Introducción: Conceptos preliminares

Giancarlo Sal y Rosas, Ph.D.

Departamento de Ciencias
Pontificia Universidad Católica del Perú
vsalyrosas@pucp.edu.pe

15 de marzo de 2017

Outline

1. Motivación
2. Notación / Modelos / Conceptos
3. Máxima Verosimilitud
4. Censura

Caso 1: Estudio HPTN 039

- **Hipótesis**: El tratamiento contra VHS-2 (virus del herpes typo 2) reduce el riesgo de infección por VIH-1.
- **Población**: mujeres y hombres que tienen sexo con otros hombres VIH-1 negativas y VHS-2 positivas en alto riesgo de adquirir VIH.
- **Grupos**: Participantes fueron aleatorizados a:
  - **Intervención**: Aciclovir un medicamento muy efectivo para tratar VHS-2 y sin efectos adversos conocidos
  - **Control**: Un placebo que lucía fisicamente similar al aciclovir

Caso 1: Estudio HPTN 039

- **Localización**: El estudio incluyó 4 ciudades de EEUU, 3 de Perú y 6 de África.
- **Visitas**: El periodo de seguimiento se dio cada 3 meses y duró entre 12 a 18 meses.
- **Respuesta**: Tiempo hasta la adquisición de VIH.
- **Objetivo científico**: El tratamiento con aciclovir reduce el riesgo de infección de VIH.
- **Objetivo estadístico**: La función de supervivencia es diferente (mayor) para el grupo que tomó aciclovir que para el que no lo tomó.
- Celum et al. [1] publicaron los resultados de este estudio.

Estructura de los datos: HPTN 039

```
> hptn <- read.csv("hptn.csv")
> head(hptn)
       ptid   t1   t2 event arm
1 203000045   NA   NA     0   1
2 203000045    0   90     0   1
3 203000045   90  175     0   1
4 203000045  175  266     0   1
5 203000045  266  357     0   1
6 203000053   NA   NA     0   1
```

- **ptid**: Codigo para identificar al participante
- **t1**, **t2**: $(t1, t2)$ intervalo de tiempo de observación
- **event**: Si el evento (infección) ocurrio (1) o no (0)
- **arm**: Grupo de intervención (1) o control (0)

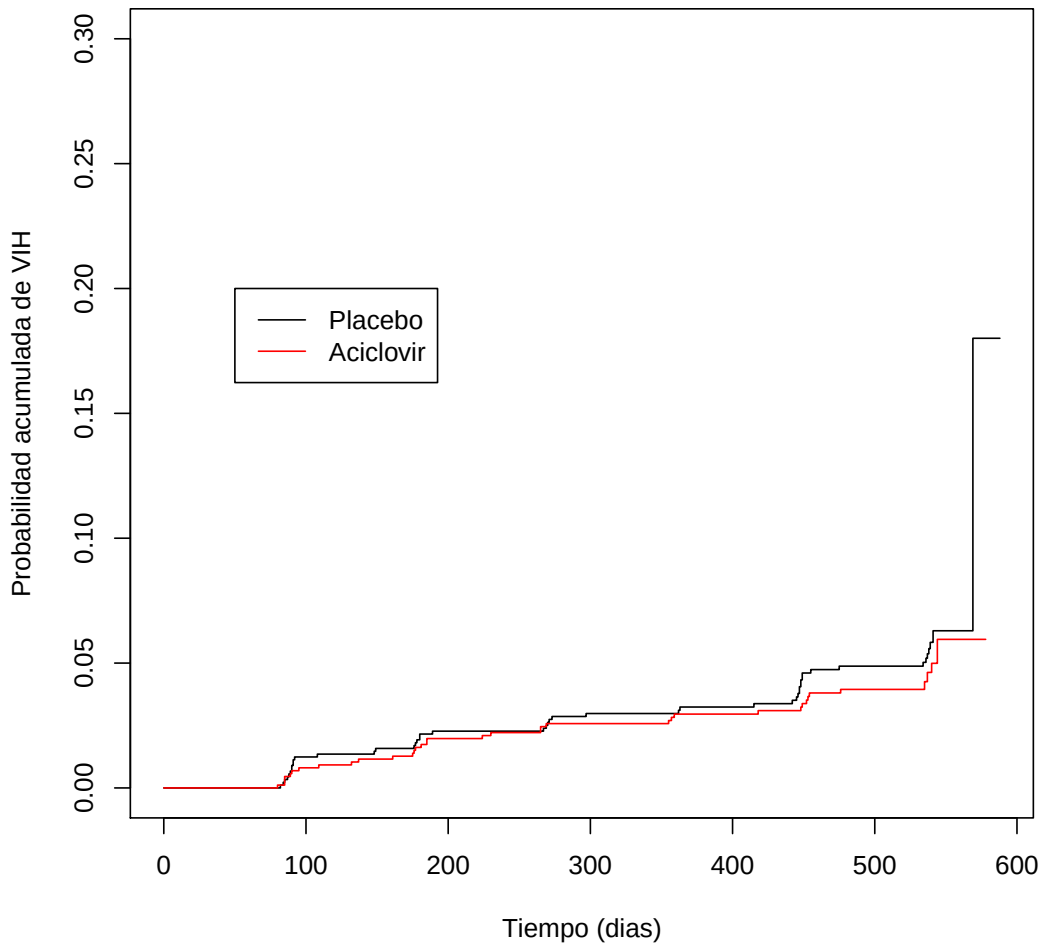


Figura : Caso 1: Estudio HPTN 039

Caso 2: Estudio de Notificación de Parejas

- **Población**: Hombres y mujeres recientemente curados de gonorrea y/o chlamydia.
- **Grupos**: Participantes fueron aleatorizados a:
  - **Intervención**: Tickets para reclamar medicación para entregar a sus parejas o si lo preferían, personal médico podría contactarlos y entegarles la medicación sin examinación médica.
  - **Control**: Tratamiento estándar
- **Respuesta**: Persistencia o recurrencia de gonorrea y/o chlamydia en el paciente enrolado.

Caso 2: Estudio de Notificación de Parejas

- **Medidas**: Cada individuo tuvo dos visitas
  - Una visita de enrolamiento
  - Una visita de seguimiento (entre las semanas 3 y 19)
- **Objetivo científico**: Mostrar que la intervención reduce la recurrencia y/o persistencia de gonorrea y/o chlamydia.
- **Objetivo estadístico**: Mostrar que la función de supervivencia es diferente para ambos grupos.
- Golden et al. [2] y Sal y Rosas and Hughes [3] publicaron los resultados de este estudio.

