USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

ANÁLISE DE DATOS CENSURADOS

Alba Candal Parafita

Curso 2023/2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

ANÁLISE DE DATOS CENSURADOS

Alba Candal Parafita

Xullo, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Análise de datos censurados
Breve descrición do contido
A Análise de Supervivencia é unha parte da Estatística que se encarga do estudo de tempos de vida, como pode ser a duración dunha enfermidade, a duración dun dispositivo electrónico, ou calquera outro tempo que transcorre entre dous eventos de interese. Habitualmente, a análise deste tipo de datos require un seguimento e se dito seguimento se interrompe por algunha causa, xorden os datos censurados. Polo tanto, cando se traballa con datos censurados dispoñerase de individuos para os cales se observa totalmente o tempo que transcorreu entre os dous eventos de interese, mentres que noutros casos só se dispón do tempo ata o momento no que se perdeu o seguimento. A modo de orientación, o traballo podería organizarse nas seguintes seccións: ▪ Introdución aos datos censurados. ▪ Estimación da función de distribución no contexto de datos censurados: o estimador de Kaplan-Meier. ▪ Propiedades do estimador de Kaplan-Meier. ▪ Estimador de Kaplan- Meier condicional: o estimador de Beran. Ademais, abordaremos a análise de datos censurados tanto sobre conxuntos de datos reais como sobre datos simulados. Para iso utilizaremos o software estatístico libre R. (https://www.r-project.org/).

Recomendacións
Outras observacións

Índice

Resumo

A censura é un fenómeno que se produce con frecuencia na Análise de Supervivencia e está asociada a unha perda parcial de información. Neste traballo, ilustramos como os estimadores non paramétricos tradicionais, como a función de distribución empírica ou o seu análogo condicional, fallan no intento de dar unha estimación da función de distribución (condicional) dunha variable aleatoria T censurada pola dereita. No contexto dos datos censurados, introducimos o estimador de Kaplan-Meier e o estimador de Beran, como estimadores da función de distribución incondicional e condicional, respectivamente. No caso do estimador de Kaplan-Meier desenvolvemos tamén algunhas das súas propiedades máis destacables, que resultarán esenciais para a construción de intervalos de confianza. Para comparar o comportamento dos distintos estimadores presentamos diferentes estudos de simulación por Montecarlo empregando o software estatístico R. Finalmente, analizamos un conxunto de datos reais provenientes dun grupo de doentes con cancro de pulmón facendo uso dos estimadores específicos para escenarios con datos censurados.

Abstract

Censoring is a common occurrence in Survival Analysis and is linked to a partial loss of information. In this work, we illustrate how the traditional nonparametric estimators, such as the empirical cumulative distribution function and its conditional version, do not present good results in order to estimate the (conditional) cumulative distribution function of a right-censored random variable T. In the context of censored data, we introduce the Kaplan-Meier estimator and the Beran estimator, as estimators of the unconditional and conditional cumulative distribution function, respectively. For the Kaplan-Meier estimator, we also discuss its most relevant properties, which are essential to construct confidence intervals. To compare the behavior of the different estimators we perform several Montecarlo simulation studies using the statistical software R. Finally, we analyze a real data set that contains information about the survival time associated with a group of lung cancer patients by using the censored-adapted estimators.

Capítulo 1

Introdución aos datos censurados

Neste capítulo, explicaremos en que consiste a Análise de Supervivencia e abordaremos unha problemática habitual neste contexto: a censura. Ademais, distinguiremos entre os distintos tipos de datos censurados e diferenciarémolos dos datos truncados.

1.1. A Análise de Supervivencia

A **Análise de Supervivencia** é unha rama da Estatística centrada no estudo dunha variable aleatoria[1] non negativa, que denotaremos por T, e que representa o tempo que transcorre ata a aparición dun certo evento de interese. Ten presenza en múltiples campos, dende a Medicina, a Bioloxía ou a Demografía, ata a Enconomía ou a Enxeñaría. No que segue, referirémonos á variable T como **tempo de supervivencia**. Vexamos algúns exemplos de tempos de supervivencia:

- No estudo da eficacia dun medicamento: considerando como evento de interese o falecemento a causa da enfermidade que se intenta combater, T representaría o tempo de vida das/os participantes dende a diagnose da enfermidade.
- Na determinación da fiabilidade dun dispositivo: se o evento de interese é o momento no que o dispositivo deixa de ser funcional, T representará o seu tempo de vida útil.
- Nun contexto de reinserción criminal: asociaremos T co tempo dende que un individuo sae do cárcere ata que comete unha nova acción delictiva.

Nun escenario como os que acabamos de expoñer, o obxectivo será estimar a función de distribución que segue a variable T. Esta información, a parte de ser chave para poder descri-

[1] Unha **variable aleatoria** X é unha función definida nun espazo de probabilidade $(\Omega, A, \mathbb{P})$ que toma valores en $\mathbb{R}$.

bir o comportamento da variable de interese, podería ser empregada para estimar o tempo de supervivencia dun hipotético novo individuo.

Lembremos que nun caso como o do primeiro exemplo, a función de distribución asociada ao tempo de supervivencia T, que denotaremos por $F(t)$, representa a probabilidade de que un suxeito/obxecto sufra o evento de interese antes dun instante dado, isto é $F(t) = \mathbb{P}(T \leq t)$. Non obstante, na Análise de Supervivencia adóptase unha visión "máis optimista" e trabállase coa función que presentamos a continuación:

Definición 1.1. Sexa $(\Omega, A, \mathbb{P})$ un espazo de probabilidade e $T : \Omega \to \mathbb{R}$ unha variable aleatoria, definiremos a **función de supervivencia** de T como $S : \mathbb{R} \to [0, 1]$ de forma que

$$S(t) = 1 - F(t) = 1 - \mathbb{P}(T \leq t) = \mathbb{P}(T > t).$$

Así, considerarase a probabilidade de que un individuo non se vexa afectado/a polo evento de interese polo menos ata un instante t. Nestas circunstancias, poderíamos tentar atopar unha función que indique o risco de que un individuo padeza o evento de interese neste instante, tendo en conta que sobreviviu ata este momento. Para iso, consideraremos seguinte límite:

$$\begin{aligned}\lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\mathbb{P}(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\mathbb{P}(T < t + \Delta t) - \mathbb{P}(T < t)}{\mathbb{P}(T \geq t)} \\ &= \frac{1}{1 - F(t-)} \lim_{\Delta t \to 0} \frac{\mathbb{P}(T < t + \Delta t) - \mathbb{P}(T < t)}{\Delta t}.\end{aligned}$$

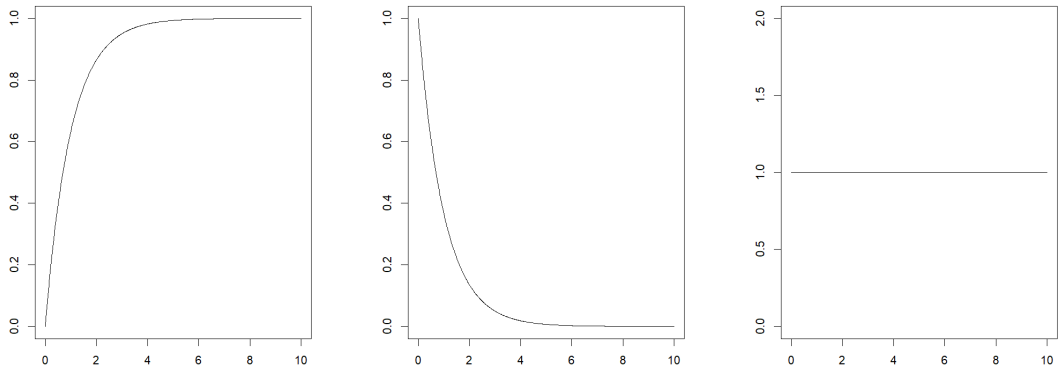
Como a función F é continua, tense que $F(t-) = F(t)$ e denotando $f(t)$ como a función de densidade[2] variable continua T, podemos definir a seguinte función:

Definición 1.2. Sexa $X : \Omega \to \mathbb{R}$ unha variable aleatoria continua, definiremos a **función de risco** de X como $\lambda(x) = f(x) \cdot (1 - F(x))^{-1}$, onde f e F son as funcións de densidade e distribución da variable X, respectivamente.

De supor que a función de risco fose constante, algo que en realidade é bastante restritivo en contextos como o médico ou o industrial que presentabamos ao comezo, e tomando por exemplo $\lambda(t) = 1$, obteríamos que $1 = \frac{F'(t)}{1-F(t)} = -\log(1 - F(t))'$, de forma que $S(t) = 1 - F(t) = e^{-t}$ e $F(t) = 1 - e^{-t}$, concluíndo que a variable T segue unha distribución expoñencial[3] de parámetro 1. Pódese observar na Figura 1.1 o comportamento destas funcións neste caso particular.

[2] Sexa $X : (\Omega, A, \mathbb{P}) \to \mathbb{R}$ unha variable aleatoria continua, denotarémo-la función $f : X \to \mathbb{R}$ como a **función de densidade** de X.

[3] Diremos que unha variable aleatoria X segue unha **distribución expoñencial de parámetro** $\lambda > 0$ ($X \in Exp(\lambda)$), se a súa función de distribución é da forma $F(t) = \lambda \cdot e^{-\lambda t}$.



(a) Función de distribución. (b) Función de supervivencia. (c) Función de risco.

Figura 1.1: Funcións de distribución, supervivencia e risco asociadas a unha variable aleatoria $T \in Exp(1)$.

1.2. Datos censurados

Supoñamos que se está a realizar un ensaio clínico[4] co obxectivo de valorar o efecto dun certo medicamento na remisión dunha enfermidade mortal. A variable de interese neste caso sería $T =$ *"tempo de supervivencia despois de comezar o tratamento co novo fármaco"*. Dada esta situación, é posible que para cando remate o estudo existan supervivintes, é dicir, individuos que aínda non sufriron o evento de interese, pero inflúen na función de supervivencia que queremos estimar. Referirémonos ás observacións incompletas asociadas a estes individuos como datos censurados.

De forma máis xeral, podemos considerar unha mostra de n variables aleatorias independentes seguindo a mesma distribución que a variable T:

$$\{T_1, \dots, T_i, \dots, T_n\}$$

Diremos que a mostra é completa se proporciona un valor concreto para cada unha das n variables correspondentes. No caso de que esta información non estea dispoñible para cada un dos T_i e obteñamos unicamente unha información parcial, diremos que estamos ante un **dato censurado**. Existen distintos tipos de datos censurados, tal e como se detalla na seguinte subsección.

[4] Un **estudo clínico** é un estudo realizado coa finalidade de investigar a seguridade ou eficacia dun certo medicamento.

1.2.1. Clasificación dos datos censurados

Poderemos representar unha mostra con datos censurados mediante un par de variables aleatorias estritamente positivas (T, C), onde T representa o tempo de supervivencia e C o tempo de censura. A clasificación que se presenta a continuación é a proporcionada en [1].

Datos censurados pola dereita

Cando o dato censurado T_i estea limitado inferiormente ($C_i < T_i$), diremos que está censurado pola dereita. Dentro desta forma de censura, podemos distinguir tres tipos:

- Tipo I
  Neste caso, a cantidade C_i está predeterminada. No exemplo do ensaio clínico, esta censura podería vir determinada pola data de finalización do mesmo; e dicir, no caso de que o tempo de supervivencia dun individuo sexa superior á data de finalización do estudo, este valor será considerado un dato censurado pois non pode ser completamente observado.

- Tipo II
  En lugar de fixar unha cantidade C_i, poderíamos fixar a proporción de individuos que deben sufrir o evento de interese. Así, no exemplo do produto industrial que presentabamos anteriormente, o estudo remataría cando fallase un número r de artiluxios.

- Tipo III
  Este último refírese a un tipo de censura aleatoria, onde a cota C_i non está predeterminada e ten unha certa compoñente aleatoria. No exemplo do ensaio clínico, podería ser que algún participante falecese por unha causa distinta á enfermidade, que decidise deixar de participar no estudo antes do remate do mesmo ou que tivese que deixar de tomar o fármaco de estudo.

Dada unha mostra $\{(T_1, C_1), (T_2, C_2), \dots, (T_n, C_n)\}$ extraída dunha distribución (T, C) con censura pola dererita, na práctica poderemos representa-los datos mediante un par de variables aleatorias (Z, δ) onde $\forall i \in \{1, \dots, n\}$,

$$Z_i = \text{mín}(T_i, C_i) \quad \text{e} \quad \delta_i = \mathbb{I}(T_i \leq C_i).$$

A variable δ coñécese como **variable de censura** e a variable Z recolle os valores observados. Desta forma, se o dato i-ésimo está censurado, $\delta_i = 0$ e $Z_i = C_i$; e no caso contrario, $\delta_i = 1$ e $Z_i = T_i$.

Na Figura 1.2 represéntanse os tempos de supervivencia de 6 doentes que padecen cancro de pulmón e que participaron nun estudo no que se levou un seguemento das/os pacientes durante

800 días. Os datos correspondentes ás/aos pacientes 2 e 4 están censurados pola dereita, o primeiro con censura Tipo III e o segundo con censura Tipo I.

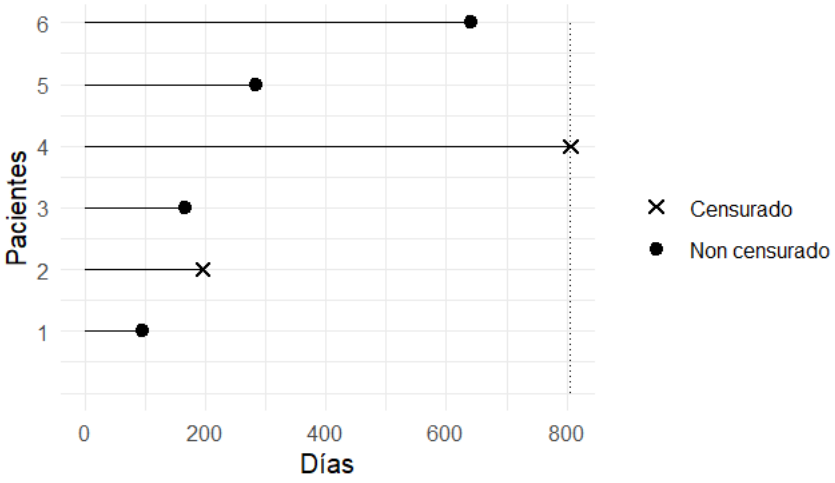


Figura 1.2: Tempo de supervivencia de 6 pacientes trala diagnose dun cancro de pulmón.

Datos censurados pola esquerda

Analogamente, un dato T_i estará censurado pola esquerda se está limitado superiormente ($T_i < C_i$). A distinción entre os Tipos I, II e III para a censura pola dereita é análoga para a censura pola esquerda. Vexamos un exemplo: no caso dun estudo pediátrico no que se está a analizar a idade á que se dan os primeiros pasos, aquelas/es nenas/os que saiban camiñar antes de entrar no estudo serán considerados datos censurados pola esquerda. De novo, representaremo-los datos cun par (Z, δ), coa diferencia de que neste caso

$$Z_i = \text{máx}(T_i, C_i) \quad \text{e} \quad \delta_i = \mathbb{I}(T_i \geq C_i).$$

Datos censurados dobremente

Chegados a este punto, resulta natural considerar un escenario no que existan datos censurados tanto pola dereita ($T_i > C_{ri}$) como pola esquerda ($T_j < C_{li}$). A representación dos datos neste caso sería mediante dúas variables aleatorias (T, δ) tales que

$$Z_i = \text{máx}[\text{mín}(T_i, C_{ri}), C_{li}] \quad \text{e} \quad \delta_i = \begin{cases} 1 & \text{se } C_{li} \leq X_i \leq C_{ri} \\ 0 & \text{se } T_i > C_{ri} \\ -1 & \text{se } T_i < C_{li} \end{cases}.$$

Datos censurados nun intervalo

Como xeralización do visto ata agora, podemos considerar un tempo de supervivencia pertencente a un intervalo ($T_i \in [L, R]$). Este tipo de censura normalmente resulta de estudos nos que se monitorea de forma periódica aos suxeitos. No caso do dispositivo industrial, se se fai unha revisión semanal do seu funcionamento e este falla na cuarta revisión, terase que o momento no que deixou de funcionar se produciu entre a terceira e cuarta semana, e a censura darase nun intervalo da forma $(L, R]$.

Os distintos tipos de censura, requiren diferentes métodos estadísticos para a súa análise. Ao longo deste Traballo de Fin de Grao centrarémonos en **datos censurados aleatoriamente pola dereita**.

1.3. Datos truncados

Os datos censurados poden ser confundidos cos **datos truncados** polas semellanzas que comparten. Estes últimos tamén son datos para os que só coñecemos unha información parcial; non obstante, no caso do truncamento a información adoita estar incompleta como consecuencia do deseño do propio estudo. Volvamos ao exemplo que mencionabamos no apartado para censura pola esquerda no que se estuda a idade á que unha/un nena/o dá os seus primeiros pasos. Supoñamos que o estudo está deseñado de forma que a idade de entrada no estudo das/os participantes sexa de 10 meses. Aquelas/es nenas/os que entren a formar parte do estudo a unha idade superior serán considerados datos truncados (pola esquerda). É habitual que en ensaios clínicos como o que acabamos de describir se produza este tipo de truncamento (truncamento pola esquerda) acompañado de censura pola dereita. Pode atoparse máis información sobre este tipo de situacións en [1].

A forma de enfrenta-los datos truncados é diferente á dos datos censurados, neste traballo abordaremos unicamente a análise con estes últimos.

1.4. Estrutura do traballo e librarías empregadas en R

Nos seguintes capítulos abordaremos as problemáticas que orixina o tratamento de datos censurados á hora de estimar a función distribución da variable aleatoria T nos casos condicional e non condicional mediante a realización de diversos estudos de simulación no software estatístico R. A continuación, presentamos brevemente cada unha das partes nas que se estrutura este Traballo de Fin de Grao e as principais librarías de R empregadas para o desenvolvemento

dos estudos de simulación e a análise de datos reais que serán levadas a cabo ao longo deste documento.

No Capítulo 2, estudaremos o comportamento da función de distribución empírica e do estimador de Kaplan-Meier como estimadores da función de distribución dunha variable T que representa o tempo de supervivencia e desenvolveremos algunhas das propiedades máis destacables deste último. Ao longo dos estudos de simulación realizados neste capítulo, faremos uso das librarías `ReIns` (ver referencia [2]) e `Bolstad2`. O primeiro, empregarémolo na computación do estimador de Kaplan-Meier, mentres que o segundo permitirá calcular aproximacións de Simpson cando sexa preciso determinar o valor dunha integral. No Capítulo 3, presentaremos dous estimadores para a función de distribución da variable T condicionada a outra variable aleatoria continua X: a función de distribución condicional empírica e o estimador de Beran. Empregaremos as librarías `npcure` (ver referencia [3]) e `VGAM` para a computación destes estimadores condicionais e para xeración dunha mostra normal bivariante, respectivamente.

Ao longo do Capítulo 4 tomaremos un efoque máis práctico e levaremos a cabo a análise dun conxunto de datos reais con censura empregando os estimadores presentados nos Capítulos 2 e 3. Para isto apoiarémonos principalmente na libraría `Survival` (ver referencia [4]) e nas librarías xa mencionadas. Finalmente, no Capítulo 5 incluiremos as principais conclusións que se derivan da realización deste Traballo de Fin de Grao.

Inclúense ademais, dous anexos coa sintaxe de R empregada nos distintos estudos de simulación (Anexo A) e na creación das diferentes representacións gráficas (Anexo B) que se presentan ao longo de todo o documento.

Capítulo 2

O estimador de Kaplan-Meier

Dedicaremos este capítulo a explicar a problemática que orixina a utilización de datos censurados na estimación da función de distribución. Ilustraremos o mal comportamento dos métodos clásicos grazas a pequenos estudos de simulación levados a cabo no programa estatístico R. Ademais, presentaremos un estimador non paramétrico que aborde a estimación da función de distribución asociada a unha mostra con datos censurados: o estimador de Kaplan-Meier. Para o desenvolvemento deste capítulo seguiremos principalemente [5] e [6].

2.1. A función de distribución empírica

Dado un fenómeno aleatorio que non se poida modelar empregando ningunha familia paramétrica coñecida, unha ferramenta útil na estimación da distribución dos datos dunha mostra é a **función de distribución empírica**. Consideremos $\{X_1, X_2, \dots, X_n\}$ unha mostra de n variables aleatorias independentes, identicamente distribuídas e coa mesma función de distribución $F(x)$ descoñecida. A función F pode estimarse a través da función de distribución empírica que vén dada por,

$$\widehat{F}_n(x) = \frac{\#\{i : X_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x),$$

é dicir, asignaralle a cada valor da mostra unha probabilidade de $1/n$. O seguinte resultado permitiranos derivar a converxencia de $\widehat{F}_n$ á función de distribución poboacional F.

Teorema 2.1 (Teorema de Glivenko-Cantelli)**.** *Sexa $X_1, X_2, ..., X_n$ unha mostra de variables aleatorias coa mesma función de distribución F e sexa $\widehat{F}_n$ a función de distribución empírica asociada. Tense que*

$$\left\|\widehat{F}_n - F\right\|_\infty = \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \longrightarrow 0$$

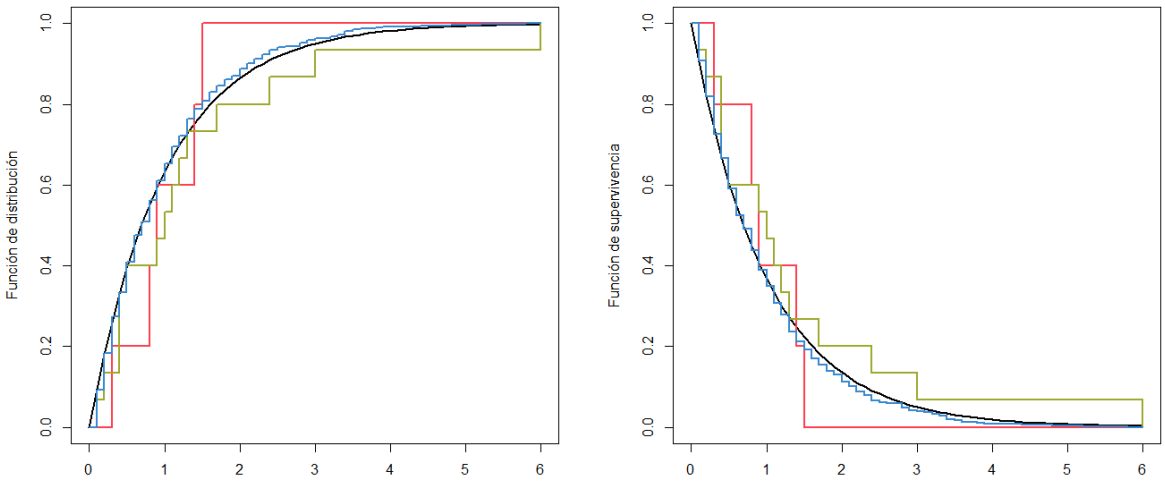
con probabilidade 1.

