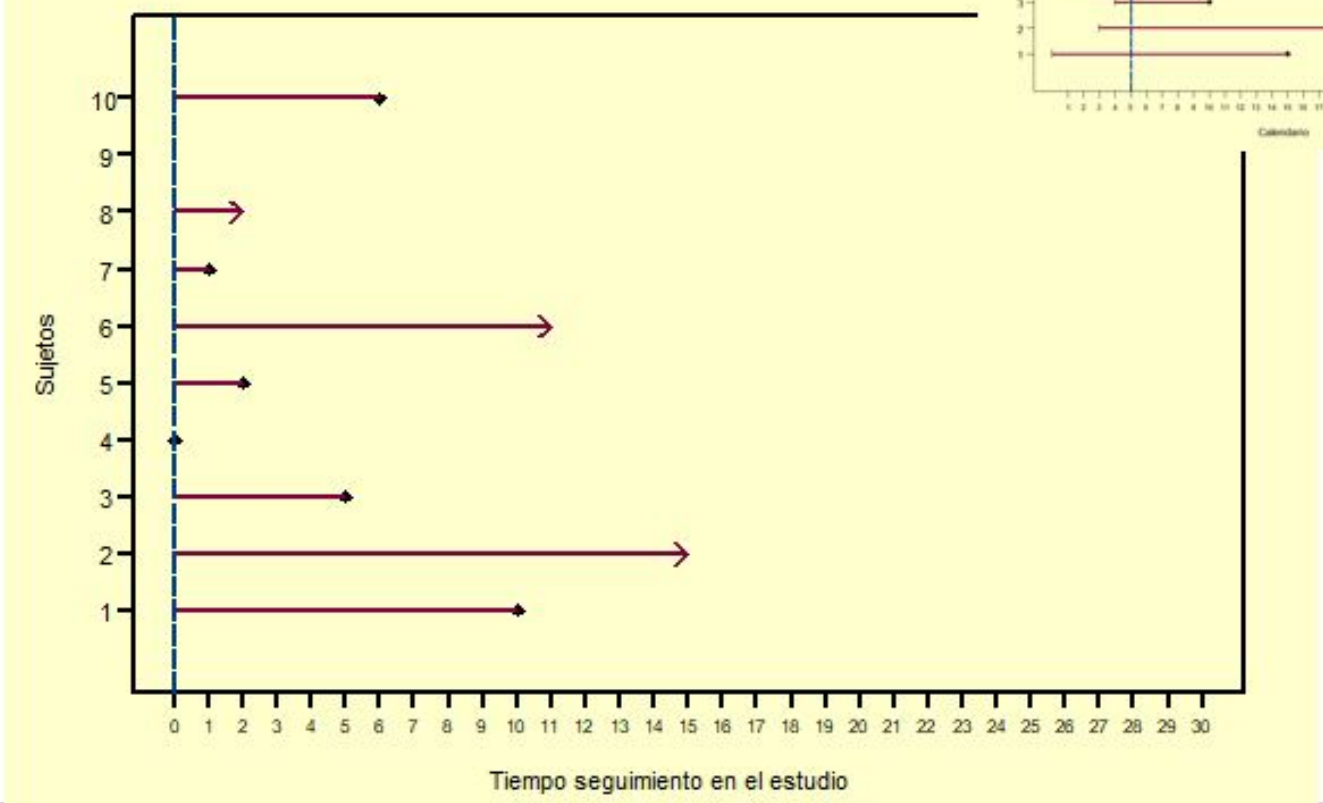




Follow up times frequently used in analysis

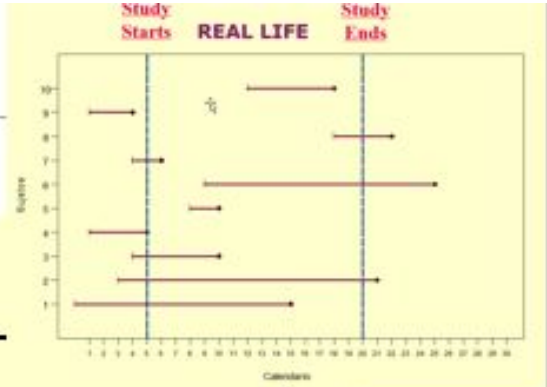
(BIASED)

