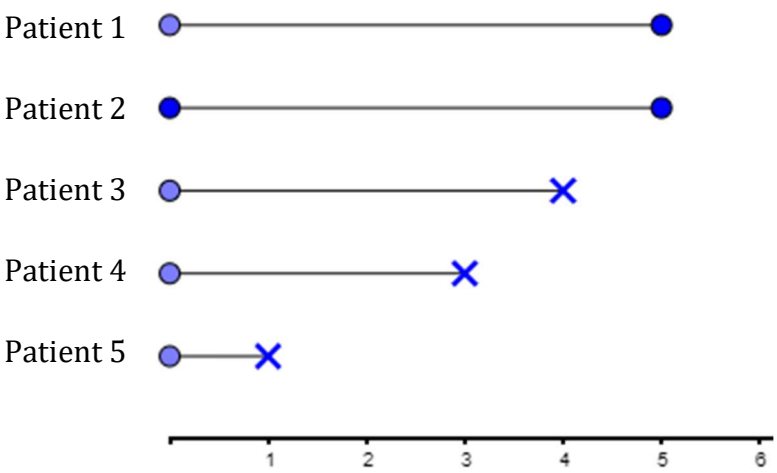
HW_1

2023-02-02

1.1 Re-write these survival times in terms of patient time, and create a simple data set listing the survival time and censoring indicator for each patient. How many patients died? How many person-years are there in this trial? What is the death rate per person-year?



in

The

Patient	Survtime (Years)	Status
1	5	0
2	5	0
3	4	1
4	3	1
5	1	1

Three patients died.

There are five patients with five person-years this trial.

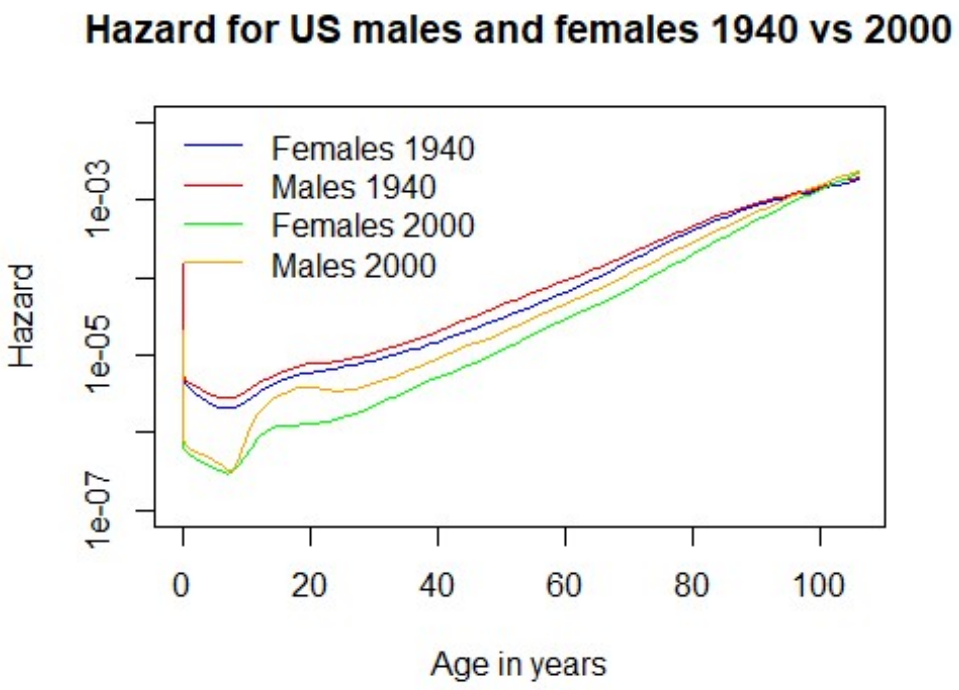
death rate is 3/5=0.6 per person-year.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
### 2.1. Using the "survexp.us" data described in Example 2.2, plot
### the hazard functions for men and women in 1940 and 2000. Comment
### on the change in mortality rates in children
library(survival)

hazMale_1940 <- survexp.us[,"male","1940"]
hazFemale_1940 <- survexp.us[,"female","1940"]
hazMale_2000 <- survexp.us[,"male","2000"]
hazFemale_2000 <- survexp.us[,"female","2000"]

