



FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Inferencia estatística con datos censurados

Álvaro Añón Dosil

Xullo, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Inferencia estatística con datos censurados

Álvaro Añón Dosil

Xullo, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Inferencia estatística con datos censurados
Breve descrición do contido
No estudo dos tempos de vida, que poden ser duracións dunha enfermidade, dun artigo de consumo (coches, teléfonos, ordenadores, etc.) ou calquera outro tempo entre dous eventos, a obtención dunha observación adoita requirir dun seguimento do individuo ou obxecto en cuestión. Se ese seguimento se interrompe, só coñeceremos que o tempo durou polo menos ata o momento da perda do seguimento. Nese caso diremos que o tempo foi censurado. Neste traballo estudaranse modelos de distribución propios de tempos de vida, como o exponencial ou o Weibull, xunto con estimadores dos seus parámetros con datos censurados. Tamén se estudará o estimador de Kaplan-Meier, que é o estimador non paramétrico da función de distribución con datos censurados. Expoñeranse os métodos existentes, estudaranse as súas propiedades mediante simulacións e ilustraranse con datos reais.
Recomendacións
Ter un coñecemento básico do programa estatístico  .
Outras observacións

Índice

Resumo	VIII
Introdución	XI
1. Funcións relevantes na análise de supervivencia	1
1.1. Función de supervivencia	1
1.2. Función de risco	3
1.3. Función de esperanza de vida residual	5
1.4. Relacións entre as funcións	6
2. Métodos de estimación paramétricos	9
2.1. Introdución	9
2.2. Función de verosimilitude baixo censura	10
2.3. Modelo exponencial	11
2.4. Modelo Weibull	13
2.5. Estudo de simulación	16
2.5.1. Propiedades dos estimadores	16
2.5.2. Propiedades dos intervalos de confianza	17
3. Métodos de estimación non paramétricos	21
3.1. Táboas de vida	21

3.2. Estimador de Kaplan-Meier	24
3.3. O paquete <i>survival</i> de  . Aplicación aos datos <i>pharmacoSmoking</i>	29
3.4. Probas de bondade de axuste	35
3.4.1. Modelo exponencial	35
3.4.2. Modelo Weibull	36
A. Código de  empregado	39
A.1. Figuras da introdución	39
A.1.1. Figuras 1 e 2.	39
A.1.2. Figura 3	40
A.2. Figuras do Capítulo 1	41
A.2.1. Figura 1.1	41
A.2.2. Figura 1.2	41
A.3. Capítulo 2	42
A.3.1. Figura 2.1	42
A.3.2. Figura 2.2	43
A.3.3. Estudo de simulación	44
A.4. Capítulo 3	48
A.4.1. Exemplo 3.2	48
A.4.2. Estudo dos datos <i>pharmacoSmoking</i>	49
Bibliografía	51

Resumo

Neste traballo estúdanse algunhas das ferramentas máis habituais na análise de datos con censura, enmarcadas dentro da rama da estatística denominada análise de supervivencia. Para comezar, introducíronse unha serie de conceptos, definicións e funcións básicas recurrentes ao longo do traballo. A continuación, explicouse o método de estimación de modelos paramétricos mediante máxima verosimilitude con datos censurados. Ademais, facendo uso de  fíxose un estudo de simulación para analizar as propiedades dos estimadores paramétricos, dos intervalos de confianza asíntóticos e por perfil de verosimilitude. Máis adiante, presentáronse os estimadores non paramétricos da distribución, primeiro o que se obtén coas táboas de vida e despois mediante o estimador de Kaplan-Meier. Para rematar, mostrouse unha proba gráfica de bondade de axuste dos modelos paramétricos baseada no estimador de Kaplan-Meier. Ilustráronse estes métodos cos datos *pharmacoSmoking* do paquete *asaur*.

Abstract

In this project some of the most common tools for the analysis of censored data are studied, framed within the branch of statistics known as survival analysis. To begin, we introduced some basic concepts, definitions and functions that appear throughout the project. Next, the estimation method based on maximum likelihood with censored data for parametric models was explained. Besides, a simulation study was made using  , analyzing properties of parametric estimates, asymptotic confidence intervals and by profile likelihood. Later, non-parametric estimates of the distribution are presented, firstly using life tables and then using the Kaplan-Meier estimate. To conclude, a graphic method to analyze the goodnes-of-fit of parametric models based on the Kaplan-Meier estimate was shown. All these methods were exemplified with *pharmacoSmoking* data from *asaur* package.

Introdución

A necesidade de analizar o tempo que transcorre entre un instante inicial e a aparición dun evento de interese está presente en numerosos campos. Por exemplo, se consideramos como instante inicial o comezo dun tratamento contra unha enfermidade e como evento de interese a remisión da mesma, analizar o tempo transcorrido entre estes dous sucesos vainos dar unha idea da eficacia do tratamento empregado. Estes tempos, aos que nos referiremos coma *tempos de supervivencia*, son moi útiles como xa vimos no ámbito da medicina ou a biomedicina (ensaos clínicos), aínda que tamén poden ser interesantes noutras áreas como a industria, a criminoloxía (tempo de liberdade condicional de delincuentes) ou a socioloxía (duración do primeiro matrimonio). Para analizar este tipo de datos introduciremos diferentes técnicas, as cales se enmarcan dentro da rama da estatística coñecida como **Análise de Supervivencia**.

Unha característica importante e diferencial da Análise de Supervivencia é a habitual presenza de datos ‘incompletos’, é dicir, aqueles para os cales non se observou o evento de interese durante o tempo de estudo. Por exemplo, se nós queremos estudar o tempo de vida dun grupo de 30 persoas cunha determinada enfermidade e ao final da investigación todavía quedan 7 persoas vivas, vai haber 7 datos (tempos de vida) que non coñezamos con exactitude. Para darlle nome a este tipo de datos introducimos as seguintes definicións:

Definición 0.1. Denominamos **censura** ao fenómeno que ten lugar cando o valor exacto dunha observación só se coñece parcialmente. Polo tanto, os **datos censurados** son aquelas observacións afectadas pola censura, é dicir, aquelas para as cales non coñecemos o seu valor exacto, soamente algunha cota (inferior e/ou superior).

Dependendo de como estean acotados os datos censurados podemos distinguir tres tipos de censura:

- **Censura pola esquerda:** as observacións están acotadas superiormente, é dicir, o evento de interese ocorreu antes do estudo. En [9] menciónase o exemplo dunha enquisa realizada a adolescentes sobre o seu primeiro consumo de droga, onde aqueles que contestan que non lembran exactamente cando foi esa primeira vez están aportando datos censurados pola

esquerda, pois enténdese que ese primeiro consumo xa ocorreu antes da data da enquisa.

- **Censura pola dereita:** as observacións están acotadas inferiormente, é dicir, o evento de interese ocorre despois do momento da censura. Volvendo ao exemplo do tratamento médico, pode ocorrer que acabado o ensaio clínico algúns pacientes aínda sigan enfermos, aportando datos censurados pola dereita, xa que o único que se coñece é que o momento da remisión foi despois de finalizar o estudo. É a máis habitual, sobre todo en contextos biomédicos, e é coa que traballaremos de aquí en diante. Formalmente, os datos virán dados por pares (t, δ) onde δ é o indicador da censura, é dicir, $\delta = 0$ se o dato é censurado e $\delta = 1$ no caso contrario.
- **Censura en intervalo:** é unha combinación das dúas anteriores, as observacións están acotadas superiormente e inferiormente, contidas nun intervalo.

Como xa se mencionou, soamente traballaremos con datos censurados pola dereita, para os cales podemos diferenciar outros tres tipos de censura:

Censura tipo I

O tempo de estudo está prefixado, ben sexa por limitacións de orzamento, temporais ou polo propio deseño da investigación. Os individuos que non experimentaran o evento de interese neste tempo constitúen as observacións censuradas. Polo tanto, os tempos de supervivencia censurados coinciden coa duración do estudo. Vexámolo claramente co seguinte exemplo procedente de [10].

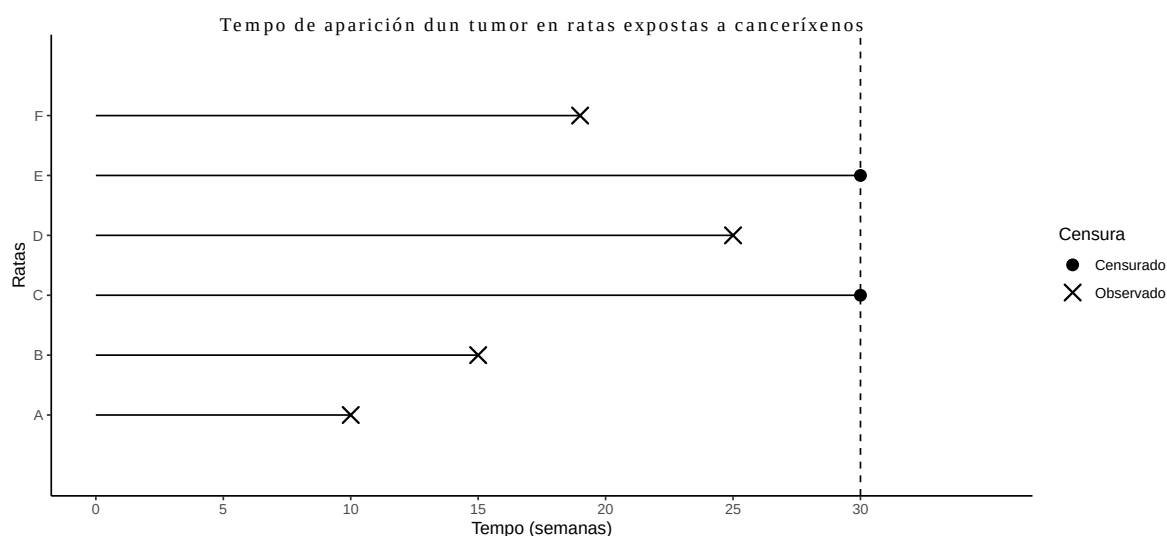


Figura 1: Exemplo de censura tipo I

Supoñamos que seis ratas son expostas a canceríxenos e queremos estudar o tempo que tarda en formarse un tumor. O investigador decide que o experimento remate despois de 30 semanas. Na figura 1 podemos ver os tempos de desenvolvemento dos tumores. Apreciamos que as ratas A, B, D e F padecen dun tumor despois de 10, 15, 25 e 19 semanas respectivamente, mentres que C e E non o padecen ao final do estudo (datos censurados).

Censura tipo II

Neste caso a investigación prolongarase ata que unha proporción fixada de obxectos de estudo experimenten o evento de interese (fallo). Un exemplo pode ser no ámbito da industria electrónica, onde pode interesar rematar un estudo cando certo número de dispositivos falle. Se consideramos o exemplo do apartado anterior, o investigador pode decidir rematar o estudo cando catro das seis ratas padezan dun tumor. Polo tanto, como podemos ver na figura 2, con este tipo de censura as observacións censuradas van ter o valor do máximo tempo de supervivencia observado.

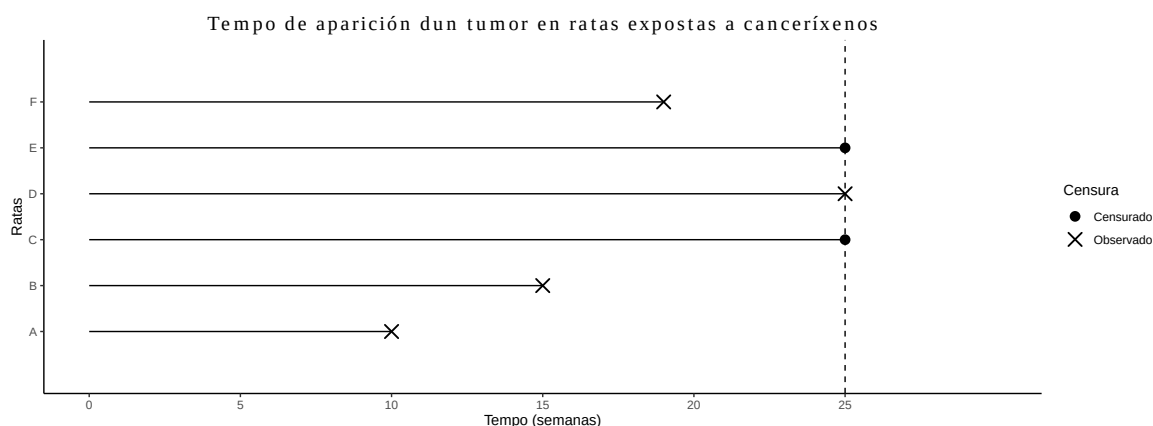


Figura 2: Exemplo de censura tipo II

Censura tipo III (censura aleatoria)

Nos dous casos anteriores temos visto que a censura vén controlada polo investigador que realiza o estudo. Neste último caso a censura vai deberse a un fenómeno aleatorio, que non está relacionado co evento de interese. Exemplos poden ser a perda de seguimento dun paciente nun ensaio clínico ou a súa morte por unha causa allea á investigación. Na figura 3 podemos ver un exemplo que mostra o tempo de remisión de pacientes tratados contra a leucemia. Nel vemos como os pacientes se incorporan ao ensaio clínico en distintos momentos e o paciente B podería ser un exemplo dunha persoa á que se lle perde o seguimento.

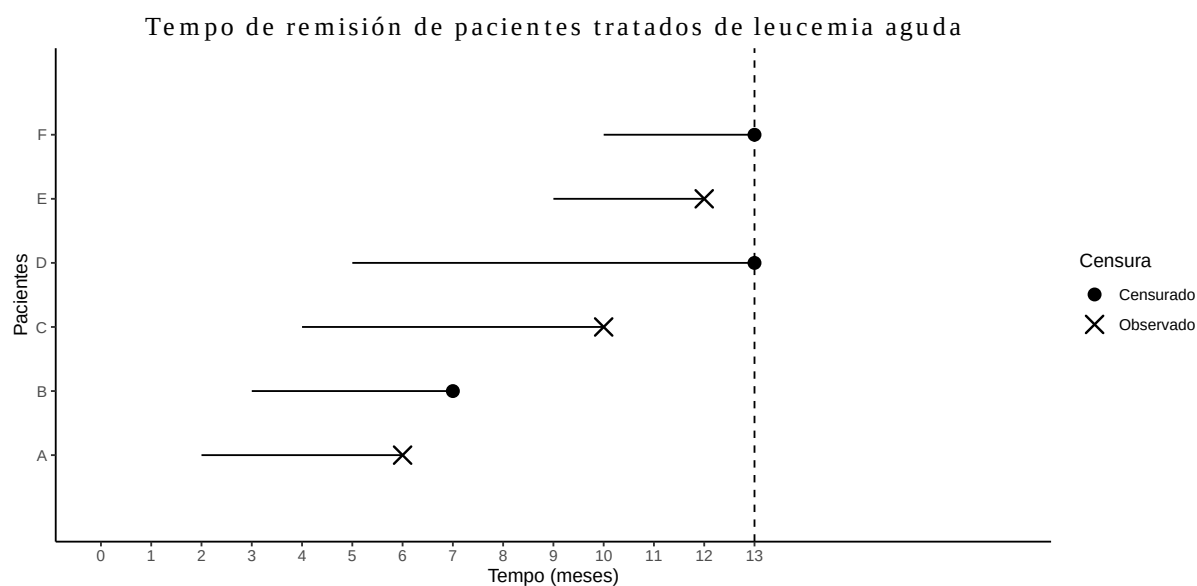


Figura 3: Exemplo de censura aleatoria

Debido a presenza deste tipo de observacións, precisamos construír novos estimadores que teñan en conta a censura para así coñecer mellor e estimar de forma máis precisa a distribución dos datos. Estes estimadores, os cales podemos clasificar en *paramétricos* e *non paramétricos*, serán os que nos permitirán facer inferencia e intentar sacar conclusións a partir das mostras. Ao longo deste traballo veremos como construír os distintos tipos de estimadores, discutiremos e compararemos as súas propiedades e as ilustraremos con exemplos empregando o *software* estatístico .

