Gabarito da 2ª Lista de Exercícios - MAE514

Professora: Gisela Tunes **Monitor:** Rodrigo Passos Martins

Exercício 1

Vamos considerar $T_1, T_2, \dots, T_n$ tempos de falha sujeitos a censura à direita, de forma que se observa $Z_i = min(T_i, C_i)$ e $\delta_i = \mathbb{1}(T_i \leq C_i)$, em que C_i são os tempos de censura, $i = 1, 2, \dots, n$. Sejam $t_1 < t_2 < ... < t_D$ os instantes em que alguma falha foi observada e defina n_j como sendo o número de indivíduos em risco em t_j (ou seja, indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j) e d_j o número de falhas observadas em t_j. O estimador de Kaplan-Meier da função de sobrevivência associada aos tempos de falha é dado por:

$$\hat{S}(t) = \begin{cases} 1, & \text{se } t < t_1 \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), & \text{se } t_1 \leq t \end{cases}$$

A variância de $\hat{S}(t)$ pode ser estimada pela fórmula de Greenwood, dada por:

$$V\hat{a}r(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j \leq t} \left(\frac{d_j}{(n_j - d_j).n_j} \right)$$

a)

Vamos mostrar que, **se não há censuras**, o estimador Kaplan-Meier se reduz a:

$$\hat{S}(t) = \frac{\text{nº obs.} > t}{n},$$

que é a função de sobrevivência empírica.

Vale observar que, para esse caso (em que não temos censuras), $\delta_i = \mathbb{1}(T_i \leq C_i) = 1, \forall i \geq 1$.

Primeiro, vamos observar que, para $t < t_1$, o nº obs. $> t$ é igual a n, logo:

$$\hat{S}(t) = \frac{\text{nº obs.} > t}{n} = \frac{n}{n} = 1$$

Agora, vamos analisar o caso em que $t_1 \leq t$. Temos que:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j}\right)$$

Quando não temos censuras, $n_j - d_j = n_{j+1}$, porque o número de indívíduos em risco no tempo t_{j+1} (n_{j+1}) é o mesmo número de indívíduos em risco no tempo t_j (n_j) retirando os indivíduos que falharam (d_j). Assim, considerando o último instante como k, temos que o estimador de Kaplan-Meier pode ser escrito como:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{t_j \leq t} \left(\frac{n_{j+1}}{n_j}\right) = \left(\frac{\cancel{n_2}}{n_1}\right) \cdot \left(\frac{\cancel{n_3}}{\cancel{n_2}}\right) \cdot \left(\frac{\cancel{n_4}}{\cancel{n_3}}\right) \cdot \dots \cdot \left(\frac{\cancel{n_k}}{\cancel{n_{k-1}}}\right) \cdot \left(\frac{n_{k+1}}{\cancel{n_k}}\right) = \frac{n_{k+1}}{n_1}$$

Como não temos censuras, $n = n_1$. Além disso, observe que n_{k+1} representa o número de observações que não falharam no instante imediatamente anterior a t. Portanto,

$$\hat{S}(t) = \frac{n_{k+1}}{n_1} = \frac{\text{n}^{\text{o}} \text{ obs.} > t}{n} \blacksquare$$

b)

Vamos mostrar que a fórmula de Greenwood se reduz a seguinte estimativa da variância de proporção:

$$V\hat{a}r(\hat{S}(t)) = n^{-1}.\hat{S}(t).(1 - \hat{S}(t))$$

A fórmula de Greenwood, é dada por:

$$V\hat{a}r(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j \leq t} \left(\frac{d_j}{(n_j - d_j).n_j} \right)$$

Como não temos censuras, sabemos pelo item **a)** que $n_j - d_j = n_{j+1}$. Além disso, é fácil ver que $d_j = n_j - n_{j+1}$. Logo, podemos escrever o termo do somatório da fórmula anterior como:

$$\frac{d_j}{(n_j - d_j).n_j} = \frac{n_j - n_{j+1}}{n_{j+1}.n_j} = \frac{1}{n_{j+1}} - \frac{1}{n_j}$$

Logo, considerando o último instante como k, temos uma soma telescópica:

$$\sum_{j:t_j \leq t} \left(\frac{d_j}{(n_j - d_j).n_j} \right) = \sum_{j:t_j \leq t} \left(\frac{1}{n_{j+1}} - \frac{1}{n_j} \right) = \cancel{\frac{1}{n_2}} - \frac{1}{n_1} + \cancel{\frac{1}{n_3}} - \cancel{\frac{1}{n_2}} + \cdots + \cancel{\frac{1}{n_k}} - \cancel{\frac{1}{n_{k-1}}} + \frac{1}{n_{k+1}} - \cancel{\frac{1}{n_k}} = \frac{1}{n_{k+1}} - \frac{1}{n_1}$$

Assim, como $n_1 = n$ (sem censuras), podemos simplificar a expressão da fórmula de Greenwood:

$$V\hat{a}r(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j \leq t} \left(\frac{d_j}{(n_j - d_j).n_j} \right) = [\hat{S}(t)]^2 . \left(\frac{1}{n_{k+1}} - \frac{1}{n_1} \right) = [\hat{S}(t)].[\hat{S}(t)]. \left(\frac{1}{n_{k+1}} - \frac{1}{n} \right)$$

Lembrando-se do item **a)**, $\hat{S}(t) = \frac{n_{k+1}}{n}$, então,

$$V\hat{a}r(\hat{S}(t)) = [\hat{S}(t)].[\hat{S}(t)]. \left(\frac{1}{n_{k+1}} - \frac{1}{n} \right) = \hat{S}(t).\frac{n_{k+1}}{n}. \left(\frac{1}{n_{k+1}} - \frac{1}{n} \right) = \frac{1}{n}.\hat{S}(t). \left(\frac{n_{k+1}}{n_{k+1}} - \frac{n_{k+1}}{n} \right) = n^{-1}.\hat{S}(t). \left(1 - \hat{S}(t) \right) \blacksquare$$

Exercício 2

Vamos considerar um estudo sobre AZT, medicamento utilizado para tratar pacientes com HIV. Os dados são de 45 pacientes que foram acompanhados desde sua entrada no estudo até a morte. Os dados contêm informação sobre a idade do paciente quando entrou no estudo, a idade que tinha quando faleceu e o indicador de falha.

Vamos importar os arquivos com os dados do estudo em R:

```
# IMPORTANDO OS DADOS (EM R)

library(readr)
dados_HIV <- read_table2("Lista2_HIV.txt",
                         col_names = FALSE)
names(dados_HIV) <- c("PACIENTE", "IDADE_ENTRADA", "IDADE_MORTE", "IND_FALHA")
```

a)

Vamos fazer a variável "tempo", que represetará o número de meses entre a entrada no estudo e óbito (ou tempo de censura):

```
# FAZENDO A VARIÁVEL "TEMPO" NO CONJUNTO DE DADOS (EM R)

dados_HIV$TEMPO <- dados_HIV$IDADE_MORTE - dados_HIV$IDADE_ENTRADA

dados_HIV$TEMPO
```

```
##  [1] 223 247 352 226 361 191 267 184   2 285 508 413 475  18  61  82 135
## [18]  31  67 249 280 558 135 302 146 193 355 251 104 350  30 254 324 355
## [35] 272 495 173 496  45 207 328 451 351 165 420
```

b)

Calculemos o estimador da tábua de vida, considerando as seguintes faixas de tempo:

- **Faixa 1:** 60 meses ou menos
- **Faixa 2:** de 60 (exclusive) a 120 meses
- **Faixa 3:** de 120 (exclusive) a 240 meses
- **Faixa 4:** de 240 (exclusive) a 360 meses
- **Faixa 5:** de 360 (exclusive) a 480 meses
- **Faixa 6:** mais de 480 meses

Ou seja, temos, $t_0 = 0$, $t_1 = 60$, $t_2 = 120$, $t_3 = 240$, $t_4 = 360$, $t_5 = 480$, $t_6 = +\infty$.

Na notação, temos que, d_j é o número de falhas (no caso, mortes) observadas em $(j-1, j]$, ω_j é o número de censuras em $(j-1, j]$ e n_j^* é o número de indivíduos em risco em t_{j-1}^-. Além disso, temos que a sobrevivência é dada por:

$$\hat{S}(t_j) = \prod_{l=1}^{j} \left(1 - \frac{d_l}{n_l^* - \frac{1}{2}\omega_l}\right)$$

Vamo então codificar a tabela no R:

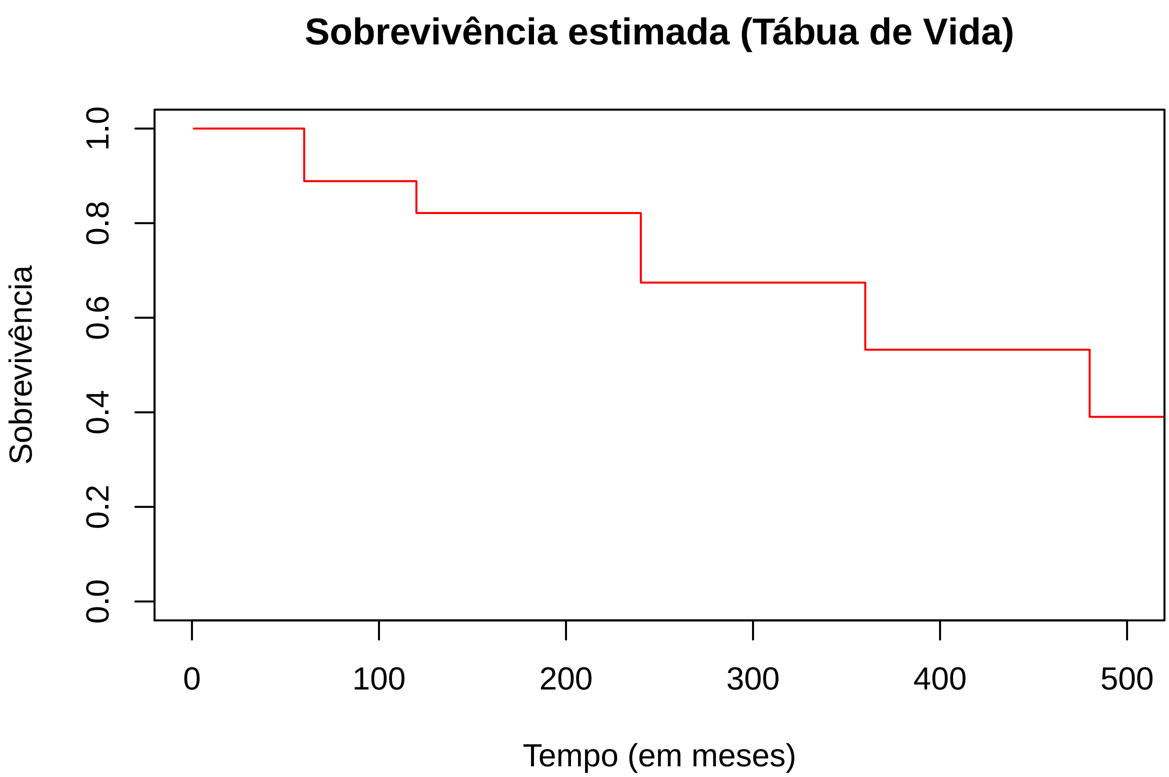
```r
# FAZENDO A TABELA (EM R)

library(dplyr)
library(magrittr)

dados_HIV %<>% mutate(FAIXA = case_when(TEMPO < 60 ~ 1,
                                        TEMPO >= 60 & TEMPO < 120 ~ 2,
                                        TEMPO >= 120 & TEMPO < 240 ~ 3,
                                        TEMPO >= 240 & TEMPO < 360 ~ 4,
                                        TEMPO >= 360 & TEMPO < 480 ~ 5,
                                        TEMPO >= 480 ~ 6))
FALHAS <- dados_HIV %>%
   group_by(FAIXA) %>%
   summarise(FALHAS = sum(IND_FALHA)) %>%
   select(FALHAS)

CENSURAS <- dados_HIV %>%
   group_by(FAIXA) %>%
   mutate(IND_CENSURA = if_else(IND_FALHA == 1, 0, 1)) %>%
   summarise(CENSURAS = sum(IND_CENSURA)) %>%
   select(CENSURAS)

TABELA <- cbind(FAIXAS = 1:6, IND_RISCO = 45, FALHAS, CENSURAS, SOBREVIVENCIA = 1)

for(i in 2:6){
   TABELA$IND_RISCO[i] <- TABELA$IND_RISCO[i-1] - TABELA$CENSURAS[i-1] - TABELA$FALHAS[i-1]
}

for(i in 2:6){
   TABELA$SOBREVIVENCIA[i] <- TABELA$SOBREVIVENCIA[i-1]*(1-TABELA$FALHAS[i-1]/
                              (TABELA$IND_RISCO[i-1]-0.5*TABELA$CENSURAS[i-1]))
}
```

