MAE0514 - Introducão a Análise de Sobrevivência - Lista 2

Bruno de Castro Paul Schultze[1]
Rubens Santos Andrade Filho[2]

Junho de 2021

Sumário

[1]Número USP: 10736862
[2]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 (capecipabina) 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

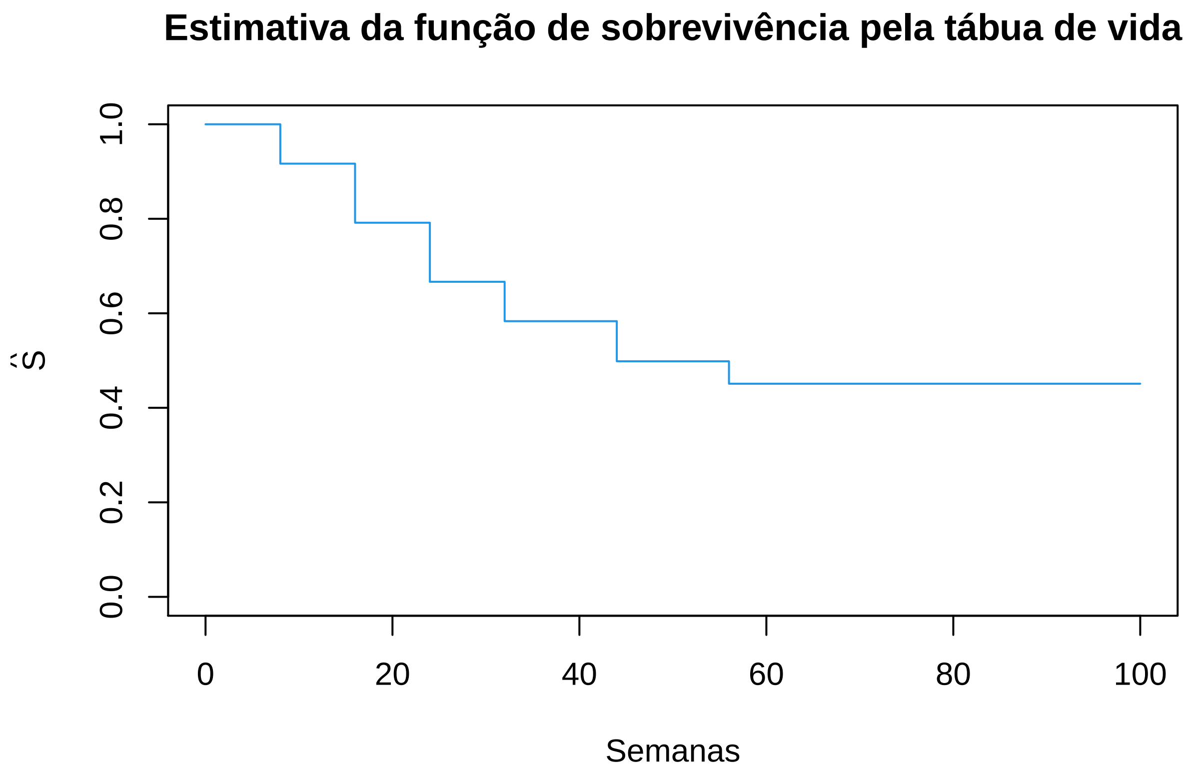
Dessa forma, consideramos os intervalos fechados à direita.

```
# QUESTAO 3a ----

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