Capítulo 1

Funcións relevantes na análise de supervivencia

Como xa se mencionou, a variable aleatoria a estudar vai ser o tempo de supervivencia, a cal denotaremos por T . Supoñemos que T é unha variable aleatoria continua e non negativa, e como toda variable aleatoria absolutamente continua ten funcións de distribución e densidade que a caracterizan. Sen embargo, no marco da Análise de Supervivencia resulta máis interesante considerar outras funcións que, dende outro punto de vista, tamén caracterizan a mesma variable aleatoria. Ditas funcións son a *función de supervivencia*, a *función de risco* e a *función de esperanza de vida residual*. Neste capítulo introduciremos estas funcións e veremos que efectivamente caracterizan a T , é dicir, a partir de calquera delas podemos determinar de forma única as restantes. Antes de continuar, definimos as funcións de distribución e densidade, que serán empregadas de aquí en diante.

Definición 1.1. Sexa T unha variable aleatoria absolutamente continua. Defínense as funcións de densidade, f , e distribución, F , de T como:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t} \quad (1.1)$$

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(t)dt \quad (1.2)$$

1.1. Función de supervivencia

A función de supervivencia, a cal denotaremos por $S(t)$, é a máis empregada para analizar e describir fenómenos que involucran tempos de supervivencia. Defínese como a probabilidade de que un individuo sobreviva (experimente o evento de interés) máis alá dun tempo t . Formalmente:

$$S(t) = \mathbb{P}(T > t) \quad \text{con } 0 \leq t < \infty$$

Polo tanto, pódese relacionar coa función de distribución F e a función de densidade f da seguinte maneira:

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t) \quad \forall t \in [0, \infty) \quad (1.3)$$

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(t)dt \Rightarrow f(t) = -S'(t) \quad \forall t \in [0, \infty) \quad (1.4)$$

A representación gráfica de $S(t)$ coñécese como curva de supervivencia. Existen moitos tipos de curvas de supervivencia, pero todas comparten unha serie de propiedades:

- $S(t)$ é monótona decrecente, é dicir, $S(t_1) \geq S(t_2)$ se $t_1 < t_2$.
- $S(0) = 1$, polo que a probabilidade de sobrevivir nun instante inicial $t = 0$ é 1.
- $\lim_{t \rightarrow \infty} S(t) = 0$, é dicir, co paso do tempo a probabilidade de supervivencia converxe a cero.

Na figura 1.1 móstrase o exemplo da curva de supervivencia dunha Exponencial(1), no que se poden apreciar claramente as propiedades mencionadas, herdadas das propiedades da función de distribución.

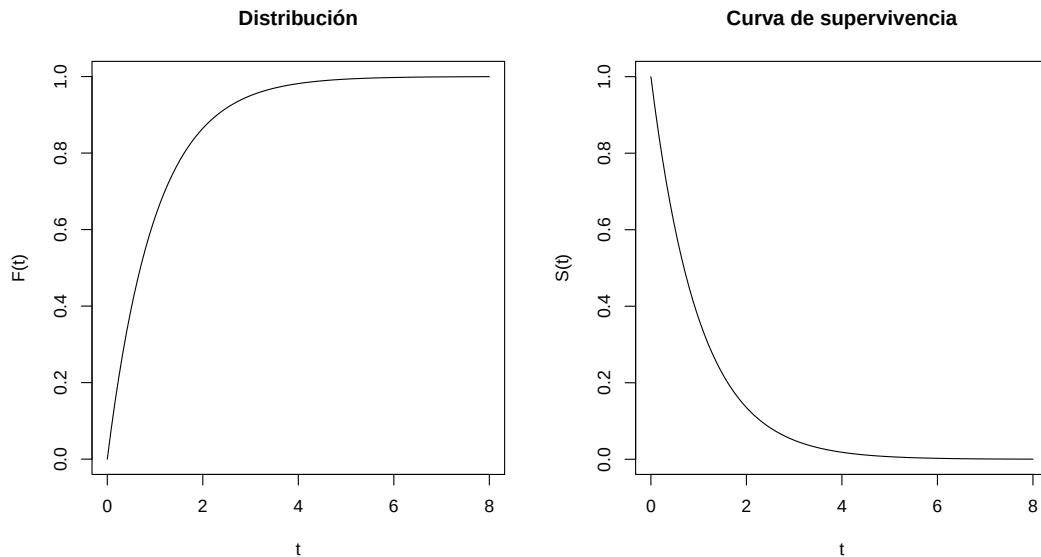


Figura 1.1: Función de distribución e supervivencia dunha Exp(1)

No caso de ter unha mostra completa, é dicir, sen datos censurados, a función de supervivencia nun instante t pode estimarse como a proporción de individuos sobrevivindo máis que t , o cal se coñece como *función de supervivencia empírica*.

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(T_i > t) \quad (1.5)$$

Sen embargo, máis adiante mostraremos que en presenza de datos censurados esta estimación non é boa, de feito non está ben definida, pois para as observacións censuradas non coñecemos se sobreviviron ou non máis aló dun tempo superior ao tempo de censura. Por este motivo construiremos outros estimadores máis apropiados que teñan en conta a censura.

1.2. Función de risco

A función de risco, $h(t)$, é a que determina o que se coñece como *taxa de fallo instantáneo*, que se define como a 'tendencia' a experimentar o evento de interese (fallo, morte...) nun intervalo de tempo moi pequeno supoñendo que o individuo sobreviviu ao comezo do mesmo. Formalmente:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \quad (1.6)$$

Podemos ver que se $\Delta t \approx 0$ entón $h(t)\Delta t \approx \mathbb{P}(t < T \leq t + \Delta t \mid T > t)$, que é aproximadamente a probabilidade de fallo no intervalo $(t, t + \Delta t]$. Polo tanto, podemos interpretar a función de risco como unha medida de propensión ao fallo nun instante t . Ademais, aplicando definicións básicas de probabilidade condicionada podemos relacionala coa función de densidade e supervivencia da seguinte forma:

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t \mid T > t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t)}{\Delta t} \frac{1}{\mathbb{P}(T > t)} = \frac{f(t)}{S(t)} \end{aligned} \quad (1.7)$$

A función de risco pode ter moitas formas, coa restricción de que ten que ser non negativa, i.e., $h(t) \geq 0$. En moitas investigacións pode ser máis doado ter unha idea de como é o risco de fallo nunha poboación, polo que esta función pode servir de guía para propoñer un modelo paramétrico para a distribución de T . Na figura 1.2 podemos ver varios exemplos de curvas de risco.

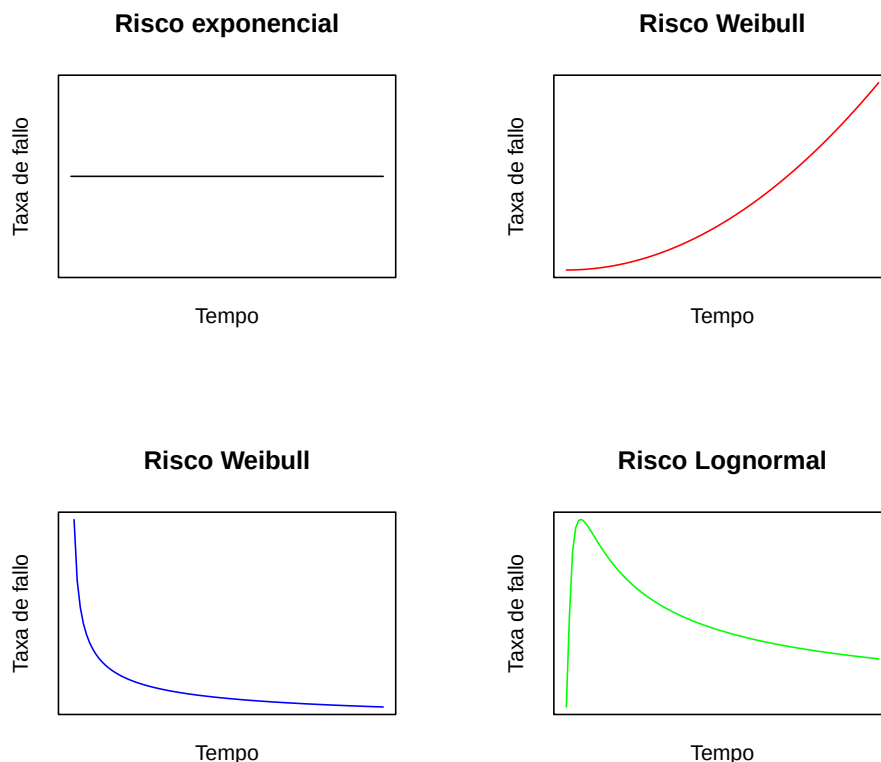


Figura 1.2: Exemplos de curvas de risco correspondentes coas distribucións $\text{Exp}(1)$, $\text{Weibull}(3,1)$ (vermella), $\text{Weibull}(\frac{1}{2},1)$ (azul) e $\text{Lognormal}(2,1)$.

Nestes exemplos podemos observar diferentes comportamentos da taxa de fallo. No primeiro caso, $\text{Exponencial}(\lambda = 1)$, vemos que a curva de risco permanece constante, o cal quere dicir que o fallo ten a mesma probabilidade de producirse ao longo do tempo. No seguinte caso, $\text{Weibull}(3,1)$, a función de risco é monótona crecente, o que indica que o fallo faise máis probable co paso do tempo. Polo tanto, esta curva de risco aporta unha idea de envellecemento ou desgaste na poboación de estudo. O caso contrario é o exemplo da $\text{Weibull}(\frac{1}{2},1)$, no que a curva de risco é monótona decrecente, e polo tanto dá unha idea de ‘rexuvenecemento’ na poboación. A última curva de risco da figura, procedente dunha $\text{Lognormal}(2,1)$, mostra o caso dunha poboación na que a probabilidade de fallar crece moi rápido nos primeiros instantes de vida, alcanzando un máximo a partir do cal este risco descende progresivamente. Ademais, na figura 1.3 móstrase unha curva de risco frecuente en moitos estudos e que describe o funcionamento de moitos procesos, incluída a vida humana, denominada ás veces pola súa forma *curva de bañeira*. Vemos que é unha ‘combinación’ dos casos anteriores. No inicio, o risco de fallo é alto (mortalidade infantil), pero co paso do tempo este risco redúcese sendo constante ata certo punto (mortes accidentales ou aleatorias), a partir do cal aumenta de novo (envellecemento).

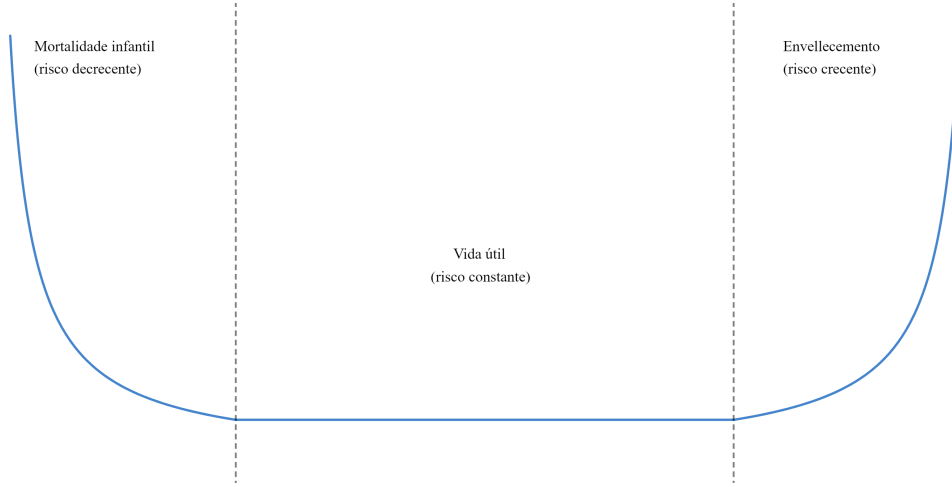


Figura 1.3: Curva de bañera

Para rematar, de forma análoga a como se define a función de distribución dunha variable aleatoria en función da súa densidade, podemos definir a *función de risco acumulado*, $H(t)$, da seguinte forma:

$$H(t) = \int_0^t h(x) dx \quad (1.8)$$

1.3. Función de esperanza de vida residual

A última función de interese na análise de supervivencia que imos ver neste capítulo é a función de esperanza de vida residual, $mrl(t)$ polas súas siglas en inglés, que mide a esperanza de vida restante para un individuo en tempo t . Polo tanto:

$$\begin{aligned}
 mrl(t) &= \mathbb{E}(T - t \mid T > t) \\
 &= \int_0^\infty (x - t) f(x \mid x > t) dx \\
 &= \int_0^\infty (x - t) \frac{f(x)}{S(t)} \mathbb{I}_{(t, \infty)}(x) dx \\
 &= \frac{\int_t^\infty (x - t) f(x) dx}{S(t)}
 \end{aligned} \quad (1.9)$$

Agora, empregando (1.4) e integrando por partes chegamos a que:

$$mrl(t) = \frac{\int_t^\infty (x-t)f(x)dx}{S(t)} = \frac{\int_t^\infty S(x)dx}{S(t)} \quad (1.10)$$

Vendo isto, podemos deducir que a esperanza de vida, $\mu = mrl(0)$, correspóndese coa área total baixo a curva de supervivencia.

$$\mu = \mathbb{E}(T) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt = mrl(0) \quad (1.11)$$

Traballando no contexto da análise de supervivencia ás veces vai resultar máis interesante calcular a mediana en lugar da media (esperanza de vida), xa que nos vai aportar máis información, sobre todo cando se traballa con distribucións moi nesgadas. Polo tanto, para calcular o cuantil p , bastaría con resolver a ecuación $S(t_p) = 1 - p$, sendo a mediana o cuantil $t_{0,5}$.

1.4. Relacións entre as funcións

Para rematar, veremos as relacións entre as funcións introducidas, que nos permitirán obter unhas a través das outras, e polo tanto caracterizarán a variable T . Se consideramos unha función de supervivencia dada, $S(t)$, poderíamos obter a densidade correspondente empregando (1.4). Agora, tendo en conta isto e a expresión (1.7) tamén podemos obter a función de risco da seguinte forma:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log S(t)}{dt} \quad (1.12)$$

Por último, a ecuación (1.10) móstranos como obter a función de esperanza de vida residual a partir da supervivencia.

Feito isto, vexamos agora como obter a función de supervivencia a partir das outras e desta forma teremos cumprido o noso obxectivo. Se coñecemos a función de risco e integramos en (1.12) podemos concluir o seguinte:

$$h(t) = -\frac{d \log S(t)}{dt} \Rightarrow S(t) = \exp\left[-\int_0^t h(x)dx\right] = \exp[-H(t)] \quad (1.13)$$

Esta expresión permítenos tamén relacionar o risco e a densidade:

$$f(t) = h(t)S(t) \Rightarrow f(t) = h(t)\exp\left[-\int_0^t h(x)dx\right] \quad (1.14)$$

Para finalizar, en [9] podemos ver a seguinte expresión que relaciona $S(t)$ con $mrl(t)$;

$$S(t) = \frac{mrl(0)}{mrl(t)} \exp\left[-\int_0^t \frac{dx}{mrl(x)}\right] \quad (1.15)$$

Feitos todos estes cálculos, xa estamos en condicións para ver que efectivamente as tres funcións vistas (ademais da densidade e da distribución) caracterizan unha variable aleatoria. De aquí en diante veremos diferentes técnicas e procedementos para construír estimadores destas funcións que teñan en conta a posible censura que poden ter os datos dunha mostra. Comezaremos vendo os métodos de estimación paramétricos no próximo capítulo.