En particular, terase que $\widehat{F}_n(x_0) \longrightarrow F(x_0)$ para calquera $x_0 \in \mathbb{R}$ fixo pertencente ao conxunto no que se dá a converxencia case segura. Ademais, posto que $\mathbb{E}[\widehat{F}_n(x_0)] = F(x_0)$, podemos concluír que $\widehat{F}_n(x_0)$ é un estimador sen nesgo[1] de $F(x_0)$.

Observación 2.2. Supoñendo que a mostra estivese ordeada e que $X_1 \leq \cdots \leq X_{j-1} \leq X_j \cdots \leq X_n$, poderemos estimar $\mathbb{P}(X \leq X_j | X > X_{j-1})$ por $1/(n-j)$.

No contexto da Análise de Supervivencia, se se considera unha mostra aleatoria simple $T_1, \ldots, T_n$ dunha certa variable T, onde T é o tempo de supervivencia, a **función de supervivencia empírica** asociada será

$$\widehat{S}_n(t) = \frac{\#\{i : T_i > t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(T_i > t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(T_i \leq t) = 1 - \widehat{F}_n(t). \tag{2.1}$$



(a) Función de distribución (b) Función de supervivencia

Figura 2.1: Funcións de distribución e supervivencia empíricas correspondentes a tres mostras de tamaño $n = 5$ (en vermello), 15 (en verde) e 500 (en azul) e asociadas a unha variable $X \in Exp(1)$. En trazo negro represéntanse as curvas poboacionais.

Para ilustrar o comportamento da función de distribución empírica e da función de supervivencia empírica, presentamos na Figura 2.1 as curvas obtidas a partires de tres mostras aleatorias dunha distribución expoñencial de parámetro 1 de diferente tamaño. Os tamaños de mostra considerados foron $n \in \{5, 15, 500\}$; presentamos as estimación para $n = 5$ en cor vermella, para

[1] Sexa $\widehat{\theta}$ un estimador dun certo parámetro descoñecido θ. O **nesgo** do estimador $\widehat{\theta}$ é a diferenza entre a esperanza do estimador $\widehat{\theta}$ e o valor do parámetro a estimar θ. Dirase que un estimador non ten nesgo no caso de que $\mathbb{E}(\widehat{\theta}) - \theta = 0$.

$n = 15$ en cor verde e para $n = 500$ en cor azul xunto coas curvas poboacionais en cor negra. Como podemos ver, ámbalas dúas son funcións escalonadas nas que a altura dos escalóns depende do tamaño da mostra e a lonxitude dos valores recollidos. Ademais, tamén se observa como a medida que aumenta o tamaño de mostra as funcións empíricas converxen ás funcións poboacionais, tal e como se deriva do Teorema 2.1.

2.1.1. Comportamento da función de distribución empírica para datos censurados

Acabamos de ver que para unha mostra con datos completos a función de distribución empírica é un estimador consistente[2] da función de distribución da variable aleatoria a estudar. O obxectivo desta sección será determinar, a través dun pequeno **estudo de simulación por Montecarlo**[3] no software estatístico R, se este comportamento se mantén no caso de que algúns dos datos da mostra estean censurados.

Procedemento de simulación

O primeiro paso consistirá en xerar unha mostra con datos censurados pola dereita e con censura aleatoria (ou de Tipo III). Para isto consideraremos dúas variables aleatorias independentes T e C, onde T é a variable a estudar e C a variable correspondente á censura. Representaremos a mostra a través dun par de variables (Z, δ), onde $Z = \text{mín}(T, C)$ e $\delta = \mathbb{I}(T \leq C)$, tal e como quedou detallado na Sección 1.2.1.

Consideraremos $T \in \mathcal{N}(0,1)$[4] e $C \in \mathcal{N}(\mu, 0.5^2)$[5], onde o valor de μ varía segundo a porcentaxe de censura que queiramos considerar. Ao tomar unha varianza para C máis baixa que a de T podemos calibrar mellor a proporción de datos censurados da mostra. Na Figura 2.2 represéntanse as funcións de densidade correspondentes ás variables T e C para porcentaxes de

[2] A **consistencia** é unha propiedade de certos estimadores pola cal ao aumentar infinitamente o tamaño da mostra, o estimador tende ao seu valor esperado de forma case segura. A consistencia de $\widehat{F}_n$ quedou vista no Teorema 2.1.

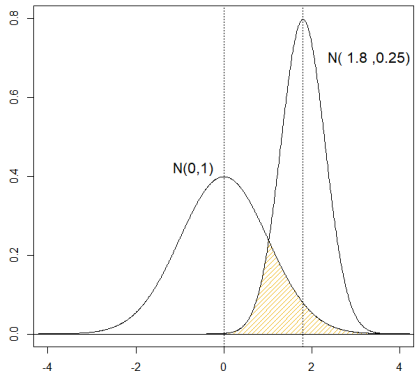
[3] O **método de Montecarlo** baséase na xeración de múltiples mostras aleatorias co fin de obter unha aproximación numérica dun certo parámetro asociado a ditas mostras, como poden ser a súa media, a súa varianza ou a súa función de distribución.

[4] Diremos que unha variable aleatoria X segue unha **distribución normal** de media μ e varianza σ^2 ($X \sim \mathcal{N}(\mu, \sigma^2)$), se a súa función de densidade é da forma

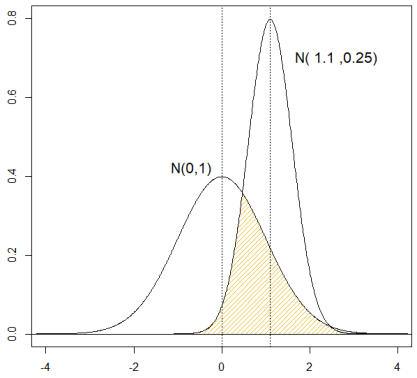
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du, \quad x \in \mathbb{R}.$$

[5] Aínda que no contexto da Análise de Supervivencia T e C son variables non negativas, nesta sección non teremos en conta esta restrición xa que non ten efecto algún sobre o obxectivo da simulación.

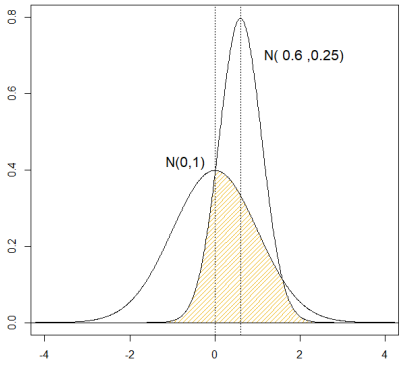
censura do 5 %, 15 % e 30 % coa idea de visualizar os diferentes escenarios a considerar no estudo de simulación.



(a) Censura do 5 %, $C \in \mathcal{N}(1.8, 0.5^2)$.



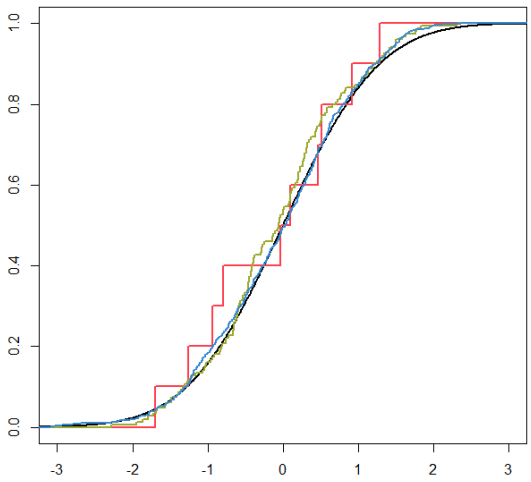
(b) Censura do 15 %, $C \in \mathcal{N}(1.1, 0.5^2)$.



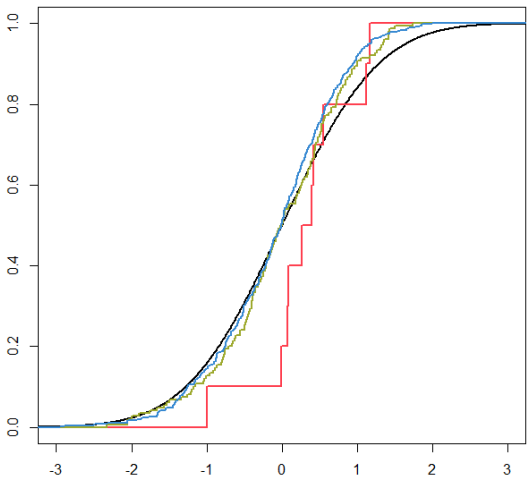
(c) Censura do 30 %, $C \in \mathcal{N}(0.6, 0.5^2)$.

Figura 2.2: Funcións de densidade das variables T e C correspondentes ás distintas porcentaxes de censura consideradas. En amarelo destácase o solapamento entre as dúas funcións de densidade. Canto maior é o solapamento, maior é a porcentaxe de censura.

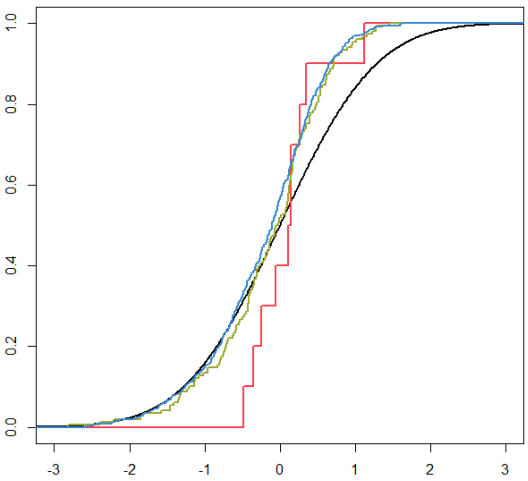
Como primeira achega ao problema, representamos na Figura 2.3 as funcións de distribución empíricas derivadas das mostras de datos censurados xunto coa función de distribución poboacional da variable T. Pode apreciarse como a medida que aumenta a porcentaxe de censura, as funcións de distribución empíricas (incluso para as mostras máis grandes) se alonxan da parte superior da función de distribución poboacional de $T \in \mathcal{N}(0, 1)$. Este feito dá pé a valorar a inconsistencia de $\widehat{F}_n$ como estimador de F na presenza de datos censurados.



(a) Censura do 5 %.



(b) Censura do 15 %.



(c) Censura do 30 %.

Figura 2.3: Funcións de distribución empíricas asociadas á variable observada Z correspondentes ás distintas porcentaxes de censura, para mostras de tamaños $n = 10, 150, 500$ (en vermello, verde e azul, respectivamente). A curva negra representa a función de distribución poboacional de $T \in \mathcal{N}(0, 1)$.

Coa idea de cuantificar estas diferenzas entre $\widehat{F}_n$ e F, recorreremos a un criterio global de erro coñecido como erro cadrático integrado e referido habitualmente polas súas siglas en inglés como

ISE[6]. No noso caso, serviranos para comparar a función de distribución de T que denotaremos por $F(x)$, coa función de distribución empírica de Z, denotada como $\widehat{F}_n(x)$. É dicir, imos calcular:

$$\text{ISE} = \int \left(\widehat{F}_n(x) - F(x)\right)^2 dx.$$

Para asegurar que o fenómeno presentado na Figura 2.3 non foi casual, realizaremos un estudo de simulación por Montecarlo no que xeraremos 1000 mostras de diversos tamaños e calcularemos unha aproximación do ISE para cada unha delas empregando a regra de Simpson. A continuación, empregaremos como valor comparativo entre as distintas porcentaxes de censura a esperanza do ISE para cada tamaño de mostra. Este valor é o que se coñece como **erro cadrático medio integrado** ou **MISE**, e vén dado por:

$$\text{MISE} = \mathbb{E}\left(\int \left(\widehat{F}_n(x) - F(x)\right)^2 dx\right).$$

Na Táboa 2.1 preséntanse os resultados do estudo de simulación para os distintos tamaños de mostra e as distintas porcentaxes de censura, incluíndo un grupo control sen censura.

	Porcentaxe de censura			
	0 %	5 %	15 %	30 %
$n = 10$	560.86	555.53	537.04	633.58
$n = 150$	35.84	39.17	78.34	229.78
$n = 500$	11.49	14.19	54.48	211.76

Táboa 2.1: Valores do MISE (multiplicados por 1000) para os distintos tamaños de mostra e as distintas porcentaxes de censura obtidos a través do estimador $\widehat{F}_n$ e sendo $T \in \mathcal{N}(0, 1)$.

Á vista dos resultados presentados na Táboa 2.1, observamos como para calquera das porcentaxes de censura o MISE diminúe a medida que aumenta o tamaño de mostra. Fixémonos no caso da mostra cun 0 % de censura (é dicir, sen presenza de datos censurados), observamos como o MISE converxe a 0 a medida que aumenta o tamaño de mostra, o cal é unha proba da consistencia do estimador presentada no Teorema 2.1. Este comportamento pode observarse tamén para as porcentaxes de censura do 5 % ou do 15 %, non obstante, vese "desacelerado" a medida que aumenta a presenza de censura. Esta desacerelación é máis salientable para o caso da porcentaxe de censura do 30 %, pois o valor do MISE asociado a $n = 500$ é case 20 veces superior ao que esperaríamos para unha mostra sen censura.

[6]O **ISE** (do inglés *integrated squared error*) é unha medida da distancia entre dúas funcións f e $\widehat{f}$, onde $\widehat{f}$ é un estimador de f:

$$\text{ISE} = \int \left(\widehat{f}(x) - f(x)\right)^2 dx$$

Estes resultados van na liña das estimacións presentadas na Figura 2.3 e ilustran a inconsistencia da función de distribución empírica como estimador da función de distribución na presenza de datos censurados. Á vista deste estudo de simulación ponse de manifesto a necesidade de presentar un novo estimador da función de distribución para datos censurados. Dito estimador é coñecido como estimador de Kaplan-Meier e será presentado na seguinte sección.

2.2. O estimador de Kaplan-Meier

Co fin de dar solución ao problema da estimación da función de distribución a partir dunha mostra con datos censurados, os investigadores Eulid L. Kaplan e Paul Meier propuxeron no ano 1958 (ver referencia [6]) o **estimador do límite-produto**, tamén coñecido como **estimador de Kaplan-Meier**. A idea principal deste estimador consiste en modificar os pesos de $1/n$ que lle asigna a función de distribución empírica a cada elemento da mostra no caso de que algúns datos estean censurados. Polo tanto o estimador de Kaplan-Meier será da forma:

$$\widehat{F}_{KM}(x) = \sum_{i=1}^{n} w_i \mathbb{I}(X_i \leq x), \tag{2.2}$$

onde as cantidades w_i respresentan estes pesos modificados, que serán coñecidos como pesos Kaplan-Meier.

Para construír formalmente o estimador, partiremos dunha mostra aleatoria simple do par (T, C) con censura pola dereita

$$\{(T_1, C_1), \ldots, (T_n, C_n)\}.$$

Suporemos ademais que as variables T e C son independentes e non negativas. Recorrendo á notación xa presentada anteriormente, poderemos representar a mostra observada a través de dúas variables aleatorias (Z, δ) onde $Z = \text{mín}(T, C)$ e $\delta = \mathbb{I}(T \leq C)$ é o indicador de censura. Suporemos tamén que as n observacións da variable Z están ordeadas e que non existen dúas observacións que tomen o mesmo valor. E teremos en conta ademais unha observación $Z_0 = 0$:

$$0 = Z_0 < Z_1 < Z_2 < \ldots < Z_n.$$

O seguinte paso consistirá en intentar estimar a probabilidade de que un individuo sofra o evento de interese antes dun tempo Z_k tendo en conta que sobreviviu ata o momento Z_{k-1}. Isto sería,

$$h_k = \mathbb{P}(T \leq Z_k | T > Z_{k-1}) = \frac{\mathbb{P}(T \in (Z_{k-1}, Z_k])}{\mathbb{P}(T > Z_{k-1})}.$$

A proposta para a estimación desta probabilidade condicionada é a seguinte:

1. Se Z_k se corresponde cunha observación censurada, tomarase $\widehat{h_k} = 0$.

2. Se Z_k se corresponde cunha observación non censurada, o estimador $\widehat{h_k}$ será $1/(n-(k-1))$ onde $n-(k-1)$ é o número de individuos que sobreviviron ata Z_{k-1} (en coherencia co mencionado na Observación 2.2), ou o que é o mesmo: $1/\#\{j : Z_j \geq Z_k\}$.

A continuación, reescribiremos a función de supervivencia da variable T avaliada nun instante Z_k empregando probabilidades condicionais como

$$\begin{aligned} S(Z_k) &= \mathbb{P}(T > Z_k), \\ &= \mathbb{P}(T > Z_k | T > Z_{k-1})\mathbb{P}(T > Z_{k-1}), \\ &= \mathbb{P}(T > Z_k | T > Z_{k-1}) \dots \mathbb{P}(T > Z_1 | T > Z_0)\mathbb{P}(T > Z_0), \\ &= \prod_{i=1}^{k} \mathbb{P}(T > Z_i | T > Z_{i-1}) = \prod_{i=1}^{k} (1 - h_k)\,. \end{aligned}$$

Así, tendo en conta os estimadores propostos anteriormente:

$$\widehat{S}_{KM}(Z_k) = \prod_{i=1}^{k} \left(1 - \widehat{h}_k\right) = \prod_{i=1}^{k} \left(1 - \frac{1}{n-(i-1)}\right)^{\delta_i} = \prod_{Z_i \leq Z_k} \left(1 - \frac{1}{\#\{j : Z_j \geq Z_i\}}\right)^{\delta_i}.$$

Como este valor será o mesmo para calquera instante t do intervalo $[Z_k, Z_{k+1})$, poderemos definir o estimador de $S(t)$ como

$$\widehat{S}_{KM}(t) = \prod_{Z_i \leq t} \left(1 - \widehat{h}_k\right) = \prod_{Z_i \leq t} \left(1 - \frac{1}{\#\{j : Z_j \geq Z_i\}}\right)^{\delta_i}.$$

Este estimador da función de supervivencia é o que se coñece como **estimador de Kaplan-Meier**, e permite construír o seguinte estimador para a función de distribución $F(t)$:

$$\begin{aligned} \widehat{F}_{KM}(t) &= 1 - \widehat{S}_{KM}(t), \\ &= 1 - \prod_{Z_i \leq t} \left(1 - \frac{1}{\#\{j : Z_j \geq Z_i\}}\right)^{\delta_i}, \\ &= \sum_{j=1}^{n} \left(\frac{\delta_j}{n-j+1} \prod_{i=1}^{j-1} \left(1 - \frac{\delta_i}{n-i+1}\right)\right) \mathbb{I}(Z_j \leq t). \end{aligned} \tag{2.3}$$

Nótese que denotando $w_j = \frac{\delta_j}{n-j+1} \prod_{i=1}^{j-1} \left(1 - \frac{\delta_i}{n-i+1}\right)$[7], chegamos a unha expresión como a de (2.2) que presentabamos ao comezo desta sección.

Observación 2.3. O estimador de Kaplan-Meier $\widehat{F}_{KM}$ coincide coa da función de distribución empírica no caso de que non existan datos censurados na mostra, pois tomando $t \in [Z_k, Z_{k+1})$, tense que

$$\widehat{F}_{KM}(t) = 1 - \widehat{S}_{KM}(t) = \prod_{Z_i \leq t} \left(1 - \frac{1}{n-(i-1)}\right)^{\delta_i} = \prod_{i=1}^{k} \left(\frac{n-i}{n-i+1}\right) = \frac{k}{n} = \widehat{F}_n(t).$$

[7] Hai un abuso de notación na escritura do produto, entenderemos que vale 1 cando $j = 1$.

Observación 2.4. Debemos prestar especial atención cando a última observación Z_n estea censurada pois nese caso, desenvolvendo a expresión (2.3) teríamos que

$$\widehat{F}_{KM}(Z_n) = 1 - \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{\delta_i} = 1 - \prod_{i=1}^{n-1} \left(\frac{n-i}{n-i+1} \right)^{\delta_i} < 1,$$

co que $\widehat{F}_{KM}$ non cumpliría unha das características propias de toda función de distribución, que é que toma valores no invervalo $[0, 1]$. Polo tanto, considerarase en calquera caso que $\widehat{F}_{KM}(t) = 1$ para $t \geq Z_n$.

Valoremos agora o caso no que se repita algún dos valores de Z_k. Nesta situación será preciso ter en conta as seguintes consideracións:

- Se nun instante Z_k se produce tanto a censura como a aparición do evento de interese para o mesmo ou distintos individuos, consideraremos que a aparición do evento de interese se produciu lixeiramente antes que Z_k e a censura lixeiramente despois de Z_k.

- Se dous ou máis elementos da mostra están censurados nun mesmo instante Z_k, non faremos cambios sobre o estimador.

- Se dous ou máis elementos da mostra sofren o evento de interese nun mesmo instante Z_k, o novo estimador para h_k será $\widehat{h}_k = d_k/n_k$, onde d_k representa o número de elementos da mostra que sufriron o evento de interese no mesmo instante Z_k, e n_k o número de individuos que sobreviviron ata polo menos Z_{k-1} e que están en risco de sufrir o evento de interese.

A forma que tomaría o estimador baixo estas condicións sería:

$$\widehat{S}_{KM}(t) = \prod_{Z_i < t} \left(1 - \frac{d_i}{n_i} \right)^{\delta_i}. \tag{2.4}$$

Exemplo 2.5. Consideremos de novo a mostra cos tempos de supervivencia (en días) de 6 pacientes trala diagnose dun cancro de pulmón que ilustrabamos na Figura 1.1 do Capítulo 1: $95, 167, 196+, 284, 641, 806+$ (o símbolo $+$ indica censura). Na Figura 2.4 representamos o estimador $\widehat{F}_{KM}$ asociado a dita mostra. Ata a segunda observación, $\widehat{F}_{KM}$ compórtase como $\widehat{F}_n$. Unha vez que aparece un dato censurado o valor de $\widehat{F}_{KM}$ non varía ata a próxima observación non censurada. Nótese ademais que para o último dato da mostra aplica o visto na Observación 2.4.

