Esenciais USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANÁLISE DE SUPERVIVENCIA
función de supervivencia e función de risco

Alberto Rodríguez Casal, Alejandro Saavedra Nieves e Paula Saavedra Nieves (Departamento de Estatística, Análise Matemática e Optimización)

A análise de supervivencia ten como obxectivo fundamental estudar a variable aleatoria non negativa X que mide o tempo que pasa dende un instante inicial (ben definido) ata que ocorre un evento de interese. Os datos mostrais proporcionados en moitos estudos clínicos miden o tempo ata a morte dun paciente dende o diagnóstico dunha enfermidade, o tempo de recorrencia dende a finalización dun tratamento ou o tempo de eficacia dunha intervención dende que foi realizada. Neste contexto, é moi común incorporar individuos ao estudo que comezan a ser observados posteriormente ao evento inicial ou pechar o estudo antes de que se dea o evento de interese (por exemplo, a morte) en todos os participantes do estudo. Por tanto, os datos de supervivencia adoitan denominarse incompletos ao estar afectados polos fenómenos de truncamento ou censura (Klein e Moeschberger, 2003).

Este documento introduce a situación máis sinxela supoñendo que dispoñemos dunha mostra aleatoria $X_1,...,X_n$ de observacións de X independentes e identicamente distribuídas (iid).

Neste contexto, unha función de gran interese é a función de supervivencia $S(x) = P(X > x)$, $x \geq 0$, que proporciona, para cada valor x, a probabilidade de supervivencia máis alá do instante de tempo x. Se F denota la función de distribución da variable X, $F(x) = P(X \geq x)$, $x \in R$. Entón, podemos deducir que $S(x) = 1 - F(x)$, $x \geq 0$. Outra función que proporciona información moi valiosa na análise de supervivencia é a función de risco h que mide o risco de que un individuo que está vivo nun instante concreto, morra nun instante inmediatamente posterior. Se f denota a función de densidade de X, $h(x) = f(x)/S(x)$, $x \geq 0$.

Exemplo 1: Sexa X a variable que mide o tempo (en anos) ata a recorrencia dun determinado tipo de cancro despois de recibir un tratamento de quimioterapia. A Figura 1 contén as funcións de supervivencia (esquerda) e de risco (centro) para o Grupo 1 (negro) onde se observou diseminación da enfermidade, e para o Grupo 2 (gris) onde esta non foi detectada. Para o Grupo 1, S(3) é igual a 0.4 e correspóndese coa probabilidade de que o tempo a recaída supere os 3 anos. Porén, h(3), que é 0.75, proporciona o risco de que un individuo sufra unha recorrencia aos 3 anos dende que recibiu o tratamento (asumindo que sobreviviu ata ese instante). Non obstante, S(3) vale 0.9 no Grupo 2 e h(3), 0.15. O prognóstico é claramente máis favorable no Grupo 2.

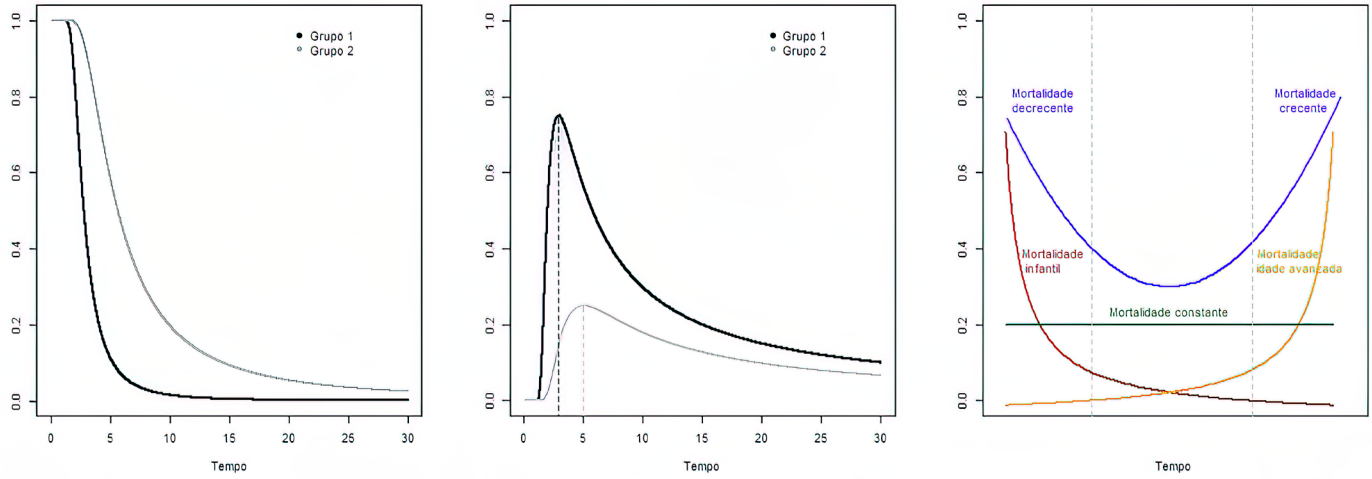

Figura 1: Funcións de supervivencia S (esquerda) e de risco h (centro) para os Grupos 1 (negro) e 2 (gris); tipos de funcións de risco (dereita).