Capítulo 2

Métodos de estimación paramétricos

2.1. Introducción

Os modelos paramétricos son aqueles que supoñen unha distribución coñecida para a variable aleatoria que se está a estudar. Polo tanto, son especialmente útiles cando a distribución de procedencia dos datos é coñecida de antemán, ben sexa grazas á realización doutros estudos previos ou a aplicación de test de bondade de axuste (veremos un exemplo no capítulo 3). As funcións que caracterizan estes modelos están determinadas por un conxunto finito de parámetros, sendo estes parámetros os que se deben estimar para atopar a distribución máis axeitada aos datos. Existen diferentes métodos de estimación dos parámetros, coma o método dos momentos ou o de máxima verosimilitude (o cal veremos nas seguintes seccións).

No contexto da análise de supervivencia, calquera distribución con soporte os reais non negativos pode ser considerada como un modelo para o tempo de supervivencia. Sen embargo, hai unha serie de modelos que son os máis frecuentes, como os modelos exponencial e Weibull que veremos neste capítulo. En canto á estimación dos parámetros, a presenza de datos censurados vailos obrigar a adaptar os métodos de estimación para ter en conta a censura. Nun primeiro razoamento, poderíamos plantexarnos eliminar os datos censurados da mostra ou interpretalos coma fallos, sen embargo, isto produciría un sesgo cara a valores máis pequenos da variable, o cal nos levaría a subestimar a distribución teórica da que proceden as observacións. Tendo isto en conta, precisamos construír novos estimadores que teñan en conta a censura, para o cal comezamos vendo como adaptar o método de máxima verosimilitude na seguinte sección.

2.2. Función de verosimilitude baixo censura

Un concepto clave na inferencia paramétrica é o de *verosimilitude*, no que se basea o método de *estimación por máxima verosimilitude*, que será o que empreguemos ao longo deste capítulo. Este método consiste en estimar o parámetro descoñecido, θ , escollendo o valor do parámetro que dá lugar á maior probabilidade para a mostra observada. Ese valor de θ será o máis verosímil, pois é o que resulta máis coherente coa mostra recollida. Se temos unha mostra de observacións independentes da variable de interese $\{T_1, \dots, T_n\}$ e denotamos a súa función de densidade f_θ , entón a función de verosimilitude adopta a forma:

$$\mathcal{L}(\theta) = f_\theta(t_1, \dots, t_n) = \prod_{i=1}^n f_\theta(t_i) \quad (2.1)$$

Así, un estimador $\hat{\theta} = \hat{\theta}(T_1, \dots, T_n)$ denomínase estimador de máxima verosimilitude (**EMV**) de θ se maximiza a función de verosimilitude, é dicir se

$$\mathcal{L}(\hat{\theta}) = \max_{\theta \in \Theta} \mathcal{L}(\theta) \quad (2.2)$$

Dado que a función logaritmo é monótona crecente, moitas veces á hora de derivar resulta máis cómodo traballar con $\ell(\theta) = \log(\mathcal{L}(\theta))$, denominada *log-verosimilitude*, xa que os valores que maximizan esta función van ser os mesmos.

Este procedemento que acabamos de introducir é sempre válido cando as mostras son completas. Sen embargo, se a mostra está formada por observacións censuradas non poderemos determinar $f_\theta(t^+)$ (sendo t^+ un dato censurado) polo que a función de verosimilitude introducida non servirá para calcular o EMV tendo en conta a censura. Para amañar este problema e xeneralizar a función de verosimilitude no caso de ter datos censurados (pola dereita), imos ter en conta que a información que nos aportan estes datos t^+ é que $T > t^+$. Polo tanto, a verosimilitude nas observacións censuradas podemos definila como $S_\theta(t^+)$, é dicir, a supervivencia nese instante, que é a $\mathbb{P}(T > t^+)$. Polo tanto, a función de verosimilitude dunha mostra censurada pola dereita defínese como

$$\mathcal{L}(\theta) = \prod_{i=1}^n \mathcal{L}_i(\theta) = \prod_{t_i \in F} f_\theta(t_i) \prod_{t_i \in C} S_\theta(t_i) \quad (2.3)$$

onde F é o conxunto de fallos observados e C é o conxunto de tempos censurados. Equivalentemente, se consideramos os datos como pares (t_i, δ_i) (sendo δ o indicador da censura) a función de verosimilitude pode expresarse da seguinte forma máis compacta (que será a que empreguemos de aquí en diante):

$$\mathcal{L}(\theta) = \prod_{i=1}^n f_{\theta}(t_i)^{\delta_i} S_{\theta}(t_i)^{1-\delta_i} \quad (2.4)$$

Feito isto, o proceso de estimación consiste en maximizar $\mathcal{L}(\theta)$ igual que no caso de traballar con mostrás completas. Algúns estimadores poderán obterse analiticamente, mentres que outros precisarán de utilizar métodos numéricos. Unha propiedade moi importante dos EMV é a invariancia funcional, que garante que o EMV de $g(\theta)$ pódese calcular aplicándolle g a $\hat{\theta}$, é dicir, $\widehat{g(\theta)} = g(\hat{\theta})$ sempre que g sexa unha función bixectiva con valores reais. Isto permítenos estimar $S(t)$, $h(t)$, a esperanza de vida, μ , ou calquera outra medida relacionada con T obtida a través dunha transformación bixectiva dos parámetros. Ademais, a normalidade e insesgadez asintóticas dos EMV permitirannos construír intervalos de confianza tanto para os parámetros como para as transformacións dos mesmos (para máis detalles sobre as propiedades asintóticas dos EMV ver [7]). A continuación introduciremos distintos modelos paramétricos habituais na Análise de Supervivencia e, aplicando o visto, estimaremos os seus parámetros e construiremos intervalos de confianza tendo en conta a posible censura.

2.3. Modelo exponencial

A distribución máis simple en estudos de supervivencia é a distribución exponencial, $Exp(\lambda)$. Está caracterizada por unha taxa de risco constante $\lambda > 0$, é dicir, a súa función de risco é constante, independente do tempo t .

$$h(t) = \lambda \quad \text{con } 0 \leq t < \infty \quad (2.5)$$

O feito de que esta función sexa constante implica que o risco de presentar o fallo non cambia ao longo da vida do individuo, polo que as mortes ou fallos son eventos aleatorios independentes do tempo transcorrido. Isto está directamente relacionado coa propiedade de *ausencia de memoria* que ten esta distribución, pola que a ‘idade’ dun individuo non afecta a súa supervivencia no futuro, é dicir, non hai envellecemento nin desgaste na poboación. Empregando as relacións vistas na sección 1.4 podemos calcular o resto de funcións que caracterizan unha variable exponencial:

$$S(t) = \exp\left(-\int_0^t h(t)dt\right) = \exp\left(-\int_0^t \lambda dt\right) = e^{-\lambda t} \quad t \in [0, \infty) \quad (2.6)$$

$$F(t) = 1 - e^{-\lambda t} \quad t \in [0, \infty) \quad (2.7)$$

$$f(t) = \lambda e^{-\lambda t} \quad t \in [0, \infty) \quad (2.8)$$

Como podemos ver na figura 2.1, canto maior sexa λ máis grande será o risco e máis corta a supervivencia. A media e a varianza dunha distribución exponencial son $\frac{1}{\lambda}$ e $\frac{1}{\lambda^2}$, respectivamente.

Aínda así, como xa se mencionou, en distribucións asimétricas á dereita a mediana vai ser unha mellor medida de centralización, sendo neste caso o seu valor $\frac{\log(2)}{\lambda}$.

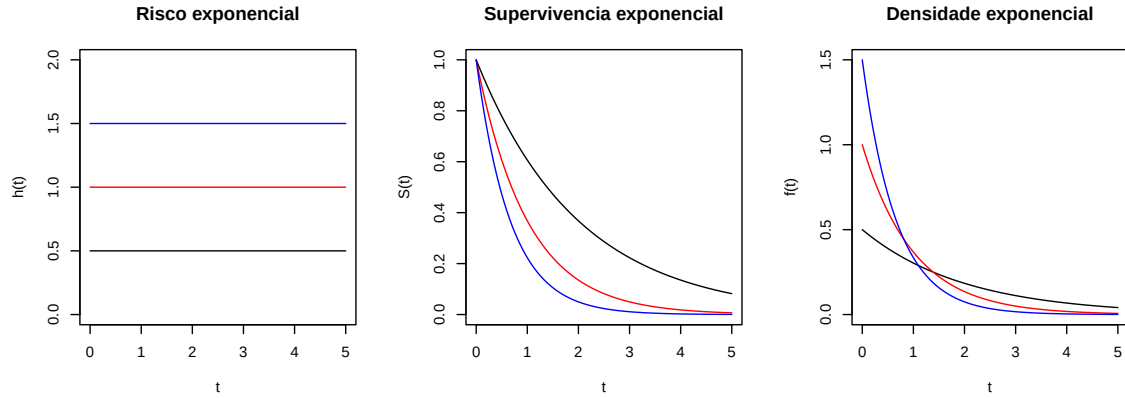


Figura 2.1: Funcións que caracterizan unha variable $Exp(0,5)$ (negro), $Exp(1)$ (vermello) e $Exp(1,5)$ (azul).

A continuación, veremos como calcular o estimador de máxima verosimilitude de λ con datos censurados. Sexa $((T_1, \delta_1), \dots, (T_n, \delta_n))$ unha mostra aleatoria simple procedente dunha distribución $Exp(\lambda)$. Polo visto anteriormente a función de verosimilitude ten a seguinte forma:

$$\mathcal{L}(\lambda) = \prod_{i=1}^n f_{\lambda}(t_i)^{\delta_i} S_{\lambda}(t_i)^{1-\delta_i} = \prod_{i=1}^n (\lambda e^{-\lambda t_i})^{\delta_i} e^{-\lambda t_i(1-\delta_i)} \quad (2.9)$$

De onde, tomando logaritmos,

$$\ell(\lambda) = \log(\mathcal{L}(\lambda)) = \sum_{i=1}^n \delta_i \log(\lambda) - \lambda t_i \delta_i - \lambda t_i (1 - \delta_i) \quad (2.10)$$

Derivando e igualando a cero obtemos o estimador de máxima verosimilitude $\hat{\lambda}$:

$$\begin{aligned} \ell'(\lambda) &= \sum_{i=1}^n \frac{\delta_i}{\lambda} - t_i \lambda_i - t_i(1 - \lambda_i) = \sum_{i=1}^n \frac{\delta_i}{\lambda} - \sum_{i=1}^n t_i = 0 \\ \hat{\lambda} &= \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i} \end{aligned} \quad (2.11)$$

Podemos ver que o EMV de λ correspóndese co número de observacións non censuradas dividido pola suma de todos os tempos (censurados e non censurados). Grazas a propiedade dos EMV de invariancia funcional podemos substituír $\hat{\lambda}$ en calquera función bixectiva de λ para obter o seu EMV. Así, por exemplo, obtéñense os estimadores $\hat{S}(t) = e^{-\hat{\lambda}t}$, $\hat{h}(t) = \hat{\lambda}$ ou $\hat{\mathbb{E}}[T] = \frac{1}{\hat{\lambda}}$.

Como xa se mencionou, os estimadores de máxima verosimilitude son asintoticamente normais e innesgados. Polo tanto, para construír intervalos de confianza (asintóticos) soamente necesitamos a estimación da varianza de $\hat{\lambda}$, a cal vén dada pola expresión $1/i(\lambda)$ onde $i(\lambda) = \frac{\sum_{i=1}^n \delta_i}{\lambda^2}$ é

a información observada de Fisher (para máis detalles sobre a distribución dos estimadores ver [1]). Substituíndo λ polo seu estimador obtemos:

$$\widehat{Var}(\hat{\lambda}) = \frac{\hat{\lambda}^2}{\sum_{i=1}^n \delta_i} \quad (2.12)$$

Polo tanto, $\hat{\lambda} \stackrel{a}{\sim} N(\lambda, \widehat{Var}(\hat{\lambda}))$ e un intervalo de confianza asintótico para λ de nivel $(1 - \alpha)$ pódese calcular da forma habitual como

$$\left(\hat{\lambda} - z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\lambda})}, \hat{\lambda} + z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\lambda})} \right) \quad (2.13)$$

onde $z_{\alpha/2}$ é o cuantil da distribución normal estándar tal que $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$. Ademais disto, tamén podemos calcular a distribución asintótica e intervalos de confianza para funcións do parámetro λ , como por exemplo a esperanza de vida $\mu = \frac{1}{\lambda}$. Grazas a propiedade de invariancia funcional sabemos que $\hat{\mu} \stackrel{a}{\sim} N(\mu, \widehat{Var}(\hat{\mu}))$, polo que soamente faltaría calcular a varianza de $\hat{\mu}$. Se denotamos $g(\lambda) = \frac{1}{\lambda}$, empregando o método delta vemos que

$$Var(\hat{\mu}) = Var[g(\hat{\lambda})] = |g'(\lambda)|^2 Var(\hat{\lambda}) = \frac{1}{\lambda^2 \sum_{i=1}^n \delta_i} = \frac{\mu^2}{\sum_{i=1}^n \delta_i} \quad (2.14)$$

E así xa podemos construír o intervalo de confianza para μ :

$$\left(\hat{\mu} - z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\mu})}, \hat{\mu} + z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\mu})} \right) \quad (2.15)$$

onde $\widehat{Var}(\hat{\mu}) = \frac{\hat{\mu}^2}{\sum_{i=1}^n \delta_i}$

A simplicidade da distribución exponencial resulta interesante para certas aplicacións, como pode ser a fiabilidade de sistemas, onde os fallos son eventos aleatorios, é dicir, a taxa de risco é constante. Aínda así, na maioría dos estudos de supervivencia a taxa de risco segue outros patróns, polo que imos necesitar maior flexibilidade nos modelos. Esta flexibilidade váinola dar a distribución Weibull, da cal a exponencial é un caso particular.

2.4. Modelo Weibull

A distribución Weibull é probablemente a máis común no modelado de datos de supervivencia, especialmente en aplicacións biomédicas, debido á súa maior flexibilidade. Foi proposta en 1939 polo enxeñeiro e matemático Waloddi Weibull nos seus traballos de fatiga de materiais, e máis tarde, en 1951, mostrou a súa aplicabilidade no campo da supervivencia [13]. Esta distribución xeneraliza a condición de risco constante que tiña a distribución exponencial, sendo a súa función de risco da seguinte forma:

$$h(t) = \alpha \lambda (\lambda t)^{\alpha-1} = \alpha \lambda^\alpha t^{\alpha-1} \quad t \geq 0 \quad \alpha, \lambda > 0 \quad (2.16)$$

Podemos ver que a distribución Weibull está caracterizada por dous parámetros positivos, o parámetro de forma α e o parámetro de escala λ . En vista da expresión (2.16) dedúcese facilmente que $h(t)$ é monótona crecente se $\alpha > 1$ e monótona decrecente se $\alpha < 1$. Cando $\alpha = 1$ o risco é constante, dando lugar á distribución $Exp(\lambda)$. Polo tanto, isto fai que a distribución Weibull sexa apropiada para modelar situacións onde o risco aumenta ou diminúe, o cal en determinados contextos podemos identificar cun envellecemento (desgaste) ou ‘rexuvenecemento’ da poboación. Por exemplo, no caso de estudar tempos de vida, se a poboación de estudo son pacientes de cancro a taxa de risco é crecente ao longo do tempo, sen embargo, no caso de que a poboación sexan os pacientes recentemente intervídos por unha cirurxía maior o risco diminúe co paso do tempo. Analogamente ao feito co modelo exponencial, podemos deducir o resto de funcións que caracterizan unha variable Weibull:

$$S(t) = \exp\left(-\int_0^t h(t)dt\right) = e^{-(\lambda t)^\alpha} \quad t \in [0, \infty) \quad (2.17)$$

$$F(t) = 1 - e^{-(\lambda t)^\alpha} \quad t \in [0, \infty) \quad (2.18)$$

$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha} \quad t \in [0, \infty) \quad (2.19)$$

A media e a varianza desta distribución son, respectivamente, $E(T) = \frac{\Gamma(1+1/\alpha)}{\lambda}$ e $Var(T) = \frac{1}{\lambda^2}[\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)]$, onde $\Gamma(x)$ é a función gamma, definida para todo $x > 0$ como $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$. Ademais, como a Weibull tamén é unha distribución asimétrica á dereita, unha medida de centralización usual é a mediana que se acada en $\frac{1}{\lambda}(\ln 2)^{1/\alpha}$. Na figura 2.2 represéntanse as funcións de densidade, risco e supervivencia dunha distribución Weibull para distintos valores de α e con $\lambda = 1$.

