Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

SME0821 - Análise de Sobrevivência e Confiabilidade

Solução para a Atividade II

Aluno: Matheus Victal Cerqueira NUSP:10276661

Exercício 1

No exercício 1, tem-se o seguinte sistema de interesse para análise:

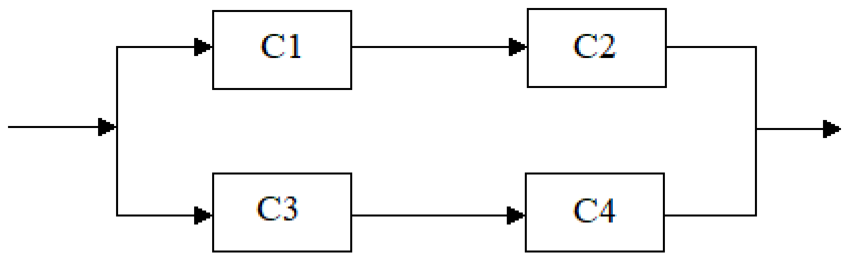

Figura 1: Sistema de interesse

Os componentes de tal sistema funcionam de forma independente e a função taxa de falha para cada um deles é dada por $h(t) = 0,1$.

a) A partir das definições de função de confiabilidade ($R(t)$), função taxa de risco ($h(t)$) e função taxa de falha acumulada ($H(t)$), pode-se obter o seguinte resultado:

$$R(t) = exp\{H(t)\} = exp\left\{-\int_0^t h(u)\,du\right\}$$

Assim sendo, pode-se obter a função de confiabilidade $R(t)$ para as componentes do sistema a aprtir dessa relação.

$$R(t) = exp\left\{-\int_0^t h(u)\,du\right\} = exp\left\{-\int_0^t 0,1\,du\right\}$$
$$\therefore R(t) = exp\left\{-0,1u\Big|_0^t\right\} = e^{-0,1t},\ t \geq 0$$

Conclui-se por fim que a função de confiabilidade para as componentes C_1, C_2, C_3 e C_4 é dada por $R(t) = e^{-0,1t}$. Com essa informação em mãos, pode-se obter a função de confiabilidade do sistema $R_s(t)$.

C_1 e C_2 estão ligados em série, assim como C_3 e C_4. Dessa forma, pode-se obter a função de confiabilidade psra os subsistemas $C_{1,2}$ e $C_{3,4}$, que podem ser visualizados na Figura 2.

Obtendo-se as funções de confiabilidade para $C_{1,2}$ e $C_{3,4}$:

$$R_{1,2}(t) = R_1(t) \cdot R_2(t) = R^2(t) = \left(e^{-0,1t}\right)^2 = e^{-0,2t}$$

De forma análoga, pode-se obter o resultado para $C_{3,4}$

$$R_{3,4}(t) = R_3(t) \cdot R_4(t) = R^2(t) = \left(e^{-0,1t}\right)^2 = e^{-0,2t}$$

Como $C_{1,2}$ e $C_{3,4}$ estão em paralelo, a função confiabilidade do sistema pode ser obtida por:

$$R_s(t) = 1 - [1 - R_{1,2}(t)]\,[1 - R_{3,4}(t)] =$$
$$1 - \left[1 - e^{-0,2t}\right]\left[1 - e^{-0,2t}\right]$$
$$1 - \left[1 - 2e^{-0,2t} + e^{-0,4t}\right]$$
$$\therefore R_s(t) = 2e^{-0,2t} - e^{-0,4t},\ t \geq 0$$

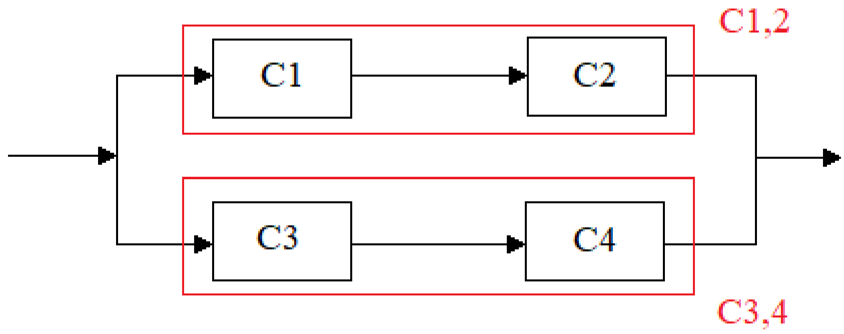

Figura 2: Sistema de interesse simplificado

b) Por desenvolvimento da definição de esperança estatística e de função de confiabilidade, tem-se que:

$$E[T] = \int_0^\infty R(t)dt,$$

Sendo $E[T]$ o valor epseraado para a variável aleatória T: "tempo até a falha", e $R(t)$ a função de confiabilidade de T. Assim:

I.) Para uma componente (T_c: "tempo até a falha de componente"):

$$E[T_c] = \int_0^\infty R(t)dt = \int_0^\infty e^{-0{,}1t}dt = -\frac{e^{-0{,}1t}}{0{,}1}\bigg|_0^\infty = 0 + \frac{1}{0{,}1} = 10$$

II.) Para o sistema (T_s: "tempo até a falha do sistema"):

$$E[T_s] = \int_0^\infty R_s(t)dt = \int_0^\infty 2e^{-0{,}2t} - e^{-0{,}4t}dt = -2\frac{e^{-0{,}2t}}{0{,}2}\bigg|_0^\infty + \frac{e^{-0{,}4t}}{0{,}4}\bigg|_0^\infty =$$
$$-2 \cdot 0 + 2 \cdot \frac{1}{0{,}2} + 0 - \frac{1}{0{,}4} = 10 - 2,5 = 7,5$$

É notório que o tempo médio até a falha do sistema ($E[T_s] = 7,5$) é menor do que o tempo médio até a falha de uma componente ($E[T_c] = 10$). Isso significa que, em média, o sistema é menos confiável do que as componentes que o compoem.

