Algunas aplicaciones en bioestadística

Luis E. Nieto Barajas

Departamento de Estadística
ITAM

UV – 20 de octubre de 2017

Contenido

Definiciones

- Bioestadística : (Wikipedia) Es la aplicación de la Estadística a la Biología. Comprende el diseño de experimentos biológicos, la recolección, resumen y análisis de los datos obtenidos de los experimentos ; e interpretación e inferencia de los resultados.

Definiciones

- Bioestadística : (Wikipedia) Es la aplicación de la Estadística a la Biología. Comprende el diseño de experimentos biológicos, la recolección, resumen y análisis de los datos obtenidos de los experimentos ; e interpretación e inferencia de los resultados.
- Ensayos clínicos : Experimentos de investigación biomédicos o de salud en humanos. Estos estudios siguen un determinado protocolo. Hay de dos tipos : intervencionales u observacionales. Para nuevos tratamientos hay 4 Fases :
  - Fase I : Identificar una dosis segura (toxicidades)
  - Fase II : Determinar la eficacia de cierta dosis
  - Fase III : Comparar tratamiento nuevo vs. control
  - Fase IV : Monitorear tratamiento (efectos secundarios)

Ideas generales

- Teoría Clásica : $T \sim F$
  - Paramétrica : $F \in \mathcal{F}_\Theta$

$$\mathcal{F}_\Theta = \{F : \ F = F_\theta, \ \theta \in \Theta\}$$

  - No paramétrica : $\dim(\theta) = \infty$, $F \in \mathcal{F}$

$$\mathcal{F} = \{F : \ F \text{ es una funcion de distribucion}\}$$

Ideas generales

- Teoría Clásica : $T \sim F$
  - Paramétrica : $F \in \mathcal{F}_\Theta$

$$\mathcal{F}_\Theta = \{F : \ F = F_\theta, \ \theta \in \Theta\}$$

  - No paramétrica : $\dim(\theta) = \infty$, $F \in \mathcal{F}$

$$\mathcal{F} = \{F : \ F \text{ es una funcion de distribucion}\}$$

- Teoria Bayesiana : Asignar una distribucion inicial a $\mathcal{F}$
  - Paramétrica : $\theta \sim \pi$, $\theta \in \Theta \Rightarrow$

$$\mathcal{F}_\Theta \sim \mathcal{P}$$

  - No paramétrica :

$$\mathcal{F} \sim \mathcal{P}$$

ITAM 70 AÑOS

Ideas generales

- Inferencia Bayesiana no paramétrica :

$$T_1, \ldots, T_n \text{ m.a.}$$

$$T_i | F \sim F \text{ definida en } (\mathbb{R}, \mathcal{B})$$

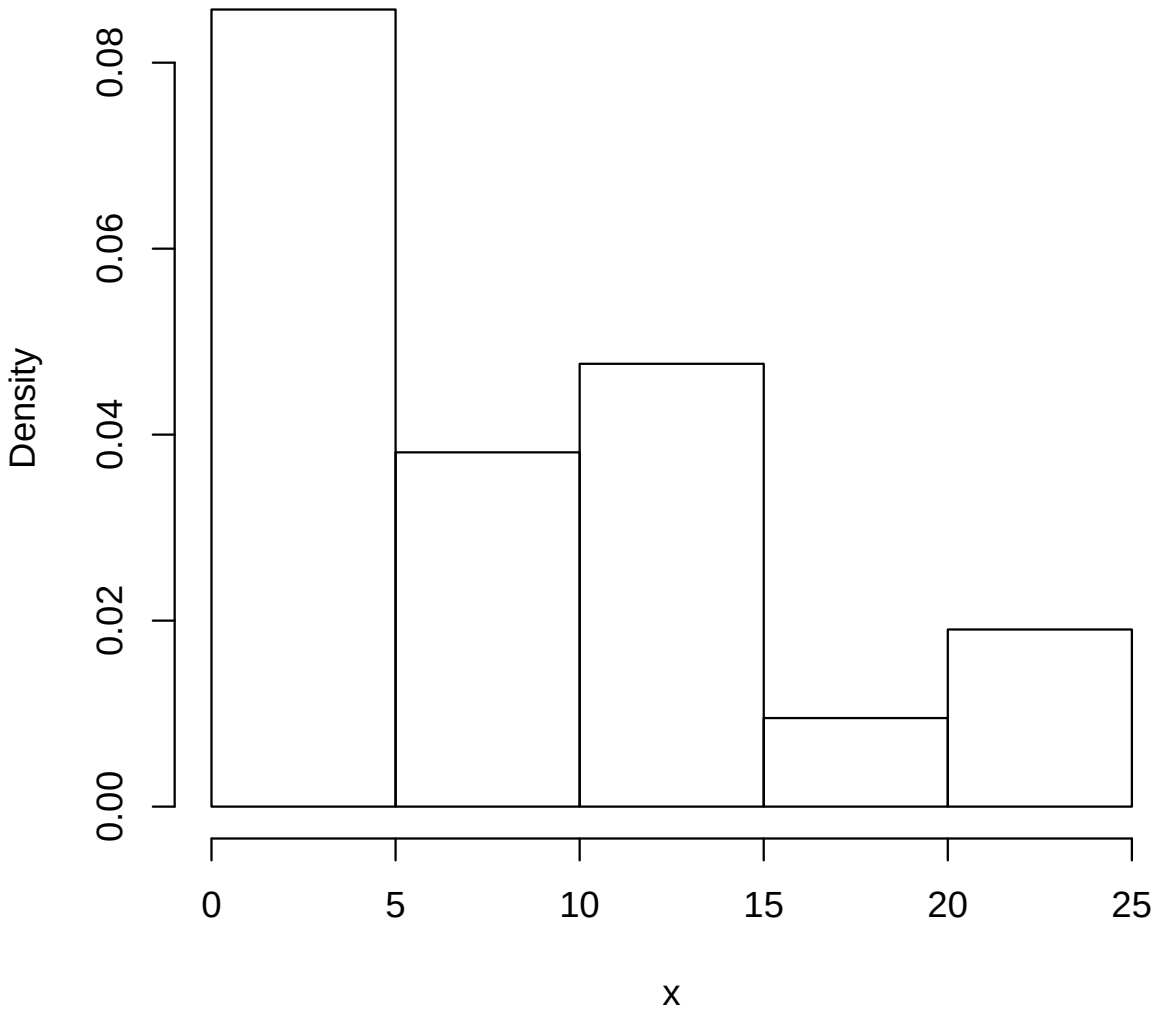
$$F \sim \mathcal{P} \text{ definida en } (\mathcal{F}, \mathcal{A})$$

Estudio

- $T =$ Tiempo de remisión (en semanas) en pacientes con leucemia
  Eventos : Inicio=remisión, fin=recaída
- Datos : (Freireich et al., 1963) $n = 21$ observaciones (0% censurados)

$$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$$

Estudio



Modelo

- Supongamos T discreta con puntos de masa $\{\tau_1, \tau_2, \ldots\}$
- Densidad : $f(\tau_j) = \mathrm{P}(T = \tau_j) = p_j$, $j = 1 \ldots, n$
  - $\Rightarrow \Theta = \{p_1, p_2, \ldots\}$ (espacio parametral desc.)
  - Inf. Bayesiana : definir dist. inicial π sobre Θ
- En A. Supervivencia es de mayor interés estudiar las funciones de riesgo y de supervivencia

$$h_j = \mathrm{P}(T = \tau_j | T \geq \tau_j), \; j = 1, 2, \ldots$$

$$S(t) = \prod_{\{j:\tau_j \leq t\}} (1 - h_j)$$

Modelo

- Sea $\Theta' = \{h_1, h_2, \ldots\}$. Como $\Theta' \iff \Theta$ entonces una inicial sobre Θ' induce inicial sobre Θ
- Inicial 1 : (Hjort, 1990). Como $h_j \in (0,1) \Rightarrow h_j \overset{\text{iid}}{\sim} \text{Be}(\alpha_j, \beta_j)$
- Inicial 2 : (Nieto-Barajas & Walker, 2002). Modelan $\{h_j\}$ mediante un proceso dependiente tal que

$$h_1 \overset{u_1}{\to} h_2 \overset{u_2}{\to} \cdots .$$

En particular, para $j = 1, 2, \ldots$

$$h_1 \sim \text{Be}(\alpha_1, \beta_1)$$

$$u_j | h_j \sim \text{Bi}(c_j, h_j)$$

$$h_{j+1} | u_j \sim \text{Be}(\alpha_{j+1} + u_j, \beta_{j+1} + c_j - u_j)$$

Modelo

- Inicial 2 : Propiedades
- Si $c_j = 0\ \forall j$ entonces h_j's son independientes
- Si $\alpha_j = \alpha$ y $\beta_j = \beta\ \forall j$, entonces
  - $\{h_j\}$ es un proceso estacionario con $h_j \sim \text{Be}(\alpha, \beta)$
  - $\text{Corr}(h_j, h_{j+1}) = \frac{c_j}{\alpha+\beta+c_j}$
- Ejemplo leucemia : especificaciones vagas en localización y dispersión con fuerte correlación inicial $\alpha_j = 0.0001$, $\beta_j = 0.01$, $c_j = 50$, $\forall j = 1, 2, \ldots$