Assim, construímos a seguinte tabela:

Faixa j	n_j^*	d_j	ω_j	$\hat{S}(t_j)$
Faixa 1	45	5	0	1,000
Faixa 2	40	3	1	0,889
Faixa 3	36	6	5	0,821
Faixa 4	25	4	12	0,674
Faixa 5	9	2	3	0,532
Faixa 6	4	0	4	0,390

Agora, vamos fazer o gráfico de Sobrevivência para esse caso:

```r
# FAZENDO O GRÁFICO DE SOBREVIVÊNCIA DE TÁBUA DE VIDA (EM R)

DADOS_GRAFICO_SOBREV <- c(rep(TABELA$SOBREVIVENCIA[1], 60), rep(TABELA$SOBREVIVENCIA[2], 61),
                          rep(TABELA$SOBREVIVENCIA[3], 121), rep(TABELA$SOBREVIVENCIA[4], 121),
                          rep(TABELA$SOBREVIVENCIA[5], 121), rep(TABELA$SOBREVIVENCIA[6], 61))

DADOS_GRAFICO_TEMPO <- c(1:60, 60:120, 120:240, 240:360, 360:480, 480:540)

plot(DADOS_GRAFICO_TEMPO, DADOS_GRAFICO_SOBREV,
     main = "Sobrevivência estimada (Tábua de Vida)",
     xlab = "Tempo (em meses)", ylab = "Sobrevivência",
     type = "l", col = "red",
     xlim = c(0, 500), ylim = c(0, 1))
```



Podemos considerar que a taxa de sobrevivência é relativamente boa, uma vez que, até a Faixa 6 (> 480 meses) temos uma sobrevivência de mais de 50%, o que indica um tempo mediano de sobrevivência alto (> 4 anos). Isso pode ser um bom indício da eficácia do medicamento AZT.

c)

Vamos agora fazer o estimador de Kaplan-Meier desses dados:

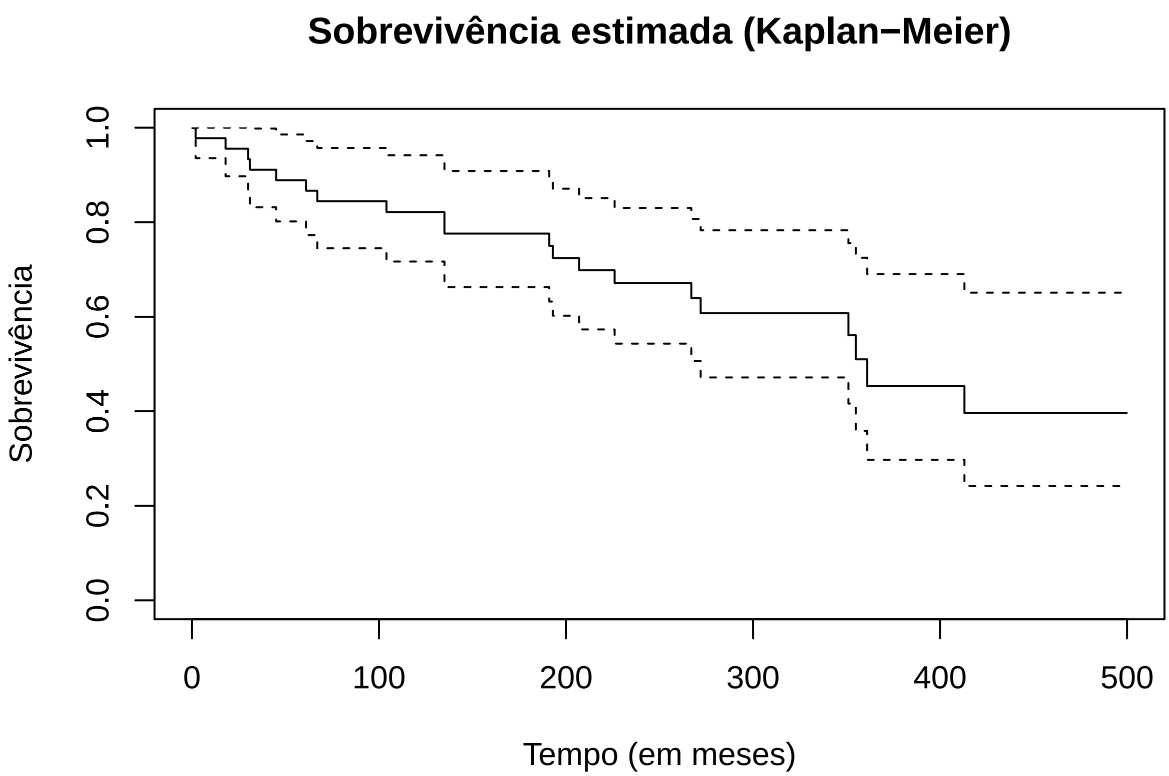
```r
# FAZENDO O GRÁFICO DE SOBREVIVÊNCIA DE KAPLAN-MEIER (EM R)

library(survival)

dados <- Surv(time = dados_HIV$TEMPO, event = dados_HIV$IND_FALHA, type = 'right')

ajuste <- survfit(dados ~ 1, data = dados_HIV)

print(ajuste)
```

```
## Call: survfit(formula = dados ~ 1, data = dados_HIV)
##
##       n  events  median 0.95LCL 0.95UCL
##      45      20     361     272      NA
```

```r
summary(ajuste)
```

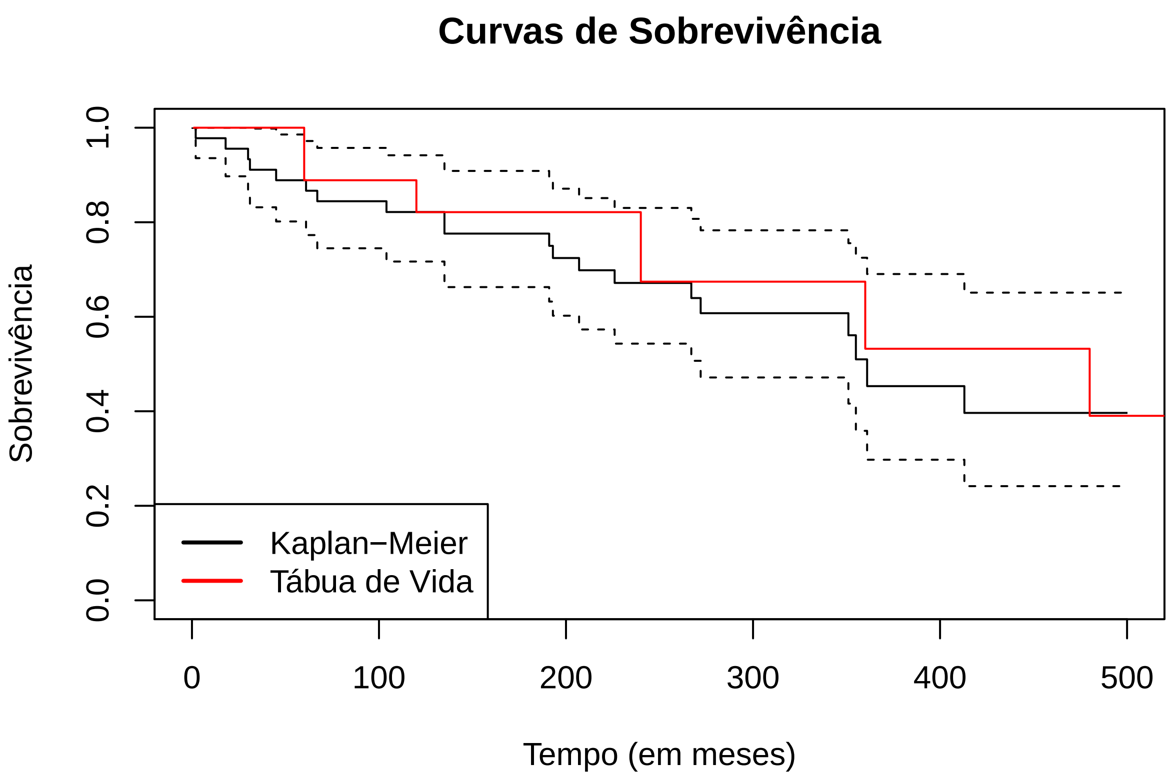
```
## Call: survfit(formula = dados ~ 1, data = dados_HIV)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2     45       1    0.978  0.0220        0.936        1.000
##    18     44       1    0.956  0.0307        0.897        1.000
##    30     43       1    0.933  0.0372        0.863        1.000
##    31     42       1    0.911  0.0424        0.832        0.998
##    45     41       1    0.889  0.0468        0.802        0.986
##    61     40       1    0.867  0.0507        0.773        0.972
##    67     39       1    0.844  0.0540        0.745        0.957
##   104     37       1    0.822  0.0572        0.717        0.942
##   135     36       2    0.776  0.0625        0.663        0.909
##   191     30       1    0.750  0.0655        0.632        0.890
##   193     29       1    0.724  0.0682        0.602        0.871
##   207     28       1    0.698  0.0705        0.573        0.851
##   226     26       1    0.672  0.0727        0.543        0.830
##   267     21       1    0.640  0.0759        0.507        0.807
##   272     20       1    0.608  0.0786        0.472        0.783
##   351     13       1    0.561  0.0853        0.416        0.756
##   355     11       1    0.510  0.0915        0.359        0.725
##   361      9       1    0.453  0.0973        0.297        0.690
##   413      8       1    0.397  0.1003        0.242        0.651
```

```r
plot(ajuste,
     main = "Sobrevivência estimada (Kaplan-Meier)",
     xlab = "Tempo (em meses)", ylab = "Sobrevivência",
     xlim = c(0, 500), ylim = c(0, 1))
```