c) A confiabilidade para um determinado período de tempo t a partir de 0 pode ser dada pela função de confiabilidade calculada no ponto t de interesse. No caso, queremos avaliar o comportamento das funções de confiabilidade do sistema $R_s(t)$ e a função de confiabilidade de uma componente $R(t)$ para $t = 2$ anos. Assim sendo:

$$R_s(t = 2) = 2e^{-0{,}2\cdot 2} - e^{-0{,}4\cdot 2} \approx 0,891$$
$$R(t = 2) = e^{-0{,}1t} = e^{-0{,}1\cdot 2} \approx 0,819$$

Assim, a confiabilidade de uma componente para uma componente em 2 anos é $R(t = 2) = P(T_c > 2) = 0,819$, menor do que a conbiabilidade do sistema para o mesmo tempo de estudo $R_s(t = 2) = P(T_s > 2) = 0,891$.

d) Com os resultados obtidos em b) e c), é notório que a comparação a confiabilidade do sistema com a confiabilidade de uma componente depende do valor de t que se está sendo analisado, pois as curvas de confiabilidade $R(t)$ e $R_s(t)$ se interceptam em $t \approx 4,8121$; que é a solução de:

$$R(t) = R_s(t) \Rightarrow e^{-0,1x} = 2e^{-0,2x} - e^{-0,4x}$$

O comportamento das funções de estudo pode ser verificado na Figura 3.

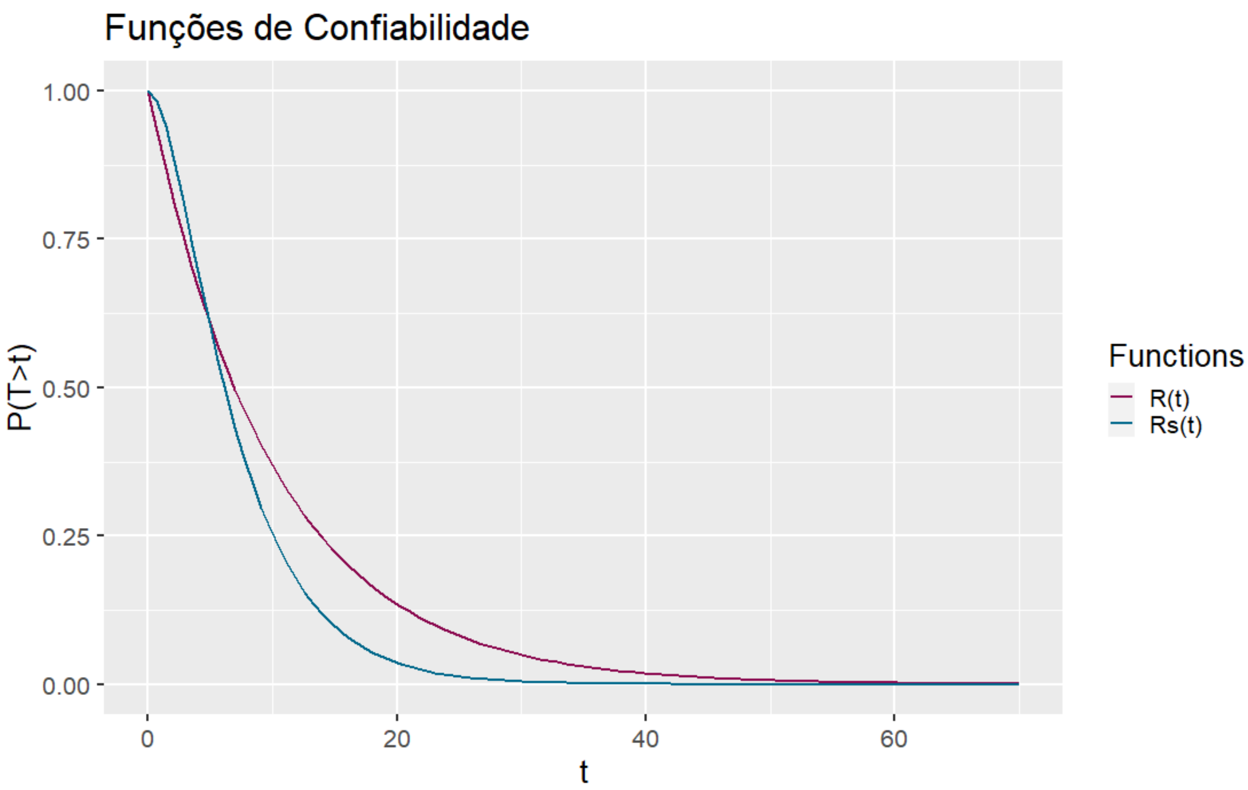

Figura 3: Gráfico para as funções $R(t)$ e $R_s(t)$

Analisando-se o gráfico, é notório que a confiabilidade do sistema $R_s(t)$ é maior do que a de uma componente $R(t)$ no intervalo $(0; 4,8121)$. Após $t \approx 4,8121$, a situação se inverte, sendo que a componente passa a ser mais confiável do que o sistema para períodos de tempo maiores que esse valor. Assim sendo, o sistema é mais confiável se o período de interesse estiver contido em $t \in (0; 4,8121)$, como podemos verificar no exemplo do item c) para $t = 2$. Porém, para outros valores de t, a componente passa a ser mais confiável.

Agora, considerando-se a confiabiliade de forma geral, ou seja, sem um valor de t até a falha determinado, a componente é mais confiável do que o sistema, pois em média, seu tempo de vida é maior, como verificado no item b) no cálculo das esperanças ($E[T_s] = 7,5$ e $E[T_c] = 10$).