Estructura de los datos

```
> head(pns)
     gender arm   ptid ct_gc ct_res gc_res      age corrctgc folltime follprim
1676      1   0 B2697     B      1      1 20.93908        1       19        0
1745      1   1 B2767     C      0     NA 22.53525        0       21        1
2263      1   1 B3286     C      1     NA 21.05133        1       21        1
445       1   0 B1455     G      0      0 23.96441        0       23        1
1693      1   1 B2714     C      1     NA 21.92745        1       23        1
1571      1   1 B2592     C      1     NA 26.35455        1       24        1
```

- **gender**: Hombre (1) o mujer (0)
- **arm**: Intervención (1) o control (0)
- **folltime**: Tiempo de observación (dias)
- **corrctgc**: Si el evento (infección) ocurrio (1) o no (0)

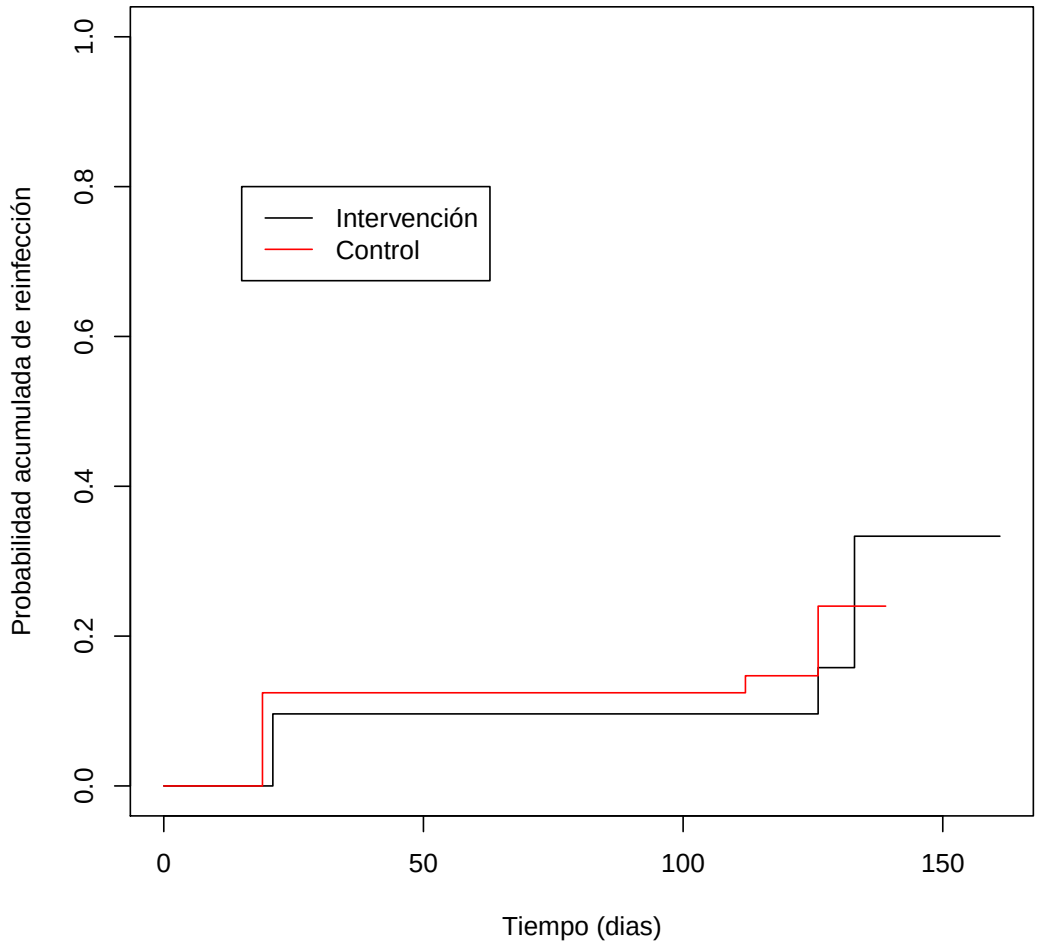


Figura : Caso 2: Estudio de Notificación de Parejas

Caso 3: Deserción universitaria

- **Motivación**: En el campo de la educación universitaria hay tres outcomes que se pueden medir:
  - Abandono forzado
  - Abandono voluntario
  - Graduación
- **Pregunta científica**: Se desea estudiar que factores están asociados con acelerar o retardar estos outcomes.
- **Población**: Estudiantes de la Pontificia Universidad Católica del Peru (PUCP).

Caso 3: Deserción universitaria

- **Muestra**: Se tiene disponible data de estudiantes que ingresaron entre los años 2002-I y 2012-I a la PUCP.
- **Respuesta**: Tiempo hasta lo que ocurra primero:
  - Abandono forzado
  - Abandono voluntario
  - Graduación
- **Características particulares**:
  - Es una variable discreta pues se mide en semestres.
  - Esta área es la que conocemos como análisis de riesgos competitivos.
- Pebes [4] analizó los factores asociados a la deserción estudiantil en la PUCP.

Estructura de los datos: Deserción

```
> head(deser)
          id time censored gender                    area                typeschool
3435    8537   15        0 female Architecture and urbanism Private / private religious
13901  19003    3        0   male                 Science Private / private religious
10788  15890   11        0   male                 Science Private / private religious
19477  24579    2        1   male                 Science Private / private religious
442     5544   16        0   male                 Science Private / private religious
10283  15385   13        0   male                 Science Private / private religious
```

- **id**: Codigo del estudiante
- **time**: Número de ciclos de observación
- **censored**: Si el estudiante abandono (1) o no (0)

