Primeira prova

Análise de Sobrevivência

Paulo Ricardo Seganfredo Campana

18 de agosto de 2024

Questão 1a: Considere as seguintes funções S(t) apresentadas abaixo, considerando as condições, vistas em sala de aula, para que S(t) seja uma função de sobrevivência, identifique quais das funções apresentadas são realmente funções de sobrevivência.

1. $S_1(t) = e^{-0.2t},\ t \geqslant 0$
2. $S_2(t) = \dfrac{1}{1+t},\ t \geqslant 0$
3. $S_3(t) = 1 - \dfrac{t}{2},\ t \geqslant 0$
4. $S_4(t) = 2e^{-0.5t},\ t \geqslant 0$

Com a relação de que $S(t) = 1 - F(t)$, temos requisitos semelhantes para que $S(t)$ seja função de sobrevivência:

- $\lim\limits_{t\to-\infty} S(t) = 1$ e $\lim\limits_{t\to\infty} S(t) = 0$
- Continuidade a direita
- Função não crescente

Todas as funções acima são contínuas pois tratam de composições de funções contínuas e não crescentes pois tem derivadas não positivas em todo o suporte $[0, \infty)$, porém temos que:

$$\lim_{t\to\infty} S_3(t) = \lim_{t\to\infty} 1 - \frac{t}{2} \neq 0$$
$$\lim_{t\to0} S_4(t) = \lim_{t\to0} 2e^{-0.5t} = 2 \neq 1$$

Assim S_3 e S_4 não são funções de sobrevivência, mas S_1 e S_2 são.

Questão 2a: Escolha um dos bancos de dados disponíveis no seguinte endereço eletrônico: http://sobrevida.fiocruz.br/dados.html. Faça uma análise exploratória do banco de dados e forneça interpretações plausíveis acerca das variáveis que encontram-se disponíveis.

```
library(tidyverse)
data <- read.table("ctinca.txt", header = TRUE)
```

O conjunto de dados `ctinca.txt` são provenientes da UTI no Instituto Nacional de Câncer, contêm informações sobre 862 pacientes internados no instituto, contendo as seguintes informações para cada um:

Variável	Tipo	Descrição
Tempo	Numérica	Tempo de sobrevivência, em dias
Status	Dicotômica	Se ocorreu óbito ou censura
Idade	Numérica	Idade do paciente em anos completos
Sexo	Dicotômica	Masculino ou Feminino
Tumor	Categórica	Tumor sólido localizado, metastático ou hematológico
Desnutrição	Dicotômica	Perda de peso recente ou IMC baixo
Comorbidade	Dicotômica	Presença de comorbidades severas
Leucopenia	Dicotômica	Se ocorreu redução do número de glóbulos brancos

