

# MAE0514 - Introdução a Análise de Sobrevivência - Lista 2

Bruno de Castro Paul Schultze<sup>1</sup>  
Rubens Santos Andrade Filho<sup>2</sup>

Junho de 2021

## Sumário

Questão 1 . . . . .	2
Questão 2 . . . . .	2
Questão 3 . . . . .	2
Questão 4 . . . . .	6
Questão 5 . . . . .	6
Questão 6 . . . . .	10
Questão 7 . . . . .	10
Questão 8 . . . . .	14
Código Completo . . . . .	14

---

<sup>1</sup>Número USP: 10736862

<sup>2</sup>Número USP: 10370336

## Questão 1

a)

A variável do estudo é o tempo compreendido da exposição a um material cancerígeno até o desenvolvimento do tumor de um tamanho determinado nos ratos. Nesse caso, a origem é a exposição a um material cancerígeno e o evento de interesse é o desenvolvimento do tumor de um tamanho determinado.

b)

Para os rato A, B e C foram observados os tempos de falha, isto é, os tempos até os ratos desenvolverem o tumor de determinado tamanho.

Para o rato D foi observado uma censura aleatória à direita na vigésima semana, sua morte. Até a semana 20 o rato não tinha desenvolvido o tumor de um tamanho determinado.

Para os ratos E e F foram observados censuras à direita do tipo I na semana 30 por ser a duração do estudo. Ademais, apenas com as informações do enunciado não é possível dizer se todos os ratos foram expostos ao material cancerígeno ao mesmo tempo para sabermos se a censura é generalizada ou não.

## Questão 2

## Questão 3

Em um estudo clínico realizado com pacientes com câncer gástrico avançado (com metástase linfodonal), uma quimioterapia com Xeloda (capecitabina) e oxaliplatina foi administrada antes da cirurgia de 48 pacientes. Nesse tipo de ensaio clínico, é de interesse estudar e avaliar o tempo livre da doença, que é o tempo que o paciente fica bem, vivo e sem a doença. Assim, um dos objetivos é estudar o tempo decorrido entre o início do tratamento e óbito ou progressão da doença (o que ocorrer primeiro). Os dados do tempo livre da doença (em semanas) dos 48 pacientes estão disponíveis no arquivo `Lista2-Xelox.csv`, sendo que a variável delta é codificada como sendo 1 se o evento ocorreu e 0 se a observação é censurada.

(a) Calculamos o estimador da tábua de vida, considerando as seguintes faixas de tempo:

Faixa 1:	8 semanas (inclusive) ou menos
Faixa 2:	de 8 a 16 semanas (inclusive)
Faixa 3:	de 16 a 24 semanas (inclusive)
Faixa 4:	de 24 a 32 semanas (inclusive)
Faixa 5:	de 32 a 44 semanas (inclusive)
Faixa 6:	de 44 a 56 semanas (inclusive)
Faixa 7:	mais de 56 semanas

Dessa forma, consideramos os intervalos fechados à direita.

```
# QUESTAO 3a ----
```

```
dados_raw <- readr::read_csv2('data/Lista2-Xelox.csv')
```

```

# limites dos intervalos
breaks <- c(0,8,16,24,32,44,56, Inf)

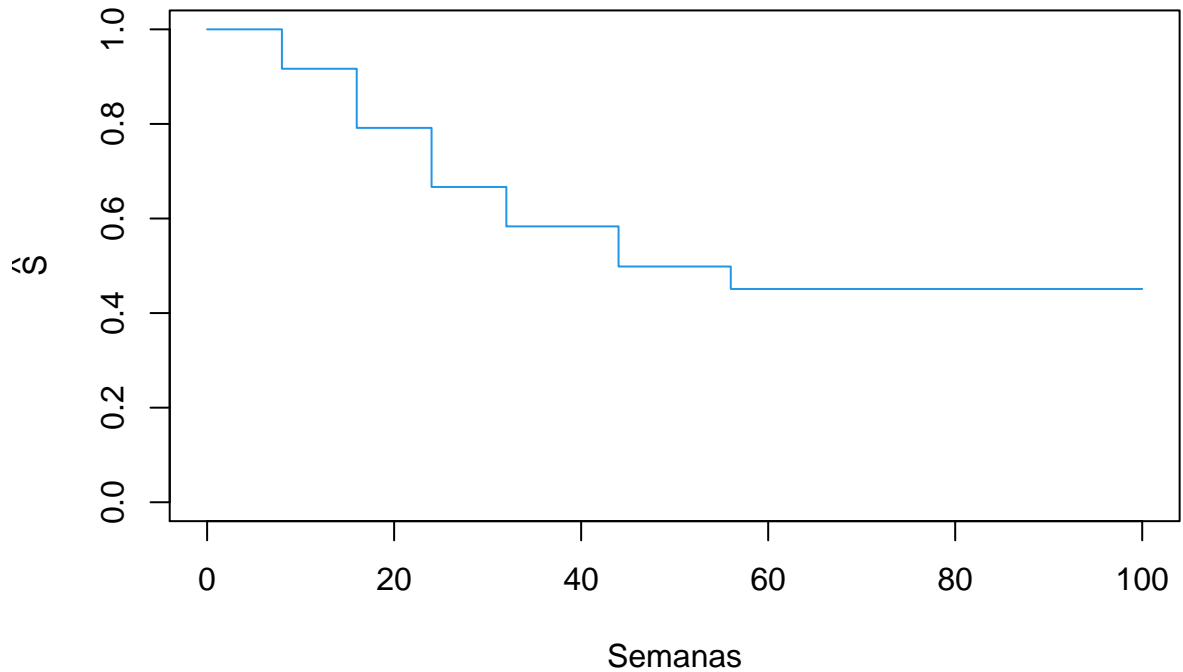
tabua <- dados_raw %>%
  mutate(
    # define as faixas
    intervalo = cut(timeWeeks, breaks=breaks, right=TRUE, include.lowest = T),
    i = as.integer(intervalo)
  ) %>%
  group_by(intervalo, i) %>%
  summarise(
    # numero de falhas no intervalo
    d = sum(delta),
    # numero de censuras no intervalo
    w = sum(1-delta)
  ) %>%
  ungroup() %>%
  mutate(
    # numero de obs em risco, que nao falharam até o fim do intervalo anterior
    n_estrela = sum(d+w) - cumsum(d+w) +w+d,
    # corrigindo o numero de ind. em risco
    n = n_estrela - w/2,
    # prop. de falhas no intervalo
    q_hat = d/n,
    # na tabua de vida, a estimativa de S do 1o intervalo = 1
    # depois o produtorio acumulado dos p_i
    s_hat = c(1, cumprod(1 - q_hat)[-n()])
  )