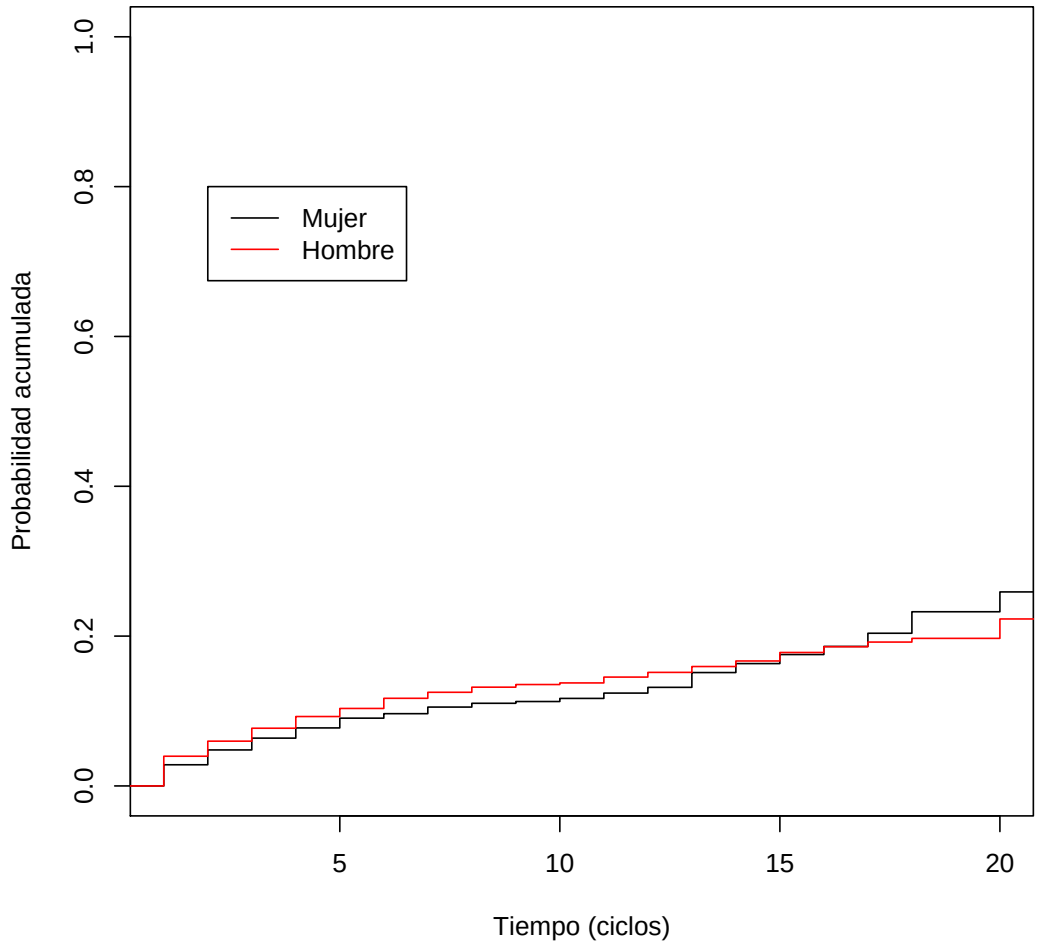


Figura : Caso 3: Deserción universitaria

Motivación

- En muchas áreas, la principal variable de respuesta es el tiempo a un determinado evento.
  - Tiempo hasta contraer una enfermedad
  - Tiempo hasta abandonar la universidad
  - Tiempo hasta terminar de pagar un préstamo
- En estos casos nuestro interés está en
  - Caracterizar la distribución del tiempo hasta que ocurra el evento para una población determinada
  - Hacer esta caracterización para dos grupos o más
  - Modelar la relación entre el tiempo a la ocurrencia de un evento y un conjunto de covariables.

Tiempo a la ocurrencia del evento

- Sea T una variable aleatoria (v.a), no negativa, que denota el tiempo a la ocurrencia de un evento de interés.
- Para evitar ambigüedad, el tiempo de inicio y el tiempo final deben ser muy bien especificados.
- Ejemplos
  - Supervivencia en general: Mide el tiempo desde el nacimiento hasta la muerte de un individuo - Estudios de esperanza de vida.
  - Tiempo de supervivencia asociado a un tratamiento para una población con determinada enfermedad: Tiempo desde el inicio del tratamiento hasta muerte.

Tiempo a la ocurrencia del evento

- En un análisis éstandar, al querer describir la variable tiempo, se presentan medidas de resumen como
  - La media, mediana
  - Desviación éstandar, rango intercuartil
- Debido a la existencia de datos censurados estas medidas de resumen pueden ser sesgadas (veremos mas adelante).
- Se necesita otra forma de presentar los datos:
  - Función de distribución
  - Función de supervivencia
  - Función de riesgo

Tiempo a la ocurrencia del evento

La distribución de T se puede describir de varias formas:

- Función acumulada de distribución:

$$F(t) = P(T \leq t) \ , \ t > 0$$

es la probabilidad de que una persona de la población seleccionada al azar **muera** antes de o en el tiempo $T = t$.

- Función de supervivencia:

$$S(t) = 1 - F(t) = P(T > t) \ , \ t > 0$$

es la probabilidad de que una persona de la población seleccionada al azar **sobreviva** hasta el tiempo $T = t$.

Propiedades

- F es una función no decreciente
- Tiende a cero por la izquierda

$$\lim_{h \to -\infty} F(h) = 0$$

- Tiende a uno por la derecha

$$\lim_{h \to \infty} F(h) = 1$$

- Si T es una variable (absolutamente) continua, entonces tiene una función de densidad, $f(\cdot)$, que se relaciona a $F(\cdot)$ via

$$f(t) = \frac{dF(t)}{dt} \quad , \quad F(t) = \int_0^t f(u)du$$

Definición

- Media del tiempo

$$\mu = E(T) = \int_0^\infty S(t)dt$$

- Mediana del tiempo

$$t_{0.5} = \inf_x \{m : S(x) \leq 0.5\}$$

Definición

- Función de riesgo

$$\lambda(t) = \lim_{h \to 0^+} \frac{P(t \leq T < t + h \mid T \geq t)}{h}$$

  es el riesgo **instantáneo** de que el evento ocurra en en el intervalo $[t, t + h]$, dado que no ha ocurrido hasta el tiempo t.

- Propiedades
  - $\lambda(t) \geq 0, \forall t$
  - $\lambda(\cdot)$ no tiene limite superior