Exercício 2

a) Sabe-se pelo desenvolvimento da definição de função de sobrevivência que:

$$S(t) = e^{-H(t)} = exp\left\{-\int_0^t h(u)du\right\}, \, t \geq 0$$

Onde $S(t)$ é a função de sobrevivência, $H(t)$ é a função de risco acumulado e $h(t)$ (a qual foi dada no exercício) é a função taxa de risco. Dessa forma, a $S(t)$ terá comportamento diferente em cada um dos intervalos que $h(t)$ está definida:

Intervalo: $0 \leq t < 2$:

$$S(t) = exp\left\{-\int_0^t h(u)du\right\} = exp\left\{-\int_0^t \lambda_1 du\right\} = e^{-\lambda_1 t}$$

Intervalo: $2 \leq t < 4$:

$$S(t) = exp\left\{-\int_0^t h(u)du\right\} = exp\left\{-\int_0^2 \lambda_1 du - \int_2^t \lambda_2 du\right\} = e^{-2\lambda_1-(t-2)\lambda_2}$$

Intervalo: $4 \leq t < \infty$:

$$S(t) = exp\left\{-\int_0^t h(u)du\right\} = exp\left\{-\int_0^2 \lambda_1 du - \int_2^4 \lambda_2 du - \int_4^t \lambda_3 du\right\} = e^{-2\lambda_1-2\lambda_2-(t-4)\lambda_3}$$

Assim sendo, pode-se definir $S(t)$ por:

$$S(t) = \begin{cases} e^{-\lambda_1 t}, 0 \leq t < 2 \\ e^{-2\lambda_1-(t-2)\lambda_2}, 2 \leq t < 4 \\ e^{-2\lambda_1-2\lambda_2-(t-4)\lambda_3}, 4 \leq t < \infty \end{cases}$$

Agora, verifiquemos se $S(t)$ satisfaz as propriedades de uma função de sobrevivência:

- $\lim_{t\to 0} S(t) = 1$:

$$\lim_{t\to 0} S(t) = \lim_{t\to 0} e^{-\lambda_1 t} = e^{-\lambda_1 \cdot 0} = 1$$

- $\lim_{t\to\infty} S(t) = 0$:

$$lim_{t\to 0} S(t) = lim_{t\to 0} e^{-2\lambda_1-2\lambda_2-(t-4)\lambda_3} =$$
$$e^{-2\lambda_1-2\lambda_2+4\lambda_3} \cdot \lim_{t\to\infty} e^{-\lambda_3 t} = e^{-2\lambda_1-2\lambda_2+4\lambda_3} \cdot 0 = 0,$$
sendo que λ_1 e λ_2 são números reais

- $S(t)$ é uma função não crescente:

  Uma função monótona não crescente é definida pela seguinte condição:

$$\forall x, y \in A, \ (x > y \Rightarrow f(x) < f(y)),$$

  Em que A é um conjunto ordenado. Sabe-se que $\mathbb{R}$ é um conjunto bem-ordenado e portanto o subconjunto dos reais definido por $[0, \infty)$, onde a função de confiabilidade $S(t)$ está definida, é um conjunto ordenado. Assim sendo, analisemos se a condição é satisfeita.

  Primeiramente, independente do intervalo analisado, é notório que $S(t)$ sempre possui o seguinte comportamento:

$$S(t) = c_1 e^{c_2 t}, \ t \geq 0 \text{ (I)},$$

  sendo c_1 e c_2 constantes não negativas pertencentes aos reais. Dessa forma, se a condição de função não crescente for satisfeita para todas as funções do tipo (I), ela será satisfeita para $S(t)$ em qualquer intervalo, já que a função de cofiabilidade sempre respeita a forma (I) independente do valor de t$\in [0, \infty)$. Assim, segue a demonstração:

$$x > y \Rightarrow -x < -y \Rightarrow -c_2 x < -c_2 y \Rightarrow$$
$$e^{-c_2 x} < e^{-c_2 y} \Rightarrow c_1 e^{-c_2 x} < c_1 e^{-c_2 y}$$
$$\Rightarrow S(x) < S(y)$$

Assim, a seguinte condição é satisfeita:

$$\forall x, y \in [0, \infty), \, (x > y \Rightarrow S(x) < S(y)),$$

E pode-se concluir que $S(t)$ é uma função monótona não crescente.

b) Pela definição de função densidade de probabilidade, dada por $f(t)$, é possível obter a sguinte relação:

$$f(t) = -\frac{\mathrm{d}}{\mathrm{d}t} S(t)$$

Assim sendo, derivando-se e multiplicando-se por (-1) a função $S(t)$ obtida no item a) deste exercício, tem-se o resultado:

$$f(t) = \begin{cases} \lambda_1 e^{-\lambda_1 t}, 0 \leq t < 2 \\ \lambda_2 e^{-2\lambda_1 - (t-2)\lambda_2}, 2 \leq t < 4 \\ \lambda_3 e^{-2\lambda_1 - 2\lambda_2 - (t-4)\lambda_3}, 4 \leq t < \infty \end{cases}$$

O qual é a função densidade de probabilidade definida para os intervalos de interesse da reta real na qual ela está definida ($t \geq 0$).

c) O tempo médio até a falha é dado pela esperança estatística da variável aleatória T de interesse, a qual pode ser obtida por sua definição e pode ser relacionada com a função de sobrevivência $S(t)$:

$$E[T] = \int_0^\infty t \cdot f(t) dt = \int_0^\infty S(t) dt$$

Dessa forma, resolvamos a integral antes de substituir os valores para λ_i, $i = 1, 2, 3$ dados pelo item:

$$E[T] = \int_0^\infty S(t)dt = \int_0^2 e^{-\lambda_1 t} dt + \int_2^4 e^{-2\lambda_1 - (t-2)\lambda_2} dt + \int_4^\infty e^{-2\lambda_1 - 2\lambda_2 - (t-4)\lambda_3} dt =$$
$$-\frac{e^{-\lambda_1 t}}{\lambda_1}\Big|_0^2 + e^{-2\lambda_1 + 2\lambda_2}\left(-\frac{e^{-\lambda_2 t}}{\lambda_2}\right)\Big|_2^4 + e^{-2\lambda_1 - 2\lambda_2 + 4\lambda_3}\left(-\frac{e^{-\lambda_3 t}}{\lambda_3}\right)\Big|_4^\infty =$$
$$\frac{1 - e^{-2\lambda_1}}{\lambda_1} + \frac{(e^{-2\lambda_1 + 2\lambda_2})(e^{-2\lambda_2} - e^{-4\lambda_2})}{\lambda_2} + \frac{(e^{-2\lambda_1 - 2\lambda_2 + 4\lambda_3})(e^{-4\lambda_3})}{\lambda_3}$$