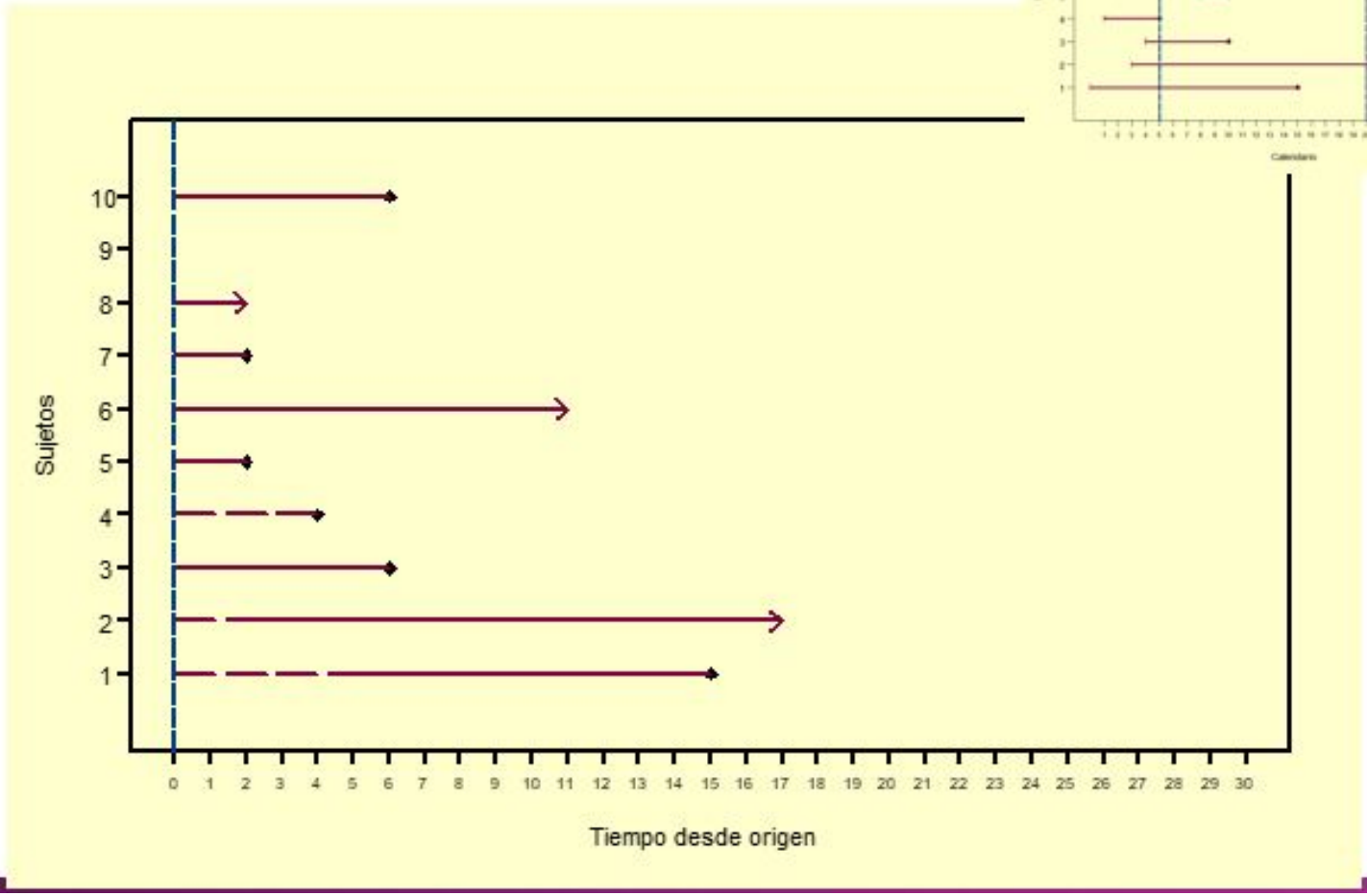




Follow up times that should be used in analysis

Truncated data

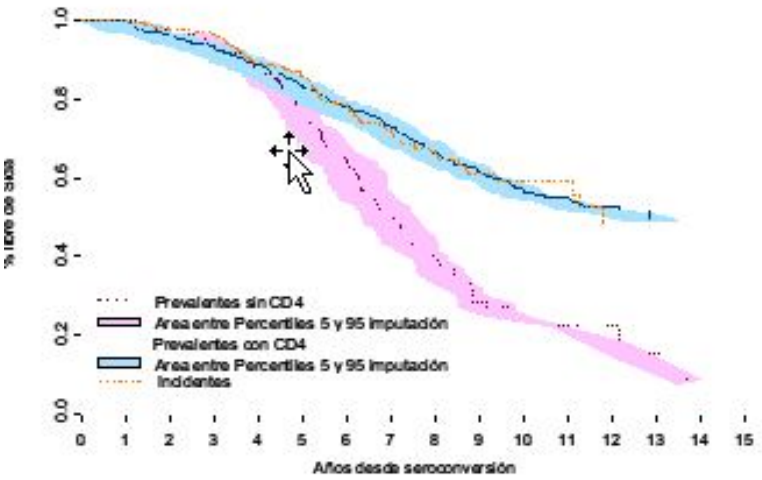


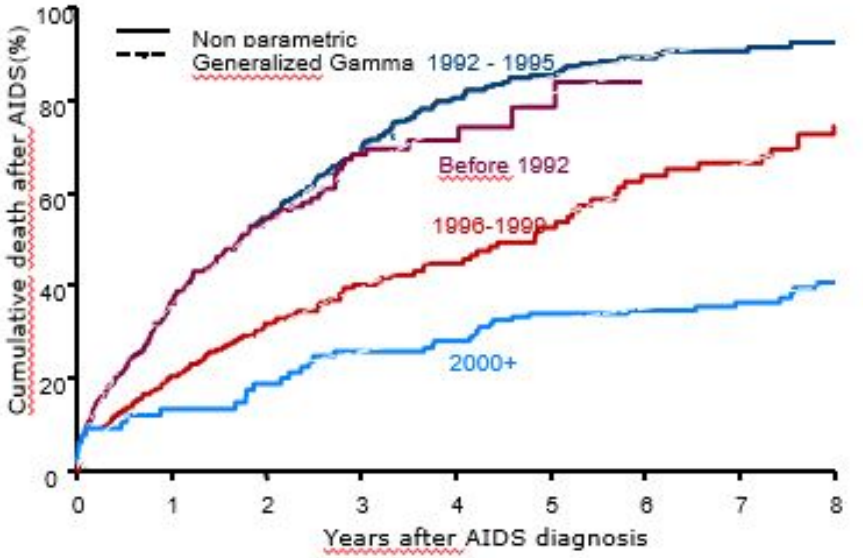


Survival Function

S(t)=Prob(Survive t)=P{T>t}=1-P(die before t)

f(t)= instantaneous probability of death (density)





Hazard Rate $\lambda(t) = h(t)$

Instantaneous probability of death in an infinitesimal interval knowing to be alive at the beginning

$$\lambda(t) = Prob(\frac{die\ \ between\ \ t,t+\triangle t)}{Alive\ \ at\ \ t} = f(t)/S(t)$$