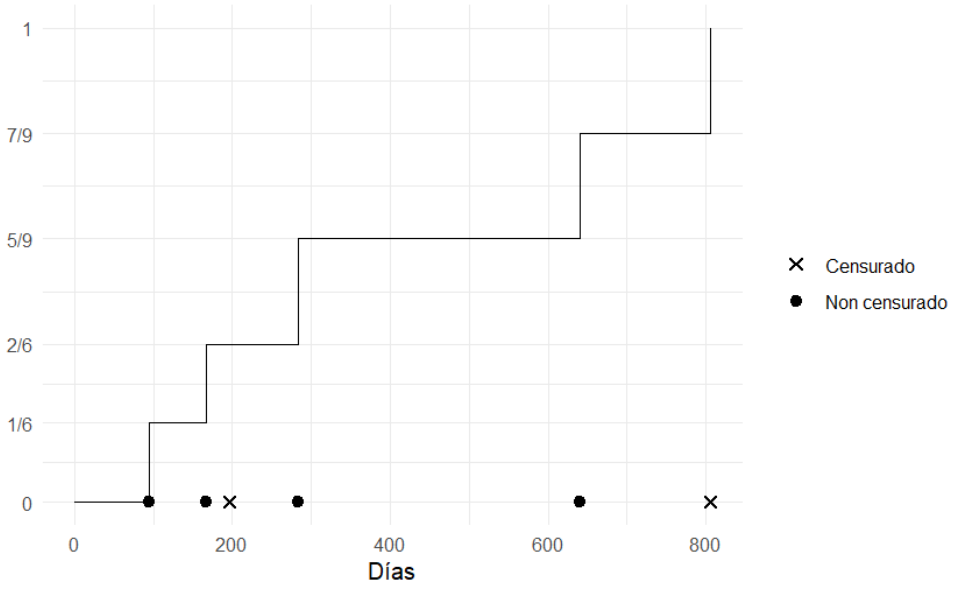


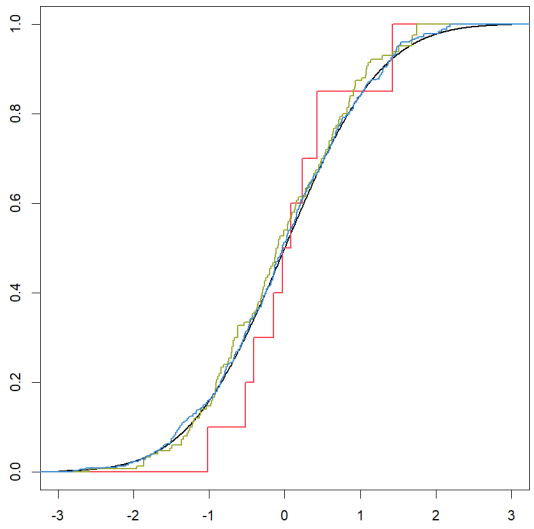
Figura 2.4: Estimador de Kaplan-Meier para a función de distribución asociada aos datos do exemplo ilustrado na Figura 1.1.

2.2.1. Comportamento do estimador de Kaplan-Meier na presenza de datos censurados

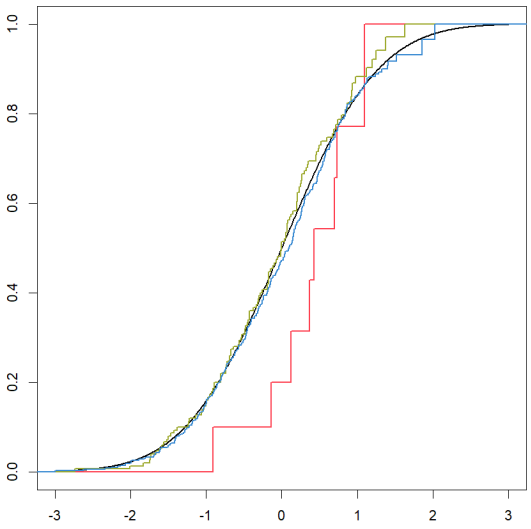
O obxectivo deste novo estudo de simulación será repetir o procedemento descrito na Sección 2.1.1 empregando agora o estimador de Kaplan-Meier como estimador da función de distribución na presenza de datos censurados. Para a computación deste estimador, empregaremos a libraría `ReIns` de R (ver referencia [2]).

Na Figura 2.5, aparece unha representación do estimador $\widehat{F}_{KM}$ paralela á da Sección 2.1.1. Obsérvase como o fenómeno de separación das $\widehat{F}_n$ con respecto á función de distribución teórica na parte superior está prácticamente corrixido. Non obstante, a medida que aumenta a censura a calidade desta corrección diminúe. Para o caso da porcentaxe de censura do 30 %, observamos un gran escalón final que semella impedir o correcto axuste á función de distribución teórica para os valores representados máis elevados. Este comportamento do estimador non é nada sorprendente pois ao aumentar o número de observacións censuradas diminúe o número de puntos nos que o estimador cambia de valor (dá un salto) e consecuentemente obtemos unha peor aproximación.

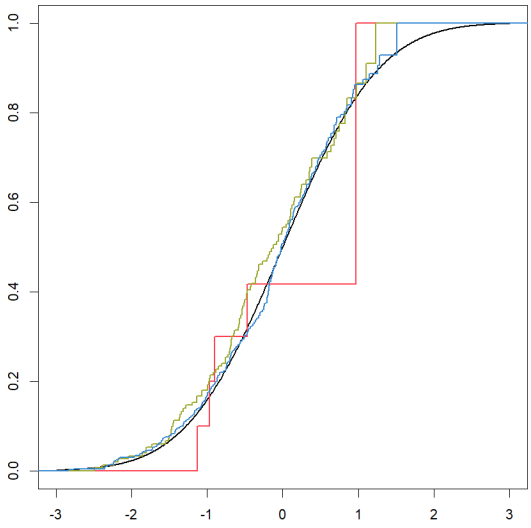
Realizamos polo tanto unha nova simulación, agora co estimador de Kaplan-Meier. Os tamaños de mostra e as variables escollidas son as mesmas que as da Sección 2.1.1: $n \in \{10, 150, 500\}$, $T \in \mathcal{N}(0,1)$ e $C \in \mathcal{N}(\mu, 0.5^2)$, onde $\mu \in \{1.8, 1.1, 0.6\}$ para porcentaxes de censura do 5 %, 15 % e 30 %, respectivamente. Os resultados poden verse na Táboa 2.2.



(a) Censura do 5 %.



(b) Censura do 15 %.



(c) Censura do 30 %.

Figura 2.5: Estimador de Kaplan-Meier da función de distribución da variable Z correspondentes ás distintas porcentexes de censura, para mostras de tamaños $n \in \{10, 150, 500\}$ (en vermello, verde e azul, respectivamente). En negro a función de distribución poboacional da variable $T \in \mathcal{N}(0, 1)$.

Comparando os novos resultados (ver Táboa 2.2) cos da anterior simulación (ver Táboa 2.1), notamos como os valores do MISE diminúen notablemente para tódalas porcentaxes de censura superiores ao 0 %. No escenario sen datos censurados, os valores son exactamente os mesmos que obtivemos na anterior simulación, en concordancia co comentado na Observación 2.3 onde explicabamos que en ausencia de censura $\widehat{F}_{KM} = \widehat{F}_n$. Para a porcentaxe de censura do 5 % os valores do MISE parecen propios dunha mostra sen presenza de datos censurados. A mellora é especialmente notable para a porcentaxe de censura máis alta, xa que no caso de $n = 500$ o MISE é case 10 veces menor que o presentado na Táboa 2.1. A priori, parece que podemos intuír certa consistencia do estimador, aínda que desenvolveremos formalmente esta cuestión nas seguintes seccións.

	Porcentaxe de censura			
	0 %	5 %	15 %	30 %
$n = 10$	560.86	576.48	583.79	677.25
$n = 150$	35.84	39.14	47.62	66.90
$n = 500$	11.49	11.69	14.89	25.76

Táboa 2.2: Valores do MISE (multiplicados por 1000) para os distintos tamaños de mostra e as distintas porcentaxes de censura obtidos a través do estimador $\widehat{F}_{KM}$, sendo $T \in \mathcal{N}(0, 1)$.

Resultados considerando outras distribucións

A baixa calidade de $\widehat{F}_n$ como estimador da función de distribución en presencia de censura non é nin moito menos algo exclusivo da distribución normal coa que estivemos traballando ata agora. Para confirmar este feito, consideraremos agora dúas variables $T \in Exp(0.5)$ e $C \in \mathcal{N}(\mu, 0.5^2)$, onde $\mu \in \{3, 2, 1.3\}$ para porcentaxes de censura aproximadas do 5 %, 15 % e 30 % respectivamente. Realizamos un estudo de simulación seguindo o procedemento anterior, onde comparabamos os estimadores $\widehat{F}_n$ e $\widehat{F}_{KM}$ coa función de distribución poboacional F da nova variable T. Na Táboa 2.3 preséntanse os resultados deste novo estudo de simulación.

En case tódolos casos, os resultados obtidos co estimador de Kaplan-Meier melloran os do estimador $\widehat{F}_n$. Os únicos que non o fan son os asociados ás observacións censuradas de tamaño $n = 10$. Resulta razoable que ocorra isto cun tamaño de mostra tan pequeno xa que o estimador de Kaplan-Meier se ve forzado a facer a estimación con 7 ou 8 valores unicamente, resultando nunha peor aproximación. Nos outros tamaños de mostra o MISE diminúe, sendo especialmente notable para a porcentaxe de censura do 30 %, pois observamos como para $n = 500$ o MISE é seis veces menor que o obtido mediante a distribución empírica. Finalmente salientar que para unha porcentaxe de censura do 0 % os valores de MISE coinciden para os dous estimadores

considerando calquera tamaño de mostra (como era de esperar posto que, sen presenza de datos censurados, ambos estimadores coinciden).

		Porcentaxe de censura			
		0 %	5 %	15 %	30 %
$\widehat{F}_n$	$n = 10$	478.99	461.49	452.46	583.08
	$n = 150$	30.92	39.17	94.88	311.38
	$n = 500$	9.52	16.80	77.53	297.11
$\widehat{F}_{KM}$	$n = 10$	478.99	473.41	516.26	630.27
	$n = 150$	30.92	36.04	54.75	109.83
	$n = 500$	9.52	10.77	23.82	57.54

Táboa 2.3: Valores do MISE (multiplicados por 1000) para os distintos tamaños de mostra e as distintas porcentaxes de censura obtidos a través dos estimadores $\widehat{F}_n$ e $\widehat{F}_{KM}$, sendo $T \in Exp(1)$.

2.3. Propiedades do estimador de Kaplan-Meier

Nesta sección, veremos que o estimador de Kaplan-Meier é tamén un estimador de máxima verosimilitude non paramétrico e presentaremos un resultado relativo á consistencia do mesmo. Ademais, deduciremos un estimador da varianza que permitirá a construción de intervalos de confianza para o estimador de Kaplan-Meier.

2.3.1. Estimador de máxima verosimilitude non paramétrico

Outra forma de afrontar o problema da estimación da función de distribución dunha variable aleatoria X sería recorrendo a un modelo paramétrico. Baixo a suposición de que a variable a estudar pertence a unha determinada familia de funcións de distribución, un podería determinar o parámetro que caracteriza a distribución de X mediante un método como o de **máxima verosimilitude**[8]. Porén, cando a suposición inicial non é adecuada, este tipo de estimación pode conducir a resultados erróneos.

[8]**Método de máxima verosimilitude** : Sexa $\mathbf{x}_n = (x_1, \ldots, x_n) \in \mathbf{X}^n$ unha posible realización de $\mathbf{X}_n$. A verosimilitude desta realización é a función $L(\cdot, \mathbf{x}_n) : \Theta \to \mathbb{R}^+$ tal que

$$L(\theta, \mathbf{x}_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Se a función $\theta \mapsto L(\theta, \mathbf{x}_n)$ acada o seu máximo global en $\theta = \theta_n(\mathbf{x}_n)$, o estimador de máxima verosimilitude é o estatístico $\widehat{\theta}_n^{MV} = \theta_n(\mathbf{X}_n)$. Este estatístico é consistente e asintóticamente normal.

Os estimadores $\widehat{F}_n$ e $\widehat{F}_{KM}$ que presentamos neste capítulo son estimadores non paramétricos que non precisan de suposicións sobre a forma que toma X. Con todo, a idea do método de máxima verosimilitude que aplica aos modelos paramétricos pode estenderse aos modelos non paramétricos. Así, en lugar de restrinxirnos a unha familia de funcións, estaremos a valorar tódalas funcións de distribución posibles. A continuación, presentaremos unha forma de obter o estimador de Kaplan-Meier a través da función de verosimiltude proposta en [7].

A función de verosimilitude para unha función de supervivencia $S = S(t)$ asociada a unha mostra de datos censurados aleatoriamente pola dereita $\{(Z_1, \delta_1), ..., (Z_n, \delta_n)\}$, ou equivalentemente $\{(T_1, C_1), ..., (T_n, C_n)\}$, poderá escribirse como

$$\begin{aligned} L(S; (Z_1, \delta_1), ..., (Z_n, \delta_n)) &= \prod_{i=1}^{n} \mathbb{P}(T_i = Z_i)^{\delta_i} \mathbb{P}(T_i > Z_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left\{ S\left(Z_i^-\right) - S(Z_i) \right\}^{\delta_i} S(Z_i)^{1-\delta_i}, \end{aligned} \tag{2.5}$$

onde se ten se Z_i é unha observación censurada contribuirá á función de verosimilitude co factor $\mathbb{P}(T_i > Z_i))$ e se é non censurada contribuirá co factor $\mathbb{P}(T_i = Z_i)$. A idea do método de máxima verosimilitude será atopar a función S que maximice a función L dada en (2.5). No caso de que o evento de interese teña lugar no instante Z_i, $S\left(Z_i^-\right) - S(Z_i)$ debe ser distinto de cero e maximizarase cando $S\left(Z_i^-\right) = S(Z_{i-1})$. No caso de que Z_i se corresponda cun dato censurado, o maior valor que pode tomar $S(Z_i)$ é o da anterior observación non censurada. Para poder continuar, denotaremos por $T_1^* < ... < T_j^* < ... < T_r^*$ a aqueles elementos da mostra $\{(Z_i, \delta_i)\}_{1 \leq i \leq n}$ para os que se observou o evento de interese ($\delta_i = 1$). Denotaremos:

- d_j: número de elementos da mostra que sufriron o evento de interese no instante T_j^*.
- n_j: número de elementos da mostra que están en risco de sufrir o evento de interese no instante T_j^*.
- c_j: número de elementos da mostra censurados no intervalo $[T_j^*, T_{j+1}^*)$.

Desta maneira e tendo en conta a nova notación, podemos reescribir a función de verosimilitude como segue,

$$L(S; (Z_1, \delta_1), ..., (Z_n, \delta_n)) = \prod_{j=1}^{r} \left\{ S\left(T_{j-1}^*\right) - S\left(T_j^*\right) \right\}^{d_j} S\left(T_j^*\right)^{c_j}.$$

Escribindo S en función das probabilidades h_i (como vimos ao comezo da Sección 2.2), a función

de verosimilitude dependerá agora dos parámetros $h_1, \dots, h_r$:

$$\begin{aligned} L(S; (Z_1, \delta_1), \dots, (Z_n, \delta_n)) &= \prod_{j=1}^{r} \left\{ h_j^{d_j} \prod_{k=1}^{j-1} (1-h_k)^{d_j} \prod_{k=1}^{j} (1-h_k)^{c_j} \right\} \\ &= \prod_{j=1}^{r} h_j^{d_j} \, (1-h_j)^{n_j - d_j} \, . \end{aligned} \tag{2.6}$$

Cada un dos factores do produto correspóndese cunha función de verosimilitude dunha distribución binomial asociada a unha mostra formada por unha única observación. Lembremos que se $Y \in \mathrm{Bin}(n, p)$, entón

$$L(p, (y_1, \dots, y_r)) = \prod_{i=1}^{r} \mathbb{P}(Y = y_i) = \prod_{i=1}^{r} \binom{n}{y_i} p^{y_i} (1-p)^{n-y_i}$$

e o estimador de máxima verosimilitude para a probabilidade p será $\widehat{p} = \sum_{i=1}^{r} y_i / r$.

Ademais, en cada un dos intervalos $[T_j^*, T_{j+1}^*)$, coñecido o número de elementos en risco n_j, o número de veces que ocorre o evento de interese segue unha distribución binomial, é dicir, $d_j \in \mathrm{Bin}(n_i, h_i)$. Polo tanto, un estimador de máxima verosimilitude para cada h_j será

$$\widehat{h}_j = \frac{d_j}{n_j}.$$

Tendo en conta todo isto, o máximo da expresión (2.6) acadarase tomando $h_j = d_j / n_j$, e desta forma a función de supervivencia que maximiza a función de verosimilitude será

$$\widehat{S}(t) = \prod_{T_j \leq t} \left(1 - \widehat{h}_j\right) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{n_j}\right),$$

que é equivalente á expresión (2.4) do estimador de Kaplan-Meier que definimos na Sección 2.2.

É coñecido que os estatísicos obtidos empregando procedementos de máxima verosimilitude son consistentes e seguen unha distribución asintóticamente normal (ver referencia [8]). Non obstante, cando moitos parámetros están involucrados na maximización (como é o caso, pois hai tantos parámetros como observacións do evento de interese), estas propiedades poden fallar e obtermos estimadores de máxima verosimilitude pouco eficientes ou inconsistentes. É por isto que desenvolveremos estas propiedades en detalle a continuación.

2.3.2. Consistencia do estimador de Kaplan-Meier

Como xa vimos, $\widehat{F}_n(t)$ é un estimador consistente de $F(t)$ como consecuencia do Teorema 2.1. O resultado que presentamos a continuación é un análogo do Teorema de Glivenko-Cantelli para o estimador $\widehat{F}_{KM}(t)$.

Teorema 2.6. *Sexa $(T_1, C_1), ..., (T_n, C_n)$ unha mostra de dúas variables aleatorias independentes T e C e sexan $F(t)$ e $G(t)$ as funcións de distribución poboacionais asociadas ás variables T e C respectivamente, tense con probabilidade 1 que*

$$\sup_{t\in[0,\sigma]} \left|\widehat{F}_{KM}(t) - F(t)\right| \longrightarrow 0,$$

onde $\sigma \in \mathcal{T} = \{t : F(t-) < 1, G(t-) < 1\}$.

A proba deste teorema require do uso de procesos estocásticos que exceden os contidos do Grao en Matemáticas e pode atoparse en [9]. Notamos como, a diferenza do que ocorría para o estimador $\widehat{F}_n$, o rango de valores nos que o estimador $\widehat{F}_{KM}$ é consistente está restrinxido.

Para comprender mellor este resultado, abandoemos por un momento a censura aleatoria que estivemos manexando ata agora e valoremos unha situación máis simple con censura pola dereita de Tipo I. Supoñamos que a variable T que queremos estudar segue unha distribución normal $\mathcal{N}(0, 1)$ e que a censura se produce, por exemplo, para todos aqueles datos que superen o valor $C = 0.67$. A porcentaxe de censura nesta situación é do 25 % e o valor do MISE (multiplicado por 1000) que obtivemos despois dun pequeno estudo de simulación por Montecarlo (con 1000 mostras de tamaño $n = 500$) no que se empregou o estimador de Kaplan-Meier foi de 251.21. Este resultado supera calquera outro valor de MISE obtido na anterior simulación para o mesmo tamaño de mostra: algo está fallando. Na Figura 2.6 represéntase a función de distribución da variable T estimada a través de $\widehat{F}_{KM}$ nunha das repeticións da simulación. Vemos como a partires da constante de censura C, o estimador deixa de ser consistente. Ao non dispor de datos máis aló de C resulta imposible poder estimar a distribución de T nesa zona. Esta idea é a esencia do Teorema 2.6 onde en lugar dunha censura fixa se traballa cunha censura aleatoria.

Volvendo aos exemplos nos que traballabamos cunha censura aleatoria, aínda que segundo o Teorema 2.6, o estimador sería consistente en todo $\mathbb{R}$ tendo en conta a distribución das variables T e C consideradas, na práctica pode non chegar a apreciarse este efecto. Isto é, en determinados casos sería preciso dispoñer de mostras realmente grandes para que o estimador $\widehat{F}_{KM}$ fose unha boa aproximación da función de distribución. Fixémonos por exemplo no modelo cun 30 % de censura que presentamos no primeiro estudo de simulación. Na situación que se reflexa na parte (c) da Figura 2.2, a probablidade de obter un valor de T fóra do intervalo $[-4, 4]$ é de $6.33 \cdot 10^{-5}$, a mesma que de obter un valor de C fóra do intervalo $[-1.4, 2.6]$. Supoñendo que para calquera mostra de tamaño razoable tódalas observacións se atopasen nestes intervalos, asumiríamos que $F(t-) = 4$ e $G(t-) = 2.6$ de maneira que a converxencia uniforme do estimador tería lugar unicamente no intervalo $(-\infty, 2.6)$.

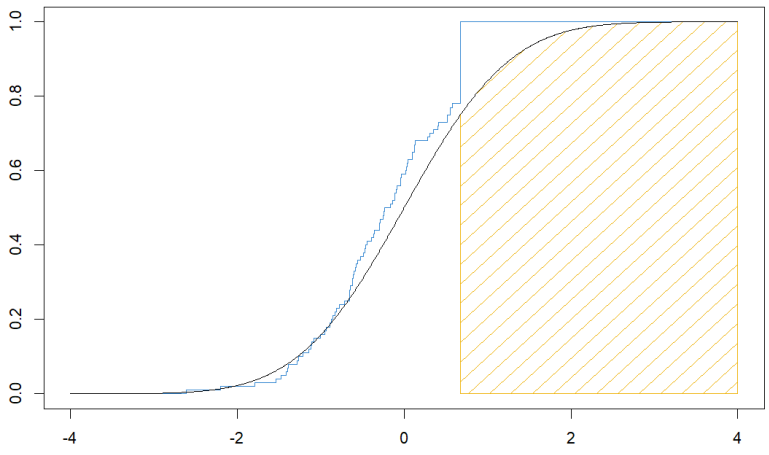


Figura 2.6: Representación do estimador de Kaplan-Meier asociado a unha mostra de tamaño $n = 500$ dunha distribución $T \in \mathcal{N}(0,1)$ con datos censurados. A zona amarela representa a área non observada como consecuencia da censura de Tipo I que sofre a variable T. Finalmente, a liña negra representa a distribución poboacional asociada á variable T.

2.3.3. Intervalos de confianza

Lembremos que un estimador é **asintoticamente normal** se para todo $\theta \in \Theta$, existe $K(\theta) > 0$ tal que

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \longrightarrow \mathcal{N}\left(0, K(\theta)\right).$$

Denotaremos $K(\theta)$ como a varianza asintótica do estimador empírico $\widehat{\theta}_n$. Para un t fixo, o estimador $\widehat{F}_n(t)$ verifica a propiedade de normalidade asintótica, pois como consecuencia do Teorema central do límite[9] tense que

$$\sqrt{n}\left(\widehat{F}_n(t) - F(t)\right) \longrightarrow \mathcal{N}\left(0, F(t)(1 - F(t))\right).$$

Baixo certas condicións de regularidade, o estimador de Kaplan-Meier verifica unha propiedade semellante, xa que para un t fixado $\widehat{F}_{KM}(t)$ (analogamente $\widehat{S}_{KM}(t)$) segue aproximadamente unha distribución normal. Para poder xustificar esta afirmación é preciso considerar o estimador $\widehat{S}_{KM}$ como un proceso estocástico en t e derivar diso que $\sqrt{n}\left(\widehat{S}_{KM}(t) - S(t)\right)$ tende a un proceso Gaussiano en t con media 0. Este tipo de consideracións afástanse do tema principal deste

[9] Sexan $X_1, \ldots, X_n$ variables independentes e identicamente distribuídas con $\mathbb{E}(X_i) = \mu$ e $\mathrm{Var}(X_i) = \sigma^2 < \infty$ e sexa $\overline{X}_n$ a media mostral, entón

$$\sqrt{n}\left(\overline{X}_n - \mu\right)) \longrightarrow \mathcal{N}\left(0, \sigma^2\right).$$

Traballo de Fin de Grao, porén pode atoparse unha explicación detallada en [9]. Para poder dar un intervalo de confianza, asumiremos simplemente que para un t fixo $\widehat{S}_{KM}(t)$ segue que unha distribución normal, isto é:

$$\sqrt{n}\left(\widehat{S}_{KM}(t) - S(t)\right) \longrightarrow \mathcal{N}\left(0, \text{Var}\left(\widehat{S}_{KM}(t)\right)\right).$$

Posto que descoñecemos o valor da varianza asintótica de $\widehat{S}_{KM}(t)$, tentaremos construír un estimador consistente de $\text{Var}\left(\widehat{S}_{KM}(t)\right)$ seguindo o procedemento descrito en [10].