d)

```r
# FAZENDO O GRÁFICO COM AS DUAS CURVAS (EM R)
plot(ajuste,
     main = "Curvas de Sobrevivência",
     xlab = "Tempo (em meses)", ylab = "Sobrevivência",
     xlim = c(0, 500), ylim = c(0, 1))

par(new = T)

plot(DADOS_GRAFICO_TEMPO, DADOS_GRAFICO_SOBREV,
     main = "", xlab = "", ylab = "", axes = F,
     type = "l", col = "red",
     xlim = c(0, 500), ylim = c(0, 1))

legend("bottomleft",
       legend = c("Kaplan-Meier",
                  "Tábua de Vida"),
       cex = 1, lwd = 2,
       col = c("black", "red"))
```



Podemos perceber que a curva de Sobrevivência estimada pela Tábua de Vida esá sempre acima da Kaplan_meier, apesar de se conservar dentro do IC dessa. Podemos dizer então que o estimador de Tábua de Vida superestima a sobrevivência dos pacientes no estudo, ou seja, ele é mais aponta resultados mais otimistas por assim dizer. Contudo, pela ténica empregada, o estimador de Kaplan-Meier tende a ser o mais factível.

Exercício 3

Vamos analisar os dados referentes a um estudo em pacientes com leucemia, que são os tempos de remissão (período em que o paciente está sem tratamento e sem a doença, ou seja, período compreendido entre o fim do tratamento e a reincidência da leucemia). Os pacientes foram submetidos a dois diferentes tratamentos e os tempos, em dias, de remissão estão abaixo:

- **Tempos de Remissão para o Tratamento 1:** 5, 5, 9, 10, 12, 12, 10, 23, 28, 28, 28, 29, 32, 32, 37, 41, 41, 57, 62, 74, 100, 139, $20^+, 258^+, 269^+$
- **Tempos de Remissão para o Tratamento 2:** 8, 10, 10, 12, 14, 20, 48, 70, 75, 99, 103, 162, 169, 195, 220, $161^+, 199^+, 217^+, 245^+$

Os tempos censurados à direita são denotados por um sinal "+".

a)

Vamos calcular **à mão** o estimador de Kaplan-Meier para esses dados. As fórmulas usadas para esses cálculos são:

- $\hat{S}(0) = 1$
- $\hat{S}(t) = \left(1 - \dfrac{d_j}{n_j}\right).\hat{S}(t-1)$
- $\hat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 . \sum_{t_j \geq t} \dfrac{d_j}{n_j.(n_j - d_j)}$

As informações e resultados estão na tabela abaixo:

Tratamento 1					**Tratamento 2**				
t_j	n_j	d_j	$\hat{S}(t_j)$	$\hat{Var}(\hat{S}(t_j))$	t_j	n_j	d_j	$\hat{S}(t_j)$	$\hat{Var}(\hat{S}(t_j))$
0	25	0	1,000	0,000	0	19	0	1,000	0,000
5	25	2	0,920	0,003	8	19	1	0,947	0,003
9	23	1	0,880	0,004	10	18	2	0,842	0,007
10	22	2	0,800	0,006	12	16	1	0,789	0,009
12	20	2	0,720	0,008	14	15	1	0,736	0,010
23	17	1	0,678	0,009	20	14	1	0,683	0,011
28	16	3	0,551	0,010	48	13	1	0,630	0,012
29	13	1	0,509	0,010	70	12	1	0,577	0,013
32	12	2	0,424	0,010	75	11	1	0,524	0,013
37	10	1	0,382	0,010	99	10	1	0,472	0,013
41	9	2	0,297	0,009	103	9	1	0,420	0,013
57	7	1	0,255	0,008	162	7	1	0,360	0,013
62	6	1	0,212	0,007	169	6	1	0,300	0,012
74	5	1	0,170	0,006	195	5	1	0,240	0,010
100	4	1	0,127	0,005	220	2	1	0,120	0,010
139	3	1	0,085	0,003	-	-	-	-	-

b)

Agora, vamos calcular as estimativas pontuais e intervalares (com coeficiente de 90%) para a <u>mediana</u> de cada tratamento.

Vamos utilizar a interpolação linear com:

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)}$$

No nosso caso, temos que p = 0,5 (quantil da mediana) e os tempos envolvidos são tais que:

- $\hat{S}(t_L) > 0,5$
- $\hat{S}(t_U) < 0,5$

Então, para o **Tratamento 1**, temos que $t_L = 29$ e $t_U = 32$, sendo as sobrevivências atribuídas $\hat{S}(t_L) = 0,509$ e $\hat{S}(t_U) = 0,424$, respectivamente. Assim,

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)} \longrightarrow \frac{32-29}{0,424-0,509} = \frac{\hat{t}_{0,5} - 29}{0,5-0,509} \Leftrightarrow \frac{-3}{0,085} = \frac{\hat{t}_{0,5} - 29}{-0,009} \Leftrightarrow \hat{t}_{0,5} \cong 29,32$$

Já para o **Tratamento 2**, temos que $t_L = 75$ e $t_U = 99$, sendo as sobrevivências atribuídas $\hat{S}(t_L) = 0,524$ e $\hat{S}(t_U) = 0,472$, respectivamente. Assim,

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)} \longrightarrow \frac{99-75}{0,472-0,524} = \frac{\hat{t}_{0,5} - 75}{0,5-0,524} \Leftrightarrow \frac{-24}{0,052} = \frac{\hat{t}_{0,5} - 75}{-0,024} \Leftrightarrow \hat{t}_{0,5} \cong 86,08$$

Essas são as estimativas pontuais. Agora, vamos calcular as estimativas intervalares com confiança de 90% com:

$$IC(t_{0,5}; \gamma = 90\%) = \hat{t}_{0,5} \pm z_{\gamma/2}.\sqrt{\hat{Var}(\hat{t}_{0,5})}$$

Sendo que:

$$\hat{Var}(\hat{t}_{0,5}) = \frac{\hat{Var}(\hat{S}(\hat{t}_{0,5}))}{[\hat{f}(\hat{t}_{0,5})]^2} = \frac{[\hat{S}(t_{0,5})]^2 . \sum_{t_j \leq t_{0,5}} \frac{d_j}{n_j.(n_j - d_j)}}{\left[\frac{\hat{S}(t_L) - \hat{S}(t_U)}{t_U - t_L}\right]^2}$$

Para o **Tratamento 1**, os cálculos estão abaixo:

$$\hat{Var}(\hat{t}_{0,5}) = \frac{[\hat{S}(t_{0,5})]^2 . \sum_{t_j \leq t_{0,5}} \frac{d_j}{n_j.(n_j - d_j)}}{\left[\frac{\hat{S}(t_L) - \hat{S}(t_U)}{t_U - t_L}\right]^2} = \frac{[0,5]^2 . 0,04006536}{0,0008027776} \cong 12,48$$

$$IC(t_{0,5}; 90\%) = 29,32 \pm 1,64.\sqrt{12,48} = 29,32 \pm 5,79 = [23,53; 35,11]$$

Enquanto que para o **Tratamento 2**, os cálculos são:

$$\hat{Var}(\hat{t}_{0,5}) = \frac{[\hat{S}(t_{0,5})]^2 . \sum_{t_j \leq t_{0,5}} \frac{d_j}{n_j.(n_j - d_j)}}{\left[\frac{\hat{S}(t_L) - \hat{S}(t_U)}{t_U - t_L}\right]^2} = \frac{[0,5]^2 . 0.04736842}{0,000004694444} \cong 2522,58$$

$$IC(t_{0,5}; 90\%) = 86,08 \pm 1,64.\sqrt{2522,58} = 86,08 \pm 82.37 = [3,71; 168,45]$$

c)

O tempo médio de sobrevivência pode ser calculado com a área de cada um dos "retângulos" formados pelo gráfico. Assim, devemos multiplcar a altura (probabilidade de sobrevivência) pela largura (diferença entre os tempos). Essa mecânica resulta na seguinte fórmula:

$$\hat{t}_{médio} = t_1 + \sum_{j=1}^{k-1} \hat{S}(t_j).(t_{j+1} - t_j)$$

Note que o primeiro tempo (t_1) não está sendo multiplicado pela probabilidade porque ela inicialmente é 1.