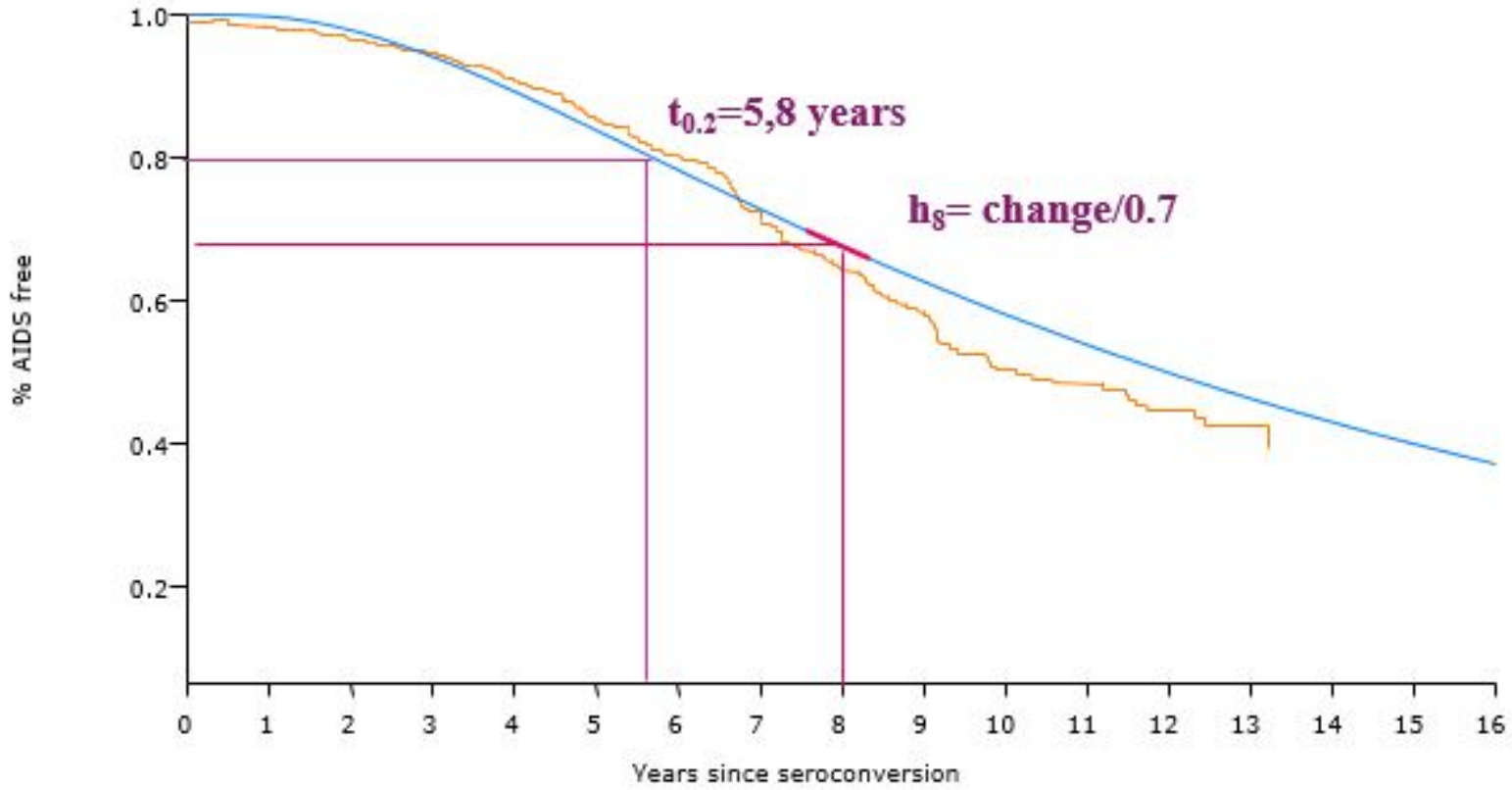
Cumulative Hazard Rate

$$\Lambda(t) = H(t) = \int \lambda(u)d(u) = -log(S(t))\ S(t) = e^{\Lambda(t)}$$

Interpretation of Survival and Hazard

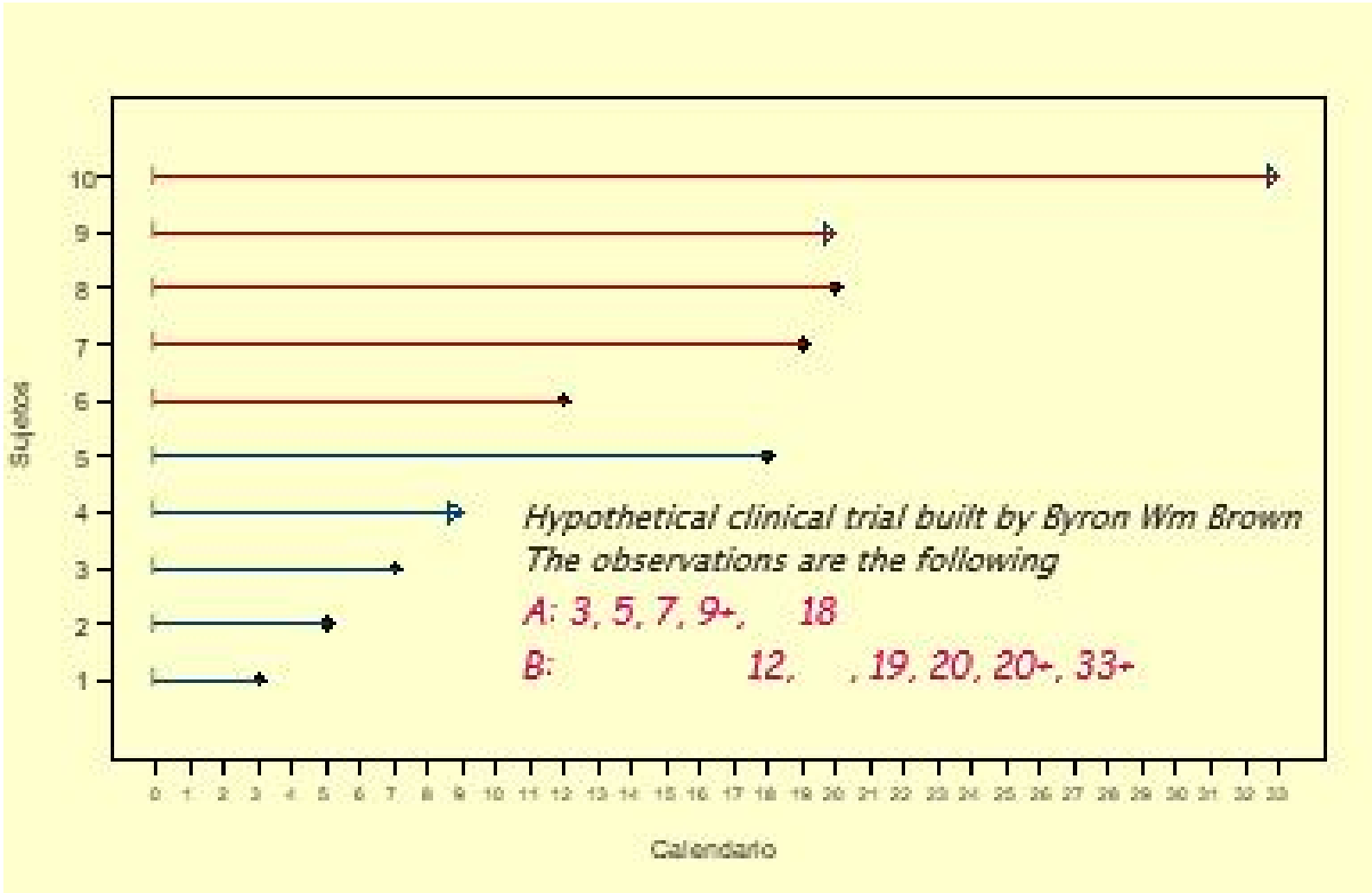
Hazard = Speed of ocorrence of events

Survival= Percentage of alive at a time



Brown Data

- Example data with 10 cases in 2 groups



Required libraries

```
library(readxl)
library(dplyr)
library(ggplot2)
library(plotly)

library(survival)
library(survminer)
library(gtsummary)
```

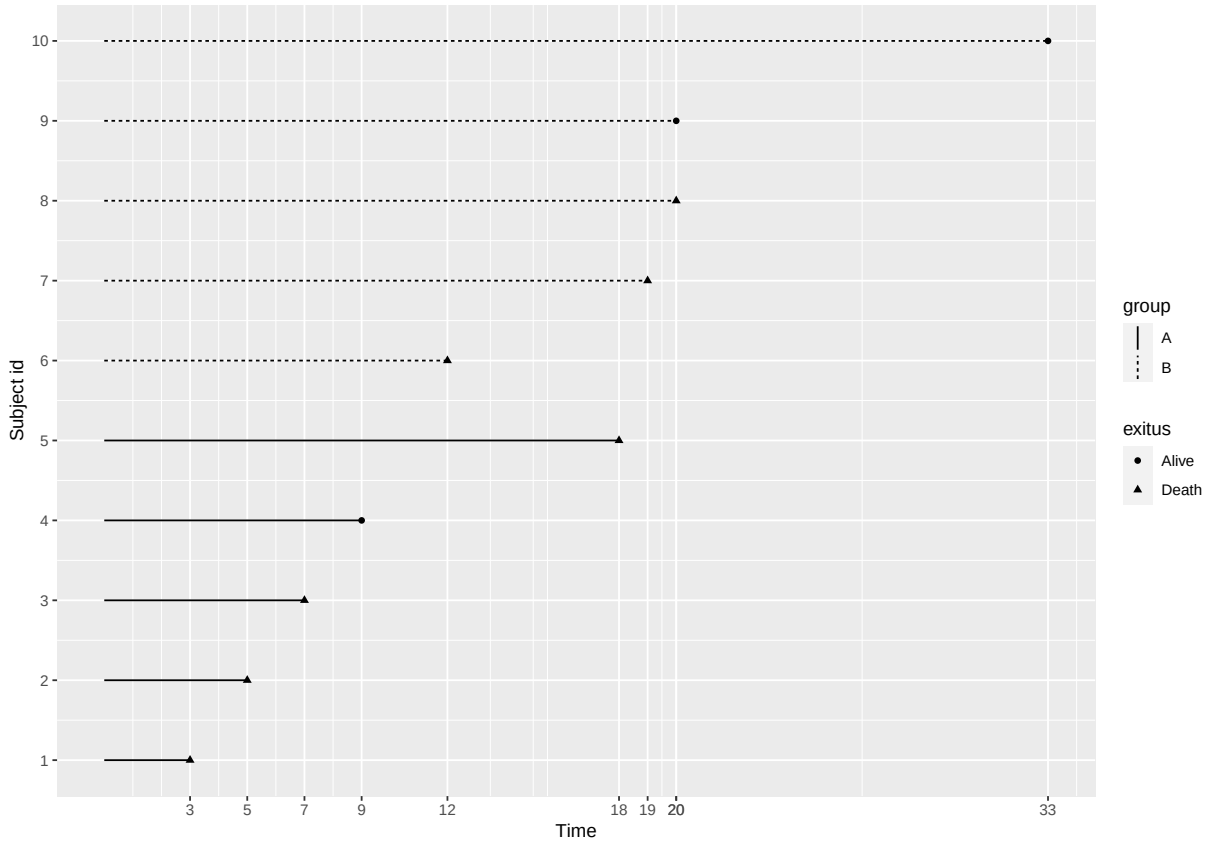
Create data

```
id<-seq(1,10,1)
time<-c(3,5,7,9,18,12,19,20,20,33)
time0<-rep(0,10)
exitus<-c("Death","Death","Death","Alive","Death"
          ,"Death","Death","Death","Alive","Alive")
group<-c("A","A","A","A","A","B","B","B","B","B")

dat<-tibble(id,time,time0,exitus,group)
```

