# Curso Demografía - Licenciatura en Estadística, UDELAR

Daniel Ciganda

Clase 15

11 de Noviembre de 2024

## Simulación de Procesos Demográficos

Pasos a seguir:

- Obtener una distribución de tasas específicas por edad *condicionales* del evento en cuestión
- Utilizar estas tasas como un modelo de riesgo constante a intervalos (edades) del evento
- Invertir la función de riesgo acumulado para obtener realizaciónes (números aleatorios) de nuestra distribución de partida
- Vincular ambos procesos en un modelo de simulación de eventos discretos que represente las trajectorias de las mujeres de una cohorte

$$m_x = \frac{D_x}{L_x}$$

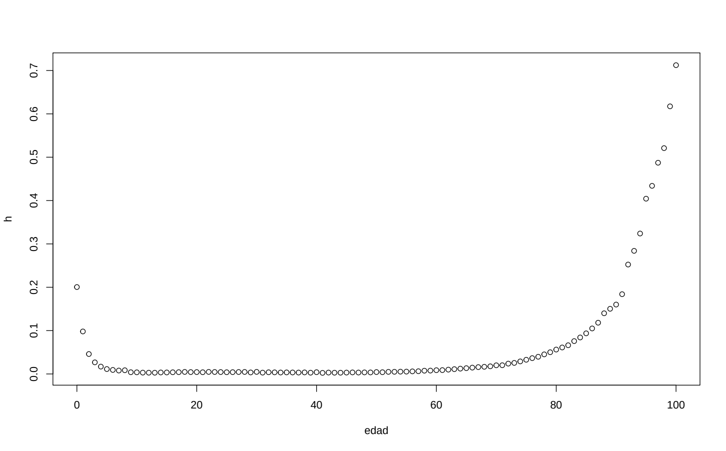



**Figure 1:** Tasas de Mortalidad por Edad de una Cohorte

$$f_{i,x} = \frac{B_{i,x}}{L_{i-1,x}}$$

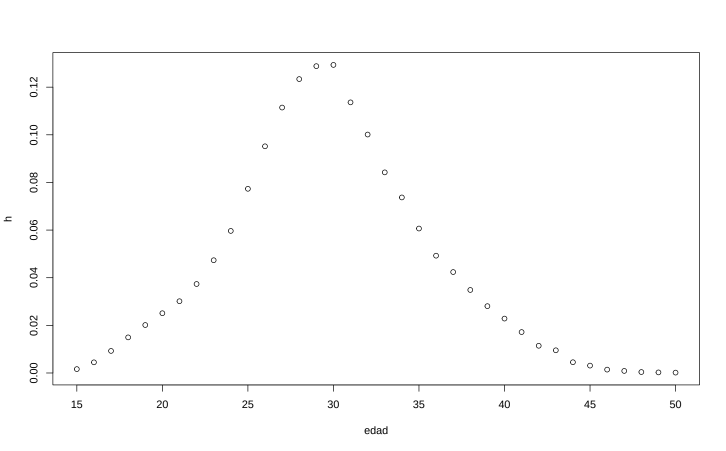



**Figure 2:** Tasas Condicionales de Fecundidad por Edad de una Cohorte - Primer Hijo

Pasos a seguir:

- Obtener una distribución de tasas específicas por edad condicionales del evento en cuestión
- Utilizar estas tasas como un modelo de riesgo constante a intervalos (edades) del evento
- Invertir la función de riesgo acumulado para obtener realizaciónes (números aleatorios) de nuestra distribución de partida
- Vincular ambos procesos en un modelo de simulación de eventos discretos que represente las trajectorias de las mujeres de una cohorte

Variable Aleatoria $T > 0$ , continua

Función de Distribución: $F(t) = P(T \leq t)$

Función de Supervivencia: $S(t) = 1 - F(t)$

Densidad: $f(t) = F'(t)$

Hazard:

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

Además:

$H(t) = \int_0^t h(s)ds$

$S(t) = exp\{-H(t)\}$

Cuando asumimos que el riesgo es constante en cada intervalo de edad, el vector de tasas de ocurrencia/exposición por edad observadas representa la función de riesgo que caracteriza a la variable aleatoria *tiempo al evento*.

Esto es así ya que, bajo el supuesto de riesgo constante, el Estimador de Máxima Verosimilitud (EMV) del riesgo de un evento es la cantidad de eventos observados sobre el tiempo de exposición al riesgo del evento, es decir la tasa de ocurrencia/exposición.

Vamos a ilustrar esta idea a través de un ejemplo.