ITAM 70 AÑOS

Resultados

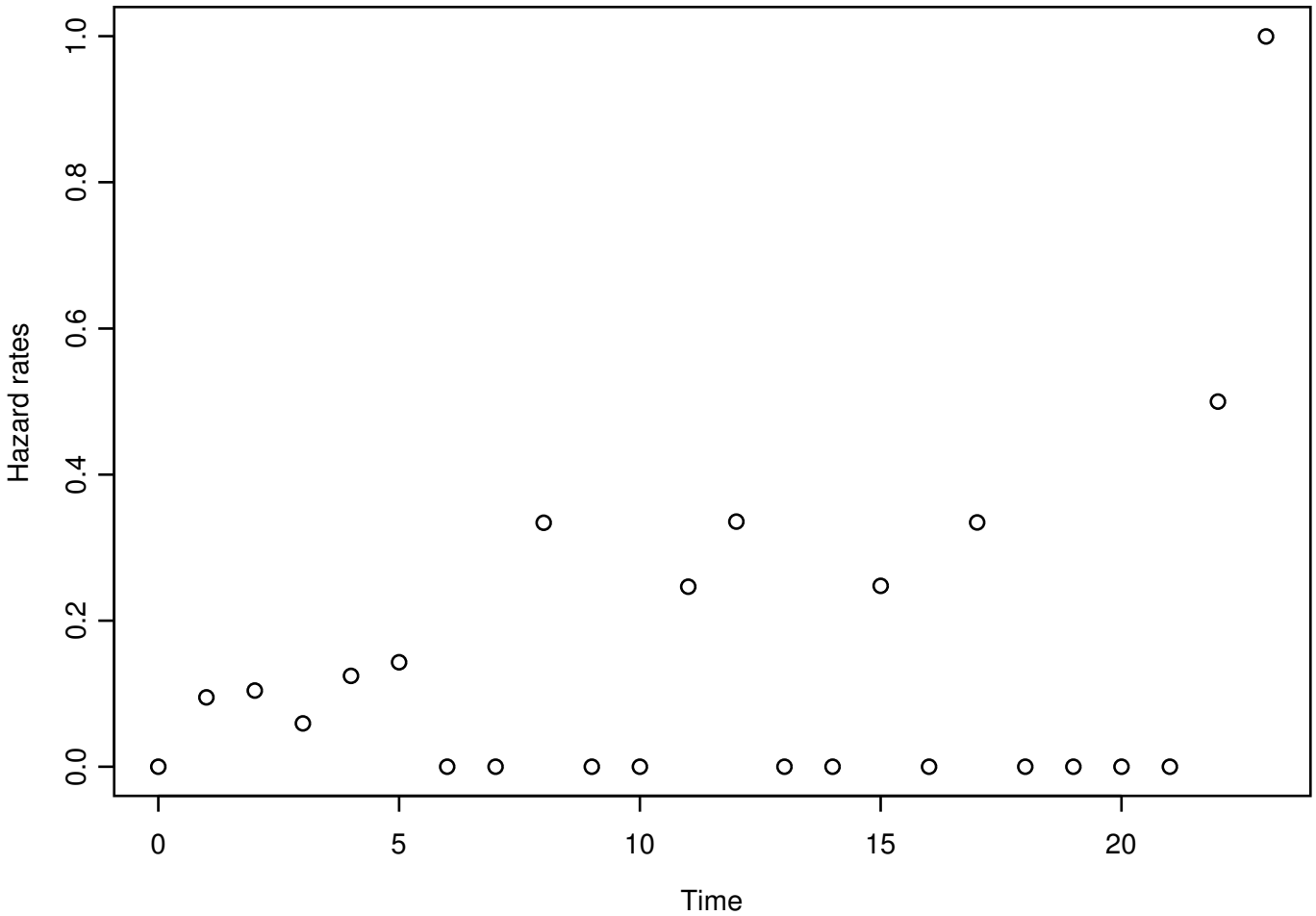


FIGURE : Estimadores Nelson-Aalen

Resultados

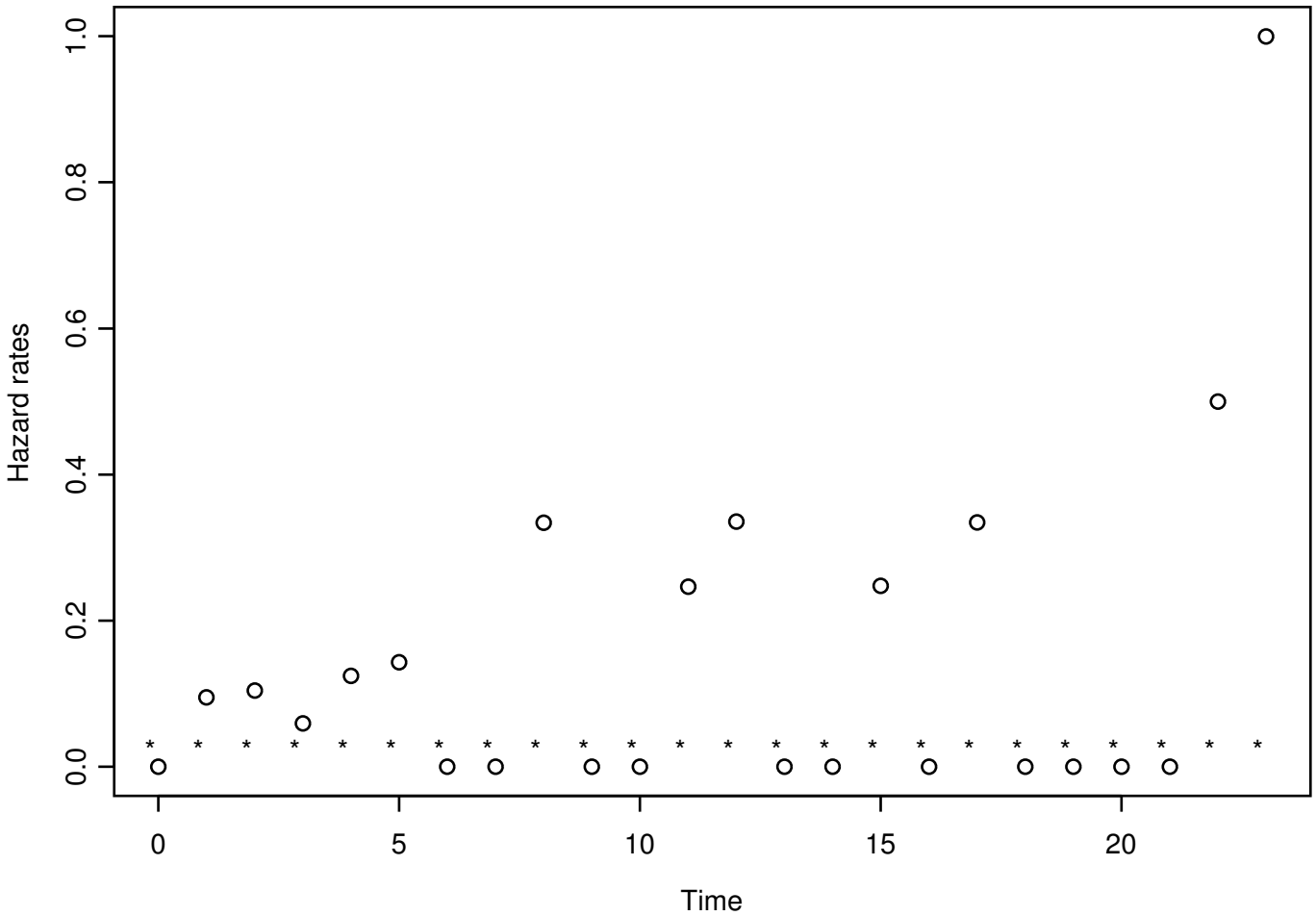


FIGURE : Estimadores : Nelson-Aalen + iniciales

Resultados

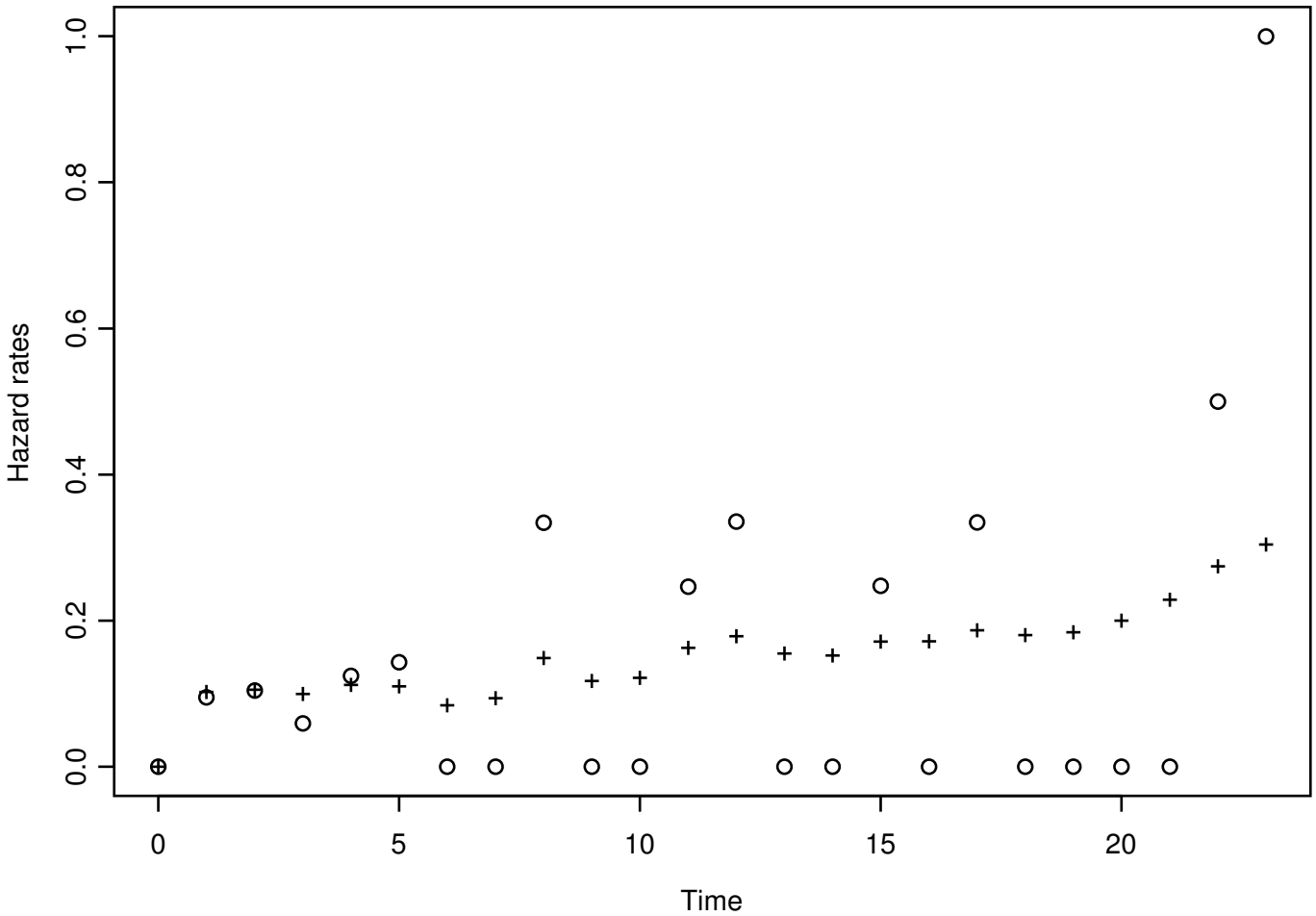


FIGURE : Estimadores : Nelson-Aalen + posteriores

Resultados

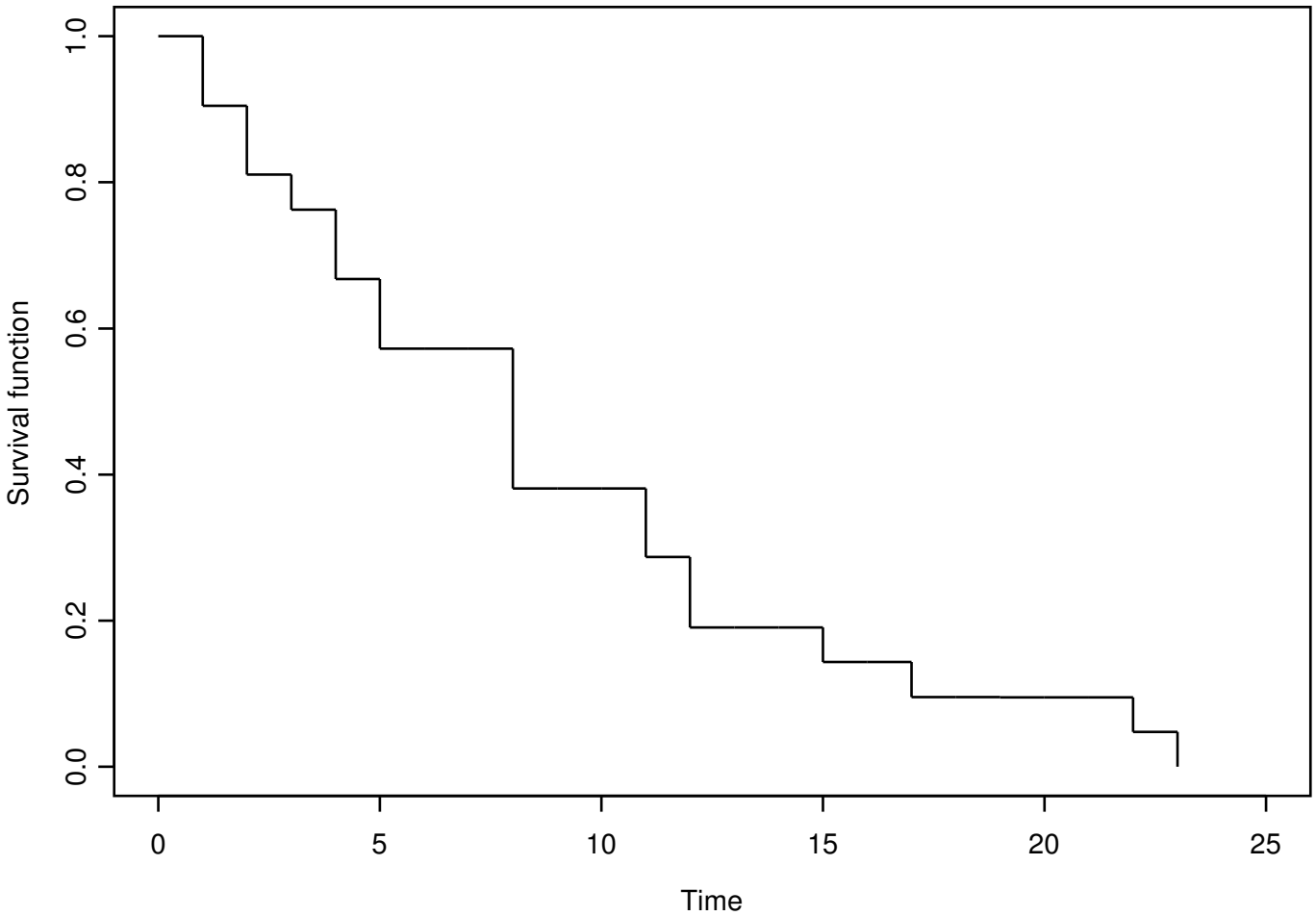


FIGURE : Estimador Kaplan-Meier

Resultados

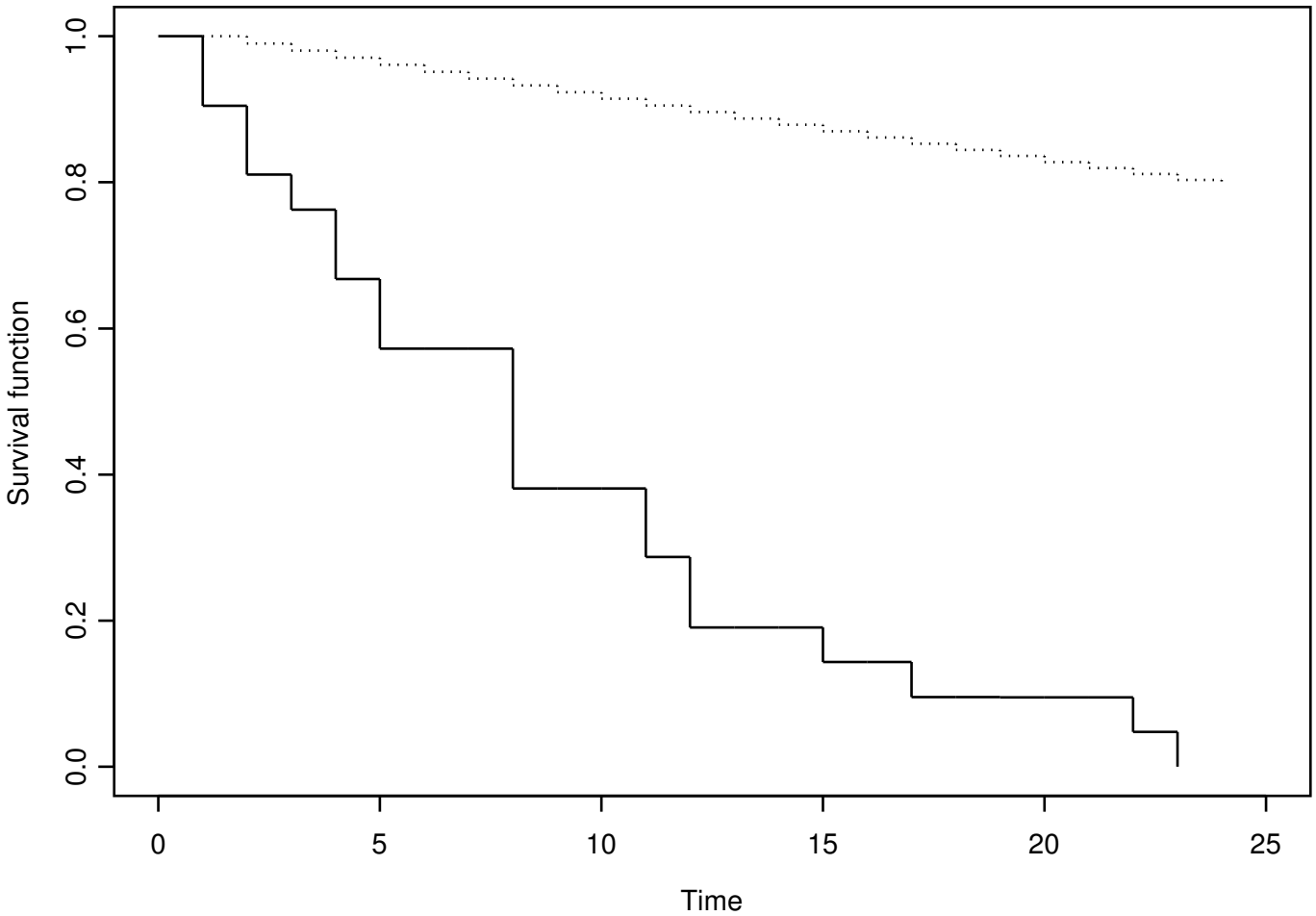


FIGURE : Estimadores : Kaplan-Meier + iniciales

Resultados

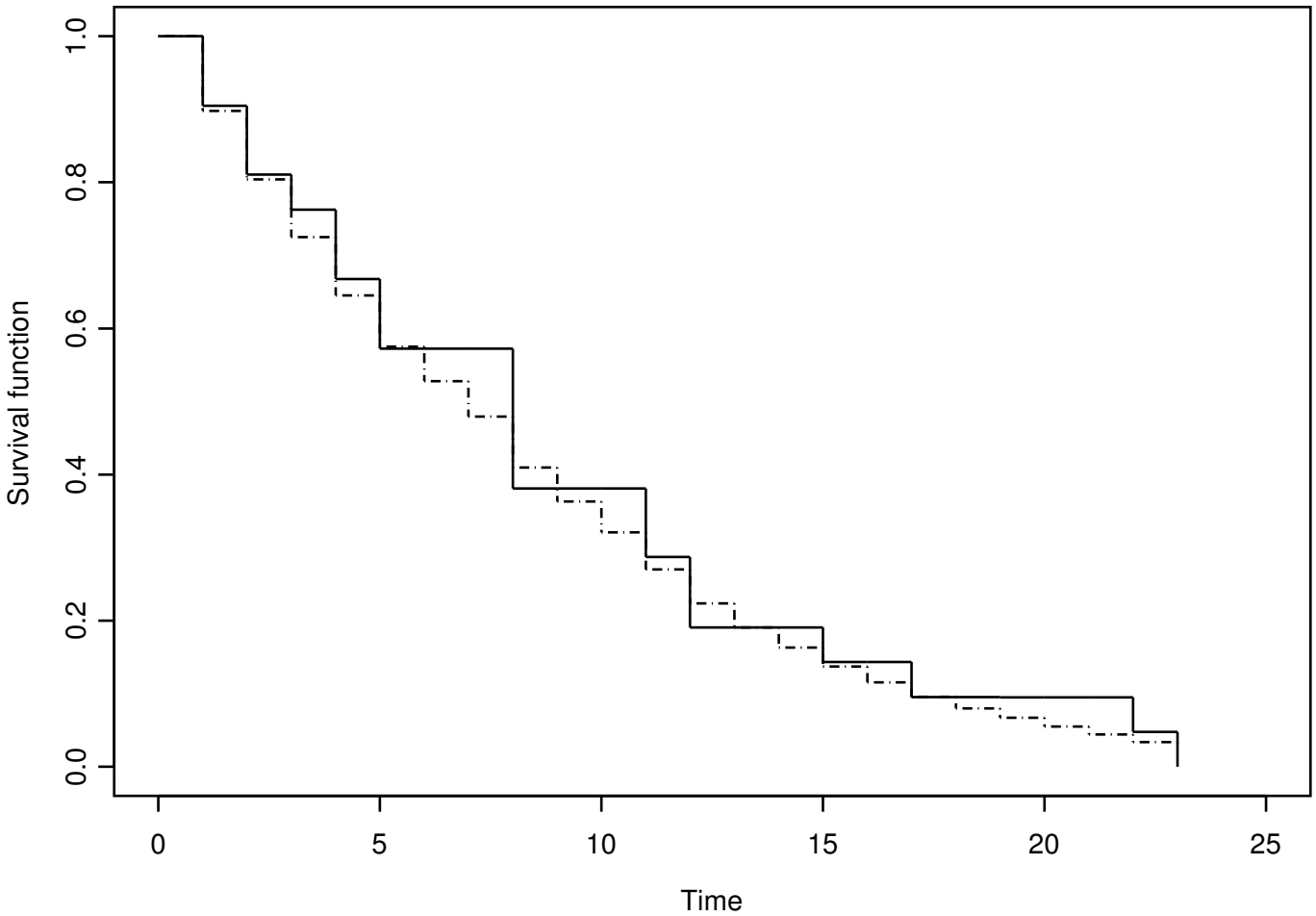


FIGURE : Estimadores : Kaplan-Meier + finales

Estudio

- Estudio : Los pacientes eran aceptados cuando eran juzgados candidatos para transplante. Cuando un donador se presentaba, los médicos seleccionaban el candidato idóneo
- $T =$ Tiempo de supervivencia (días) de un paciente con problemas de corazón
- Datos : (Cox & Oakes, 1984) $n = 249$ pacientes aceptados, de los cuales sólo 184 recibieron transplante (74%)

Tiempo espera	*Indicador transplante*	*Superviv. total*	*Status final*
49	0	49	1
112	0	112	0
50	1	673	1
.	.	.	.
22	1	3716	0

Estudio

- Existe información de covariables $\mathbf{W}_i(t) \Rightarrow$ mod. regresión
- Cada individuo tiene$(T_i, \mathbf{W}_i)$, i.e., $f_i(t)$, $F_i(t)$, $h_i(t)$ y $S_i(t)$
- Modelo de riesgos proporcionales : (Cox, 1972)

$$h_i(t) = h_0(t) e^{\mathbf{W}_i(t)' \boldsymbol{\theta}}$$

  $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_p)$ vector de coeficientes
  $h_0(t)$ función de riesgo base (común)

- Inf. Bayesiana : Definir iniciales sobre $\boldsymbol{\theta}$ y $h_0(t)$
- Como $h_0(t)$ es cualquier función de riesgo $\Rightarrow$ inicial NP

Modelo

- Para una v.a. T continua :

$$f(t) = h(t)S(t), \quad S(t) = e^{-\int_0^t h(s)ds}$$

donde $h(t)$ es una función de riesgo :
  - No negativa
  - El área bajo su curva es ∞