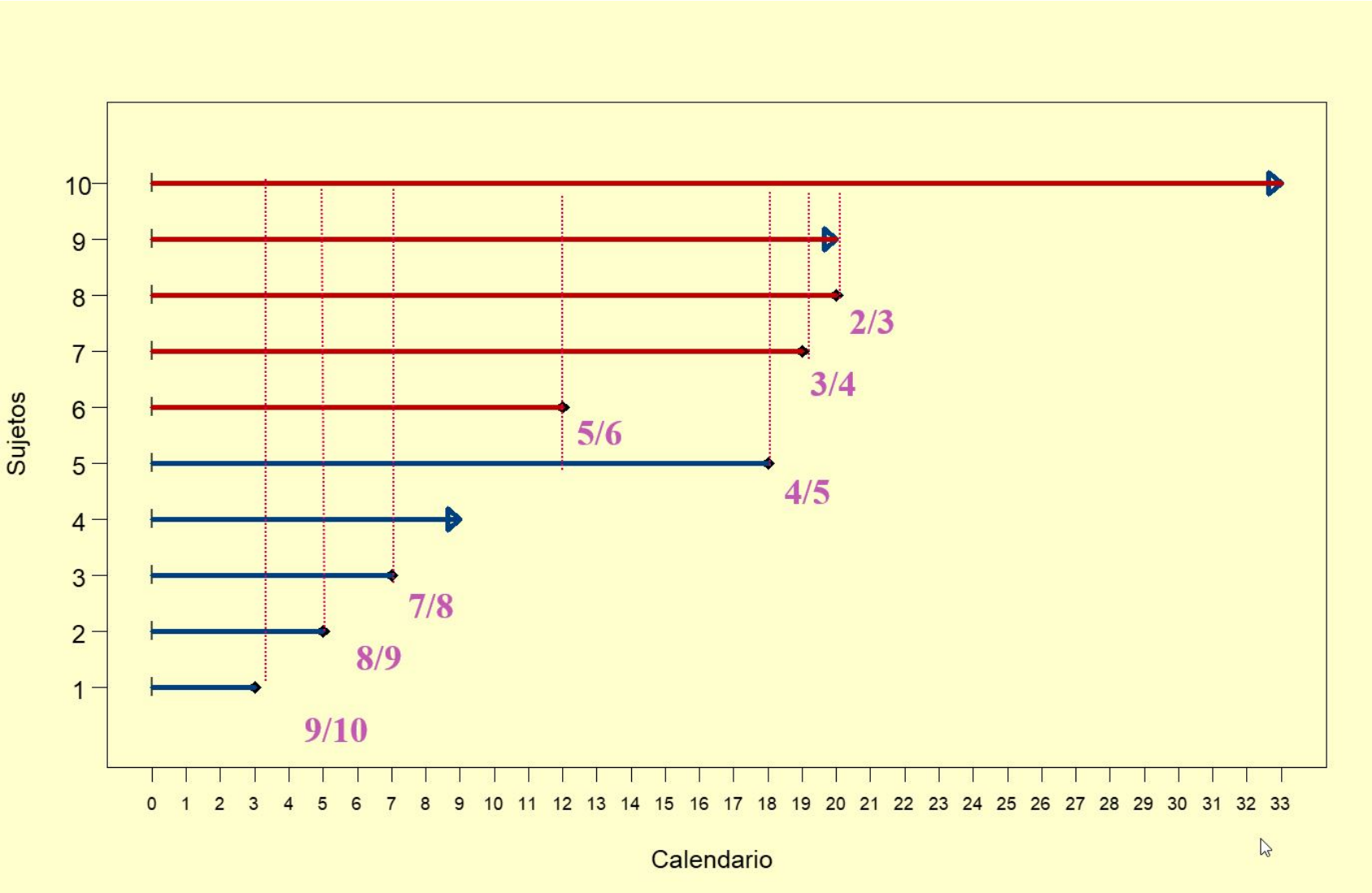
Graph the Data in R

```
ggplot(dat, aes(x = id))+
            geom_linerange(aes(ymin = time0, ymax = time,linetype=group)) +
            geom_point(aes(y = time, shape =exitus))+
            coord_flip()+
            labs( y="Time", x="Subject id")+
            scale_x_continuous(breaks=seq(1,10,1))+
            scale_y_continuous(breaks=c(3,5,7,9,18,12,19,20,20,33))
```



Kaplan-Meier Estimator

Probability to pass t / alive at t-



Kaplan-Meier Estimator

- *No censoring*

  S(t)= Nº survivors after t/individuals at the beginning of follow-up

- *Censoring. Kaplan-Meier Estimator*

$$S_{K\text{-}M}(t) = \prod_{t_i<t} (1 - \frac{1_i}{R_i})^{\delta_i}$$

Time	At Risk at t R_i	P(event t/ alive t) q_i	P(alive after t/ alive t) $p_i=1-q_i$	Kaplan Meier at t KM
3	10	1/10	9/10	9/10
5	9	1/9	8/9	9/10*8/9
7	8	1/8	7/8	9/10*8/9*7/8
12	6	1/7	6/7	9/10*8/9*7/8*6/7
18	5	1/5	4/5	9/10*8/9*7/8*6/7*4/5
19	4	1/4	4/3	9/10*8/9*7/8*6/7*4/5*4/3
20	3	1/3	2/3	9/10*8/9*7/8*6/7*4/5*4/3*2/3

Define Survival data

```
KM_fit <- survfit(Surv(time, exitus=="Death") ~ 1, data =dat)
KM_fit
```

