Lista 2 - MAE0514

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

Exercício 1

Sejam $T_1, T_2, ..., T_n$ tempos de falha sujeitos a censura à direita, de forma que se observa $Z_i = min(T_i, C_i)$ e $\delta_i = I(T_i \leq C_i)$, em que C_i são os tempos de censura, $i = 1, 2, ..., n$. 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

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

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

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

(a) Mostre que o estimador de Kaplan-Meier se reduz à função de sobrevivência empírica se não há censuras, ou seja,

$$\widehat{S}(t) = \frac{n^\circ \ obs. > t}{n}$$

Resolução

Para o primeiro instante:

$$\hat{S}(t_1) = \left(1 - \frac{d_1}{n_1}\right) = \frac{n_1 - d_1}{n_1}$$

Como $n = n_1$ e $n_t - d_t$ é o número de observações até o instante t_1

$$\hat{S}(t_1) = \frac{n^\circ \ obs. > t_1}{n}$$

Para o segundo instante:

$$\hat{S}(t_2) = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) = \frac{n_1 - d_1}{n_1}\frac{n_2 - d_2}{n_2}$$

Como $n = n_1$, $n_2 = n_1 - d_1$ sempre pois não há censura e $n_t - d_t$ é o número de observações até o instante t_2

$$\hat{S}(t_2) = \frac{n_2}{n}\frac{n_2 - d_2}{n_2} = \frac{n_2 - d_2}{n} = \frac{n^\circ \ obs. > t_2}{n}$$

Para o terceiro instante:

$$\hat{S}(t_3) = \Big(1 - \frac{d_1}{n_1}\Big)\Big(1 - \frac{d_2}{n_2}\Big)\Big(1 - \frac{d_3}{n_3}\Big) = \frac{n_1 - d_1}{n_1}\frac{n_2 - d_2}{n_2}\frac{n_3 - d_3}{n_3}$$

Como $n = n_1$, $n_2 = n_1 - d_1$,$n_3 = n_2 - d_2$ sempre pois não há censura e $n_t - d_t$ é o número de observações até o instante t_3

$$\hat{S}(t_3) = \frac{n_2}{n}\frac{n_3}{n_2}\frac{n_3 - d_3}{n_3} = \frac{n_3 - d_3}{n} = \frac{n^\circ\ obs. > t_3}{n}$$

Assim sendo, supondo para um instante t_j com $j = 1...n$, temos:

$$\hat{S}(t_j) = \prod_{t_j \leq t}\Big(1 - \frac{d_j}{n_j}\Big) = \frac{n_1 - d_1}{n_1}...\frac{n_j - d_j}{n_j}$$

Como $n = n_1$, $n_j = n_{j-1} - d_{j-1}$ sempre pois não há censura e $n_t - d_t$ é o número de observações até o instante t_j

$$\hat{S}(t_j) = \frac{n_2}{n}...\frac{n_j - d_j}{n_{j-1}} = \frac{n_j - d_j}{n} = \frac{n^\circ\ obs. > t_j}{n}$$

Logo para um $t > 0$ fixado, temos que o estimador de Kaplan-Meier se reduz à função de sobrevivência empírica, logo:

$$\hat{S}(t) = \frac{n^\circ\ obs. > t}{n}\ \blacksquare$$

(b) Mostre que a fórmula de Greenwood se reduz à estimativa da variância de uma proporção, ou seja,

$$\widehat{Var}(\hat{S}(t)) = n^{-1}\widehat{S}(t)(1 - \widehat{S}(t))$$

Resolução

Para o primeiro instante, utilizando o resultado do exercício anterior, temos:

$$\hat{Var}(\hat{S}(t_1)) = [\hat{S}(t_1)]^2\Big(\frac{d_1}{(n_1 - d_1)n_1}\Big) = \Big(\frac{n_1 - d_1}{n_1}\Big)^2\Big(\frac{d_1}{(n_1 - d_1)n_1}\Big) = \Big(\frac{n_1 - d_1}{n_1}\Big)\Big(\frac{d_1}{n_1}\Big)\Big(\frac{1}{n_1}\Big)$$

Como $n = n_1$, $\hat{S}(t_1) = \frac{n_1 - d_1}{n_1}$ e $1 - \hat{S}(t_1) = 1 - \frac{n_1 - d_1}{n_1} = \frac{d_1}{n_1} \Rightarrow \hat{Var}(\hat{S}(t_1)) = n^{-1}\hat{S}(t_1)(1 - \hat{S}(t_1))$

Para o segundo instante:

$$\hat{Var}(\hat{S}(t_2)) = [\hat{S}(t_2)]^2\Big(\frac{d_1}{(n_1 - d_1)n_1} + \frac{d_2}{(n_2 - d_2)n_2}\Big) = \Big(\frac{n_1 - d_1}{n_1}\Big)^2\Big(\frac{n_2 - d_2}{n_2}\Big)^2\Big(\frac{d_1(n_2 - d_2)n_2 + d_2(n_1 - d_1)n_1}{(n_2 - d_2)n_2(n_1 - d_1)n_1}\Big)$$

Como $n_2 = n_1 - d_1$

$$\Big(\frac{n_1 - d_1}{n_1}\Big)^2\Big(\frac{n_2 - d_2}{n_2}\Big)^2\Big(\frac{d_1(n_2 - d_2)n_2 + d_2(n_1 - d_1)n_1}{(n_2 - d_2)n_2(n_1 - d_1)n_1}\Big) = \Big(\frac{n_1 - d_1}{n_1^2}\Big)\Big(\frac{n_2 - d_2}{n_2^2}\Big)\Big(\frac{d_1(n_2 - d_2)n_2 + d_2n_2n_1}{n_2n_1}\Big) =$$

$$= \Big(\frac{n_1 - d_1}{n_1}\Big)\Big(\frac{n_2 - d_2}{n_2}\Big)\Big(\frac{d_1n_2 - d_1d_2 + d_2n_1}{n_2n_1}\Big)\frac{1}{n_1}$$

Como $n = n_1$ e $\hat{S}(t_2) = \frac{n_1-d_1}{n_1}\frac{n_2-d_2}{n_2}$ e $1-\hat{S}(t_2) = 1 - \frac{n_1-d_1}{n_1}\frac{n_2-d_2}{n_2} = \frac{n_1d_2+n_2d_1-d_1d_2}{n_1n_2}$

Logo,

$$\hat{Var}(\hat{S}(t_2)) = n^{-1}\hat{S}(t_2)(1-\hat{S}(t_2))\hat{Var}(\hat{S}(t_1)) = n^{-1}\hat{S}(t_1)(1-\hat{S}(t_1))$$

Generalizando:

$$\hat{Var}(\hat{S}(t_j)) = [\hat{S}(t_j)]^2\Big(\frac{d_1}{(n_1-d_1)n_1} + ... + \frac{d_j}{(n_j-d_j)n_j}\Big) =$$

$$\Big(\frac{n_1-d_1}{n_1}\Big)^2...\Big(\frac{n_j-d_j}{n_j}\Big)^2\Big(\frac{d_1(n_2-d_2)...(n_j-d_j)n_2...n_j + ... + d_j(n_1-d_1)...(n_{j-1}-d_{j-1})n_1...n_{j-1}}{(n_1-d_1)...(n_j-d_j)n_1...n_j}\Big)$$