```r
# 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



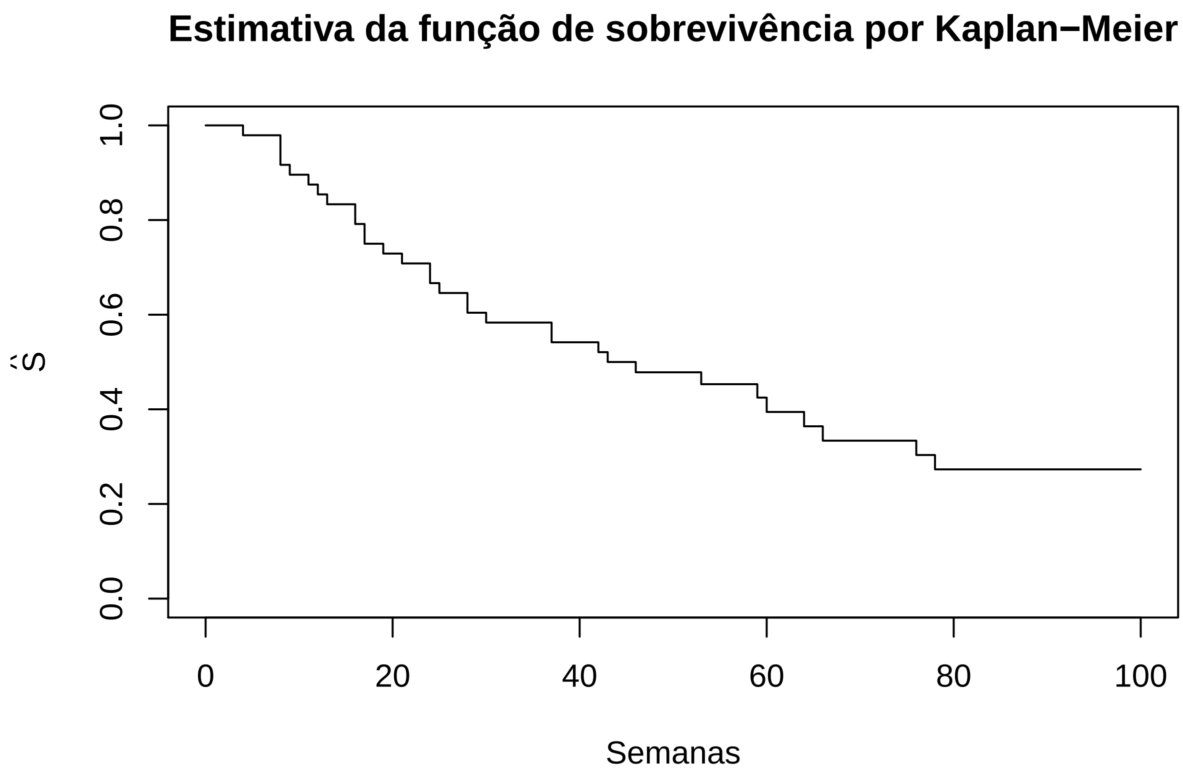
Chama a atenção o fato da estimativa da função de sobrevivência não se aproximar de 0 à medida que aumentam o número de semanas. Isso acontece principalmente devido às 11 observações censuradas que ficaram no último intervalo, isto é, temos menos muito menos informação a respeito das falhas nesse intervalo.