$\Rightarrow$ Dist. inicial NP sobre $h(t)$: a través de un proceso estocástico

- Inicial : (Nieto-Barajas & Walker, 2004). Basada en procesos de Lévy (mezcla)

$$h(t) = \int_{\mathbb{R}} k(t,s)\mathrm{d}L(s)$$

Modelo

- $L(t) =$ Proceso de Lévy o proceso aditivo creciente (PAC)
- PAC = proceso de saltos puros

$$L(t) = \sum_j V_j I(\mu_j \leq t)$$

  V_j y μ_j son alturas y localizaciones aleatorias
- $L(t)$ se caracteriza por la medida de Lévy $N_t(\nu)$
- Dependiendo del kernel $k(t, s)$ se pueden obtener distintas formas del proceso mezcla

$$\int k(t, s) \mathrm{d}L(s) = \sum_j V_j k(t, \mu_j)$$

Modelo

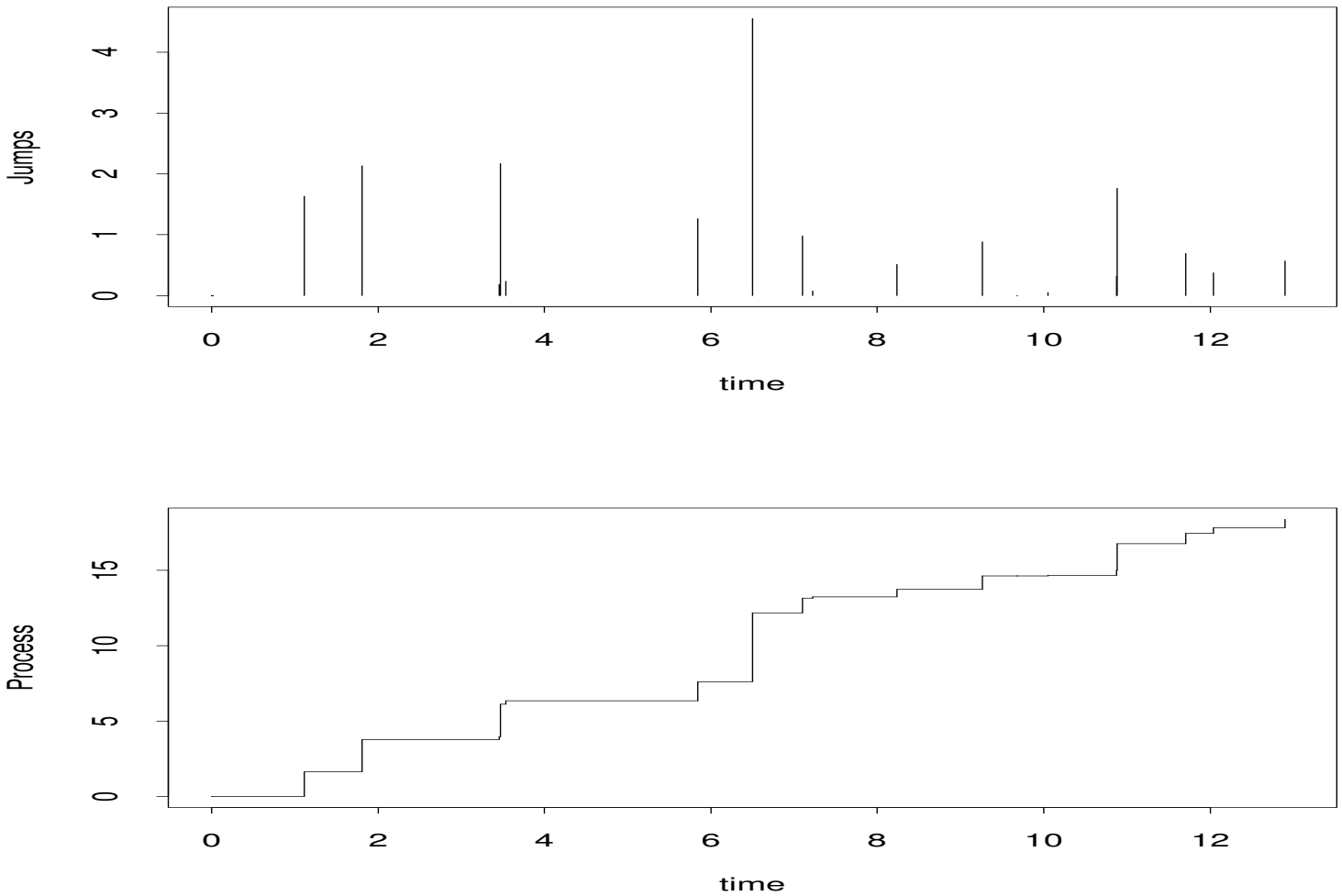


FIGURE : Ejemplo de PAC $L(t)$

Modelo

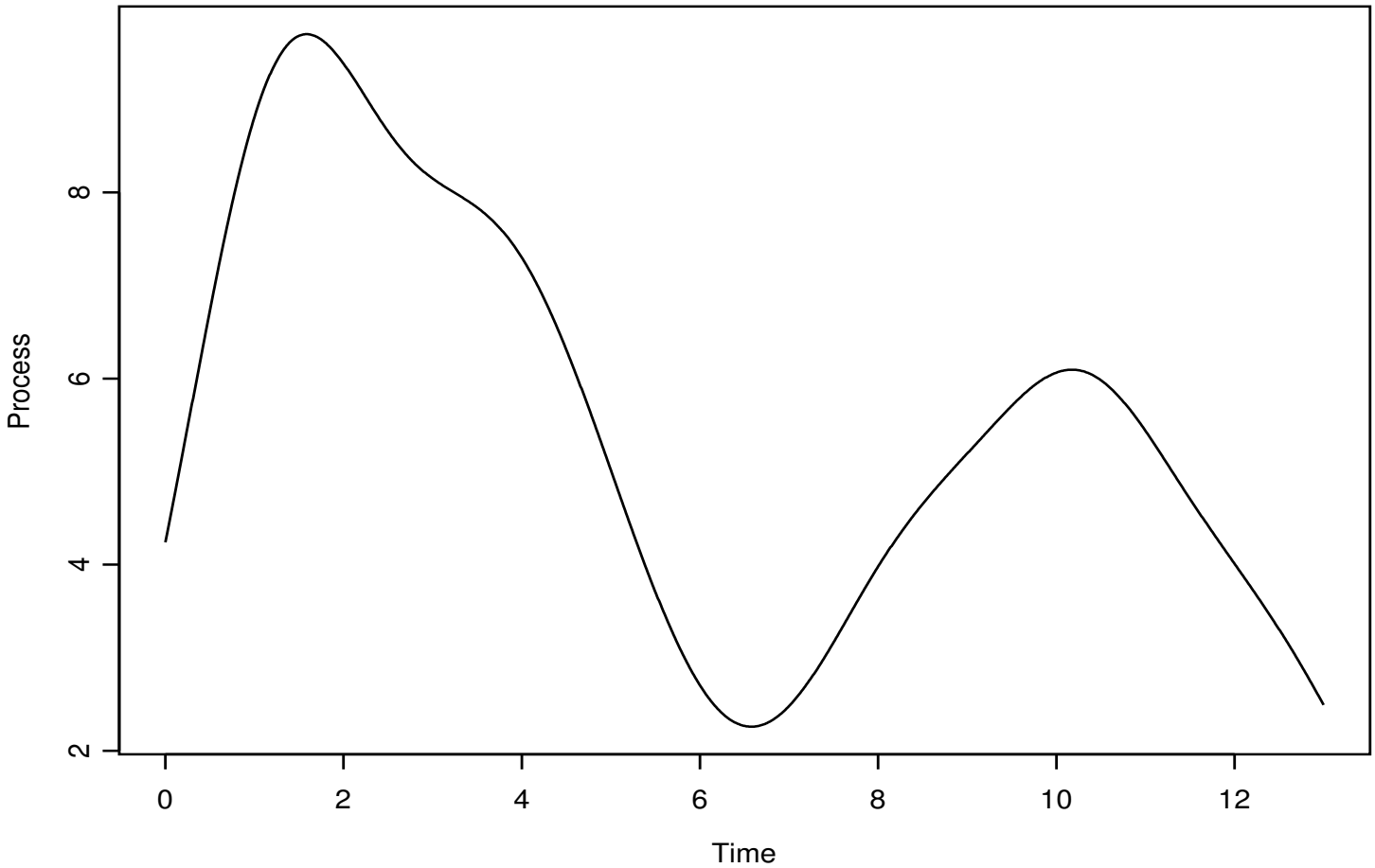


FIGURE : Ejemplo de mezcla con PAC $\int k(t, s)dL(s)$

Modelo

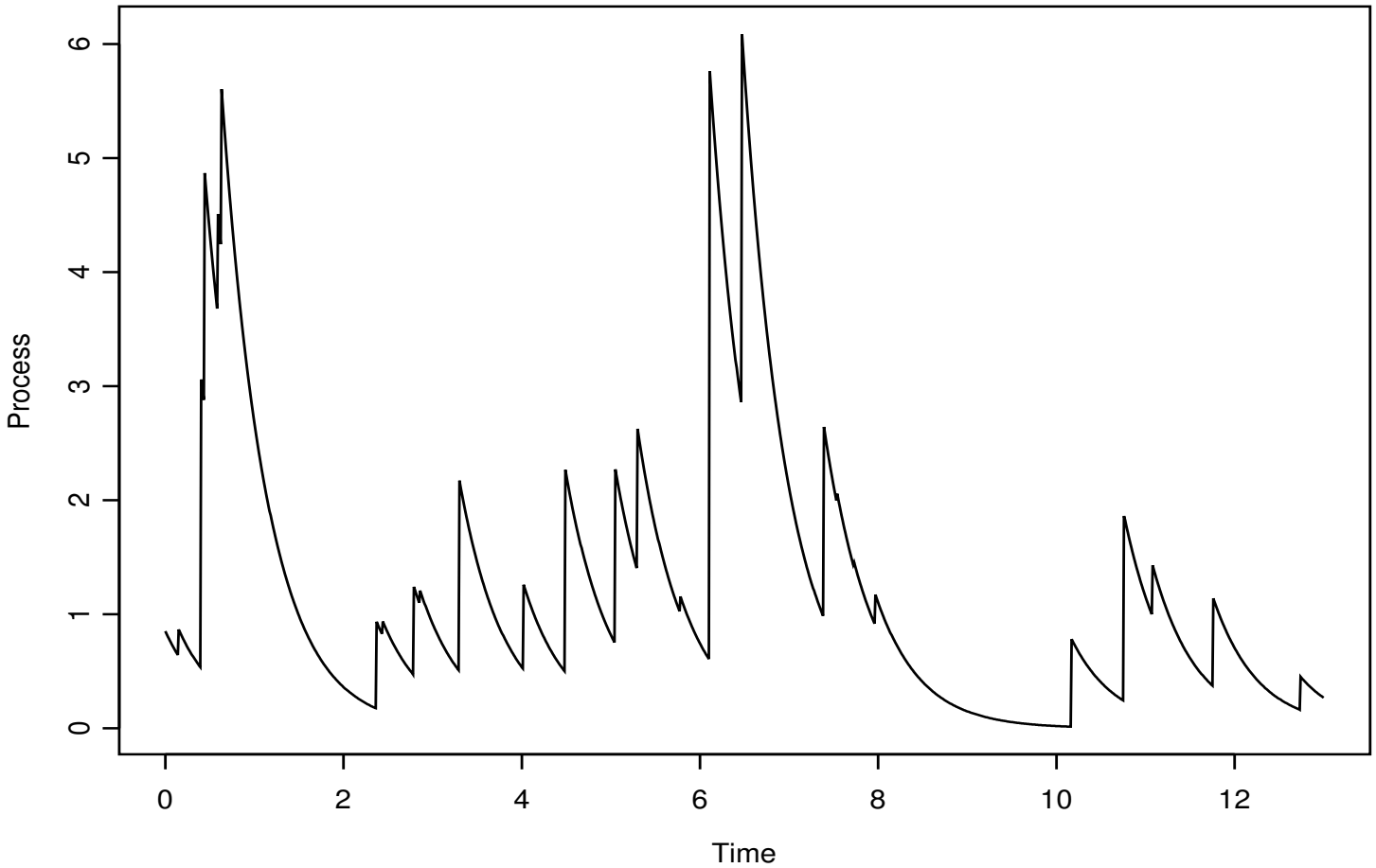


FIGURE : Ejemplo de mezcla con PAC $\int k(t, s)dL(s)$

Modelo

