Análisis de datos longitudinales

Grado en Estadística

Bloque 2 – Sesión 3

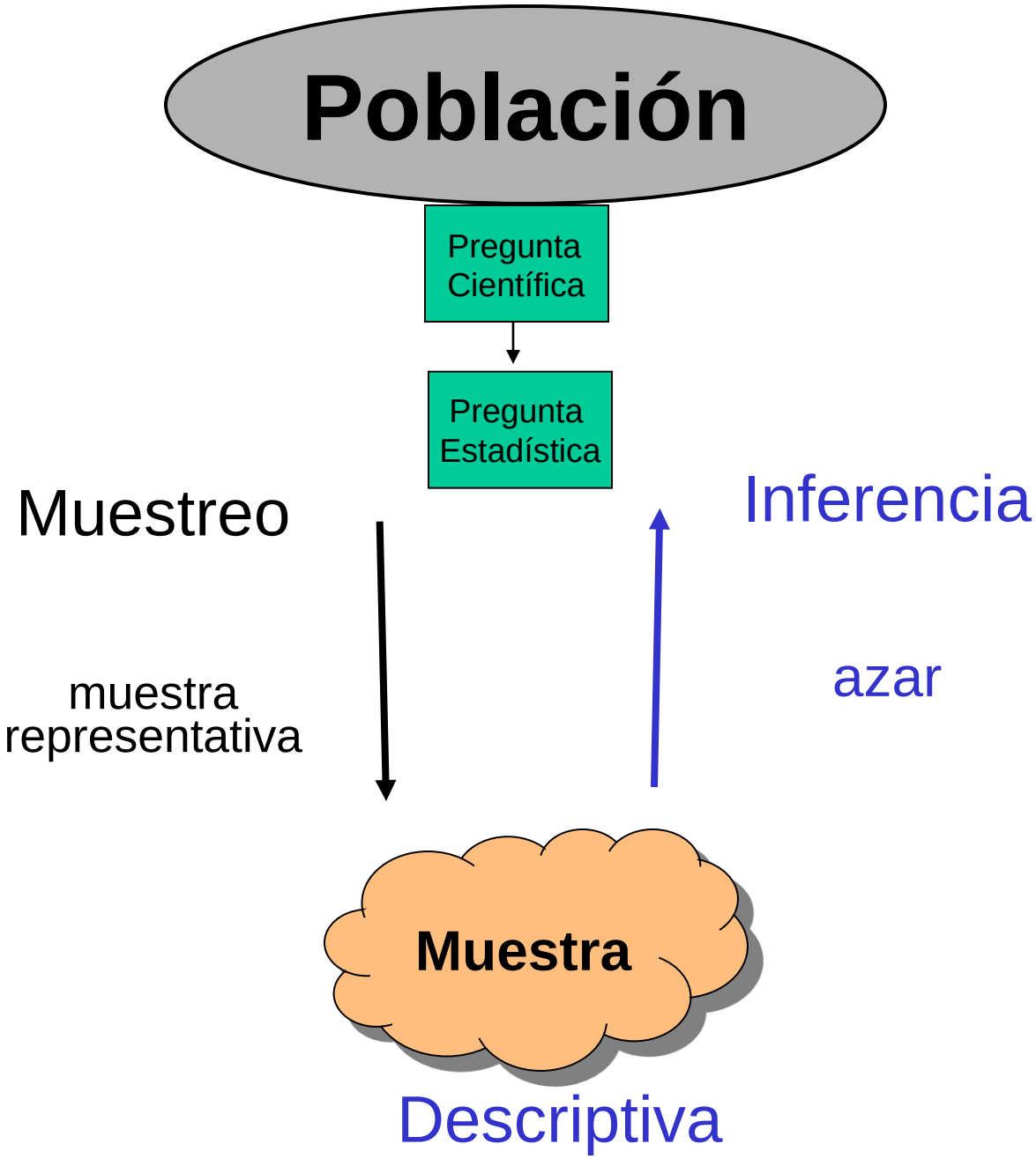
Análisis de Supervivencia (I)