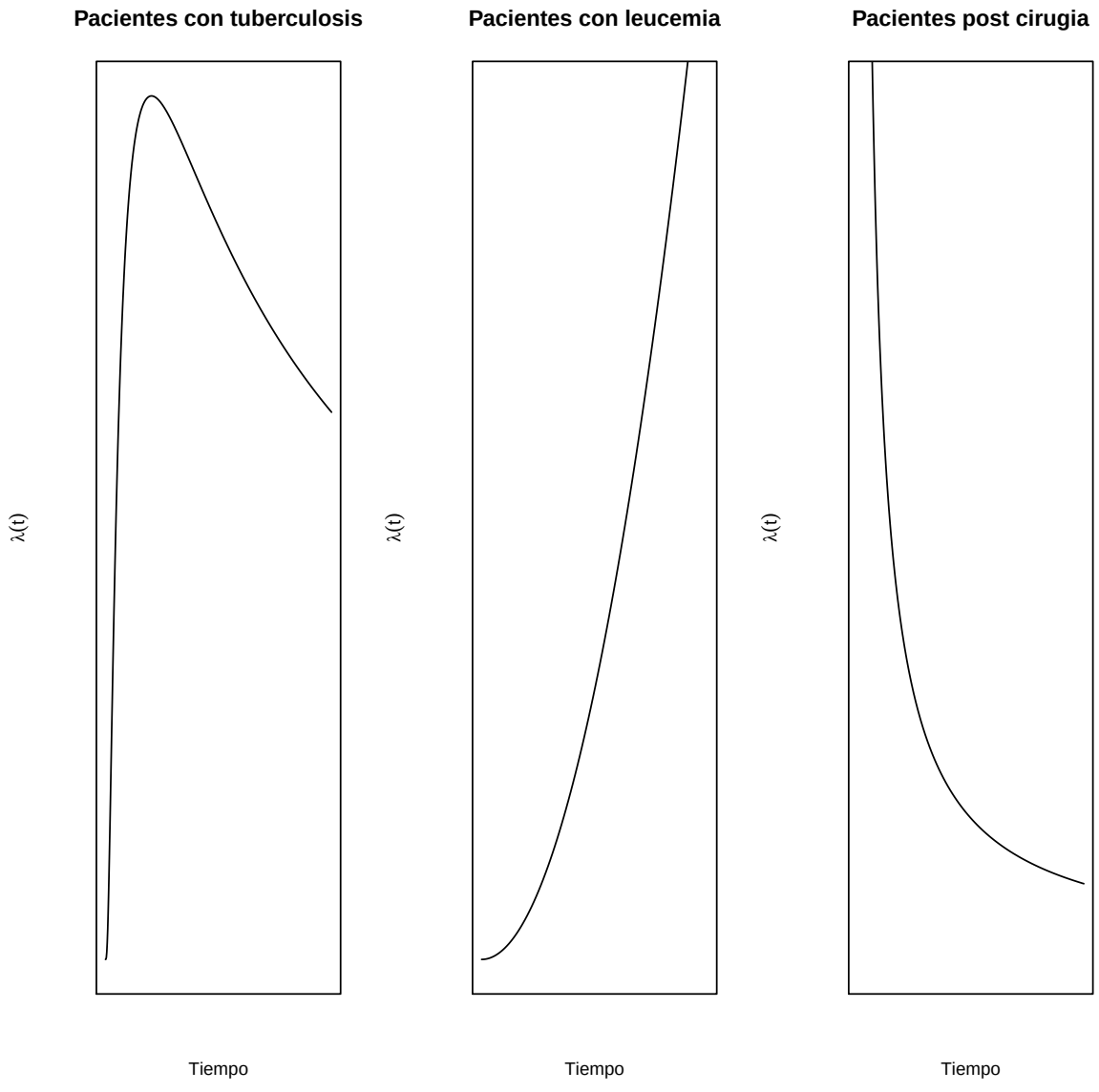


Figura : Función de riesgo instantaneo para diferentes situaciones

Definición

- Función de riesgo acumulado

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

que define el riesgo acumulado de que acurra el evento hasta el tiempo $T = t$

Definición

En el caso $f(\cdot)$ exista, las siguientes relaciones se cumplen

- $\lambda(\cdot)$ puede ser expresada como

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log\left(S(t)\right)$$

- La función de supervivencia puede ser expresada como

$$S(t) = \exp\left[-\Lambda(t)\right]$$

Modelo exponencial

Sea T una variable aleatoria continua que se comporta como un modelo exponencial con parámetro η, $T \sim Exp(\eta)$. Entonces

$$\begin{aligned} f(t) &= \eta \exp(-\eta t) \\ F(t) &= 1 - \exp(-\eta t) \\ \lambda(t) &= \eta \\ \Lambda(t) &= \eta t \end{aligned}$$

para $t > 0$

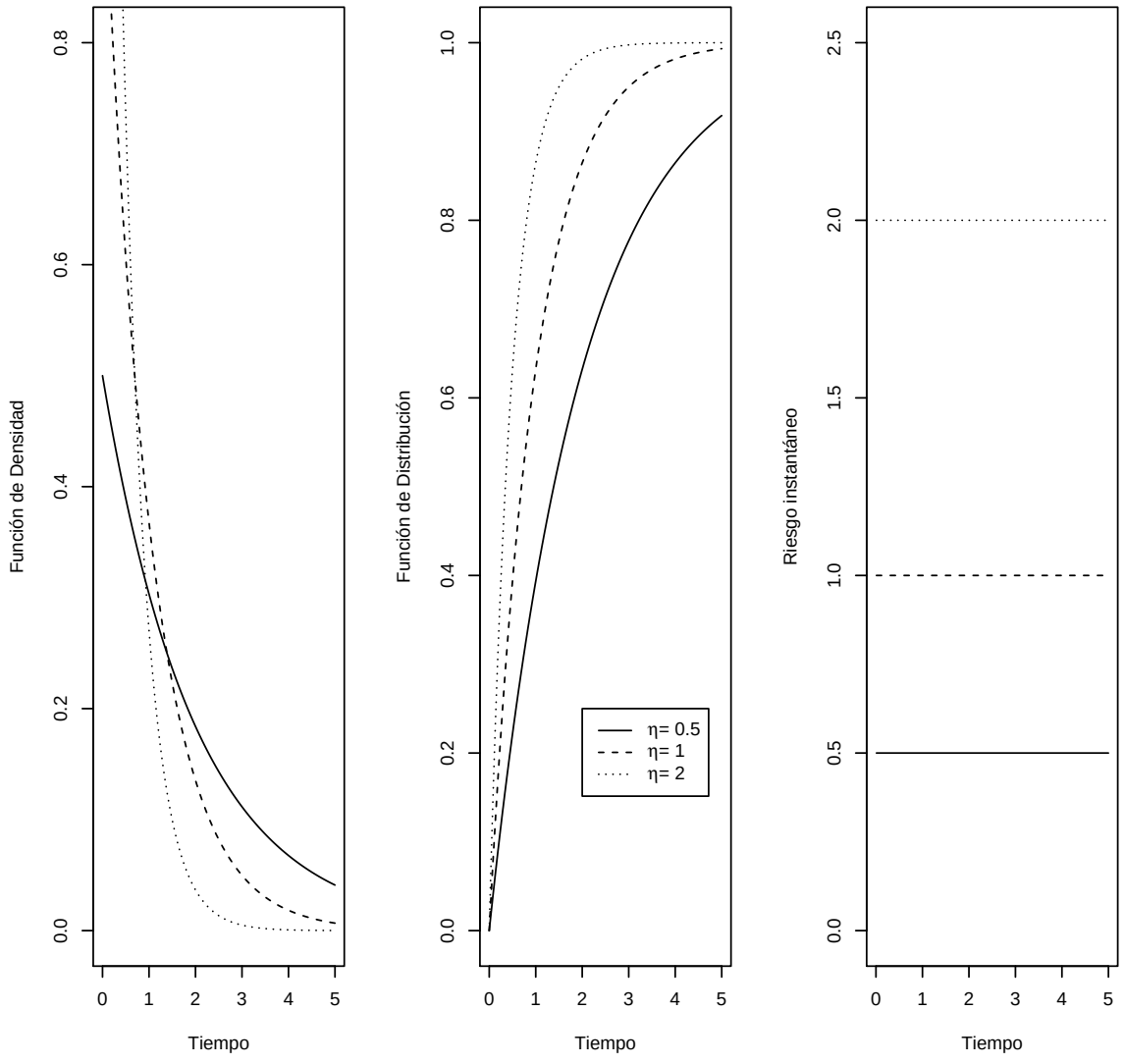


Figura : Modelo Exponencial

Modelo Weibull

Sea T una variable aleatoria continua que se comporta como un modelo Weibull con parámetros de forma y escala a y b, $T \sim Weibull(a, b)$, entonces