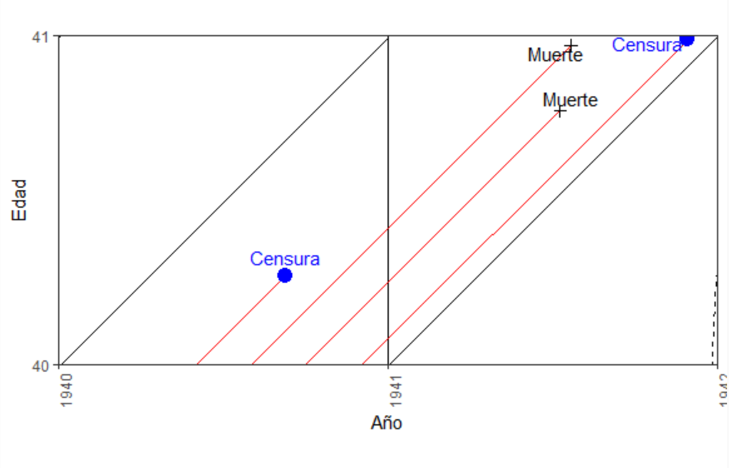



**Figure 3:** Seguimiento de 4 individuos en un estudio longitudinal. La primer persona abandona el estudio antes de morir of cumplir 41 años. La segunda y tercera persona fallecen antes de cumplir 41 años de edad. La cuarta observación también es censurada, pero en este caso porque la persona alcanza su 41 cumpleaños.

# Ejemplo: Mortalidad de una cohorte en edades 40-41

En un estudio, se observó la edad exacta de muerte de individuos en el intervalo de edad de 40 a 41 años:

- $x_i$ = edad hasta la que se observa a cada individuo $i$.
- $d_i$ es un indicador: $d_i = 1$ si el individuo murió, $d_i = 0$ si no tenemos información sobre su muerte en ese intervalo (datos censurados).

Nuestro objetivo es derivar el EMV para el riesgo de muerte bajo la distribución exponencial en este contexto específico.

# Ejemplo: Mortalidad de una cohorte en edades 40-41

Bajo el supuesto de que nuestros tiempos a la muerte se distribuyen de acuerdo a la distribución exponencial, la función de riesgo es:

$$h(x) = \lambda$$

La función de supervivencia es:

$$S(x) = e^{-\lambda x}$$

Y la función de densidad para el tiempo a la muerte (luego de los 40) es:

$$f(x) = \lambda e^{-\lambda x}$$

# Función de Verosimilitud

Para cada individuo:

- Si $d_i = 1$: contribución a la verosimilitud = $f(x_i)$
- Si $d_i = 0$: contribución a la verosimilitud = $S(x_i)$

Por lo tanto, la función de verosimilitud está dada por:

$$L(\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)^{d_i} \left(e^{-\lambda x_i}\right)^{1-d_i}$$

# Log-verosimilitud y EMV

Tomando la log-verosimilitud:

$$\ell(\lambda) = \sum_{i=1}^{n} [d_i \log(\lambda) - \lambda x_i]$$

Diferenciando respecto a $\lambda$ e igualando a cero:

$$\frac{d\ell}{d\lambda} = \sum_{i=1}^{n} \left[\frac{d_i}{\lambda} - x_i\right] = 0$$

Esto nos da:

$$\hat{\lambda} = \frac{\sum d_i}{\sum x_i}$$

EMV del riesgo de muerte:

$$\frac{\text{numero total de muertes}}{\text{tiempo total de exposición al riesgo de morir}} = \frac{D_x}{L_x} = m_x$$

# Simulación de Procesos Demográficos

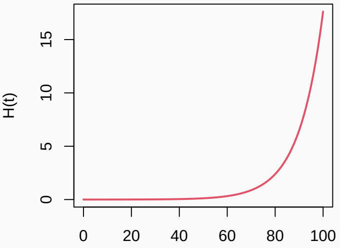
Pasos a seguir:

- Obtener una distribución de tasas específicas por edad condicionales del evento en cuestión
- Utilizar estas tasas como un modelo de riesgo constante a intervalos (edades) del evento
- Obtener la función de riesgo acumulado $H(t)$. Invertir $H(t)$ para obtener realizaciónes (tiempos de espera) de nuestra distribución de partida
- Vincular ambos procesos en un modelo de simulación de eventos discretos que represente las trajectorias de las mujeres de una cohorte

# Método de Inversión - Función de Riesgo Acumulado

Podemos obtener realizaciones de T a partir de la función de riesgo acumulado $H(t)$ con:

$$T = H^{-1}(-\log U)$$




Podemos obtener realizaciones de X:

- Generando valores aleatorios $u_1, \ldots, u_n \sim U(0,1)$
- Transformarlos $H^{-1}(-\log u_i) = t_i$

Podemos invertir $H(t)$

$$H(t_i) = -\log u_i \qquad \Rightarrow \qquad H(t_i) + \log u_i = 0$$

En R: `uniroot()`

# El Estimador Kaplan-Meier de la Función Supervivencia

- Método no paramétrico para estimar la función de supervivencia.
- Tiene en cuenta datos censurados.
- La probabilidad de supervivencia en el tiempo $t$ es un producto de probabilidades condicionales.