Como $n_j = n_{j-1} - d_{j-1}$

$$\Big(\frac{n_1-d_1}{n_1}\Big)^2...\Big(\frac{n_j-d_j}{n_j}\Big)^2\Big(\frac{d_1(n_2-d_2)...(n_j-d_j)n_2...n_j + ... + d_j(n_1-d_1)...(n_{j-1}-d_{j-1})n_1...n_{j-1}}{(n_1-d_1)...(n_j-d_j)n_1...n_j}\Big) =$$

$$= \Big(\frac{n_1-d_1}{n_1^2}\Big)...\Big(\frac{n_j-d_j}{n_j^2}\Big)\Big(\frac{d_1(n_2-d_2)...(n_j-d_j)n_2...n_j + ... + d_jn_2...n_jn_1...n_{j-1}}{n_1...n_j}\Big)$$

$$\Rightarrow \Big(\frac{n_1-d_1}{n_1^2}\Big)...\Big(\frac{n_j-d_j}{n_j^2}\Big)\Big(\frac{d_1(n_2-d_2)...(n_j-d_j)n_2...n_j + ... + d_jn_2...n_jn_1...n_{j-1}}{n_1...n_j}\Big) =$$

$$= \Big(\frac{n_1-d_1}{n_1}\Big)...\Big(\frac{n_j-d_j}{n_j}\Big)\Big(\frac{d_1(n_2-d_2)...(n_j-d_j) + ... + d_jn_1...n_{j-1}}{n_1...n_j}\Big)\frac{1}{n_1}$$

Como $n = n_1$ e $\hat{S}(t_j) = \frac{n_1-d_1}{n_1}...\frac{n_j-d_j}{n_j}$ e $1-\hat{S}(t_j) = 1 - \frac{n_1-d_1}{n_1}...\frac{n_j-d_j}{n_j} = \frac{d_1(n_2-d_2)...(n_j-d_j)+...+d_jn_1...n_{j-1}}{n_1...n_j}$

Logo

$$\hat{Var}(\hat{S}(t_j)) = n^{-1}\hat{S}(t_j)(1-\hat{S}(t_j))$$

Ou seja, para um instante $t > 0$ fixado, temos que a variância do estimador de Kaplan-Meier obtida pela fórmula de Greenwood se reduz à estimativa da variância de uma proporção, logo:

$$\hat{Var}(\hat{S}(t)) = n^{-1}\hat{S}(t)(1-\hat{S}(t))\blacksquare$$

Exercício 2

Considere 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 e a idade que tinha quando faleceu. Os dados estão disponíveis no arquivo **Lista2-HIV.txt** e a descrição dos dados está em **Lista2-HIV.des.**

(a) Com base nos dados apresentados, calcule a variável tempo: número de meses entre a entrada no estudo e óbito (ou tempo de censura).

Resolução

Para a criação da variável tempo segue o código em R:

```
# Carregando dados
dados_HIV <- read.table('Lista2_HIV.txt',header = F)

# Atribuindo nomes as variáveis
names(dados_HIV) <- c("Paciente", "Idade_entrada", "Idade_morte_censura","Falha")

# Criando variável tempo
dados_HIV$Tempo <- (dados_HIV$Idade_morte_cen - dados_HIV$Idade_entr)
```

(b) Calcule 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

Apresente o estimador de duas formas: através de tabelas e gráfico. O que pode ser dito?

Resolução

O estimador tábua de vida é dado por:

$$\hat{S}(t) = \begin{cases} 1, \ t \in [\xi_0, \xi_1) \\ \prod_{l=1}^{j} \left(1 - \frac{d_l}{n_l^* - 1/2w_l}\right), \ j = 2, .., K+1, \ t \in [\xi_{j-1}, \xi_j) \end{cases}$$

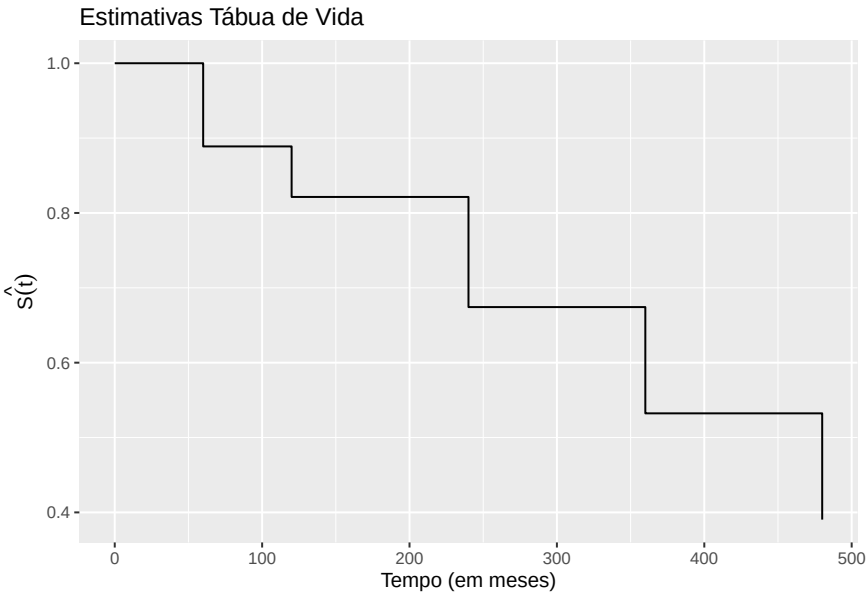
Fixando intervalos $\xi_1, ..\xi_K$, em cada intervalo $[\xi_{j-1}, \xi_j)$, d_j é o número de eventos, w_j o número de observações censuradas e n_j o número de observação em risco em ξ_{j-1}

Estimativas em forma de tabela

Table 1: Tabela com estimativas Tábua de Vida

	nº individuos	nº cens	nº individuos em risco	nº de eventos	sobrevivência	desv.pad sobrevivência
0-60	45	0	45.0	5	1.000	0.000
60-120	40	1	39.5	3	0.889	0.047
120-240	36	5	33.5	6	0.821	0.057
240-360	25	12	19.0	4	0.674	0.072
360-480	9	3	7.5	2	0.532	0.085
480-600	4	4	2.0	0	0.390	0.106

Estimativas em forma de gráfico:



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

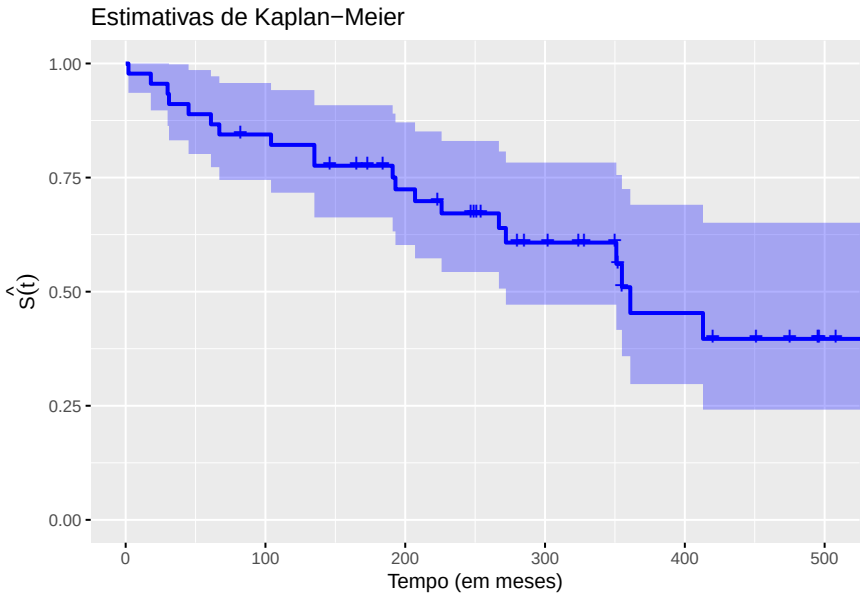
Resolução

Utilizando o R como software, a tabela com as estimativas de Kaplan-Meier:

Tempo	nº em risco	nº de eventos	censura	sobreviv.	desv.pad sobrev.	IC(95%) sup.	IC(95%) inf.
2	45	1	0	0.9777778	0.0224733	1.0000000	0.9356444
18	44	1	0	0.9555556	0.0321495	1.0000000	0.8972020
30	43	1	0	0.9333333	0.0398410	1.0000000	0.8632252
31	42	1	0	0.9111111	0.0465620	0.9981711	0.8316444
45	41	1	0	0.8888889	0.0527046	0.9856205	0.8016508
61	40	1	0	0.8666667	0.0584705	0.9719016	0.7728263
67	39	1	0	0.8444444	0.0639810	0.9572642	0.7449213
82	38	0	1	0.8444444	0.0639810	0.9572642	0.7449213
104	37	1	0	0.8216216	0.0696011	0.9417084	0.7168483
135	36	2	0	0.7759760	0.0804879	0.9085728	0.6627303
146	34	0	1	0.7759760	0.0804879	0.9085728	0.6627303
165	33	0	1	0.7759760	0.0804879	0.9085728	0.6627303
173	32	0	1	0.7759760	0.0804879	0.9085728	0.6627303
184	31	0	1	0.7759760	0.0804879	0.9085728	0.6627303
191	30	1	0	0.7501101	0.0873369	0.8901564	0.6320970
193	29	1	0	0.7242442	0.0941236	0.8709701	0.6022362
207	28	1	0	0.6983784	0.1009059	0.8511029	0.5730592
223	27	0	1	0.6983784	0.1009059	0.8511029	0.5730592
226	26	1	0	0.6715177	0.1082611	0.8302511	0.5431320
247	25	0	1	0.6715177	0.1082611	0.8302511	0.5431320
249	24	0	1	0.6715177	0.1082611	0.8302511	0.5431320
251	23	0	1	0.6715177	0.1082611	0.8302511	0.5431320
254	22	0	1	0.6715177	0.1082611	0.8302511	0.5431320

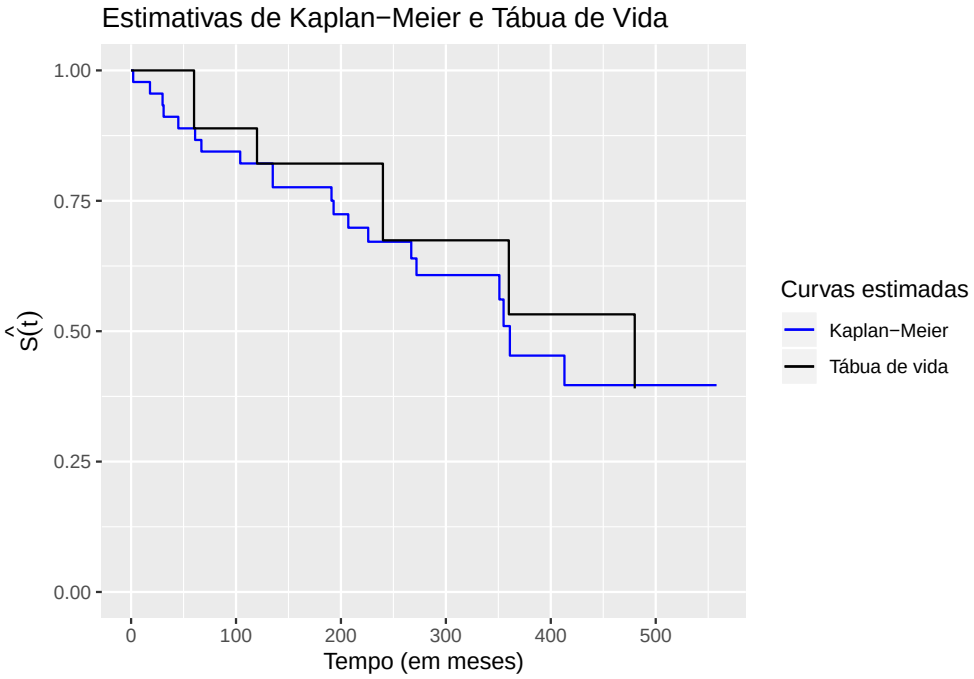
Tempo	nº em risco	nº de eventos	censura	sobreviv.	desv.pad sobrev.	IC(95%) sup.	IC(95%) inf.
267	21	1	0	0.6395406	0.1187494	0.8071381	0.5067438
272	20	1	0	0.6075636	0.1293561	0.7828884	0.4715021
280	19	0	1	0.6075636	0.1293561	0.7828884	0.4715021
285	18	0	1	0.6075636	0.1293561	0.7828884	0.4715021
302	17	0	1	0.6075636	0.1293561	0.7828884	0.4715021
324	16	0	1	0.6075636	0.1293561	0.7828884	0.4715021
328	15	0	1	0.6075636	0.1293561	0.7828884	0.4715021
350	14	0	1	0.6075636	0.1293561	0.7828884	0.4715021
351	13	1	0	0.5608279	0.1521291	0.7556526	0.4162336
352	12	0	1	0.5608279	0.1521291	0.7556526	0.4162336
355	11	1	1	0.5098436	0.1795388	0.7248708	0.3586025
361	9	1	0	0.4531943	0.2147628	0.6903842	0.2974939
413	8	1	0	0.3965450	0.2529431	0.6510255	0.2415388
420	7	0	1	0.3965450	0.2529431	0.6510255	0.2415388
451	6	0	1	0.3965450	0.2529431	0.6510255	0.2415388
475	5	0	1	0.3965450	0.2529431	0.6510255	0.2415388
495	4	0	1	0.3965450	0.2529431	0.6510255	0.2415388
496	3	0	1	0.3965450	0.2529431	0.6510255	0.2415388
508	2	0	1	0.3965450	0.2529431	0.6510255	0.2415388
558	1	0	1	0.3965450	0.2529431	0.6510255	0.2415388

O estimador Kaplan-Meier com intervalo de confiança de 95% segue a curva abaixo



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

Resolução



Ao obter o gráfico acima, nota-se que a curva de Tábua de vida não possui tantos decaimentos como a curva de Kaplan-Meier. A curva de Tábua de vida está sobrestimando o tempo estudado.

Exercício 3

Os dados deste exercício referem-se a um estudo em pacientes com leucemia. Os dados são referentes a 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 apresentados na tabela abaixo. Os tempos censurados à direita são denotados por um sinal "+".

Tratam.	Tempo de remissão								
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^+		
2	8	10	10	12	14	20	48	70	75
	99	103	162	169	195	220	161^+	199^+	217^+
	245^+								

Usando esses dados, calcule à mão o seguinte:

(a) O estimador Kaplan-Meier para cada tratamento.