$$\begin{aligned} f(t) &= \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \exp\left[-(t/b)^a\right] \\ F(t) &= 1 - \exp\left[-(t/b)^a\right] \\ \lambda(t) &= \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \\ \Lambda(t) &= \left(\frac{t}{b}\right)^a \end{aligned}$$

donde $t > 0$

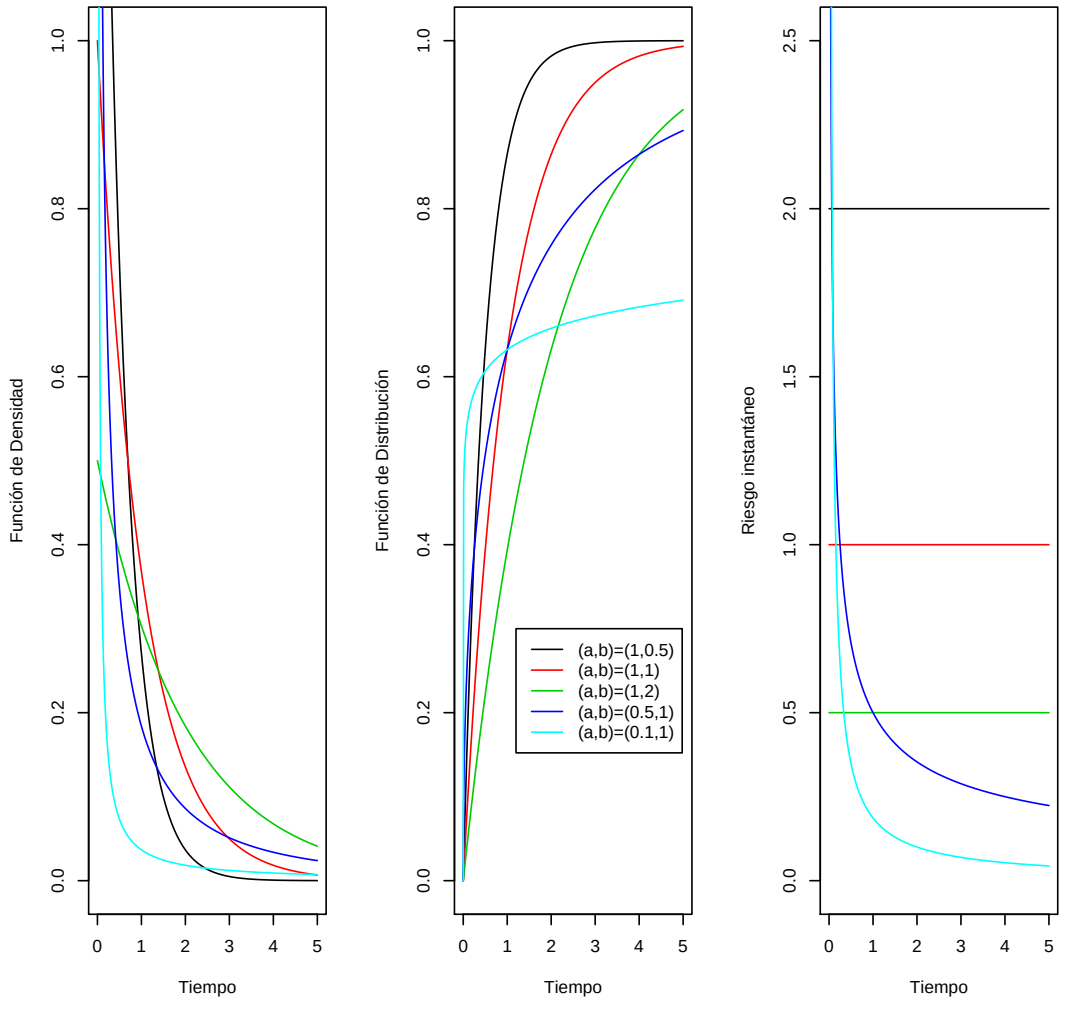


Figura : Modelo Weibull

Tiempos discretos

Si T es una v.a. discreta que toma los valores $a_1 < a_2 < \cdots$

- La función de probabilidad asociada es

$$f(a_i) = P(T = a_i) \quad , \quad i = 1, 2, \ldots$$

- La función acumulada de probabilidad es

$$F(t) = P(T \leq t) = \sum_{j|a_j \leq t} f(a_j)$$

- La función de supervivencia es

$$S(t) = P(T > t) = \sum_{j|a_j > t} f(a_j)$$

Tiempos discretos

- La función de riesgo esta dada por

$$\lambda_i = P(T = a_i \mid T \geq a_i) = \frac{f(a_i)}{S(a_i^-)} = 1 - \frac{S(a_i)}{S(a_{i-1})}$$

- La función de supervivencia se puede caracterizar como

$$S(t) = \prod_{j|a_j \leq t} (1 - \lambda_j)$$

Interpretación: La probabilidad de sobrevivir hasta el instante $T = a_i$ es igual a la probabilidad de que el evento no ocurra en los instantes $a_1, a_2, \ldots, a_i$.

Tiempos discretos

- La función de probabilidad se puede caracterizar por

$$f(a_i) = \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j)$$

Interpretación: La probabilidad de que ocurra el evento en el instante a_i es igual a la probabilidad de que ocurra en a_i (λ_i) y no ocurra en $a_{i-1}, a_{i-2}, \ldots, a_2, a_1$

Ejemplo

- Los siguientes datos son una muestra aleatoria de los tiempos (en meses) hasta la ocurrencia de un evento:

$$14,2 \quad 7,3 \quad 21,6 \quad 0,5 \quad 13,5 \quad 3,0 \quad 9,7 \quad 7,4 \quad 3,1 \quad 1,9$$

- Ojbetivo: Construir un modelo para describir este conjunto de datos
- Modelos potenciales: Exponencial, Weibull, no paramétrico

Estimación: Modelo exponencial

- Supongamos que observamos una muestra

$$T_1, T_2, \ldots, T_n$$

de tiempos que son v.a. independientes e identicamente distribuidas.