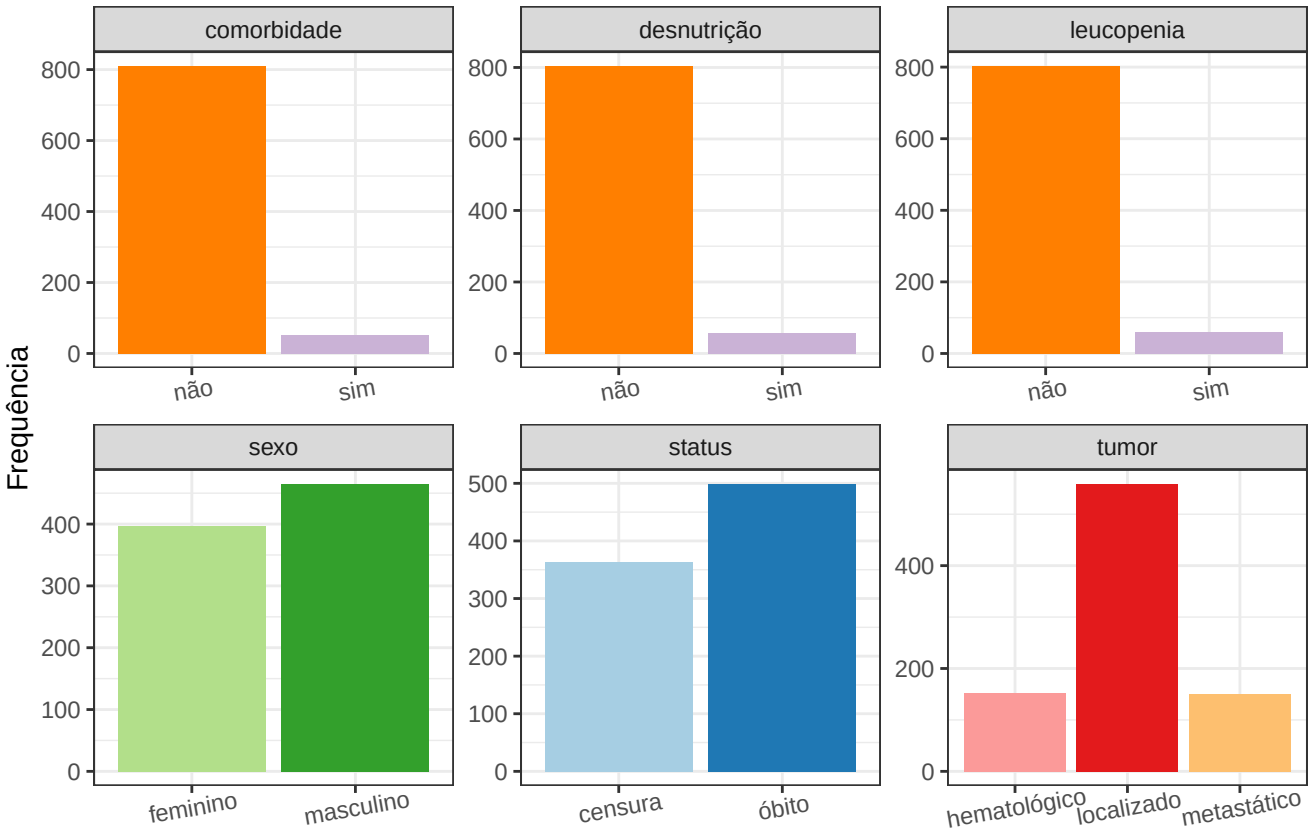

Figura 1: Distribuição das variáveis dicotômicas e categóricas do conjunto de dados

Segundo a Figura 1, em geral, a maioria dos pacientes não tem comorbidades severas e não sofreram perda de peso e leucopenia durante o acompanhamento. A quantidade de homens e mulheres internados é equilibrada, porém este conjunto de dados apresenta muita censura, cerca de 42%. Entre os tipos de tumores classificados, 65% deles são do tipo localizado sólido, o menos violento pois em geral não se espalha para outras partes do corpo. 499 dos 862 pacientes morreram por complicações do câncer durante o estudo, os outros 363 viveram por mais de 182 dias, quando o estudo teve conclusão.

Questão 2b: Considerando a variável de tempo até ocorrência do evento de interesse na base de dados escolhida, forneça as seguintes informações:

1. É possível montar uma tabela para descrever os dados de acordo com o número de ocorrências do evento de interesse registradas em intervalos de tempo da pesquisa? Se sim, apresente-a.

Sim, esta tabela apresenta os tempos únicos de falha em t_j, o número de falhas neste instante em d_j e o número de pacientes sobre risco em n_j.

Por brevidade, será mostrado apenas as primeiras 15 linhas das tabelas.

```
tbl <- table(data$tempo)
tj <- as.numeric(names(tbl))
dj <- as.numeric(tbl)
nj <- length(data$tempo) - c(0, cumsum(dj))[1:length(tbl)]
amplitude <- c(diff(as.numeric(tj)), 1)
empirico <- data.frame(
    intervalo = paste0("[", tj, ", ", tj + amplitude, ")"),
    tj = tj,
    dj = dj,
    nj = nj
)
```

Tabela 2: Eventos de interesse pelo estimador empírico

Intervalo	t_j	d_j	n_j
[1, 2)	1	23	862
[2, 3)	2	41	839
[3, 4)	3	38	798
[4, 5)	4	24	760
[5, 6)	5	22	736
[6, 7)	6	22	714
[7, 8)	7	18	692
[8, 9)	8	16	674
[9, 10)	9	14	658
[10, 11)	10	15	644
[11, 12)	11	7	629
[12, 13)	12	14	622
[13, 14)	13	10	608
[14, 15)	14	13	598
[15, 16)	15	8	585

2. Apresente o cálculo da função de sobrevivência empírica (pela definição apresentada na aula 1). Apresente também as estimativas empíricas das seguintes quantidades: função densidade, função de risco, função de risco acumulada