```

Tabela 1: Estimativas da tábua de vida.

Semanas	$i$	$d_i$	$w_i$	$n^*$	$n$	$q_i$	$\hat{S}(t)$
[0,8]	1	4	0	48	48.0	0.0833333	1.0000000
(8,16]	2	6	0	44	44.0	0.1363636	0.9166667
(16,24]	3	6	0	38	38.0	0.1578947	0.7916667
(24,32]	4	4	0	32	32.0	0.1250000	0.6666667
(32,44]	5	4	1	28	27.5	0.1454545	0.5833333
(44,56]	6	2	4	23	21.0	0.0952381	0.4984848
(56,Inf]	7	6	11	17	11.5	0.5217391	0.4510101

## Estimativa da função de sobrevivência pela tábua de vida



Chama a atenção o fato da estimativa da função de sobrevivência não se aproximar de 0 à medida que aumentam o número de semanas. Isso acontece principalmente devido às 11 observações censuradas que ficaram no último intervalo, isto é, temos menos muito menos informação a respeito das falhas nesse intervalo.

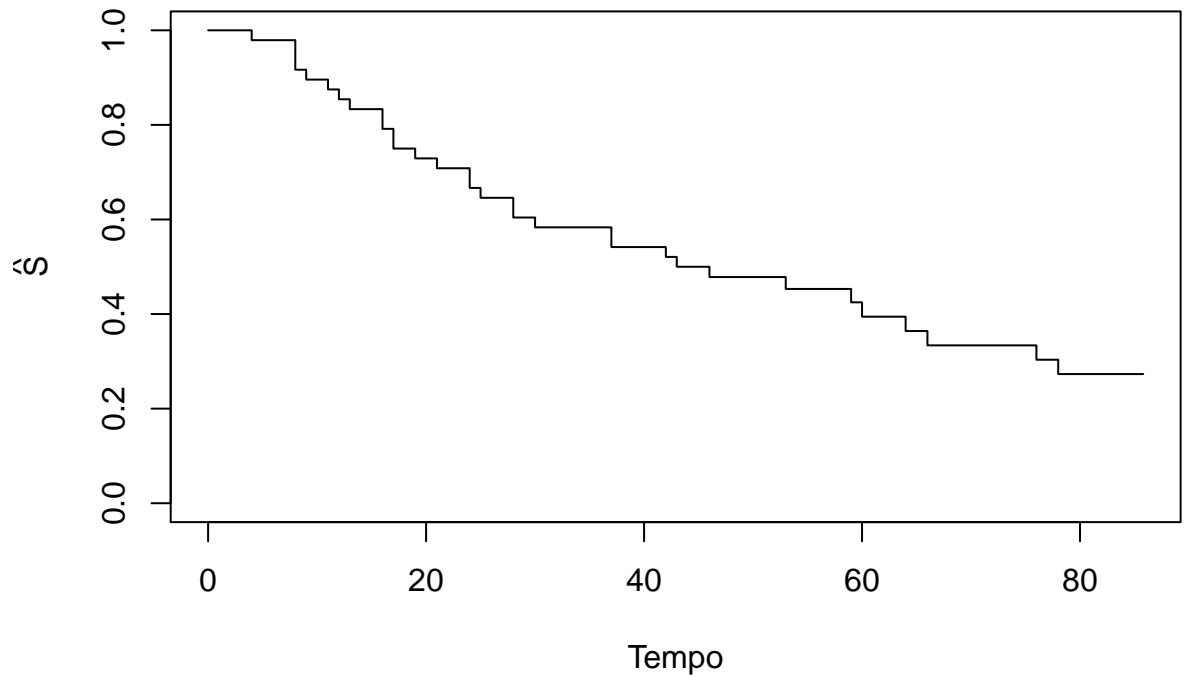
(b) Calcule o estimador Kaplan-Meier para os dados (você pode utilizar um software).

Tabela 2: Estimativas de Kaplan-Meier.

Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
4	1	0	48	0.0208333	0.9791667
8	3	0	47	0.0638298	0.9166667
9	1	0	44	0.0227273	0.8958333
11	1	0	43	0.0232558	0.8750000
12	1	0	42	0.0238095	0.8541667
13	1	0	41	0.0243902	0.8333333
16	2	0	40	0.0500000	0.7916667
17	2	0	38	0.0526316	0.7500000
19	1	0	36	0.0277778	0.7291667
21	1	0	35	0.0285714	0.7083333
24	2	0	34	0.0588235	0.6666667
25	1	0	32	0.0312500	0.6458333
28	2	0	31	0.0645161	0.6041667
30	1	0	29	0.0344828	0.5833333

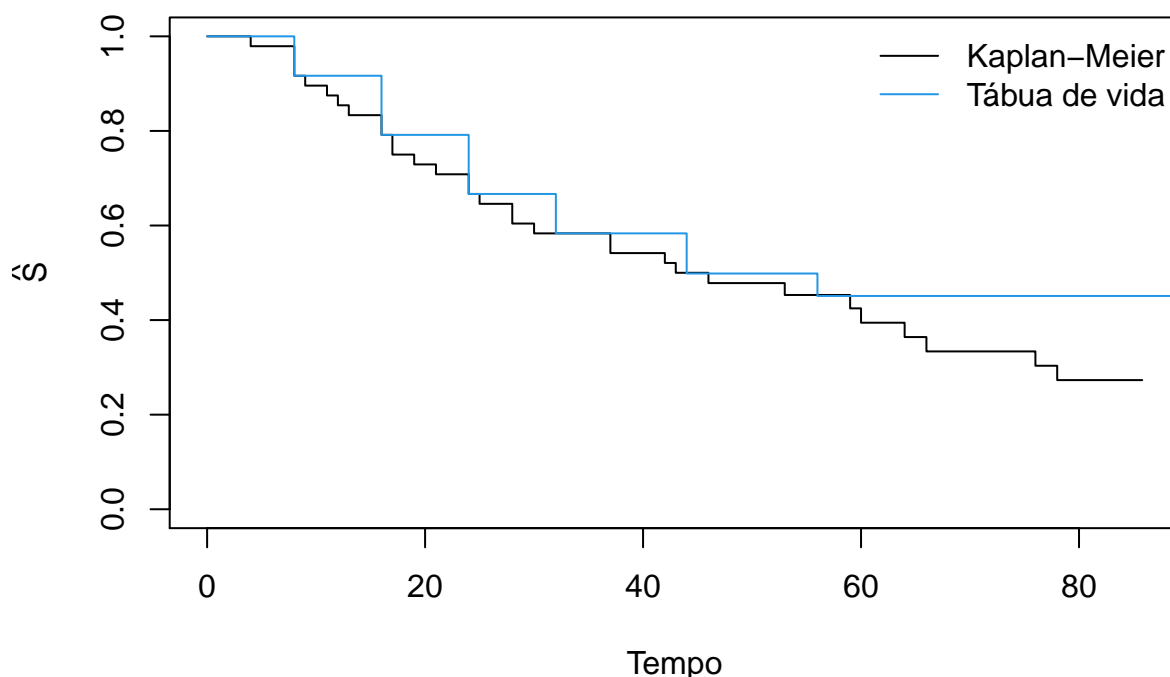
Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
37	2	0	28	0.0714286	0.5416667
42	1	0	26	0.0384615	0.5208333
43	1	1	25	0.0400000	0.5000000
46	1	0	23	0.0434783	0.4782609
53	1	0	19	0.0526316	0.4530892
59	1	1	16	0.0625000	0.4247712
60	1	0	14	0.0714286	0.3944304
64	1	0	13	0.0769231	0.3640896
66	1	0	12	0.0833333	0.3337488
76	1	0	11	0.0909091	0.3034080
78	1	0	10	0.1000000	0.2730672

### Estimativa da função de sobrevivência por Kaplan–Meier



- (c) Coloque em um mesmo gráfico as duas curvas estimadas nos itens anteriores. Compare as curvas e comente.

## Estimativas da função de sobrevivência



O método de Kaplan-Meier é o método da tábua de vida quando o número de intervalos é o número de instantes únicos nos dados e os tamanhos dos intervalos são os tempos. Nota-se que o método de Kaplan-Meier melhora visualmente a estimativa da curva da função de sobrevivência, principalmente após a semana 56, onde o número de censuras é maior e ficaram todas no último intervalo da tábua de vida todas

## Questão 4

## Questão 5

Os dados mostrados a seguir representam o tempo até a ruptura de um tipo de isolante elétrico sujeito a uma tensão de estresse de 35 Kvolts. O teste consistiu em deixar 25 destes isolantes funcionando até que 15 deles falhassem (censura tipo II), obtendo-se os seguintes resultados (em minutos):

0,19	0,78	0,96	1,31	2,78	3,16	4,67	4,85
6,50	7,35	8,27	12,07	32,52	33,91	36,71	

Observe que 10 observações foram censuradas. Para este exercício, os cálculos podem ser feitos à mão ou com auxílio computacional, porém a ideia é não utilizar uma função pronta que calcule o que for pedido. Você deve usar uma planilha ou escrever o código que faça as contas no R ou outro software de sua preferência. A partir desses dados amostrais, deseja-se obter:

- (a) a função de sobrevivência estimada por Kaplan-Meier;

```

# QUESTAO 5a ----

# dados
dados <- tibble(
  t = c(0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.67, 4.85,
        6.50, 7.35, 8.27, 12.07, 32.52, 33.91, 36.71),
  delta = 1
) %>%
# censuras
add_row(t= rep(36.71, 10), delta = 0)

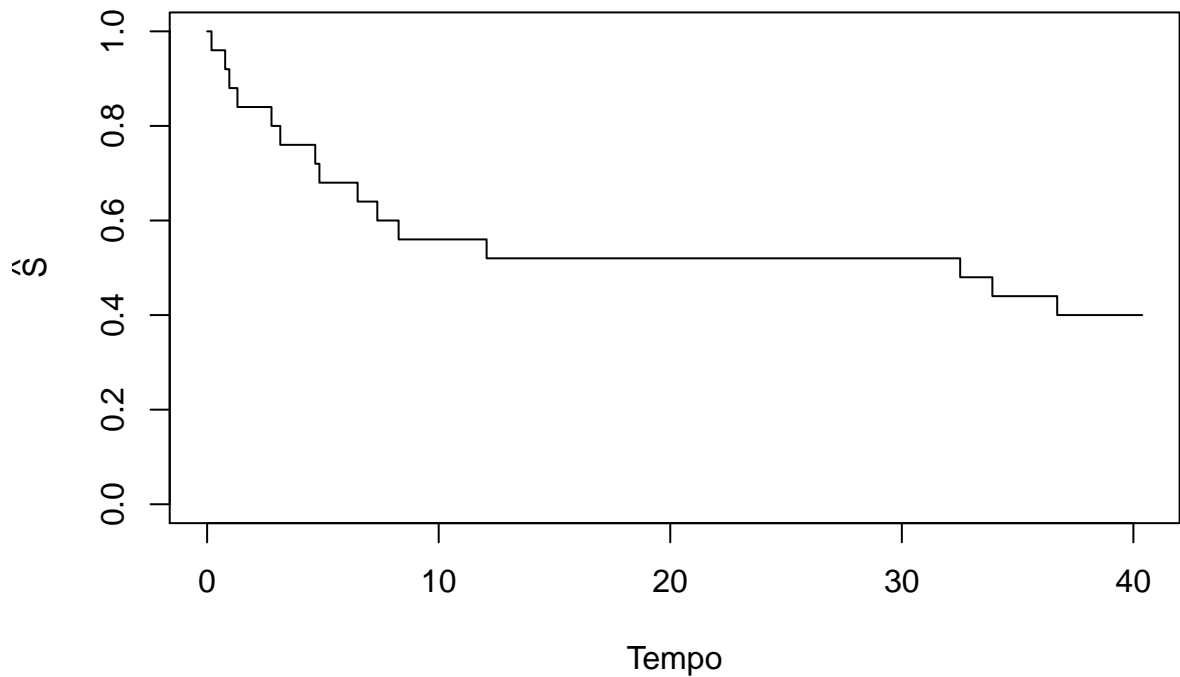
estimativas <- estimar_km(dados, t, delta)

```

Tabela 3: Estimativas de Kaplan-Meier.

Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
0.19	1	0	25	0.0400000	0.96
0.78	1	0	24	0.0416667	0.92
0.96	1	0	23	0.0434783	0.88
1.31	1	0	22	0.0454545	0.84
2.78	1	0	21	0.0476190	0.80
3.16	1	0	20	0.0500000	0.76
4.67	1	0	19	0.0526316	0.72
4.85	1	0	18	0.0555556	0.68
6.50	1	0	17	0.0588235	0.64
7.35	1	0	16	0.0625000	0.60
8.27	1	0	15	0.0666667	0.56
12.07	1	0	14	0.0714286	0.52
32.52	1	0	13	0.0769231	0.48
33.91	1	0	12	0.0833333	0.44
36.71	1	10	11	0.0909091	0.40

## Estimativa da função de sobrevivência por Kaplan–Meier



(b) uma estimativa para o tempo mediano de vida deste tipo de isolante elétrico funcionando a essa tensão;

Para a vida mediana, nós vemos que  $\hat{S}(12.07) = 0.52$  e  $\hat{S}(32.52) = 0.48$ , então o tempo mediano se encontra entre esses dois tempos. Por interpolação linear,

```
# QUESTAO 5b ----
```

```
t0 = 12.07
t1 = 32.52
s0 = estimativas$s_hat[which(estimativas$t==t0)]
s1 = estimativas$s_hat[which(estimativas$t==t1)]

s=0.5
t_mediano = t0 + (t1-t0)*(s-s0)/(s1-s0)
t_mediano
```

```
## [1] 22.295
```

encontramos um tempo de vida mediano igual a 22.295.

(c) uma estimativa (pontual e intervalar) para a fração de defeituosos esperada nos dois primeiros minutos de funcionamento;



Observemos que  $t_0 = 1.31 < 2 < 2.78 = t_1$ . Com isso, usando interpolação linear,

```
# QUESTAO 5c ----

t0 = 1.31
t1 = 2.78
s0 = estimativas$s_hat[which(estimativas$t==t0)]
s1 = estimativas$s_hat[which(estimativas$t==t1)]

# do item anterior:
# t(s) = t0 + (t1-t0)*(s-s0)/(s1-s0)
# então s(t):
# s = (t-t0)*(s1-s0)/(t1-t0) + s0
t=2
s_hat = (t-t0)*(s1-s0)/(t1-t0) + s0
c(s_hat, 1-s_hat)
```

```
## [1] 0.8212245 0.1787755
```

Obtemos que  $\hat{S}(2) \approx 0.82$ , logo  $1 - \hat{S}(2) = 0.18$  é uma estimativa pontual para a fração de defeituosos esperada nos dois primeiros minutos de funcionamento.