Aplicando o **Método delta**[10] para a función $g(x) = log(x)$, verifícase que

$$\sqrt{n}\left(\log\left(\widehat{S}_{KM}(t)\right) - \log\left(S(t)\right)\right) \longrightarrow \mathcal{N}\left(0, \text{Var}\left(\widehat{S}_{KM}(t)\right)\frac{1}{S(t)^2}\right).$$

e polo tanto

$$\text{Var}\left(\widehat{S}_{KM}(t)\right) \approx \text{Var}\left(\log\left(\widehat{S}_{KM}(t)\right)\right)\widehat{S}_{KM}(t)^2. \tag{2.7}$$

Baixo a hipótese de que as variables $Z_1, \ldots, Z_n$ son independentes tense ademais que

$$\text{Var}\left(\log\left(\widehat{S}_{KM}(t)\right)\right) = \sum_{T_i<t} \text{Var}\left(\log\left(1-\widehat{h}_i\right)\right).$$

Polo tanto para construír un estimador para $\text{Var}\left(\widehat{S}_{KM}(t)\right)$ chegará con atopar un estimador para $\text{Var}\left(\log\left(1-\widehat{h}_i\right)\right)$.

Como xa mencionamos na Sección 2.3.1 o número de veces que ocorre o evento de interese nun intervalo $[T_i^*, T_{i+1}^*)$ coñecido n_i, segue unha distribución binomial $d_i \in \text{Bin}(n_i, h_i)$. Posto que $d_i = \widehat{h}_i \cdot n_i$,

$$\sqrt{n}\left(\widehat{h}_i - h_i\right) \longrightarrow \mathcal{N}\left(0, \frac{h_i(1-h_i)}{n_i}\right),$$

deducimos que

$$\text{Var}(\widehat{h}_i) = \frac{\text{Var}(d_i)}{n_i^2} = \frac{h_i(1-h_i)}{n_i},$$

e como consecuencia, podemos empregar o seguinte estimador da varianza:

$$\widehat{\text{Var}}(\widehat{h}_i) = \frac{\widehat{h}_i(1-\widehat{h}_i)}{n_i} = \frac{d_i(n_i-d_i)}{n_i^3}. \tag{2.8}$$

Así que aplicando de novo o Método Delta con $g(x) = \log(1-x)$, tense que

$$\sqrt{n}\left(\log(1-\widehat{h}_i) - \log(1-h_i)\right) \longrightarrow \mathcal{N}\left(0, \frac{\text{Var}(\widehat{h}_i)}{(1-h_i)^2}\right),$$

[10] Sexa $(Y_n)_{n\in\mathbb{N}}$ unha sucesión de variables asintóticamente normal tal que $\sqrt{n}\left(Y_n - \mu\right) \to \mathcal{N}\left(0, \sigma^2\right)$ e $g : \mathbb{R} \to \mathbb{R}$ sexa unha función derivable en μ, entón

$$\sqrt{n}\left(g\left(Y_n\right) - g(\mu)\right) \to \mathcal{N}\left(0, \sigma^2 g'(\mu)^2\right).$$

e polo tanto

$$\mathrm{Var}\left(\log\left(1-\widehat{h_i}\right)\right)=\frac{\mathrm{Var}(\widehat{h_i})}{(1-h_i)^2}.$$

Volvendo á expresión (2.7) e recapitulando as distintas aproximacións, podemos escribir

$$\begin{aligned}\mathrm{Var}\left(\widehat{S}_{KM}(t)\right) &\approx \widehat{S}_{KM}(t)^2\cdot\mathrm{Var}\left(\log\left(\widehat{S}_{KM}(t)\right)\right)\\ &= \widehat{S}_{KM}(t)^2\cdot\sum_{T_i<t}\mathrm{Var}\left(\log\left(1-\widehat{h_i}\right)\right)\\ &\approx \widehat{S}_{KM}(t)^2\cdot\sum_{T_i<t}\widehat{\mathrm{Var}}(\widehat{h_i})\frac{1}{(1-\widehat{h_i})^2}.\end{aligned}$$

Finalmente tendo en conta a expresión (2.8), a varianza de $\widehat{S}_{KM}$ pode ser estimada mediante

$$\widehat{\mathrm{Var}}\left[\widehat{S}_{KM}(t)\right]=\widehat{S}_{KM}(t)^2\sum_{T_i<t}\frac{d_i}{n_i(n_i-d_i)}. \tag{2.9}$$

Esta expresión coñécese como **fórmula de Greenwood** e foi proposta en 1926 (ver referencia [11]), antes incluso que de que Kaplan e Meier presentasen o estimador para a función de supervivencia. O epidemiólogo Major Greenwood chegou a esta expresión traballando con táboas de vida clásicas nas que se recollían datos de supervivencia e falecemento.

A construción dos intervalos de confianza do estimador de Kaplan-Meier deriva dos resultados que acabamos de presentar. No que segue, $(1-\alpha)$ referirase ao nivel de confianza do intervalo e introduciremos a seguinte notación:

$$\sigma_S^2(t)=\frac{\widehat{\mathrm{Var}}\left[\widehat{S}_{KM}(t)\right]}{\widehat{S}_{KM}(t)^2}.$$

O chamado **intervalo de confianza linear** é dos máis empregados e está implementado na maioría de paquetes estatísticos. Se $q_{1-\alpha/2}$ representa cuantil de orde $1-\alpha/2$ dunha distribución normal estándar, o intervalo de confianza linear vén dado por:

$$\left(\widehat{S}_{KM}(t)-q_{1-\alpha/2}\sigma_S(t)\,\widehat{S}_{KM}(t),\widehat{S}_{KM}(t)+q_{1-\alpha/2}\sigma_S(t)\,\widehat{S}_{KM}(t)\right)$$

Un intervalo deste tipo pode incluír valores imposibles que estean fóra do rango $[0,1]$. Aplicando certas transformacións sobre $\widehat{S}_{KM}(t)$ podemos evitar esta restrición; as que presentamos a continuación foron propostas por Borgan e Liestøl no ano 1990 (ver [12]):

- Transformación logarítmica
  Facendo unha doble transformación logarítmica do estimador $\widehat{v}(t)=\log\left(-\log(\widehat{S}_{KM}(t))\right)$, obtemos o intervalo

$$\left[\widehat{S}_{KM}(t)^{1/\theta},\widehat{S}_{KM}(t)^{\theta}\right],\ \text{onde}\ \theta=\exp\left\{\frac{q_{1-\alpha/2}\sigma_S(t)}{\ln\left(\widehat{S}_{KM}(t)\right)}\right\}.$$

A diferenza do caso anterior, este intervalo non é simétrico pero soluciona o problema do rango.

- Transformación arcseno-raíz cadrada
  Aplicando unha transfomación do tipo $\widehat{v}(t) = \operatorname{arcsen}\left(\widehat{S}_{KM}(t)^{1/2}\right)$, obtemos o intervalo

$$\operatorname{sen}^2\left\{\operatorname{máx}\left[0, \operatorname{arcsen}\left(\widehat{S}_{KM}(t)^{1/2}\right) - 0.5 q_{1-\alpha/2}\sigma_S(t)\left(\frac{\widehat{S}_{KM}(t)}{1-\widehat{S}_{KM}(t)}\right)^{1/2}\right]\right\}$$
$$\leq S(t) \leq$$
$$\operatorname{sen}^2\left\{\operatorname{mín}\left[\frac{\pi}{2}, \operatorname{arcsen}\left(\widehat{S}_{KM}(t)^{1/2}\right) + 0.5 q_{1-\alpha/2}\sigma_S(t)\left(\frac{\widehat{S}_{KM}(t)}{1-\widehat{S}_{KM}(t)}\right)^{1/2}\right]\right\}.$$

  Neste caso recupérase a simetría e ademais solvéntase a cuestión do rango.

Observación 2.7. En determinadas situacións, pode resultar de interese atopar unha **banda de confianza** na que a función de supervivencia estea contida para todo instante t dun intervalo dado e cun certo nivel de confianza. Nótese que este problema é distinto ao que presentabamos ao comezo da sección, onde nos interesabamos por atopar intervalos de confianza para un instante t fixo. Neste caso o obxectivo consiste en atopar dúas funcións aleatorias $L(t)$ e $U(t)$ de forma que $\mathbb{P}(L(t) \leq S(t) \leq U(t), \forall t \in [t_L, t_U]) = 1-\alpha$, onde $1-\alpha$ e o nivel de confianza. Non nos adentraremos máis nesta cuestión pois na práctica traballaremos unicamente con intervalos de confianza, para máis información ver [1].

Capítulo 3

O estimador de Beran

É natural imaxinar que poidan existir outras variables aleatorias que estean relacionadas (no senso de relación de dependencia) cun certo tempo de supervivencia T. Desta forma, cando estean presentes outras variables aleatorias ademais da variable censurada T na que nos centramos ata o momento, resultará de interese a estimación da función de supervivencia condicionada a estas novas covariables. Nesta sección prestaremos especial atención ao caso cunha única covariable continua e presentaremos un estimador da función de supervivencia condicional coñecido como estimador de Beran ou estimador de Kaplan-Meier condicional.

3.1. Función de distribución condicional

Dado un par de variables aleatorias (Y, X), referirémonos á variable Y como variable aleatoria resposta ou dependente e á variable X como variable explicativa ou independente. A continuación, definiremos a **función de distribución condicional**, que denotaremos por $F_{Y|X}$, tendo en conta se a variable explicativa X é discreta ou continua. A varaible resposta Y considerarase continua en calquera caso.

Se X é unha **variable discreta** e $x \in X$ un posible valor da variable X, entón a función de distribución de Y condicionada a que $X = x$ será

$$\begin{aligned} F_{Y|X} : \mathbb{R} &\longrightarrow [0,1] \\ y &\longmapsto F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x). \end{aligned}$$

Supoñendo que $\mathbb{P}(X = x) > 0$, por medio da fórmula de Bayes poderemos escribir

$$F_{Y|X}(y \mid x) = \frac{F_{Y,X}(y, x)}{\mathbb{P}(X = x)},$$

onde $F_{Y,X}$ é a función de distribución conxunta das variables X e Y.

No contexto da Análise de Supervivencia, a variable resposta será a variable T que representa o tempo de supervivencia. Consideraremos en primeiro lugar unha situación con datos completos, isto é, en ausencia de datos censurados. A estimación da función de distribución da variable T condicionada á variable X no caso de que esta última sexa discreta é sinxela, pois bastará con facer uso da función de distribución empírica para cada unha das categorías de X. Así, dada unha mostra $\{(T_1, X_1), \dots, (T_n, X_n)\} \in (T, X)$ de observacións independentes, a **función de distribución condicional empírica** será:

$$\widehat{F}_{T|X}(t \mid x) = \frac{\#\{i : T_i \leq t, X_i = x\}}{\#\{j : X_j = x\}} = \frac{1}{\sum_{j=1}^{n} \mathbb{I}(X_j = x)} \sum_{i=1}^{n} \mathbb{I}(T_i \leq t, X_i = x).$$

De existiren observacións censuradas, a dupla anterior pasará a ser unha terna (T, C, X) onde C é a variable censura, ou equivalentemente (Z, δ, X), onde Z indica o tempo de supervivencia observado e δ é a variable indicadora da censura. Así, considerando unha mostra $\{(Z_1, \delta_1, X_1), \dots, (Z_n, \delta_n, X_n)\} \in (Z, \delta, X)$ de observacións aleatorias, bastaría con restrinxir o estimador de Kaplan-Meier presentado no Capítulo 2 a cada unha das categorías da variable X para obter unha boa estimación de $\widehat{F}_{T|X}$. É dicir, teríamos que:

$$\widehat{F}^{KM}_{T|X}(t \mid x) = 1 - \prod_{Z_i \leq t} \left(1 - \frac{1}{\#\{j : Z_j \geq Z_i, X_j = x\}}\right)^{\delta_i}.$$

No caso de que X se trate dunha **variable continua** non será posible considerar a mesma expresión que antes, pois para calquera valor x tense que $\mathbb{P}(X = x) = 0$. Definiremos entón a función de distribución condicional en termos da función de densidade condicional como

$$\begin{aligned} F_{Y|X} : \mathbb{R} &\longrightarrow [0, 1] \\ y &\longmapsto F_{Y|X}(y \mid x) = \int_{-\infty}^{y} f_{Y|X}(u \mid x) du, \end{aligned}$$

onde a función de densidade $f_{Y|X}(y \mid x)$, sempre que $f_X(x) > 0$, se pode escribir como

$$f_{Y|X}(y \mid x) = \frac{f_{Y,X}(y, x)}{f_X(x)}.$$

Deducimos a seguinte expresión para a función de distribución condicional:

$$F_{Y|X}(y \mid x) = \frac{1}{f_X(x)} \int_{-\infty}^{y} f_{Y,X}(u, x) du.$$

Observemos que se Y e X son independentes, entón $F_{Y|X}(y \mid x) = F_Y(y)$ e estaremos na situación presentada no Capítulo 2. No caso contrario, atopar un estimador para a función de distribución condicional non resulta tan evidente como no caso discreto. Non obstante, a seguinte igualdade danos unha pista sobre o enfoque co que resolveremos este problema:

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x) = \mathbb{E}(\mathbb{I}(Y \leq y) \mid X = x),$$

onde para cada $y \in Y$, $\mathbb{I}(Y \leq y)$ é unha nova variable aleatoria. Desta forma, a función de distribución condicional poderá verse como unha función de regresión que depende dos valores da variable aleatoria X e onde a variable resposta é $\mathbb{I}(Y \leq y)$. Antes de continuar, recordemos a definición de función de regresión e algún dos seus estimadores:

Definición 3.1. Dadas unha variable aleatoria explicativa X e unha variable aleatoria resposta Y, o **modelo de regresión** que busca explicar Y en función de X será $Y = m(X) + \varepsilon$, onde

$$m(x) = \mathbb{E}(Y \mid X = x), \quad x \in X.$$

é a **función de regresión** e ε é unha variable de erro que verifica $\mathbb{E}(\varepsilon \mid X = x) = 0$, $\forall x \in X$.

Seguindo coa notación da Definición 3.1, podemos intentar estimar a función $m(X)$ partindo dunha mostra de observacións independentes $(Y_1, X_1), \ldots, (Y_n, X_n) \in (Y, X)$. Como primeira idea, poderíamos considerar unha partición $B_1, \ldots, B_r$ do rango da variable explicativa X e asociar a cada un dos intervalos da partición o valor medio da variable Y:

$$\widehat{m}_P(x) = \frac{\sum_{i=1}^n \mathbb{I}\left(X_i \in B_l\right) Y_i}{\sum_{i=1}^n \mathbb{I}\left(X_i \in B_l\right)}, \quad x \in B_l.$$

Este estimador permite construír o que se coñece como regresograma (unha adaptación do histograma ao contexto de regresión) e presenta certos inconvenientes como a falta de continuidade e unha forte dependencia da partición escollida. Para perfeccionar esta idea, en lugar dunha partición, consideraremos para cada un dos valores x nos que queremos estimar a función de regresión un intervalo da forma $(x - h, x + h)$, onde h denota un **parámetro ventá**. Ademais, para que sexan os $X_i \in (x - h, x + h)$ máis próximos a x os que teñan un maior peso sobre a súa estimación podemos recorrer a unha **función núcleo**, isto é, unha función $K : \mathbb{R} \to \mathbb{R}$ que verifica $K(u) = K(-u)$ e $\int_{-\infty}^{\infty} K(u)du = 1$. Entre as máis usadas están a función núcleo uniforme $K(u) = \frac{1}{2} \cdot \mathbb{I}\left(|u| < 1\right)$, a Gaussiana $K(u) = \exp\left(-u^2/2\right)/(\sqrt{2\pi})$ ou a de Epanechnikov $K(u) = \frac{3}{4}\left(1 - u^2\right) \cdot \mathbb{I}\left(|u| < 1\right)$. Así, tendo todo isto en conta podemos definir o estimador coñecido como **estimador de Nadaraya-Watson**

$$\widehat{m}_{NW}(x) = \sum_{i=1}^{n} W_{i,h}\left(x\right) Y_i = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)}.$$

Como no caso anterior, este estimador non paramétrico é en esencia unha media ponderada dos valores da variable resposta, onde o valor $W_{i,h}$ representa o peso de cada observación e depende do parámetro ventá. A escolla deste parámetro non é trivial: unha ventá moi pequena leva estimacións pouco suaves e con moito ruído, mentres que unha ventá demasiado grande pode chegar a ignorar as relacións entre as variables que intentamos determinar. Existen distintos métodos que permiten estimar un valor de ventá óptimo, entre os máis empregados está o de **validación cruzada**, que busca minimizar as diferenzas entre os valores observados da variable

resposta e os estimados tendo en conta a variable explicativa. É dicir, a ventá estimada mediante validación cruzada, que denotaremos por $\widehat{h}_{CV}$, será aquela que minimice a expresión

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}_{-i,h}\left(X_i\right)\right)^2,$$

onde $\widehat{m}_{-i,h}$ é o estimador da función de regresión xerado sen ter en conta o par (Y_i, X_i) e que se coñece habitualmente como estimador *"leave-one-out"*. Poden verse máis detalles sobre a estimación deste tipo de modelos de regresión en [13].

Retomando o problema da estimación a función de distribución condicional e trasladando a idea do estimador de Nadaraya-Watson empregado en modelos de regresión, podemos definir o seguinte estimador para a función de distribución condicional:

$$\widehat{F}_{Y|X}(y \mid x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\mathbb{I}(Y_i \leq y)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

Chamaremos a este estimador **función de distribución condicional empírica**, pois comparte a mesma esencia que a función de distribución empírica, coa diferenza de que as contribucións de cada valor da mostra neste caso virán influenciadas pola proximidade da covariable X ao valor x para o que queremos estimar $F_{Y|X}$. Computacionalmente, o cálculo deste novo estimador é máis costoso, xa que para cada valor y debemos estimar unha función de regresión $\widehat{F}_{Y|X}$ e devolver a avaliación en $X = x$. Ademais, a estimación da ventá óptima tamén se volve especialmente costosa. Esta estimación da ventá óptima refina a idea da validación cruzada que viamos no caso do estimador de Nadaraya-Watson e lévase a cabo en dous pasos. En primeiro lugar, para cada un dos X_i trataremos de estimar $F_{Y|X}(y \mid X = X_i)$. Para estimar o parámetro de suavizado correspondente debemos minimizar a función

$$\mathrm{CV}(h) = \sum_{j=1}^{n}\left(\mathbb{I}(Y_i \leq Y_j) - \widehat{F}_{Y|X}^{-i,h}\left(Y_j \mid X_i\right)\right)^2,$$

onde $\widehat{F}_{Y|X}^{-i,h}$ é a función de distribución de Y condicionada a $X = X_i$ estimada sen o dato i-ésimo. Obtense así unha ventá local óptima $\widehat{h}_i$ e tendo en conta tódolos elementos da mostra terase:

$$\widehat{\mathbf{h}} = (\widehat{h}_1, \ldots, \widehat{h}_i, \ldots, \widehat{h}_n),$$

O segundo paso consistirá en tomar como ventá óptima global $\widehat{h}_{CV}$ á maior das ventás óptimas locais estimadas, isto é

$$\widehat{h}_{CV} = \max\left(\widehat{\mathbf{h}}\right) = \max\left(\widehat{h}_1, \ldots, \widehat{h}_i, \ldots, \widehat{h}_n\right).$$

Da mesma forma que no Capítulo 2 viamos que a función de distribución empírica non era un bo estimador da función de distribución poboacional en presenza de censura, podemos intuír

que o estimador aquí presentado vai ter un comportamento semellante. É por isto que na Sección 3.2 presentaremos un estimador da función de distribución condicional adaptado a un escenario con censura: o estimador de Beran. Ademais, na Sección 3.3 compararemos ambos estimadores a través dun estudo de simulación.

3.2. O estimador de Beran

Nunha publicación do ano 1980 (ver [14]), Rudolf Beran propuxo un estimador para a función de distribución condicional baixo censura para un caso xeral con varias covariables e asumindo unicamente a independencia das variables T e C. Na expresión que aquí presentamos teremos en conta unicamente unha covariable continua X. O estimador de Beran comparte semellanzas tanto co estimador da función de distribución empírica como co estimador de Kaplan-Meier e é por isto que tamén se coñece polo nome de **estimador de Kaplan-Meier condicional**.

Partindo dunha mostra $(T_1, C_1, X_1), \dots, (T_n, C_n, X_n)$ de observacións independentes de distribución unha tripla (T, C, X), onde T e C son variables aleatorias positivas asociadas a tempos de supervivencia e X unha variable aleatoria continua. Baixo a condición de que as variables T e C sexan independentes, poderemos identificar univocamente a distribución (T, C, X) cunha distribución (Z, δ, X) (en concordancia coa notación xa introducida). Desta maneira, a mostra anterior observada na práctica será da forma

$$(Z_1, \delta_1, X_1), \dots, (Z_n, \delta_n, X_n).$$

Nestas condicións, o estimador da función de distribución da variable T condicionada a X, ao que nos referiremos como **estimador de Beran**, será:

$$\widehat{F}_B(t \mid x) = 1 - \widehat{S}_B(t \mid x) = 1 - \prod_{i=1}^{n} \left(1 - \frac{\mathbb{I}\left(Z_i \leq t, \delta_i = 1\right) W_{i,h}(x)}{1 - \sum_{j=1}^{n} \mathbb{I}\left(Z_j < Z_i\right) W_{j,h}(x)} \right),$$

onde $\widehat{S}_B(t \mid x)$ representa o estimador da función de supervivencia condicional e $W_{i,h}$ os pesos de Beran, que dependen dun parámetro ventá h e da función núcleo escollida

$$W_{i,h}(x) = \frac{K\left(\left(x - X_i\right)/h\right)}{\sum_{j=1}^{n} K\left(\left(x - X_j\right)/h\right)}.$$

A estimación da ventá óptima faise de maneira análoga á vista para a función de distribución condicional empírica, coa única diferenza de que neste caso a función a minimizar será

$$\mathrm{CV}(h) = \sum_{j=1}^{n} \Delta_{i,j} \left(\mathbb{I}(T_i \leq T_j) - \widehat{F}_B^{-i,h}\left(T_j \mid X_i\right) \right)^2,$$

onde $\Delta_{i,j}$ indica a chamada "utilidade" de cada par (T_i, T_j) e vale 1 se o par é considerado útil ou 0 no caso contrario. Dirase que un par (T_i, T_j) é útil se verifica algunha das seguintes condicións: $(\delta_i, \delta_j) = (1,1)$, $(\delta_i, \delta_j) = (1,0)$ e $T_i \leq T_j$, $(\delta_i, \delta_j) = (0,1)$ e $T_i \geq T_j$ ou $i = j$.

Desta forma, retrínxese en gran medida a influencia das observacións censuradas na estimación da ventá óptima sen obviarse completamente, permitindo o bo funcionamento do estimador incluso para porcentaxes de censura elevadas.