- Si estas v.a. siguen un modelo **exponencial**, la probabilidad de observar esta muestra es

$$\begin{aligned} P(T = t_1, \ldots, T = t_n) &= \prod_{i=1}^{n} P(T = t_i) = \prod_{i=1}^{n} f(t_i) \\ &= \prod_{i=1}^{n} \eta \exp\left[-\eta t_i\right] = \eta^n \exp\left[-\eta \sum_{i=1}^{n} t_i\right] \end{aligned}$$

Estimación: Modelo exponencial

- La verosimilitud está dada por

$$L_n(\eta) = \eta^n \exp\left[-\eta \sum_{i=1}^{n} t_i\right]$$

y su logaritmo

$$l_n(\eta) = \log\left(L_n(\eta)\right) = n\log\left(\eta\right) - \eta \sum_{i=1}^{n} t_i \qquad (1)$$

- El estimador de máxima verosimilitud se define como el valor que máximiza (1), es decir

$$\hat{\eta} = \arg_\eta \max l_n(\eta)$$

Estimación: Modelo exponencial

- En este caso en particular tiene una solución explícita

$$\hat{\eta} = \frac{n}{\sum_{i=1}^{n} t_i}$$

- Los estimadores de $F(\cdot)$, $S(\cdot)$ y $\lambda(\cdot)$ estan dados por

$$\begin{aligned} \hat{F}_n(t) &= 1 - e^{-\hat{\eta}t} \\ \hat{S}_n(t) &= e^{-\hat{\eta}t} \\ \hat{\lambda}_n(t) &= \frac{n}{\sum_{i=1}^{n} t_i} \end{aligned}$$

para $t > 0$

Estimación: Modelo exponencial

- La matriz de información, $I(\eta)$ está dada por:

$$I(\eta) = -E\left(\frac{\partial^2 L(\eta)}{\eta^2}\right) = \frac{n}{\hat{\eta}^2}$$

- El error estándar es la raíz de la inversa de la matriz de información:

$$SE(\hat{\eta}) = \sqrt{I(\eta)^{-1}} = \frac{\hat{\eta}}{\sqrt{n}}$$

Estimación: Modelo exponencial

- El estimador de máxima verosimilitud de η es

$$\hat{\eta}_n = 0,12$$

- El error estándar de $\hat{\eta}$ es

$$SE(\hat{\eta}) = 0,04$$

- La probabilidad estimada de sobrevivir 5 meses es

$$\hat{S}_n(5) = \exp\left[-0,12 \times 5\right] = 0,55$$

- La probabilidad estimada de sobrevivir 10 meses es

$$\hat{S}_n(10) = \exp\left[-0,12 \times 10\right] = 0,30$$

Estimación: Modelo exponencial

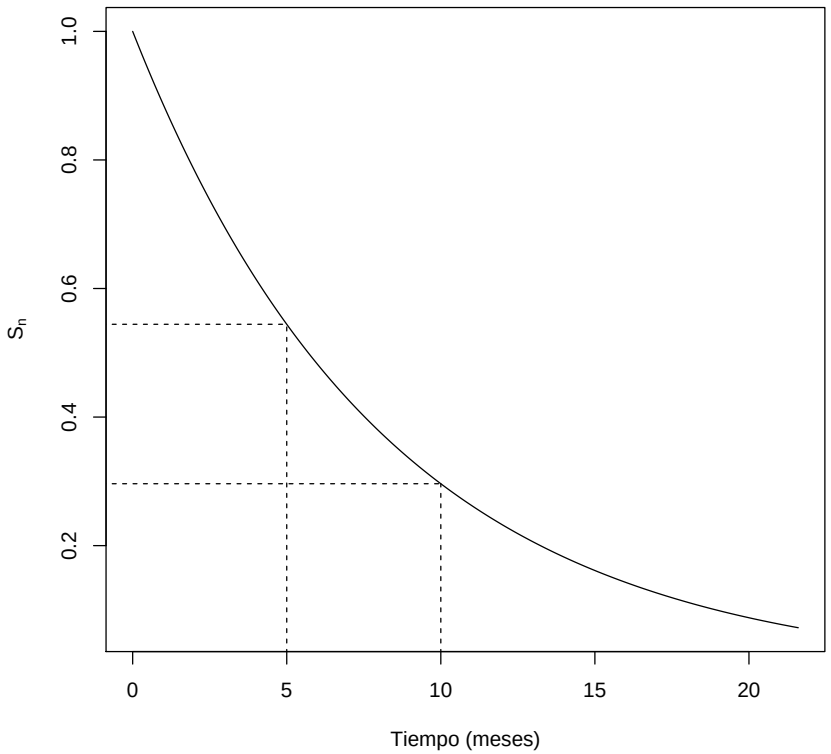


Figura : Función estimada de supervivencia

Estimación: Modelo Weibull

- Supongamos que observamos una muestra

$$T_1, T_2, \ldots, T_n$$

de tiempos que son v.a. independientes e identicamente distribuidas.

- Si estas v.a. siguen un modelo Weibull, la probabilidad de observar esta muestra es

$$\begin{aligned} P(T = t_1, \ldots, T = t_n) &= \prod_{i=1}^{n} P(T = t_i) = \prod_{i=1}^{n} f(t_i) \\ &= \prod_{i=1}^{n} \frac{a}{b} \left(\frac{t_i}{b}\right)^{a-1} \exp\left[-(t_i/b)^a\right] \end{aligned}$$

Estimación: Modelo Weibull

- La verosimilitud está dada por

$$L_n(a,b) = \left(\frac{a}{b}\right)^n \left(\prod_{i=1}^{n} t_i\right)^{a-1} \left(\frac{1}{b}\right)^{na-n} \exp\left[-\frac{1}{b^a}\sum_{i=1}^{n} t_i^a\right]$$

y su logaritmo por

$$\begin{aligned} l_n(a,b) &= n[\log(a) - \log(b)] + (a-1)\sum_{i=1}^{n}\log(t_i) \\ &\quad -n(a-1)\log(b) - \frac{1}{b^a}\sum_{i=1}^{n} t_i^a \end{aligned}$$

- En este caso, el estimador de máxima verosimilitud se debe calcular de manera numérica.

Estimación: Modelo Weibull

```
> # Weibull model
> lweibull <- function(x,t)
+ {
+   n   <- length(t)
+   a   <- x[1]
+   b   <- x[2]
+   res <- n*log(a) - n*log(b) + (a-1)*sum(log(t)) - sum(t^a)/(b^a) - (n*a-n)*log(b)
+   return(-res)
+ }
> mwei     <- nlminb(c(1,1),lweibull,t=t)
> mwei
$par
[1] 1.221615 8.756190

> ee <- solve(hessian(lweibull,x=mwei$par,t=t))
> ee
           [,1]       [,2]
[1,] 0.09880791  0.2306766
[2,] 0.23067658  5.6761494
```

Estimación: Modelo Weibull

- Usando R, el estimador de máxima verosimilitud de a y b son

$$\hat{a}_n = 1.22 \; y \; \hat{b}_n = 8.76$$

- La matriz de varianza es

$$SE(\hat{a}_n, \hat{b}_n) = \begin{pmatrix} 0.1 & 0.23 \\ 0.23 & 5.68 \end{pmatrix}$$

es decir la desviación éstandar de $\hat{a}_n$ y $\hat{b}_n$ es 0.31 y 2.38, respectivamente.