(b) Estimativas pontuais e intervalares (use coeficiente de 90%) para a mediana de cada tratamento.

(c) O tempo médio de sobrevivência para cada tratamento.

(d) Faça (à mão) um gráfico com as funções de sobrevivência estimadas (para cada tratamento).

(e) O estimador de Nelson-Aalen. Faça um gráfico com as duas curvas estimadas.

(f) Utilizando a estatística de log-rank, teste a igualdade dos tratamentos. Apresente os cálculos realizados para a obtenção da estatística em forma de uma tabela (você pode utilizar alguma planilha eletrônica para os cálculos, mas todos os passos devem estar bem explicados).

Resolução dos exercícios a,b,c,d,e,f

Em ANEXO

Exercício 4

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 façam 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

Resolução

Da definição do estimador de Kaplan-Meier, temos:

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

Assim é construida a tabela:

```
# dados exercício 4
df_ex4 <- data.frame(tempos=c(0.19,0.78,0.96,1.31,2.78,3.16,4.67,4.85,6.5,7.35,
                              8.27,12.07,32.52,33.91,36.71),
                     falhas=rep(1,15))
```

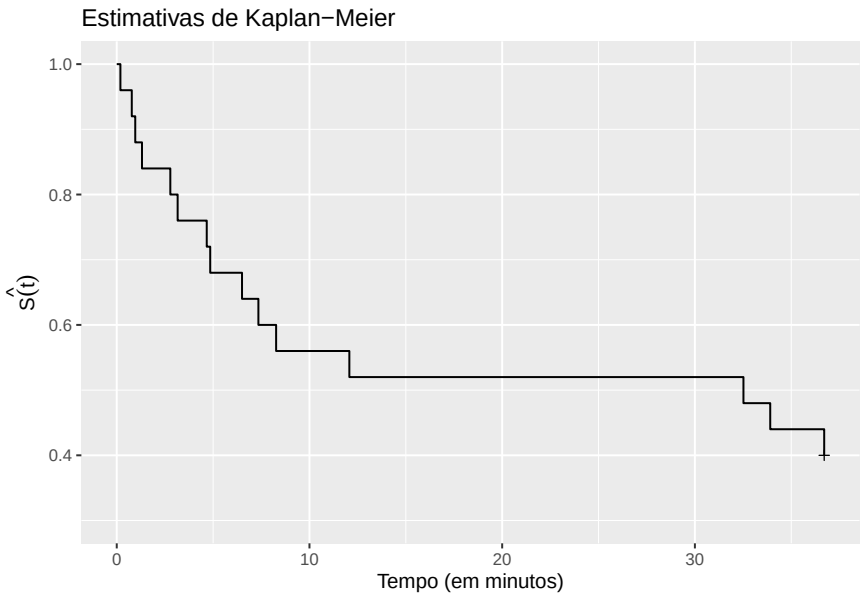
```r
# função que estima a função de sobrevivência via KM
S_km <- function(tempos,falhas,n){

  k <- length(tempos)
  km_matrix <- matrix(0,k+1,4)

  km_matrix[1,] <- c(0,0,n,1)

  km_matrix[1:k+1,1] <- tempos
  km_matrix[1:k+1,2] <- falhas

  for (i in 3:k){
    km_matrix[2,3] <- km_matrix[1,3]
    km_matrix[i,3] <- km_matrix[i-1,3] - km_matrix[i,2]
  }

  km_matrix[k+1,3] <- km_matrix[k-1,3] - km_matrix[k,2]

  for(j in 1:k+1){
    km_matrix[j,4] <- round((1-km_matrix[j,2]/km_matrix[j,3])*km_matrix[j-1,4],2)
  }

  df <- data.frame(km_matrix)
  colnames(df) <- c(expression(t[j]),expression(d[j]),expression(n[j]),"Surv")
  return(df)
}

S <- S_km(df_ex4[,1],df_ex4[,2],25)

knitr::kable(S_km(df_ex4[,1],df_ex4[,2],25))
```

t[j]	d[j]	n[j]	Surv
0.00	0	25	1.00
0.19	1	25	0.96
0.78	1	24	0.92
0.96	1	23	0.88
1.31	1	22	0.84
2.78	1	21	0.80
3.16	1	20	0.76
4.67	1	19	0.72
4.85	1	18	0.68
6.50	1	17	0.64
7.35	1	16	0.60
8.27	1	15	0.56
12.07	1	14	0.52
32.52	1	13	0.48
33.91	1	12	0.44
36.71	1	12	0.40

Logo,

$$\hat{S}(t) = \begin{cases} 1, \ se \ t < 0.19 \\ 0.96, \ se \ 0.19 \leq t < 0.78 \\ 0.92, \ se \ 0.78 \leq t < 0.96 \\ 0.88, \ se \ 0.96 \leq t < 1.31 \\ 0.84, \ se \ 1.31 \leq t < 2.78 \\ 0.80, \ se \ 2.78 \leq t < 3.16 \\ 0.76, \ se \ 3.16 \leq t < 4.67 \\ 0.72, \ se \ 4.67 \leq t < 4.85 \\ 0.68, \ se \ 4.85 \leq t < 6.50 \\ 0.64, \ se \ 6.50 \leq t < 7.35 \\ 0.60, \ se \ 7.35 \leq t < 8.27 \\ 0.56, \ se \ 8.27 \leq t < 12.07 \\ 0.52, \ se \ 12.07 \leq t < 32.52 \\ 0.48, \ se \ 32.52 \leq t < 33.91 \\ 0.44, \ se \ 33.91 \leq t < 36.71 \\ 0.40, \ se \ t \geq 36.71 \end{cases}$$

E a curva estimada:



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

Resolução

Observando o gráfico da estimativa de Kaplan-Meier, podemos estimar a mediana através de uma interpolação lienar simples:

$$\frac{32.52 - 12.07}{0.48 - 0.52} = \frac{\widehat{MED} - 12.07}{0.5 - 0.52} \Rightarrow -\frac{20.45}{0.04} = -\frac{\widehat{MED} + 12.07}{0.02} \Rightarrow \widehat{MED} = \frac{20.45 \cdot 0.02}{0.04} + 12.07 = 22.295$$

Assim, o tempo mediano de vida deste tipo de isolante elétrico funcionando a essa tensão é de 22,3 minutos aproximadamente.