Sendo assim, para o **Tratamento 1** temos que:

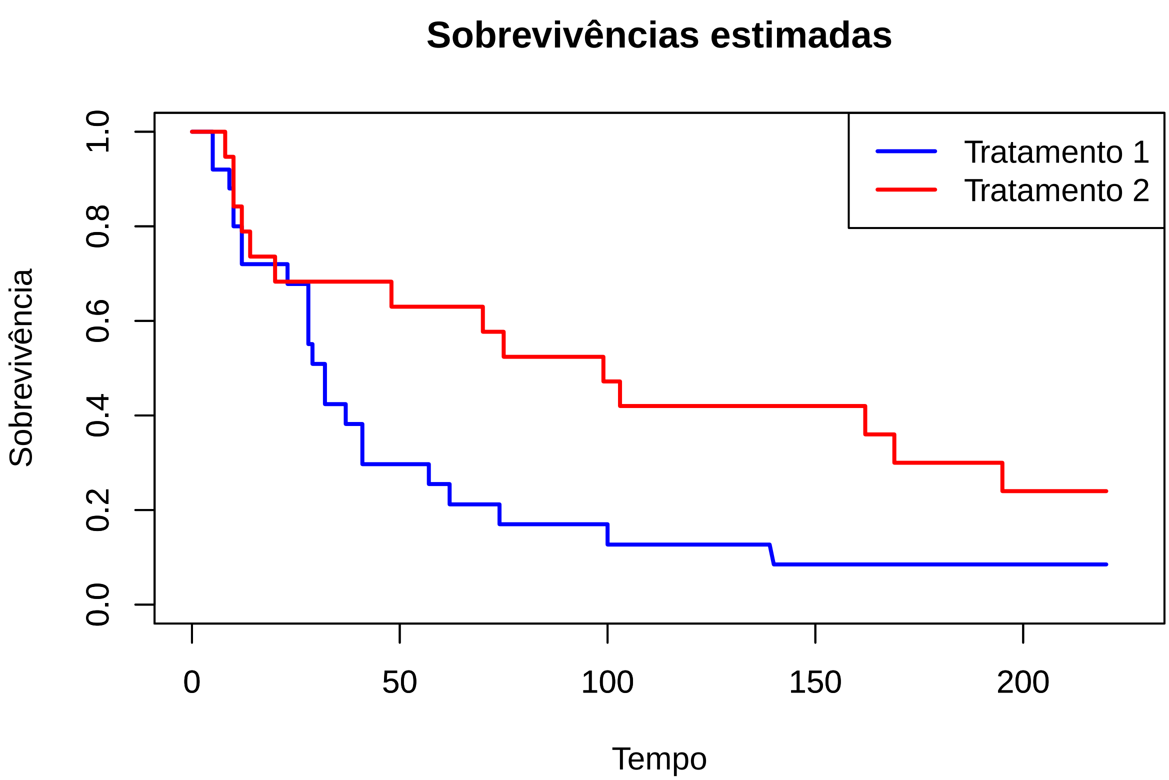
$$\hat{t}_{médio} = t_1 + \sum_{j=1}^{k-1} \hat{S}(t_j).(t_{j+1} - t_j) = 5 + 0,920.(9-5) + 0,880.(10-9) + \cdots + 0,127.(139-100) = 46,14 \cong 46$$

Já para o **Tratamento 2** temos que:

$$\hat{t}_{médio} = t_1 + \sum_{j=1}^{k-1} \hat{S}(t_j).(t_{j+1} - t_j) = 8 + 0,947.(10-8) + 0,842.(12-10) + \cdots + 0,240.(220-195) = 109,005 \cong 109$$

d)

Abaixo, temos o gráfico com as curvas estimadas para cada tratamento (feito apenas com a função *plot* do R):



e)

Agora, vamos calcular o estimador de Nelson-Aalen. Para tal, usaremos as seguintes expressões:

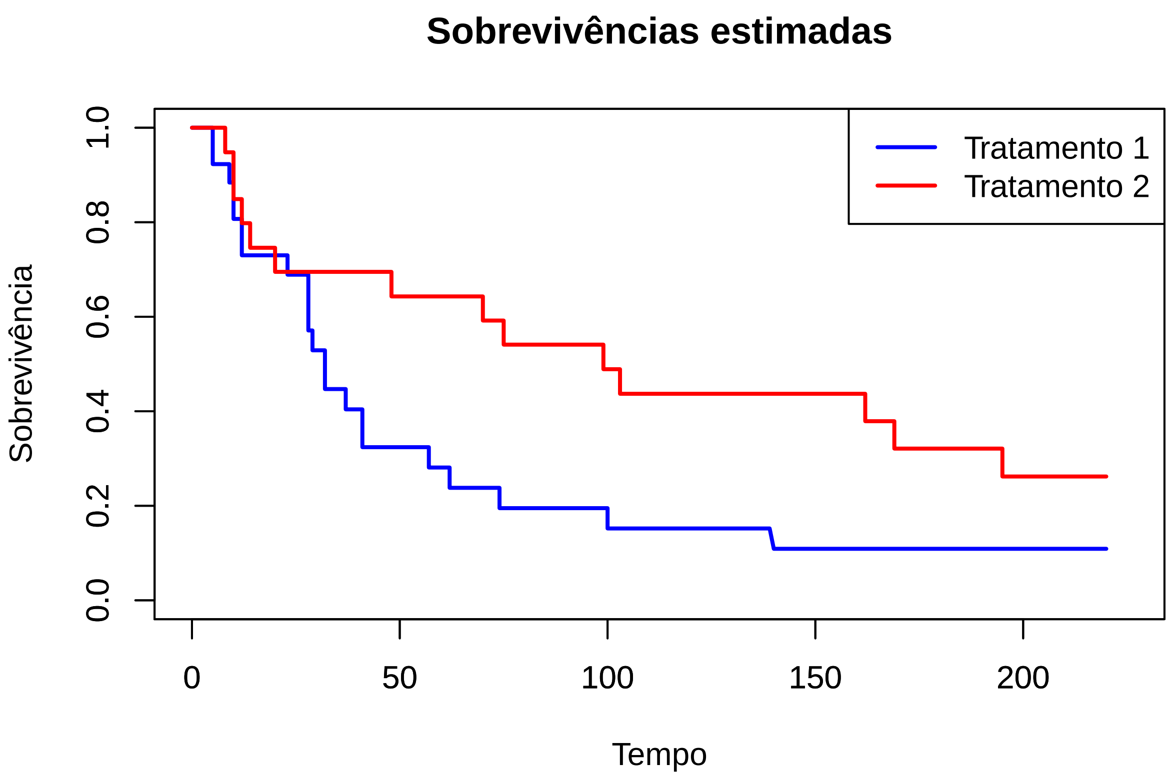
- $\hat{\Lambda}_{NA}(t) = \hat{\Lambda}_{NA}(t-1) + \dfrac{d_j}{n_j}$
- $\hat{\Lambda}_{NA}(0) = 0$
- $\hat{S}_{NA}(t) = exp(-\hat{\Lambda}_{NA}(t))$

Note que d_j e n_j são definidos da mesma forma que anteriormente no estimador de Kaplan-Meier.

Os resultados das estimativas estão na tabela abaixo:

Tratamento 1					**Tratamento 2**				
t_j	n_j	d_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}_{NA}(t_j)$	t_j	n_j	d_j	$\hat{\Lambda}_{NA}(t)$	$\hat{S}_{NA}(t_j)$
0	25	0	0,000	1,000	0	19	0	0,000	1,000
5	25	2	0,080	0,923	8	19	1	0,053	0,948
9	23	1	0,123	0,884	10	18	2	0,164	0,849
10	22	2	0,214	0,807	12	16	1	0,226	0,798
12	20	2	0,314	0,730	14	15	1	0,293	0,746
23	17	1	0,373	0,689	20	14	1	0,364	0,695
28	16	3	0,561	0,571	48	13	1	0,441	0,643
29	13	1	0,637	0,529	70	12	1	0,524	0,592
32	12	2	0,804	0,447	75	11	1	0,615	0,541
37	10	1	0,904	0,404	99	10	1	0,715	0,489
41	9	2	1,126	0,324	103	9	1	0,827	0,437
57	7	1	1,269	0,281	162	7	1	0,969	0,379
62	6	1	1,436	0,238	169	6	1	1,136	0,321
74	5	1	1,636	0,195	195	5	1	1,336	0,262
100	4	1	1,886	0,152	220	2	1	1,836	0,159
139	3	1	2,219	0,109	-	-	-	-	-

Assim, é possível fazer o próximo gráfico com as curvas estimadas para cada tratamento (feito apenas com a função *plot* do R):



f)

Vamos calcular a estatística de *log-rank* com as seguintes expressões:

- n_{1j}: Números de indivíduos em risco no tempo j no Tratamento 1
- n_{2j}: Números de indivíduos em risco no tempo j no Tratamento 2
- n_j: Números de indivíduos em risco no tempo j (em ambos os tratamentos)
- d_{1j}: Números de falhas no tempo j no Tratamento 1
- d_{2j}: Números de falhas no tempo j no Tratamento 2
- d_j: Números de falhas no tempo j (em ambos os tratamentos)
- $\omega_{2j} = \dfrac{(d_{1j} + d_{2j}).n_{2j}}{n_j}$
- $v_{2j} = d_j.\dfrac{n_{1j}.n_{2j}.(n_j - d_j)}{n_j^2.(n_j - 1)}$

Tendo isso em vista, a estatística de *log-rank* é dada por:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - \omega_{2j})\right]^2}{\sum_{j=1}^{k} v_{2j}}$$

Então, vamos colocar todos os valores envolvendo n_{1j}, n_{2j}, n_j, d_{1j}, d_{2j}, d_j, ω_{2j} e v_{2j} na seguinte tabela:

t_j	n_{1j}	n_{2j}	n_j	d_{1j}	d_{2j}	d_j	ω_{2j}	v_{2j}
5	25	19	44	2	0	2	0,864	0,479
8	23	19	42	0	1	1	0,452	0,248
9	23	19	42	1	0	1	0,452	0,248
10	22	18	40	2	2	4	1,800	0,914
12	20	16	36	2	1	3	1,333	0,698
14	20	15	35	0	1	1	0,429	0,245
20	20	14	34	0	1	1	0,412	0,242
23	17	14	31	1	0	1	0,452	0,248
28	16	14	30	3	0	3	1,400	0,695
29	13	14	27	1	0	1	0,519	0,250
32	12	14	26	2	0	2	1,077	0,477
37	10	14	24	1	0	1	0,583	0,243
41	9	14	23	2	0	2	1,217	0,455
48	9	13	22	0	1	1	0,591	0,242
57	7	13	20	1	0	1	0,650	0,228
62	6	13	19	1	0	1	0,684	0,216
70	6	12	18	0	1	1	0,667	0,222
74	5	12	17	1	0	1	0,706	0,208
75	5	11	16	0	1	1	0,688	0,215
99	5	10	15	0	1	1	0,667	0,222
100	4	10	14	1	0	1	0,714	0,204
103	4	9	13	0	1	1	0,692	0,213
139	3	9	12	1	0	1	0,750	0,188
162	3	7	10	0	1	1	0,700	0,210
169	3	6	9	0	1	1	0,667	0,222
195	3	5	8	0	1	1	0,625	0,234
220	3	2	5	0	1	1	0,400	0,240

Com isso, temos que a estatística T de *log-rank* é 3,17. Sob algumas condições, razoáveis para o nosso caso, T possui converge em distribuição para uma Qui-quadrado com 1 grau de liberdade. Assim, o valor-p referente é de, aproximadamente, 0,08, o que indica uma diferença estatisticamente significativa entre as duas curvas. Logo, há diferença entre os dois tratamentos.