DAYS_BY_YEAR <- 365.25
tm <- c(0,
             1/DAYS_BY_YEAR,
             7/DAYS_BY_YEAR,
             28/DAYS_BY_YEAR,
             1:(dim(hazMale_1940)-4))

plot(x = tm, log = "y", y = hazFemale_1940, type = "l", col="blue", lwd
=1,
     ylab ="Hazard", xlab = "Age in years",
     main ="Hazard for US males and females 1940 vs 2000", ylim=c(1e-07
,1e-02))
lines(x = tm,  y = hazMale_1940, type = "l", col="red",lwd=1)
lines(x = tm,  y = hazFemale_2000, type = "l", col="green",lwd=1)
lines(x = tm,  y = hazMale_2000, type = "l", col="orange",lwd=1)
legend("topleft",col=c("blue","red","green","orange"),
       legend =c("Females 1940","Males 1940","Females 2000","Males 2000
"), lwd=1, bty = "n")
```

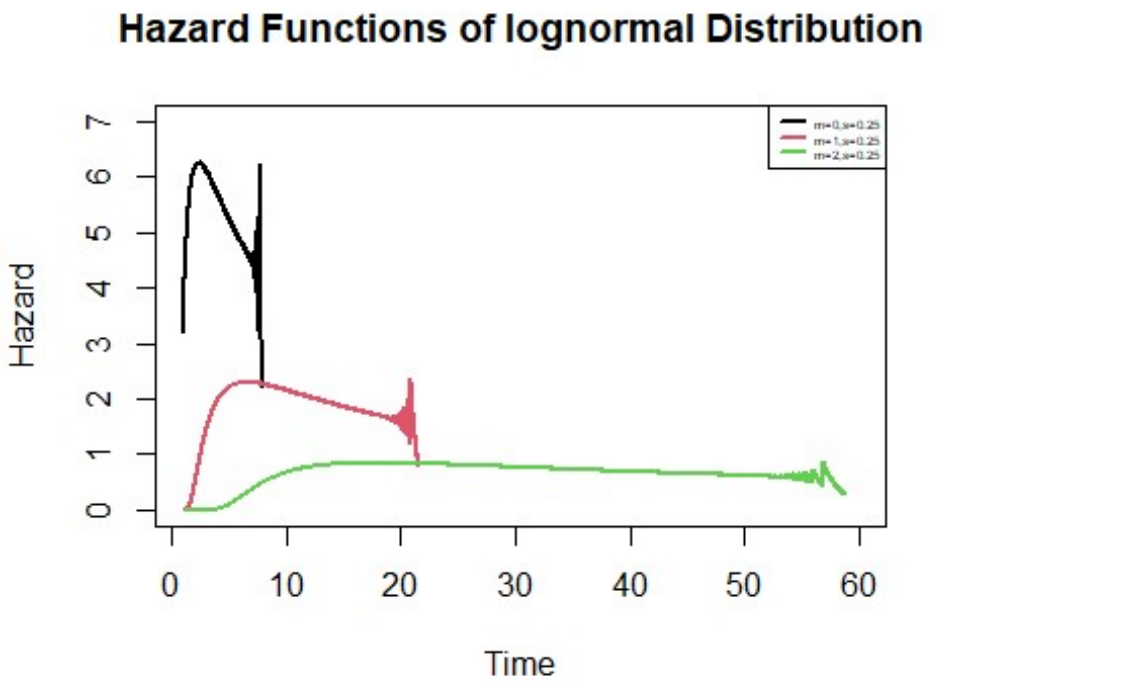


The mortality rates in children decreased from 1940 to 2000.

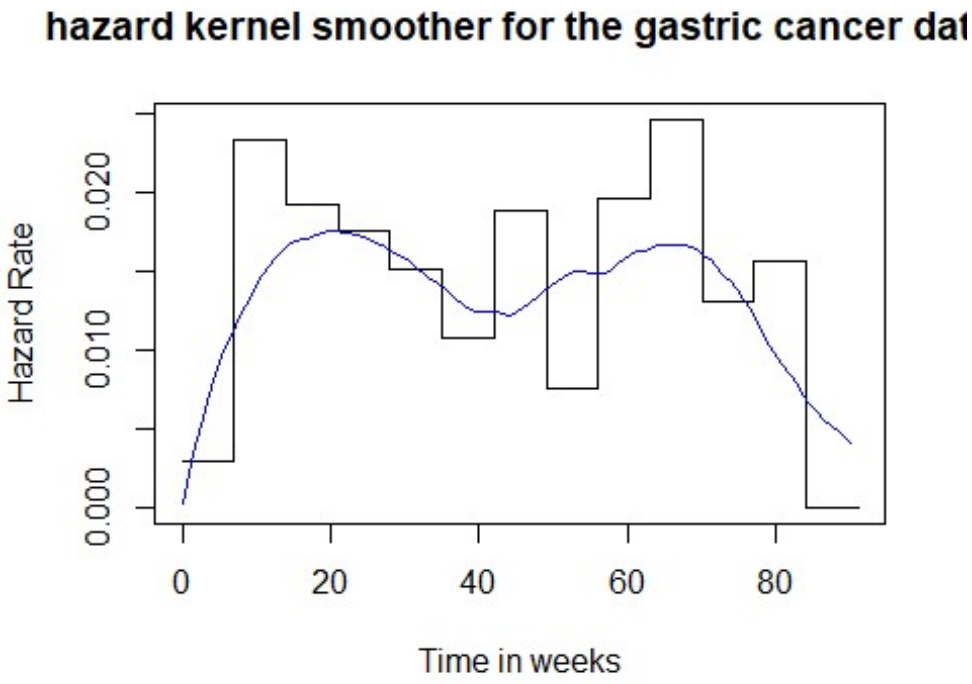
```
### 2.4. Consider the survival data in Exercise 1.1 Assuming that
### these observations are from an exponential distribution, find LAMDA
### ESTIMATE and an estimate of var(lambda ESTIMATE)