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

Resolução

Queremos estimar $\hat{S}(2)$, utilizando o item anterior, fazendo uma interpolação linear, dado que no livro texto diz-se ser a estimativa mais apropriada, logo:

$$\frac{2.78-1.31}{0.80-0.84}=\frac{2-1.31}{\hat{S}(2)-0.84}\Rightarrow-\frac{1.47}{0.04}=-\frac{0.69}{\hat{S}(2)-0.84}\Rightarrow\hat{S}(2)=-\frac{0.04\cdot0.69}{1.47}+0.84=0.821$$

Agora, calulando a variância de $\hat{S}(2)$ pela fórmula de Greenwood e utilizando a tabela do item a, temos:

$$\widehat{Var}(\hat{S}(2))=[\hat{S}(2)]^2\sum_{j:t_j<2}\frac{d_j}{n_j(n_j-d_j)}=(0.821)^2*\Big[\frac{1}{25(25-1)}+\frac{1}{24(24-1)}+\frac{1}{23(23-1)}+\frac{1}{22(22-1)}\Big]=0.00691$$

Por fim, calculando o intervalo de confiança com nível de confiança de 95%, temos:

$$\hat{S}(2)\pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{S}(2))}$$

Em que $z_{\alpha/2}$ é o percentil 1.96 da distribuição $N(0,1)$, logo:

$$\hat{S}(2)\pm1.96\cdot\sqrt{\widehat{Var}(\hat{S}(2))}\Rightarrow IC(S(2),95\%)=0.821\pm1.96\cdot0.083147$$

Assim:

$$IC(S(2),95\%)=0.821\pm0.163$$

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

Resolução

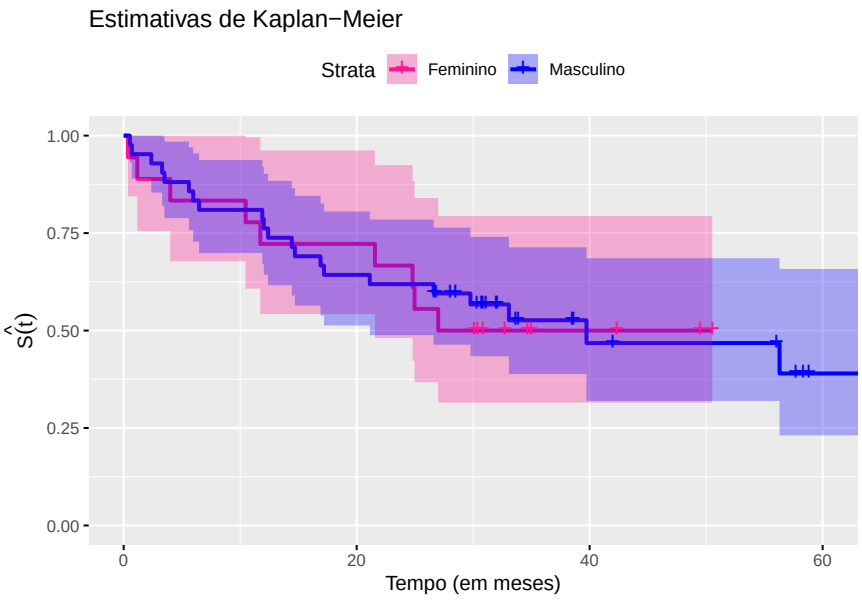
Assim como nos itens anteriores, basta fazer uma interpolação linear, porém da tabela do item a, podemos concluir que o tempo necessário para 20% dos isolantes estarem fora de operação funcionando a essa tensão é de 2,78 minutos aproximadamente.

Exercício 5

Os dados disponíveis no arquivo **Lista2-hodgkins.xlsx** são referentes 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. Em todos os itens a seguir, apresente os resultados em forma de relatório (você pode utilizar o software de sua preferência), explicando e interpretando os resultados. Acrescente o código do programa utilizado no final, como um apêndice.

(a) Construa, no mesmo gráfico, as curvas de Kaplan-Meier para pacientes do sexo masculino e feminino. Teste a igualdade das curvas de sobrevivência (obtenha duas estatísticas de teste, sendo uma delas a de log-rank). Comente.

Resolução



Queremos testar a igualdade das curvas, assim:

$$\begin{cases} H_0 : S_1(t) = S_2(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : S_1(t) \neq S_2(t) \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Em que τ é o maior instante observado tal que os dois grupos possuem pelo menos um indivíduo em risco.

Sob a hipótese nula, a estatística do teste Log-Rank é:

$$L_r = \frac{[\sum_{j=1}^{L}(d_{2j} - e_{2j})]^2}{\sum_{j=1}^{L} V_j^2}$$

Em que d_{2j} é o # de indivíduos observados no grupo 2, e_{2j} é o # de indivíduos esperadados no grupo 2 e V_j é a variância de d_{2j} que é dada por:

$$V_j = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_{1_j} - 1)}$$

Em que n_{1j} e n_{2j} são o número de indivíduos nos grupo 1 e 2 respectivamente. Assim sendo, sob a hipótese nula,

$$L_r \overset{a}{\sim} \chi^2_{(1)}$$

Utilizando o teste log-rank, temos:

variable	pval	method
sex	0.8618044	Log-rank

Utilizando o teste Tarone-Ware, cuja a estatística de teste é:

$$T_w = \frac{[\sum_{j=1}^{L} \sqrt{n_j}(d_{2j} - e_{2j})]^2}{\sum_{j=1}^{L} n_j V_j^2}$$

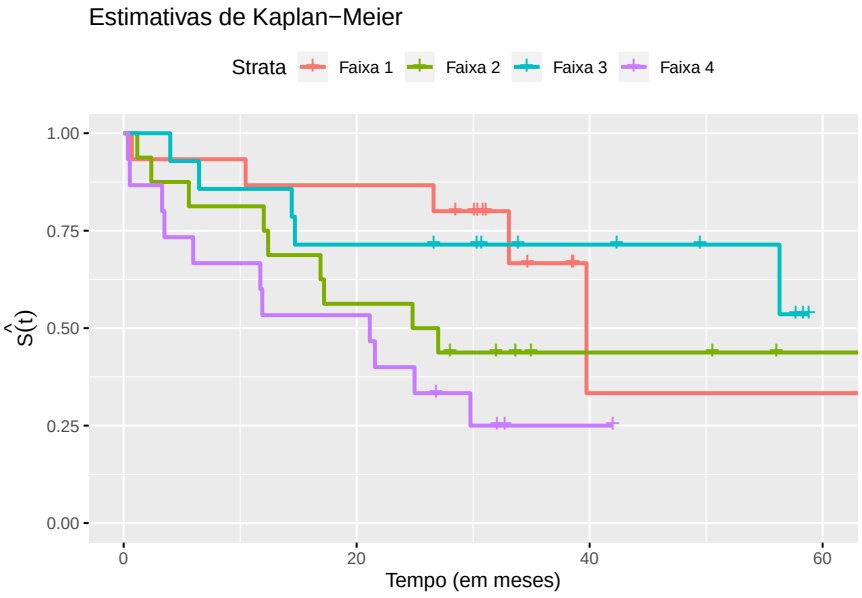
obtemos o seguinte resultado:

variable	pval	method
sex	0.8124	Tarone-Ware

A partir do gráfico obtido, observa-se que as curvas aparentemente são próximas entre si e a banda de confiança de um tratamento engloba o outro tratamento, logo aparentemente não possui efeito de tratamento de sexo no tempo da doença de Hodgkings. Isso se confirma com os testes, com *p-values* próximos de 80%, não rejeita-se as hipóteses nulas, a um nível de significância de 5%.

(b) Divida os pacientes em quatro grupos etários: menos de 25 anos; de 25 anos (inclusive) até menos de 38 anos; de 38 anos (inclusive) até 53 anos; 53 anos ou mais. Obtenha as curvas de Kaplan-Meier e teste a igualdade das curvas de sobrevivência (obtenha duas estatísticas de teste, sendo uma delas a de log-rank). Comente.