Uma estimativa intervalar pode ser obtida com

$$IC(S(t); \gamma) = \left( \hat{S}(t) - z_{\gamma/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))}; \hat{S}(t) + z_{\gamma/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))} \right)$$

$$\Rightarrow IC(1 - S(t); \gamma) = \left( 1 - \hat{S}(t) - z_{\gamma/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))}; 1 - \hat{S}(t) + z_{\gamma/2} \sqrt{\widehat{\text{Var}}(\hat{S}(t))} \right)$$

onde

$$\widehat{\text{Var}}(\hat{S}(t)) = \left[ \hat{S}(t) \right]^2 \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

```
var_hat <- estimativas %>% filter(t<2) %>%
  summarise(
    a = (!s_hat)^2 * sum(d/(Y*(Y-d)))
  ) %>% pull()
var_hat
```

```
## [1] 0.005138359
```

Obtemos que  $\widehat{\text{Var}}(\hat{S}(2)) = \widehat{\text{Var}}(1 - \hat{S}(2)) \approx 0.0051$ . Portanto, uma estimativa intervalar, com 90% de confiança, para a fração de defeituosos esperada nos dois primeiros minutos de funcionamento é

```
z <- qnorm(0.95) # 1 - (1-gamma)/2
round(c((1-s_hat) - z * sqrt(var_hat), (1-s_hat) + z * sqrt(var_hat)), 3)
```

```
## [1] 0.061 0.297
```

$$IC(1 - S(2)); 90\%) = [0.061; 0.297]$$

(d) o tempo necessário para 20% dos isolantes estarem fora de operação.

Observando a tabela, vemos que uma estimativa do tempo necessário para 20% dos isolantes estarem fora de operação é 2.78 quando  $\hat{S}(t) = 0.80$ . Caso não estivesse na tabela, poderíamos obter por interpolação linear como nos outros itens.

## Questão 6

## Questão 7

```
# QUESTAO 7 ----

dados_raw <- readr::read_csv('data/pharmacoSmoking.csv')

dados_grp <- dados_raw %>% group_by(grp)
lst_dados <- dados_grp %>% group_split()
names(lst_dados) <- group_keys(dados_grp) %>% pull()

# calcula estimativas para cada grupo
lst_estimativas <- lst_dados %>% purrr::map(~estimar_km(., ttr, relapse) )
```

a) Obtenha as curvas de Kaplan-Meier dos dois tratamentos;

```
tabela_km(lst_estimativas[["combination"]],
          caption="Estimativas de Kaplan-Meier para combination.")
```

Tabela 4: Estimativas de Kaplan-Meier para combination.

Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
0	4	0	61	0.0655738	0.9344262
2	3	0	57	0.0526316	0.8852459
4	1	0	54	0.0185185	0.8688525
5	2	0	53	0.0377358	0.8360656
8	2	0	51	0.0392157	0.8032787
10	1	0	49	0.0204082	0.7868852
12	1	0	48	0.0208333	0.7704918
14	1	0	47	0.0212766	0.7540984
15	2	0	46	0.0434783	0.7213115
16	1	0	44	0.0227273	0.7049180
20	1	0	43	0.0232558	0.6885246
21	1	0	42	0.0238095	0.6721311
30	2	0	41	0.0487805	0.6393443

Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
42	1	0	39	0.0256410	0.6229508
50	1	0	38	0.0263158	0.6065574
56	2	0	37	0.0540541	0.5737705
60	2	0	35	0.0571429	0.5409836
63	2	0	33	0.0606061	0.5081967
65	1	0	31	0.0322581	0.4918033
75	1	0	30	0.0333333	0.4754098
110	1	0	29	0.0344828	0.4590164
140	3	0	28	0.1071429	0.4098361
170	1	0	25	0.0400000	0.3934426

```
tabela_km(lst_estimativas[["patchOnly"]],
caption="Estimativas de Kaplan-Meier para patchOnly.")
```