survival.data <- data.frame(patient = c(1,2,3,4,5),
                            survtime = c(5,5,4,3,1),
                            status = c(0,0,1,1,1))

d <- sum(survival.data$status)
V <- sum(survival.data$survtime)

lambda_estimado <- d/V
lambda_estimado

## [1] 0.1666667

var_lambda_estimado <- d/V^2
var_lambda_estimado

## [1] 0.009259259
```

```
###2.6 Another parametric survival distribution is the log-normal distr
ibution. Use the density and cumulative distribution R functions “dlnor
m” and “plnorm” to compute and plot the lognormal hazard functions with
 the parameter “meanlog” taking the values 0, 1, and 2, and with “sdlog
” fixed at 0.25. Describe the risk profile a disease would have if it f
ollowed one of these hazard functions.
```

```r
t=seq(1,60,by=0.01)
hlnorm=function(m,s)
{

h=(dlnorm(t,meanlog=m,sdlog=s))/(1-plnorm(t,meanlog=m,sdlog=s))

return(h)
}

plot(t,hlnorm(0,0.25),type="l",ylim=c(0,7),lwd=2,col=1,xlab="Time",ylab
="Hazard",main="Hazard Functions of lognormal Distribution")
lines(t,hlnorm(1,0.25),lwd=2,type="l",col=2)
lines(t,hlnorm(2,0.25),lwd=2,type="l",col=3)
legend("topright",legend=c("m=0,s=0.25","m=1,s=0.25","m=2,s=0.25"),lwd=
c(2,2,2),col=c(1,2,3),cex=0.4)
```



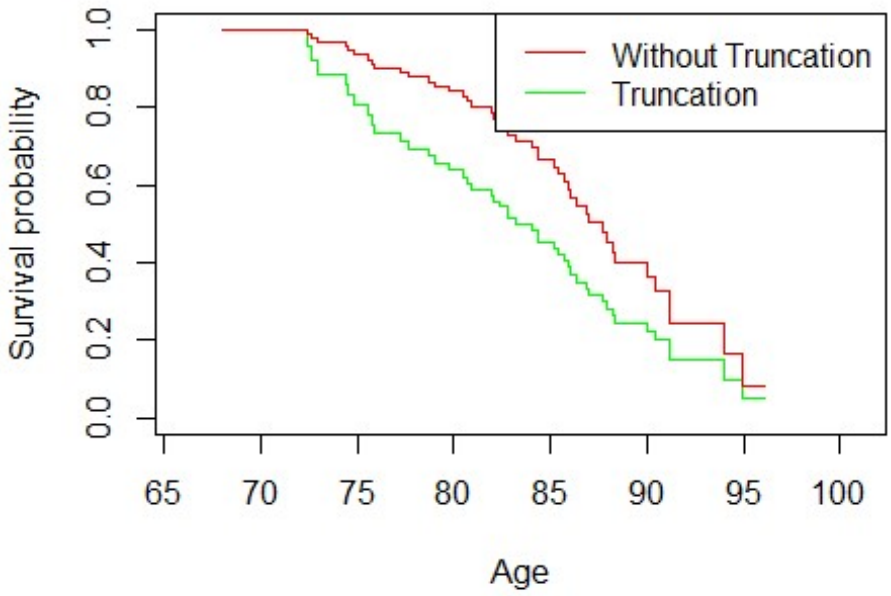
```
###
# 3.3. Find a smooth hazard function estimate for the gastric    #
# cancer data using kernel width "bw.grid = 20". Explain reason  #
# for the multiple peaks in the estimate.                        #
###