Resolução



Sob a hipótese:

$$\begin{cases} H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : pelo\ menos\ uma\ funcao\ diferente\ para\ algum\ t \in \ [0, \tau] \end{cases}$$

Sendo a estatística de teste é:

$$L_r = v' V^{-1} v$$

Em que V é a matriz de variância-covariância $(r_1 x r_1)$ da distribuição hipergeométrica associada a tabela definida pelo tempo de falha t_j e v' e v são os vetores da diferença entre o esperado e o observado em cada grupo, sendo r grupos.

Assim, sob a hipótese nula,

$$L_r \overset{a}{\sim} \chi^2_{(r-1)}$$

Utilizando o teste log-rank, temos:

variable	pval	method
Faixa_idade	0.026225	Log-rank

Utilizando o teste Tarone-Ware:

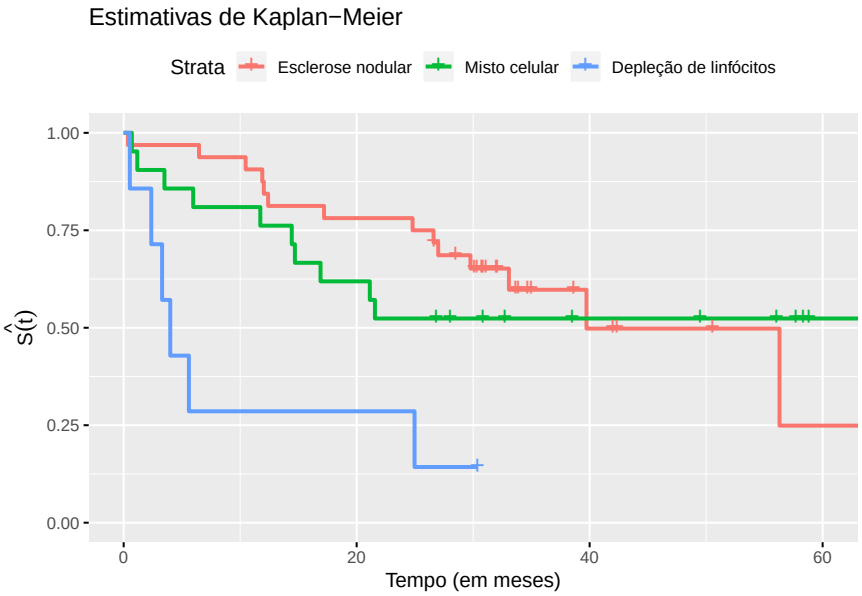
variable	pval	method
Faixa_idade	0.0213	Tarone-Ware

Pelo gráfico, nota-se que a faixa 4 possui um menor tempo de sobrevivência, em seguida pela faixa 2. A faixa 1, inicialmente aparenta ter um melhor cenário, entretanto após mais de 30 meses, nota-se uma piora. Ao realizar os testes, rejeita-se as hipóteses nulas, logo a indícios que as faixas possuem tempos de sobrevivência diferentes entre si, a um nível de significância de 5%.

(c) Repita o item (a) para as variáveis estágio da doença e histologia. Comente os resultados.

Resolução

Fazendo o cálculo para Histologia, temos:



Sob a hipótese:

$$\begin{cases} H_0 : S_1(t) = S_2(t) = S_3(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : pelo \ menos \ uma \ funcao \ diferente \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Sendo a estatística de teste é:

$$L_r = v' V^{-1} v$$

Utilizando o teste log-rank, temos:

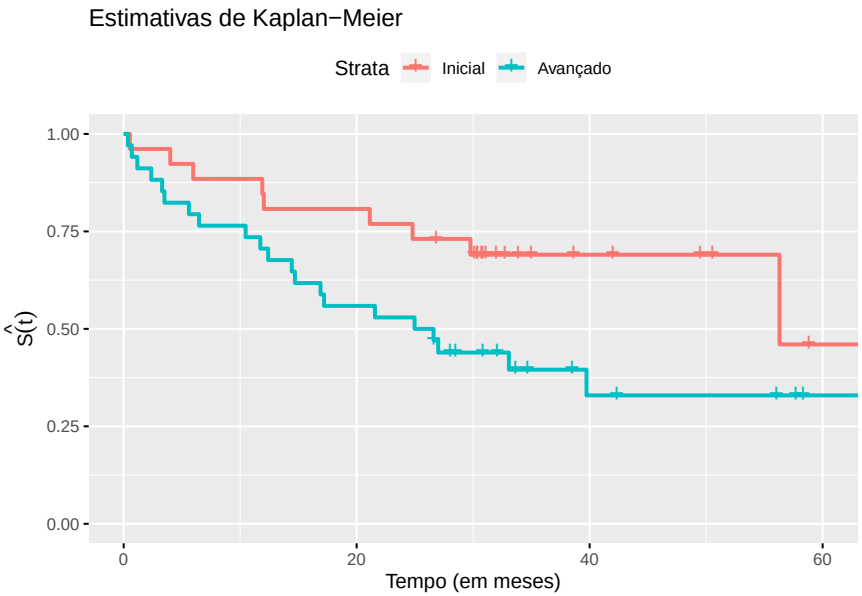
variable	pval	method
hist	0.0016638	Log-rank

Utilizando o teste Tarone-Ware:

variable	pval	method
hist	9e-04	Tarone-Ware

No gráfico, nota-se que a histologia depleção de linfócitos possui aparentemente um tempo de sobrevivência menor que as demais histologias. A esclerose nodular apresenta um melhor cenário até 40 meses de estudo, quando é substituída pelo misto celular, entretanto essa interpretação pode estar equivocada devido o grande número de censuras no final do estudo para a histologia misto celular. Ao realizar os testes, a um nível de significância de 5%, rejeita-se as hipóteses nulas, logo há indícios que os tempos de sobrevivência por histologias são diferentes.

Fazendo o cálculo para estágio da doença, temos:



$$\begin{cases} H_0 : S_1(t) = S_2(t), \ \forall\, t \in \ [0, \tau] \\ H_1 : S_1(t) \neq S_2(t) \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Utilizando o teste log-rank, temos:

variable	pval	method
stage	0.041817	Log-rank

Utilizando o teste Tarone-Ware:

variable	pval	method
stage	0.0374	Tarone-Ware

Pelo gráfico de estágio da doença, é possível observar que o estágio inicial aparenta possuir um tempo de sobrevivência superior ao estágio avançado da doença. Essa primeira interpretação é confirmada com os testes, que a um nível de significância de 5%, rejeita-se as hipóteses nulas, logo há indícios que os tempos de sobrevivência pelo estágio da doença são diferentes.