Á vista da **Figura 1** (centro), a función de risco no Grupo 1 acada o pico máximo de recidivas arredor dos 3 anos. Para o Grupo 2, o máximo acádase aos 5 anos pero a probabilidade de recorrencia é, en xeral, claramente inferior ca no primeiro Grupo. Esta información ten un valor clínico chave. Un seguimento máis exhaustivo dos pacientes con diseminación antes do terceiro ano podería detectar as recaídas nunha fase precoz con máis posibilidades terapéuticas. Moitas veces, unha ferramenta tan útil coma esta é infrautilizada nos ambientes médicos.

Finalmente, a **Figura 1** (dereita) mostra exemplos clásicos de funcións de risco na análise de supervivencia. A curva vermella asóciase a enfermidades con elevada mortalidade infantil con riscos de falecemento elevados nos primeiros anos de vida; a curva amarela soe corresponderse con enfermidades asociadas á vellez; a curva azul (*bathtub curve)* aparece comunmente cando a variable X mide o tempo ata o fallo dun sistema ou máquina; claramente, distínguense tres etapas: (1) risco inicial de fallo elevado (equipo defectuoso ou mala instalación), (2) fallos normais: etapa cunha taxa de erros menor que pode ser case constante e (3) fallos de desgaste ao final da vida útil da máquina. En enxeñaría, este tipo particular de estudos desenvolveuse de xeito acelerado a partir da Segunda Guerra Mundial antes de que estas técnicas foran aplicadas no campo das ciencias da saúde cara aos anos setenta.

Na práctica, f é descoñecida e, en consecuencia, tamén o son F, S e h. O habitual é dispoñer só da mostra de observacións $X_1,\ldots,X_n$ de X. En certas ocasións, o investigador pode ter, de xeito adicional, información sobre o comportamento da variable X na poboación baixo estudo. Ese coñecemento permitiría asociar f con algunha familia de funcións de densidade. Aínda que calquera distribución con soporte nos números reais non negativos podería considerarse (queda excluída a distribución normal estándar), os modelos de supervivencia paramétricos soen considerar as distribucións exponencial ou Weibull.

En particular, se asumimos que X segue unha distribución exponencial con parámetro a > 0, $f(x) = ae^{-ax}$ para x ≥ 0. Neste caso, a media de X é 1/a, $S(x)= e^{-ax}$ e h(x)=a. Xeralmente, o valor do parámetro é descoñecido. Por iso, pode estimarse a partir dos datos $X_1,\ldots,X_n$ empregando o método de máxima verosimilitude (Evans et al., 2011). Nótese que a función de risco exponencial é constante (ver Figura 1, dereita, función de risco verde). Este feito implica que o risco non cambia ao longo do tempo. Coloquialmente, dise que a distribución exponencial sofre *perda de memoria* porque o risco de que ocorra o evento de interese nun intervalo de tempo concreto non depende da idade do individuo. Por iso, o seu uso en aplicacións biomédicas é bastante escaso. Habitualmente, a distribución que se emprega para modelar datos de supervivencia é a Weibull. Se asumimos que X segue unha distribución Weibull con parámetros da forma b > 0 e da escala c > 0, $f(x) = (b/c)(x/c)^{b-1}\exp(-(x/c)^b)$ para x ≥ 0. Neste caso, $S(x) = \exp(-(x/c)^b)$ e $h(x)= (b/c)(x/c)^{b-1}$. Agora, a función de risco poder ser non constante. Como $h'(x) = [b(b-1)/c^2](x/c)^{b-2}$, se b > 1 entón h'(x) > 0 e h modelará riscos monótonos crecentes; se b <1, h'(x) < 0 e h modelará riscos monótonos decrecentes; finalmente, se b = 1, a distribución de X coincide cunha exponencial de parámetro 1/c e, en consecuencia, a función de risco é constante. A Figura 2 (esquerda e centro) contén as funcións de supervivencia e de risco para a distribución Weibull para tres valores de b diferentes (b = 0.5, 1 e 3) e c = 3. A diferenza do modelo exponencial, non é posible calcular de forma exacta os estimadores de máxima verosimilitude $\hat{b}$ e $\hat{c}$. É preciso empregar un método iterativo de aproximación como Newton-Raphson. Para os datos do Exemplo 2, a Figura 2 (dereita) contén as representacións gráficas das funcións de supervivencia ($\hat{S}$, verde) e risco ($\hat{h}$, laranxa) estimadas.