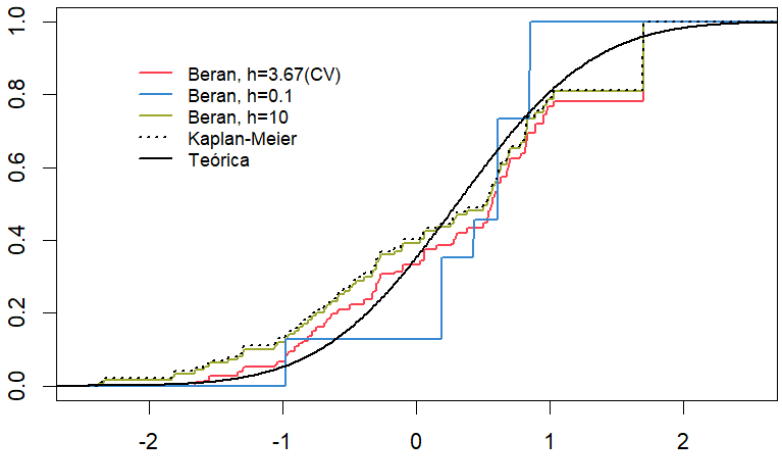


Figura 3.1: Funcións de distribución condicionais $F_{T|X}(t, X = 1)$ estimadas mediante o estimador de Beran empregando distintos valores de ventá ($h = 3.67$ en trazo vermello, $h = 0.1$ en trazo azul e $h = 10$ en trazo verde) e a partires dunha mostra de tamaño $n = 100$ dunha distribución normal bivariante (T, X) onde $T \in \mathcal{N}(0,1)$ e $X \in \mathcal{N}(0, 2^2)$ cun coeficiente de correlación de $\rho_{T,X} = 0.6$. A porcentaxe de censura é do 20 %, sendo a variable de censura $C \in \mathcal{N}(0.9, 0.5^2)$. Represéntanse tamén o estimador de Kaplan-Meier (con liña negra punteada) e a función de distribución condicional poboacional (en trazo negro).

Se nos alonxamos da ventá óptima e tomamos valores de h moi grandes ou moi pequenos, ocorre o que se pode ver na Figura 3.1, onde se representan as funcións de distribución condicionais obtidas co estimador de Beran considerando distintos valores de ventá para as variables $T \in \mathcal{N}(0,1)$, $C \in \mathcal{N}(0.9, 0.5^2)$ e $X \in \mathcal{N}(0, 2^2)$. Nótese que neste caso a porcentaxe de datos censurados é do 20 % e que trataremos de estimar a distribución de T condicionada a $X = 1$. A curva en trazo vermello é a asociada ao valor de ventá estimado co método de validación cruzada e tamén a que parece minimizar as distancias coa función de distribución condicional teórica. Xusto o contrario ocorre para a curva azul que representa o estimador de Beran cunha ventá

moi pequena. Neste segundo caso, ao ser tan estritos, o número de observacións que se teñen en conta para a estimación é moi baixo e como consecuencia os escalóns da función de distribución estimada vólvense demasiado grandes. En contraposición, ao tomar un valor de ventá moi grande (en trazo verde) acaban por terse en conta practicamente tódolos valores da mostra e o resultado é unha curva moi semellante á que obteríamos co estimador de Kaplan-Meier (en trazo punteado).

De maneira análoga, un valor de ventá moi grande fará que a función de distribución condicional empírica tome exactamente a mesma forma que a función de distribución empírica $\widehat{F}_n$. En conclusión, este feito simplemente serve de apoio na visión dos estimadores condicionais non paramétricos que presentamos como unha restrición dos estimadores non condicionais en función duns determinados valores da covariable.

3.3. Comportamento da función de distribución condicional empírica e do estimador de Beran

Presentamos un novo estudo de simulación que nos permite comparar a función de distribución condicional empírica co estimador de Beran. Consideraremos tres variables aleatorias (T, C, X) tales que o vector aleatorio[1] (T, X) contendo a variable dependente T suxeita a estudo e a covariable independente X segue unha distribución normal bivariante[2],

$$(T, X) \in \mathcal{N}(\mu, \mathbf{\Sigma}), \quad \text{onde} \quad \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{e} \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 1.2 \\ 1.2 & 4 \end{pmatrix},$$

con coeficiente de correlación $\rho_{T,X} = 0.6$. Por outra parte, a distribución da variable de censura $C \in \mathcal{N}(m, 0.5^2)$ varía en función da porcentaxe de censura escollida, tomarase $m \in \{1.4, 0.9, 0.2\}$ para porcentaxes de censura do 10 %, 20 % e 40 %, respectivamente. O feito de que (T, X) siga unha distribución normal bivariante facilita a determinación distribución de T condicionada a $X = x$, pois terase que $T \mid (X = x) \in \mathcal{N}\left(\mu_T + \frac{\sigma_T}{\sigma_X}\rho_{T,X}\left(x - \mu_X\right), \left(1 - \rho_{T,X}^2\right)\sigma_T^2\right)$.

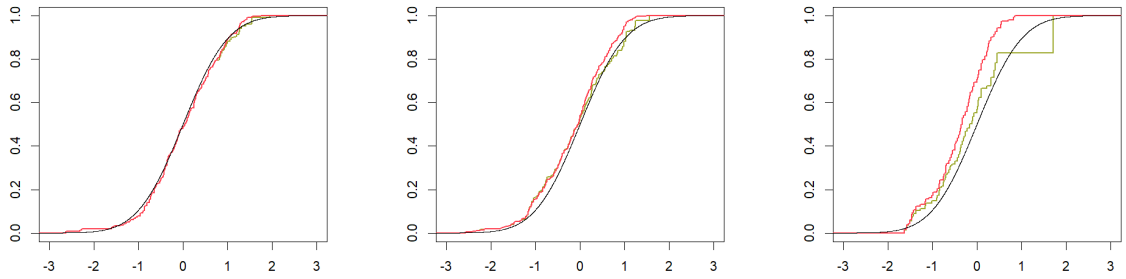
[1] Un **vector aleatorio** $(X_1, \ldots, X_n)$ é un vector formado polas variables aleatorias $X_1, \ldots, X_n$.

[2] Unha **distribución normal bivariante** (Y, X) de vector de medias e matriz de covarianza

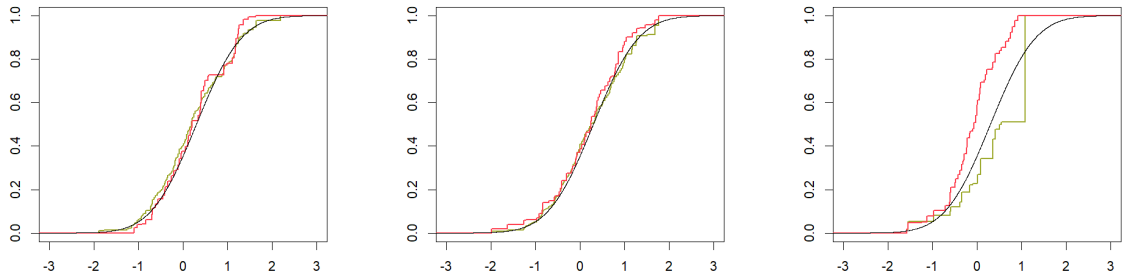
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} \quad \mathbf{e} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_Y^2 & \rho_{Y,X}\sigma_Y\sigma_X \\ \rho_{Y,X}\sigma_Y\sigma_X & \sigma_X^2 \end{pmatrix},$$

onde $\rho_{Y,X}$ é o coeficiente de correlación que toma valores entre -1 e 1 e indica a correlación lineal entre as variables, ten como función de densidade

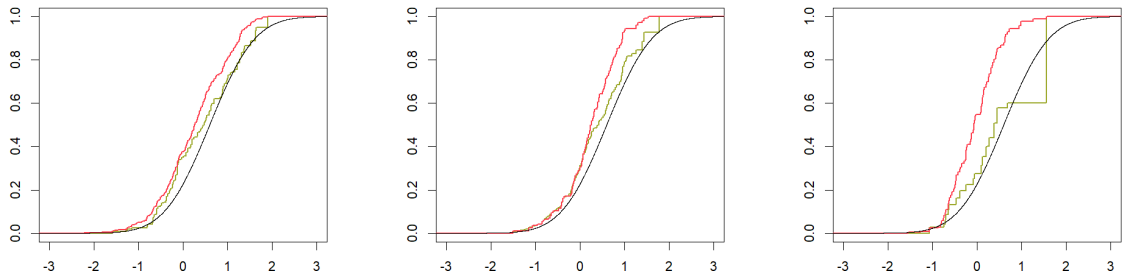
$$f(x, y) = \frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1 - \rho_{Y,X}^2}} \exp\left(-\frac{1}{2\left[1 - \rho_{Y,X}^2\right]}\left[\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2 - 2\rho_{Y,X}\left(\frac{y - \mu_Y}{\sigma_Y}\right)\left(\frac{x - \mu_X}{\sigma_X}\right) + \left(\frac{x - \mu_X}{\sigma_X}\right)^2\right]\right).$$



(a) $x = 0$, 10 % de censura. (b) $x = 0$, 20 % de censura. (c) $x = 0$, 40 % de censura.



(d) $x = 1$, 10 % de censura. (e) $x = 1$, 20 % de censura. (f) $x = 1$, 40 % de censura.



(g) $x = 2$, 10 % de censura. (h) $x = 2$, 20 % censura. (i) $x = 2$, 40 % de censura.

Figura 3.2: Representacións da función de distribución condicional empírica (en vemello), o estimador de Beran (en verde) e a función de distribución condicional teórica (en negro). As estimacións foron xeradas a partires dunha mostra de tamaño $n = 500$ do vector aleatorio (T, C, X), con $T \in \mathcal{N}(0, 1)$, $X \in \mathcal{N}(0, 2^2)$ e $C \in \mathcal{N}(m, 0.5^2)$ con m variando en función da porcentaxe de censura e sendo $\rho_{T,X} = 0.6$ o coeficiente de correlación entre as variables T e X.

Levamos a cabo un estudo de simulación por Montecarlo de 1000 repeticións coas variables anteriores tendo en conta tamaños de mostra de $n = 25, 200$ e 500. En cada repetición, está-

mase o erro cadrático medio integrado entre a función de distribución condicional empírica ou o estimador de Beran e a función de distribución condicional teórica. Téñense tamén en conta distintas porcentaxes de censura, incluída a do 0 % para abordar o caso de datos completos. Para a computación de ambos estimadores empregouse a libraría `npcure` (ver máis detalles en [3]) do software estatístico R, permitindo a estimación por validación cruzada do parámetro de ventá. Ademais, realizamos estes cálculos para tres posibles valores da covariable X: $x = 0, 1$ e 2. Os valores do MISE obtidos en cada caso poden verse na Táboa 3.1. Porén, antes de comentar estes resultados e para comprender mellor o que está a acontecer no estudo de simulación, presentamos na Figura 3.2 unha representación gráfica dos estimadores $\widehat{F}_{T|X}$ (en vermello) e $\widehat{F}_B$ (en verde) xunto coa función de distribución condicional poboacional $F_{T|X}$ (en negro).

En primeiro lugar, debemos destacar que en tódolos casos, a estimación feita co estimador de Beran se axusta mellor á función de distribución condicional teórica que a feita coa versión empírica. As diferenzas entre os dous estimadores son máis notables canto maior é a porcentaxe de censura ou maior é o valor da variable explicativa no que se avalía $F_{T|X}$. Lembremos que variable X ten como función de distribución marxinal unha distribución normal centrada en 0 de varianza 4, de maneira que a medida que nos alonxamos do 0 obtemos peores aproximacións debido a unha diminución dos datos considerados relevantes polos estimadores e ao efecto de x na distribución condicional. Ademais, a este efecto súmase o feito de que a censura é pola dereita, provocando un incremento na porcentaxe de observacións censuradas con peso na estimación a medida que aumenta o valor de x. Este fenómeno pode observarse con claridade nas mostras cun 40 % de censura, onde o estimador de Beran toma dous grandes escalóns na metade superior por mor da alta concentración de observacións censuradas. Se nos fixamos nas porcentaxes de censura do 10 % e do 20 %, podemos ver tamén como para $x = 0$ ou $x = 1$ ambos estimadores se aproximan moi ben á función de distribución condicional, mentres para $x = 2$ a función de distribución empírica non é capaz de xestionar o crecente volume de datos censurados.

Para confirmar se o comportamento descrito na Figura 3.2 é debido ao azar da mostra considerada, ou se en efecto é representativo dos dous estimadores, comentamos os resultados da Táboa 3.1. Se recordamos, nesta táboa inclúense os valores do MISE (multiplicados por 1000) obtidos no estudo de simulación por Montecarlo. Á vista dos resultados, podemos derivar en primeiro lugar o mal comportamento da función de distribución condicional empírica $\widehat{F}_{T|X}$ posto que observamos como a medida que aumenta a porcentaxe de censura, o erro cadrático medio integrado aumenta incluso ao aumentar o tamaño de mostra, e independientemente do punto x considerado. Estes resultados van na liña dos presentados no Capítulo 2 e reforzan a necesidade de empregar procedementos específicos para o tratamento de datos censurados.

Por outra banda, os resultados presentados na Táboa 3.1 sobre o estimador de Beran son coherentes dende un punto de vista teórico. En primeiro lugar, o erro cadrático medio integrado

diminúe ao aumentar o tamaño de mostra para calquera dos niveis de censura ou valores da covariable considerados. Por exemplo, se nos fixamos nas porcentaxes de censura máis elevadas, do 20 % e do 40 %, vemos como as diferenzas entre os dous estimadores son cada vez máis notables, sobre todo ao considerar $x = 2$. Así, para unha mostra de tamaño $n = 500$ e unha porcentaxe de censura do 40 % o valor do MISE é sete veces menor para o estimador de Beran, pasando de 186.71 a 26.89. É dicir, o MISE obtido mediante a función de distribución condicional empírica duplica os valores obtidos co estimador de Beran para os tamaños de mostra máis grandes.

	Variable explicativa	Tamaño da mostra	Porcentaxe de censura			
			0 %	10 %	20 %	40 %
$\widehat{F}_{T\|X}$	$x = 0$	25	58.70	51.10	49.84	76.11
		200	10.01	8.65	11.1	47.68
		500	4.71	4.62	7.25	42.9
	$x = 1$	25	69.35	60.14	66.16	138.32
		200	11.20	11.67	21.93	100.84
		500	5.65	6.94	16.64	95.72
	$x = 2$	25	78.89	86.15	107.44	243.67
		200	19.28	25.57	51.10	196.73
		500	9.13	15.64	42.41	186.71
$\widehat{F}_B$	$x = 0$	25	58.70	54.65	50.93	62.93
		200	10.01	9.49	10.01	16.83
		500	4.71	4.55	5.22	8.80
	$x = 1$	25	69.35	59.32	59.79	89.02
		200	11.20	11.00	12.77	23.17
		500	5.65	6.08	6.96	13.53
	$x = 2$	25	78.89	80.14	85.94	130.63
		200	19.28	21.04	25.72	45.58
		500	9.13	10.38	13.94	26.89

Táboa 3.1: Valores do MISE (multiplicado por 1000) para distintos valores da variable X como resultado de comparar a función de distribución condicional empírica e o estimador de Beran coa función de distribución condicional teórica. Valores obtidos para mostras de tamaño $n \in \{25, 200, 500\}$ do vector aleatorio (T, C, X), con $T \in \mathcal{N}(0, 1)$, $X \in \mathcal{N}(0, 2^2)$ e $C \in \mathcal{N}(m, 0.5^2)$ con m variando en función da porcentaxe de censura e sendo $\rho_{T,X} = 0.6$ o coeficiente de correlación entre as variables T e X.

A modo de conclusión, o pequeno estudo de simulación presentado ilustra o bo comportamento do estimador de Beran baixo datos censurados en contraposición ao estimador clásico empírico que non amosa bos resultados no caso de que non poidamos observar completamente os valores da variable de interese.

Capítulo 4

Análise sobre unha base de datos reais

Neste capítulo, empregaremos as ferramentas estatísticas vistas nos Capítulos 2 e 3 para analizar un conxunto de datos reais. A base de datos da que provén a mostra coa que traballaremos é de libre acceso e pode atoparse na páxina web https://seer.cancer.gov/statfacts/html/lungb.html. Na base de datos considerada inclúense distintos valores recollidos durante o seguemento dun grupo de máis de medio millón de doentes con cancro de pulmón nos Estados Unidos realizado entre os anos 2000 e 2015.

4.1. Presentación e análise descritiva

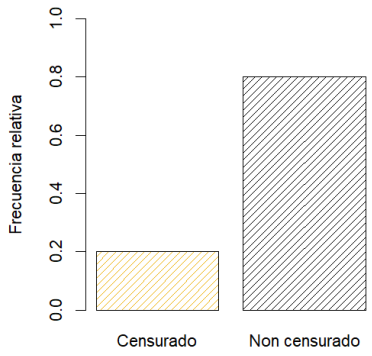
A mostra coa que imos traballar é unha submostra de tamaño $n = 1000$ obtida de forma aleatoria a partires da base de datos orixinal. Na nosa análise, teremos en conta unicamente as seguintes variables:

- `time`: tempo de supervivencia observado, en meses. Representa o tempo transcorrido entre a diagnose do cancro e momento de falecemento ou perda do seguemento da/o doente.
- `status`: variable indicadora de censura, vale 0 se o dato está censurado e 1 no caso contrario.
- `sex`: sexo da/o doente, `Muller` ou `Home`.
- `loc`: estadío no que se atopa o cancro no momento da diagnose. As categorías consideradas son `Localizado` se o tumor está limitado ao lugar no que comezou, `Rexional` se afecta a tecidos ou órganos próximos, `Distante` se acada partes do corpo alonxadas da orixe e `Descoñecido` no caso de que non se dispoña desta información[1].

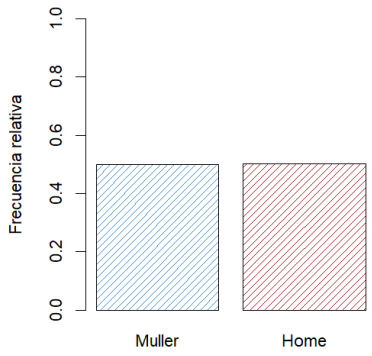
[1]Para máis información pode consultarse a seguinte ligazón: https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

- `age`: idade da/o doente, en anos.
- `size`: tamaño do tumor na escala CS[2]. Nesta escala, un valor entre 001 e 989 representa o tamaño en milímetros do tumor, empregándose a cifra 989 no caso de que mida 989 mm ou máis.

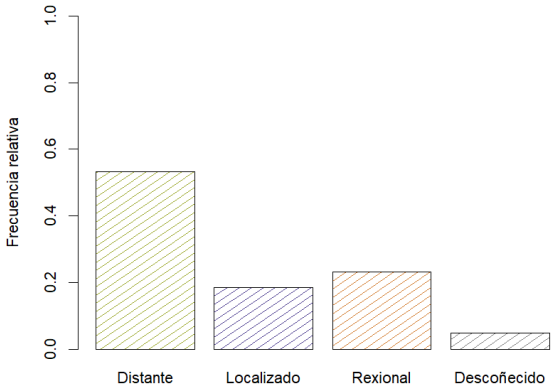
A continuación, presentamos unha análise da Estatística descritiva cuxo obxectivo é ilustrar a forma das variables que imos empregar. Nas Figuras 4.1 e 4.2 preséntanse gráficos de barras e histogramas asociados ás diferentes variables incluídas na base de datos que vimos de presentar.



(a) Gráfico de barras asociado á variable `status`.



(b) Gráfico de barras asociado á variable `sex`.

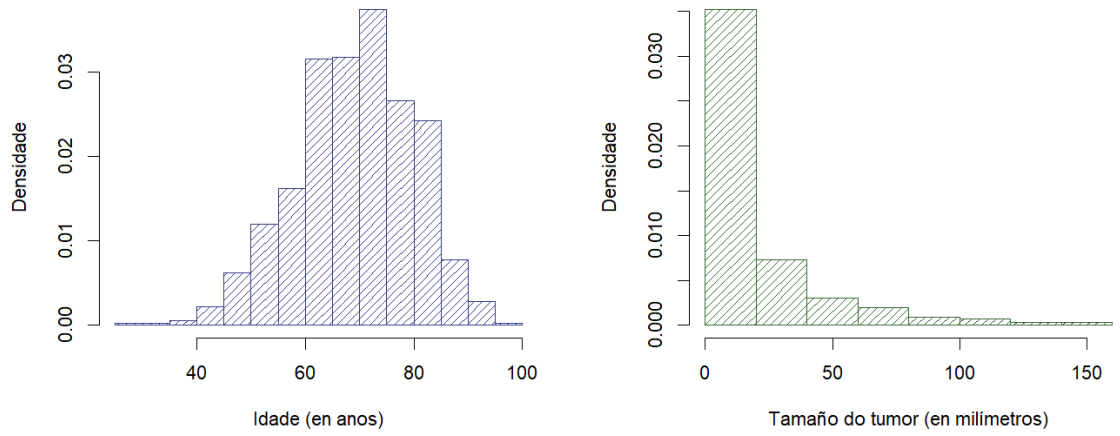


(c) Gráfico de barras asociado á variable `loc`.

Figura 4.1: Gráficos de barras e ás distintas variables cualitativas consideradas no estudo: `status`, `sex` e `loc`.

[2]Para máis información pode consultarse a seguinte ligazón: https://training.seer.cancer.gov/collaborative/system/tnm/t/size/.

A distribución das variables categóricas `status`, `sex` e `loc` ilústrase mediante un gráfico de barras. En primeiro lugar, represéntase na parte (a) da Figura 4.1 a proporción de datos censurados da mostra. É importante coñecer esta proporción posto que canto máis elevada sexa, máis necesario será empregar ferramentas adaptadas ao contexto de datos censurados. Neste caso, a porcentaxe de censura na mostra (é dicir, a porcentaxe de individuos para os que non se rexistrou unha data de falecemento) é exactamente do 20 %, semellante ás porcentaxes coas que xa traballamos nos estudos de simulación realizados ao longo do traballo. En segundo lugar, na parte (b) da Figura 4.1 representamos o gráfico de barras asociado á variable `sex`. Observamos que a proporción de mulleres e homes participantes no estudo é moi equilibrada; en concreto, participan 499 mulleres e 501 homes. En último lugar, na parte (c) da Figura 4.1 represéntanse as proporcións de tumores clasificados como distantes, localizados, rexionais ou de estado descoñecido no momento de diagnose. Como vemos, ao comezo do seguemento máis da metade dos tumores se atopaban espallados por distintas partes do corpo, un cuarto afectaba xa a tecidos próximos e algo menos dun 25 % se atopaban localizados nunha área concreta do pulmón. O número de tumores para os que non se ten información sobre a súa distribución no organismo representan unha minoría, con 49 observacións.



(a) Histograma asociado á variable `age`.

(b) Histograma asociado á variable `size`.

Figura 4.2: Histogramas asociados ás distintas variables continuas consideradas no estudo: `age` e `size`.