Usando os estimadores empíricos:

$$\hat{f}(t) = \frac{\#\text{ falhas em } t}{\#\text{ indivíduos} \times \text{amplitude do intervalo}} \qquad \hat{S}(t) = \frac{\#\text{ indivíduos sob risco}}{\#\text{ indivíduos}}$$

$$\hat{h}(t) = \frac{\#\text{ falhas em } t}{\#\text{ indivíduos sob risco} \times \text{amplitude do intervalo}} \qquad \hat{H}_x(t) = \sum_{k=1}^{x-1} \hat{h}_k(t)\Delta_k$$

```
empirico$densidade <- dj / (sum(dj) * amplitude)
empirico$sobrevivencia <- nj / sum(dj)
empirico$risco <- dj / (nj * amplitude)
empirico$risco_acumulado <- c(0, cumsum(empirico$risco * amplitude))[1:length(tbl)]
```

Tabela 3: Funções de sobrevivência pelo estimador empírico

Intervalo	t_j	d_j	n_j	$\hat{f}(t)$	$\hat{S}(t)$	$\hat{h}(t)$	$\hat{H}(t)$
[1, 2)	1	23	862	0.0267	1.0000	0.0267	0.0000
[2, 3)	2	41	839	0.0476	0.9733	0.0489	0.0267
[3, 4)	3	38	798	0.0441	0.9258	0.0476	0.0755
[4, 5)	4	24	760	0.0278	0.8817	0.0316	0.1232
[5, 6)	5	22	736	0.0255	0.8538	0.0299	0.1547
[6, 7)	6	22	714	0.0255	0.8283	0.0308	0.1846
[7, 8)	7	18	692	0.0209	0.8028	0.0260	0.2155
[8, 9)	8	16	674	0.0186	0.7819	0.0237	0.2415
[9, 10)	9	14	658	0.0162	0.7633	0.0213	0.2652
[10, 11)	10	15	644	0.0174	0.7471	0.0233	0.2865
[11, 12)	11	7	629	0.0081	0.7297	0.0111	0.3098
[12, 13)	12	14	622	0.0162	0.7216	0.0225	0.3209
[13, 14)	13	10	608	0.0116	0.7053	0.0164	0.3434
[14, 15)	14	13	598	0.0151	0.6937	0.0217	0.3599
[15, 16)	15	8	585	0.0093	0.6787	0.0137	0.3816

3. Apresente o cálculo da função de sobrevivência $S(t)$ considerando os seguintes estimadores: Kaplan-Meier, Nelson-Aalen e Tabela de Vida. Para cada versão desses estimadores, apresente também as estimativas das seguintes quantidades: função densidade, função de risco, função de risco acumulada. Interprete os resultados.

A estimação das funções de sobrevivência e risco acumulado são feitas pelo `R`, as funções de risco e densidade podem ser obtidas pelas relações entre elas.

```
library(survival)
surv <- Surv(data$tempo, data$status)

km <- survfit(surv ~ 1)
kaplan_meier <- empirico
kaplan_meier$sobrevivencia <- km$surv
```

```r
kaplan_meier$risco_acumulado <- km$cumhaz
kaplan_meier$risco <- c(diff(kaplan_meier$risco_acumulado), 1)
kaplan_meier$densidade <- kaplan_meier$risco * kaplan_meier$sobrevivencia
```

Tabela 4: Funções de sobrevivência pelo estimador Kaplan-Meier

Intervalo	t_j	d_j	n_j	$\hat{f}_{\text{KM}}(t)$	$\hat{S}_{\text{KM}}(t)$	$\hat{h}_{\text{KM}}(t)$	$\hat{H}_{\text{KM}}(t)$
[1, 2)	1	23	862	0.0476	0.9733	0.0489	0.0267
[2, 3)	2	41	839	0.0441	0.9258	0.0476	0.0755
[3, 4)	3	38	798	0.0278	0.8817	0.0316	0.1232
[4, 5)	4	24	760	0.0255	0.8538	0.0299	0.1547
[5, 6)	5	22	736	0.0255	0.8283	0.0308	0.1846
[6, 7)	6	22	714	0.0209	0.8028	0.0260	0.2155
[7, 8)	7	18	692	0.0186	0.7819	0.0237	0.2415
[8, 9)	8	16	674	0.0162	0.7633	0.0213	0.2652
[9, 10)	9	14	658	0.0174	0.7471	0.0233	0.2865
[10, 11)	10	15	644	0.0081	0.7297	0.0111	0.3098
[11, 12)	11	7	629	0.0162	0.7216	0.0225	0.3209
[12, 13)	12	14	622	0.0116	0.7053	0.0164	0.3434
[13, 14)	13	10	608	0.0151	0.6937	0.0217	0.3599
[14, 15)	14	13	598	0.0093	0.6787	0.0137	0.3816
[15, 16)	15	8	585	0.0081	0.6694	0.0121	0.3953

```r
na <- survfit(surv ~ 1, stype = 2)
nelson_aalen <- empirico
nelson_aalen$sobrevivencia <- na$surv
nelson_aalen$risco_acumulado <- na$cumhaz
nelson_aalen$risco <- c(diff(nelson_aalen$risco_acumulado), 1)
nelson_aalen$densidade <- nelson_aalen$risco * nelson_aalen$sobrevivencia
```