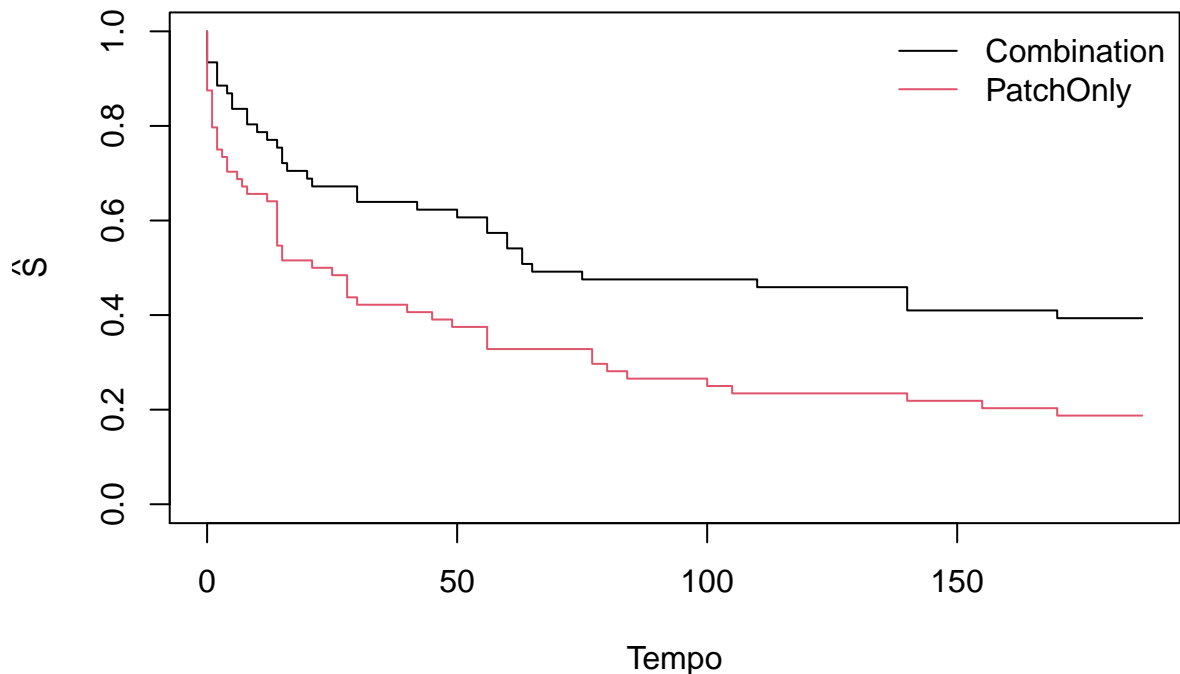
Tabela 5: Estimativas de Kaplan-Meier para patchOnly.

Tempo	$d_i$	$w_i$	$Y_i$	$q_i = d_i/Y_i$	$\hat{S}(t)$
0	8	0	64	0.1250000	0.875000
1	5	0	56	0.0892857	0.796875
2	3	0	51	0.0588235	0.750000
3	1	0	48	0.0208333	0.734375
4	2	0	47	0.0425532	0.703125
6	1	0	45	0.0222222	0.687500
7	1	0	44	0.0227273	0.671875
8	1	0	43	0.0232558	0.656250
12	1	0	42	0.0238095	0.640625
14	6	0	41	0.1463415	0.546875
15	2	0	35	0.0571429	0.515625
21	1	0	33	0.0303030	0.500000
25	1	0	32	0.0312500	0.484375
28	3	0	31	0.0967742	0.437500
30	1	0	28	0.0357143	0.421875
40	1	0	27	0.0370370	0.406250
45	1	0	26	0.0384615	0.390625
49	1	0	25	0.0400000	0.375000
56	3	0	24	0.1250000	0.328125
77	2	0	21	0.0952381	0.296875
80	1	0	19	0.0526316	0.281250
84	1	0	18	0.0555556	0.265625
100	1	0	17	0.0588235	0.250000
105	1	0	16	0.0625000	0.234375
140	1	0	15	0.0666667	0.218750
155	1	0	14	0.0714286	0.203125
170	1	0	13	0.0769231	0.187500

```
plot_km(lst_estimativas[["combination"]])
plot_km(lst_estimativas[["patchOnly"]], col=2, add=T)

legend("topright", legend=c("Combination", "PatchOnly"), lty = c(1, 1),
col = c(1,2), bty="n")
```

## Estimativa da função de sobrevivência por Kaplan–Meier



As estimativas de Kaplan-Meier são bem diferentes para os dois grupos como podemos notar pelas tabelas e gráfico.

- (b) Compare os tratamentos utilizando o teste de log-rank e também utilizando diferentes ponderações (escolha pelo menos três diferentes);

Para esse item, vamos usar os pacotes `survival` para o modelo de sobrevivência e `survminer` para realizar os testes. Realizamos o teste de log-rank e também usando diferentes ponderações: Gehan-Breslow (Wilcoxon generalizado), Peto-Peto e Fleming-Harrington( $p=1$ ,  $q=1$ ).

```
# QUESTAO 7b ----

# install.packages(c("survival", "survminer"))
library(survival)
library(survminer)

# modelo
sfit <- survfit(Surv(ttr, relapse)~grp, data=dados_raw)

bind_rows(
  # teste de log rank
  surv_pvalue(sfit, method = "log-rank"),

  # usando Gehan-Breslow (Wilcoxon generalizado)
  surv_pvalue(sfit, method = "gehan-breslow"),
```

```
# usando Peto-Peto
surv_pvalue(sfit, method = "S1"),

# usando Fleming-Harrington(p=1, q=1)
surv_pvalue(sfit, method = "FH_p=1_q=1"),
) %>%
dplyr::relocate(method) %>%
dplyr::select(-pval.txt)
```

```
##                method variable      pval
## 1                Log-rank      grp 0.004606898
## 2              Gehan-Breslow      grp 0.005100000
## 3                Peto-Peto      grp 0.005200000
## 4 Fleming-Harrington (p=1, q=1)      grp 0.020600000
```