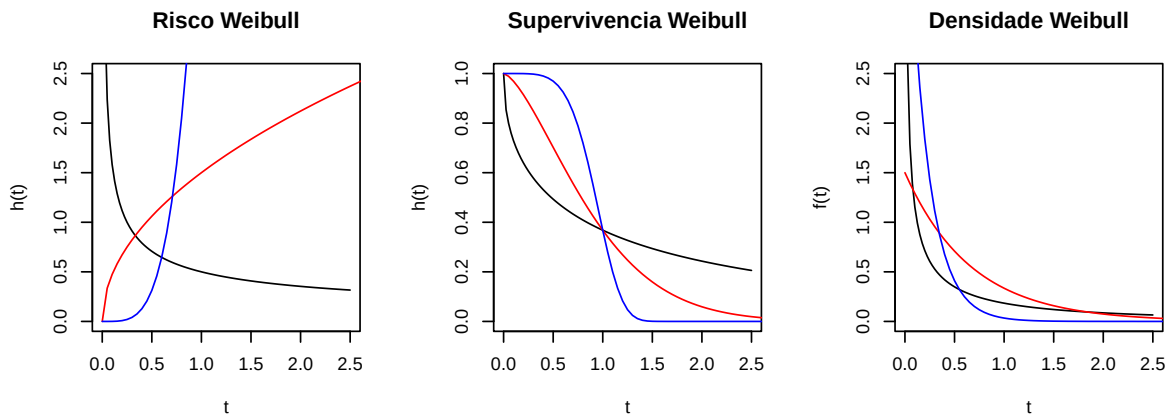


Figura 2.2: Funcións que caracterizan unha variable $Weibull(\alpha, 1)$, con $\alpha = 0,5$ (negro), $\alpha = 1,5$ (vermello) e $\alpha = 5$ (azul).

A continuación vexamos como obter os estimadores de máxima verosimilitude dos parámetros α e λ baixo censura. Sexa $((T_1, \delta_1), \dots, (T_n, \delta_n))$ unha mostra aleatoria simple procedente dunha distribución $Weibull(\alpha, \lambda)$, calculamos a súa función de verosimilitude (que neste caso vai ser función de dous parámetros):

$$\mathcal{L}(\alpha, \lambda) = \prod_{i=1}^n (\alpha \lambda^\alpha t_i^{\alpha-1} e^{-(\lambda t_i)^\alpha})^{\delta_i} e^{-(\lambda t_i)^\alpha (1-\delta_i)} = \prod_{i=1}^n (\alpha \lambda^\alpha t_i^{\alpha-1})^{\delta_i} e^{-(\lambda t_i)^\alpha} \quad (2.20)$$

Agora, tomando logaritmos obtemos a log-verosimilitude:

$$\begin{aligned} \ell(\alpha, \lambda) &= \sum_{i=1}^n \delta_i \log(\alpha \lambda^\alpha t_i^{\alpha-1}) - (\lambda t_i)^\alpha \\ &= \log \alpha \sum_{i=1}^n \delta_i + \alpha \log \lambda \sum_{i=1}^n \delta_i + (\alpha - 1) \sum_{i=1}^n \delta_i \log t_i - \lambda^\alpha \sum_{i=1}^n t_i^\alpha \end{aligned} \quad (2.21)$$

Ao estar estimando máis dun parámetro, a diferencia do caso exponencial, para maximizar a verosimilitude teremos que traballar coas derivadas parciais con respecto a α e λ e ver onde se anulan. Polo tanto, hai que resolver o sistema formado polas denominadas *ecuacións de verosimilitude*:

$$\left. \begin{aligned} \frac{\partial \ell(\alpha, \lambda)}{\partial \alpha} &= 0 \\ \frac{\partial \ell(\alpha, \lambda)}{\partial \lambda} &= 0 \end{aligned} \right\} \quad (2.22)$$

Se intentamos resolver o sistema neste caso, temos que

$$\frac{\partial \ell(\alpha, \lambda)}{\partial \alpha} = \frac{1}{\alpha} \sum_{i=1}^n \delta_i + \log \lambda \sum_{i=1}^n \delta_i + \sum_{i=1}^n \delta_i \log t_i - \lambda^\alpha \log \lambda \sum_{i=1}^n t_i^\alpha - \lambda^\alpha \sum_{i=1}^n t_i^\alpha \log t_i \quad (2.23)$$

$$\frac{\partial \ell(\alpha, \lambda)}{\partial \lambda} = \frac{\alpha}{\lambda} \sum_{i=1}^n \delta_i - \alpha \lambda^{\alpha-1} \sum_{i=1}^n t_i^\alpha \quad (2.24)$$

Igualando a 0 a expresión (2.24) obtemos o EMV de λ en función de α

$$\begin{aligned} \frac{\alpha}{\lambda} \sum_{i=1}^n \delta_i - \alpha \lambda^{\alpha-1} \sum_{i=1}^n t_i^\alpha &= 0 \implies \sum_{i=1}^n \delta_i = \lambda^\alpha \sum_{i=1}^n t_i^\alpha \implies \lambda^\alpha = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i^\alpha} \\ \implies \hat{\lambda} &= \sqrt[\alpha]{\frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i^\alpha}} \end{aligned} \quad (2.25)$$

Vemos que a expresión de $\hat{\lambda}$ depende de α . Polo tanto, substituíndo en (2.23) chegamos a,

$$\frac{1}{\alpha} \sum_{i=1}^n \delta_i + \sum_{i=1}^n \delta_i \log t_i - \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i^\alpha} \sum_{i=1}^n t_i^\alpha \log t_i = 0 \quad (2.26)$$

unha ecuación que soamente depende de α , pero que non é posible resolver analiticamente. Entón, para calcular $\hat{\alpha}$ é necesario empregar métodos numéricos coma o método de Newton-Raphson ou o de Nelder-Mead. Unha vez estimado o parámetro de forma, o de escala obtémolo substituindo na expresión (2.25). Estimados ambos parámetros, tamén podemos construír os seus respectivos intervalos de confianza empregando as propiedades asintóticas dos EMV. Desta forma, os intervalos de confianza asintóticos para os parámetros de escala e forma dunha distribución Weibull (baixo censura) de nivel $(1-\alpha)$ son da forma

$$\left(\hat{\alpha} - z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\alpha})}, \hat{\alpha} + z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\alpha})} \right) \quad (2.27)$$

$$\left(\hat{\lambda} - z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\lambda})}, \hat{\lambda} + z_{\alpha/2} \sqrt{\widehat{Var}(\hat{\lambda})} \right) \quad (2.28)$$

onde $\widehat{Var}(\hat{\alpha})$ e $\widehat{Var}(\hat{\lambda})$ se extraen dos elementos da diagonal da inversa da matriz de información de Fisher. Na seguinte sección faremos un estudo de simulación co obxectivo de ilustrar as propiedades vistas dos estimadores.

2.5. Estudo de simulación

Vexamos cun estudo de simulación realizado con  o cálculo de todas estas estimacións e como afecta a porcentaxe de censura. Para elo, empregamos a función **mle** da librería *stats4*. Esta función resolve o problema de minimizar a log-verosimilitude negativa (o cal é equivalente a maximizar a log-verosimilitude). Mediante 1000 mostras censuradas (censura aleatoria) para cada tamaño da mostra n procedentes dunha *Weibull*(5,1) imos analizar as propiedades dos estimadores dos parámetros, dos intervalos de confianza e o efecto que ten aumentar a censura (de 25 % a 50 %) sobre estas estimacións (ver código no Apéndice A.3.3):

2.5.1. Propiedades dos estimadores

En base ao cadro 2.1 podemos comprobar como a estimación dos parámetros converge aos valores reais dos mesmos ao aumentar o tamaño da mostra, xa que vemos que tanto o nesgo coma o erro típico diminúen, tendendo a 0. Este comportamento refléxase tamén no erro cadrático medio (ECM), polo que dicimos que os estimadores son consistentes en media cadrática. Por outra banda, podemos apreciar que ao aumentar a censura os estimadores teñen nesgo e erro típico máis grandes (e polo tanto tamén ECM), producindo peores estimacións debido ao maior descoñecemento dos datos (sobre todo en mostras pequenas).

25 % censura						
n	Nesgo($\hat{\alpha}$)	$\sqrt{\widehat{Var}(\hat{\alpha})}$	ECM($\hat{\alpha}$)	Nesgo($\hat{\lambda}$)	$\sqrt{\widehat{Var}(\hat{\lambda})}$	ECM($\hat{\lambda}$)
10	0.97526	1.95216	4.76209	0.00972	0.07198	0.00528
30	0.31002	0.94399	0.98724	0.00147	0.04278	0.00183
300	0.02708	0.25307	0.06478	0.00060	0.01379	0.00019
3000	0.00353	0.08274	0.00686	0.00003	0.00421	0.00002
50 % censura						
n	Nesgo($\hat{\alpha}$)	$\sqrt{\widehat{Var}(\hat{\alpha})}$	ECM($\hat{\alpha}$)	Nesgo($\hat{\lambda}$)	$\sqrt{\widehat{Var}(\hat{\lambda})}$	ECM($\hat{\lambda}$)
10	1.44182	3.33090	13.17373	0.00844	0.09681	0.00944
30	0.36805	1.21499	1.61166	0.00120	0.05040	0.00254
300	0.03730	0.32491	0.10696	0.00140	0.01703	0.00029
3000	0.00581	0.10381	0.01081	0.00021	0.00551	0.00003

Cadro 2.1: Estimacións do nesgo, erro típico e erro cadrático medio dos estimadores dos parámetros cunha censura do 25 % e 50 %

2.5.2. Propiedades dos intervalos de confianza

Ao longo deste capítulo temos construído intervalos de confianza para os estimadores mediante o método asintótico, o cal se basea na distribución asintótica do estimador de máxima verosimilitude. Sen embargo, existen outros métodos para construír intervalos de confianza, como pode ser o método que emprega o que se coñece como *perfil de verosimilitude*, o cal resulta máis robusto traballando con mostras pequenas. Sexa $\Theta = (\theta_1, \dots, \theta_k)$ o espazo de parámetros, defínese o perfil de verosimilitude (profile likelihood en inglés) dun parámetro θ_j como

$$PL(\theta_j) = \max_{\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k} \mathcal{L}(\theta_1, \dots, \theta_{j-1}, \theta_j, \theta_{j+1}, \dots, \theta_k) \quad (2.29)$$

onde $\mathcal{L}$ fai referencia á función de verosimilitude 2.4. Polo tanto, estamos a maximizar a verosimilitude co valor do parámetro θ_j fixado. Mediante o perfil de verosimilitude podemos construír intervalos de confianza, os cales conterán os valores máis verosímiles dos parámetros, é dicir, os que teñan un perfil de verosimilitude máis grande. Polo tanto, tómasse como intervalo de confianza para θ_j de nivel $(1 - \alpha)$ o seguinte conxunto:

$$\{\theta_j : 2(\mathcal{L}(\hat{\Theta}) - PL(\theta_j)) < \chi_{1,\alpha}^2\} \quad (2.30)$$

onde $\hat{\Theta}$ é o estimador de máxima verosimilitude dos parámetros e $\chi_{1,\alpha}^2$ é o cuantil $(1 - \alpha)$ da distribución chi-cadrado cun grao de liberdade. En xeral, estes intervalos serán asimétricos, a diferenza dos construídos mediante o método asintótico. A función *confint* de  permítenos obter este tipo de intervalos, os cales incluímos xunto aos asintóticos no noso estudo de simulación.

Intervalos por perfil de verosimilitude				
25 %	$\hat{\alpha}$		$\hat{\lambda}$	
n	Cobertura	Lonx. media	Cobertura	Lonx. media
10	0.937	6.88890	0.935	0.32393
30	0.930	3.46309	0.940	0.17199
300	0.959	1.02313	0.939	0.05264
3000	0.947	0.32164	0.954	0.01659
50 %	$\hat{\alpha}$		$\hat{\lambda}$	
n	Cobertura	Lonx. media	Cobertura	Lonx. media
10	0.914	9.46189	0.929	0.44844
30	0.943	4.29943	0.964	0.22312
300	0.935	1.25773	0.945	0.06584
3000	0.941	0.39513	0.947	0.02070

Cadro 2.2: Cobertura e lonxitude media dos intervalos por perfil de verosimilitude construídos para α e λ cunha censura do 25 % e 50 %

Intervalos asintóticos				
25 %	$\hat{\alpha}$		$\hat{\lambda}$	
n	Cobertura	Lonx. media	Cobertura	Lonx. media
10	0.959	6.86132	0.917	0.27602
30	0.949	3.46437	0.937	0.16311
300	0.962	1.02327	0.938	0.05237
3000	0.947	0.32164	0.954	0.01658
50 %	$\hat{\alpha}$		$\hat{\lambda}$	
n	Cobertura	Lonx. media	Cobertura	Lonx. media
10	0.953	9.36289	0.891	0.36832
30	0.956	4.30045	0.967	0.20713
300	0.938	1.25800	0.941	0.06534
3000	0.943	0.39513	0.947	0.02068

Cadro 2.3: Cobertura e lonxitude media dos intervalos asintóticos construídos para α e λ cunha censura do 25 % e 50 %

A partir dos dous tipos de intervalos calculamos a cobertura (proporción de intervalos que conteñen o valor real do parámetro) e a lonxitude media dos mesmos para os distintos tamaños da mostra. Nos cadros 2.2 e 2.3 podemos ver os resultados obtidos. En ambos casos, vemos como

os resultados son coherentes coa converxencia dos estimadores, xa que a cobertura está sempre preto do 95 % e a lonxitude media redúcese cando aumenta n , apreciándose moi pouca diferenza entre os dous tipos de intervalos. Ademais, da mesma forma que nas estimacións puntuais, o aumento da censura tamén afecta aos intervalos de confianza, para os cales se pode observar unha maior lonxitude media, debida á menor información que nos aportan os datos.

Se ben é certo que a distribución Weibull é moi útil e empregada na análise de supervivencia, existen outros modelos alternativos que tamén teñen aplicación neste contexto. Este é o caso da distribución Lognormal, da cal se mostrou o comportamento do seu risco na figura 1.2, ou a distribución Gompertz. Estes e outros modelos adicionais poden verse máis a fondo en [10]. No seguinte capítulo introducimos os métodos de estimación non paramétricos.

Capítulo 3

Métodos de estimación non paramétricos

Cando os tempos de supervivencia non seguen unha distribución teórica ou non é posible axustarlle unha, os métodos non paramétricos son máis eficientes. Ademais, empregar estes métodos pode ser de axuda para posteriormente asignarlle unha distribución aos datos.

Como xa se mencionou no capítulo 2, a función de supervivencia empírica (1.5) non é aplicable no caso de traballar con mostras censuradas. Ademais, ao igual que no caso paramétrico, se eliminamos os datos censurados ou os interpretamos como fallos estaremos subestimando a curva de supervivencia, o cal non é adecuado. Polo tanto, ao longo deste capítulo veremos os dous métodos de estimación non paramétrica da función de supervivencia censurada máis habituais, as táboas de vida e o estimador de Kaplan-Meier.

3.1. Táboas de vida

A estimación da función de supervivencia mediante táboas de vida, tamén coñecidas como táboas actuariais, é unha das técnicas máis antigas para analizar e describir a supervivencia e mortalidade dunha poboación de estudo. Estas táboas son ferramentas comúns en moitos ámbitos, como as investigacións médicas, demográficas, sociolóxicas e mesmo políticas. Cando se está a traballar con mostras cun gran número de observacións este método é o máis recomendable, xa que proporciona unha curva de supervivencia clara e sintetiza nunha táboa de fácil interpretación toda a información.