Tabela 5: Funções de sobrevivência pelo estimador Nelson-Aalen

Intervalo	t_j	d_j	n_j	$\hat{f}_{\text{NA}}(t)$	$\hat{S}_{\text{NA}}(t)$	$\hat{h}_{\text{NA}}(t)$	$\hat{H}_{\text{NA}}(t)$
[1, 2)	1	23	862	0.0476	0.9737	0.0489	0.0267
[2, 3)	2	41	839	0.0442	0.9272	0.0476	0.0755
[3, 4)	3	38	798	0.0279	0.8841	0.0316	0.1232
[4, 5)	4	24	760	0.0256	0.8566	0.0299	0.1547
[5, 6)	5	22	736	0.0256	0.8314	0.0308	0.1846
[6, 7)	6	22	714	0.0210	0.8062	0.0260	0.2155
[7, 8)	7	18	692	0.0186	0.7855	0.0237	0.2415
[8, 9)	8	16	674	0.0163	0.7671	0.0213	0.2652
[9, 10)	9	14	658	0.0175	0.7509	0.0233	0.2865
[10, 11)	10	15	644	0.0082	0.7336	0.0111	0.3098
[11, 12)	11	7	629	0.0163	0.7255	0.0225	0.3209
[12, 13)	12	14	622	0.0117	0.7093	0.0164	0.3434
[13, 14)	13	10	608	0.0152	0.6978	0.0217	0.3599
[14, 15)	14	13	598	0.0093	0.6828	0.0137	0.3816
[15, 16)	15	8	585	0.0082	0.6735	0.0121	0.3953

As estimativas da tabela de vida podem ser feitas agregando os tempos de falha em intervalos e aplicando o estimador de Kaplan-Meier.

```
vida_tempo <- cut(data$tempo, breaks = 10)
vida_surv <- Surv(as.numeric(vida_tempo), data$status)
vida <- survfit(vida_surv ~ 1)

tabela_vida <- data.frame(
    intervalo = levels(vida_tempo),
    dj = vida$n.event,
    nj = vida$n.risk,
    densidade = 1,
    sobrevivencia = 1,
    risco = 1
)

tabela_vida$sobrevivencia <- vida$surv
tabela_vida$risco_acumulado <- vida$cumhaz
tabela_vida$risco <- c(diff(tabela_vida$risco_acumulado), 1)
tabela_vida$densidade <- tabela_vida$risco * tabela_vida$sobrevivencia
```

Tabela 6: Funções de sobrevivência pelo estimador Tabela de Vida

Intervalo	d_j	n_j	$\hat{f}_{\text{TV}}(t)$	$\hat{S}_{\text{TV}}(t)$	$\hat{h}_{\text{TV}}(t)$	$\hat{H}_{\text{TV}}(t)$
(0.819,19.1]	314	862	0.0742	0.6357	0.1168	0.3643
(19.1,37.2]	64	548	0.0499	0.5615	0.0888	0.4811
(37.2,55.3]	43	484	0.0255	0.5116	0.0499	0.5699
(55.3,73.4]	22	441	0.0128	0.4861	0.0263	0.6198
(73.4,91.5]	11	419	0.0174	0.4733	0.0368	0.6460
(91.5,110]	15	408	0.0081	0.4559	0.0178	0.6828
(110,128]	7	393	0.0070	0.4478	0.0155	0.7006
(128,146]	6	386	0.0116	0.4408	0.0263	0.7162
(146,164]	10	380	0.0081	0.4292	0.0189	0.7425
(164,182]	7	370	0.4211	0.4211	1.0000	0.7614

Dessa forma, vemos que o estimador de Kaplan-Meier coincide exatamente com a estimação empírica, exceto por uma defasagem: $\hat{S}_{\text{KM}}(t) = \hat{S}(t+1)$. Devido ao grande número de amostras, os estimadores Kaplan-Meier e Nelson-Aalen obteram estimativas muito parecidas, com os primeiros dois dígitos significantes iguais.

A tabela de vida deixa mais fácil a visualização do comportamento da sobrevivência pois resume 113 falhas em tempos distintos em 10 intervalos, assim podemos ver que o maior risco é no começo do acompanhamento pois apenas 63% dos pacientes sobrevivem os primeiros 19 dias pela função de sobrevivência e possui valor mais alto na função de rísco.