Exercício 4

Vamos analisar 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 isolante 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

Como a quantidade de falhas preterida foi alcançada, tivemos 10 censuras.

a)

Vamos calcular a função de sobrevivência estimada por Kaplan-Meier no R, sem utilizar os pacotes. Para tal, vale destacar que a função de sobrevivência por Kaplan-Meier é dada por:

$$\hat{S}(t) = \begin{cases} 1, & \text{se } t < t_1 \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), & \text{se } t_1 \leq t \end{cases}$$

Tendo isso em vista, vamos para os cálculos:

```
# CALCULANDO AS ESTIMATIVAS DE KAPLAN-MEIER (NO R)

TEMPOS <- c(0, 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)

IND_RISCO <- c(25, 25:11)

FALHAS <- c(0, rep(1, 15))

TABELA <- cbind(TEMPOS, IND_RISCO, FALHAS, SOBREVIVENCIA = 1) %>% as.data.frame()

for(i in 2:16){
   TABELA$SOBREVIVENCIA[i] <- TABELA$SOBREVIVENCIA[i-1]*(1-TABELA$FALHAS[i]/TABELA$IND_RISCO[i])
}
```

Os resultados constam na tabela abaixo:

t_j	n_j	d_j	$\hat{S}(t)$
0,00	25	0	1,00
0,19	25	1	0,96
0,78	24	1	0,92
0,96	23	1	0,88
1,31	22	1	0,84
2,78	21	1	0,80
3,16	20	1	0,76
4,67	19	1	0,72
4,85	18	1	0,68
6,50	17	1	0,64
7,85	16	1	0,60
8,27	15	1	0,56
12,07	14	1	0,52
32,52	13	1	0,48
33,91	12	1	0,44
36,71	11	1	0,40

b)

Agora, vamos calcular uma estimativa para o tempo mediano de vida (sobrevivência) deste tipo de isolante elétrico funcionando a essa tensão.

Para tal, vamos utilizar a interpolação linear com:

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)}$$

No nosso caso, temos que p = 0,5 (quantil da mediana) e os tempos envolvidos são tais que:

- $\hat{S}(t_L) > 0,5$
- $\hat{S}(t_U) < 0,5$

Temos que $t_L = 12,07$ e $t_U = 32,52$, sendo as sobrevivências atribuídas $\hat{S}(t_L) = 0,52$ e $\hat{S}(t_U) = 0,48$, respectivamente. Assim,

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)} \longrightarrow \frac{32,52 - 12,07}{0,48 - 0,52} = \frac{\hat{t}_{0,5} - 12,07}{0,5 - 0,52} \Leftrightarrow \frac{-20,45}{0,04} = \frac{\hat{t}_{0,5} - 12,07}{-0,02} \Leftrightarrow \hat{t}_{0,5} \cong 22,30$$

c)

Vamos calcular uma estimativa (pontual e intervalar) para a fração de defeituosos esperada nos 2 primeiros minutos de funcionamento.

Nesse caso, vamos utilizar a interpolação linear para os tempos, sendo um menor e um maior que 2 minutos. Sendo assim, temos que $t_L = 1,31$ e $t_U = 2,78$, sendo as sobrevivências atribuídas $\hat{S}(t_L) = 0,84$ e $\hat{S}(t_U) = 0,8$, respectivamente. Então,

$$\frac{t_U - t_L}{\hat{S}(t_U) - \hat{S}(t_L)} = \frac{\hat{t}_p - t_L}{(1-p) - \hat{S}(t_L)} \longrightarrow \frac{2,78 - 1,31}{0,8 - 0,84} = \frac{2 - 1,31}{\hat{S}(2) - 0,52} \Leftrightarrow \frac{-1.47}{0,04} = \frac{0.69}{\hat{S}(2) - 0,52} \Leftrightarrow \hat{S}(2) = 0,821$$

Essa é a estimativa da fração de não defeituosos (aqueles que sobrevivem). Para os defeituosos temos que:

$$1 - \hat{S}(2) = 1 - 0,821 = 0,179$$

Para estimar o IC, vamos calcular a variância de $\hat{S}(2)$ pela fórmula de Greenwood:

$$V\hat{a}r(\hat{S}(2)) = [\hat{S}(2)]^2. \sum_{j:t_j<2} \frac{d_j}{n_j.(n_j - d_j)} = [0,821]^2. \left[\frac{1}{25.(25-1)} + \frac{1}{24.(24-1)} + \frac{1}{23.(23-1)} + \frac{1}{22.(22-1)}\right] = 0,00691$$

Como o $1 - \hat{S}(2)$ possui a mesma variância de $\hat{S}(2)$ e como não foi especificado o nível de confiança, vamos usar o de 90% para o cálculo do IC, que é dado por:

$$IC(1 - S(2); \gamma = 90\%) = 1 - \hat{S}(2) \pm z_{\gamma/2}.\sqrt{V\hat{a}r(\hat{S}(2))} = 0,179 \pm 1,64.\sqrt{0,00691} = 0,179 \pm 0,136 = [0,043; 0,315]$$

d)

O tempo necessário para 20% dos isolantes estarem fora de operação pode ser estimado via interpolação linear, como nos itens anteriores. Contudo, observando-se a tabela presente no item **a)** é fácil notar que o tempo é 2,78 (pois a sobrevivência estimada nesse tempo é de 80%).

Exercício 5

Os dados desse exercício são referentes a um estudo com 60 pacientes com doença de Hodgkins que receberam tratamento padrão para a doença. O tempo de vida (em meses), bem como idade, sexo, histologia e estágio da doença de cada paciente foi observado.

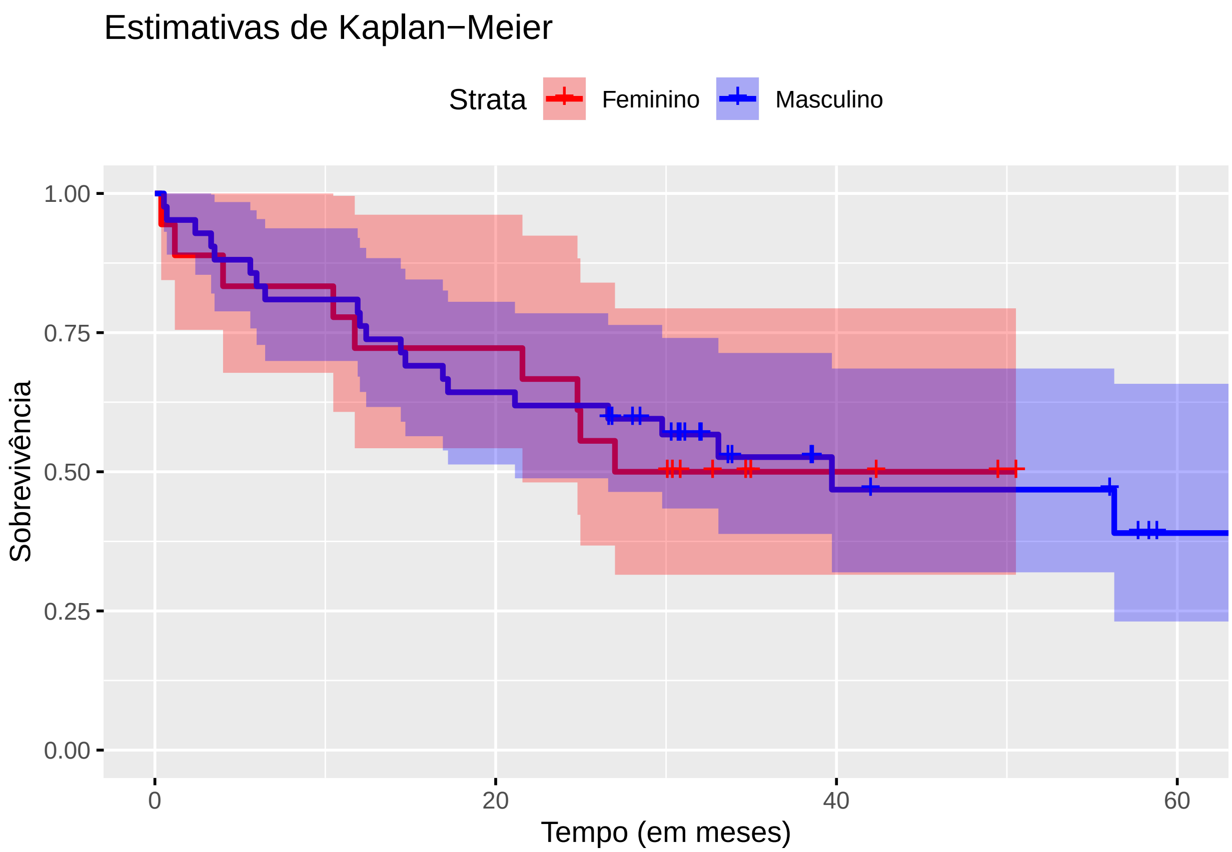
Vamos importar os dados com o R:

```
# IMPORTANDO OS DADOS (EM R)

library(readxl)
dados_Hodgkins <- read_excel("Lista2_Hodgkins.xlsx")
```

a)

Vamos construir, no mesmo gráfico, as curvas de Kaplan-Meier para pacientes do sexo masculino e feminino:

```
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

library(survival)
library(survminer)

S_KM <- survfit(Surv(survivaltime, dead) ~ sex, data = dados_Hodgkins)

ggsurvplot(S_KM, data = dados_Hodgkins, palette = c("red", "blue"), conf.int = T,
           ggtheme = theme_gray(), legend.labs = c("Feminino", "Masculino")) +
   labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```



Além disso, vamos testar a igualdade das duas curvas de sobrevivência, com a estatística *log-rank* e a de Tarone-Ware:

- *Log-rank*: $\dfrac{\left[\sum_{j=1}^{k}(d_{2j} - \omega_{2j})\right]^2}{\sum_{j=1}^{k} v_{2j}}$
- Tarone-Ware: $\dfrac{\left[\sum_{j=1}^{k}\sqrt{n_j}.(d_{2j} - \omega_{2j})\right]^2}{\sum_{j=1}^{k} n_j.v_{2j}}$
- Com n_j, d_j, ω_{2j} e v_{2j} definidos como no **Exercício 3**

```r
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM, method = c("1"))[,1:3] # log-rank
```