Substituindo-se na equação acima $\lambda_1 = 0,01$, $\lambda_2 = 0,02$ e $\lambda_3 = 1$ e realizando-se os cálculos, obtém-se:

$$E[T] \approx 4,7682$$

Sendo T a variável aleatória de interesse, o tempo mediano $t_{0,5}$ é definido como o valor de t tal que:

$$F(t_{0,5}) = 0,5 \Rightarrow S(t_{0,5}) = 1 - 0,5 = 0,5$$

Como $S(t)$ é uma função não crescente, a qual é segmentada, analisemos os valores que ela assume nas fronteiras dos segmentos (para $\lambda_1 = 0,01$, $\lambda_2 = 0,02$ e $\lambda_3 = 1$):

$$S(0) = 1$$
$$S(2) = e^{-2\lambda_1-(t-2)\lambda_2} = e^{-\lambda_1 \cdot 2-(2-2)\lambda_2} = e^{-0,02} \approx 0,980$$
$$S(4) = e^{-2\lambda_1-2\lambda_2-(t-4)\lambda_3} = e^{-2\lambda_1-2\lambda_2-(4-4)\lambda_3} = e^{-0,06} \approx 0,942$$

Assim, pode-se concluir que $t_{0,5}$ é maior do que 4, conjunto de valores na qual $S(t)$ assume o formato:

$$S(t) = e^{-2\lambda_1-2\lambda_2-(t-4)\lambda_3}$$

E para obter o valor de interesse, basta resolver a equação para $t_{0,5}$, considerando-se que $\lambda_1 = 0,01$, $\lambda_2 = 0,02$ e $\lambda_3 = 1$:

$$S(t_{0,5}) = e^{-2\lambda_1-2\lambda_2-(t_{0,5}-4)\lambda_3} = 0,5 \Rightarrow$$
$$e^{-2\lambda_1-2\lambda_2+4\lambda_3} \cdot e^{\lambda_3 t_{0,5}} = 0,5 \Rightarrow$$
$$e^{3,94} \cdot e^{-t_{0,5}} = 0,5 \Rightarrow t_{0,5} = -\log(0,5/e^{3,94}) \rightarrow$$
$$t_{0,5} \approx 4,63315$$

Exercício 3

a) No estudo em questão há dois grupos de interesse envolvidos: pacientes que apresentam células cancerígenas com aneuploidia (anormal) e pacientes que apresentam células cancerígenas com diploidia (normal). O objetivo é estudar o comportamento da curva de sobrevivência para os pacientes de cada grupo. O gráfico apresentado, corresponde à estimativa de Kaplan-Meier (K-M) para as funções de sobrevivência desses dois grupos. Sejam as curvas de sobrevivência dadas por:

$S_a(t)$: curva de sobrevivência para grupo com aneuploidia (anormal)
$S_n(t)$: curva de sobrevivência para grupo com diploidia (normal)

As estimativas K-M podem ser observadas no gráfico da Figura 4, sendo $\hat{S}_a(t)$ a estimativa K-M para $S_a(t)$ e $\hat{S}_n(t)$ para $S_n(t)$. É notório que $\hat{S}_a(t) > \hat{S}_n(t)$ no intervalo de tempo estudado, o que leva a crer que a probabilidade de um paciente que apresente aneuploidia sobreviver mais do que um determinado tempo $t \in [0, \infty)$ é maior do que a probabilidade de um paciente que apresente diploidia sobreviver mais do que o mesmo tempo t. mesmo que a análise gráfica leve a crer que $S_a(t) > S_n(t)$ para $t \in [0, \infty)$, como $\hat{S}_a(t)$ e $\hat{S}_n(t)$ são estimativas e possuem seus respectivos erros padrão, é necessário um teste estatístico para que uma conclusão possa ser feita com determinado nível de significância sobre a igualdade ou não igualdade entre $S_a(t)$ e $S_n(t)$. Assim, como é mostrado no exercício, um teste Logrank foi realizado para a comparação das curvas de sobrevivência de interesse. O resultado de tal teste é discutido no item b) desta solução.

b) O teste Logrank se trata de um teste não paramétrico para a comparação de funções de sobrevivência a partir de uma amostra de tempos de sobrevivência de dois tratamentos de interesse. No caso do exercício, quer-se comparar se as funções de sobrevivência para pacientes que apresentavam células cancerígenas com aneuploidia (anormal) e pacientes que apresentavam células cancerígenas com diploidia (normal) podem ser consideradas iguais ou não. Considerando-se duas funções de sobrevivência de interesse, $S_0(t)$ e $S_1(t)$, as hipóteses que são comparadas neste teste estatístico são as seguintes:

$$H_0\text{: } S_0(t) = S_1(t) \text{ vs } H_1\text{: } S_0(t) = [S_1(t)]^{\phi},\ \phi > 0, \text{ para todo } t > 0.$$