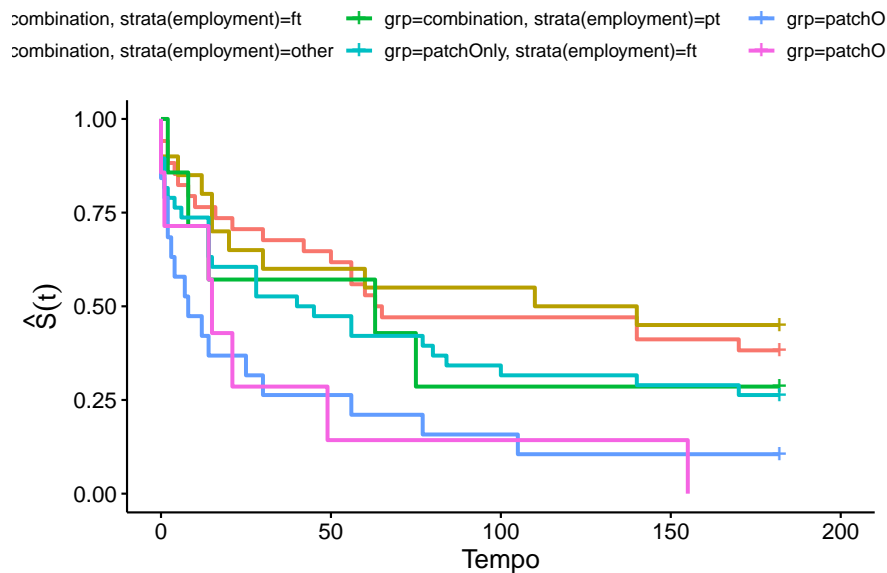
Em todos os testes o valor p deu significativo a um nível de 5%. Com isso, rejeitamos a hipótese nula de que não existe diferença na curva de sobrevivência dos dois grupos.

- (c) Compare as curvas de Kaplan-Meier utilizando o teste de log-rank, estratificado por situação de trabalho (variável employment) e discuta os resultados.

```
# QUESTAO 7c ----

# modelo
sfit <- survfit(Surv(ttr, relapse)~grp+strata(employment), data=dados_raw)

ggsurvplot(sfit,dados_raw)+
  labs(x="Tempo", y=expression(hat(S)(t)))
```



Fizemos o teste de log-rank do modelo estratificado por employment:

```
survdif(Surv(ttr, relapse)~grp+strata(employment), data=dados_raw)
```

```
## Call:
## survdiff(formula = Surv(ttr, relapse) ~ grp + strata(employment),
##          data = dados_raw)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## grp=combination 61      37    50.3      3.50      8.58
## grp=patchOnly   64      52    38.7      4.54      8.58
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

Obtivemos um valor p de 0.003, com isso, a um nível de 5%, rejeitamos a hipótese nula de diferença nas curvas de sobrevivência entre os grupos, controlando pela situação de trabalho.

## Questão 8

### Código Completo

```
knitr::opts_chunk$set(warning=FALSE,
                       # fig.dim = c(5,5),
                       # out.height = '40%',
                       # fig.align = 'center',
                       message=FALSE
                       )

library(tidyverse)
library(ggplot2)
library(knitr)
library(readr)
library(dplyr)

# QUESTAO 3a ----

dados_raw <- readr::read_csv2('data/Lista2-Xelox.csv')

# limites dos intervalos
breaks <- c(0,8,16,24,32,44,56, Inf)

tabua <- dados_raw %>%
  mutate(
    # define as faixas
    intervalo = cut(timeWeeks, breaks=breaks, right=TRUE, include.lowest = T),
    i = as.integer(intervalo)
  ) %>%
  group_by(intervalo, i) %>%
  summarise(
```

```

# numero de falhas no intervalo
d = sum(delta),
# numero de censuras no intervalo
w = sum(1-delta)
) %>%
ungroup() %>%
mutate(
  # numero de obs em risco, que nao falharam até o fim do intervalo anterior
  n_estrela = sum(d+w) - cumsum(d+w) +w+d,
  # corrigindo o numero de ind. em risco
  n = n_estrela - w/2,
  # prop. de falhas no intervalo
  q_hat = d/n,
  # na tabua de vida, a estimativa de S do 1o intervalo = 1
  # depois o produtorio acumulado dos p_i
  s_hat = c(1, cumprod(1 - q_hat)[-n()])
)

tabua %>%
kable(
  caption = "Estimativas da tábua de vida.",
  col.names = c(
    "Semanas", "$i$", "$d_i$", "$w_i$",
    "$n*$", "$n$",
    "$q_i$",
    "$\\hat{S}(t)$")
)

x = rep(breaks, each=2)[2:15]
x[length(x)] <- 100 # substitui infinito
y = rep(tabua$s_hat, each=2)

plot(x, y, type="l", col=4, xlab="Semanas", ylab=expression(hat(S)), ylim = c(0,1),
     main = "Estimativa da função de sobrevivência pela tábua de vida", cex=.6)

# QUESTAO 3b ----

# vamos criar uma funcao para calcular as estimativas de Kaplan-Meier
estimar_km <- function(dados, tempo, delta) {
  dados %>%
    # agrupa para cada tempo unico
    group_by(t={{ tempo }}) %>%
    # numero de eventos e censuras em cada t
    summarise(d = sum({{ delta }}), w=sum(1-{{ delta }})) %>%
    ungroup() %>%
    mutate(
      # numero de individuos vivos até antes de cada instante t
      Y = sum(d+w) - (cumsum(d+w) - (d+w)),
      # estimate of the conditional probability that an individual who survives
      # to just prior to time ti experiences the event at time ti
      q = d/Y,

```