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

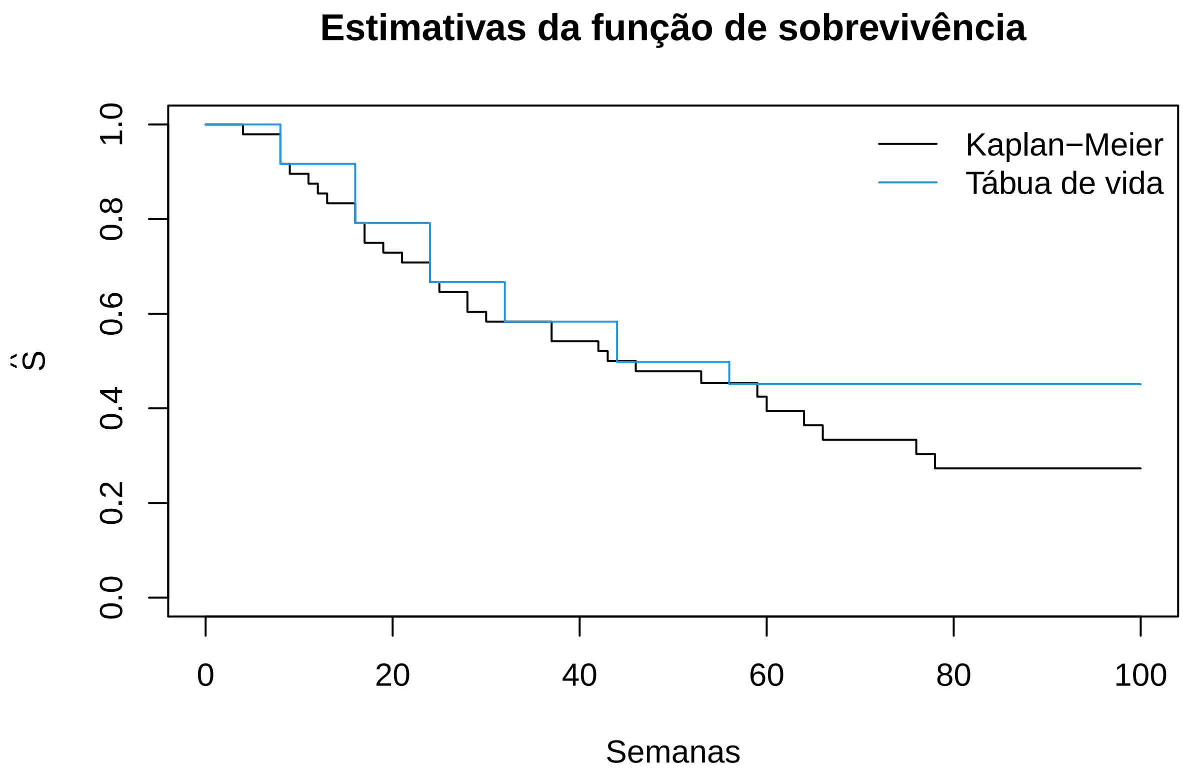
Tabela 2: Estimativas de Kaplan-Meier.

Semana	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

Semana	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



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



O método de Kaplan-Meier é o método da tábua de vida quando o 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 sobrevivencia, 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	