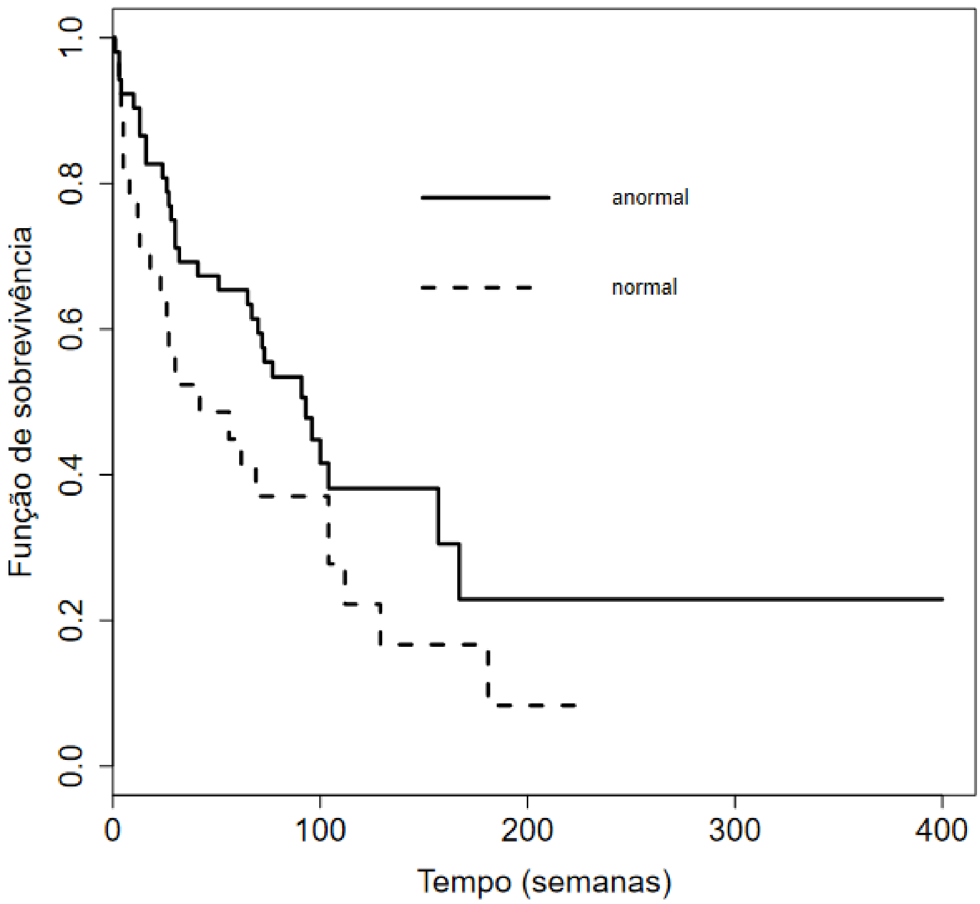

Figura 4: Estimativa K-M para as funções de sobrevivência de interesse.

Assim, ao rejeitar-se a hipósete nula H_0, há evidências nas amostras coletadas de que as funções de confiabilidade não são iguais para determinado nível de significância α.

Considerando-se o valor do enunciado de $\alpha = 0,1$ e o resultado do deste Logrank realizado com a linguagem R para os dados de interesse, pode-se concluir que há evidências para rejeitar-se a hipótese nula. O nível descritivo (p-valor) do teste resultou em $p = 0,0949$, ou seja, dado que a hipótese nula é verdadeira, há uma probabilidade de $0,0949$ de observar-se um valor para a estatística de teste igual ou mais extremo do que o observado na amostra. Assim sendo, se tomarmos $\alpha = 0,1 > 0,0949$, rejeitamos a hipótese nula e concluimos que existem evidências estatisticamente significativas de que as curvas de sobrevivência para pacientes que apresentavam células cancerígenas com aneuploidia (anormal) e pacientes que apresentavam células cancerígenas com diploidia (normal) não são iguais, considerando-se $\alpha = 0,1$.

Exercício 4

Exercício 2, lista 3: Os dados de estudo do exercício são tempos (censurados e nao censurados) de remissão, sem semanas, para 30 pacientes com leucemia em um determinado tipo de tratamento de interesse. Os dados podem ser observados na Figura 5. Como estamos lidando com uma variável aleatória de interesse que representa o tempo e há a presença de dados censurados e não censurados, será realizada uma análise de sobrevivência para o estudo em questão.

1,	1,	2,	4,	4,	6,	6,	7,	8,	9
9,	10,	12,	13,	14,	18,	19,	24,	26,	29
31+,	42,	45+	,50+,	57,	60,	71+,	85+	91	

Figura 5: Dados observados de remissão para pacientes com leucemia em um tipo de tratamento.

a) Aqui iremos obter as curvas para os estimadores Kaplan-Meyer e Nelson-Aalen utilizando-se de uma rotina de R e das bibliotecas *survival* e *survminer*. Primeiramente, lembremos de como é feita a obtenção de cada estimador:

Kaplan-Meyer: Supondo que n unidades experimentais foram coletadas de um problema de análise de sobrevivência:

- $t_1 < \cdots < t_k$: os k tal que $k < n$ tempos distintos e ordenados de falha;
- d_j: o número de falhas no tempo t_j;
- n_j: o número de indivíduos sob risco em t_j.

O estimador de Kaplan-Meyer (K-M), ou do produto-limite, é definido por:

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

Nelson Aalen: O estimador de Nelson-Aalen (N-A) é um estimador para a função de risco acumulado $H(t)$ da variável aleatória de interesse. Assim, supondo que n unidades experimentais foram coletadas de um problema de análise de sobrevivência:

- $t_1 < \cdots < t_k$: os k tal que $k < n$ tempos distintos e ordenados de falha;
- d_j: o número de falhas no tempo t_j;
- n_j: o número de indivíduos sob risco em t_j.