O primeiro paso na construción dunha táboa de vida é dividir o tempo de supervivencia nunha serie de intervalos de tempo predeterminados $[t_j, t_{j+1})$, empregando como extremos $0 \leq$

$t_0 < t_1 \dots < t_J \leq \infty$. Feito isto, a probabilidade de supervivencia en cada extremo t_j calcúlase mediante o produto das seguintes probabilidades condicionadas:

$$S(t_j) = \mathbb{P}(T > t_j) = \mathbb{P}(T > t_0) \cdot \mathbb{P}(T > t_1 \mid T > t_0) \cdot \dots \cdot \mathbb{P}(T > t_j \mid T > t_{j-1}) \quad (3.1)$$

Polo tanto, vemos que a idea que está detrás deste estimador baséase en que a probabilidade de sobrevivir a un intervalo pode ser vista como a probabilidade de sobrevivir ao intervalo anterior e entón sobrevivir un intervalo máis. Desta forma, a posible censura estase tendo en conta á hora de estimar. As diferentes columnas das táboas de vida conteñen a seguinte información:

- Intervalos de tempo (prefixados) $I_j = [t_j, t_{j+1})$. Nestes intervalos agrúpanse todas as observacións da mostra, censuradas e non censuradas.
- Número de datos censurados no intervalo j -ésimo, c_j . As observacións censuradas inclúen aquelas ás que se lles perde o seguemento e as que aínda non fallaron ao rematar o estudo.
- Número de fallos ou mortes en cada intervalo, d_j .
- Individuos vivos no comezo do intervalo, N_j . O número de individuos no comezo do primer intervalo, N_0 , coincide co tamaño da mostra. Podemos obter este valor para o resto de intervalos tendo en conta que $N_j = N_{j-1} - c_{j-1} - d_{j-1}$.
- Individuos a risco, n_j . Este dato correspóndese co número de datos susceptibles de fallar ou ser censurados no intervalo I_j . Coincide con N_j , salvo que se asume que as observacións censuradas distribúense de forma uniforme ao longo do intervalo, polo tanto considérase que soamente están expostas ao risco a metade do intervalo. Desta forma $n_j = N_j - \frac{c_j}{2}$.
- Proporción de fallos condicionada $\hat{q}_j$. É a estimación da probabilidade de fallar no intervalo I_j condicionada a estar vivo ao comezo do mesmo. Vén dada por $\hat{q}_j = \frac{d_j}{n_j}$.
- Proporción de superviventes condicionada $\hat{p}_j = 1 - \hat{q}_j$.
- Proporción de superviventes acumulada $\hat{S}(t_j)$. Trátase dun estimador da función de supervivencia no instante t_j . Tense que $\hat{S}(t_0) = 1$ e $\hat{S}(t_j) = \hat{p}_{j-1} \hat{S}(t_{j-1})$.

No cadro 3.1 vemos un exemplo dunha táboa de vida extraída de [4]. Nela reflíctense os datos de supervivencia de 2418 varóns doentes de anxina de peito e as estimacións realizadas a partir destes datos. Os tempos de supervivencia, calculados dende o momento do diagnóstico, foron agrupados en 16 intervalos.

$[t_j, t_{j+1})$	c_j	d_j	N_j	n_j	$\hat{q}_j$	$\hat{p}_j$	$\hat{S}(t_j)$
[0, 1)	0	456	2418	2418	0,1886	0,8114	1,0000
[1, 2)	39	226	1962	1942,5	0,1163	0,8837	0,8114
[2, 3)	22	152	1697	1686	0,0902	0,9098	0,7170
[3, 4)	23	171	1523	1511,5	0,1131	0,8869	0,6524
[4, 5)	24	135	1329	1317	0,1025	0,8975	0,5786
[5, 6)	107	125	1170	1116,5	0,1120	0,8880	0,5193
[6, 7)	133	83	938	871,5	0,0952	0,9048	0,4611
[7, 8)	102	74	722	671	0,1103	0,8897	0,4172
[8, 9)	68	51	546	512	0,0996	0,9004	0,3712
[9, 10)	64	42	427	395	0,1063	0,8937	0,3342
[10, 11)	45	43	321	298,5	0,1441	0,8559	0,2987
[11, 12)	53	34	233	206,5	0,1646	0,8354	0,2557
[12, 13)	33	18	146	129,5	0,1390	0,8610	0,2136
[13, 14)	27	9	95	81,5	0,1104	0,8896	0,1839
[14, 15)	23	6	59	47,5	0,1263	0,8737	0,1636
[15, ∞)	0	30	30	30	1,0000	0,0000	0,1429

Cadro 3.1: Táboa de vida de 2418 homes con anxina de peito

A expresión máis empregada para a estimación da varianza deste estimador é a dada pola fórmula de Greenwood:

$$\widehat{Var}[\hat{S}(t_j)] = \hat{S}(t_j)^2 \sum_{i=1}^{j-1} \frac{\hat{q}_i}{n_i \hat{p}_i} \quad (3.2)$$

Con isto, poderíamos construír intervalos de confianza (asumindo normalidade asintótica) de forma análoga á vista para os métodos de estimación paramétricos. Polo tanto, un intervalo de confianza asintótico de nivel $(1-\alpha)$ para a función de supervivencia nun instante t é da forma

$$\left(\hat{S}(t) - z_{\alpha/2} \sqrt{\widehat{Var}[\hat{S}(t)]}, \hat{S}(t) + z_{\alpha/2} \sqrt{\widehat{Var}[\hat{S}(t)]} \right) \quad (3.3)$$

onde $\hat{S}(t)$ é o estimador obtido a partir da táboa de vida.

Cando cada un dos intervalos nos que está dividido o tempo de supervivencia contén unha única observación temos o caso *límite* das táboas de vida (a lonxitude dos intervalos tende a 0). Na seguinte sección estudamos este caso, que é de especial interese xa que dá lugar ao coñecido como estimador de Kaplan-Meier.

3.2. Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, tamén coñecido como o estimador *produto-límite*, foi proposto en 1958 [8] polos matemáticos Edward L. Kaplan e Paul Meier e trátase do estimador non paramétrico máis habitual para estimar a función de supervivencia. Antes de introduci-lo formalmente, vexamos cun exemplo sinxelo a idea que hai detrás deste estimador, inspirándonos no exemplo que os propios Kaplan e Meier empregan no seu artigo.

Exemplo 3.1. Supoñamos que 10 pacientes se unen a un ensaio clínico a comezos do ano 2018; durante ese ano 6 dos pacientes falecen e 4 sobreviven. A finais de ano 20 pacientes adicionais se incorporan ao estudo. En 2019, 3 dos doentes que comezaron en 2018 e 15 dos que entraron máis tarde falecen, deixando polo tanto 1 e 5 superviventes no momento de rematar o estudo. Vexamos como estimar $S(2)$.

Unha posible estimación sería empregando a mostra reducida, $S^*(2) = \frac{1}{10} = 0,1$, que ignora os 20 pacientes incorporados en 2019 que soamente foron seguidos durante un ano. Sen embargo, Kaplan e Meier pensaron que este grupo de pacientes tamén podían contribuír na estimación, facendo a seguinte interpretación:

$$S(2) = \mathbb{P}(T > 2) = \mathbb{P}(T > 2 \mid T > 1) \cdot \mathbb{P}(T > 1)$$

É dicir, a probabilidade de sobrevivir dous anos pode ser vista como a probabilidade de sobrevivir un ano e entón sobrevivir un ano máis. Polo tanto, vemos que a idea que está detrás deste estimador é a mesma que a do estimador das táboas de vida, xa que a única diferenza é que neste caso a lonxitude dos intervalos tende a 0 (de ahí a denominación *produto-límite*), contendo cada un un único dato. Estimando en (3.4) as probabilidades mediante as proporcións mostrais obtense o estimador de Kaplan-Meier (KM) $\hat{S}(2)$. Polo tanto, neste caso temos que $\hat{S}(2) = \frac{1}{4} \cdot \frac{4+5}{10+20} = 0,25 \cdot 0,3 = 0,075$. Desta forma, os datos censurados están sendo tidos en conta para estimar a curva de supervivencia.

En vista do exemplo, este procedemento pódese xeneralizar da seguinte forma: en base a unha mostra de n individuos que contén $k \leq n$ tempos non censurados $t_{(1)} < t_{(2)} < \dots < t_{(k)}$, a estimación da supervivencia nun instante t é o produto das estimacións das probabilidades de supervivencia ata o tempo t , é dicir,

$$\hat{S}(t) = \prod_{\{i \mid t_{(i)} \leq t\}} \hat{\mathbb{P}}(T > t_{(i)} \mid T > t_{(i-1)}) = \prod_{\{i \mid t_{(i)} \leq t\}} \frac{\hat{\mathbb{P}}(T > t_{(i)})}{\hat{\mathbb{P}}(T > t_{(i-1)})} \quad (3.4)$$

Para tempos anteriores a $t_{(1)}$ o estimador vale 1, mentres que en $t_{(1)}$ temos que $\hat{S}(t_{(1)}) = \hat{\mathbb{P}}(T > t_{(1)})$. Partindo de aquí e introducindo algo de notación chegaremos á expresión máis habitual do estimador de Kaplan-Meier de $S(t)$. Para cada fallo $t_{(i)}$ definimos:

- d_i : número de individuos que experimentan o fallo nese instante. Vale 1 salvo que haxa empates en $t_{(i)}$.
- n_i : número de individuos a risco nese instante, é dicir, aqueles que aínda non fallaron nin foron censurados en tempos anteriores. As observacións con tempos exactamente iguais a $t_{(i)}$ cóntanse coma individuos a risco en n_i .

Polo tanto, o estimador de Kaplan-Meier da función de supervivencia defínese como:

$$\hat{S}(t) = \prod_{\{i \mid t_{(i)} \leq t\}} \frac{n_i - d_i}{n_i} = \prod_{\{i \mid t_{(i)} \leq t\}} \left(1 - \frac{d_i}{n_i}\right) \quad (3.5)$$

Á vista da expresión (3.5) vemos que o estimador de Kaplan-Meier é unha función escalonada, constante entre dous tempos observados e continua pola dereita. Ademais, vemos que o valor do estimador non cambia nas observacións censuradas, as cales si inflúen nos n_i e polo tanto no tamaño das discontinuidades de salto. Ilustraremos cun exemplo simple o cálculo e representación deste estimador.

Exemplo 3.2. Supoñamos un ensaio clínico no que se quere estudar o tempo de remisión de enfermos tratados contra o cancro, no cal participan 10 persoas. Ao longo do estudo seis pacientes recaen aos 3, 6.5, 6.5, 10, 12 e 15 meses; a outro pérdeselle o seguemento aos 8.4 meses; e os tres restantes seguen en remisión ao final do estudo despois de 4, 5 e 10 meses. No cadro 3.1 móstrase o cálculo do estimador de Kaplan-Meier a partir da mostra censurada mencionada. Neste cadro distínguense 4 columnas: na primeira colócanse os tempos de remisión, nas dúas seguintes calcúlanse os n_i e d_i e na última calcúlase o valor do estimador de Kaplan-Meier. Vemos que nas filas correspondentes a tempos censurados (sinalados cun + no expoñente) omítense os cálculos xa que $d_i = 0$ e polo tanto o valor do estimador non varía neses tempos. Na figura 3.1 represéntase graficamente o estimador de Kaplan-Meier calculado.

A partir do estimador de Kaplan-Meier poderíamos estimar outras medidas de interese de forma doada. Por exemplo, poderíamos estimar a esperanza de vida, $\hat{\mu} = \int_0^\infty \hat{S}(t)dt$, calculando a área baixo a curva de Kaplan-Meier que non é máis que a suma de áreas de rectángulos (no exemplo 3.2 obtemos $\hat{\mu} = 7,416$ meses). Outro exemplo sería a estimación da mediana, $\hat{t}_{med} = \inf \{t : \hat{S}(t) \leq 0,5\}$, que no exemplo anterior é $\hat{t}_{med} = 10$ meses. Como último exemplo, poderíamos estimar funcións relacionadas coa supervivencia, como é o caso do risco acumulado, $\hat{H}(t) = -\ln \hat{S}(t)$, aínda que para esta función é máis habitual empregar directamente o estimador de Nelson-Aalen (ver [9, páx 86]).

Tempo de reimisión t	n_i	d_i	$\hat{S}(t) = \prod_{\{i \mid t_{(i)} \leq t\}} \frac{n_i - d_i}{n_i}$
3	10	1	$\frac{9}{10} = 0,900$
4 ⁺	-	-	-
5 ⁺	-	-	-
6,5	7	2	$\frac{9}{10} \cdot \frac{5}{7} = 0,643$
8,4 ⁺	-	-	-
10	4	1	$\frac{9}{10} \cdot \frac{5}{7} \cdot \frac{3}{4} = 0,482$
10 ⁺	-	-	-
12	2	1	$\frac{9}{10} \cdot \frac{5}{7} \cdot \frac{3}{4} \cdot \frac{1}{2} = 0,241$
15	1	1	0

Cadro 3.2: Cálculo do estimador de Kaplan-Meier para os datos do exemplo 3.2

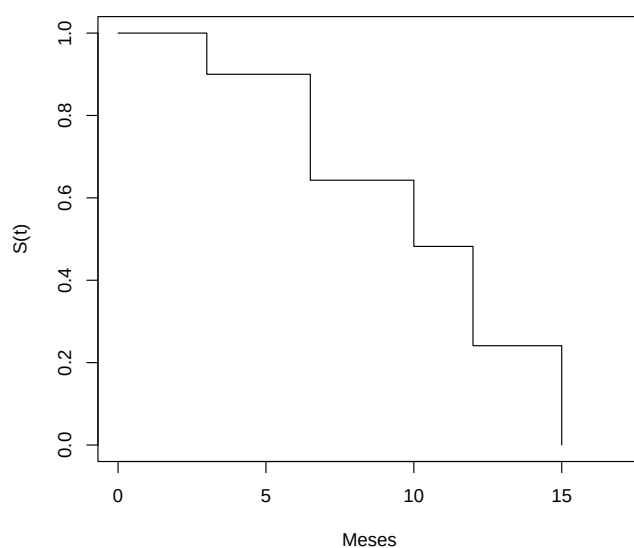


Figura 3.1: Representación gráfica do estimador de Kaplan-Meier para os datos do exemplo 3.2

Pódese demostrar que o estimador de Kaplan-Meier é o estimador non paramétrico de máxima verosimilitude da función de supervivencia [8, páx. 475], e polo tanto, coma todo estimador de máxima verosimilitude, baixo certas condicións de regularidade $\hat{S}(t)$ é asintoticamente innesgado e normal para cada t . Ao igual que para o estimador construído a partir das táboas de vida, mediante a fórmula de Greenwood podemos dar unha aproximación da varianza do estimador de Kaplan-Meier:

$$\widehat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{\{i \mid t_{(i)} \leq t\}} \frac{d_i}{n_i(n_i - d_i)} \quad (3.6)$$

onde os $t_{(i)}$ recordamos que son tempos de fallo.

Tendo isto en conta, estamos en condicións de construír intervalos de confianza para $S(t)$ dunha forma análoga á vista para os métodos de estimación paramétricos. Polo tanto, un intervalo de confianza asíntótico de nivel $(1 - \alpha)$ para a función de supervivencia nun instante t é da forma

$$\left(\hat{S}(t) - z_{\alpha/2} \sqrt{\widehat{Var}[\hat{S}(t)]}, \hat{S}(t) + z_{\alpha/2} \sqrt{\widehat{Var}[\hat{S}(t)]} \right) \quad (3.7)$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier.