Para a análise das variables continuas `time`, `age` e `size` empregamos un histograma unicamente para a representación das dúas últimas variables. En presenza de censura, un histograma da variable `time` representaría a variable que ata agora denotamos como $Z = \text{mín}(T, C)$ e non o tempo de supervivencia T que queremos estudar. Máis adiante, presentamos outro tipo de

gráficos máis adecuados para a visualización da súa distribución. Continuando coa descrición das outras dúas variables continuas, centrarémonos en primeiro lugar no histograma asociado á variable `age` presentado na parte (a) da Figura 4.2. As idades mínima e máxima das/os doentes participantes son de 26 e 96 anos, respectivamente. Debemos destacar que case dous terzos das/os participantes teñen ente 60 e 80 anos, sendo a idade media de case 69.6 anos. Por outra banda, no relativo á variable `size` representante do tamaño do tumor (ver parte (b) da Figura 4.2), vemos como a gran maioría de tumores miden menos de 50 mm. En concreto, o 75 % dos tumores miden menos de 25 mm e un 12.2 % do total mide un só milímetro. O valor medio é de 20.07 mm, mentres que o máximo rexistrado é de 156 mm.

Posto que será de utilidade máis adiante no estudo función de distribución condicional da variable `time`, presentamos na Figura 4.3 un gráfico de caixas que permite ver a distribución do tamaño do tumor en función de cada un dos estadíos iniciais. Como podemos observar, a gran maioría de tumores distantes ou de estado descoñecido miden apenas uns milímetros; en efecto, o 75 % dos tumores distantes miden menos de 12 mm, sendo o tamaño medio de 10 mm e a mediana de 5 mm. Simultaneamente, o 75 % dos tumores localizados miden máis de 12 mm, cun tamaño medio de 49 mm; mentres que os tumores rexionais parecen situarse entre os localizados e os distantes en termos de tamaño, sendo o valor medio de 28 mm. Polo tanto, aínda que en tódolos casos (agás no de localización descoñecida) o tamaño dos tumores se distribúe nun amplo rango de valores, podemos intuír como unha maior dispersión do tumor no organismo vai ligada un menor tamaño do mesmo. Así mesmo, as conclusións que se derivan da Figura 4.3 poñen de manifesto unha certa relación de dependencia entre as variables `loc` e `size`.

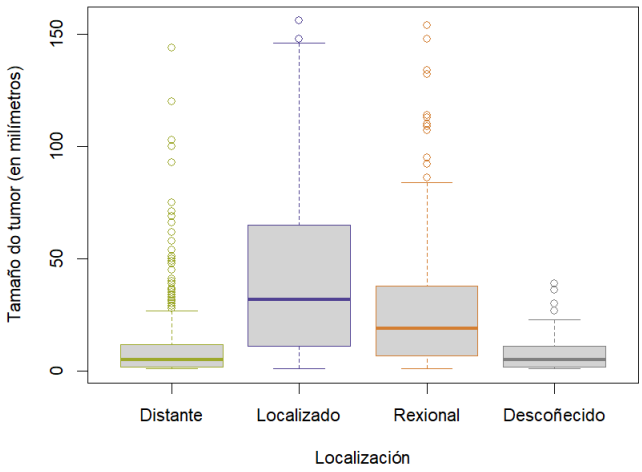


Figura 4.3: Gráfico de caixas asociado á variable `size` en función da variable `loc`.

Retomando o comentario sobre a representación das observacións da variable `time`, na Figura 4.4 ilustramos os tempos de supervivencia observados distinguindo aquelas observacións que están censuradas en cor amarela. Como se pode ver, a censura é de tipo aleatorio e tódalas observacións son inferiores a 156 meses (que posiblemente coincida coa duración do estudo). O 83.7 % dos tempos de supervivencia rexistrados están comprendidos nos primeiros 4 anos despois do comezo do seguemento; e destes, un 14 % son observacións censuradas, unha porcentaxe inferior á da mostra total, que era do 20 %. A partires dos primeiros 4 anos, o 41.7 % das observacións están censuradas; na Figura 4.4 vemos como os puntos amarelos están máis dispersos que os puntos negros, que están moi concentrados na parte inferior. Este cambio na porcentaxe de censura é coherente coa censura aleatoria pola dereita á que nos enfrontamos.

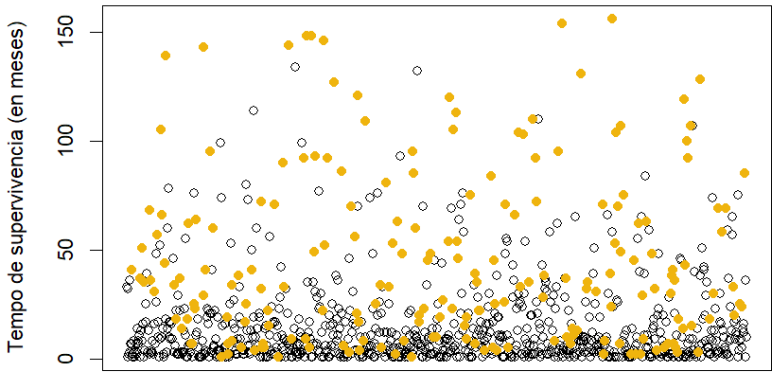


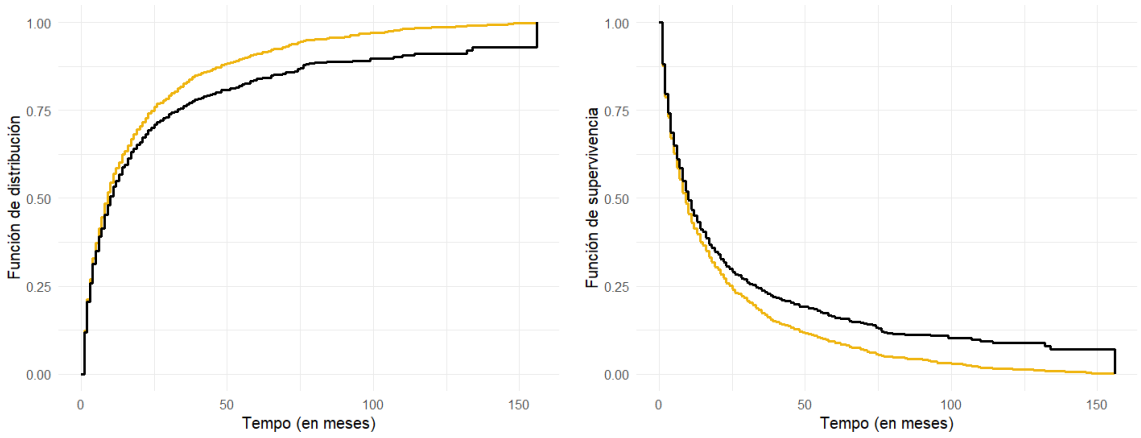
Figura 4.4: Representación do tempo de supervivencia observado (representado pola variable `time`). En amarelo destácanse aquelas observacións censuradas (indicadas grazas á variable `status`).

4.2. Estimando a función de supervivencia

Ao longo desta sección, densenvolveremos as técnicas estatísticas estudadas para a estimación da función de distribución da variable representante do tempo transcorrido entre a diagnose do cancro e o falecemento da/o doente.

Nesta sección, centrarémonos unicamente no par de variables `time` e `status` que representan o tempo de supervivencia observado e a posible censura. Na Figura 4.5 preséntanse a función

de distribución (parte (a)) e función de supervivencia (parte (b)) obtidas empregando tanto o estimador de Kaplan-Meier (en negro) como a función de distribución empírica (en amarelo). Para a interpretación de ambas funcións, tomaremos a modo de exemplo a perspectiva da segunda gráfica, correspondente á función de supervivencia estimada. Ademais, aínda que só a curva xerada co estimador de Kaplan-Meier (en negro) permite extraer conclusións razoables, representamos tamén a función de distribución empírica (en amarelo) coa idea de ver o efecto que tería ignorar a existencia de censura na mostra. Vemos que en efecto, unha interpretación da función de supervivencia empírica levaría a conclusións moito máis pesimistas e erradas ao considerar certas perdas no seguemento como decesos. Tomando por exemplo a probabilidade de supervivencia aos 100 meses despois da diagnose da enfermidade, a probabilidade estimada pola función $\widehat{S}_n$ sería do 3 % mentres que a estimada por $\widehat{S}_{KM}$ sería do 10 %.



(a) Función de distribución estimada.

(b) Función de supervivencia estimada.

Figura 4.5: Representación das funcións de distribución e supervivencia estimadas. En trazo negro represéntase o estimador de Kaplan-Meier, mentres que en trazo amarelo se representa a función de distribución empírica.

Continuando cos comentarios en relación á Figura 4.5, centrarémonos unicamente na estimación obtida co estimador de Kaplan-Meier posto que, en presenza de censura, proporciona unha mellor aproximación da función de supervivencia poboacional (como xa xustificamos grazas aos estudos de simulación presentados no Capítulo 2). Como podemos ver, a probabilidade de supervivencia diminúe drasticamente nos dous anos inmediatamente posteriores á diagnose do cancro, sintuándose nun 30 % aproximadamente. Non obstante, a función parece ralentizarse a partires deste momento, sendo máis claro este fenómeno a partires dos 75 meses, xa que entre os 75 e os 150 meses a función apenas varía nun 5 %. Lembremos que a maioría de datos están concentrados nos primeiros 25 meses e que a censura é maior nos últimos anos, de maneira que a función de supervivencia toma escalóns máis grandes a medida que se aproxima a cero.

Co fin de extraer máis información sobre a función de supervivencia que acabamos de estimar, podemos facer uso do comando `survfit` da libraría `survival` (ver referencia [4]). Mostramos a continuación o código de R empregado, así como os resultados obtidos:

```
> fit <- survfit(Surv(time, status) ~ 1, data=lungs, conf.type="plain")
> summary(fit, times = c(1,10,20,30,40,50,75,100,150))
Call: survfit(formula = Surv(time, status) ~ 1, data = lungs)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
1      1000     119    0.8810  0.0102        0.8609        0.901
10      484     377    0.4935  0.0161        0.4620        0.525
20      305     140    0.3398  0.0155        0.3095        0.370
30      216      67    0.2601  0.0146        0.2315        0.289
40      151      32    0.2170  0.0140        0.1895        0.244
50      118      17    0.1910  0.0137        0.1641        0.218
75       57      33    0.1288  0.0129        0.1034        0.154
100      30      10    0.1022  0.0128        0.0771        0.127
150       2       5    0.0701  0.0155        0.0397        0.100
```

Neste resumo, aparecen baixo a categoría `Survival` as probabilidades de supervivencia estimadas polo estimador de Kaplan-Meier en cada un dos instantes que se recollen na columna `time`. Nas columnas `n.risk` e `n.event` indícase o número de doentes en risco de falecer e o número de falecementos que tiveron lugar en cada un destes instantes temporais. Tomando como exemplo o primeiro mes de estudo, vemos que o número de participantes en risco coincide co tamaño da mostra e que o número de decesos foi de 119. Finalmente, na columna `std.err` aparece a desviación típica estimada mediante a fórmula de Greenwood que vimos na Sección 2.3.3 e en `lower 95% IC` e `upper 95% IC` os intervalos de confianza lineares derivados directamente de `std.err`. Como xa vimos, existen expresións alternativas que refinan os intervalos de confianza lineares, como son a transformación log-log ou a transformación arcsen-raíz cadrada. Na Táboa 4.1 amosamos os valores da probabilidade de supervivencia estimada co estimador de Kaplan-Meier para $t = 10, 50$ e 150, xunto cos distintos intervalos de confianza detallados na Sección 2.3.3 considerando un nivel de confianza do 95 %.

	$\widehat{S}_{KM}$	Linear	Log-log	Arcseno
$t = 10$	0.493	$(0.462, 0.525)$	$(0.461, 0.525)$	$(0.462, 0.525)$
$t = 50$	0.191	$(0.164, 0.218)$	$(0.164, 0.219)$	$(0.164, 0.219)$
$t = 150$	0.087	$(0.061, 0.114)$	$(0.063, 0.116)$	$(0.062, 0.116)$

Táboa 4.1: Estimacións asociadas á función de supervivencia obtidas empregando o estimador de Kaplan-Meier xunto cos respectivos intervalos confianza obtidos para un nivel de confianza do 95 %.

Como vemos, as diferenzas son practicamente imperceptibles, con variacións de apenas unha ou dúas milésimas. Nestes exemplos, non chegamos a apreciar as correccións de rango e simetría das transformacións log-log ou arcsen-raíz debido ao amplo tamaño de mostra que estamos considerando. Destacamos unicamente como os intervalos obtidos mediante a última das transformacións son sempre máis grandes que os intervalos de confianza lineares. Ademais, nótese que os intervalos de confianza obtidos son intervalos que conteñen á probabilidade de supervivencia poboacional cunha probabilidade do 95 %.

4.3. Efecto de variables categóricas sobre a estimación da función de supervivencia

Teñamos en conta agora as variables `sex` e `loc` referidas ao sexo da/o doente e á localización do tumor nas/os participantes do estudo. Neste novo escenario, o noso obxectivo será presentar unha estimación da función de supervivencia da variable representante do tempo transcorrido entre a diagnose da enfermidade e o falecemento da/o doente para cada unha das categorías das variables cualitativas a ter en conta. Na Figura 4.6 e na Figura 4.7 aparecen representadas tales funcións.

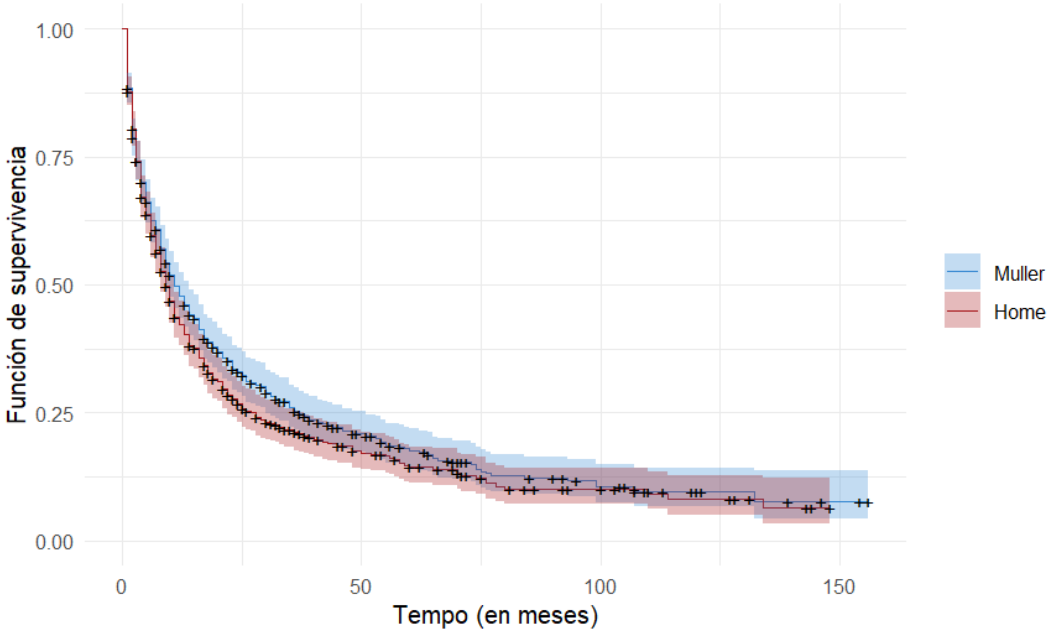


Figura 4.6: Función de supervivencia asociada á variable `time` estimada en función da variable `sex` acompañada de intervalos de confianza de nivel 95 %.

Á vista da Figura 4.6 non podemos concluír que existan diferenzas substanciais nas proba-

bilidades de supervivencia entre homes e mulleres. A curva de supervivencia para os homes está case sempre lixeiramente por debaixo da correspondente ás mulleres e a distancia máxima entre as curvas semella producirse arredor dos 25 meses da diagnose da enfermidade. Non obstante, vemos como as partes sombreadas (que representan os correspondentes intervalos de confianza) se solapan en todo momento, impedindo concluír a existencia dalgún tipo de influencia da variable `sex` sobre o tempo de supervivencia despois da diagnose do cancro de pulmón.

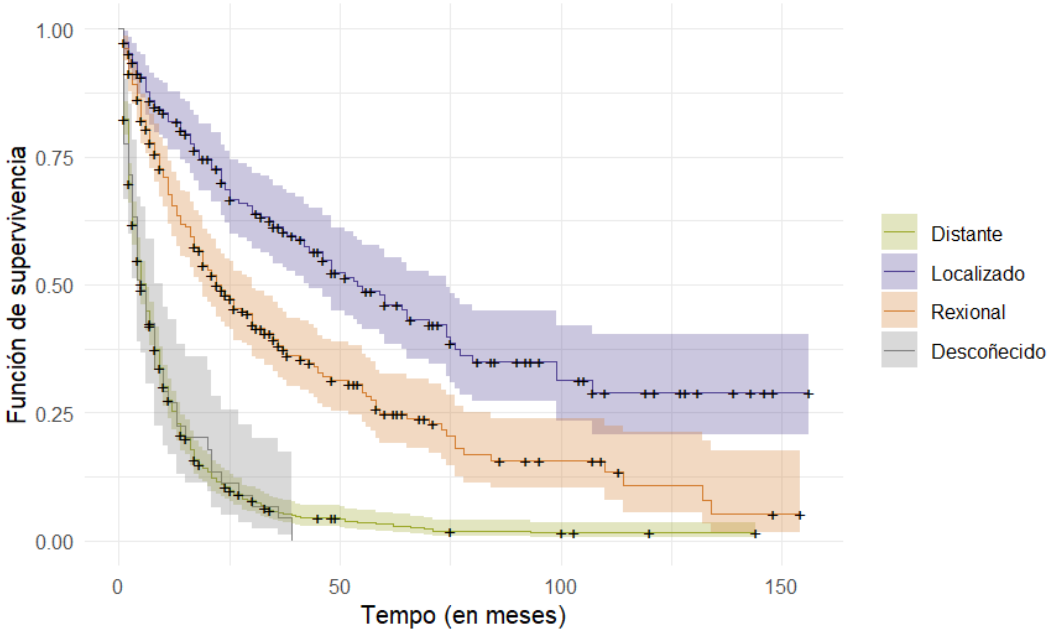


Figura 4.7: Función de supervivencia asociada á variable `time` estimada en función da variable `loc`, xunto cos correspondentes intervalos de confianza de nivel 95 %.

Pola contra, na Figura 4.7 observamos un claro efecto da variable `loc` sobre a variable `time`. A probabilidade de supervivencia vese realmente influída pola localización do tumor. Se analizamos, por exemplo, a probabilidade de supervivencia estimada aos 4 anos da diagnose, será do 52 % no caso de que o tumor estea localizado, do 31 % no caso de que sexa de tipo rexional e do 4 % no caso de que sexa distante. É polo tanto clara a diminución da probabilidade de supervivencia en función da dispersión do tumor no organismo, tal e como era de esperar. No caso dos tumores para os que non se tiña coñecemento da súa localización, parece que á vista da Figura 4.7 se trataba de tumores distantes ou dispersos, xa que as dúas curvas de supervivencia se solapan constantemente. Porén, posto que o número deste tipo de observacións na mostra é moi baixo, abstémonos de levar a cabo conclusións rotundas.

4.4. Efecto das variables continuas sobre a estimación da función de supervivencia

Procederemos a continuación a analizar a función de supervivencia da variable que representa tempo transcorrido entre a diagnose do cancro e o falecemento da/o doente na súa versión condicional e tendo en conta as variables continuas `age` e `size` referidas á idade das/os doentes e ao tamaño do tumor, facendo uso das técnicas vistas ao longo do Capítulo 3. Nas Figuras 4.8 e 4.9 presentamos as curvas de supervivencia condicionais ás distintas idades e aos distintos tamaños do tumor, respectivamente. Ademais, aparecen representadas tanto a función de supervivencia empírica condicional (en trazo punteado) como o estimador de Beran (en trazo liso) coa idea de ver o erro que implicaría o feito de ignorar a existencia de datos censurados na mostra.

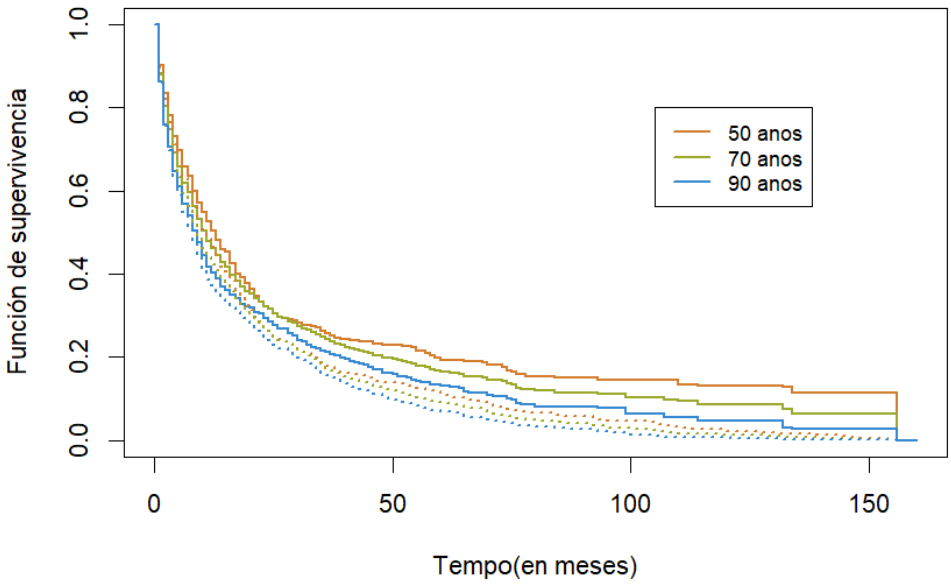


Figura 4.8: Representación da función de supervivencia condicionada a valores de idade de 50, 70 e 90 anos. En trazo liso represéntase o estimador de Beran e en trazo punteado represéntase a función de distribución condicional empírica.

Pode verse na Figura 4.8 a estimación da variable `time` condicionada a tres posibles valores da variable `age`: 50, 60 e 70 anos. Á vista da Figura 4.8 vemos como función de distribución condicional empírica (en trazo punteado) proporciona aproximacións máis pesimistas ao obviar o efecto da censura, estimando probabilidades de supervivencia que son superadas polo estimador de Beran ata en 10 puntos porcentuais. Nótese que, de considerar idades máis extremas, as

estimacións serían pouco interpretables por mor do baixo volume de datos. Á vista da Figura 4.8, as probabilidades de supervivencia obtidas empregando o estimador de Beran semellan diminuír canto maior é a idade das/os doentes no momento da diagnose. Non obstante, as curvas están demasiado próximas entre si como para poder extraer conclusións claras e asumiremos que a idade é un factor pouco influínte no tempo de supervivencia.