# Fórmula de Kaplan-Meier

Para cada tiempo $t_i$:

$$S(t_i) = \frac{n_i - d_i}{n_i}$$

Donde:

- $S(t)$: Probabilidad estimada de supervivencia en el tiempo $t$.
- $d_i$: Número de eventos en el tiempo $t_i$.
- $n_i$: Número en riesgo justo antes del tiempo $t_i$.

La probabilidad acumulada de supervivencia hasta el tiempo $t$ es:

$$S(t) = \prod_{i:t_i \leq t} \left( \frac{n_i - d_i}{n_i} \right)$$

# Ejemplo: Datos de tiempo a la Muerte

Consideremos un conjunto de datos de tiempo hasta la muerte o último seguimiento:

| Individuo | Tiempo (meses) | Evento |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 4 | 1 |
| 3 | 6 | 0 |
| 4 | 8 | 0 |
| 5 | 10 | 1 |
| 6 | 11 | 1 |
| 7 | 12 | 0 |
| 8 | 15 | 1 |
| 9 | 18 | 0 |
| 10 | 20 | 1 |

Donde ”Evento” es una variable binaria: 1 indica muerte y 0 indica que la observación fue censurada en el último seguimiento.

# Cálculo

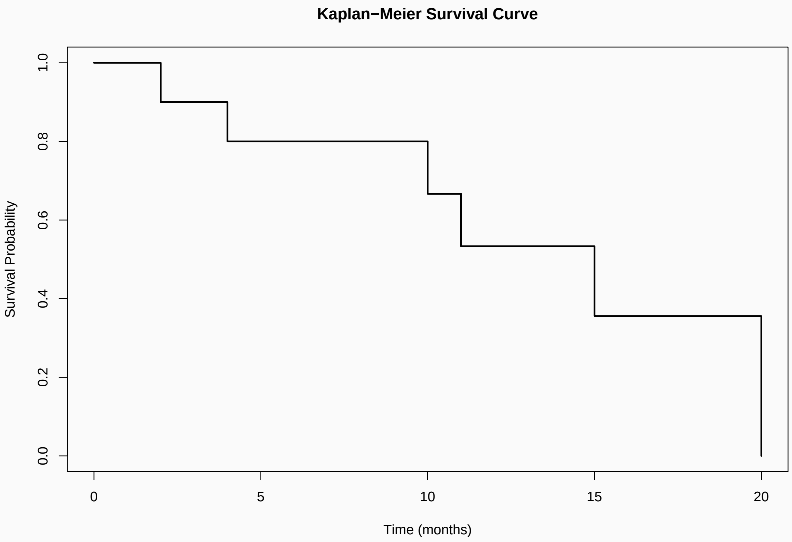
Para cada tiempo $t$ único donde ocurre un evento:

- Determine $n$, el número de individuos en riesgo justo antes de $t$.
- Determine $d$, el número de eventos (muertes) en $t$.
- Calcule $S(t)$ usando:

$$S(t) = \prod_{i:t_i \leq t} \left( \frac{n_i - d_i}{n_i} \right)$$

| Tiempo (meses) | Evento | Cálculo | S(t) |
|---|---|---|---|
| 2 | muerte | $1.0000 \times (1 - \frac{1}{10})$ | 0.9000 |
| 4 | muerte | $0.9000 \times (1 - \frac{1}{9})$ | 0.8000 |
| 6 | censurado | $0.8000 \times (1 - \frac{0}{8})$ | 0.8000 |
| 8 | censurado | $0.8000 \times (1 - \frac{0}{7})$ | 0.8000 |
| 10 | muerte | $0.8000 \times (1 - \frac{1}{6})$ | 0.6667 |
| 11 | muerte | $0.6667 \times (1 - \frac{1}{5})$ | 0.5333 |
| 12 | censurado | $0.5333 \times (1 - \frac{0}{4})$ | 0.5333 |
| 15 | muerte | $0.5333 \times (1 - \frac{1}{3})$ | 0.3556 |
| 18 | censurado | $0.3556 \times (1 - \frac{0}{2})$ | 0.3556 |
| 20 | muerte | $0.3556 \times (1 - \frac{1}{1})$ | 0.0000 |

**Table 1:** Estimador Kaplan-Meier.


Kaplan−Meier Survival Curve
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time (months)