O estimador N-A para a função de sobreviv|ência da variável aleatória T é:

$$\hat{S}_{NA}(t) = exp\left\{-\hat{H}(t)\right\}, \text{ onde}$$

$$\hat{H}(t) = \sum_{j:t_j<t}\frac{d_j}{n_j}$$

Tanto o estimador K-M quanto o estimador N-A podem ser obtidos por meio de funções presentes no pacote *survival* do R. A rotina para sua obtenção pode ser verificada abaixo.

```r
rm(list=ls(all=TRUE))

# Bibliotecas
library(survminer)
library(survival)
library(ggfortify)

####################################################################

#a) Obtencao das curvas de sobrevivencia estimadas no N-A e por K-M

# Dados fornecidos pelo exercicio
tempos<- c(1,1,2,4,4,6,6,7,8,9,9,10,12,13,14,18,19,24,26,29,31,42,45,
           50,57,60,71,85,91)
censuras <- c(rep(1,20),0,1,0,0,1,1,0,0,1)


# Funcao do pacote survival para obter o estimador K-M
fit1 <- survfit(Surv(tempos,censuras)~1)
# Funcao do pacote survival para obter o estimadorN-A
fit2 <- survfit(coxph(Surv(tempos,censuras) ~ 1))


# Alguns links interessantes para comandos em graficos de sobrevivencia:

# https://rpkgs.datanovia.com/survminer/
# https://github.com/kassambara/survminer/issues/195
# https://rpkgs.datanovia.com/survminer/survminer_cheatsheet.pdf
# https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_surv.html


# Utilizacao do pacote survminer para a obtencao das curvas de sobrevivencia

fit <- list("Keplan-Meyer" = fit1, "Nelson-Aalen" = fit2)

ggsurvplot(fit, combine = TRUE,          # Combinar curvas no grafico
           legend.labs =
             c("Keplan-Meyer",           # Legendas para as curvas
               "Nelson-Aalen"),
           conf.int = T,                 # Apresenta o IC[95%]
           #conf.int.style = "step",     # Estilo grafico para o IC
           censor = TRUE,                # Mostrar censuras
           palette = "jco",
           xlab = "Tempo t (semanas)",   # Nomes para os eixos
           ylab = "P(T>t)",
           ggtheme = theme_gray()        # Estilo de plotagem do ggplot
           #risk.table = TRUE,           # Appresentar tabela de risco
           #risk.table.col = "strata",   # Cores na tabela de risco
           #risk.table.height = 0.25     # Altura da tabela de risco
           )
```

Após a obtenção dos objetos que representão as curvas de sobrevivência estimadas por K-M e N-A, utilizou-s da função *ggsurvplot* da biblioteca *survminer* para obter-se o gráfico da Figura 6.

Analisando-se tal gráfico, é notório que a curva obtida pelo estimador K-M assume sempre valores iguais ou menores quando comparada com a estimação N-A para o mesmo valor de t. Ou seja:

$$\hat{S}_{NA}(t) \geq \hat{S}_{KM}(t), \text{ para } t > 0$$

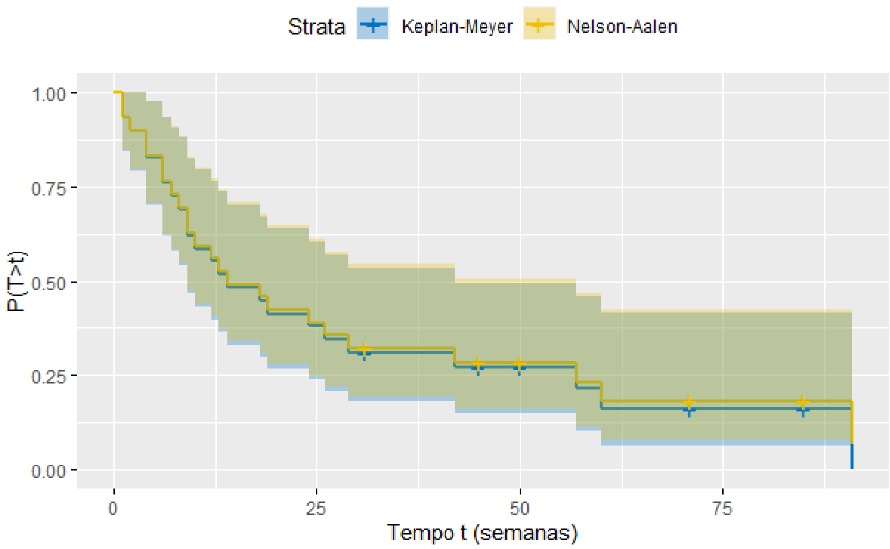

Figura 6: Curvas de sobrevivência obtidas por estimação K-M e N-A sobre o tempo de remissão para pacientes de leucemia.

Assim sendo, a probabilidade $P(T > t)$ é menor na estimação K-M do que na estimação N-A para um mesmo valor de t.

b) O tempo de remissão médio (t_{rm}) é o tempo médio de vida (TMV) para a variável T de interesse no exercício. E um estimador para a sua obtenção é o seguinte, como apresentado e dsicutido no capítulo de estimadores nãi paramétricos de [1] é o seguinte:

$$T\hat{M}V = \hat{t_{rm}} = t_{(1)} + \sum_{j=1}^{k-1} \hat{S}_{KM}(t_j) \cdot (t_{(j+1)} - t_{(j)}),$$

onde $t_{(1)} \leq \cdots \leq t_{(k)}$ são os tempos de falha distintos e ordenados. Abaixo tem-se a rotina em R implementada para obter-se o valor pontual desse estimador para o problema em questão.

```
####################################################################


#b) Obtencao de uma estimativa para o TMV a partir da estimacao K-M:

# Pode-se obter os valores de t onde ha degraus na funcao de sobrevivencia e
# as respectivas probabilidades de sobrevivencia para cada patamar a partir
# das colunas da funcao summary.

tempos <- summary(fit1)[[2]] #segunda coluna apresenta os tempos
sobrev <- summary(fit1)[[6]] #sexta coluna apresenta P(T>t) para cada
   patamar


length(tempos)
length(sobrev)

#Funcao para a obtencao do TMV:
tempo_medio_est <- function(S,t){ # recebe S e t

  tmv <- S[1] # Valor inicial para o processo iterativo (t1)

  for(i in 1:(length(t)-1)){ # Soma de produtos do segundo termo do TMV
   estimado
```

```r
    tmv <- tmv + S[i] * (t[i+1] - t[i])
  }

  return(tmv) # retorna o valor obtido
}


tempo_medio_est(sobrev,tempos) #calculando-se para o problema
# Resposta: 30.39655 semanas
```