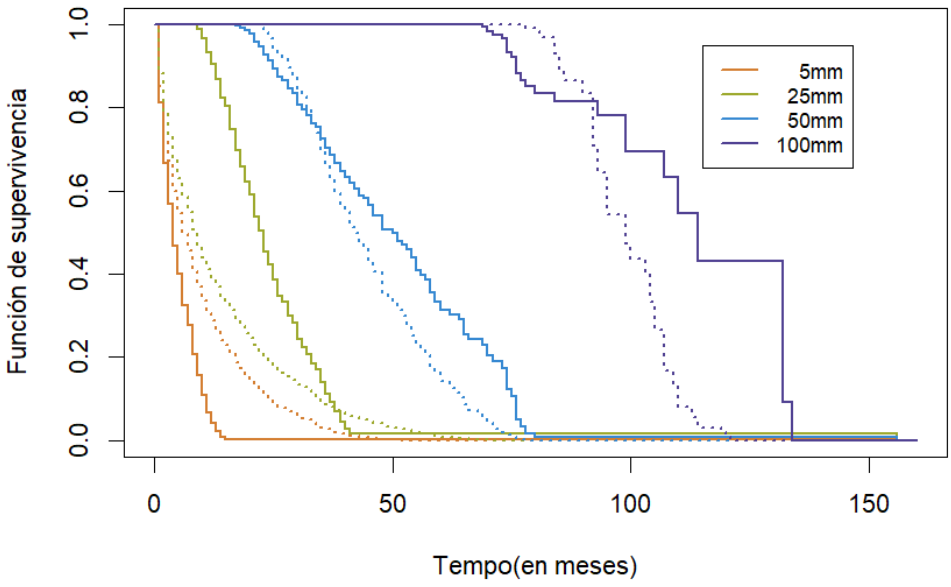


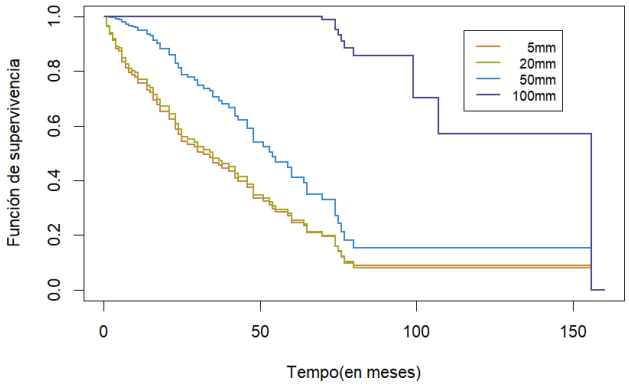
Figura 4.9: Representación da función de supervivencia condicionada a valores de tamaño de tumor de 5, 20, 50 e 100 mm. En trazo liso represéntase o estimador de Beran e en trazo punteado represéntase a función de distribución condicional empírica.

Por outra parte na Figura 4.9, pode apreciarse que a forma das funcións de supervivencia estimadas varía substancialmente en función do tamaño do tumor. Os tamaños considerados son de 5, 20, 50 e 100 mm, resultando claro o aumento das probabilidades de supervivencia canto maior é o tamaño do tumor. Así, tense que a probabilidade de supervivencia estimada aos 50 meses despois da diagnose da enfermidade é do 100 % no caso de que o tamaño do tumor sexa de 100 mm, do 50 % no caso de que sexa de de 50 mm e practicamente nulas para os tamaños de 20 mm e 5 mm. Se nos fixamos agora no comportamento da función de distribución condicional empírica con respecto ao estimador de Beran, vemos que de novo proporciona estimacións máis pesimistas agás no caso de que o tamaño do tumor sexa de 5 mm. Para rematar, debemos lembrar que o 75 % dos tumores miden menos de 25 mm, polo que as estimacións condicionadas a tamaños de 50 e 100 mm son menos significativas que as de 5 e 20 mm.

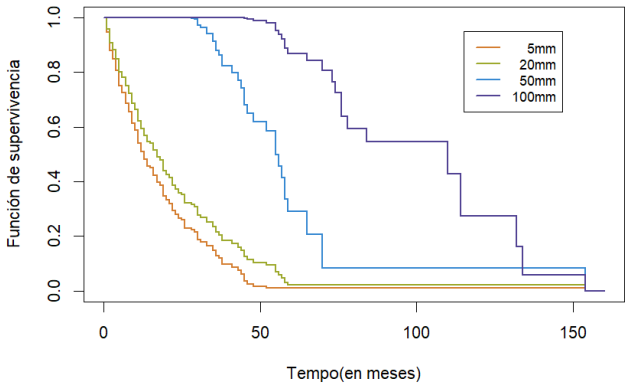
Ata o momento, a única variable que parece ter un efecto notable sobre o tempo de supervivencia, a parte da correspondente ao tamaño do tumor que acabamos de mencionar, é a variable asociada á localización do mesmo. Xa vimos durante a análise descritiva e na Figura 4.3 como se pode intuír unha certa relación entre ambas. No caso da variable relacionada coa localización do tumor os resultados non foron demasiado sorprendentes, pois é de esperar que canto máis espallada no corpo se atope a enfermidade maior sexa o risco de deceso. A hipótese que plantexamos neste momento é que os tumores máis grandes están menos dispersos e que por iso levan asociada unha maior probabilidade de supervivencia. Para tratar de avaliar dita hipótese, na Figura 4.10 representamos a estimación da función de supervivencia condicionada a distintos tamaños do tumor e tamén a distintas localizacións dos mesmos. Nótese que neste novo escenario só imos presentar os resultados asociados ao estimador de Beran.

O primeiro que destacamos da Figura 4.10 é a forma que toman as estimacións das funcións de supervivencia para tamaños de tumor de 100 mm. A supervivencia parece ser moi elevada pero as funcións son moi escalonadas; no caso da parte (c) da Figura 4.10 obtemos unha función constante igual a 1 que se vai a cero pouco antes dos 150 meses. Como xa sabiamos, o número de observacións para as que o tamaño do tumor é tan grande é moi pequeno e ao ter en conta a variable categórica `loc` a cantidade de datos da que se dispón para construír o estimador diminúe aínda máis, resultando en estimacións pouco ou nada significativas. Algo semellante, aínda que menos acusado, ocorre para o tamaño de 50 mm, sendo a curva da parte (a) da Figura 4.10 a máis suave. Lembremos que a mediana do tamaño dos tumores localizados está en 32 mm mentres que o tamaño medio é de 41 mm, resultando así nunha mellor estimación para o tamaño de 50 mm ao restrinxirnos a este tipo de tumor. Nótese que poderíamos mellorar estas estimacións aumentando o tamaño de mostra que estamos a considerar.

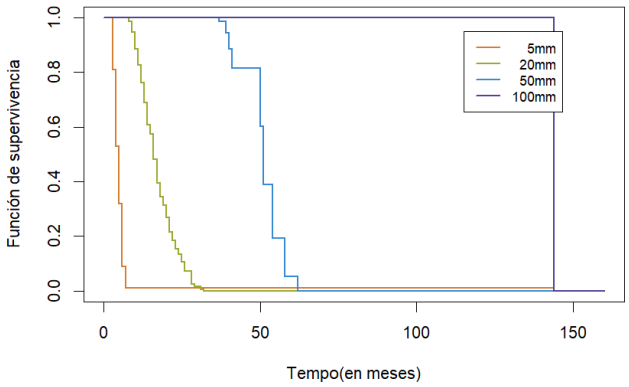
Prestando atención aos tamaños de 5 e 20 mm, que son os asociados ás estimacións máis significativas debido ao elevado número de observacións condicionadas a ditos valores, vemos claras diferenzas en función do estadío do tumor. Canto máis disperso está o cancro, menores son as probabilidades de supervivencia. No caso do tumor distante, representado na parte (c) da Figura 4.1, sorprende a forma que toma o estimador da función de supervivencia para o tamaño de 5 mm, onde as probabilidades de supervivencia decaen rapidamente despois da diagnose da enfermidade. Recordemos ademais, que a mediana do tamaño tumoral para os cancros de tipo distante (que representa algo máis da metade dos tumores) é de precisamente 5 mm polo que a estimación pode considerarse significativa. Por outra banda, na parte (a) da Figura 4.1 ocorre tamén un fenómeno interesante, pois as curvas estimadas aos tamaños de 5 e 20 mm coinciden case totalmente. Isto débese a que o estimador está a usar pesos moi semellantes en ámbalas dúas estimacións como consecuencia do maior tamaño dos tumores localizados.



(a) Tumor localizado.



(b) Tumor rexional.



(c) Tumor distante.

Figura 4.10: Representacións das funcións de supervivencia condicionadas a valores de tamaño de tumor de 5, 20, 50 e 100 mm e ás distintas localizacións do tumor (localizado, rexional ou distante). En trazo liso represéntase o estimador de Beran.

Capítulo 5

Conclusións

Este Traballo de Fin de Grao céntrase no estudo dunha variable aleatoria non negativa T que representa o tempo ata a aparición dun certo evento de interese. Esta variable, á que nos referimos como tempo de supervivencia, pode verse afectada por un fenómeno coñecido como censura. Isto é, ao extraer unha mostra dos tempos de supervivencia observados, algúns datos estarán incompletos e só coñeceremos unha información parcial en lugar dun valor concreto. No Capítulo 1, explicamos en detalle este efecto e presentamos distintos tipos de censura posibles, dando especial importancia á censura coa que traballamos ao longo de todo o documento: a censura aleatoria pola dereita.

Nas primeiras páxinas do Capítulo 2, analizamos os problemas que orixina a censura na estimación da función de distribución da variable T ao empregar a función de distribución empírica, que é un estimador non paramétrico clásico no contexto de datos completos (non censurados). De seguido, presentamos un estimador que adapta o anterior a unha situación con datos censurados e que se coñece como estimador de Kaplan-Meier. Este estimador modifica os pesos outorgados a cada observación na construción da estimación da función de distribución e permanece constante frente a unha observación censurada. A través de pequenos estudos de simulación por Montecarlo e botando man dunha medida global de erro coñecida como MISE, comparamos os dous estimadores mencionados en escenarios con distintas porcentaxes de censura. Os resultados destes estudos de simulación poñen de manifesto as vantaxes do estimador de Kaplan-Meier como estimador da función de distribución en presenza de datos censurados. Na segunda metade do Capítulo 2, centrámonos nalgunhas das propiedades máis interesantes do estimador de Kaplan-Meier e propoñemos unha estimación do mesmo como máximo da función de verosimilitude. Presentamos ademais un resultado de consistencia e obtemos un estimador para a varianza do mesmo que permite a construción dos intervalos de confianza.

No Capítulo 3 temos en conta unha posible variable aleatoria que denotamos por X e que

presentamos como unha variable explicativa do comportamento da variable T. Interesámonos entón pola estimación da función de distribución de T condicionada a unha variable X continua e presentamos a función de distribución condicional empírica como estimador clásico no contexto de datos completos. Este estimador non paramétrico é a extensión natural do estimador de Nadaraya-Watson, empregado no contexto de modelos de regresión, á estimación dunha función de distribución condicional. Como era de agardar vendo o comportamento da función de distribución empírica, este estimador non ten un bo comportamento en presenza de datos censurados. É por isto que densenvolvemos un novo estimador adecuado para a análise de datos censurados e baseado nas mesmas ideas que o estimador de Kaplan-Meier: o estimador de Beran. Da mesma forma que fixemos no Capítulo 2, comparamos o comportamento do estimador de Beran e da función de distribución condicional empírica mediante un novo estudo de simulación, quedando patente o mellor comportamento na práctica do estimador de Beran.

Ao longo do Capítulo 4, aplicamos as ferramentas anteriores sobre un conxunto de datos reais provenientes dun grupo de doentes con cancro de pulmón cunha porcentaxe de censura do 20 %. Neste caso, a variable T que queremos estudar representa o tempo transcorrido entre a diagnose do cancro e o falecemento da/o doente. Empregamos o estimador de Kaplan-Meier na estimación da función de distribución da variable T, tanto considerando todos os datos da mostra, como en función do sexo da/o doente e o estadío do tumor no momento da diagnose. As outras variables explicativas tidas en conta son a idade da/o doente e o tamaño do tumor e posto que se trata de variables continuas, empregamos o estimador de Beran para determinar a función de distribución condicional da variable T.

A modo de resumo, as diferentes análises levadas a cabo ao longo deste traballo poñen de manifesto que é necesario deseñar procedementos de estimación específicos cando estamos a traballar cunha base de datos censurados. Para rematar, na referencia [15] pode atoparse unha interesante revisión de diversas técnicas estatísticas clásicas que se poden adaptar ao contexto de datos censurados. Por exemplo, a consideración de modelos de Cox, que son modelos de regresión onde a variable resposta está censurada pola dereita, sería a continuación natural da metodoloxía presentada neste Traballo de Fin de Grao.

Anexo A

Código asociado aos diferentes estudos de simulación

Ao longo deste anexo preséntase o código de R empregado para levar a cabo os diferentes estudos de simulación presentados ao longo deste Traballo de Fin de Grao para ilustrar os distintos procedementos considerados.

```
#######################################################################
# 0. Determinación das porcentaxes de censura
#######################################################################

# Cálculo da porcentaxe de censura modificando a distribución da variable de
# censura.

n=1000000
for (m in c(1.8,1.1,0.6)){
  Te=rnorm(n)                  #distribución teórica
  C=rnorm(n,mean=m,sd=0.5)     #distribución de censura
  Z=pmin(Te,C)                 #tempo de supervivencia observado
  delta=1*(Z==Te)              #variable indicadora de censura
  print((n-sum(delta))/n)      #porcentaxe de censura
}

#######################################################################
# 1. Estudo 1: función distribución empírica vs. datos censurados
#######################################################################

# Estudo de simulación por Montecarlo para a comparación entre a función de
# distribución empírica (f.d.e.) e a función de distribución teórica na presenza
# de datos censurados.
# Inclúese unha simulación control sen censura.
```

```
# Devólvese o valor do MISE para cada un dos tamaños de mostra e das porcentaxes
# de censura.
# Aproximación da integral con Simpson.

set.seed(12345)
rep=1000                          #número de repeticións da simulación
t = seq(-4,4,length=10000)        #instantes nos que avaliamos f.d.e.

print("MISE para porcentaxe de censura do 0%")
for (n in c(10,150,500)){         #distintos tamaños de mostra
  print(paste("n=",n))
  ise=c()
  for (i in 1:rep){
    Te = rnorm(n)
    Femp = ecdf(Te)               #función de distribución empírica
    ise = append(sintegral(t,(Femp(t)-pnorm(t))^2)$int,ise)
  }
  print(mean(ise))                #MISE
}

print("MISE para porcentaxes de censura do 5%, 15% e 30%")
for (n in c(10,150,500)){         #distintos tamaños de mostra
  print(paste("n=",n))
  vmise = c()
  for (m in c(1.8,1.1,0.6)){      #distintas porcentaxes de censura
    ise=c()
    for (i in 1:rep){
      Te = rnorm(n)
      C = rnorm(n,mean=m,sd=0.5)
      Z = pmin(Te,C)
      Femp = ecdf(Z)              #función de distribución empírica
      ise = append(sintegral(t,(Femp(t)-pnorm(t))^2)$int,ise)
    }
    mise = mean(ise)
    vmise = append(vmise,mise)  #vector cos valores do MISE para as
    #distintas porcentaxes de censura
  }
  print(vmise)
}

#########################################################################
# 2. Estudo 2: estimador Kaplan-Meier vs. datos censurados
#########################################################################

# Estudo de simulación por Montecarlo para a comparación entre a o estimador de
# Kaplan-Meier (KM) e a función de distribución teórica na presenza de datos
# censurados.
```

```
# Inclúese unha simulación control sen censura.
# Devólvese o valor do MISE para cada un dos tamaños de mostra e das
# porcentaxes de censura.
# Aproximación da integral con Simpson.
# A función empregada para a estimación KM considera 0=non censurado e
# 1=censurado e devolve unha función de supervivencia.

set.seed(12345)
rep=1000                          #número de repeticións da simulación
t = seq(-4,4,length=10000)        #instantes nos que avaliamos f.d.e.

print("MISE para porcentaxe de censura do 0%")
for (n in c(10,150,500)){         #distintos tamaños de mostra
  print(paste("n=",n))
  ise=c()
  for (i in 1:rep){
    Te = rnorm(n)
    km = 1-KaplanMeier(t,data=Te,censored=1-rep(1,n))$surv
    ise = append(sintegral(t,(km-pnorm(t))^2)$int,ise)
  }
  print(mean(ise))                #MISE
}

print("MISE para porcentaxes de censura do 5%, 15% e 30%")
for (n in c(10,150,500)){             #distintos tamaños de mostra
  print(paste("n=",n))
  vmise = c()
  for (m in c(1.8,1.1,0.6)){          #distintas porcentaxes de censura
    ise=c()
    for (i in 1:rep){
      Te = rnorm(n)
      C = rnorm(n,mean=m,sd=0.5)
      Z = pmin(Te,C)
      delta = 1*(Z==Te)

      aux=sort(Z,index.return=T)
      Z.ord=aux$x                     #datos observados ordeados de menor a maior
      delta.ord=delta[aux$ix]         #variable indicadora de censura ordeada
      delta.ord[n]=1                  #última observación non censurada

      km = 1-KaplanMeier(t,data=Z.ord,censored=1-delta.ord)$surv
      ise = append(sintegral(t,(km-pnorm(t))^2)$int,ise)
    }
    mise = mean(ise)
    vmise = append(vmise,mise)        #vector cos valores do MISE para as
    #distintas porcentaxes de censura
  }
```

```
  print(vmise)
}

#########################################################################
# 3. Estudo 3: f.d.e. & KM vs. datos censurados con outras distribucións
#########################################################################

# 3.1 Cálculo das porcentaxes de censura
n=1000000
for (m in c(3,2,1.3)){
  Te=rexp(n)                    #distribución teórica
  C=rnorm(n,mean=m,sd=0.5)      #distribución de censura
  Z=pmin(Te,C)                  #tempo de supervivencia observado
  delta=1*(Z==Te)               #variable indicadora de censura
  print((n-sum(delta))/n)       #porcentaxe de censura
}

# 3.2 Simulación: función distribución empírica vs. datos censurados

set.seed(12345)
rep=1000                        #número de repeticións da simulación
t = seq(-4,4,length=10000)      #instantes nos que avaliamos f.d.e.

print("MISE para porcentaxe de censura do 0%")
for (n in c(10,150,500)){       #distintos tamaños de mostra
  print(paste("n=",n))
  ise=c()
  for (i in 1:rep){
    Te = rexp(n)
    Femp = ecdf(Te)             #función de distribución empírica
    ise = append(sintegral(t,(Femp(t)-pexp(t))^2)$int,ise)
  }
  print(mean(ise))              #MISE
}

print("MISE para porcentaxes de censura do 5%, 15% e 30%")
for (n in c(10,150,500)){           #distintos tamaños de mostra
  print(paste("n=",n))
  vmise = c()
  for (m in c(3,2,1.3)){            #distintas porcentaxes de censura
    ise=c()
    for (i in 1:rep){
      Te = rexp(n)
      C = rnorm(n,mean=m,sd=0.5)
      Z = pmin(Te,C)
      Femp = ecdf(Z)                #función de distribución empírica
      ise = append(sintegral(t,(Femp(t)-pexp(t))^2)$int,ise)
```

```
    }
    mise = mean(ise)
    vmise = append(vmise,mise)      #vector cos valores do MISE para as
    #distintas porcentaxes de censura
  }
  print(vmise)
}

# 3.2 Simulación: estimador Kaplan-Meier vs. datos censurados

set.seed(12345)
rep=1000                          #número de repeticións da simulación
t = seq(-4,4,length=10000)        #instantes nos que avaliamos f.d.e.

print("MISE para porcentaxe de censura do 0%")
for (n in c(10,150,500)){         #distintos tamaños de mostra
  print(paste("n=",n))
  ise=c()
  for (i in 1:rep){
    Te = rexp(n)
    km = 1-KaplanMeier(t,data=Te,censored=1-rep(1,n))$surv
    ise = append(sintegral(t,(km-pexp(t))^2)$int,ise)
  }
  print(mean(ise))
}

print("MISE para porcentaxes de censura do 5%, 15% e 30%")
for (n in c(10,150,500)){         #distintos tamaños de mostra
  print(paste("n=",n))
  vmise = c()
  for (m in c(1.8,1.1,0.6)){      #distintas porcentaxes de censura
    ise=c()
    for (i in 1:rep){
      Te = rexp(n)
      C = rnorm(n,mean=m,sd=0.5)
      Z = pmin(Te,C)
      delta = 1*(Z==Te)

      aux=sort(Z,index.return=T)
      Z.ord=aux$x                 #datos observados ordeados de menor a maior
      delta.ord=delta[aux$ix]     #variable indicadora de censura ordeada
      delta.ord[n]=1              #última observación non censurada

      km = 1-KaplanMeier(t,data=Z.ord,censored=1-delta.ord)$surv
      ise = append(sintegral(t,(km-pexp(t))^2)$int,ise)
    }
    mise = mean(ise)
```

```
    vmise = append(vmise,mise)      #vector cos valores do MISE para as
    #distintas porcentaxes de censura
  }
  print(vmise)
}

#####################################################################
# 4. Estudo 4: consistencia do estimador Kaplan-Meier
#####################################################################
#Considérase unha situación con censura pola dereita de Tipo 1.

cte=qnorm(1-0.75,lower.tail = FALSE)     #censura do 25%
C = rep(cte,n)                           #variable de censura
t = seq(-4,4,length=1000)

rep = 1000
ise=c()
for (i in 1:rep){
  Te=rnorm(n)
  Z = pmin(C,Te)
  delta = 1*(Z==Te)
  t = seq(-4,4,length=10000)
  km = 1-KaplanMeier(t,data=Z,censored=1-delta)$surv
  ise = append(sintegral(t,(km-pnorm(t))^2)$int,ise)
}
mise = mean(ise)
print(mise)                              #valor MISE


#####################################################################
# 5. Estudo 5: f.d.c. empírica e estimador de Beran vs. datos censurados
#####################################################################
# Estudo de simulación por Montecarlo para a comparación entre a función de
# distribución condicional empírica e o estimador de Beran coa función de
# distribución condicional teórica en presenza de datos censurados.
# Inclúese unha simulación control sen censura.
# Devólvese o valor do MISE para cada un dos tamaños de mostra, das
# porcentaxes de censura e dos valores da variable explicativa.
# Aproximación da integral con Simpson.
# A obtención da ventá óptima lévase a cabo por validación cruzada.
# Emprégase a función de beran para a estimación da función de distribución
# condicional empírica tomando como variable indicadora de censura un vector
# constante igual a 1.

# 5.1. Modificación da función de Beran para que tome o valor 1 a partires
# da última observación

```

```
ff_beran <- function(ft,fY,fZ,fdelta,fx0,fh){
  b = 1-unlist(beran(x=fY,t=fZ,d=fdelta,x0=fx0,h=fh,testimate=ft)$S)
  b[ft>fZ[length(fZ)]]=1
  return(b)
}

# 5.2. Simulación con con valores de x=0,1,2.
set.seed(12345)
rep = 1000
ro=0.6
t = seq(-4,4,length=10000)
for (x0 in c(0,1,2)){                #valores da explicativa
  print(paste("x=",x0))
  teorica=pnorm(t,0.5*ro*x0,1-ro^2) #distrib. teórica
  for (n in c(25,200,500)){          #tamaño mostra
    print(n)
    vmisea = c()
    vmiseb = c()
    for (mC in  c(1.4,0.9,0.2)){     #porcentaxes de censura 5%, 15%, 30%
      isea=c()
      iseb=c()
      for (i in 1:rep){              #repeticións da simulación
        TX = rbinorm(n,0,0,1,4,ro*2)
        Te = TX[,1]
        X= TX[,2]
        C = rnorm(n,mean=mC,sd=0.5)
        Z = pmin(Te,C)
        delta = 1*(Z==Te)

        aux=sort(Z,index.return=T)
        Z.ord=aux$x                  #datos observados ordeados de menor a maior
        delta.ord=delta[aux$ix]      #variable indicadora de censura ordeada
        delta.ord[n]=1               #última observación non censurada
        X.ord=X[aux$ix]              #covariable X ordeada
        isea = append(sintegral(t,(ff_beran(t,X.ord,Z.ord,rep(1,n),x0)
        -teorica)^2)$int,isea) #f.d.c. empírica
        iseb = append(sintegral(t,(ff_beran(t,X.ord,Z.ord,delta.ord,x0)
        -teorica)^2)$int,iseb) #beran
      }
      vmisea = append(vmisea,mean(isea))
      vmiseb = append(vmiseb,mean(iseb))
    }
    print(paste("Empir",vmisea))
    print(paste("Beran",vmiseb))
  }
}