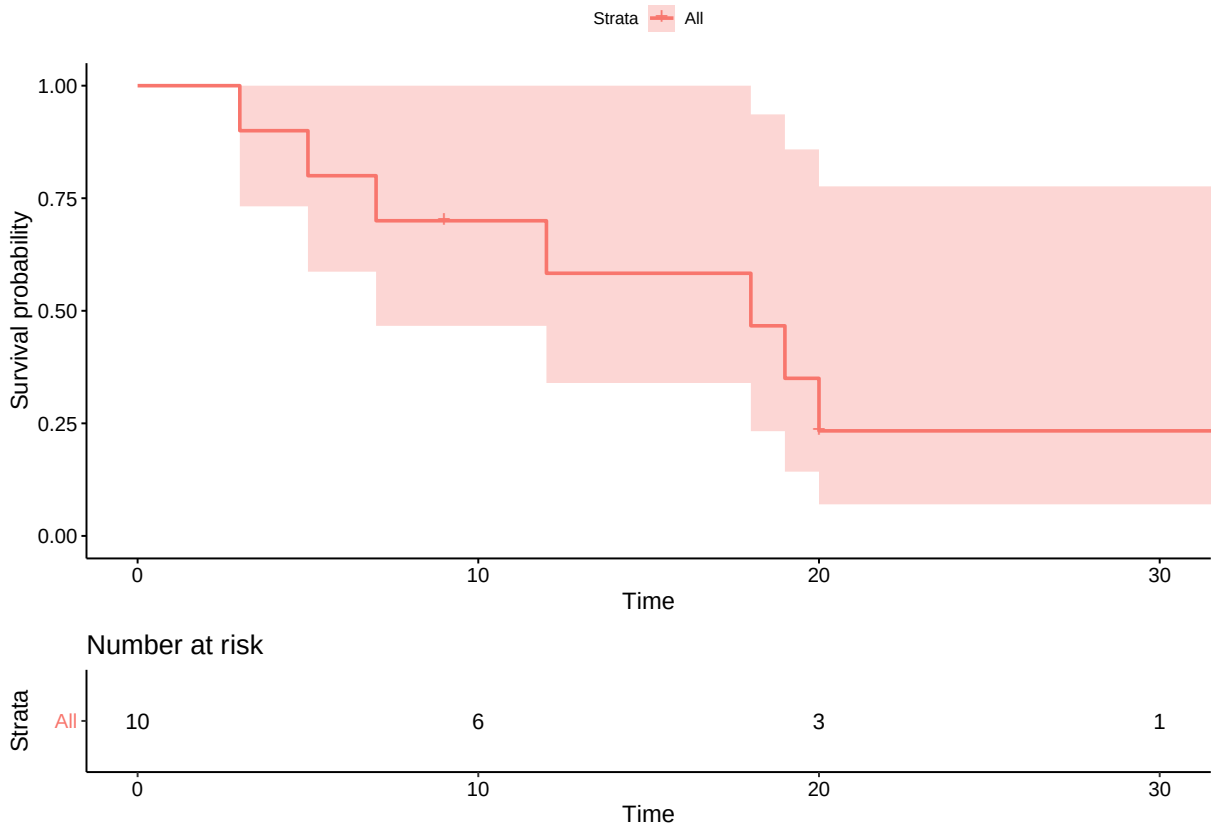
```
## Call: survfit(formula = Surv(time, exitus == "Death") ~ 1, data = dat)
##
##       n  events  median 0.95LCL 0.95UCL
##      10       7      18       7      NA
```

```
summary(KM_fit)
```

```
## Call: survfit(formula = Surv(time, exitus == "Death") ~ 1, data = dat)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3     10       1    0.900  0.0949       0.7320        1.000
##     5      9       1    0.800  0.1265       0.5868        1.000
##     7      8       1    0.700  0.1449       0.4665        1.000
##    12      6       1    0.583  0.1610       0.3396        1.000
##    18      5       1    0.467  0.1658       0.2326        0.936
##    19      4       1    0.350  0.1602       0.1427        0.858
##    20      3       1    0.233  0.1431       0.0701        0.776
```

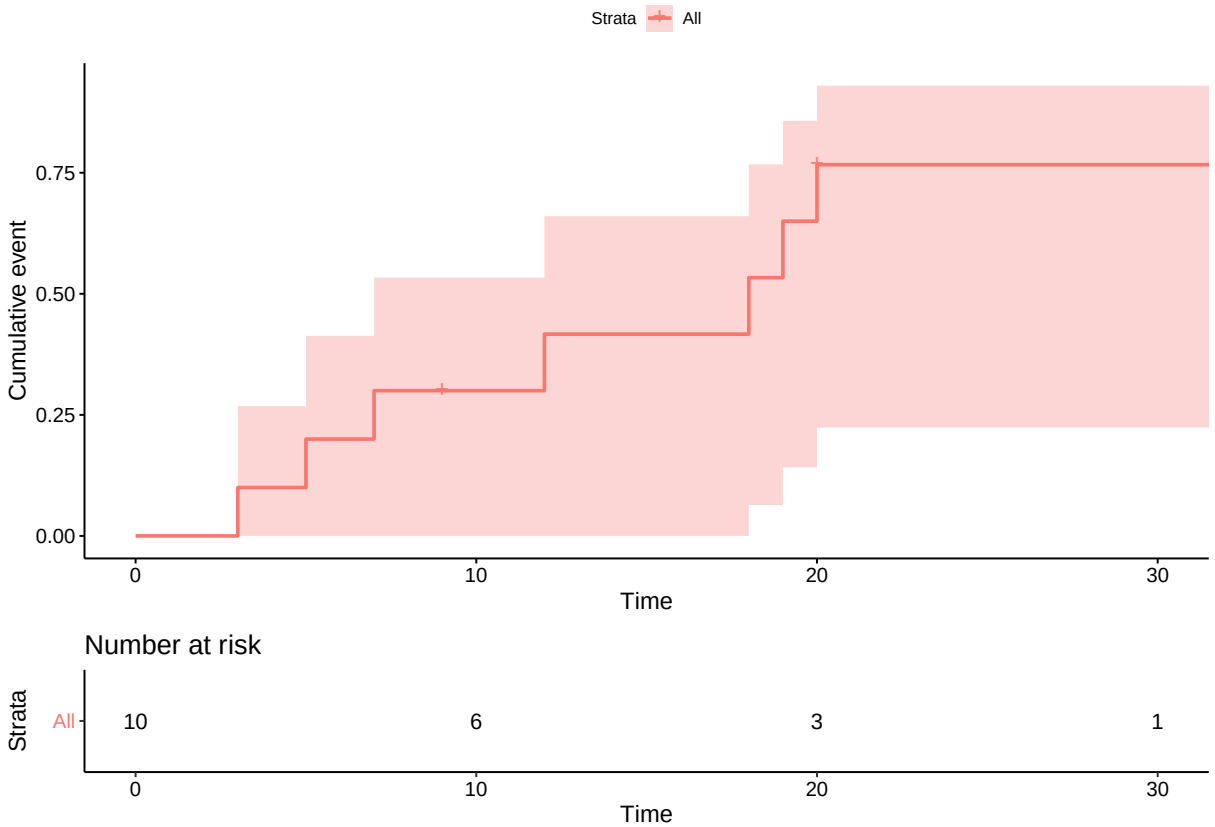
Plot Kaplan-Meier curve

```
ggsurvplot(KM_fit, data = dat,risk.table = TRUE )
```