Assim sendo, o tempo de remissão méidio t_{rm} é de 30,396 semanas, como obtido após a compilação da rotina descrita.

c) Podemos estimar a variância do estimador para o tempo médio de remissão estudado no item anterior por meio do estimador, tembém apresentado e estudado em [1], mostrado a seguir:

$$\hat{Var}(\hat{t_m}) = \frac{r}{r-1} \sum_{j=1}^{r-1} \frac{A_j^2}{n_j(n_j - d_j)}$$

Onde:

- $A_j = S(\hat{t}_{(j)})(t_{(j+1)} - t_{(j)}) + \cdots + S(\hat{t_{(r-1)}})(t_{(r)} - t_{(r-1)})$
- $t_1 < \cdots < t_r$: os r tal que $r < n$ tempos de falha (observações não censuradas);
- d_j: o número de falhas no tempo t_j;
- n_j: o número de indivíduos sob risco em t_j.

Com esse resultado, pode-se estimar o erro padrão do estimador por:

$$\hat{ep} = \sqrt{\hat{Var}(\hat{t_m})}$$

Assim, foi utilizada da rotina em R apresentada abaixo para a o cálculo da medida para o estimador da variância de $t_r m$ e depois, de seu erro padrão estimado $\hat{ep}$.

```r
##################################################################

#c) Abaixo temos a rotina para a obtencao do erro padrao

#O valor r corresponde de falhas (observacoes nao censuradas)
r = length(censuras[censuras == 1])

# Obtencao dos dados da curva de sobrevivencia como no item anterior
tempos = summary(fit1)[[2]]
sobrev = summary(fit1)[[6]]

# Utilizando-se dos metodos do objeto fit1, pode-se obter as seguintes
   listas:
n_ur = fit1$n.risk #lista de unidades em risco ate determinado t
n_uf = fit1$n.event#lista de unidades que falharam ate determinado t

# A correspondencia entre esses valores e os tempos podem ser verificadas na
# tabela gerada pelo comando summary(fit1)

```

```
var_tmv = 0#inicializacao do valor da variancia estimada do tempo medio de vida

A = rep(0, length(sobrev)) # inicializacao do vetor de valores de Aj

#Laco para a obtencao dos valores dos termos Aj
for(i in 1:(length(sobrev) - 1) ){
  A[i] = sobrev[i]*(tempos[i + 1] - tempos[i])
}

#Laco para a obtencao da somatoria de Aj^2/nj(nj-dj), de j = 1 ate j = r-1
for(i in 1:(length(sobrev) - 1)){
  var_tmv =
    var_tmv + sum(A[i:length(A2)])^2/(n_ur[i] * (n_ur[i] - n_uf[i]))
  }

var_tmv = (r/(r-1))*var_tmv #obtencao do valor final para a variancia do
# estimador do tmv


ep_tmv <- sqrt(var_tmv) #obtencao do erro padrao estimado do estimador de tmv

ep_tmv

# Resposta: 5.373217
```

Dessa forma, o estimador para o erro padrão do estimador do tempo remissão médio é dado por:

$$\hat{ep} = 5,373$$

d) Como descrito no capítulo 2 do trabalho de Colosimo e colaboradores [1], é possível obter os quantis para a curva de sobrevivência de interessse por meio da técnica de interpolação linear. Assim, podemos obter a mediana $t_{0,5}$ interpolada por meio da seguinte equação:

$$\frac{13-14}{0,517-0,483} = \frac{13-t_{0,5}}{0,517-0,5} \Rightarrow t_{0,5} = 13,5 semanas$$

Os valores para realizar a interpolação linear podem ser obtidos da tablela gerada pelo comando *summary(fit1)* na rotina em R que foi trabalhada neste exercício:

```
#time  n.risk  n.event  survival  std.err  lower 95% CI  upper 95% CI
#    1     29        2     0.931   0.0471       0.8432         1.000
#    2     27        1     0.897   0.0566       0.7923         1.000
#    4     26        2     0.828   0.0701       0.7009         0.977
#    6     24        2     0.759   0.0795       0.6178         0.932
#    7     22        1     0.724   0.0830       0.5784         0.907
#    8     21        1     0.690   0.0859       0.5403         0.880
#    9     20        2     0.621   0.0901       0.4670         0.825
#   10     18        1     0.586   0.0915       0.4318         0.796
#   12     17        1     0.552   0.0923       0.3974         0.766
#   13     16        1     0.517   0.0928       0.3639         0.735
#   14     15        1     0.483   0.0928       0.3312         0.704
#   18     14        1     0.448   0.0923       0.2994         0.671
#   19     13        1     0.414   0.0915       0.2683         0.638
#   24     12        1     0.379   0.0901       0.2381         0.604
#   26     11        1     0.345   0.0883       0.2088         0.569
#   29     10        1     0.310   0.0859       0.1804         0.534
#   42      8        1     0.272   0.0835       0.1487         0.496
```

```
#    57        5        1      0.217   0.0826        0.1031          0.458
#    60        4        1      0.163   0.0778        0.0639          0.415
#    91        1        1      0.000      NaN            NA             NA
```

e) O mesmo procediemnto do item anterior pode ser obtido para obter $t_{0,1}$:

$$\frac{2-1}{0{,}897-0{,}931} = \frac{t_{0,1}-1}{\hat{S}_{KM}(t_{0,1})-0{,}931} \Rightarrow \frac{2-1}{0{,}897-0{,}931} = \frac{t_{0,1}-1}{0{,}9-0{,}931} \Rightarrow t_{0,1} \approx 1,912 \text{ semana}$$