4. Explique como o teste de LogRank deve ser aplicado. Escolha uma variável qualitativa de sua base e realize o teste de comparação de curvas de sobrevivência. Interprete adequadamente os resultados.

O teste LogRank é o mais comum para comparar curvas de sobrevivências entre grupos, é um tipo de teste Chi-quadrado de bondade do ajuste sobre as funções de risco, ou seja, a estatística deste teste trabalha

em cima da diferença entre os valores observados e esperados sob a hipótese de que os grupos possuem o mesmo risco, assim se estas diferenças forem muito grandes, podemos rejeitar a hipótese de igualdade de risco entre os grupos.

```
survdiff(surv ~ tumor, data)
## Call:
## survdiff(formula = surv ~ tumor, data = data)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## tumor=hematológico 152      120     71.4      33.1      39.7
## tumor=localizado   559      270    355.7      20.7      74.2
## tumor=metastático  151      109     71.9      19.1      22.9
##
##  Chisq= 75.2  on 2 degrees of freedom, p= <2e-16
```

Temos uma estatística de teste de $\chi^2 = 75.2$ e p-valor extremamente baixo, assim podemos concluir que há diferença significativa entre os tipos de tumores na sobrevivência dos pacientes como pode ser visto na Figura 2, os pacientes com tumores localizados tem uma sobrevivência maior que os demais tipos, pois esse tipo de tumor é mais concentrado em uma região do corpo, não afetando as demais.