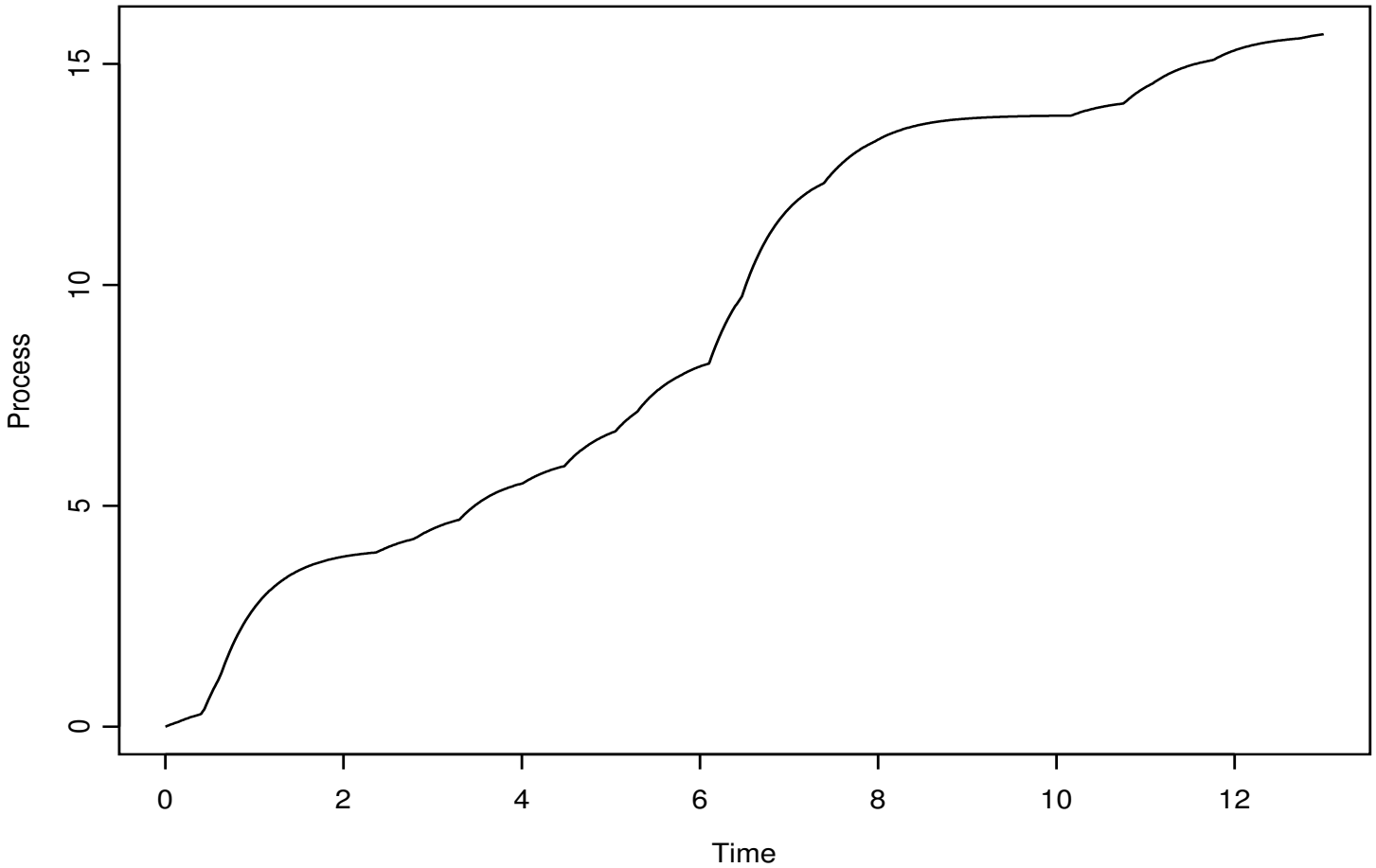


FIGURE : Ejemplo de mezcla con PAC $\int k(t, s)dL(s)$

Modelo

En particular, la inicial para $h_0(t)$ es de la forma

$$h_0(t) = \int k(t,s)\mathrm{d}L(s)$$

- $\mathrm{d}N_s(\nu) = \mathrm{d}\nu \int_0^s \exp\{-\nu\beta(u)\}\mathrm{d}\alpha(u)$
- $k(t,s) = e^{-a(t-s)}I(s \leq t)$
- Para el ejemplo : $W_i(t) = I(t \geq w_i)$
  - $\alpha(\cdot)$ and $\beta(\cdot)$ t.q. $\mathrm{E}\{h_0(t)\} = 0.02$ and $\mathrm{Var}\{h_0(t)\} = 20$
  - $a \sim \mathrm{Ga}(1/2, 1/2)$
  - $\theta \sim N(0, 9)$
- Resultados : $\hat{\theta} = -0.88 \Rightarrow$ se reduce el riesgo de morir en casi 60% si se realiza el transplante de corazón

Resultados

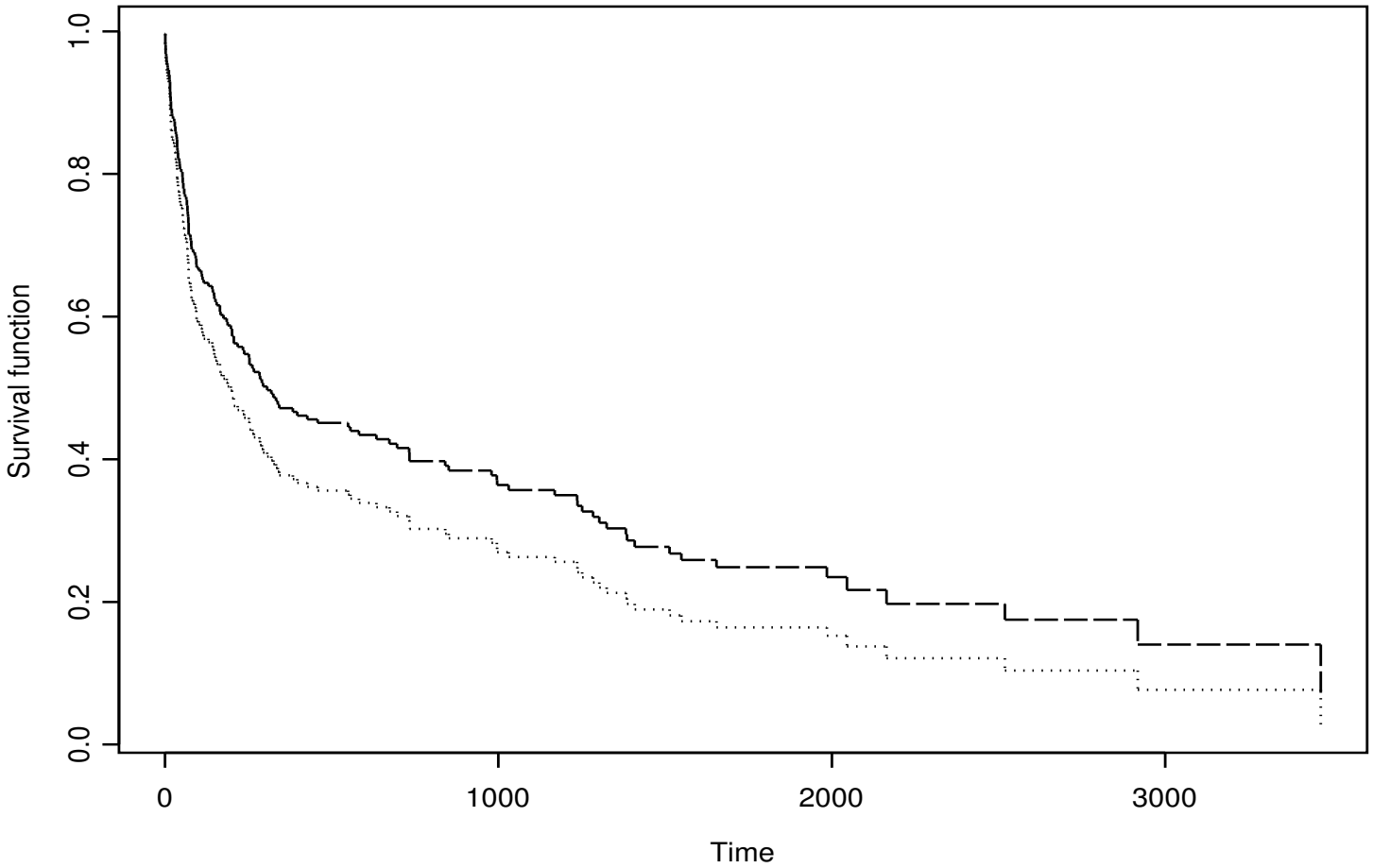


FIGURE : Estimadores de Breslow ($w_i = \infty$ & $w_i = 0$)

Resultados

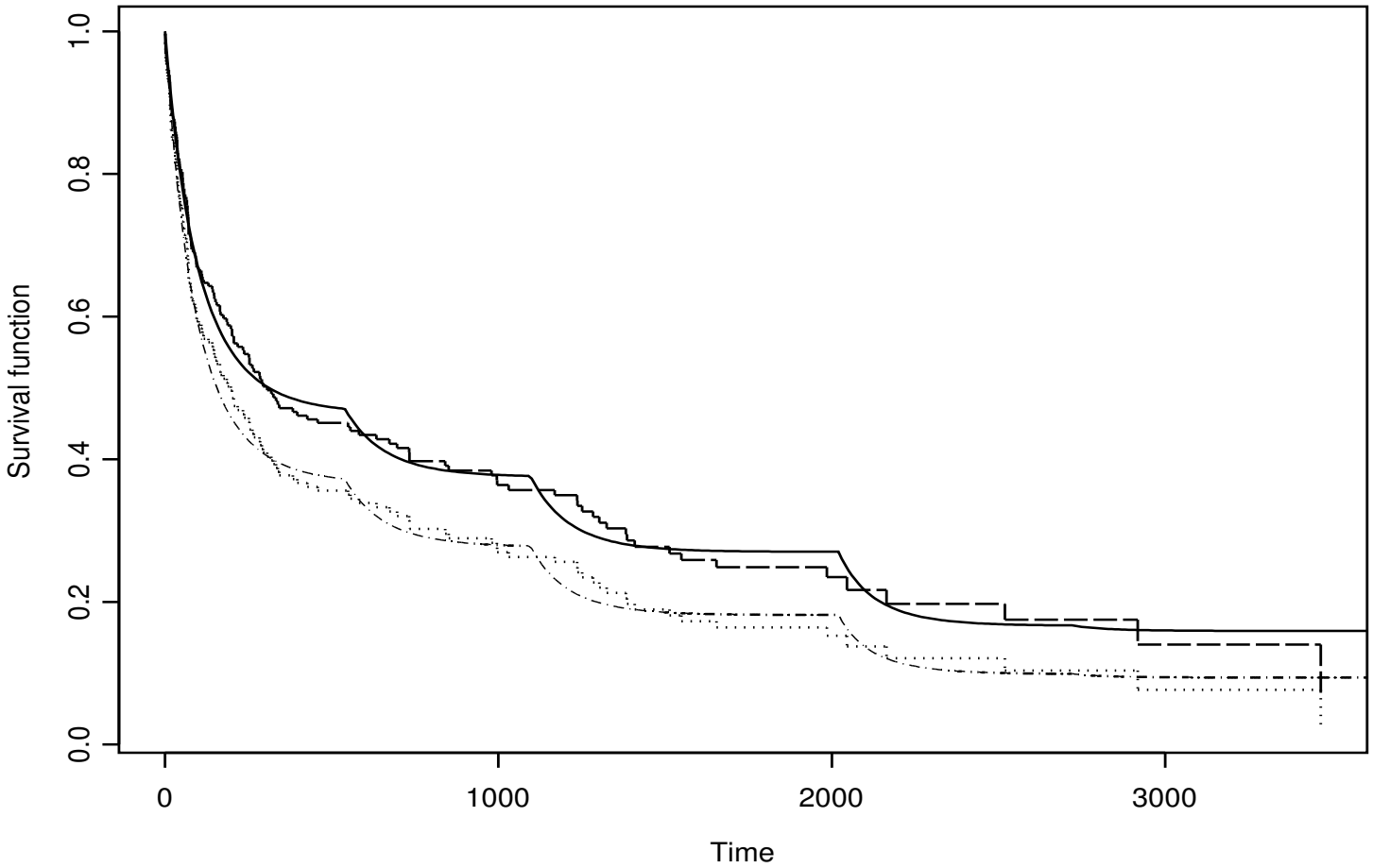


FIGURE : Estimadores de Breslow + estimadores posteriores ($w_i = \infty$ & $w_i = 0$)