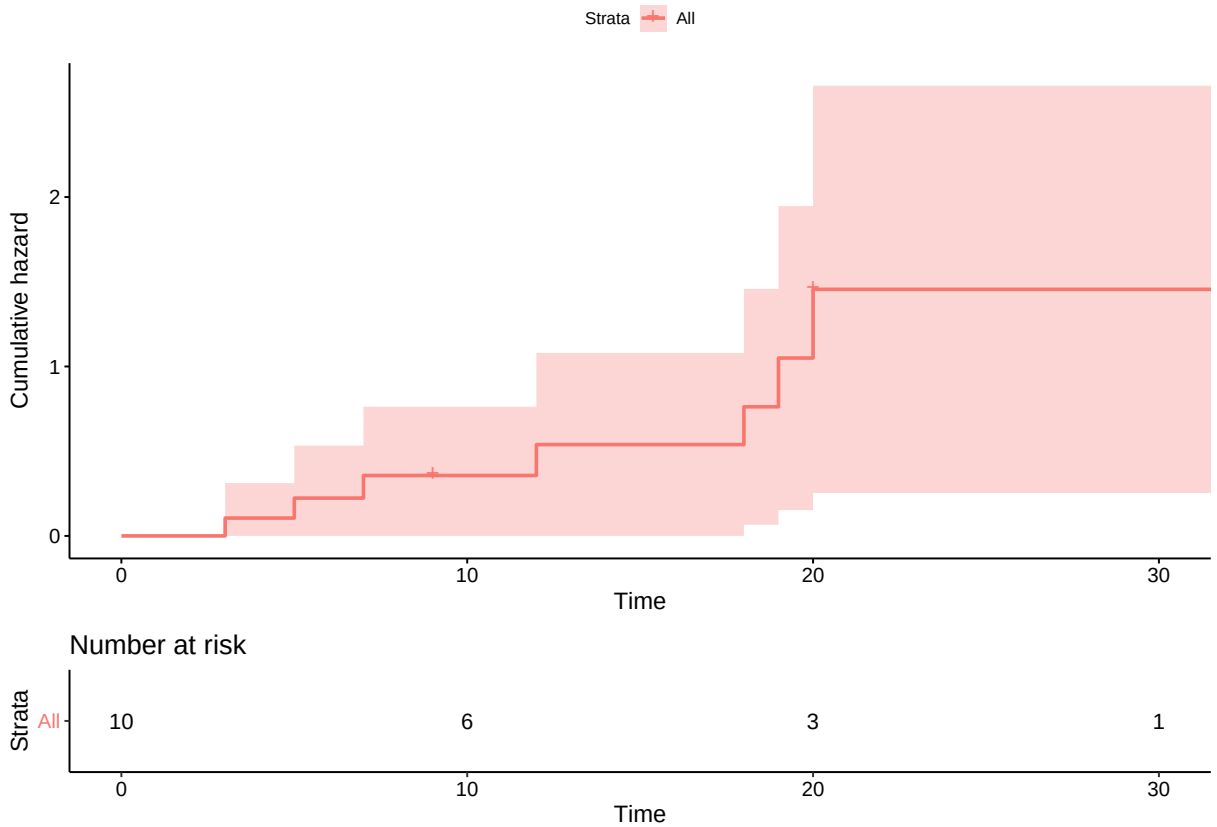
Plot Kaplan-Meier curve (Cumulative events)

```
ggsurvplot(KM_fit, data = dat,risk.table = TRUE ,fun="event")
```



Plot Kaplan-Meier curve (Cumulative hazard)

```
ggsurvplot(KM_fit, data = dat,risk.table = TRUE ,fun="cumhaz")
```



Compare Curves

```
KM_fit2 <- survfit(Surv(time, exitus=="Death") ~ group, data =dat)
KM_fit2
```

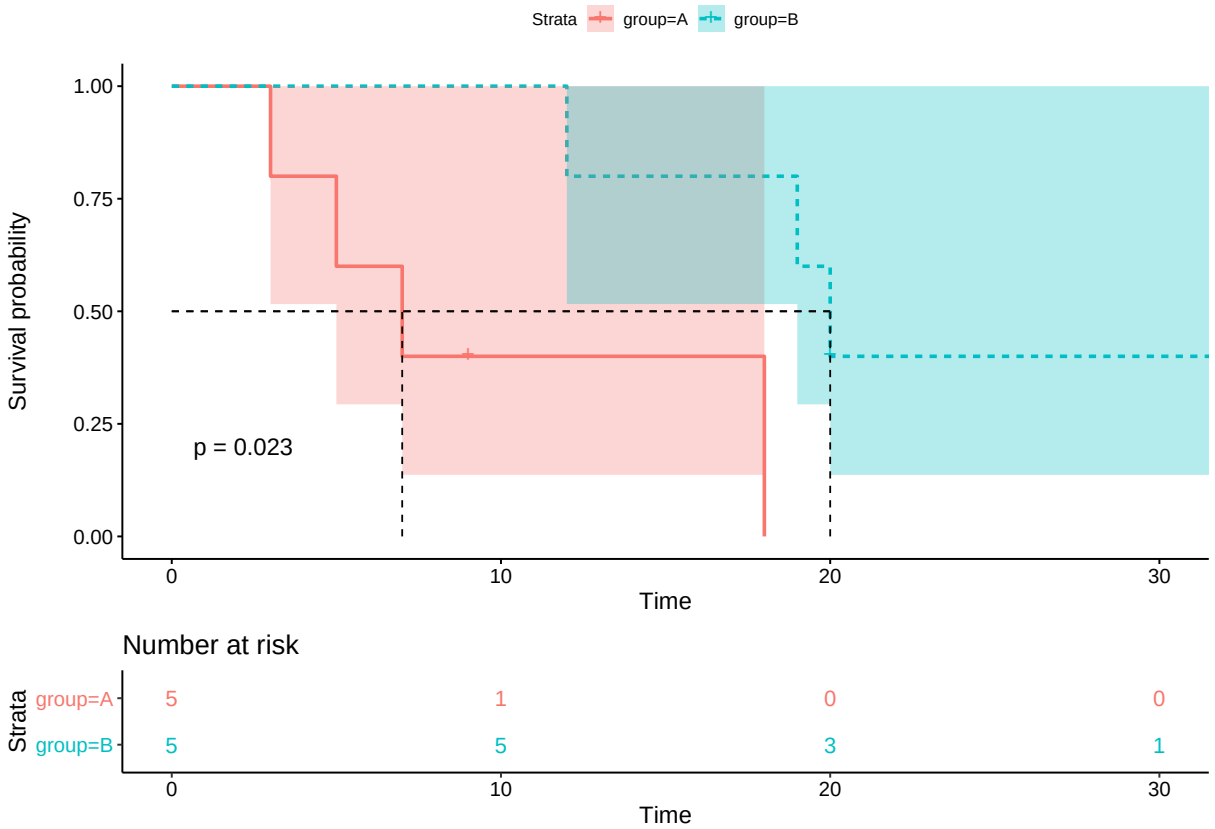
```
## Call: survfit(formula = Surv(time, exitus == "Death") ~ group, data = dat)
##
##         n events median 0.95LCL 0.95UCL
## group=A 5      4      7       5      NA
## group=B 5      3     20      19      NA
```

```
summary(KM_fit2)
```

```
## Call: survfit(formula = Surv(time, exitus == "Death") ~ group, data = dat)
##
##                 group=A
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3      5       1      0.8   0.179        0.516            1
##     5      4       1      0.6   0.219        0.293            1
##     7      3       1      0.4   0.219        0.137            1
##    18      1       1      0.0     NaN           NA           NA
##
##                 group=B
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    12      5       1      0.8   0.179        0.516            1
##    19      4       1      0.6   0.219        0.293            1
##    20      3       1      0.4   0.219        0.137            1
```

Plot Kaplan-Meier curves and compare

```
ggsurvplot(KM_fit2, data = dat,
           pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # Add risk table)
             risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv") # Specify median survival
```



Exercise 1

- Read Diabetes data
- Calculate global survival curve (time=tempsviu, death=mort)
- Plot Kaplan Meier curve
- Are differences between ecg and chd?
- Calculate tables and plots

Read Diabetes Data

```
diabetes <- read_excel("datasets/diabetes.xls")
sapply(diabetes, class)
```

```
##    numpacie        mort    tempsviu        edat         bmi    edatdiag
##   "numeric" "character"   "numeric"   "numeric"   "numeric"   "numeric"
##       tabac         sbp         dbp         ecg         chd
## "character"   "numeric"   "numeric" "character" "character"
```

```
diabetes_factor <- diabetes %>%
  mutate_if(sapply(diabetes, is.character), as.factor) %>%
  select (-numpacie)
sapply(diabetes_factor, class)
```

```
##      mort  tempsviu      edat       bmi  edatdiag     tabac       sbp       dbp
##  "factor" "numeric" "numeric" "numeric" "numeric"  "factor" "numeric" "numeric"
##       ecg       chd
##  "factor"  "factor"
```

```
dat2<-diabetes_factor
```