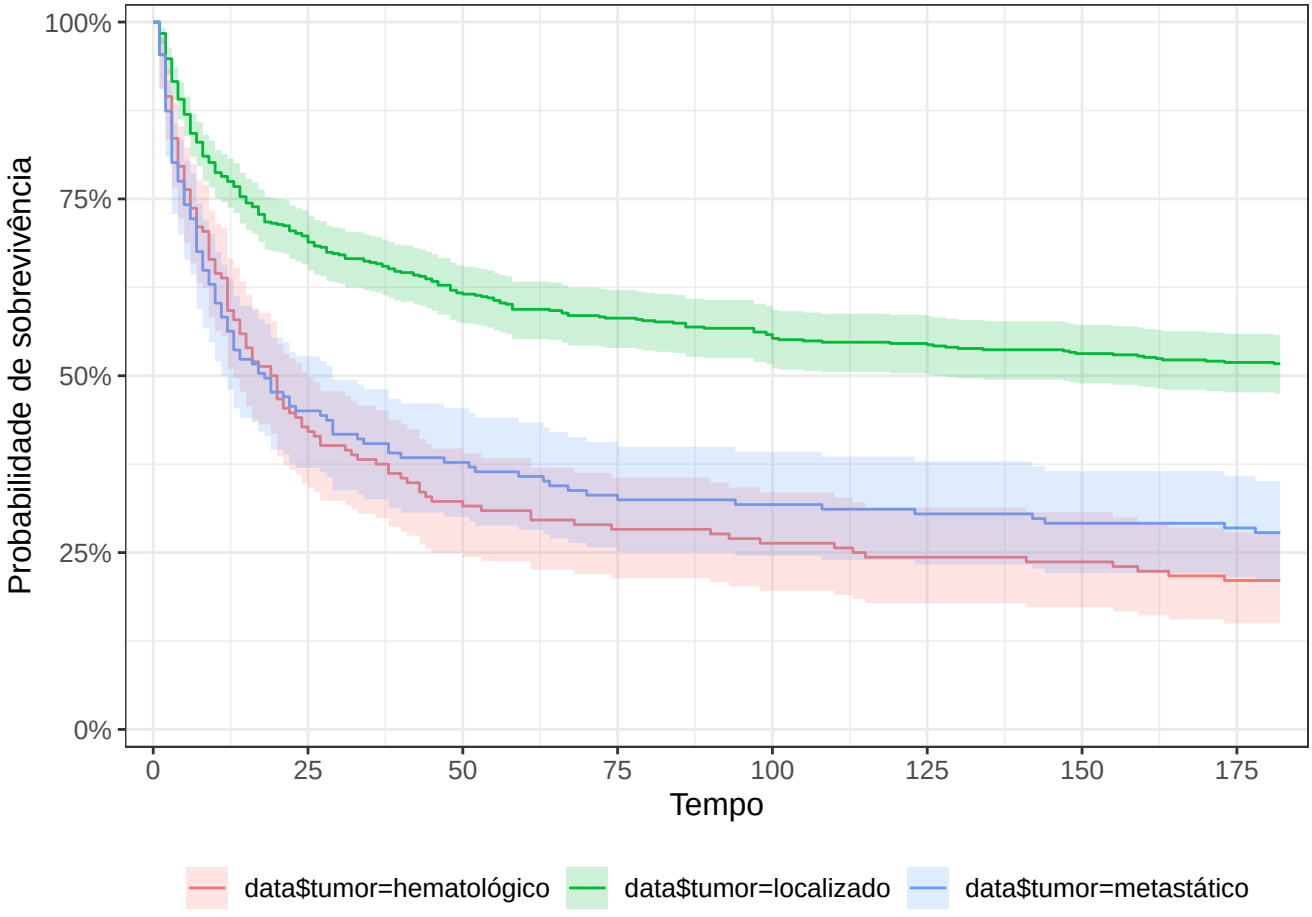

Figura 2: Curvas de sobrevivência dos diferentes tipos de tumores

5. Apresente o cálculo da função de sobrevivência $S(t)$ considerando as seguintes distribuições de probabilidade: Exponencial, Weibull, Gama, Log-Normal, Gama Generalizada e as duas

distribuições da questão 1 que você identificou. Apresente os valores do AIC e BIC apenas para os ajustes baseados nas distribuições Exponencial, Weibull, Gama, Log-Normal e Gama Generalizada. Como você pode comparar as estimativas geradas por essas distribuições a partir do teste da razão de verossimilhanças? Interprete os resultados.

```
exp <- flexsurv::flexsurvreg(surv ~ 1, dist = "exp")
wei <- flexsurv::flexsurvreg(surv ~ 1, dist = "weibull")
gam <- flexsurv::flexsurvreg(surv ~ 1, dist = "gamma")
gga <- flexsurv::flexsurvreg(surv ~ 1, dist = "gengamma")
lnm <- flexsurv::flexsurvreg(surv ~ 1, dist = "lnorm")

parametrico <- data.frame(
    intervalo = paste0("[", tj, ", ", tj + amplitude, ")"),
    dj = dj,
    nj = nj,
    exp = summary(exp)[[1]]$est,
    wei = summary(wei)[[1]]$est,
    gam = summary(gam)[[1]]$est,
    gga = summary(gga)[[1]]$est,
    lnm = summary(lnm)[[1]]$est,
    S1 = (function (t) exp(-0.2 * t))(tj),
    S2 = (function (t) 1 / (1 + t))(tj)
)
```

A tabela com os valores das funções de sobrevivência paramétricas se encontram na próxima página.

```
stat <- sapply(
    list(exp, wei, gam, gga, lnm), function(x) list(
        x$AIC, x$BIC,
        lmtest::lrtest(gga, x)$Chisq[2],
        lmtest::lrtest(gga, x)$Pr[2]
    )
)
```

Tabela 7: Critérios de informação e teste de razão de verossimilhança contra a distribuição Gama generalizada

Modelo	AIC	BIC	Estatística	p-valor
Exponencial	6069.4	6074.2	728.7	0
Weibull	5599.6	5609.1	256.9	0
Gama	5651.5	5661.0	308.8	0
Gama generalizado	5344.7	5359.0	0.0	1
Log-normal	5488.7	5498.3	146.0	0

Apresentando menor AIC e BIC é o modelo baseado na distribuição Gama generalizada, o grande número de amostra permite utilizar a maior liberdade da distribuição para obter estimativas melhores. Realizando testes de razão de verossimilhanças contra o modelo Gama generalizado mostram que os demais modelos são estatísticamente diferentes e inferiores.