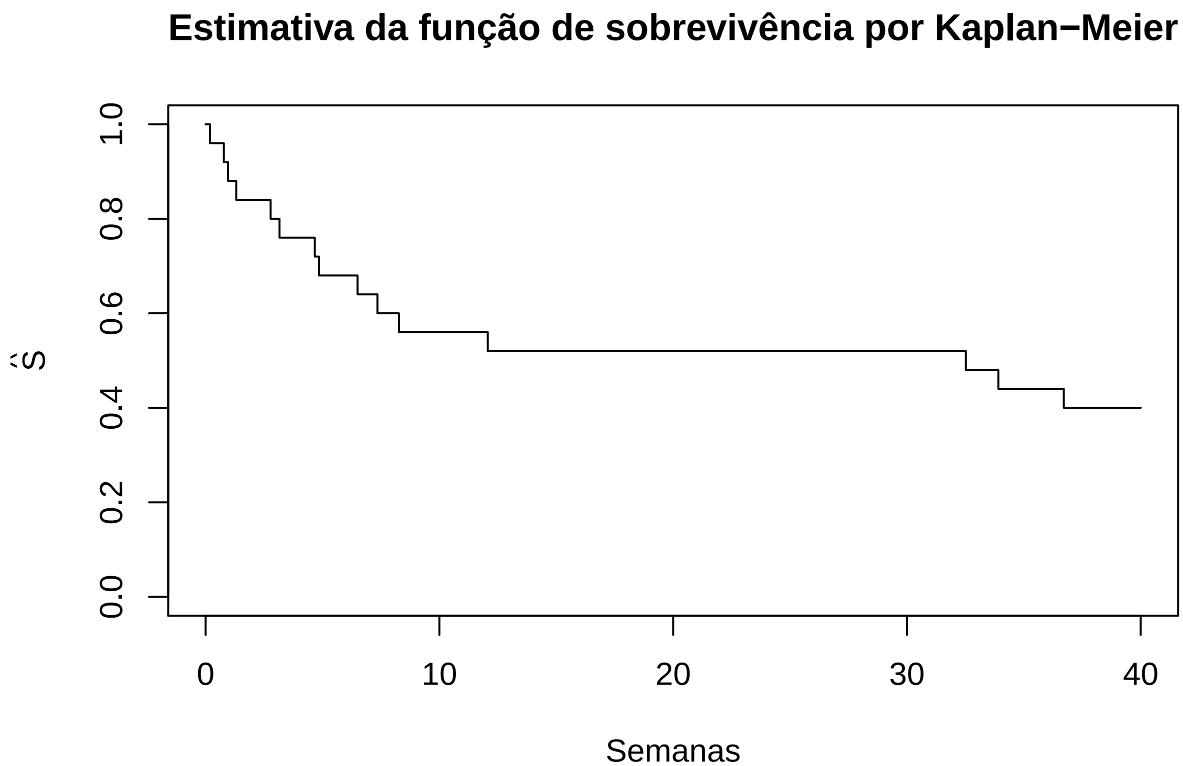
Observe que 10 observações foram censuradas. Para este exercício, os cálculos podem ser feitos *à* mão ou com auxílio computacional, porém a ideia é não utilizar uma função pronta que calcule o que for pedido. Você deve usar uma planilha ou escrever o código que faça as contas no R ou outro software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

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

```r
# 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)

kmeier <- dados %>%
  group_by(t) %>%
  # 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 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



(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 = kmeier$s_hat[which(kmeier$t==t0)]
s1 = kmeier$s_hat[which(kmeier$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,

```r
# QUESTAO 5c ----

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

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

onde

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

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

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

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

```r
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, poderiamos obter por interpolação linear como nos outros itens.

Questão 6

Questão 7

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 ----
kmeier <- dados_raw %>%
  group_by(t=timeWeeks) %>%
  # 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)
```

```r
kmeier %>%
  kable(
    caption = "Estimativas de Kaplan-Meier.",
    col.names = c(
      "Semana","$d_i$","$w_i$",
      "$Y_i$",
      "$q_i=d_i/Y_i$",
      "$\\hat{S}(t)$"))


x_km <- c(0, rep(kmeier$t, each=2), 100)
y_km <- c(1, 1, rep(kmeier$s_hat, each=2))
plot(x_km, y_km, type="l", col=1, xlab="Semanas", ylab=expression(hat(S)),ylim = c(0,1),
     main = "Estimativa da função de sobrevivência por Kaplan-Meier", cex=.6)


# QUESTAO 3c ----
plot(x_km, y_km, type="l", col=1, xlab="Semanas", ylab=expression(hat(S)),ylim = c(0,1),
     main = "Estimativas da função de sobrevivência", cex=.6)
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)

kmeier <- dados %>%
  group_by(t) %>%
  # 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)
```

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


x_km <- c(0, rep(kmeier$t, each=2), 40)
y_km <- c(1, 1, rep(kmeier$s_hat, each=2))
plot(x_km, y_km, type="l", col=1, xlab="Semanas", ylab=expression(hat(S)),ylim = c(0,1),
     main = "Estimativa da função de sobrevivência por Kaplan-Meier", cex=.6)


# QUESTAO 5b ----

t0 = 12.07
t1 = 32.52
s0 = kmeier$s_hat[which(kmeier$t==t0)]
s1 = kmeier$s_hat[which(kmeier$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 = kmeier$s_hat[which(kmeier$t==t0)]
s1 = kmeier$s_hat[which(kmeier$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 <- kmeier %>% 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)
```