```

    # estimate of surv function
    s_hat = cumprod(1 - q)
  ) %>%
  filter(d!=0)
}

tabela_km <- function(estimativa_km, caption="Estimativas de Kaplan-Meier.",
                      col.names = c("Tempo", "$d_i$", "$w_i$", "$Y_i$",
                                     "$q_i=d_i/Y_i$", "$\\hat{S}(t)$"),
                      ...) {
  estimativa_km %>% kable(caption=caption, col.names=col.names, ...)
}

# calcula estimativas
estimativas <- dados_raw %>% estimar_km(timeWeeks, delta)

estimativas %>% tabela_km()

plot_km <- function(
  estimativa_km, x_max=NULL, type="l", col=1, xlab="Tempo",
  ylab=expression(hat(S)),ylim = c(0,1),
  main = "Estimativa da função de sobrevivência por Kaplan-Meier", cex=.6, add=FALSE) {

  if(is.null(x_max)){
    x_max <- max(estimativa_km$t) * 1.10
  }

  x_km <- c(0, rep(estimativa_km$t, each=2), x_max)
  y_km <- c(1, 1, rep(estimativa_km$s_hat, each=2))

  if(add){
    lines(x_km, y_km, type=type, col=col, xlab=xlab, ylab=ylab,
          ylim=ylim, main=main, cex=cex)
  }else{
    plot(x_km, y_km, type=type, col=col, xlab=xlab, ylab=ylab,
         ylim=ylim, main=main, cex=cex)
  }
}

plot_km(estimativas)

# QUESTAO 3c ----
plot_km(estimativas,main = "Estimativas da função de sobrevivência")
lines(x, y, type="l", col=4)
legend("topright",legend=c("Kaplan-Meier","Tábua de vida"),lty = c(1, 1),
      col = c(1,4), bty="n")

# QUESTAO 5a ----

# dados

```



```

dados <- tibble(
  t = c(0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.67, 4.85,
        6.50, 7.35, 8.27, 12.07, 32.52, 33.91, 36.71),
  delta = 1
) %>%
  # censuras
  add_row(t= rep(36.71, 10), delta = 0)

estimativas <- estimar_km(dados, t, delta)

estimativas %>% tabela_km()

estimativas %>% plot_km()

# QUESTAO 5b ----

t0 = 12.07
t1 = 32.52
s0 = estimativas$s_hat[which(estimativas$t==t0)]
s1 = estimativas$s_hat[which(estimativas$t==t1)]

s=0.5
t_mediano = t0 + (t1-t0)*(s-s0)/(s1-s0)
t_mediano

# QUESTAO 5c ----

t0 = 1.31
t1 = 2.78
s0 = estimativas$s_hat[which(estimativas$t==t0)]
s1 = estimativas$s_hat[which(estimativas$t==t1)]

# do item anterior:
#  $t(s) = t0 + (t1-t0)*(s-s0)/(s1-s0)$ 
# então  $s(t)$ :
#  $s = (t-t0)*(s1-s0)/(t1-t0) + s0$ 
t=2
s_hat = (t-t0)*(s1-s0)/(t1-t0) + s0
c(s_hat, 1-s_hat)

var_hat <- estimativas %>% filter(t<2) %>%
  summarise(
    a = (!s_hat)^2 * sum(d/(Y*(Y-d)))
  ) %>% pull()
var_hat

z <- qnorm(0.95) # 1 - (1-gamma)/2

```

```

round(c((1-s_hat) - z * sqrt(var_hat), (1-s_hat) + z * sqrt(var_hat)), 3)
# QUESTAO 7 ----

dados_raw <- readr::read_csv('data/pharmacoSmoking.csv')

dados_grp <- dados_raw %>% group_by(grp)
lst_dados <- dados_grp %>% group_split()
names(lst_dados) <- group_keys(dados_grp) %>% pull()

# calcula estimativas para cada grupo
lst_estimativas <- lst_dados %>% purrr::map(~estimar_km(., ttr, relapse) )

tabela_km(lst_estimativas[["combination"]],
          caption="Estimativas de Kaplan-Meier para combination.")
tabela_km(lst_estimativas[["patchOnly"]],
          caption="Estimativas de Kaplan-Meier para patchOnly.")

plot_km(lst_estimativas[["combination"]])
plot_km(lst_estimativas[["patchOnly"]], col=2, add=T)

legend("topright", legend=c("Combination", "PatchOnly"), lty = c(1, 1),
      col = c(1,2), bty="n")

# QUESTAO 7b ----

# install.packages(c("survival", "survminer"))
library(survival)
library(survminer)

# modelo
sfit <- survfit(Surv(ttr, relapse)~grp, data=dados_raw)

bind_rows(
  # teste de log rank
  surv_pvalue(sfit, method = "log-rank"),

  # usando Gehan-Breslow (Wilcoxon generalizado)
  surv_pvalue(sfit, method = "gehan-breslow"),

  # usando Peto-Peto
  surv_pvalue(sfit, method = "S1"),

  # usando Fleming-Harrington(p=1, q=1)
  surv_pvalue(sfit, method = "FH_p=1_q=1"),
) %>%
  dplyr::relocate(method) %>%
  dplyr::select(-pval.txt)

# QUESTAO 7c ----

```

```
# modelo
sfit <- survfit(Surv(ttr, relapse)~grp+strata(employment), data=dados_raw)

ggsurvplot(sfit,dados_raw)+
  labs(x="Tempo", y=expression(hat(S)(t)))

survdif(Surv(ttr, relapse)~grp+strata(employment), data=dados_raw)
```