Define Survival data

```
KM_fit <- survfit(Surv(tempsviu, mort=="Muerto") ~ 1, data =dat2)
KM_fit
```

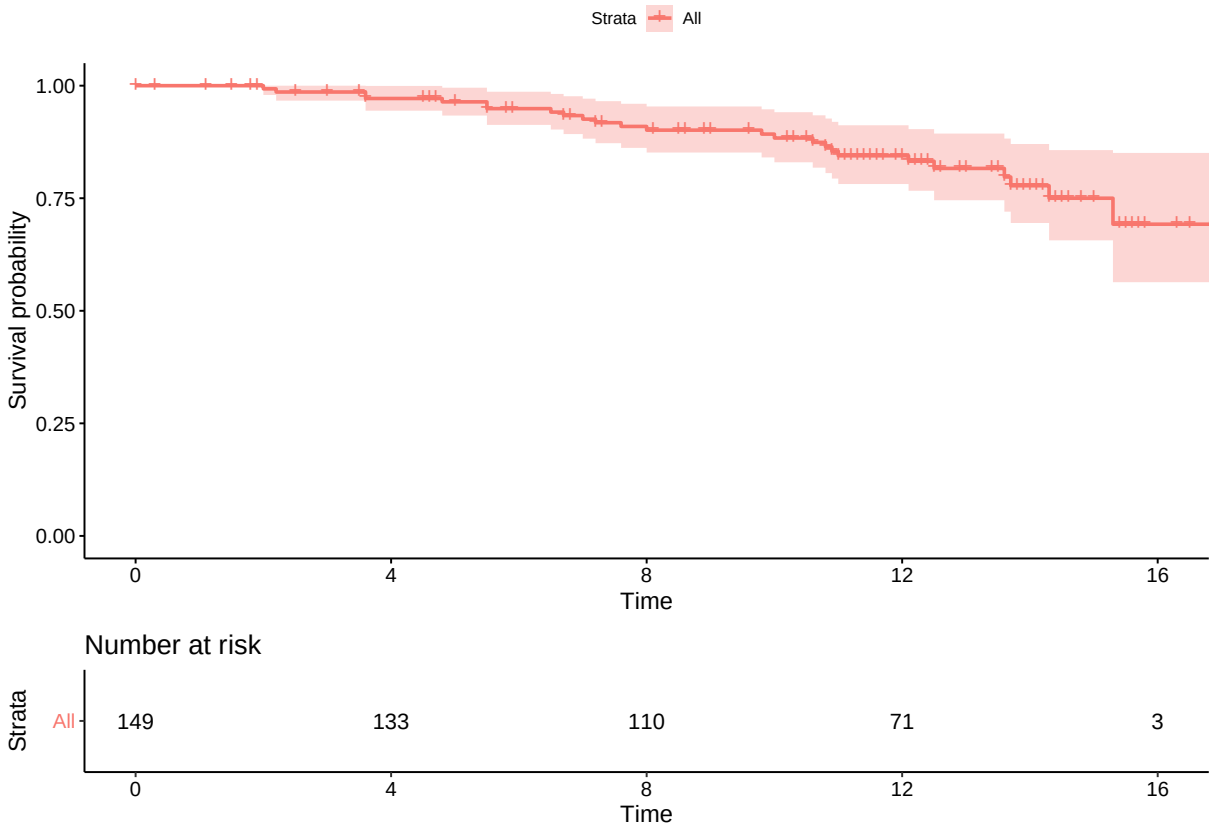
```
## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)
##
##       n  events  median 0.95LCL 0.95UCL
##     149      25      NA      NA      NA
```

```
summary(KM_fit,times=c(0,2,4,6,8,10))
```

```
## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ 1, data = dat2)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    149       0    1.000 0.00000        1.000        1.000
##     2    142       1    0.993 0.00702        0.979        1.000
##     4    133       3    0.972 0.01404        0.944        0.999
##     6    123       3    0.949 0.01880        0.913        0.987
##     8    110       6    0.901 0.02609        0.851        0.954
##    10    102       2    0.884 0.02837        0.830        0.941
```

Plot Kaplan-Meier curve

```
ggsurvplot(KM_fit, data = dat2,risk.table = TRUE )
```



Compare Curves

```
KM_fit2 <- survfit(Surv(tempsviu, mort=="Muerto") ~ ecg, data =dat2)
KM_fit2
```

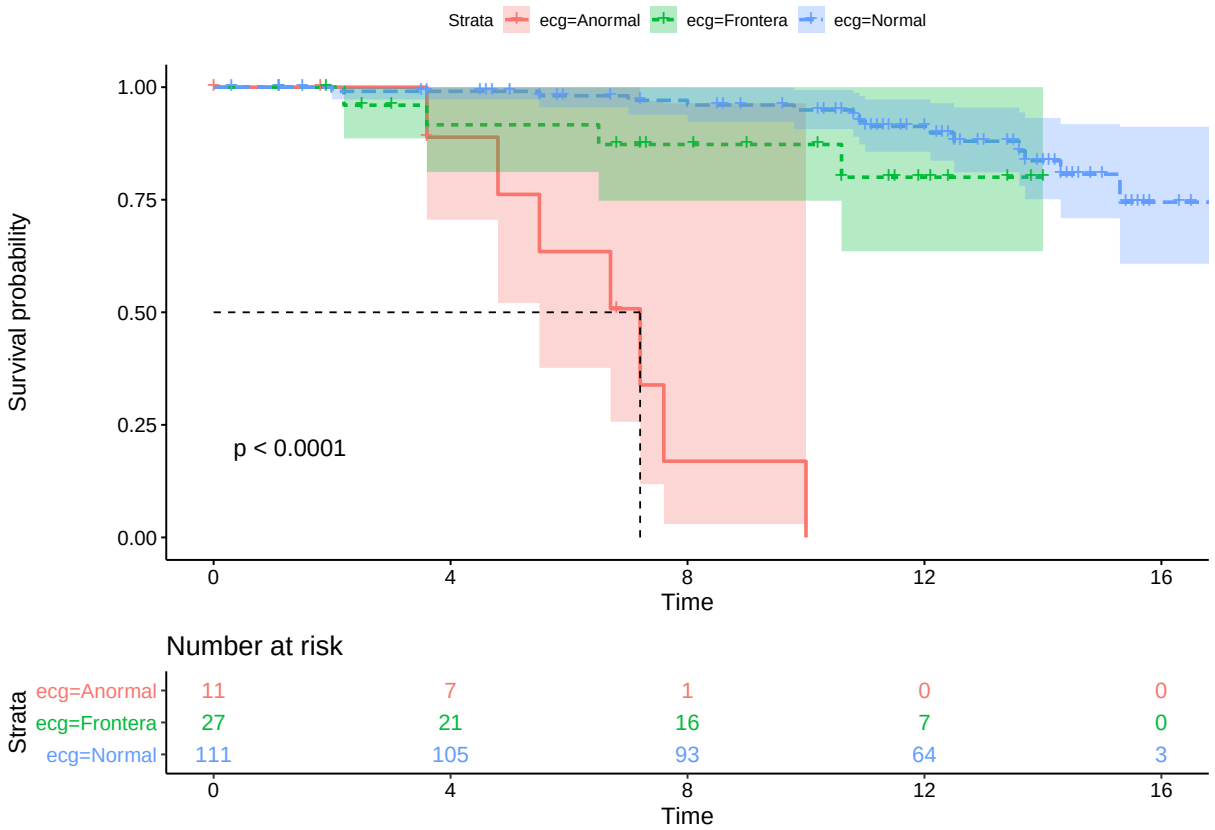
```
## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ ecg, data = dat2)
##
##                n events median 0.95LCL 0.95UCL
## ecg=Anormal   11      7    7.2     5.5      NA
## ecg=Frontera  27      4     NA      NA      NA
## ecg=Normal   111     14     NA      NA      NA
```

```
summary(KM_fit2, times=c(0,5,10))
```

```
## Call: survfit(formula = Surv(tempsviu, mort == "Muerto") ~ ecg, data = dat2)
##
##                 ecg=Anormal
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0     11       0    1.000   0.000        1.000            1
##     5      6       2    0.762   0.148        0.521            1
##    10      1       5    0.000     NaN           NA           NA
##
##                 ecg=Frontera
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0     27       0    1.000  0.0000        1.000            1
##     5     21       2    0.916  0.0567        0.812            1
##    10     14       1    0.873  0.0688        0.748            1
##
##                 ecg=Normal
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    111       0    1.000 0.00000        1.000        1.000
##     5    102       1    0.991 0.00922        0.973        1.000
##    10     87       4    0.949 0.02212        0.907        0.994
```