Estudio

- Estudio : Pacientes con leucemia eran sometidos a dos tipos de tratamiento BMT
  - Alogénico (infusión de médula de otra persona compatible)
  - Autogénico (reinfusión de la médula del mismo paciente previa extracción y limpieza)
- $T =$ Tiempo de supervivencia (días) libre de leucemia

  Eventos : Inicio=remisión, fin=recaída o muerte
- Datos : $n = 43$ pacientes, 16- T.Alogénico y 27-T.Autogénico
- Objetivo : Evaluar la diferencia en la supervivencia libre de leucemia de los dos tipos de BMT

Estudio

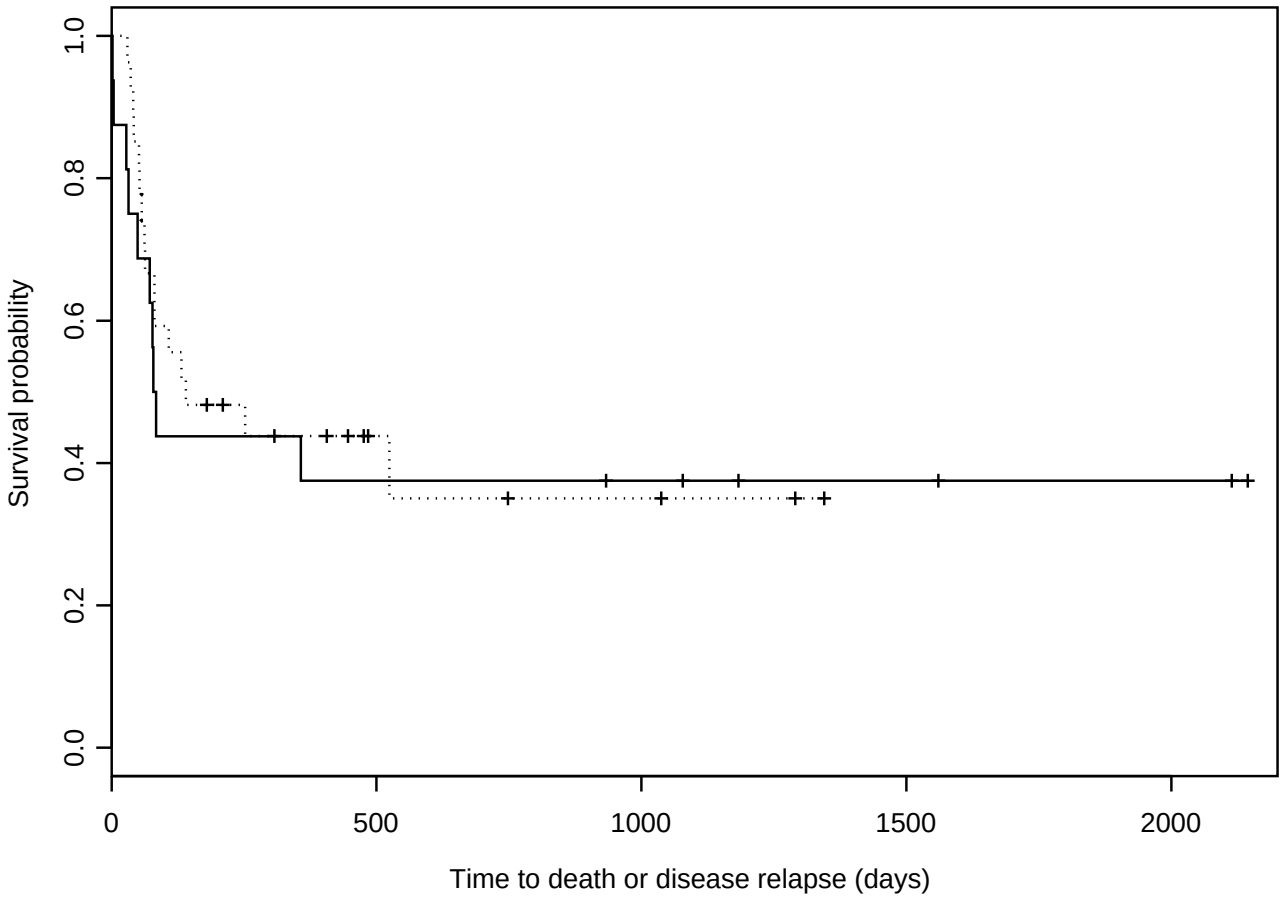


FIGURE : Estimadores K-M de las supervivencias estratificadas por grupo BMT

Modelo

- La gráfica muestra que una parte de los pacientes no presentarán el evento de interés (recaída o muerte), por lo que se pueden considerar curados de la enfermedad

$\Rightarrow$ Necesitamos un modelo que permita una tasa de cura ($T = \infty$) con probabilidad positiva

- Modelo 1 : (Berkson & Gage, 1952)

$$S_{pop}(t) = \pi + (1 - \pi)S(t),$$

donde π =prop. de cura, y $S(t)$ =función de superv. propia

Modelo

- Modelo 2 : (Yakolev & Tsodikov, 1996)

$$S_{pop}(t) = e^{-\theta F(t)},$$

  donde $e^{-\theta}$ =prop. de cura, y $F(t)$ =fn. de dist. propia

- Estos modelos consideran la prop. de cura pero no el tiempo de cura

Modelo

- Modelo 2 : (Yakolev & Tsodikov, 1996)

$$S_{pop}(t) = e^{-\theta F(t)},$$

donde $e^{-\theta}$ =prop. de cura, y $F(t)$ =fn. de dist. propia

- Estos modelos consideran la prop. de cura pero no el tiempo de cura
- Modelo 3 : (Nieto-Barajas & Yin, 2008)

$$h_{pop}(t) = h(t) I(t \leq \tau),$$

  - $h(t)$ una función de riesgo
  - $\exp\{-\int_0^\tau h(s)ds\}$=prop. de cura
  - τ=tiempo de cura

Modelo

- Modelo 3 + covariables :

$$h_i(t|\mathbf{x}_i, z_i) = h(t|z_i)e^{\gamma' \mathbf{x}_i(t)},$$

- Inicial : Modelamos $h(t|z_i)$ que permita determinar una prop. de cura y un tiempo de cura

$$h(t|z_i) = \sum_{k=1}^{\infty} \lambda_k I(k \leq z_i) I(\tau_{k-1} < t \leq \tau_k),$$

  - $\{\lambda_k\}$ es un proceso de Markov gamma en tiempo discreto común a todos los individuous
  - z_i es un índice de cura para el individuo i t.q.

$$z_i \sim \text{Po}^+(e^{\delta' \mathbf{y}_i}),$$

  con δ un vector de coeficientes y $\mathbf{y}_i$ variables explicativas

Modelo

- El proceso de Markov gamma en tiempo discreto $\{\lambda_k\}$ se define como :

$$\begin{aligned} \lambda_1 &\sim \text{Ga}(\alpha_1, \beta_1) \\ u_k|\lambda_k &\sim \text{Po}(c_k\lambda_k) \\ \lambda_{k+1}|u_k &\sim \text{Ga}(\alpha_{k+1} + u_k, \beta_{k+1} + c_k) \end{aligned}$$

para $k = 1, 2, \ldots$

Resultados

- Covariables :
  - X_1=Tipo de transplante (0-autogénico, 1-alogénico)
  - X_2=Tumor Hodgkin (0-ausente, 1-presente)
  - X_3=Coeficiente de Karnofsky (100-muy bien, 0-muerte)
  - X_4=Tiempo de espera al transplante (en meses)
- Las covariables aparecen en el modelo via dos diferentes caminos :
  1. En una forma multiplicativa afectando el riesgo base, y
  2. En la media inicial de z_i afectando el tiempo de cura
- Especificaciones : $\alpha_k = \beta_k = 2$, $c_k = 50$

ITAM 70 AÑOS

Resultados

TABLE : Estimaciones del efecto de las covariables en el riesgo base

	Our model	
Covariate	Mean	95% CI
Trans. type	-0.03	$(-0.92, 0.86)$
Hodgkin	1.17	(0.14, 2.15)
Karnofsky	-0.07	$(-0.08, -0.05)$
Waiting	-0.01	$(-0.03, 0.008)$

- El tener tumor Hodgkin presente es dos veces más riesgoso
- Una disminución de 10 unidades del coef. de Karnofsky aumenta el riesgo en 50%

ITAM 70 AÑOS

Resultados

TABLE : Estimaciones del efecto de las covariables en el tiempo de cura

Covariate	Post. Mean	95% CI
Transplant type	0.10	$(-0.73, 0.88)$
Hodgkin's disease	-0.74	$(-1.70, -0.001)$
Karnofsky score	0.03	(0.02, 0.04)
Waiting time	0.004	$(-0.01, 0.02)$

- El tener tumor Hodgkin presente disminuye el tiempo de cura a la mitad
- Una disminución de 10 unidades del coef. de Karnofsky disminuye el tiempo de cura en 35%

Resultados

TABLE : Estimaciones del tiempo de cura y de la prop. de cura para nuevos pacientes con transplante autogénico con y sin enfermedad Hodgkin

Patient (x_1, x_2, x_3, x_4)	τ_z 95% quantile	π Post. Mean
$(0, 0, 90, 36)$	34 months	0.64
$(0, 1, 90, 36)$	24 months	0.50
$(0, 0, 60, 36)$	15 months	0.27
$(0, 1, 60, 36)$	10 months	0.15

Resultados

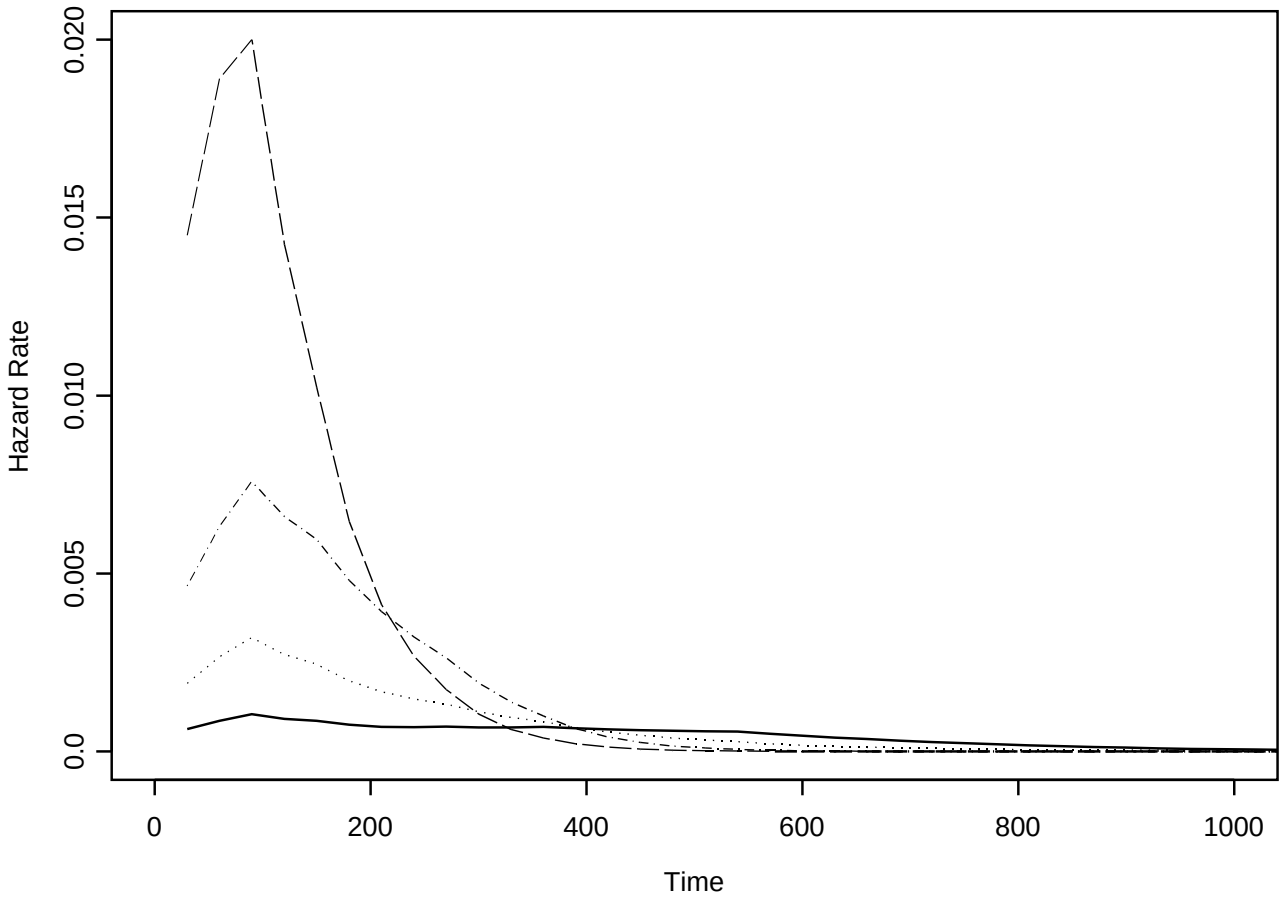


FIGURE : Estimaciones de la función de riesgo para pacientes con covariables (0, 0, 90, 36) linea contínua, (0, 1, 90, 36) linea punteada, (0, 0, 60, 36) linea punteada-rayada, y (0, 1, 60, 36) linea rayada