Tabela 8: Funções de sobrevivência dos modelos paramétricos

Intervalo	d_j	n_j	$\hat{S}_{Exp}(t)$	$\hat{S}_{Wei}(t)$	$\hat{S}_{Gam}(t)$	$\hat{S}_{GenGam}(t)$	$\hat{S}_{LogNorm}(t)$	$S_1(t)$	$S_2(t)$
[1, 2)	23	862	0.9938	0.9251	0.9243	0.9923	0.9541	0.8187	0.5000
[2, 3)	41	839	0.9876	0.8972	0.8994	0.9476	0.9225	0.6703	0.3333
[3, 4)	38	798	0.9815	0.8766	0.8812	0.9019	0.8975	0.5488	0.2500
[4, 5)	24	760	0.9754	0.8598	0.8663	0.8642	0.8765	0.4493	0.2000
[5, 6)	22	736	0.9694	0.8453	0.8536	0.8333	0.8583	0.3679	0.1667
[6, 7)	22	714	0.9634	0.8324	0.8422	0.8077	0.8421	0.3012	0.1429
[7, 8)	18	692	0.9574	0.8208	0.8320	0.7860	0.8275	0.2466	0.1250
[8, 9)	16	674	0.9514	0.8102	0.8226	0.7672	0.8142	0.2019	0.1111
[9, 10)	14	658	0.9455	0.8004	0.8138	0.7508	0.8020	0.1653	0.1000
[10, 11)	15	644	0.9397	0.7913	0.8057	0.7362	0.7906	0.1353	0.0909
[11, 12)	7	629	0.9338	0.7827	0.7980	0.7232	0.7800	0.1108	0.0833
[12, 13)	14	622	0.9280	0.7746	0.7907	0.7114	0.7700	0.0907	0.0769
[13, 14)	10	608	0.9223	0.7669	0.7838	0.7007	0.7607	0.0743	0.0714
[14, 15)	13	598	0.9166	0.7596	0.7772	0.6908	0.7518	0.0608	0.0667
[15, 16)	8	585	0.9109	0.7526	0.7709	0.6817	0.7434	0.0498	0.0625
[16, 17)	7	577	0.9052	0.7459	0.7648	0.6733	0.7354	0.0408	0.0588
[17, 18)	9	570	0.8996	0.7395	0.7590	0.6655	0.7278	0.0334	0.0556
[18, 19)	7	561	0.8940	0.7334	0.7534	0.6582	0.7205	0.0273	0.0526
[19, 20)	6	554	0.8885	0.7274	0.7479	0.6514	0.7136	0.0224	0.0500
[20, 21)	6	548	0.8830	0.7217	0.7427	0.6449	0.7069	0.0183	0.0476
[21, 22)	4	542	0.8775	0.7162	0.7375	0.6388	0.7004	0.0150	0.0455
[22, 23)	7	538	0.8721	0.7108	0.7326	0.6331	0.6943	0.0123	0.0435
[23, 24)	4	531	0.8667	0.7056	0.7278	0.6276	0.6883	0.0101	0.0417
[24, 25)	4	527	0.8613	0.7006	0.7231	0.6224	0.6825	0.0082	0.0400
[25, 26)	6	523	0.8559	0.6957	0.7185	0.6175	0.6770	0.0067	0.0385
[26, 27)	4	517	0.8506	0.6909	0.7140	0.6128	0.6716	0.0055	0.0370
[27, 28)	4	513	0.8453	0.6863	0.7097	0.6083	0.6664	0.0045	0.0357
[28, 29)	5	509	0.8401	0.6818	0.7054	0.6040	0.6613	0.0037	0.0345
[29, 30)	4	504	0.8349	0.6774	0.7013	0.5999	0.6564	0.0030	0.0333
[30, 31)	1	500	0.8297	0.6731	0.6972	0.5959	0.6516	0.0025	0.0323