Na seguinte saída de  (a súa obtención será explicada nunha sección posterior) podemos ver os intervalos de confianza do 95 % para a supervivencia nos tempos de remisión do exemplo 3.2 (ademais doutra información coma o erro típico ou o propio valor do estimador).

```
> summary(datos.km)

Call: survfit(formula = datos ~ 1, type = "kaplan-meier")

   time n.risk n.event survival std.err lower 95% CI upper 95% CI
   3.0     10      1    0.900  0.0949    0.7320         1
   6.5      7      2    0.643  0.1679    0.3852         1
  10.0      4      1    0.482  0.1877    0.2248         1
  12.0      2      1    0.241  0.1946    0.0496         1
  15.0      1      1    0.000   NaN      NA         NA
```

Aínda así, este tipo de intervalos non son os máis adecuados, especialmente traballando con mostrás pequenas, xa que os límites poden extenderse máis aló de 0 e 1. Unha primeira e rápida solución pode ser reemplazar por 0 ou 1 aqueles límites que foran sobrepasados, que é o que está a facer  cos extremos superiores dos intervalos de confianza da saída anterior. Outra solución máis adecuada é aplicarlle ao noso estimador unha transformación bixectiva h , calcular o intervalo de confianza para $h(S(t))$ e despois aplicarlle h^{-1} aos extremos deste respectivo intervalo. A transformación bixectiva máis empregada é $h(x) = \ln[-\ln x]$ (que non é máis que o logaritmo do risco acumulado) coñecida como transformación *log-log*. Podemos pedirlle a  que empregue este procedemento para calcular os intervalos de confianza, dando o seguinte resultado:

```
> summary(datos.km)

Call: survfit(formula = datos ~ 1, type = "kaplan-meier", conf.type = "log-log")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3.0	10	1	0.900	0.0949	0.4730	0.985
6.5	7	2	0.643	0.1679	0.2447	0.871
10.0	4	1	0.482	0.1877	0.1254	0.774
12.0	2	1	0.241	0.1946	0.0132	0.626
15.0	1	1	0.000	NaN	NA	NA

Na figura 3.2 podemos ver unha comparativa gráfica das distintas rexións de confianza para a curva de supervivencia.

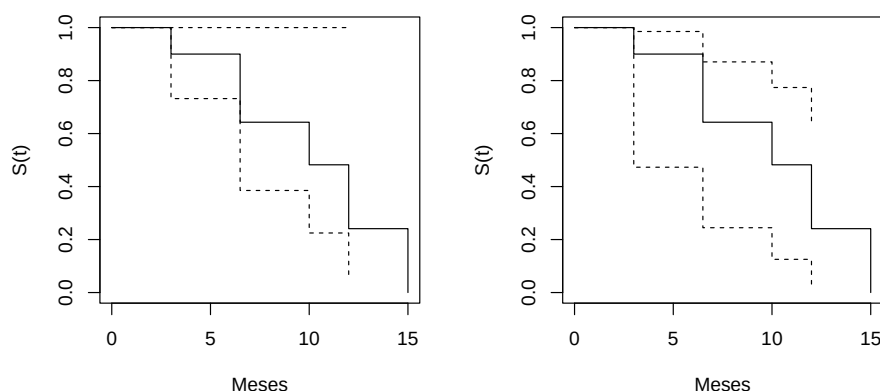


Figura 3.2: Estimador de Kaplan-Meier e rexións de confianza, sen transformar (esquerda) e empregando a transformación *log-log* (dereita)

O estimador de KM pode calcularse empregando unha formulación alternativa e equivalente á expresión (3.5), a cal pode resultar máis cómoda para traballar ou programar. Sexa n o número total de individuos na mostra, se ordeamos os seus tempos de supervivencia, censurados ou non, $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(n)}$, temos que o estimador de Kaplan-Meier pode expresarse como

$$\hat{S}(t) = \prod_{t_{(r)} \leq t} \left(\frac{n-r}{n-r+1} \right)^{\delta_{(r)}} \quad (3.8)$$

onde $\delta_{(r)}$ é o indicador de censura do tempo $t_{(r)}$. En ausencia de censura ambas expresións dan lugar á función de supervivencia empírica (1.5). Aínda que o método de Kaplan-Meier é o máis

empregado na análise de datos de supervivencia, ten certas características críticas que cómpre ter en conta:

- En primeiro lugar, se a maior observación t_{max} está censurada, $\hat{S}(t)$ non está ben definido para $t > t_{max}$, xa que non coñecemos cando este último supervivente habería experimentado o fallo, polo que $\lim_{t \rightarrow \infty} \hat{S}(t) \neq 0$. Este feito pode dar lugar a problemas á hora de estimar a esperanza de vida, a cal neste caso calcúlase en intervalos acotados. Distintas propostas foron feitas para amañar esta imprecisión, coma as non paramétricas de Efron [3] e Gill [5], ou a paramétrica de Brown, Hollander e Kowar [2].
- Por outra banda, na construción do estimador KM estase a asumir a independencia entre os tempos de censura e os de supervivencia, é dicir, o motivo da censura non está relacionado coa causa do fallo ou da morte. Cando non se verifica esta hipótese o estimador KM non é o máis apropiado, xa que pode dar lugar a estimacións incorrectas.

A continuación introduciremos e veremos como funciona o paquete *survival* de , o cal empregamos para realizar todas as saídas vistas anteriormente e que conta con moitas ferramentas para o análise de supervivencia.

3.3. O paquete *survival* de . Aplicación aos datos *pharmacoSmoking*

Dentro do software estatístico  podemos atopar unha gran cantidade de paquetes e librerías. Un destes é o paquete *survival* [12], que nos permite realizar dunha forma intuitiva moitas tarefas da análise de supervivencia, como a estimación da función de supervivencia mediante curvas Kaplan-Meier (cargámolo mediante a orde `library(survival)`).

En primeiro lugar, a maioría das funcións deste paquete traballan cos que se denominan obxectos de supervivencia, que o que fan é asociar a cada dato de supervivencia un indicador de censura. Estes obxectos constrúense mediante o comando `Surv()`, o cal precisa de dous argumentos; un vector de tempos de supervivencia e un vector binario que contén os indicadores de fallo (1) ou censura (0). Empregando os datos do exemplo 3.2 obtemos esta saída (vemos que as observacións censuradas aparecen indicadas cun +):

```
> t=c(3,4,5.7,6.5,6.5,8.4,10,10,12,15) # tempos de supervivencia
> c=c(1,0,0,1,1,0,1,0,1,1) # indicador de censura
> Surv(t,c)
[1] 3.0 4.0+ 5.7+ 6.5 6.5 8.4+ 10.0 10.0+ 12.0 15.0
```

Co noso obxecto de supervivencia definido xa podemos calcular o estimador de Kaplan-Meier da función de supervivencia e os seus respectivos intervalos de confianza. Isto realízase mediante a función `survfit()`, a cal admite os seguintes argumentos:

- **formula**: é unha expresión do tipo $y \sim x$, onde y debe ser un obxecto `Surv`. Se temos interese en calcular o estimador dependendo doutras covariables, estas irían no lado dereito da expresión. De non ser así colócase un `1` simplemente.
- **data**: *data frame* de onde proveñen os datos, no caso de que y e x non estean directamente cargados na memoria.
- **type**: tipo de estimador a calcular. No noso caso `'kaplan-meier'` (que é a opción por defecto), aínda que tamén se poderían calcular outros coma `'fleming-harrington'` ou `'fh2'`.
- **error**: tipo de estimación da varianza. Por defecto emprega a fórmula de Greenwood (3.6).
- **conf.type**: tipo de transformación utilizada para construír os intervalos de confianza, `'plain'`, `'log'` (por defecto) ou `'log-log'`. A primeira calcula os intervalos de confianza asíntóticos orixinais. A segunda emprega a transformación $g(x) = -\ln x$. A terceira emprega a transformación mencionada $h(x) = \ln[-\ln x]$. Coa opción `'none'` os intervalos de confianza non serán calculados.
- **conf.int**: nivel dos intervalos de confianza (95 % por defecto).

Vexamos o uso desta función aplicándolla a unha base de datos. Para iso empregaremos o *data set* `pharmacoSmoking`, que vén incluído no paquete `asaaur`, o cal recolle todas as bases de datos empregadas en [11]. Estes datos foron tomados co obxectivo principal de evaluar a eficacia de dous tratamentos deseñados para axudar a deixar de fumar, un consistente en combinar tres medicamentos e outro en empregar parches de nicotina. Por isto, neste estudo participaron 125 fumadores, que foron seguidos durante 6 meses, para os/as cales se rexistrou o tempo transcorrido en tratamento ata que volvían a fumar. Como non todos os participantes recaeron, a mostra contén datos censurados. Aínda que o conxunto de datos orixinal está formado por 14 variables, para o noso propósito seranos suficiente traballar con tres:

- **ttr**: tempo ata a recaída (en días).
- **relapse**: indicador de censura.
- **grp**: tipo de tratamento seguido (`combination` ou `patchOnly`).

Antes de comezar a análise, modificamos o conxunto de datos orixinal para quedarnos soamente coas variables que nos interesan. Ademais, aqueles tempos de recaída rexistrados coma 0 redefínímolos coma 0.5. Coa función `head()` mostramos as primeiras observacións.

```
> data(pharmacoSmoking)
> datos=pharmacoSmoking
> attach(datos)
> ttr[ttr==0]=0.5
> datos$ttr=ttr
> datos=datos[,c(2,3,4)]
> head(datos)
  ttr relapse      grp
1 182.0      0 patchOnly
2  14.0      1 patchOnly
3   5.0      1 combination
4  16.0      1 combination
5   0.5      1 combination
6 182.0      0 combination
```

A continuación, creamos o obxecto de supervivencia correspondente e estimamos a función de supervivencia xeral (de toda a mostra) con `survfit()`, que na saída por pantalla nos proporciona información coma o número de mortes (`events`), a mediana (`median`) e o seu intervalo de confianza do 95 % (para ver como se obtén consultar [11, páx. 30]).

```
> datos_s=Surv(datos$ttr,datos$relapse)
> datos.km=survfit(datos_s~1,conf.type='log-log')
> datos.km
Call: survfit(formula = datos_s ~ 1, data = datos, conf.type = "log-log")

      n events median 0.95LCL 0.95UCL
[1,] 125     89    49      21     65
```

Empregando agora a función `summary()` obtemos toda a información do estimador de Kaplan-Meier construído; os tempos de fallo (`time`), individuos a risco (`n.risk`), número de fallos (`n.event`), a estimación da función de supervivencia (`survival`), o seu erro típico (`std.error`) e os intervalos de confianza (coa configuración elixida, neste caso `log-log`). Ademais, empregando a función `plot()` podemos representar graficamente este estimador.

```
> summary(datos.km)

Call: survfit(formula = datos_s ~ 1, data = datos, conf.type = "log-log")

   time  n.risk  n.event  survival  std.err  lower 95% CI  upper 95% CI
0.5      125      12    0.904    0.0263    0.837    0.944
1.0      113       5    0.864    0.0307    0.790    0.913
2.0      108       6    0.816    0.0347    0.736    0.874
3.0      102       1    0.808    0.0352    0.727    0.867
4.0      101       3    0.784    0.0368    0.701    0.846
5.0       98       2    0.768    0.0378    0.684    0.833
6.0       96       1    0.760    0.0382    0.675    0.826
...
170.0      38       2    0.288    0.0405    0.212    0.369
```

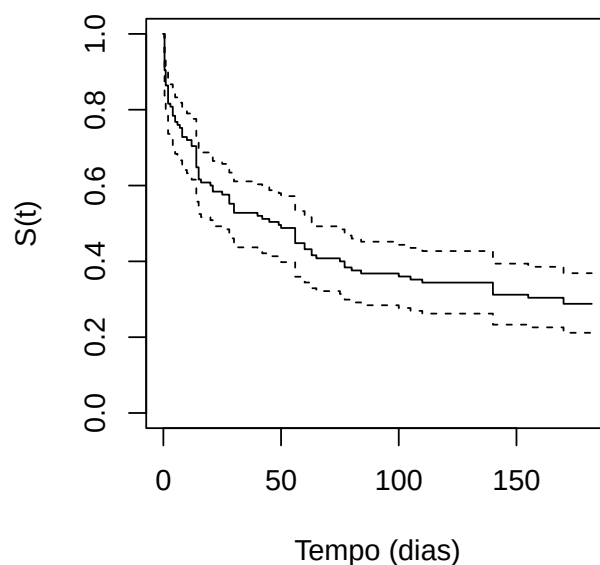


Figura 3.3: Estimador de Kaplan-Meier construído a partir dos datos *pharmacoSmoking*

Como podemos ver, estamos a traballar cun caso no que a observación máis grande está censurada, polo que o estimador de Kaplan-Meier non está ben definido a partir dese tempo (neste caso 182 días). Como xa se mencionou, isto pode dar problemas para estimar por exemplo a esperanza de vida μ . Unha alternativa pode ser calcular a esperanza de vida restrinxida ao máximo tempo supervivencia coa opción `print.rmean=T`. O resultado son 77.5 días para os

pacientes que volven a fumar antes dos 182 días.

```
> print(datos.km, print.rmean=T)
Call: survfit(formula = datos_s ~ 1, data = datos, conf.type = "log-log")

      n events rmean* se(rmean) median 0.95LCL 0.95UCL
[1,] 125      89  77.5      6.8     49      21      65
* restricted mean with upper limit = 182
```

Outra característica da función `survfit()` é a súa capacidade de estimar distintas curvas de supervivencia en función doutras covariables. Por exemplo, con este conxunto de datos poderíamos estimar dúas curvas de supervivencia distintas, dependendo do tratamento empregado. Vemos que a información das saídas é a mesma, só que calculada para cada caso.

```
> datost.km=survfit(datos_s~grp,data=datos,conf.type='log-log')
> summary(datost.km)
Call: survfit(formula = datos_s ~ grp, data = datos, conf.type = "log-log")
```

```
      grp=combination
time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.5    61      4    0.934  0.0317    0.835    0.975
2.0    57      3    0.885  0.0408    0.774    0.944
4.0    54      1    0.869  0.0432    0.755    0.932
5.0    53      2    0.836  0.0474    0.717    0.908
8.0    51      2    0.803  0.0509    0.680    0.883
...
170.0   25      1    0.393  0.0625    0.272    0.513
      grp=patchOnly
time n.risk n.event survival std.err lower 95% CI upper 95% CI
0.5    64      8    0.875  0.0413    0.766    0.935
1.0    56      5    0.797  0.0503    0.676    0.877
2.0    51      3    0.750  0.0541    0.625    0.839
3.0    48      1    0.734  0.0552    0.608    0.826
4.0    47      2    0.703  0.0571    0.575    0.799
...
170.0   13      1    0.188  0.0488    0.103    0.291
```