Plot Kaplan-Meier curves and compare

```
ggsurvplot(KM_fit2, data = dat2,
           pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # Add risk table)
             risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv") # Specify median survival
```



Cox Regression Model

- Semiparametric model that fit hazard rate $\lambda(t)$

$$(t|X_i) = h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Assume non-informative censoring
- Assume proportional hazards

Cox Regression for Brown data

```
Coxfit<-coxph(Surv(time, exitus=="Death") ~ group, data = dat)
Coxfit
```

```
## Call:
## coxph(formula = Surv(time, exitus == "Death") ~ group, data = dat)
##
##          coef exp(coef) se(coef)      z      p
## groupB -2.254     0.105    1.155 -1.951 0.0511
##
## Likelihood ratio test=4.95  on 1 df, p=0.02608
## n= 10, number of events= 7
```

```
summary(Coxfit)
```

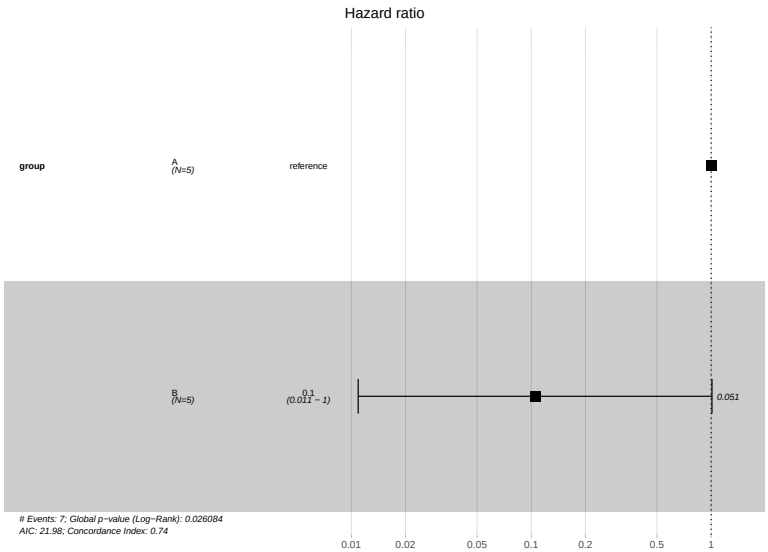
```
## Call:
## coxph(formula = Surv(time, exitus == "Death") ~ group, data = dat)
##
##   n= 10, number of events= 7
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## groupB -2.254     0.105    1.155 -1.951   0.0511 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## groupB     0.105      9.524   0.01091      1.01
##
## Concordance= 0.737  (se = 0.038 )
## Likelihood ratio test= 4.95  on 1 df,   p=0.03
## Wald test            = 3.81  on 1 df,   p=0.05
## Score (logrank) test = 5.2  on 1 df,   p=0.02
```

Cox Regression for Brown data (cont)

```
coxph(Surv(time, exitus=="Death") ~ group, data = dat)%>%
  gtsummary::tbl_regression(exp = TRUE)
```

Characteristic	HR	95% CI	p-value
group			
A			
B	0.10	0.01, 1.01	0.051

```
ggforest(coxph(Surv(time, exitus=="Death") ~ group, data = dat),data=dat)
```



Exercise 1 (cont)

- Fit a Cox model for ECG in diabetes data
- Add age to this model
- Plot hazard ratios

Cox Regression Diabetes data

```
Coxfit<- coxph(Surv(tempsviu, mort=="Muerto") ~ ecg, data =dat2)
summary(Coxfit)
```

```
## Call:
## coxph(formula = Surv(tempsviu, mort == "Muerto") ~ ecg, data = dat2)
##
##   n= 149, number of events= 25
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## ecgFrontera -2.59348   0.07476  0.67609 -3.836 0.000125 ***
## ecgNormal   -3.45046   0.03173  0.57787 -5.971 2.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## ecgFrontera   0.07476      13.38   0.01987   0.28129
## ecgNormal     0.03173      31.52   0.01022   0.09848
##
## Concordance= 0.711  (se = 0.053 )
## Likelihood ratio test= 28.29  on 2 df,   p=7e-07
## Wald test            = 35.92  on 2 df,   p=2e-08
## Score (logrank) test = 76.45  on 2 df,   p=<2e-16
```

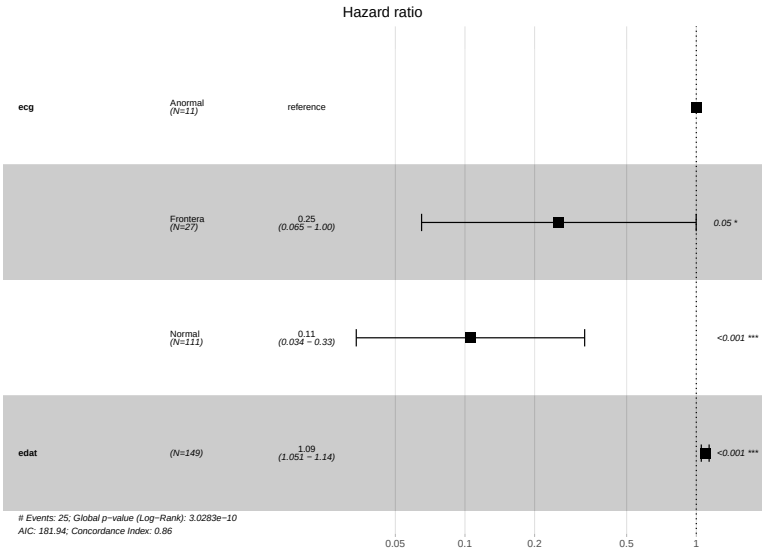
Cox Regression Diabetes data (cont)

```
coxph(Surv(tempsviu, mort=="Muerto") ~ ecg+edat, data =dat2)%>%
  gtsummary::tbl_regression(exp = TRUE)
```

Characteristic	HR	95% CI	p-value
ecg			
Anormal			
Frontera	0.25	0.06, 1.00	0.050
Normal	0.11	0.03, 0.33	<0.001
edat	1.09	1.05, 1.14	<0.001

Cox Regression Diabetes data (cont)

```
ggforest(coxph(Surv(tempsviu, mort=="Muerto") ~ ecg+edat, data =dat2) ,data=dat2)
```



Some useful Web pages

- https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html
- http://www.sthda.com/english/wiki/survival-analysis-basics
- http://www.sthda.com/english/wiki/survminer-r-package-survival-data-analysis-and-visualization