```
##   variable      pval   method
## 1      sex 0.8618044 Log-rank
```

```r
surv_pvalue(S_KM, method = c("sqrtN"))[,1:3] # Tarone-Ware
```

```
##   variable   pval      method
## 1      sex 0.8124 Tarone-Ware
```

Ambos os valores-p estão acima de 80%, o que nos leva a inferir que não há motivos para recusar a igualdade entre as curvas.

b)

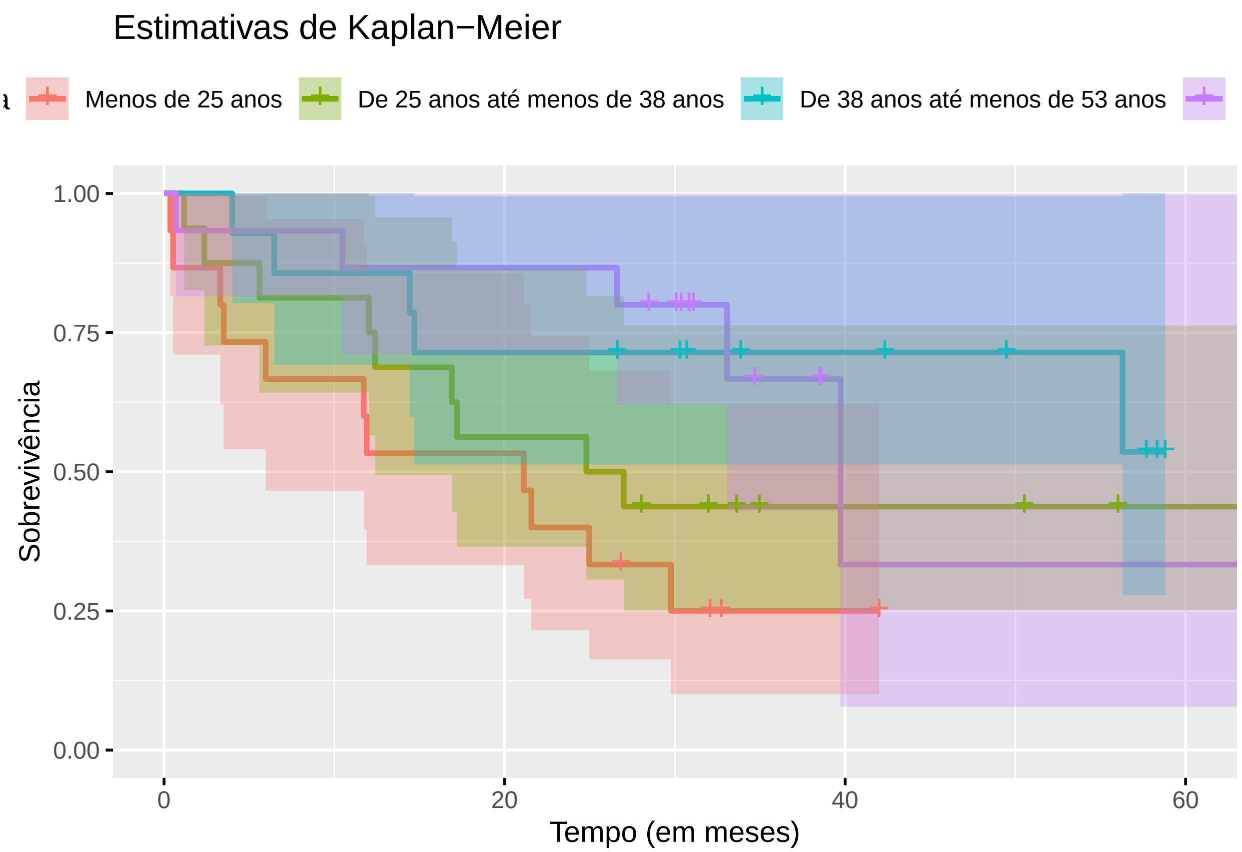
Vamos dividir os pacientes em quatro grupos etários:

- Menos de 25 anos;
- De 25 anos (inclusive) até menos de 38 anos;
- De 38 anos (inclusive) até menos de 53 anos;
- 53 anos ou mais.

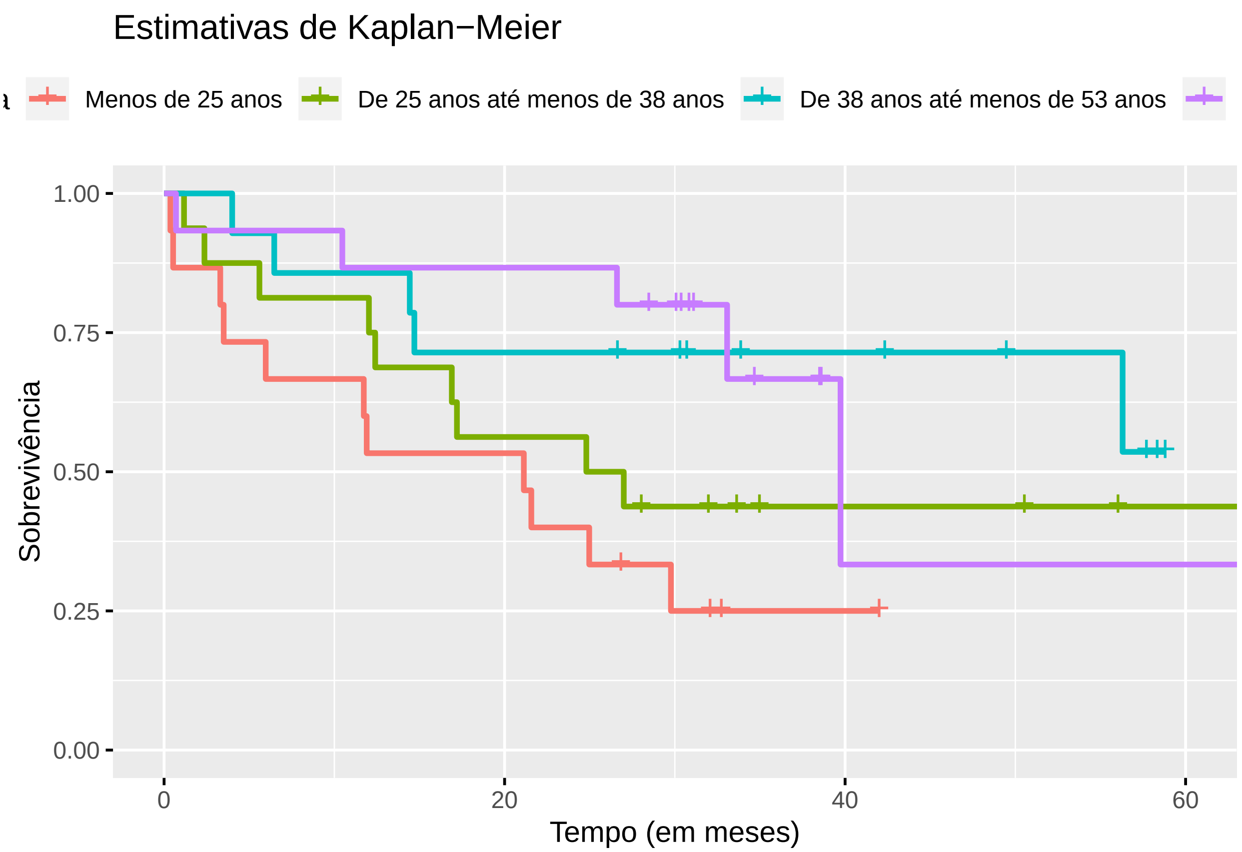
```r
# DIVIDINDO EM GRUPOS ETÁRIOS (EM R)

library(dplyr)

dados_Hodgkins %<>% mutate(GRUPO_ETARIO =
                              case_when(age < 25 ~ "Menos de 25 anos",
                                        age >= 25 & age < 38 ~ "De 25 anos até menos de 38 anos",
                                        age >= 38 & age < 53 ~ "De 38 anos até menos de 53 anos",
                                        age >= 53 ~ "53 anos ou mais"))
```

Agora, vamos construir, no mesmo gráfico, as curvas de Kaplan-Meier para pacientes do sexo masculino e feminino:

```r
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

S_KM_ETARIO <- survfit(Surv(survivaltime, dead) ~ GRUPO_ETARIO, data = dados_Hodgkins)

# Com as bandas de confiança
ggsurvplot(S_KM_ETARIO, data = dados_Hodgkins, conf.int = T, ggtheme = theme_gray(),
           legend.labs = c("Menos de 25 anos", "De 25 anos até menos de 38 anos",
                           "De 38 anos até menos de 53 anos", "53 anos ou mais")) +
   labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```




```r
# Sem as bandas de confiança
ggsurvplot(S_KM_ETARIO, data = dados_Hodgkins, conf.int = F, ggtheme = theme_gray(),
           legend.labs = c("Menos de 25 anos", "De 25 anos até menos de 38 anos",
                           "De 38 anos até menos de 53 anos", "53 anos ou mais")) +
  labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```



Além disso, vamos testar a igualdade das duas curvas de sobrevivência, com a estatística *log-rank* e a de Tarone-Ware:

```
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM_ETARIO, method = c("1"))[,1:3] # log-rank
```

```
##       variable     pval   method
## 1 GRUPO_ETARIO 0.026225 Log-rank
```

```
surv_pvalue(S_KM_ETARIO, method = c("sqrtN"))[,1:3] # Tarone-Ware
```

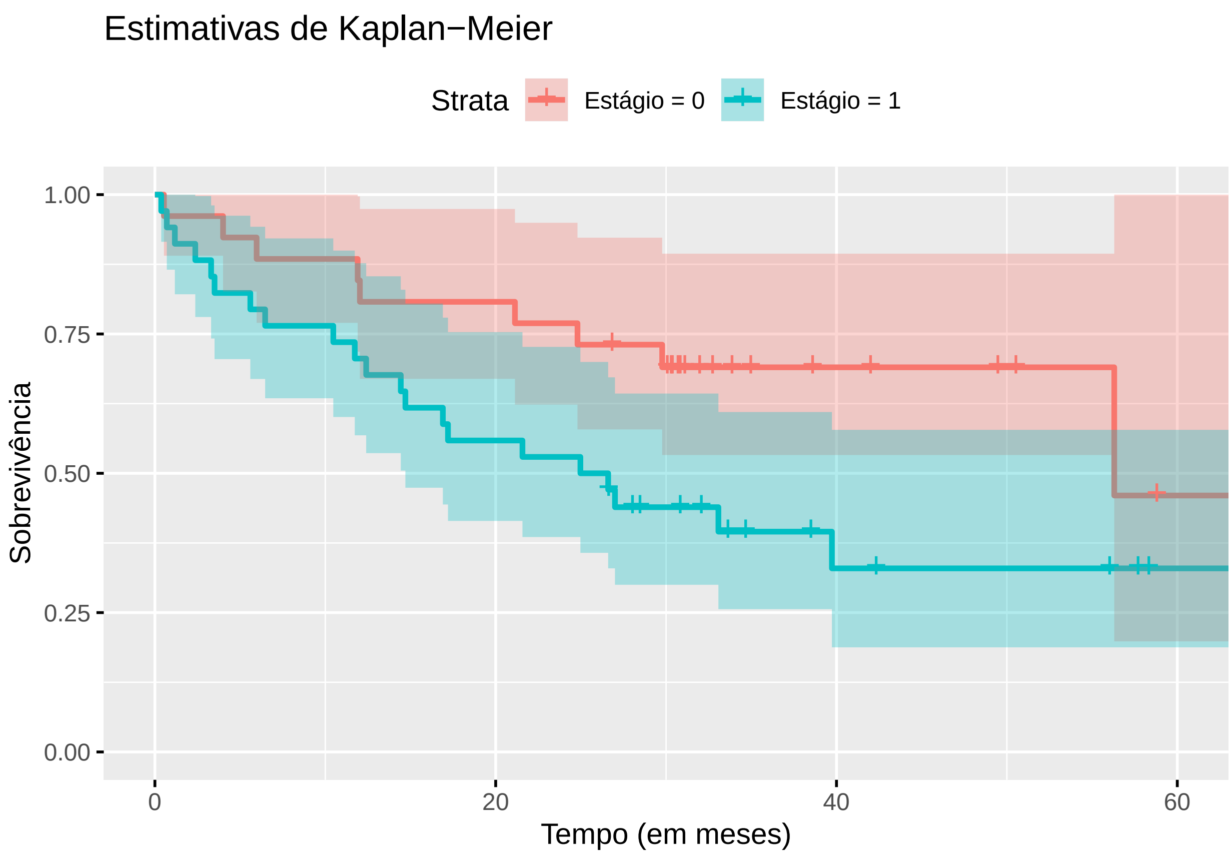
```
##       variable   pval      method
## 1 GRUPO_ETARIO 0.0213 Tarone-Ware
```

Ambos os valores-p estão abaixo de 5%, então, a esse nível de significância, temos evidência para rejeitar a hipótese de que todas as curvas de sobrevivência são iguais.

c)

Vamos repetir os passos do item **a)** para as variáveis "estágio da doença" e "histologia". Primeiro, comecemos pelo "estágio da doença":