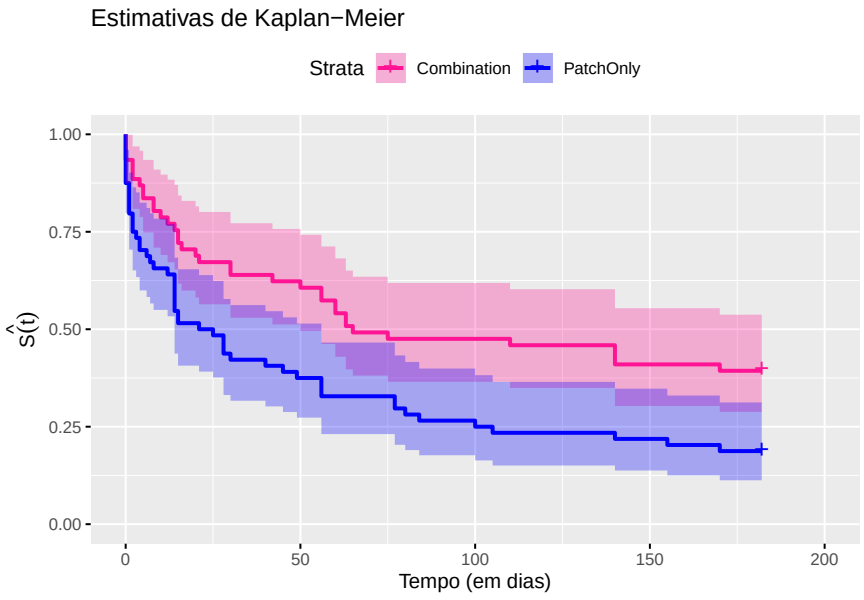
Exercício 6

Considere os dados do arquivo **pharmacoSmoking.csv** com 125 pacientes e 14 variáveis. Esse arquivo está disponível na biblioteca **asaur** do **R**, cuja documentação está disponibilizada. 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. Utilizando o *software* de sua preferência:

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

Resolução

A curva de Kaplan-Meier para os tratamentos:



Observa-se que, aparentemente, o tratamento “Combination” tem um tempo de sobrevivência maior que o tratamento “PatchOnly”

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

Resolução

Sob a hipótese de:

$$\begin{cases} H_0 : S_1(t) = S_2(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : S_1(t) \neq S_2(t) \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Utilizando o teste log-rank, temos:

variable	pval	method
grp	0.0046069	Log-rank

Utilizando o teste Tarone-Ware

variable	pval	method
grp	0.0046	Tarone-Ware

Utilizando o teste Familia Fleming-Harrington

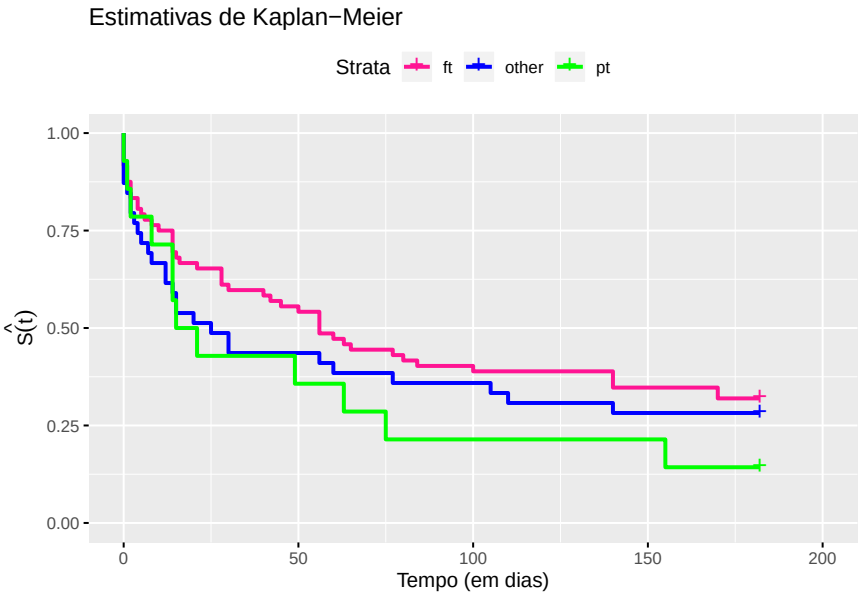
variable	pval	method
grp	0.0206	Fleming-Harrington (p=1, q=1)

A partir dos testes, rejeita-se a hipótese nula que o tempo de sobrevivência dos grupos são identicos.

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

Observação: Consulte o livro do Klein e Moeschberger, página 219, para mais detalhes sobre testes estratificados.

Resolução



Sob a hipótese:

$$\begin{cases} H_0 : S_1(t) = S_2(t) = S_3(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : pelo\ menos\ uma\ funcao\ diferente\ para\ algum\ t \in \ [0, \tau] \end{cases}$$

Utilizando o teste log-rank, temos:

variable	pval	method
employment	0.3271981	Log-rank

Nota-se pelo gráfico que os tempos de sobrevivência para as situações de trabalho são parecidas inicialmente e no fim do estudo nota-se que o tratamento ft tem um maior tempo de sobrevivência aparentemente, em seguida os tratamentos other e pt, respectivamente. Ao observarmos o gráfico com a banda de confiança, observa-se que os tratamentos estão todos englobados nas bandas de confiança dos outros tratamentos, seguindo a mesma conclusão do teste Log-rank, do qual foi rejeitado a um nível de significância de 5%, logo há indícios que os tratamentos possuem tempos de sobrevivência identicos.