library(asaur)
library(survival)
library(muhaz)
head(gastricXelox)

##   timeWeeks delta
## 1         4     1
## 2         8     1
## 3         8     1
## 4         8     1
## 5         9     1
## 6        11     1

timeMonths <- gastricXelox$timeWeeks*7/30.25
time <- gastricXelox$timeWeeks

result.pe7 <- pehaz(time, gastricXelox$delta, width=7, max.time=90)

##
## max.time= 90
## width= 7
## nbins= 13

result.pe1 <- pehaz(time, gastricXelox$delta, width=1, max.time=90)

##
## max.time= 90
## width= 1
## nbins= 90

result.smooth <- muhaz(time, gastricXelox$delta, bw.smooth=20,
b.cor="left", max.time=90)
result.smooth.g <- muhaz(time, gastricXelox$delta, bw.grid = 20, bw.met
hod
= "global", b.cor="left", max.time=90)
plot(result.pe7, xlab="Time in weeks",main ="hazard kernel smoother for
 the gastric cancer data")
lines(result.smooth.g, col='blue')
```



The kernel smoothing technique involves weighting nearby observations more heavily in the estimation process, which can result in multiple peaks if there are regions of the data where the hazard rate changes rapidly. Additionally, the choice of kernel function and bandwidth used in the smoothing process can impact the shape of the hazard function estimate and the number of peaks that are present. Therefore, multiple peaks in a kernel-smoothed hazard function estimate may reflect underlying features in the data or be an artifact of the smoothing process.

```
#######
###3.4. Estimate the survival distribution for men, conditional on reac
hing the age of 68, ignoring the left truncation times. Discuss the bia
s of this estimate by comparing to the estimate presented in Sect. 3.4.
#######


head(ChanningHouse)

##    sex entry exit time cens
## 1 Male   782  909  127    1
## 2 Male  1020 1128  108    1
## 3 Male   856  969  113    1
## 4 Male   915  957   42    1
```

```
## 5 Male   863  983  120    1
## 6 Male   906 1012  106    1

ChanningHouse$entryYears <- ChanningHouse$entry/12
ChanningHouse$exitYears <- ChanningHouse$exit/12
ChanningMales <- ChanningHouse[ChanningHouse$sex == "Male",]
head(ChanningHouse)

##    sex entry exit time cens entryYears exitYears
## 1 Male   782  909  127    1   65.16667  75.75000
## 2 Male  1020 1128  108    1   85.00000  94.00000
## 3 Male   856  969  113    1   71.33333  80.75000
## 4 Male   915  957   42    1   76.25000  79.75000
## 5 Male   863  983  120    1   71.91667  81.91667
## 6 Male   906 1012  106    1   75.50000  84.33333

result.km.68 <- survfit(Surv(entryYears, exitYears, cens,
type="counting")~1,start.time=68,data=ChanningMales)
result.km2 <- survfit(Surv(exitYears, cens) ~ 1,start.time=68,
data=ChanningMales)
plot(result.km.68, xlim=c(66, 101), xlab="Age",
ylab="Survival probability", conf.int=F)
 lines(result.km.68, col="green", conf.int=F)
 lines(result.km2, col="red", conf.int=F)
 legend("topright", legend=c("Without Truncation", "Truncation"),
 lty=1, col=c("red", "green"))
```



Ignoring the left truncation time means only the time of exit or death is used to estimate the survival function. The left truncation time model takes into account the timing of entry and exit from the study and may provide more accurate estimates. Therefore, left t

"result.km.68" is a counting process survival model, where the time at which the study participant enters and exits the study is taken into account when estimating the survival function.

"result.km0" is a standard survival model, where only the time of exit or death is used to estimate the survival function. Censored observations are still taken into account.

In summary, "result.km.68" is a more detailed and nuanced model of survival, while "result.km0" is a simpler model.