```
> datost.km
```

```
Call: survfit(formula = datos_s ~ grp, data = datos, conf.type = "log-log")
```

	n	events	median	0.95LCL	0.95UCL
grp=combination	61	37	65	42	NA
grp=patchOnly	64	52	23	14	49

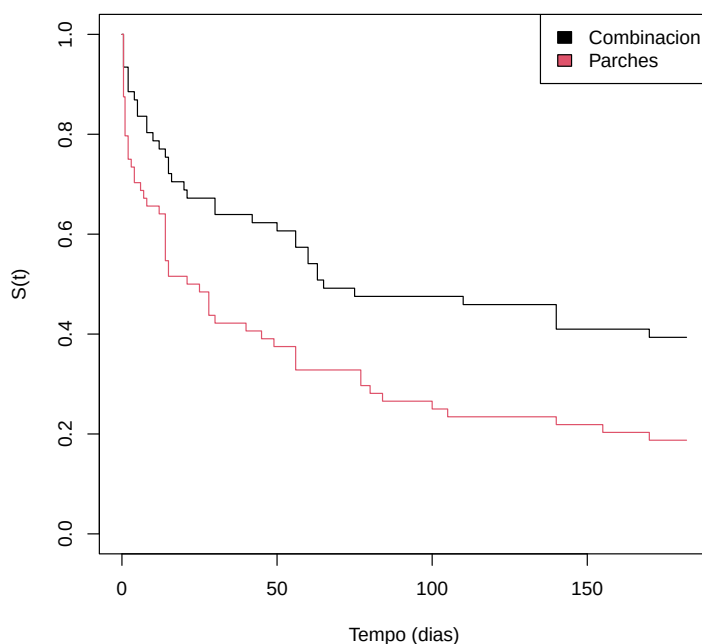


Figura 3.4: Estimador de Kaplan-Meier diferenciando os tratamentos

Desta forma, poderíamos comparar distintas curvas de supervivencia. Neste caso, vemos claramente como o tratamento consistente en empregar unha combinación de tres medicamentos é máis efectivo que os parches de nicotina. En efecto, a figura 3.4 indícanos que o tempo ata a recaída é menor nos pacientes que empregan os parches, que se corresponden coa curva vermella. Este mesmo procedemento poderíamolo empregar con outras covariables coma o xénero, a raza ou a condición de ter deixado de fumar con anterioridade, por exemplo. É por isto que este paquete tamén conta con máis ferramentas que axudan a realizar contrastes de hipóteses ou comparacións entre poboacións, mediante diferentes tests coma o log-rank. Ademais, tamén conta con funcións que permiten modelar o que se coñecen como modelos Cox, que permiten ver que factores ou covariables inflúen máis na supervivencia dunha poboación (en [11] poden verse todas estas ferramentas con detalle).

3.4. Probas de bondade de axuste

Como xa se mencionou, os estimadores non paramétricos poden axudar a axustarlle unha distribución paramétrica aos datos. Neste capítulo veremos un método gráfico para comprobar a bondade de axuste dos datos aos modelos exponencial e Weibull a partir do estimador de Kaplan-Meier da función de supervivencia.

3.4.1. Modelo exponencial

Como xa temos visto, a función de supervivencia exponencial é $S(t) = e^{-\lambda t}$. Polo tanto, tomando logaritmos chegamos á seguinte expresión, que non é máis que a relación da función de supervivencia coa función de risco acumulado:

$$-\log[S(t)] = H(t) = \lambda t \quad (3.9)$$

Tendo isto en conta, a partir dun conxunto de tempos de supervivencia podemos representar graficamente $-\log[\hat{S}(t_i)]$ fronte aos t_i , onde $\hat{S}(t_i)$ é o estimador de Kaplan-Meier. Desta forma, se os datos cos que estamos a traballar proceden efectivamente dun modelo exponencial, a nube de puntos resultante debería seguir unha liña recta (a cal pode axustarse por mínimos cadrados). Ademais, a pendente desta recta pode proporcionar un estimador do parámetro λ , $\hat{\lambda}$. Vexamos este procedemento cun exemplo, empregando a base de datos `pharmacoSmoking` e as estimacións vistas na sección 3.3.

Exemplo 3.3. A partir do estimador de Kaplan-Meier xa calculado extraemos a información que nos interesa, os diferentes tempos de supervivencia observados e o valor do estimador, para despois aplicarlles as transformacións correspondentes.

```
> EstSup=datos.km$surv # valores do estimador Kaplan-Meier
> TSurv=datos.km$time # diferentes tempos observados
> logEstSup=-log(EstSup)
```

Feito isto, temos que axustar o modelo de regresión linear relativo á expresión (3.9), o cal se realiza coa función `lm`. Este modelo vainos permitir representar a recta de regresión xunto á nube de puntos. Na figura 3.5 vemos como os datos non se axustan ben á recta calculada, polo que a hipótese de procedencia dun modelo exponencial non parece certa. De feito, no exemplo 3.4 veremos que os datos axústanse ben a unha Weibull.

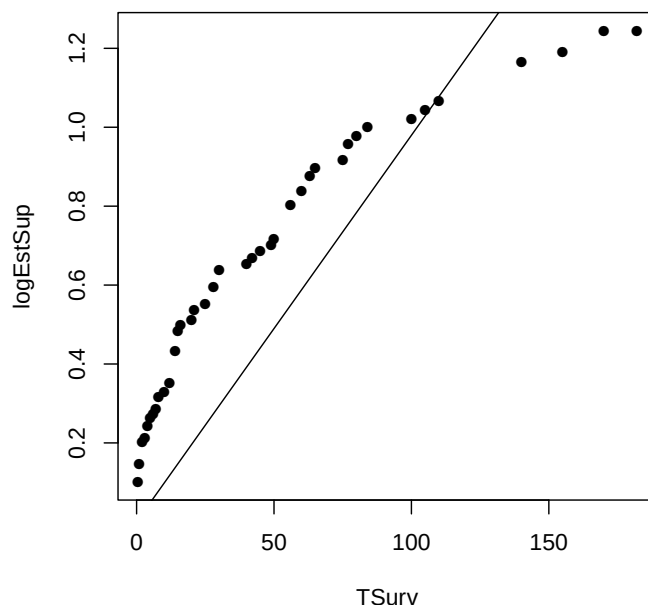


Figura 3.5: Nube de puntos de $-\log[\hat{S}(t_i)]$ fronte a t_i xunto coa recta de regresión axustada

3.4.2. Modelo Weibull

No caso Weibull, recordamos que a función de supervivencia é $S(t) = e^{-(\lambda t)^\alpha}$. De aquí, se tomamos dúas veces logaritmos e operamos no segundo membro obtemos:

$$\log[-\log[S(t)]] = \log[H(t)] = \alpha \log \lambda + \alpha \log t \quad (3.10)$$

Polo tanto, de forma análoga ao caso exponencial, podemos representar graficamente $\log[-\log[\hat{S}(t_i)]]$ fronte a $\log t_i$, obtendo unha nube de puntos lineal se a hipótese de procedencia dunha Weibull é certa. Ademais, ao axustar unha recta de regresión podemos obter estimadores dos parámetros da Weibull; se $\hat{\beta}_0$ e $\hat{\beta}_1$ representan os estimadores do intercepto e da pendente da recta, respectivamente, temos que $\hat{\alpha} = \hat{\beta}_1$ e $\hat{\lambda} = \exp(\frac{\hat{\beta}_0}{\hat{\beta}_1})$. Vexamos este procedemento aplicándoo aos mesmos datos do exemplo 3.3.

Exemplo 3.4. A partir dos tempos de supervivencia e os valores do estimador de Kaplan-Meier empregados no exemplo 3.3, calculamos as transformacións necesarias e plantexamos o modelo de regresión linear relativo á expresión (3.10), obtendo o seguinte resultado:

```
> loglogEstSup=log(-log(EstSup))
> logTSurv=log(TSurv)
```

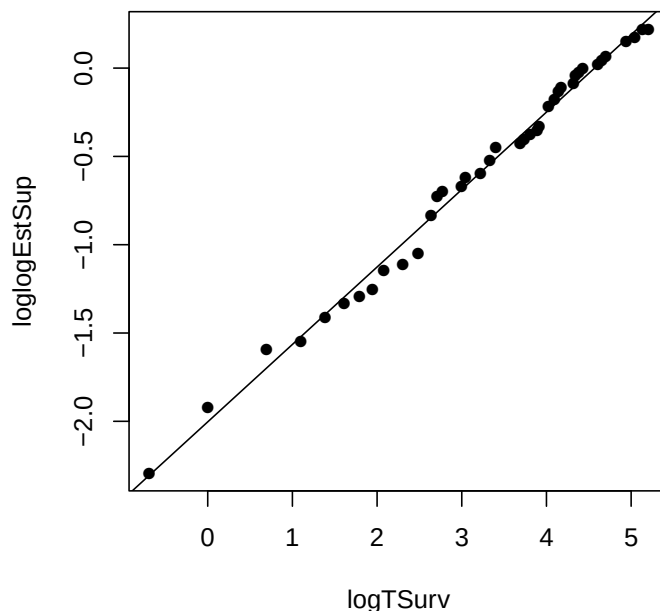


Figura 3.6: Nube de puntos de $\log[-\log[\hat{S}(t_i)]]$ fronte a $\log t_i$ xunto coa recta de regresión axustada

Na figura 3.6 vemos como os datos se axustan bastante ben á recta calculada, o cal pode ser un indicio de que efectivamente a distribución dos datos é unha Weibull. Ademais, podemos extraer información como a pendente ($\hat{\beta}_1 = 0,4385$) e o intercepto da recta ($\hat{\beta}_0 = -2,0032$), que nos permiten estimar os parámetros de forma e escala da Weibull.

```
> axuste=lm(loglogEstSup~logTSurv)
> axuste
Call:
lm(formula = loglogEstSup ~ logTSurv)

Coefficients:
(Intercept)      logTSurv
   -2.0032         0.4385
```

Desta forma, temos que $\hat{\alpha} = \hat{\beta}_1 = 0,4385$ e $\hat{\lambda} = \exp(\frac{\hat{\beta}_0}{\hat{\beta}_1}) = 0,0103$. Aínda así, ao coñecer que os datos se axustan ben a unha Weibull, poderíamos calcular os estimadores dos parámetros empregando o método de estimación por máxima verosimilitude visto no capítulo 2. De feito, se calculamos os estimadores de máxima verosimilitude obtemos que $\hat{\alpha}_{emv} = 0,4899$ e $\hat{\lambda}_{emv} = 0,0095$, valores similares aos obtidos mediante o método gráfico. Para rematar, na figura 3.7

móstrase como ambas curvas de supervivencia estimadas teñen unha forma moi semellante ao estimador de Kaplan-Meier, sendo a curva estimada por máxima verosimilitude lixeiramente maior que a estimada polo método gráfico.

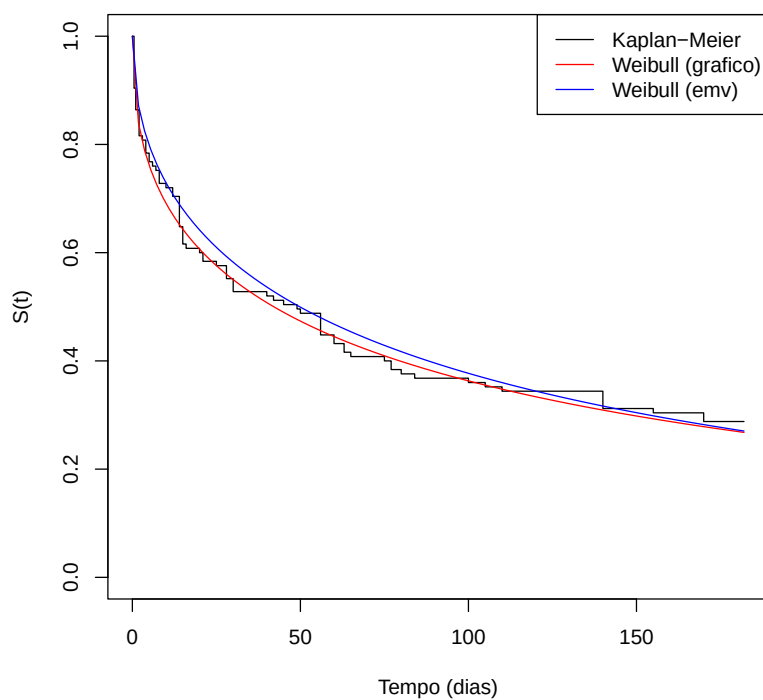


Figura 3.7: Comparativa do estimador de Kaplan-Meier xunto ás curvas de supervivencia Weibull estimadas mediante o método gráfico (vermella) e máxima verosimilitude (azul)

Anexo A

Código de empregado

A.1. Figuras da introdución

Para a realización das figuras 1, 2 e 3 empregamos o paquete *ggplot2*, que incorpora diversas ferramentas para a representación e edición dos gráficos.

```
library(ggplot2)
```

A.1.1. Figuras 1 e 2.

```
## Función para crear gráficos de censuras tipo I y II ##
```

```
graf_censura<-function(cen,t1,titulo,xlabel,ylabel){  
  m<-data.frame(t1,cen)  
  censura<-as.factor(m$cen)  
  ggplot(m)+  
    geom_point(aes(x=t1,y=1:length(t1),shape=censura),size=4,stroke=1)+  
    scale_shape_manual(values=c(20,4),name = "Censura",labels=  
      c(quote(Censurado), quote(Observado)))+  
    geom_linerange(aes(xmin=0, xmax=t1, y=1:length(t1)))+  
    geom_vline(xintercept=max(t1),linetype='dashed')+  
    coord_cartesian(ylim = c(0,length(t1)+1),xlim=c(0,max(t1)+5))+  
    labs(x=xlabel, y=ylabel, title=titulo)+  
    theme_classic()+  
    theme(plot.title = element_text(family = "Bookman",  
      face = "bold",
```

```

        color = 1,
        size = 15,
        hjust = 0.5,
        vjust = 2.5,
        angle = 0,
        lineheight = 3,
        margin = margin(0, 0, 0, 0)))+
  scale_y_continuous(breaks=c(1:length(t1)),labels=c('A','B','C',
'D','E','F'))+scale_x_continuous(breaks=seq(0,max(t1),5))
}
graf_censura(c(1,1,0,1,0,1),c(10,15,30,25,30,19),'Tempo de aparición dun tumor
  en ratas expostas a canceríxenos','Tempo (semanas)','Ratas')
graf_censura(c(1,1,0,1,0,1),c(10,15,25,25,25,19),'Tempo de aparición dun tumor
  en ratas expostas a canceríxenos','Tempo (semanas)','Ratas')
```

A.1.2. Figura 3

Función para crear gráfico de censura tipo III

```

graf_censura3<-function(cen,t1,t0,titulo,xlabel,ylabel){
  m<-data.frame(t1,cen)
  censura<-as.factor(m$cen)
  ggplot(m)+
    geom_point(aes(x=t0+t1,y=1:length(t1),shape=censura),size=4,stroke=1)+
    scale_shape_manual(values=c(20,4),name = "Censura",
    labels=c(quote(Censurado), quote(Observado)))+
    geom_linerange(aes(xmin=t0, xmax=t0+t1, y=1:length(t1)))+
    geom_vline(xintercept=max(t0+t1),linetype='dashed')+
    coord_cartesian(ylim = c(0,length(t1)+1),xlim=c(0,max(t0+t1)+5))+
    labs(x=xlabel, y=ylabel, title=titulo)+
    theme_classic()+
    theme(plot.title = element_text(family = "Bookman",
        face = "bold",
        color = 1,
        size = 15,
        hjust = 0.5,
        vjust = 2.5,
        angle = 0,
```

```

                                lineheight = 3,
                                margin = margin(0, 0, 0, 0)))+
scale_y_continuous(breaks=c(1:length(t1)),labels=c('A','B','C',
'D','E','F'))+scale_x_continuous(breaks=seq(0,max(t1+t0),1))
}

graf_censura3(c(1,0,1,0,1,0),c(4,4,6,8,3,3),c(2,3,4,5,9,10),'Tempo de remisión
de pacientes tratados de leucemia aguda','Tempo (meses)','Pacientes')

```

A.2. Figuras do Capítulo 1

A.2.1. Figura 1.1

```

## Gráfica da función de distribución e supervivencia (exponencial) ##

par(mfrow=c(1,2))
curve(pexp(x,1),0,8,main='Distribucion',xlab='t',ylab='F(t)')
curve(pexp(x,1,lower.tail=F),0,8,main='Curva de supervivencia',
      xlab='t',ylab='S(t)')

```

A.2.2. Figura 1.2

```

## Curvas de risco ##

par(mfrow=c(2,2)) # creamos a ventá gráfica

# Risco exponencial(1)
curve(dexp(x)/pexp(x,lower.tail=F),0,12,main='Risco exponencial',xlab='',
      ,ylab='',xaxt='n',yaxt='n')
title(ylab='Taxa de fallo',line=1,cex.lab=1)
title(xlab='Tempo',line=1,cex.lab=1)

# Risco Weibull(3,1) 3*x^2=dweibull(x,3,1)/pweibull(x,3,1,lower.tail=F)
# para evitar problemas de precisión.
curve(3*x^2,0,12,col='red',main='Risco Weibull',xlab='',ylab='',
      ,xaxt='n',yaxt='n')

```

```

title(ylab='Taxa de fallo',line=1,cex.lab=1)
title(xlab='Tempo',line=1,cex.lab=1)