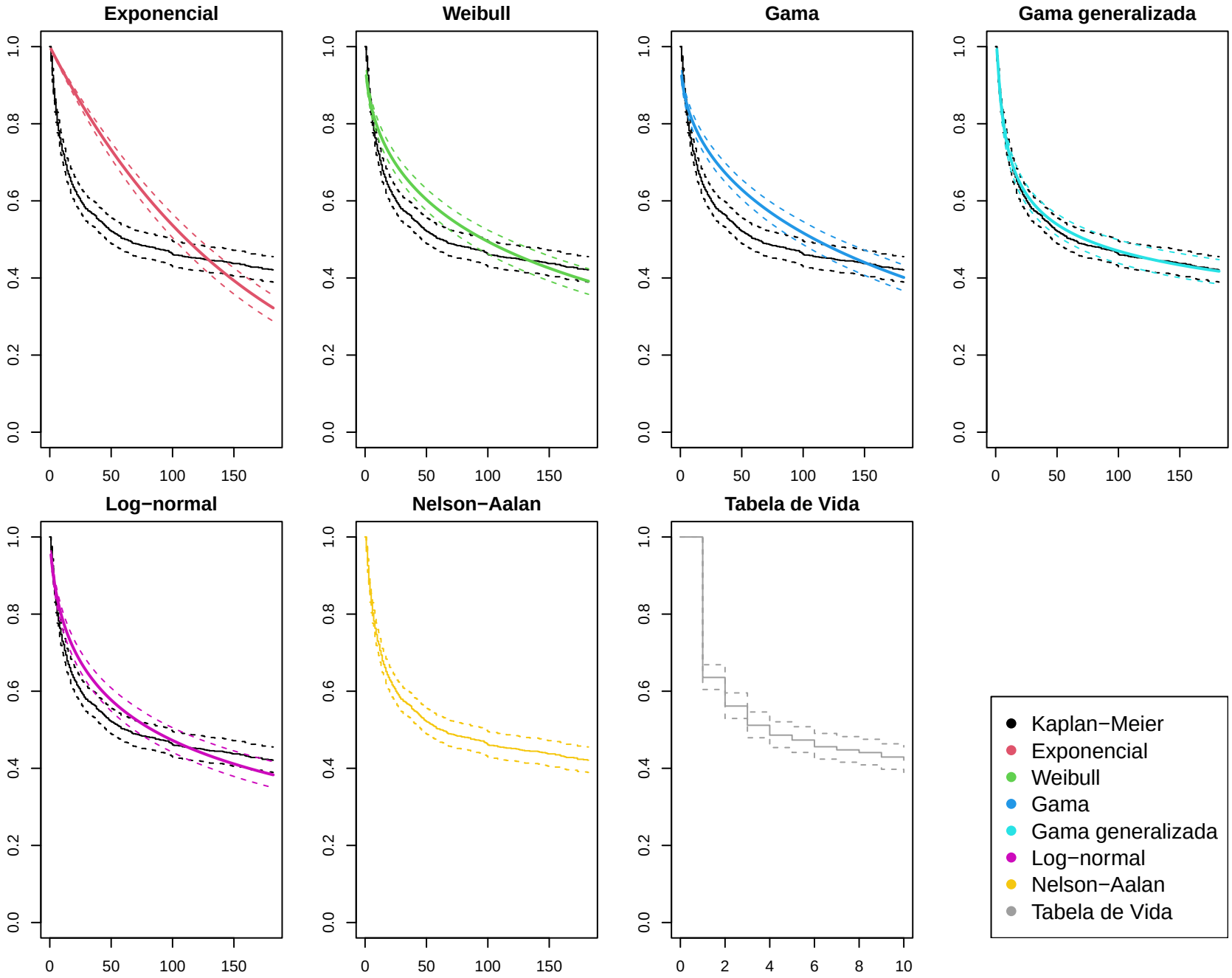

Figura 3: Funções de sobrevivência e intervalos de confiança para os estimadores não paramétricos e paramétricos

6. Para cada um dos resultados apresentados nos itens 2., 3. e 5., apresente os respectivos intervalos de confiança, a partir da definição. Interprete os resultados.

7. Apresente métodos gráficos de comparação das funções de sobrevivência consideradas no item 5. com as obtidas pelo estimador de Kaplan-Meier. Interprete os resultados

Segundo a Figura 3 o único modelo que se adequa corretamente dados, acompanhando o estimador de Kaplan-Meier é o modelo Gama generalizado, os modelos Exponencial, Weibull, Gama e Log-normal superestimam a sobrevivência na maior parte do acompanhamento do paciente e sobestimam no final do estudo.

Os intervalos de confiança da são em geral pequenos devido ao tamanho da amostra de 862 pacientes, porém é possível notar que o modelo exponencial tem uma margem de erro menor comparado com o Gama generalizado, devido a estimação de menos parâmetros.

O estimador de Nelson-Aalan como visto anteriormente é semelhante ao de Kaplan-Meier, com os primeiros digitos decimais iguais, o estimador da tabela de vida é quantizada pelos intervalos.

8. Considerando todos os resultados, o que é possível concluir sobre a sobrevivência dos pacientes analisados na sua base de dados?

O acompanhamento do estudo em pacientes com câncer permitiu quantificar a sobrevivência destes pacientes, com isso vimos que o maior risco do câncer é no começo do acompanhamento, onde de 862 pacientes, mais de 20 faleceram a cada dia da primeira semana. O risco da doênça decresce bastante com e tempo, ocorrendo a conclusão do estudo com 42% dos pacientes vivos após 182 dias.

Entre os tipos de tumores de câncer, verificamos que os tumores sólidos localizados são mais frequêntes e também acarretam significantemente menos risco pelo teste LogRank, cerca de 50% dos pacientes com tumores sólidos localizados tiveram desfecho de óbito no estudo, comparado com cerca de 25% para os tipos de tumores metastáticos e hematológicos.

Modelando a sobrevivência do câncer de maneira paramétrica, o modelo de distribuição Gama generalizado obteve excelete ajuste, com melhores critérios de seleção de modelo e significantemente diferente dos demais atravéz do teste de razão de verossimilhança.

Os intervalos de confiança calculados e vistos em gráficos tem margem de erro em geral pequena, dando maior credibilidade para as estimativas pontuais de sobrevivência do estimador Kaplan-Meier e do modelo Gama Generalizado.