Estimación: Modelo Weibull

- La probabilidad estimada de sobrevivir 5 meses es

$$\hat{S}_n(5) = \exp\left[-\left(\frac{5}{8.76}\right)^{1.22}\right] = 0.60$$

- La probabilidad estimada de sobrevivir 10 meses es

$$\hat{S}_n(10) = \exp\left[-\left(\frac{10}{8.76}\right)^{1.22}\right] = 0.31$$

Estimación: Modelo weibull

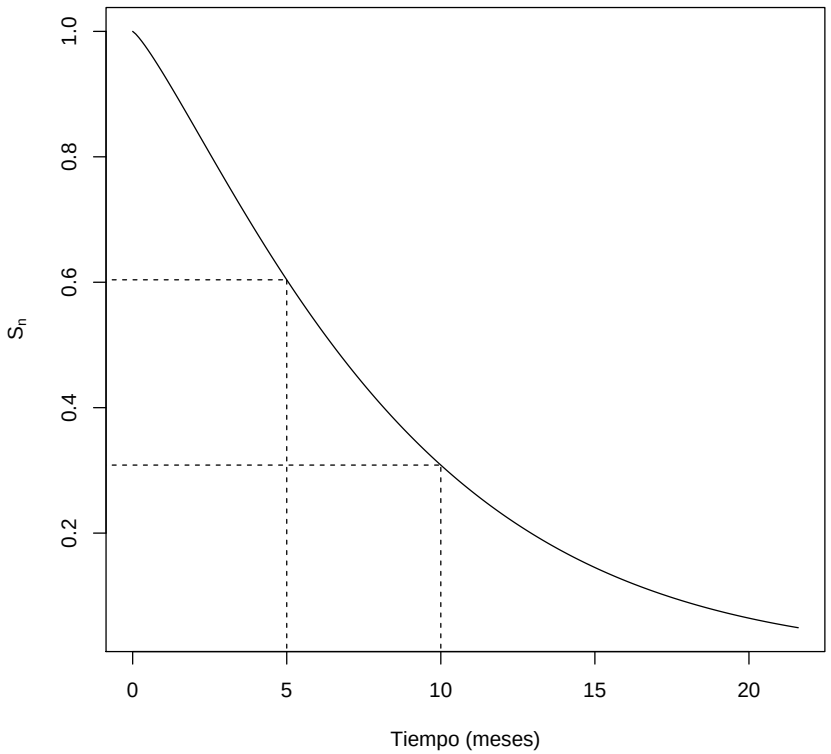


Figura : Función estimada de supervivencia

Estimación: No paramétrica

- ¿ Qué opción tenemos si queremos ser flexibles y deseamos no asumir un modelo exponencial ?
- Sin información adicional, el mejor estimador de $F(\cdot)$ tiene la forma

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t) = \frac{\#\ observaciones \leq t}{n}$$

donde

$$I(T_i \leq t) = \begin{cases} 1 & T_i \leq t \\ 0 & T_i > t \end{cases}$$

es decir $\hat{F}_n(t)$ es el promedio del número de observaciones, en la muestra, que son menores o iguales a t.

Estimación: No paramétrica

- Este estimador asigna masa $\frac{1}{n}$ a cada observación
- La probabilidad estimada de sobrevivir 5 meses es

$$\hat{S}_n(5) = 1 - \hat{F}_n(5) = 1 - \frac{4}{10} = 0,6$$

- La probabilidad estimada de sobrevivir 10 meses es

$$\hat{S}_n(10) = 1 - \hat{F}_n(10) = 1 - \frac{7}{10} = 0,3$$

Estimación: No paramétrica

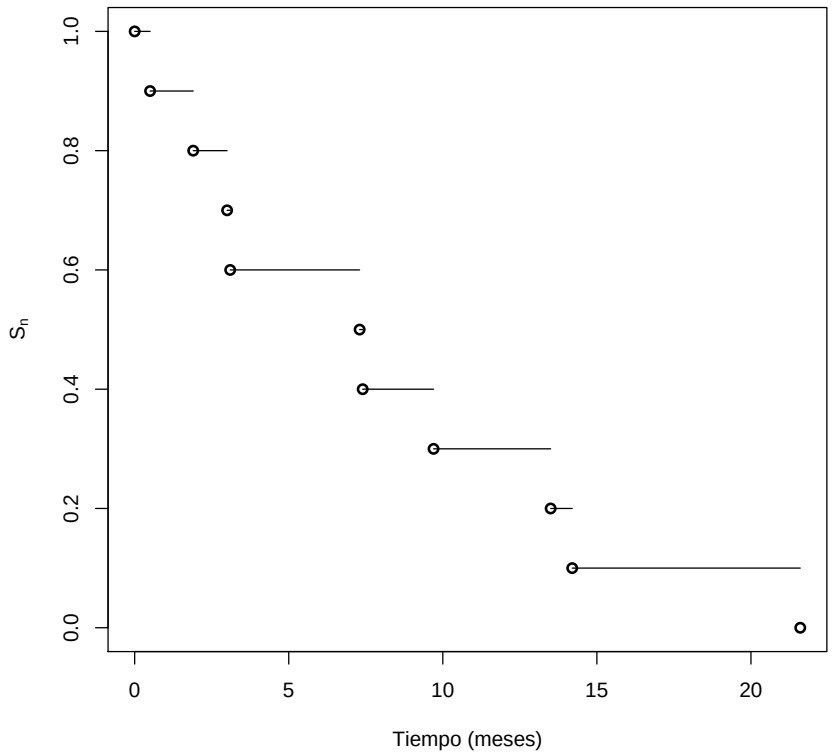


Figura : Función acumulada de probabilidad estimada

Estimación: No paramétrica vs. paramétrica

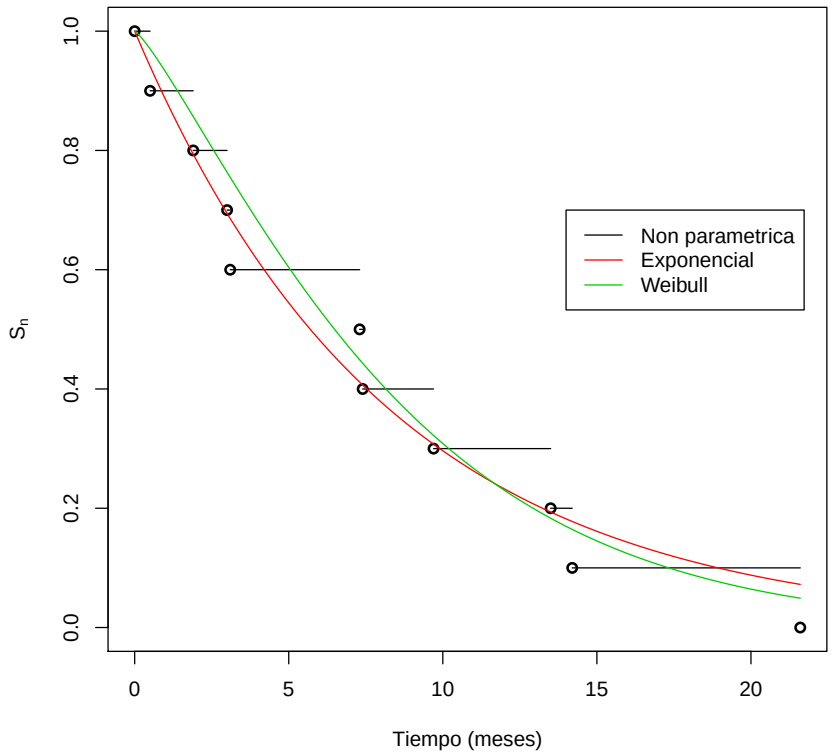


Figura : Función acumulada de probabilidad estimada en base a tres diferentes modelos