Com tal resultado, pode-s concluir que em média 10% dos pacientes com leucemia para o contexto estudado terão remissão após 1,912 semana.

f) Utilizando-se da estimativa K-M obtida e dos valores presentes na tabela *summary(fit1)*, podemos novamente utilizar de interpolação linear para obter os valores estimados nos pontos de interesse, e, na mesma tabela, encontram-se os valores inferiores e superiores para o intervalode confiança de $\gamma = 95\%$ para cada patamar da função (basta verificar o intervalo de confiança no intervalo que o valor de t de interesse se encontra), o qual iremos incluir ao lado do resultado da estimação:

$$\frac{2-1}{0{,}897-0{,}931} = \frac{1{,}5-1}{\hat{S}_{KM}(1{,}5)-0{,}931} \Rightarrow \hat{S}_{KM}(1,5) \approx 0,914,\ IC[95\%] = [0,843; 1,000]$$

De forma análoga, pode-se fazer a estimação para os outros dois valores de interesse de t:

$$\hat{S}_{KM}(11) \approx 0,568,\ IC[95\%] = [0,432; 0,796]$$

$$\hat{S}_{KM}(40) \approx 0,278,\ IC[95\%] = [0,180; 0,534]$$

Quando obtemos uma estimação para a curva de sobrevivência em um determinado ponto t, estamos estimando a probabilidade de um indivíduo não ter remissão até o tempo t. O intervalo de confiança de 95% nos diz que para tal valor t que pretende-se estimar, o intervalo aleatório obtido contém o verdadeiro valor de $S(t)$ com 95% de confiança. Por exemplo, estimamos a probabilidade de que não haja remissão $P(T > 1,5) = 0,914$ e o valor verdadeiro de $P(T > 1,5)$ está contido em $[0,843; 1,000]$ com 95% de confiança.

Exercício 4, lista 3: O exercício apresenta os dados presentes na Figura 7 e propõe que seja realizado um teste para verificar se as curvas de sobrevivênicia para o aparecimento da primeira alteração de saúde pós-cirurgica em pacientes tratados com a droga *Compath* ($S_c(t)$) e a droga *Zena* ($S_z(t)$) são as mesmas. Os dados amostrados podem ser verificados na Figura 7.

Aqui optou-se pelo uso do teste não paramétrico Logrank, o qual é amplamente utilizado para a comparação de funções de sobrevivência. As hipóteses a serem comparadas neste tipo de teste são:

$$H_0\colon\ S_0(t) = S_1(t) \text{ vs } H_1\colon\ S_0(t) = [S_1(t)]^{\phi},\ \phi > 0, \text{ para todo } t > 0.$$

Então, para o caso de interesse, iremos comparar as seguintes hipóteses por meio de um teste Logrank:

$$H_0\colon\ S_c(t) = S_z(t) \text{ vs } H_1\colon\ S_c(t) = [S_z(t)]^{\phi},\ \phi > 0, \text{ para todo } t > 0.$$

O teste foi performado em uma implementação na linguagem R utilizando-se de funções do pacote *survival*.

Droga	Tempos (em dias) até a ocorrência da 1^a alteração pós-cirúrgica
Compath	8 11 19 24* 28 33 36* 38 44 96 124 130 250 250* 250*
Zena	7 8 10 12 13 14* 19 23 25* 26 27 31 31* 49 59* 64* 87 89 107 117 119 130 148 153 156 159 191 222 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250* 250*

Figura 7: Dados de tempo até o aparecimento de alterações no estado de saúde de pacientes tratados com as drogas *Compath* e *Zena* após procedimento cirúrgico no intestino. Os tempos assinalados com (*) são censurados. O acompanhamento foi realizado no decorrer de 250 dias.

Resultados: A seguinte rotina em R foi aplicada para a obtenção da resposta do teste Logrank:

```r
# Limpeza do ambiente de trabalho
rm(list=ls(all=TRUE))

#Bibliotecas
library(survminer)
library(survival)
library(ggfortify)


# Tempos e censuras firnecidas pelo exercicio:
temposCompath <- c(8,11,19,24,28,33,36,38,44,96,124,130,250,250,250)
censuraCompath <- c(1,1,1,0,1,1,0,1,1,1,1,1,1,0,0)

temposZena <- c(7,8,10,12,13,14,19,23,25,26,27,31,31,49,59,64,87,89,107,117,119,
                130,148,153,156,159,191,222,rep(250,16))
censuraZena <- c(rep(1,5),0,1,1,0,1,1,1,0,1,0,0,rep(1,12),rep(0,16))


# Vetor contendo os tempos para as duas drogas (concatenado)
tempos <- c(temposCompath,temposZena)
# Vetor contendo as censuras para as duas drogas (concatenado)
censura <- c(censuraCompath,censuraZena)

# Divisor para os grupos de estudo (grupo 1: Compath, grupo 2: Zena)
grupo <- c(rep(1,length(temposCompath)), rep(2,length(temposZena)))

#Teste Logrank:
fit <- survdiff(Surv(tempos,censura)~grupo)
fit

            #######SAIDA OBSERVADA#######


#          N Observed Expected (O-E)^2/E (O-E)^2/V
#grupo=1 15       11     6.76     2.656      3.38
#grupo=2 44       23    27.24     0.659      3.38

```

```
#    Chisq= 3.4  on 1 degrees of freedom, p= 0.07
```

Assim sendo, é notório que o nível descritivo para o teste Logrank performado é $p = 0,07$. Assim sendo, tomando-se um nível de significância de $\alpha = 0,1$; rejeita-se a hipótese nula H_0 e conclui-se que há evidências amostrais significativas de que as curvas de sobrevivência $S_c(t)$ e $S_z(t)$ não são iguais.

Referência: [1] COLOSIMO, Enrico Antonio; GIOLO, Suely Ruiz. Análise de sobrevivência aplicada. Editora Blucher, 2006.