# Risco Weibull(1/2,1)
curve(dweibull(x,1/2,1)/pweibull(x,1/2,1,lower.tail=F),0,12,col='blue',
      main='Risco Weibull',xlab='',ylab='',xaxt='n',yaxt='n')
title(ylab='Taxa de fallo',line=1,cex.lab=1)
title(xlab='Tempo',line=1,cex.lab=1)

# Risco LogNormal(2,1)
curve(dlnorm(x,2,1)/plnorm(x,2,1,lower.tail=F),0,100,col='green',
      main='Risco Lognormal',xlab='',ylab='',xaxt='n',yaxt='n')
title(ylab='Taxa de fallo',line=1,cex.lab=1)
title(xlab='Tempo',line=1,cex.lab=1)

```

A.3. Capítulo 2

A.3.1. Figura 2.1

```

## Modelo exponencial ##

par(mfrow=c(1,3))
lambda=c(0.5, 1, 1.5) # vector de parámetros a emplegar
color=c('black','red','blue')

# Función de riesgo #
curve(dexp(x,lambda[1])/pexp(x,lambda[1],lower.tail=F),0,5,ylim=c(0,2),
      xlab='t',ylab='h(t)')
title('Risco exponencial')
for (i in 2:length(lambda)){
  curve(dexp(x,lambda[i])/pexp(x,lambda[i],lower.tail=F),0,5,
        col=color[i],add=TRUE)}

# Función de supervivencia #
curve(pexp(x,lambda[1],lower.tail=F),0,5,ylim=c(0,1),xlab='t',ylab='S(t)')
title('Supervivencia exponencial')
for (i in 2:length(lambda)){

```

```

curve(pexp(x,lambda[i],lower.tail=F),0,5,col=color[i],add=TRUE)}

# Función de densidade #
curve(dexp(x,lambda[1]),0,5,ylim=c(0,1.5),xlab='t',ylab='f(t)')
title('Densidade exponencial')
for (i in 2:length(lambda)){
  curve(dexp(x,lambda[i]),0,5,col=color[i],add=TRUE)}

```

A.3.2. Figura 2.2

```

## Modelo Weibull ##

# Definimos funcións para calcular o risco e a supervivencia
weibSurv=function(x, forma, escala) {
  pweibull(x, shape=forma,scale=escala, lower.tail=F)}
weibRies=function(x, forma, escala) {
  dweibull(x, shape=forma,scale=escala)/pweibull(x, shape=forma,
scale=escala,lower.tail=F)}

par(mfrow=c(1,3))
forma=c(0.5,1.5,5) # parámetros de forma a empregar
color=c('black','red','blue')

# Función de risco #
curve(weibRies(x,forma[1],1),0,2.5,ylim=c(0,2.5),xlab='t',ylab='h(t)')
title('Risco Weibull')
for (i in 2:length(forma)){
  curve(weibRies(x,forma[i],1),0,5,col=color[i],add=TRUE)}

# Función de supervivencia #
curve(weibSurv(x,forma[1],1),0,2.5,ylim=c(0,1),xlab='t',ylab='S(t)')
title('Supervivencia Weibull')
for (i in 2:length(forma)){
  curve(weibSurv(x,forma[i],1),0,5,col=color[i],add=TRUE)}

# Función de densidade #
curve(dweibull(x,forma[1],1),0,2.5,ylim=c(0,2.5),xlab='t',ylab='f(t)')
title('Densidade Weibull')

```

```
for (i in 2:length(forma)){
  curve(dexp(x,forma[i]),0,5,col=color[i],add=TRUE)}
```

A.3.3. Estudo de simulación

```
## Exemplo de simulacion e estimacion Weibull ##

library(stats4) # libreria necesaria para empregar a funcion mle
set.seed(2022) # fixamos a semente
s=1000 #numero de simulacións
n=10 #30 300 3000 tamanho das mostras
alfa=5
lambda=1

# 25% de censura %
# Construción das mostras censuradas #
lambdac=1.25
est_alfa=numeric(s)
est_lambda=numeric(s)
cob_alfa_pv=0
cob_lambda_pv=0
lon_alfa_pv=numeric(s)
lon_lambda_pv=numeric(s)
cob_alfa_as=0
cob_lambda_as=0
lon_alfa_as=numeric(s)
lon_lambda_as=numeric(s)
est_cens=numeric(s)
for (i in 1:s){
  m=rweibull(n,alfa,lambda)
  cen=rweibull(n,alfa,lambdac)
  mcen=pmin(m,cen)
  icen=as.numeric(m<=cen)
  s1=subset(mcen,icen=='1')
  s0=subset(mcen,icen=='0')
  est_cens[i]=1-sum(icen)/n

  # Calculo dos estimadores #
```

```

lvn=function(a,l){-sum(dweibull(s1,a,1/l,log=TRUE))-sum(pweibull(s0,a,1/l,
                      log=TRUE,lower=FALSE))}
emv=mle(lvn,list(a=5,l=1))
est_alfa[i]=coef(emv)[1]
est_lambda[i]=coef(emv)[2]

# Intervalos por perfil de verosimilitud
int=confint(emv)
alfa_iz_pv=int[1]
alfa_der_pv=int[3]
lambda_iz_pv=int[2]
lambda_der_pv=int[4]
lon_alfa_pv[i]=alfa_der_pv-alfa_iz_pv
lon_lambda_pv[i]=lambda_der_pv-lambda_iz_pv
if (alfa_iz_pv<=alfa & alfa_der_pv>=alfa){
  cob_alfa_pv=cob_alfa_pv+1}
if (lambda_iz_pv<=lambda & lambda_der_pv>=lambda){
  cob_lambda_pv=cob_lambda_pv+1}

# Intervalos asintoticos
err=sqrt(diag(vcov(emv)))
alfa_iz_as=est_alfa[i]-qnorm(0.975)*err[1]
alfa_der_as=est_alfa[i]+qnorm(0.975)*err[1]
lambda_iz_as=est_lambda[i]-qnorm(0.975)*err[2]
lambda_der_as=est_lambda[i]+qnorm(0.975)*err[2]
if (alfa_iz_as<=alfa & alfa_der_as>=alfa){
  cob_alfa_as=cob_alfa_as+1}
if (lambda_iz_as<=lambda & lambda_der_as>=lambda){
  cob_lambda_as=cob_lambda_as+1}
lon_alfa_as[i]=alfa_der_as-alfa_iz_as
lon_lambda_as[i]=lambda_der_as-lambda_iz_as
}

nesgo=c(mean(est_alfa)-alfa,mean(est_lambda)-lambda)
varianza=c(var(est_alfa)*(n-1)/n,var(est_lambda)*(n-1)/n)
ECM=nesgo^2+varianza

# Resultados por perfil

```

```

cob_alfa_pv=cob_alfa_pv/s
cob_lambda_pv=cob_lambda_pv/s
mean(lon_alfa_pv)
mean(lon_lambda_pv)

# Resultados asintoticos
cob_alfa_as=cob_alfa_as/s
cob_lambda_as=cob_lambda_as/s
mean(lon_alfa_as)
mean(lon_lambda_as)

# 50% de censura #
# Construcion das mostras censuradas #
lambdac=1
est_alfa=numeric(s)
est_lambda=numeric(s)
cob_alfa_pv=0
cob_lambda_pv=0
lon_alfa_pv=numeric(s)
lon_lambda_pv=numeric(s)
cob_alfa_as=0
cob_lambda_as=0
lon_alfa_as=numeric(s)
lon_lambda_as=numeric(s)
est_cens=numeric(s)
for (i in 1:s){
  m=rweibull(n,alfa,1/lambda)
  cen=rweibull(n,alfa,1/lambdac)
  mcen=pmin(m,cen)
  icen=as.numeric(m<=cen)
  s1=subset(mcen,icen=='1')
  s0=subset(mcen,icen=='0')
  est_cens[i]=1-sum(icen)/n

  # Calculo dos estimadores #
  lvn=function(a,l){-sum(dweibull(s1,a,1/l,log=TRUE))-sum(pweibull(s0,a,1/l,
    log=TRUE,lower=FALSE))}
  emv=mle(lvn,list(a=5,l=1))

```

```

est_alfa[i]=coef(emv)[1]
est_lambda[i]=coef(emv)[2]

# Intervalos por perfil de verosimilitude
int=confint(emv)
alfa_iz_pv=int[1]
alfa_der_pv=int[3]
lambda_iz_pv=int[2]
lambda_der_pv=int[4]
if (is.na(alfa_iz_pv)){
  alfa_iz_pv=0}
if (is.na(lambda_iz_pv)){
  lambda_iz_pv=0}
if (alfa_iz_pv<=alfa & alfa_der_pv>=alfa){
  cob_alfa_pv=cob_alfa_pv+1}
if (lambda_iz_pv<=lambda & lambda_der_pv>=lambda){
  cob_lambda_pv=cob_lambda_pv+1}
lon_alfa_pv[i]=alfa_der_pv-alfa_iz_pv
lon_lambda_pv[i]=lambda_der_pv-lambda_iz_pv

# Intervalos asintoticos
err=sqrt(diag(vcov(emv)))
alfa_iz_as=est_alfa[i]-qnorm(0.975)*err[1]
alfa_der_as=est_alfa[i]+qnorm(0.975)*err[1]
lambda_iz_as=est_lambda[i]-qnorm(0.975)*err[2]
lambda_der_as=est_lambda[i]+qnorm(0.975)*err[2]
if (alfa_iz_as<=alfa & alfa_der_as>=alfa){
  cob_alfa_as=cob_alfa_as+1}
if (lambda_iz_as<=lambda & lambda_der_as>=lambda){
  cob_lambda_as=cob_lambda_as+1}
lon_alfa_as[i]=alfa_der_as-alfa_iz_as
lon_lambda_as[i]=lambda_der_as-lambda_iz_as
}

nesgo=c(mean(est_alfa)-alfa,mean(est_lambda)-lambda)
varianza=c(var(est_alfa)*(n-1)/n,var(est_lambda)*(n-1)/n)
ECM=nesgo^2+varianza

```

```
# Resultados por perfil
cob_alfa=cob_alfa/s          # cobertura para alfa
cob_lambda=cob_lambda/s      # cobertura para lambda
mean(lon_alfa_pv)            # lonxitude media para alfa
mean(lon_lambda_pv)          # lonxitude media para lambda

# Resultados asymptoticos
cob_alfa_as=cob_alfa_as/s
cob_lambda_as=cob_lambda_as/s
mean(lon_alfa_as)
mean(lon_lambda_as)
```

A.4. Capítulo 3

A.4.1. Exemplo 3.2

```
## Exemplo estimador de Kaplan-Meier ##

library(survival) # librería necesaria para facer análise de supervivencia

# Introducimos os datos #
t=c(3,4,5.7,6.5,6.5,8.4,10,10,12,15) # tempos de supervivencia
c=c(1,0,0,1,1,0,1,0,1,1) # indicador de censura
datos=Surv(t,c)

# Estimador KM con intervalos de confianza por defecto
datos.km=survfit(datos~1,type='kaplan-meier')
# Estimador KM empregando a transformación log-log nos i.c.
datosl.km=survfit(datos~1,type='kaplan-meier',conf.type='log-log')
summary(datos.km)
summary(datosl.km)

# Representación sen rexións de confianza
plot(datos.km,xlab='Meses',ylab='S(t)',conf.int=F)
# Representación cos intervalos de confianza calculados
```

```
par(mfrow=c(1,2))
plot(datos.km,xlab='Meses',ylab='S(t)')
plot(datosl.km,xlab='Meses',ylab='S(t)')
```

A.4.2. Estudio dos datos pharmacoSmoking

```
## Aplicación a datos pharmacoSmoking ##

library(survival)
library(asauro)
library(stats4)

# Preparamos os datos
data(pharmacoSmoking)
datos=pharmacoSmoking
attach(datos)
ttr[ttr==0]=0.5
datos$ttr=ttr
datos=datos[,c(2,3,4)]
head(datos)

# Creamos obxecto de supervivencia
datos_s=Surv(datos$ttr,datos$relapse)

# Estimador de Kaplan-Meier
datos.km=survfit(datos_s~1,conf.type='log-log')
summary(datos.km)
plot(datos.km,xlab='Tempo (dias)',ylab='S(t)')
print(datos.km,print.rmean=TRUE)

# Estimador de Kaplan-Meier por tratamento
datost.km=survfit(datos_s~grp,data=datos,conf.type='log-log')
summary(datost.km)
plot(datost.km,xlab='Tempo (dias)',ylab='S(t)',col=c(1,2))
legend(x = "topright", legend = c("Combinacion", "Parches"), fill = c(1, 2))

# Bondade de axuste exponencial
EstSup=datos.km$surv # valores do estimador Kaplan-Meier
```

```

TSurv=datos.km$time # diferentes tempos observados
logEstSup=-log(EstSup)
axuste_exp=lm(logEstSup~0+TSurv) # modelo de regresion con intercepto=0
axuste_exp
plot(logEstSup~TSurv,pch=16)
abline(axuste_exp)

# Bondade de axuste Weibull
loglogEstSup=log(logEstSup)
logTSurv=log(TSurv)
axuste_weibull=lm(loglogEstSup~logTSurv) # modelo de regresion
axuste_weibull
plot(loglogEstSup~logTSurv,pch=16)
abline(axuste_weibull)

# Calculo de estimadores por maxima verosimilitude
s1=subset(datos,relapse=='1')
s0=subset(datos,relapse=='0')
lvn=function(a,l){
  -sum(dweibull(s1$ttr,a,1/l,
    log=TRUE))-sum(pweibull(s0$ttr,a,1/l,log=TRUE,lower=FALSE))}
emv=mle(lvn,list(a=0.4385,l=0.0103),method='Nelder-Mead')
# Figura 3.7
plot(datos.km,xlab='Tempo (dias)',ylab='S(t)',conf.int=FALSE)
curve(pweibull(x,0.4385,1/0.0103,lower.tail=FALSE),0,182,col='red',add=TRUE)
curve(pweibull(x,coef(emv)[1],1/coef(emv)[2],lower.tail=FALSE),0,182,
  col='blue',add=TRUE)
legend(x='topright',legend=c('Kaplan-Meier','Weibull (grafico)','Weibull (emv)'),
  lty=c(1,1,1),col=c('black','red','blue'))

```

Bibliografía

- [1] D. J. Bartholomew, *A problem in life testing*, Journal of the American Statistical Association **52** (1957), no. 279, 350–355.
- [2] Byron W. Brown, Myles Hollander, and Ramesh M. Korwar, *Nonparametric tests of independence for censored data with application to heart transplant studies*, 1973.
- [3] B. Efron, *The two-sample problem with censored data*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **4** (1967), 831–852.
- [4] Edmund A. Gehan, *Estimating survival functions from the life table*, Journal of Chronic Diseases **21** (1969), no. 9, 629–644.
- [5] Richard Gill, *Censoring and stochastic integrals*, vol. 125, 01 1980.
- [6] Myles Hollander and Frank Proschan, *Testing to determine the underlying distribution using randomly censored data*, Biometrics **35** (1979), no. 2, 393–401.
- [7] Ricardo Vélez Ibarrola and Alfonso García Pérez, *Principios de inferencia estadística*, 1^a ed., 2^a reimp ed., Universidad Nacional de Educación a Distancia, Madrid, 1997.
- [8] E. L. Kaplan and Paul Meier, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association **53** (1958), no. 282, 457–481.
- [9] John P. Klein and Melvin L. Moeschberger, *Survival analysis: techniques for censored and truncated data*, Springer, New York, 1997.
- [10] Elisa T. Lee, *Statistical methods for survival data analysis*, 2nd ed ed., Wiley, New York, 1992.
- [11] Dirk Foster Moore, *Applied survival analysis using r*, Springer, Switzerland, 2016.
- [12] Terry M. Therneau, *Survival analysis [r package survival version 3.3-1]*, -03-03 2022.
- [13] W. Weibull, *A statistical distribution of wide applicability*, Journal of Applied Mechanics **18**, 293–297.