```r
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

library(survival)
library(survminer)

S_KM_ESTAGIO <- survfit(Surv(survivaltime, dead) ~ stage, data = dados_Hodgkins)

ggsurvplot(S_KM_ESTAGIO, data = dados_Hodgkins, conf.int = T,
           ggtheme = theme_gray(), legend.labs = c("Estágio = 0", "Estágio = 1")) +
   labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```




```r
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM_ESTAGIO, method = c("1"))[,1:3] # log-rank
```

```
##   variable       pval   method
## 1    stage 0.04181697 Log-rank
```

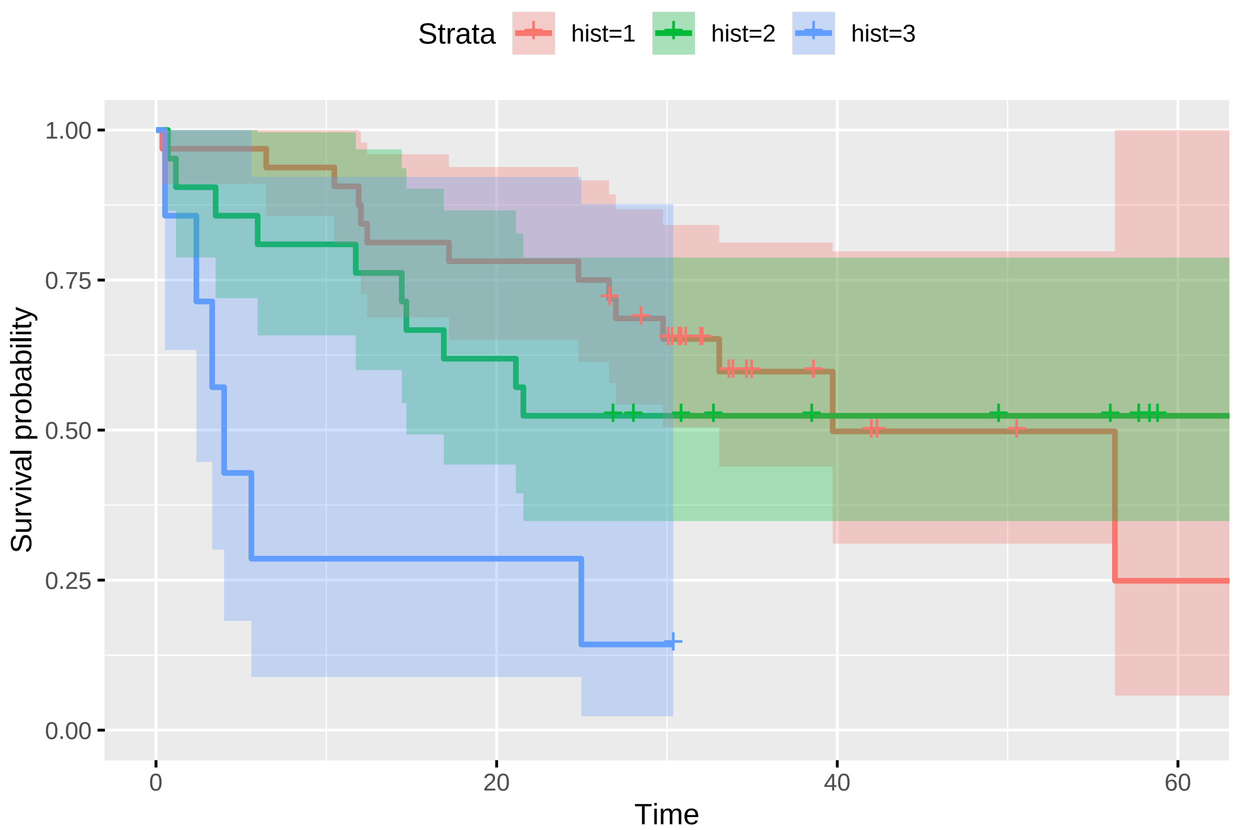
```r
surv_pvalue(S_KM_ESTAGIO, method = c("sqrtN"))[,1:3] # Tarone-Ware
```

```
##   variable   pval      method
## 1    stage 0.0374 Tarone-Ware
```

Ambos os valores-p estão abaixo de 5%, então, a esse nível de significância, temos evidência para rejeitar a hipótese de que as curvas de sobrevivência para os diferentes estágios da doença são iguais.

Agora, vamos analisar dividindo entre a "histologia":

```r
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

library(survival)
library(survminer)

S_KM_HISTOLOGIA <- survfit(Surv(survivaltime, dead) ~ hist, data = dados_Hodgkins)

ggsurvplot(S_KM_HISTOLOGIA, data = dados_Hodgkins, conf.int = T, ggtheme = theme_gray())
```




```r
   labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```

```
## $x
## [1] "Tempo (em meses)"
##
## $y
## [1] "Sobrevivência"
##
## $title
## [1] "Estimativas de Kaplan-Meier"
##
## attr(,"class")
## [1] "labels"
```

```r
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM_HISTOLOGIA, method = c("1"))[,1:3] # log-rank
```

```
##   variable        pval   method
## 1     hist 0.001663778 Log-rank
```

```r
surv_pvalue(S_KM_HISTOLOGIA, method = c("sqrtN"))[,1:3] # Tarone-Ware
```

```
##   variable  pval      method
## 1     hist 9e-04 Tarone-Ware
```

Ambos os valores-p estão (muito) abaixo de 5%, então, a esse nível de significância, temos evidência para rejeitar a hipótese de que as curvas de sobrevivência para os diferentes níveis de histologia são iguais.

Exercício 6

Vamos analisar o arquivo **pharmacoSmoking.csv** que contém os dados de um estudo com 125 pacientes e 14 variáveis. Esse arquivo está disponível na biblioteca *asaur* do R. A descrição dos dados está na documentação e um dos principais objetivos do estudo era comparar o tempo até o fumante voltar a fumar após o início de um dentre dois diferentes tratamentos.

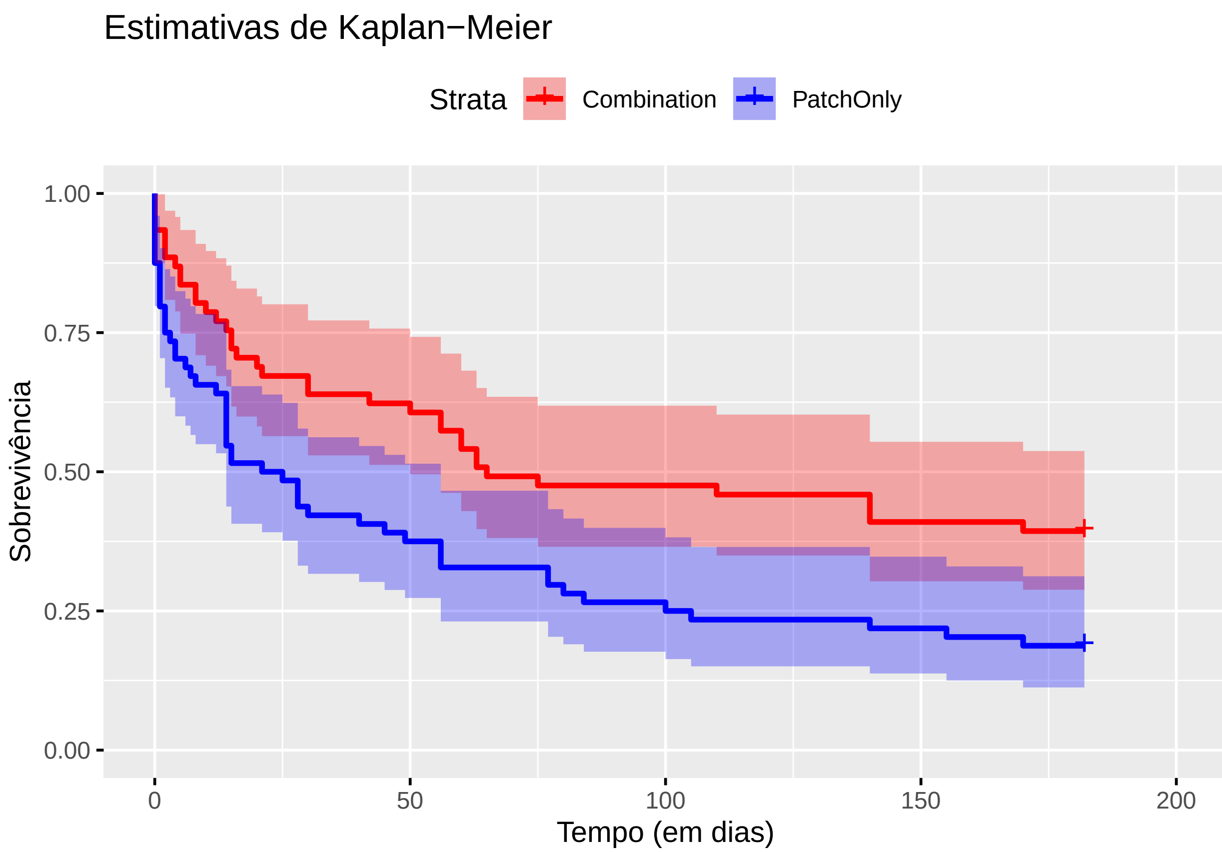
Vamos carregar as informações do pacote no R:

```r
# IMPORTANDO OS DADOS (EM R)

library(asaur)

dados_PS <- pharmacoSmoking
```

a)

Vamos fazer as estimativas de Kaplan-Meier para os dois tratamentos *"Combination"* e *"PatchOnly"*:

```r
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

S_KM <- survfit(Surv(ttr, relapse) ~ grp, data = dados_PS)

ggsurvplot(S_KM, data = dados_PS, palette = c("red", "blue"), conf.int = T,
           ggtheme = theme_gray(), legend.labs = c("Combination", "PatchOnly")) +
   labs(x = "Tempo (em dias)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```



b)

Vamos testar a igualdade entre as curvas de sobrevivência (deve ser interpretado como a igualdade do desempenho dos tratamentos). Além dos testes utilizados no **Exercício 5**, *log-rank* e Tarone-Ware, também utilizaremos o teste de Fleming-Harrington, com $\rho = 1$ e $q = 1$:

```
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM, method = c("1"))[,1:3] # log-rank
```

```
##   variable        pval   method
## 1      grp 0.004606898 Log-rank
```

```
surv_pvalue(S_KM, method = c("sqrtN"))[,1:3] # Tarone-Ware
```

```
##   variable   pval      method
## 1      grp 0.0046 Tarone-Ware
```

```
surv_pvalue(S_KM, method = c("FH_p=1_q=1"))[,1:3] # Fleming-Harrington
```

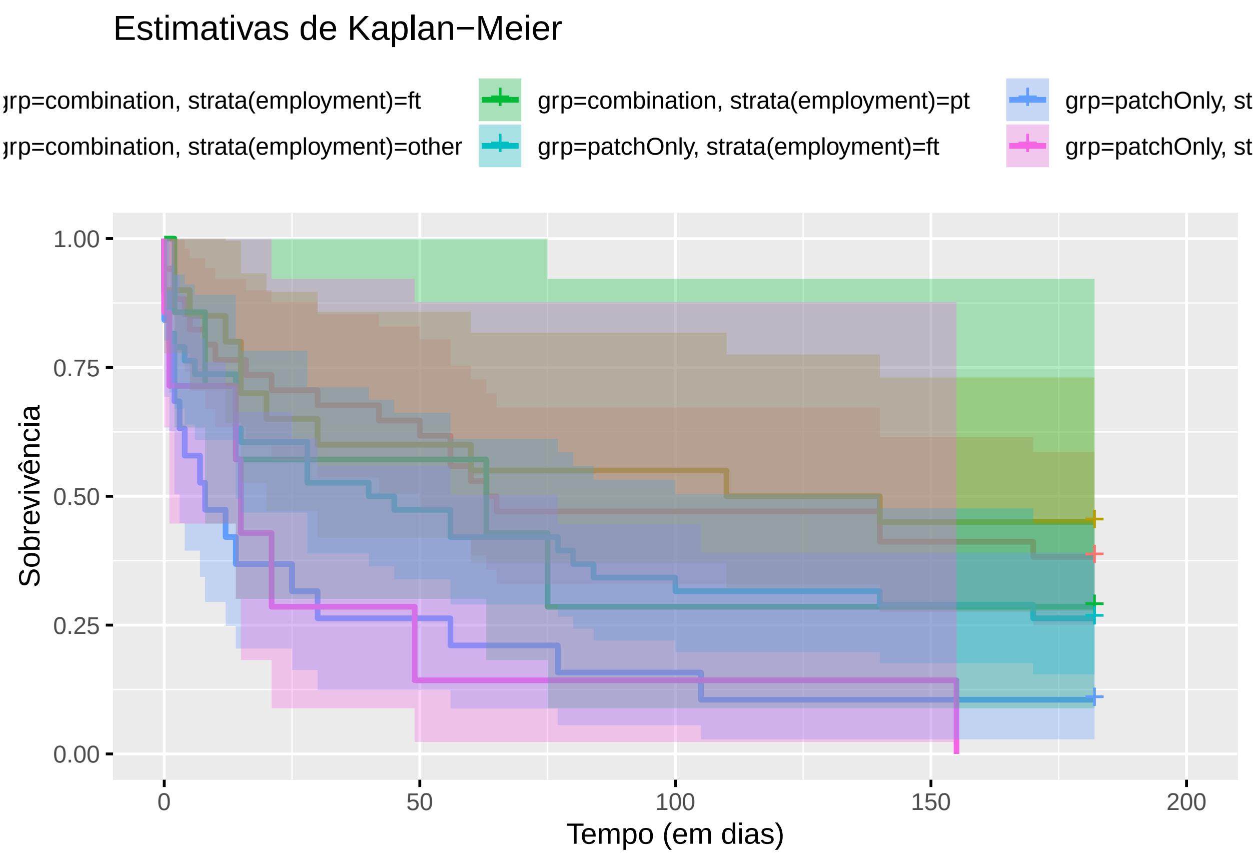
```
##   variable   pval                        method
## 1      grp 0.0206 Fleming-Harrington (p=1, q=1)
```

Ambos os valores-p estão abaixo de 5%, então, a esse nível de significância, temos evidência para rejeitar a hipótese de que as curvas de sobrevivência para os tratamentos são iguais. Assim, é válido afirmar que o tratamento *"Combination"* possui um desempenho melhor.

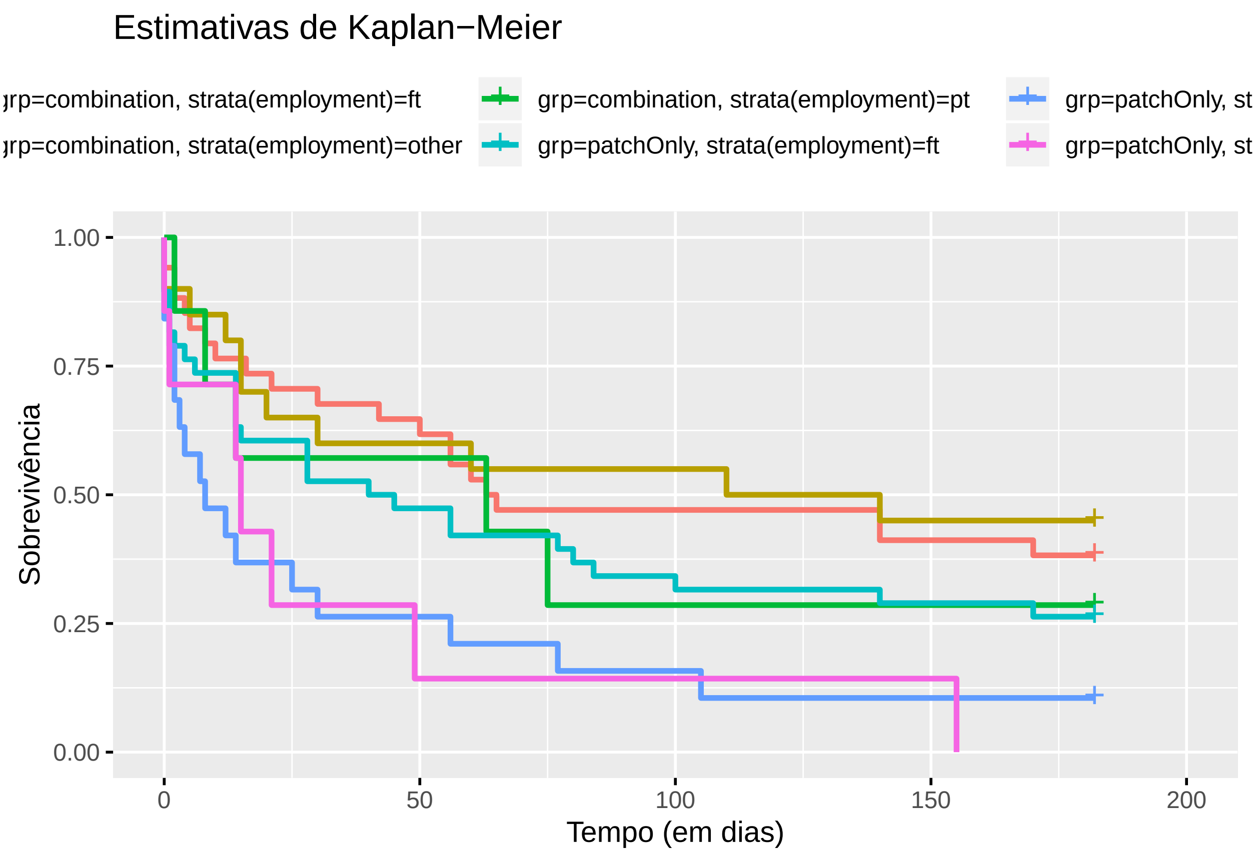
c)

Vamos comparar as curvas de Kaplan-Meier utilizando a estatística de *log-rank*, estratificado por situação de trabalho (variável *employment*). Abaixo, estão as curvas estimadas de Kaplan-Meier e o cálculo do teste para igualdade da sobrevivência entre os estratos:

```r
# FAZENDO AS ESTIMATIVAS DE KM E O GRÁFICO (EM R)

S_KM_EMPLOYMENT <- survfit(Surv(dados_PS$ttr,
                                 dados_PS$relapse) ~ dados_PS$grp + strata(dados_PS$employment),
                           data = dados_PS)

# Com as bandas de confiança
ggsurvplot(S_KM_EMPLOYMENT, data = dados_PS, conf.int = T,
           ggtheme = theme_gray()) +
   labs(x = "Tempo (em dias)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```




```r
# Sem as bandas de confiança
ggsurvplot(S_KM_EMPLOYMENT, data = dados_PS, conf.int = F,
           ggtheme = theme_gray()) +
   labs(x = "Tempo (em dias)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```




```r
# TESTANDO A IGUALDADE DAS CURVAS (EM R)

surv_pvalue(S_KM_EMPLOYMENT, method = c("1"))[,1:3] # log-rank
```

```
##                                       variable        pval   method
## 1 dados_PS$grp+strata(dados_PS$employment) 0.003392996 Log-rank
```

O valor-p está abaixo de 5%, então, a esse nível de significância, temos evidência para rejeitar a hipótese de que as curvas de sobrevivência são iguais.