Resultados

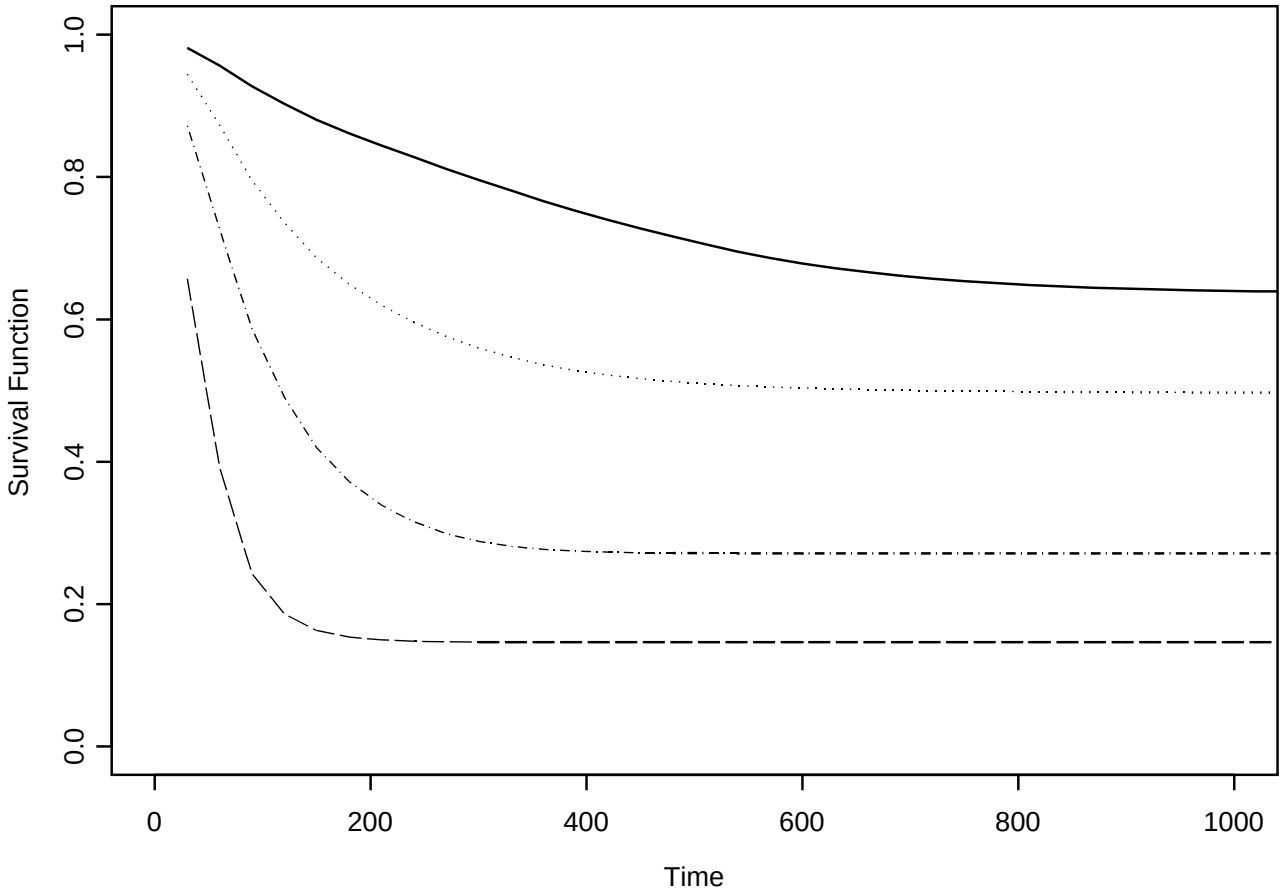


FIGURE : Estimaciones de la función de supervivencia para pacientes con covariables (0, 0, 90, 36) linea contínua, (0, 1, 90, 36) linea punteada, (0, 0, 60, 36) linea punteada-rayada, y (0, 1, 60, 36) linea rayada

Estudio

- Estudio : Muestras de pacientes con cáncer de mama son analizadas mediante la técnica de arreglos de Hibridación Genómica Comparativa (aCGH). Se contabiliza el número de copias de cadenas del ADN
  - 2 = normal
  - 1 = pérdida
  - $\geq$3 = ganancia
- $Y_{ij} =$ Logaritmo (base 2) del cociente del número de copias de cadenas de ADN con respecto a 2 en la sonda (base nitrogenada) i e individuo j
- Datos : $n = 549$ sondas, $J = 111$ muestras clasificadas en dos tipos : ER+ (60 muestras) y TN (51 muestras)
- Objetivo : Determinar las regiones del cromosoma con mismo número copiado y determinar si existen regiones con aberraciones distintas en los dos tipos de cáncer ER+ y TN

Estudio

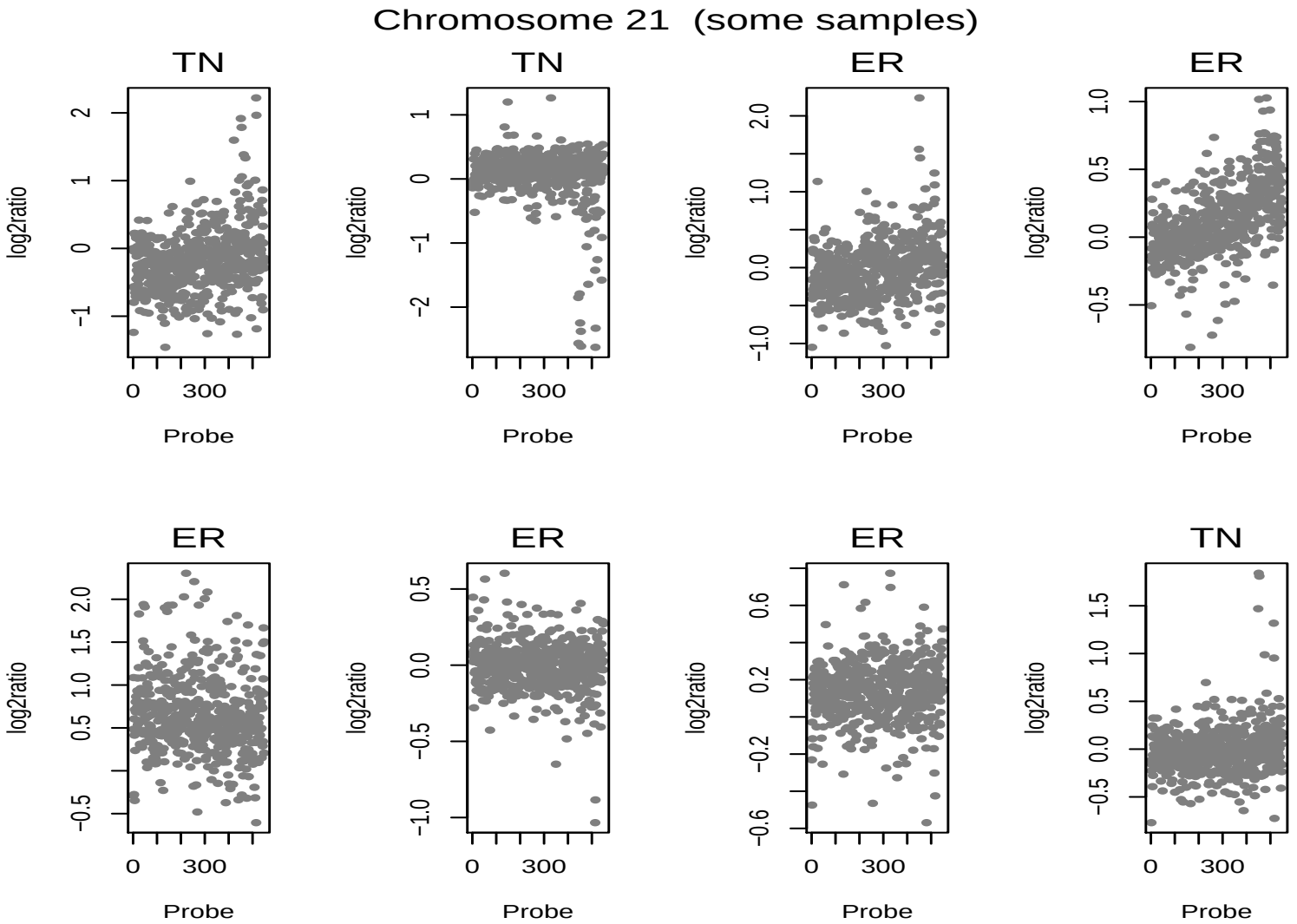


FIGURE : 8 primeras muestras

Modelo

- Modelo semiparamétrico : (Nieto-Barajas et al., 2016)

$$Y_{ij} = \sum_{k=1}^{K} \mu_{kg_j} I(i \in \Delta_k) + \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj}) + \epsilon_{ij}$$

donde Δ_k=part. pob. y Δ_{lj}=part. indiv.