```

```
#5.3. Mostra control

set.seed(12345)
rep = 1000
ro=0.6
t = seq(-3,3,length=10000)
for (x0 in c(0,1,2)){              #valores da covariable
  print(paste("x=",x0))
  teorica=pnorm(t,0.5*ro*x0,1-ro^2)
  for (n in c(25,200,500)){        #tamaño mostra
    print(n)
    isea=c()
    for (i in 1:rep){              #repeticións da simulación
      TX = rbinorm(n,0,0,1,4,ro*2)
      Te = TX[,1]
      X= TX[,2]
      Z = Te
      delta = 1*(Z==Te)
      aux=sort(Z,index.return=T)
      Z.ord=aux$x                  #datos observados ordeados de menor a maior
      delta.ord=delta[aux$ix]      #variable indicadora de censura ordeada
      delta.ord[n]=1               #última observación non censurada
      X.ord=X[aux$ix]              #covariable ordeada
      isea = append(sintegral(t,(ff_beran(t,X.ord,Z.ord,rep(1,n),x0)
      -teorica)^2)$int,isea)
    }
    print(paste("Emp-Beran",mean(isea)))
  }
}
```

Anexo B

Código asociado ás representacións gráficas

Ao longo deste anexo preséntase o código de R empregado para levar a cabo as representacións gráficas presentadas ao longo deste Traballo de Fin de Grao.

```r
#####################################################################
#FIGURA 1.1.
#####################################################################
t=seq(0, 10, length=1000 )
plot(t,pexp(t,rate=1),type="l",xlab=" ",ylab=" ",ylim=c(0,1),xlim=c(0,10))
plot(t,1-pexp(t,rate=1),type="l",xlab=" ",ylab=" ",ylim=c(0,1),xlim=c(0,10))
plot(t,rep(1,1000),type="l",xlab=" ",ylab=" ",ylim=c(0,2),xlim=c(0,10))


#####################################################################
#FIGURA 1.2.
#####################################################################
ggplot(data=data_frame(index = c(1:6),cancer[82:87,2:3]),
mapping = aes(x = time, y = index, shape=as.factor(status)))+
geom_point(size=2,stroke=1)+
scale_y_continuous(breaks = c(1:6),labels = as.character(c(1:6)))+
geom_linerange(aes(xmin = 0, xmax = time))+
geom_linerange(aes(ymin = 0, ymax = 6,x=max(time)),lty="dotted")+
scale_shape_manual(name =' ', values = c(4, 19),
labels = c("Censurado", "Non censurado"))+
ylab("Pacientes")+ xlab("Días")+
theme_minimal()

#####################################################################
#FIGURA 2.1.
```

```
26 ######################################################################
27 y1 = rexp(5,1)
28 y2 = rexp(15,1)
29 y3 = rexp(500,1)
30 windows()
31 plot(t,pexp(t,rate=1),type="l",ylim=c(0,1),xlim=c(0,6),lwd=2,
32 ylab="Función de distribución",xlab=" ")
33 lines(c(0,t,6),c(0,ecdf(y1)(t),1),type="s",col="#FE3F4F",lwd=2)
34 lines(c(0,t,6),c(0,ecdf(y2)(t),1),type="s",col="#9DA92D",lwd=2)
35 lines(c(0,t,6),c(0,ecdf(y3)(t),1),type="s",col="#3789D2",lwd=2)
36 windows()
37 plot(t,1-pexp(t,rate=1),type="l",ylim=c(0,1),xlim=c(0,6),lwd=2,
38 ylab="Función de supervivencia",xlab=" ")
39 lines(c(0,t,6),c(1,1-ecdf(y1)(t),0),type="s",col="#FE3F4F",lwd=2)
40 lines(c(0,t,6),c(1,1-ecdf(y2)(t),0),type="s",col="#9DA92D",lwd=2)
41 lines(c(0,t,6),c(1,1-ecdf(y3)(t),0),type="s",col="#3789D2",lwd=2)
42
43
44 ######################################################################
45 #FIGURA 2.2.
46 ######################################################################
47 t=seq(-6, 6, .01)
48 for (mC in c(1.8,1.1,0.6)){
49   windows()
50   Te=dnorm(t)
51   C=dnorm(t,mean=mC,sd=0.5)
52   plot(t,Te,type="l",xlab=" ",ylab=" ",ylim=c(0,0.8), xlim=c(-4,4))
53   lines(t,C)
54   abline(v=0,lty=3)
55   text(x=mC+1.5,y=0.7, paste("N(",mC,",0.5)"), cex = 1.3)
56   polygon(c(t[t>=min(t)], max(t), 0), c(pmin(Te,C)[t>=min(t)], 0, 0),
57   col="#EFB40E",density = 20, angle = 45,border="black")
58   abline(v=mC,lty=3)
59 }
60
61 ######################################################################
62 #FIGURA 2.3.
63 ######################################################################
64 t=seq(-6, 6, .01)
65 f_emp <- function(n,color,mTe,sdTe,mCe,sdCe){
66   Te = rnorm(n,mean=mTe,sd=sdTe)
67   C = rnorm(n,mean=mCe,sd=sdCe)
68   Z = pmin(Te,C)
69   delta = 1*(Te==Z)
70   mostra = sort(Z)
71   mostra_emp = seq(1,n)/n
72   lines(c(-6,mostra,6),c(0,mostra_emp,1),type="s",col=color,lwd=2)
```

```
}
j=1
for (mC in c(1.8,1.1,0.6)){
  windows()
  plot(t,pnorm(t,sd=sdT),type="l",xlab=" ",ylab=" ",ylim=c(0,1),xlim=c(-3,3),
  lwd=2)
  j=1
  for (i in c(10,150,500)){
    f_emp(i,cores_i[j],0,1,mC,0.5)
    j=j+1
  }
  j=j+1
}

#########################################################################
#FIGURA 2.4.
#########################################################################
can=cancer[82:87,2:3]
can=can[order(can$time),]
dd=data_frame(f=c(1/6,2/6,2/6,5/9,7/9,1),can)
ggplot()+
geom_linerange(data=dd,aes(xmin=0, xmax=min(time), y=0))+
geom_linerange(data=dd,aes(ymin=0, ymax=1/6, x=min(time)))+
scale_y_continuous(breaks = c(0,1/6,2/6,5/9,7/9,1),
labels = c('0','1/6','2/6','5/9','7/9','1'))+
scale_shape_manual(name =' ', values = c(4, 19),
labels = c("Censurado", "Non censurado"))+
geom_step(data=dd,mapping=(aes(x=time,y=f)))+
geom_point(data=dd,mapping=(aes(x=time,y=rep(0,6),shape=as.factor(status))),
size=2,stroke=1)+
xlab("Días")+ylab(" ")+
theme_minimal()

#########################################################################
#FIGURA 2.5.
#########################################################################
t=seq(-3, 3, .01)
cores_i = c("#FE3F4F","#9DA92D","#3789D2")
for (mC in c(1.8,1.1,0.6)){
  windows()
  plot(t,pnorm(t),type="l",xlab=" ",ylab=" ",ylim=c(0,1),xlim=c(-3,3),lwd=2)
  j=1
  for (n in c(10,150,500)){
    Te = rnorm(n)                  #a teórica
    C = rnorm(n,mean=mC,sd=0.5)
    Z = pmin(Te,C)                 #a observada con censura
    delta = 1*(Te==Z)
```

```
120       aux=sort(Z,index.return=T)
121       Z.ord=aux$x                    #datos observados ordeados de menor a maior
122       delta.ord=delta[aux$ix]        #variable indicadora de censura ordeada
123       delta.ord[n]=1                 #última observación non censurada
124
125       km = 1-KaplanMeier(t,data=Z.ord,censored=1-delta.ord)$surv
126       lines(c(-6,t,6),c(0,km,1),type="s",col=cores_i[j],lwd=2)
127       j = j+1
128     }
129   }
130
131   ##########################################################################
132   #FIGURA 2.6.
133   ##########################################################################
134   n=500
135     Te=rnorm(n)
136   Z = pmin(C,Te)
137   delta = 1*(Z==Te)
138   t = seq(-4,4,length=10000)
139   plot(t,1-KaplanMeier(t,data=Z,censored=1-delta)$surv ,type="s",
140   col="#3789D2",ylim=c(0,1),ylab="",xlab="")
141   polygon(c(t[t>=cte], max(t), cte), c(y[t>=cte], 0, 0), col="#EFB40E",
142   density = 10, angle = 45)
143   lines(t,pnorm(t),type="l")
144
145   ##########################################################################
146   #FIGURA 3.1.
147   ##########################################################################
148   set.seed(12345)
149   ro=0.6
150   x0=1
151   n=100
152   TX = rbinorm(n,0,0,1,4,ro*1*2)
153   Te = TX[,1]
154   X= TX[,2]
155   C = rnorm(n,mean=0.9,sd=0.5)
156   Z = pmin(Te,C)
157   delta = 1*(Z==Te)
158   aux=sort(Z,index.return=T)
159   Z.ord=aux$x                    #datos observados ordeados de menor a maior
160   delta.ord=delta[aux$ix]        #variable indicadora de censura ordeada
161   delta.ord[n]=1                 #última observación non censurada
162   X.ord=X[aux$ix]
163   print(beran(X.ord,Z.ord,delta.ord,x0=x0)$h)
164   plot(t,ff_beran(t,X.ord,Z.ord,delta.ord,x0),xlim=c(-2.5,2.5),ylim=c(0,1),
165   type="l",col="#FE3F4F",xlab="",ylab="",lwd=2)
166   lines(t,ff_beran(t,X.ord,Z.ord,delta.ord,x0,10),col="#9DA92D",lwd=2)
```

```
lines(t,ff_beran(t,X.ord,Z.ord,rep(1,n),x0,0.1),col="#3789D2",lwd=2)
lines(t,1-KaplanMeier(t,data=Z.ord,censored=1-delta.ord)$surv,lty=3,lwd=2)
lines(t,pnorm(t,0.5*ro*x0,sqrt(1-ro^2)),lwd=2)
legend(-2.2, 0.92, legend=c("Beran, h=3.67(CV)","Beran, h=0.1","Beran, h=10",
"Kaplan-Meier","Teórica"),col=c("#FE3F4F","#3789D2","#9DA92D","black",
"black"),lty=c(1,1,1,3,1),lwd=c(2,2,2,2,2),cex=0.8,bty = "n")

#########################################################################
#FIGURA 3.2.
#########################################################################
set.seed(123456)
ro=0.6
n=500
windows()
par(mfrow=c(3,3))
t = seq(-4,4,length=10000)
for (x0 in c(0,1,2)){
  cont = 1
  censura=c(10,20,40)
  for (mC in c(1.4,0.9,0.2)){
    TX = rbinorm(n,0,0,1,4,ro*2)
    Te = TX[,1]
    X= TX[,2]
    C = rnorm(n,mean=mC,sd=0.5)
    Z = pmin(Te,C)
    delta = 1*(Z==Te)
    print(sum(delta))
    aux=sort(Z,index.return=T)
    Z.ord=aux$x                    #datos observados ordeados de menor a maior
    delta.ord=delta[aux$ix]        #variable indicadora de censura ordeada
    delta.ord[n]=1
    X.ord=X[aux$ix]
    plot(t,ff_beran(t,X.ord,Z.ord,delta.ord,x0),xlim=c(-3,3),ylim=c(0,1),
    type="l",col="#9DA92D",xlab="",ylab="",lwd=2)#,main=paste("x=",x0,", ",
    censura[cont],"% cens."))
    lines(t,ff_beran(t,X.ord,Z.ord,rep(1,n),x0),col="#FE3F4F",lwd=2)
    lines(t,pnorm(t,0.5*ro*x0,sqrt(1-ro^2)),type="l")
    cont = cont+1
  }
}

#########################################################################
#FIGURA 4.1.
#########################################################################
barplot(table(lungs$status)/1000,col=c("#EFB40E","black"),density=20,angle=45,
ylab="Frecuencia relativa", ylim=c(0,1))
barplot(table(lungs$sex)/1000,col=c("#3789D2","#A7181C"),density=20,angle=45,
```

```
214 ylab="Frecuencia relativa", ylim=c(0,1))
215 barplot(table(lungs$loc)/1000,col=c("#9DA92D","#514293","#D47E31","#808080"),
216 density=20,angle=45,ylab="Frecuencia relativa", ylim=c(0,1))
217 hist(lungs$age,main=" ",xlab="Idade (en anos)",col="white",freq=F,
218 ylab="Densidade")
219 hist(lungs$size,main=" ",,xlab="Tamaño do tumor (en milímetros)",col="white",
220 freq=F,ylab="Densidade")
221
222 ######################################################################
223 #FIGURA 4.2.
224 ######################################################################
225 boxplot(lungs$size~lungs$loc,border=c("#9DA92D","#514293","#D47E31","#808080")
      ,
226 xlab="Localización",ylab="Tamaño do tumor (en milímetros)", density=20,
227 angle=45)
228
229 ######################################################################
230 #FIGURA 4.3.
231 ######################################################################
232 plot(seq(1,1000),lungs$time,xlab="",ylab="Tempo de supervivencia",xaxt="n")
233 points(seq(1,1000)[lungs$status=="Censurado"],
234 lungs$time[lungs$status=="Censurado"],col="#EFB40E",pch=19)
235
236 ######################################################################
237 #FIGURA 4.4.
238 ######################################################################
239 levels(lungs$status) = c(0,1)
240 lungs$status=as.numeric(lungs$status)
241 aux=sort(lungs$time,index.return=T)
242 time.ord=aux$x                    #datos observados ordeados de menor a maior
243 delta.ord=lungs$status[aux$ix]
244 age.ord=lungs$age[aux$ix]
245 size.ord=lungs$size[aux$ix]
246 delta.ord[1000]=1
247 meses=seq(0,time.ord[1000],0.001)
248
249 km=1-KaplanMeier(meses,data=time.ord,censored=1-delta.ord)$surv
250 empir=ecdf(time.ord)(meses)
251 datsurv = data.frame(meses,km,empir)
252
253 ggplot()+
254 geom_line(data = datsurv, aes(x =meses, y=1-empir), color = "#EFB40E")+
255 geom_line(data = datsurv, aes(x =meses, y=1-km))+
256 ylab("Función de supervivencia")+ xlab("Tempo (en meses)")+
257 theme_minimal()
258
259 ggplot()+
```

```
260  geom_line(data = datsurv, aes(x =meses, y=empir), color = "#EFB40E")+
261  geom_line(data = datsurv, aes(x =meses, y=km))+
262  ylab("Función de distribución")+ xlab("Tempo (en meses)")+
263  theme_minimal()
264
265  ########################################################################
266  #FIGURA 4.5.
267  ########################################################################
268  autoplot(survfit(Surv(time, status) ~ sex, data=lungs))+
269  scale_color_manual(values=c("#3789D2","#A7181C"))+
270  scale_fill_manual(values=c("#3789D2","#A7181C"))+
271  ylim(0,1)+
272  ylab("Función de supervivencia")+ xlab("Tempo (en meses)")+
273  theme_minimal()+
274  theme(legend.title=element_blank())
275
276  ########################################################################
277  #FIGURA 4.5.
278  ########################################################################
279  autoplot(survfit(Surv(time, status) ~ loc, data=lungs))+
280  scale_color_manual(values=c("#9DA92D","#514293","#D47E31","#808080"))+
281  scale_fill_manual(values=c("#9DA92D","#514293","#D47E31","#808080"))+
282  ylim(0,1)+
283  ylab("Función de supervivencia")+ xlab("Tempo (en meses)")+
284  theme_minimal()+
285  theme(legend.title=element_blank())
286
287  ########################################################################
288  #FIGURA 4.6.
289  ########################################################################
290  plot(meses,1-ff_beran(meses,age.ord,time.ord,delta.ord,50),lwd=2,xlim=c(0,160)
       ,
291  ylim=c(0,1),type="l",xlab="Tempo(en meses)",
292  ylab="Función de supervivencia",col="#D47E31")
293  lines(meses,1-ff_beran(meses,age.ord,time.ord,delta.ord,70),lwd=2,type="l"
294  ,col="#9DA92D")
295  lines(meses,1-ff_beran(meses,age.ord,time.ord,delta.ord,90),lwd=2,type="l"
296  ,col="#3789D2")
297  lines(meses,1-ff_beran(meses,age.ord,time.ord,rep(1,1000),50),lty=3,lwd=2
298  ,type="l",col="#D47E31")
299  lines(meses,1-ff_beran(meses,age.ord,time.ord,rep(1,1000),70),lty=3,lwd=2
300  ,type="l",col="#9DA92D")
301  lines(meses,1-ff_beran(meses,age.ord,time.ord,rep(1,1000),90),lty=3,lwd=2
302  ,type="l",col="#3789D2")
303  legend(105, 0.8, legend=c("50 anos", "70 anos","90 anos"),
304  col=c("#D47E31", "#9DA92D","#3789D2"),lty=c(1,1,1),lwd=c(2,2,2),cex=0.8)
305
```

```
306 ######################################################################
307 #FIGURA 4.7.
308 ######################################################################
309 plot(meses,1-ff_beran(meses,size.ord,time.ord,delta.ord,5),lwd=2,xlim=c(0,160)
      ,
310 ylim=c(0,1),type="l",xlab="Tempo(en meses)"
311 ,ylab="Función de supervivencia",col="#D47E31")
312 lines(meses,1-ff_beran(meses,size.ord,time.ord,delta.ord,25),lwd=2,type="l",
313 col="#9DA92D")
314 lines(meses,1-ff_beran(meses,size.ord,time.ord,delta.ord,50),lwd=2,type="l",
315 col="#3789D2")
316 lines(meses,1-ff_beran(meses,size.ord,time.ord,delta.ord,100),lwd=2,type="l",
317 col="#514293")
318 lines(meses,1-ff_beran(meses,size.ord,time.ord,rep(1,1000),5),lty=3,lwd=2,
319 type="l",col="#D47E31")
320 lines(meses,1-ff_beran(meses,size.ord,time.ord,rep(1,1000),25),lty=3,lwd=2,
321 type="l",col="#9DA92D")
322 lines(meses,1-ff_beran(meses,size.ord,time.ord,rep(1,1000),50),lty=3,lwd=2,
323 type="l",col="#3789D2")
324 lines(meses,1-ff_beran(meses,size.ord,time.ord,rep(1,1000),100),lty=3,lwd=2,
325 type="l",col="#514293")
326 legend(115, 0.95, legend=c("    5mm", "  25mm","  50mm","100mm"),
327 col=c("#D47E31", "#9DA92D","#3789D2","#514293"), lty=c(1,1,1,1),
328 lwd=c(2,2,2,2), cex=0.8)
329
330 ######################################################################
331 #FIGURA 4.8.
332 ######################################################################
333 loc.times=lungs$time[lungs$loc=="Distante"]
334 loc.status=lungs$status[lungs$loc=="Distante"]
335 loc.size=lungs$size[lungs$loc=="Distante"]
336 laux=sort(loc.times,index.return=T)
337 loc.time.ord=laux$x
338 loc.delta.ord=loc.status[laux$ix]
339 loc.size.ord=loc.size[laux$ix]
340 plot(meses,1-ff_beran(meses,loc.size.ord,loc.time.ord,loc.delta.ord,5),lwd=2,
341 xlim=c(0,160),ylim=c(0,1),type="l",xlab="Tempo(en meses)"
342 ,ylab="Función de supervivencia",col="#D47E31")
343 lines(meses,1-ff_beran(meses,loc.size.ord,loc.time.ord,loc.delta.ord,20),lwd=2
344 ,type="l",col="#9DA92D")
345 lines(meses,1-ff_beran(meses,loc.size.ord,loc.time.ord,loc.delta.ord,50),lwd
      =2,
346 type="l", col="#3789D2")
347 lines(meses,1-ff_beran(meses,loc.size.ord,loc.time.ord,loc.delta.ord,100),
348 lwd=2,type="l",col="#514293")
349 legend(115, 0.95, legend=c("    5mm", "  20mm","  50mm","100mm"),
350 col=c("#D47E31", "#9DA92D","#3789D2","#514293"), lty=c(1,1,1,1),
```

```
351   lwd=c(2,2,2,2), cex=0.8)
```

Bibliografía

[1] Klein, J.P. e Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data (Statistics for Biology and Health).* Springer.

[2] Reynkens, T., Verbelen, R., Bardoutsos, A., Cornilly, D., Goegebeur, Y. e Herrmann, K. (2023). *ReIns: Functions from Reinsurance: Actuarial and Statistical Aspects"*,versión 1.0.14, https://CRAN.R-project.org/package=ReIns.

[3] López-de-Ullibarri, I., López-Cheda, A. e Jácome, M.A. (2020). *npcure: Nonparametric Estimation in Mixture Cure Models*, versión 0.1-5, https://CRAN.R-project.org/package=npcure.

[4] Therneau, T.M., Lumley, T., Atkinson, E. e Crowson C. (2024). *survival: Survival Analysis*, versión 3.7-0, https://CRAN.R-project.org/package=survival.

[5] Gijbels, I. (2010). *Censored data.* Wiley Interdisciplinary Reviews: Computational Statistics, 2, 178-188.

[6] Kaplan, E.L. e Meier P. (1958). *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, 53 (282), 457-481.

[7] Kalbfleisch, J.D. e Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* 2º Edición, John Wiley & Sons.

[8] Vélez-Ibarrola, R. e García-Pérez, A. (1997). *Principios de inferencia estadística.* Universidad Nacional de Educación a Distancia, Madrid.

[9] Gill, R.D. (1992). *Lectures on Probability Theory.* École d'Eté de Probabilités de Saint Flour XXII, Springer.

[10] Sawyer, S. (2003). *The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis.* Washington University in St. Louis.

[11] Greenwood, M.J. (1926). *The Natural Duration of Cancer.* Reports of Public Health and Related Subjects, Volumen 33, HMSO, Londres.

[12] Borgan O. e Liestøl K. (1990). *A note on confidence intervals and bands for the survival function based on transformations.* Scandinavian Journal of Statististics 17, 35-41.

[13] Wand, M.P. e Jones, M.C. (1994). *Kernel Smoothing.* CRC press, Nova York.

[14] Beran, R. (1981). *Nonparametric regression with randomly censored survival data.* Technical report, University of California, Berkeley.

[15] Amico, M. e Van Keilegom, I. (2018). *Cure models in survival analysis.* Annual Review of Statistics and Its Application, 5(1), 311-342.