Datos censurados

- Usualmente, cuando los datos son recolectados durante un periodo de tiempo determinado, puede pasar que el tiempo a la ocurrencia del evento no sea observado para algunos individuos.
  - Este resultado es lo que se conoce como dato censurado
- Se pueden presentar tres tipos de datos censurados:
  - Por la derecha: Solo se conoce que el evento no ha ocurrido hasta el tiempo t.
  - Por la izquierda: Se conoce que el evento ha ocurrido previo a un tiempo t.
  - Intervalo: Se conoce que el evento ocurrio entre dos puntos en el tiempo $[t_1, t_2]$

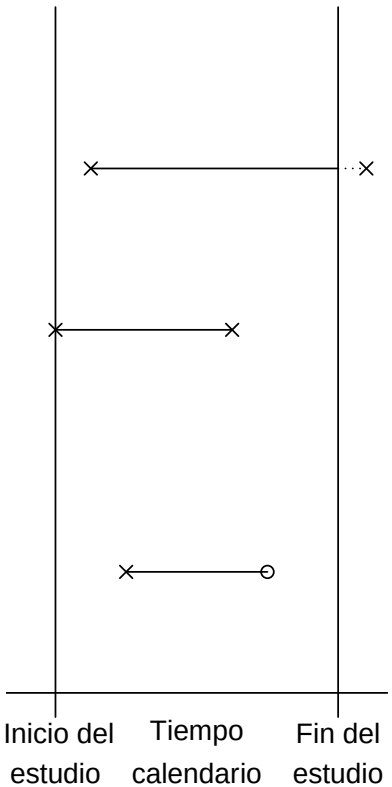
Datos censurados: Derecha

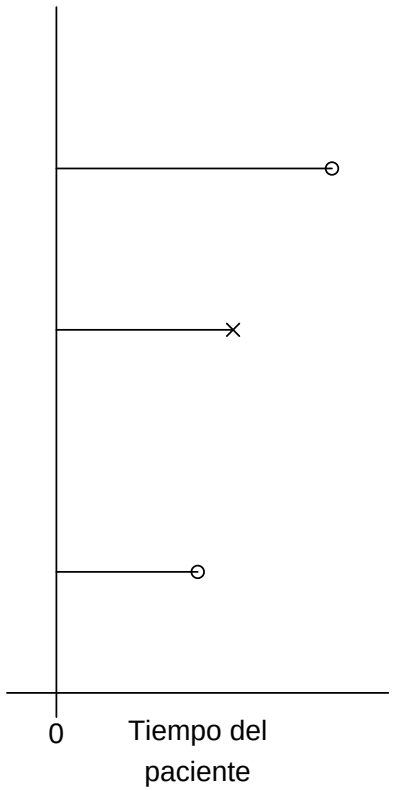
- Datos del participante con codigo 203000045

```
> head(hptn)
       ptid  t1  t2 event arm
1 203000045  NA  NA     0   1
2 203000045   0  90     0   1
3 203000045  90 175     0   1
4 203000045 175 266     0   1
5 203000045 266 357     0   1
```

- Nuestra información es que si la infección ocurre sera en algun momento despues del dia 357

Datos censurados: Derecha





Datos censurados: Intervalo

- Datos del participante con codigo 628001747

```
> hptn[hptn$ptid=="628001747",]
             ptid  t1  t2 event arm
10028 628001747  NA  NA     0   0
10029 628001747   0  90     0   0
10030 628001747  90 180     1   0
```

- La infección por VIH, para este participante, se dio en el intervalo (90,180], pero no sabemos en que dia exactamente.

Datos censurados

- Datos censurados por la derecha
  - El participante del estudio se mudó de ciudad.
  - El paciente falleció de una causa ajena a al enfermedad estudiada.
  - El estudio terminó y el evento de interés no se observó.
- Datos censurados por la izquierda
  - Tiempo hasta el primer resfrío de un bebe
- Datos censurados por intervalo
  - Se registra anualmente si una persona se ha contagiado (o no) de determinada enfermedad.

Libreria survival

- Esta data es un estudio del tiempo de vida de 100 personas con cancer de pulmon

```
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0

> Surv(lung$time,lung$status)[1:10]
 [1]  306   455  1010+  210   883  1022+  310   361   218   166
```

- El primer y segundo paciente murieron a los 306 y 455 dias del diagnóstico, respectivamente
- El tercer paciente se mantuvo con vida hasta el ultimo dia de observación (1010)

Libreria survival

```
> head(hptn,n=8)
         ptid  t1  t2 event arm
2   203000045   0  90     0   1
3   203000045  90 175     0   1
4   203000045 175 266     0   1
5   203000045 266 357     0   1
7   203000053   0  86     0   1
8   203000053  86 175     0   1
9   203000053 175 265     0   1
10  203000053 265 358     0   1

> Surv(time = hptn$t1,time2 = hptn$t2,event = hptn$event)[1:10]
 [1] (  0, 90+] ( 90,175+] (175,266+] (266,357+] (  0, 86+] ( 86,175+] (175,265+] (265,358+]
 [9] (  0, 88+] ( 88,182+]
```

- Esta data es del estudio HPTN 039 que busca prevenir la infección de VIH
- En todos los intervalos observados tenemos data censurada

Referencias I

[1] C. Celum, A. Wald, J. Hughes, J. Sanchez, S. Reid, S. Delany-Moretlwe, F. Cowan, M. Casapia, A. Ortiz, J. Fuchs, S. Buchbinder, B. Koblin, S. Zwerski, S. Rose, J. Wang, and L. Corey. Effect of aciclovir on hiv-1 acquisition in herpes simplex virus 2 seropositive women and men who have sex with men: a randomised, double-blind, placebo-controlled trial. *Lancet*, 371(9630): 2109–2119, 2008.

[2] M. R. Golden, W. L. Whittington, H. H. Handsfield, J. P. Hughes, W. E. Stamm, M. Hogben, A. Clark, C. Malinski, J. R. Helmers, K. K. Thomas, and K. K. Holmes. Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine*, 352(7):676–685, 2005.

Referencias II

[3] V. Sal y Rosas and J. Hughes. Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. *Stat Commun Infect Dis*, 2010 (364), 2010.