- $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2_\epsilon)$, $\sigma^2_\epsilon \sim \text{Iga}(2, 1)$
- $m_{lj} \overset{\text{iid}}{\sim} \text{N}(0, \tau^2_j)$, $\tau^2_j \sim \text{Iga}(2, 1)$
- $(\mu_{k1}, \mu_{k2})|G_0, G_1 \overset{\text{iid}}{\sim} (1-\pi)G_0 + \pi G_1$,
  $G_0 \sim \mathcal{DP}(a_0, F_0)$, $G_1 \sim \mathcal{DP}(a_1, F_1)$, con $\pi = 0.5$
  $a_0 \sim \text{Ga}(1, 1)$, $a_1 \sim \text{Ga}(1, 1)$
- $F_0(\mu_{k1}, \mu_{k2}) = \text{N}(\mu_{k1}|0, \lambda^2_0) I(\mu_{k1} = \mu_{k2})$,
  $F_1(\mu_{k1}, \mu_{k2}) = \text{N}_2(\mu_{k1}, \mu_{k2}|\mathbf{0}, \Lambda_1)$

Resultados

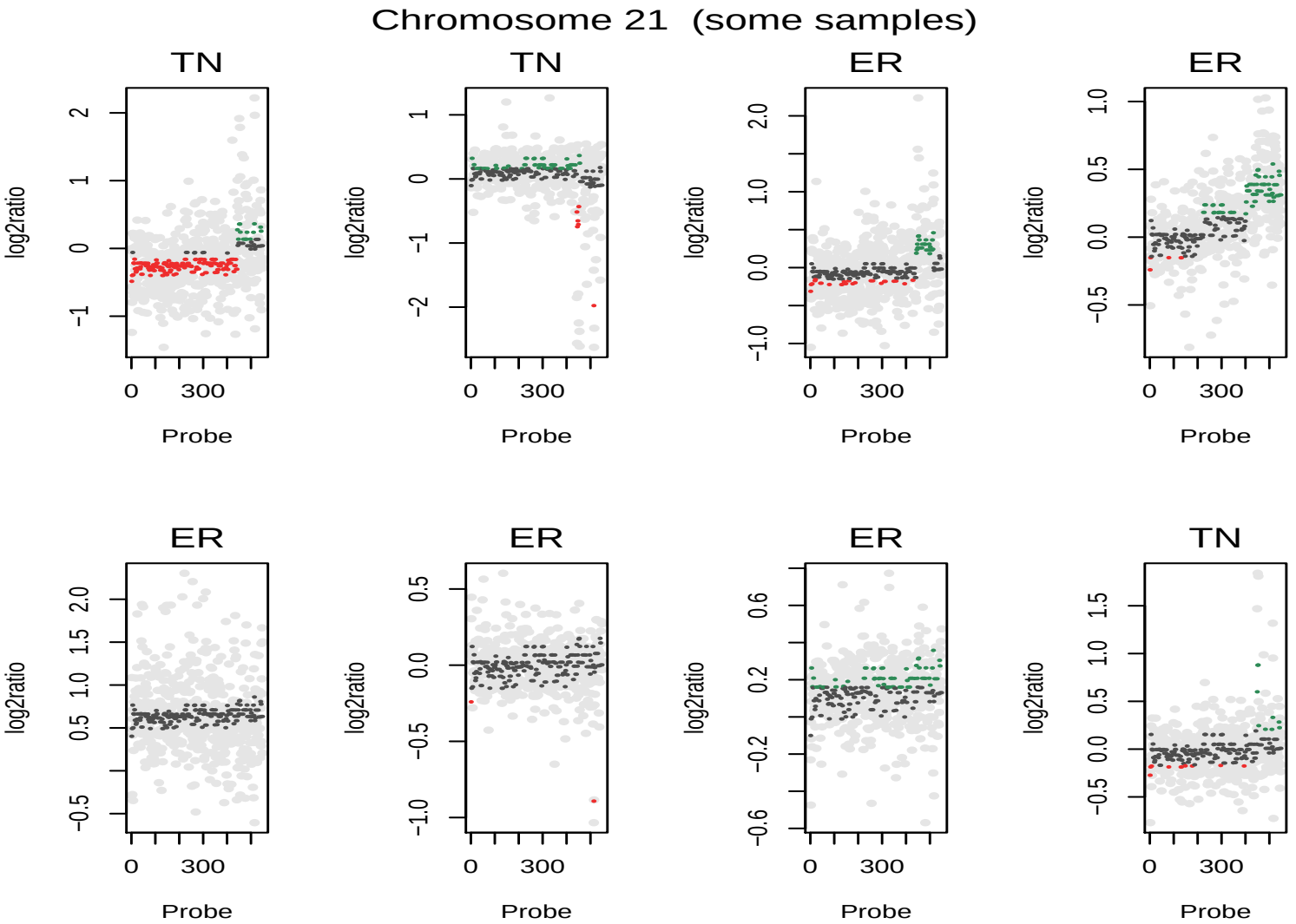


FIGURE : Inferencia para las 8 primeras muestras

Estudio

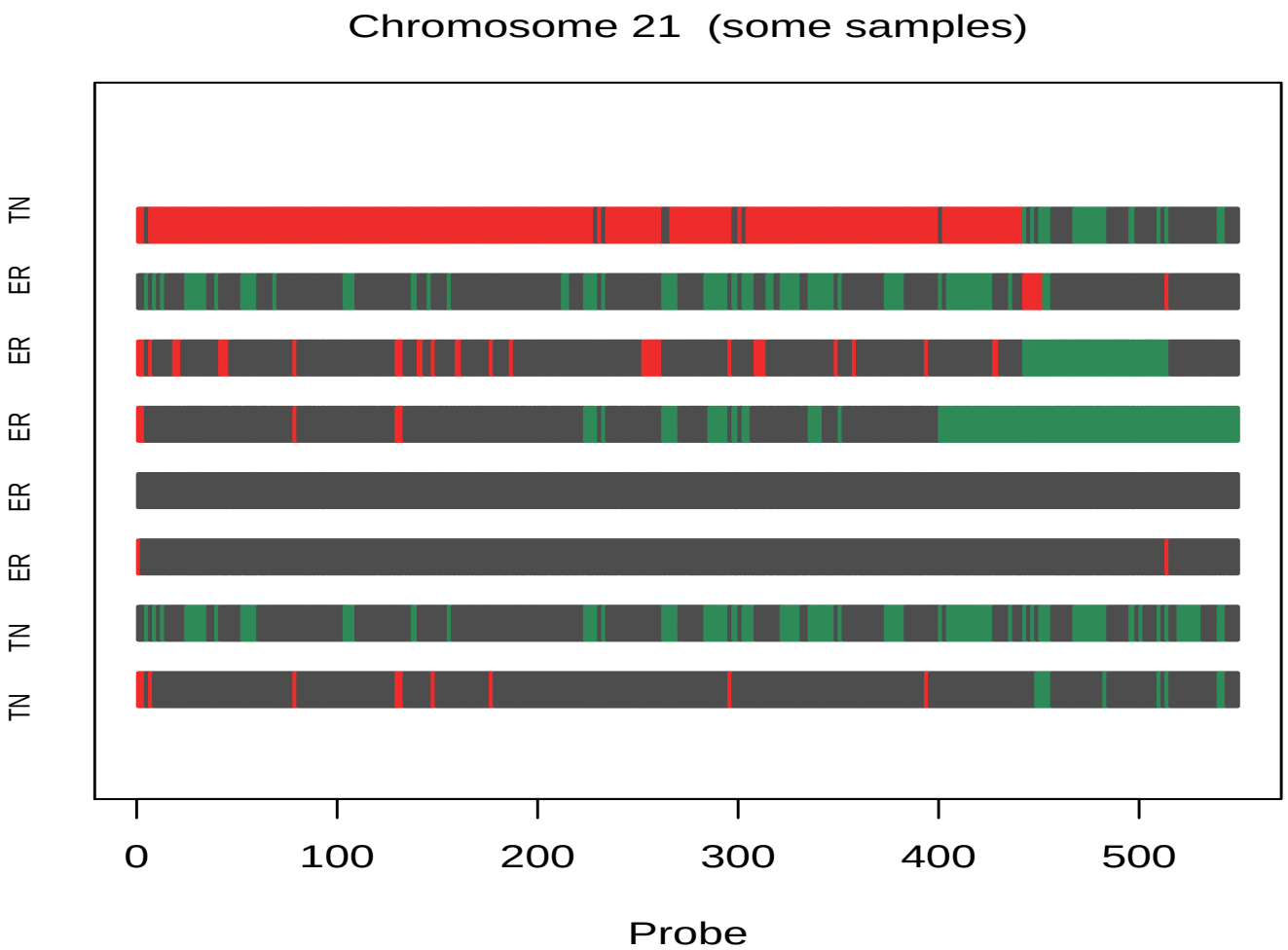


FIGURE : Inferencia para las 8 primeras muestras

Estudio

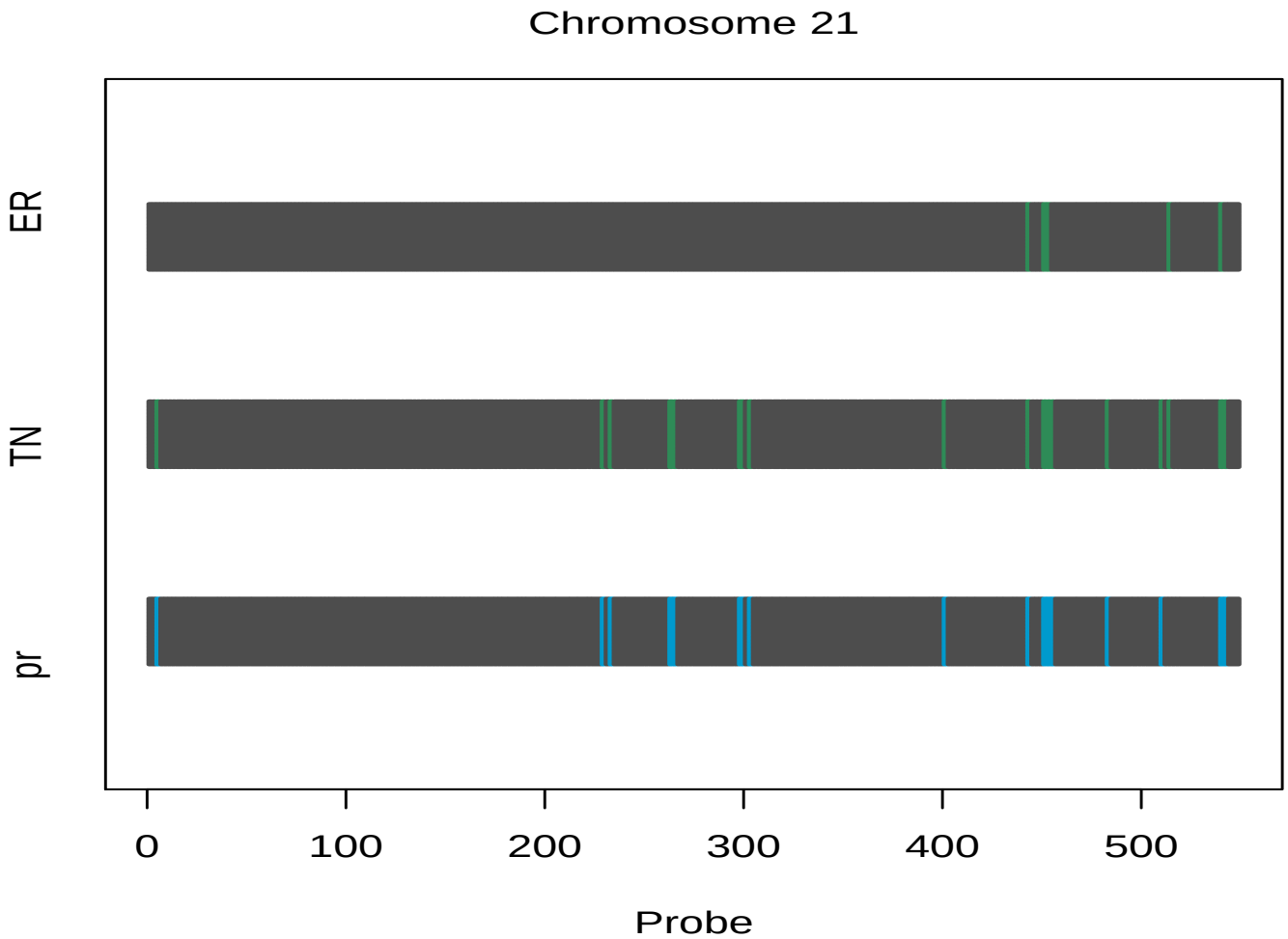


FIGURE : Resumen inferencial para las medias poblacionales

Referencias

- Nieto-Barajas, L.E. and Walker, S.G. (2002). Markov beta and gamma processes for modeling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.
- Nieto-Barajas, L.E. and Walker, S.G. (2004). Bayesian nonparametric survival analysis via Lévy driven Markov processes. *Statistica Sinica* **14**, 1127–1146.
- Nieto-Barajas, L.E. and Walker, S.G. (2005). A semiparametric Bayesian analysis of survival data based on Lévy-driven processes. *Lifetime data analysis* **11**, 529–543.
- Nieto-Barajas, L.E. and Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian Journal of Statistics* **35**, 540–556.
- Nieto-Barajas, L. E., Ji, Y. and Baladandayuthapani, V. (2016). A semiparametric Bayesian model for comparing DNA copy numbers. *Brazilian Journal of Probability and Statistics* **30**, 345–365.

ITAM 70 AÑOS