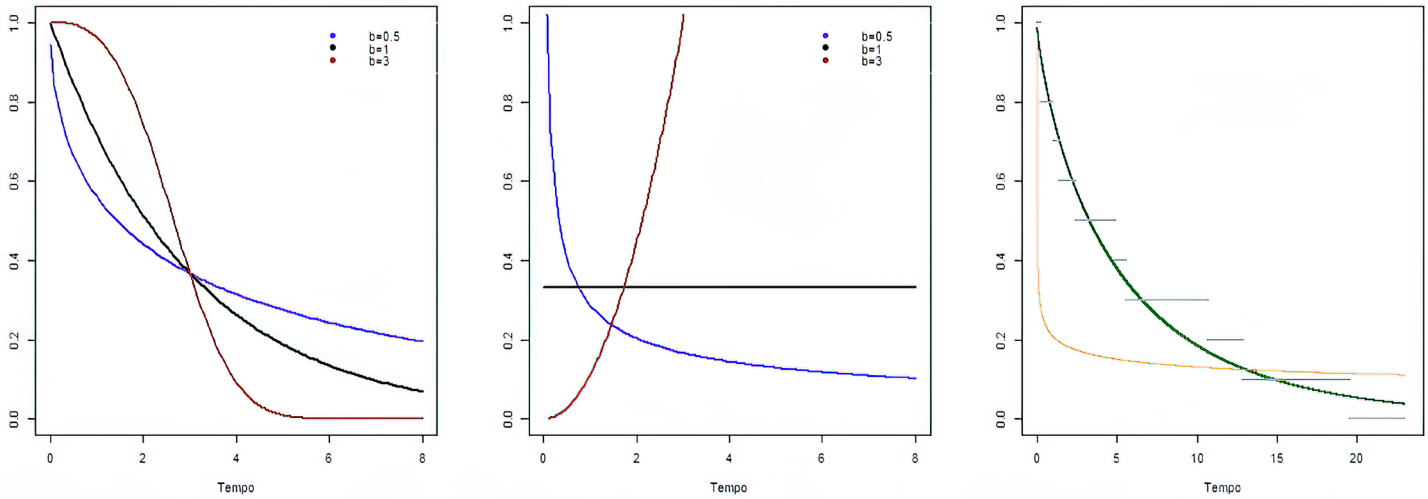

Figura 2: Funcións de supervivencia (esquerda) e de risco h (centro) para a distribución Weibull con c = 3. Estimadores de máxima verosimilitude $\hat{S}$ (verde) e $\hat{h}$ (laranxa) para as funcións de supervivencia e risco calculados a partir dos datos do Exemplo 2; estimador empírico $\hat{S}_n$ (gris) da función de supervivencia (dereita).

Exemplo 2: En 10 doentes con esclerose múltiple remitente-recorrente mediuse o tempo (en anos) entre o primeiro brote producido pola enfermidade e o segundo. Os valores rexistrados foron: 5.6, 19.5, 12.9, 1.4, 2.4, 0.2, 1, 10.7, 0.2, 4.9. Se asuminos que a variable tempo transcorrido dende o primeiro ao segundo brote segue unha distribución Weibull, o software estatístico R (R Core Team, 2022), permítenos estimar os valores dos parámetros b e c empregando o paquete `EnvStats`:

```
> datos <- c(5.6,19.5,12.9,1.4,2.4,0.2,1,
10.7,0.2,4.9)
> eweibull(datos)
```

De acordo cos resultados obtidos, $\hat{b} = 0.8$ e $\hat{c} = 5.2$. En particular, a Figura 2 (dereita) contén a representación gráfica das funcións de supervivencia e risco estimadas: $\hat{h}(x)= \exp(-(x/5.2)^{0.8})$ (verde) e $\hat{S}(x) = 0.15(x/5.2)^{-0.2}$ (laranxa). A continuación, móstrase o código R que permite obter a función de supervivencia empírica S_n e, como exemplo, a súa avaliación en 3.

```
> Sn <- 1-ecdf(sample); Sn(3)
```

Baixo a hipótese iid, os modelos de supervivencia non paramétrica estiman S de forma empírica. Tendo en conta a súa definición, para cada x ≥ 0, calculan a proporción de individuos onde a variable X supera o valor x. Matematicamente, $\hat{S}_n(x) = (1/n) \sum_{i=1}^{n} I(X_i > x)$ onde I denota a función indicadora. Para os datos do Exemplo 2, a Figura 2 (dereita, gris) contén a representación gráfica de $\hat{S}_n$. Esta aproximación discontinua é ademais o estimador non paramétrico de máxima verosimilitude. Desafortunadamente, na maioría das aplicacións con datos de supervivencia non se cumpre o suposto iid, e precísanse estimadores alternativos aos propostos para a función de supervivencia coma o proposto en Kaplan e Meier (1958) pola presenza de datos censurados.

Referencias

Evans, M., Hastings, N., Peacock, B., & Forbes, C. (2011). Statistical distributions. John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data (Vol. 1230). New York: Springer.

R Core Team (2021). R: A language and environment for statistical computing. En: https://www.R-project.org/

https://dx.doi.org/10.15304/9788410142022