Apêndice

```
# set
setwd("~/Área de Trabalho/Lista 2")
```

```r
# Lista 2 - MAE 514

library(ggplot2)
library(asaur)
library(survival)
library(survminer)
library(sqldf)
library(KMsurv)

# Exercício 2

# item a

# Carregando dados
dados_HIV <- read.table('Lista2_HIV.txt',header = F)

# Atribuindo nomes as variáveis
names(dados_HIV) <- c("Paciente", "Idade_entrada", "Idade_morte_censura","Falha")

# Criando variável tempo
dados_HIV$Tempo <- (dados_HIV$Idade_morte_cen - dados_HIV$Idade_entr)

# item b
dados_HIV$Faixa <- sapply(dados_HIV$Tempo,
                          function(x){
                            if (x <= 60) x = 'Faixa_1'
                            else if (x > 60 & x <= 120)  x = 'Faixa_2'
                            else if (x > 120 & x <= 240) x = 'Faixa_3'
                            else if (x > 240 & x <= 360) x = 'Faixa_4'
                            else if (x > 360 & x <= 480) x = 'Faixa_5'
                            else x = 'Faixa_6'
                          })

cont_HIV <- sqldf('SELECT Faixa,
                          count(Faixa),
                          sum(Falha),
                          count(Faixa)-sum(Falha)
                          FROM dados_HIV GROUP BY Faixa')

# Dados para contrução do estimador tábua de vida
intervalos <- c(0,60,120,240,360,480,600)
nindiv <- 45
ncens <- c(0,1,5,12,3,4)
neventos <- c(5,3,6,4,2,0)
HIV.tv <- lifetab(intervalos,nindiv,ncens,neventos)

# Tabela com estimativas de tábua de vida
HIV.tv_imp <- round(HIV.tv[,c("nsubs","nlost","nrisk","nevent","surv","se.surv")],3)
names(HIV.tv_imp) <- c("nº individuos","nº cens","nº individuos em risco",
                       "nº de eventos","sobrevivência","desv.pad sobrevivência")
knitr::kable(caption = "Tabela com estimativas Tábua de Vida", HIV.tv_imp)

# Gráfico do estimador tábua de vida
```

```r
dt <- data.frame(x=intervalos[1:6],y= HIV.tv[,5])

ggplot() +
  geom_step(aes(x=x,y=y),data=dt) +
  labs(title = "Estimativas Tábua de Vida",
       y = expression(hat(S(t))),
       x = "Tempo (em meses)")

# item c
ekm_ex2 <- survfit(Surv(dados_HIV$Tempo, dados_HIV$Falha)~1)

# Tabela com estimativas de Kaplan-Meier
knitr::kable(surv_summary(ekm_ex2),col.names = c("Tempo","nº em risco","nº de eventos",
                                                 "censura","sobreviv.","desv.pad sobrev.",
                                                 "IC(95%) sup.","IC(95%) inf."))

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex2, data = dados_HIV,palette = c('blue'),
           ggtheme=theme_gray(), legend = 'none') +
      labs(x="Tempo (em meses)",
           y=expression(hat(S(t))),
           title = "Estimativas de Kaplan-Meier")

# item d

m <- 1 # numero de grupos
df <- data.frame(time = c(rep(0, m), ekm_ex2$time),
                 surv = c(rep(1, m), ekm_ex2$surv))

colors <- c("Tábua de vida" = "Black", "Kaplan-Meier" = "Blue")

ggplot(data = df) +
  geom_step(aes(x = time, y = surv,color = "Kaplan-Meier")) +
  scale_y_continuous(limits = c(0,1)) +
  geom_step(aes(x=x,y=y,color = "Tábua de vida"),data=dt) +
  scale_x_continuous(breaks = seq(0,500,100)) +
  scale_color_manual(values = colors) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier e Tábua de Vida",
       color = "Curvas estimadas")


# Exercício 4

# dados exercício 4
df_ex4 <- data.frame(tempos=c(0.19,0.78,0.96,1.31,2.78,3.16,4.67,4.85,6.5,7.35,
                              8.27,12.07,32.52,33.91,36.71),
                     falhas=rep(1,15))


# função que estima a função de sobrevivência via KM
S_km <- function(tempos,falhas,n){
```

```
  k <- length(tempos)
  km_matrix <- matrix(0,k+1,4)

  km_matrix[1,] <- c(0,0,n,1)

  km_matrix[1:k+1,1] <- tempos
  km_matrix[1:k+1,2] <- falhas

  for (i in 3:k){
    km_matrix[2,3] <- km_matrix[1,3]
    km_matrix[i,3] <- km_matrix[i-1,3] - km_matrix[i,2]
  }

  km_matrix[k+1,3] <- km_matrix[k-1,3] - km_matrix[k,2]

  for(j in 1:k+1){
    km_matrix[j,4] <- round((1-km_matrix[j,2]/km_matrix[j,3])*km_matrix[j-1,4],2)
  }

  df <- data.frame(km_matrix)
  colnames(df) <- c(expression(t[j]),expression(d[j]),expression(n[j]),"Surv")
  return(df)
}

S <- S_km(df_ex4[,1],df_ex4[,2],25)

# Gráfico do estimador tábua de vida
dt_ex4 <- data.frame(x=S[,1],y= S[,4])

ggplot() +
  geom_step(aes(x=x,y=y),data=dt_ex4) +
  ylim(c(.3,1)) +
  geom_point(shape=3,aes(x=36.71,y=0.4)) +
  labs(x="Tempo (em minutos)",
      y=expression(hat(S(t))),
      title = "Estimativas de Kaplan-Meier")

# item b

# (32.52-12.07)/(0.48-0.52)=(med-12.07)/(0.5-0.52)
med <- (32.52-12.07)/(0.48-0.52)*(0.5-0.52) + 12.07
med

# item c

# (2.78-1.31)/(0.80-0.84)=(2-1.31)/(x-0.84)
x <- (0.80-0.84)/(2.78-1.31)*(2-1.31)+0.84
x

# Exercício 5

# Carregando os dados
data_Hodgkins <- readxl::read_excel("Lista2_Hodgkins.xlsx")
```

```r
# item a

ekm_ex5 <- survfit(Surv(survivaltime, dead)~ sex,data = data_Hodgkins)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex5, data = data_Hodgkins, palette = c('deeppink','blue'),conf.int = T,
           ggtheme=theme_gray(), legend.labs = c("Feminino", "Masculino")) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# teste log-rank
surv_pvalue(ekm_ex5, method = c("1"))[,1:3]

# teste Tarone-Ware
surv_pvalue(ekm_ex5, method = c("sqrtN"))[,1:3]

# item b

# criando faixas de idade
data_Hodgkins$Faixa_idade <- sapply(data_Hodgkins$age,
                            function(x){
                              if (x < 25) x = 'Faixa_1'
                              else if (x >= 25 & x < 38)  x = 'Faixa_2'
                              else if (x >= 38 & x < 53) x = 'Faixa_3'
                              else x = 'Faixa_4'
                            })

# Kaplan-Meier por faixa de idade
ekm_ex5_age <- survfit(Surv(survivaltime, dead)~ Faixa_idade,data = data_Hodgkins)

ggsurvplot(ekm_ex5_age, data = data_Hodgkins,conf.int = F,
           ggtheme=theme_gray(), legend.labs = c("Faixa 1", "Faixa 2",
                                                 "Faixa 3", "Faixa 4")) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# teste log-rank

surv_pvalue(ekm_ex5_age, method = c("1"))[,1:3]

# teste Tarone-Ware
surv_pvalue(ekm_ex5_age, method = c("sqrtN"))[,1:3]

# item c

# Para histologia
ekm_ex5_hist <- survfit(Surv(survivaltime, dead)~ hist,data = data_Hodgkins)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex5_hist, data = data_Hodgkins,conf.int = F,
           ggtheme=theme_gray(), legend.labs = c("Esclerose nodular",
```

```
                                                  "Misto celular","Depleção de linfócitos")) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# teste log-rank
surv_pvalue(ekm_ex5_hist, method = c("1"))[,1:3]

# teste Tarone-Ware
surv_pvalue(ekm_ex5_hist, method = c("sqrtN"))[,1:3]

# Para doença
ekm_ex5_est_doe <- survfit(Surv(survivaltime, dead)~ stage,data = data_Hodgkins)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex5_est_doe, data = data_Hodgkins,conf.int = F,
           ggtheme=theme_gray(), legend.labs = c("Inicial","Avançado")) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# teste log-rank
surv_pvalue(ekm_ex5_est_doe, method = c("1"))[,1:3]

# teste Tarone-Ware
surv_pvalue(ekm_ex5_est_doe, method = c("sqrtN"))[,1:3]

# Exercício 6

# Carregando dados
data_PS <- asaur::pharmacoSmoking

# item a
ekm_ex6 <- survfit(Surv(ttr, relapse)~ grp,data = data_PS)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex6, data = data_PS, palette = c('deeppink','blue'),conf.int = T,
           ggtheme=theme_gray(), legend.labs = c("Combination", "PatchOnly")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# item b

# teste log-rank
surv_pvalue(ekm_ex6, method = c("1"))[,1:3]

# teste Tarone-Ware
surv_pvalue(ekm_ex6, method = c("sqrtN"))[,1:3]

# teste Familia Fleming-Harrington
surv_pvalue(ekm_ex6, method = c("FH_p=1_q=1"))[,1:3]
```

```r
# item c

ekm_ex6_2 <- survfit(Surv(ttr, relapse)~ employment,data = data_PS)

ggsurvplot(ekm_ex6_2, data = data_PS, palette = c('deeppink','blue',"green"),conf.int = F,
           ggtheme=theme_gray(), legend.labs = c("ft", "other","pt")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# teste log-rank
surv_pvalue(ekm_ex6_2, method = c("1"))[,1:3]
```