Juan R González

Departamento de Matemáticas, UAB

Insitituto de Salud Global de Barcelona, ISGlobal

juanr.gonzalez@isglobal.org


Población
Pregunta Científica
Pregunta Estadística
Muestreo
muestra representativa
Inferencia
azar
Muestra
Descriptiva

Tipos de preguntas

Científica	Estadística
• ¿Tomar sal (si/no) influye en la probabilidad de sufrir un infarto de miocardio en pacientes hipertensos?	$H_0 : p_{\text{tomar sal}} = p$
• ¿El nivel de colesterol (continua) influye en la probabilidad de sufrir un infarto de miocardio en pacientes hipertensos?	$H_0 : \bar{x}_{\text{infarto}} = \bar{x}$ Modelo: Regresión lineal
• ¿El nivel de colesterol influye en la probabilidad de sufrir un infarto de miocardio (si/no) en pacientes hipertensos teniendo en cuenta el consumo de sal?	Modelo: Regresión logística
• ¿El nivel de colesterol influye en el riesgo de sufrir un infarto de miocardio (tiempo hasta que se observa el infarto) en pacientes hipertensos teniendo en cuenta el consumo de sal?	Modelo: Análisis de supervivencia **Censura!**

Guión

- Descripción del problema: la censura
- Parte I: La función de Supervivencia
  - Estimación e interpretación: Kaplan-Meier
  - Comparación de funciones de Supervivencia: log-rank
- Parte II: Introducción de covariables:
  - El modelo de Cox
  - Interpretación del modelo
  - Elección del modelo (ejemplo artículo científico)
  - Validación del Modelo

Análisis de supervivencia

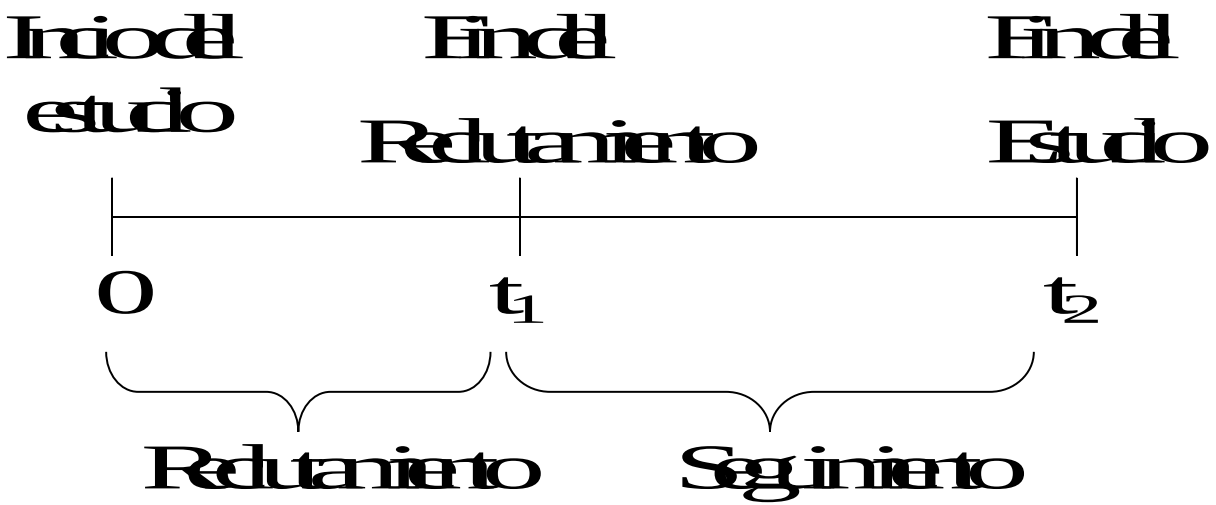
Parte I. La función de Supervivencia

Esquema

- Diseño de estudios de seguimiento
- Supervivencia: tiempo hasta un evento
- Censuras
- Funciones estadísticas
- Estimación de la probabilidad de sobrevivir
- Comparación de curvas de supervivencia

Diseño de un estudio prospectivo

- Estudio de cohortes (registros)
- Ensayo clínico

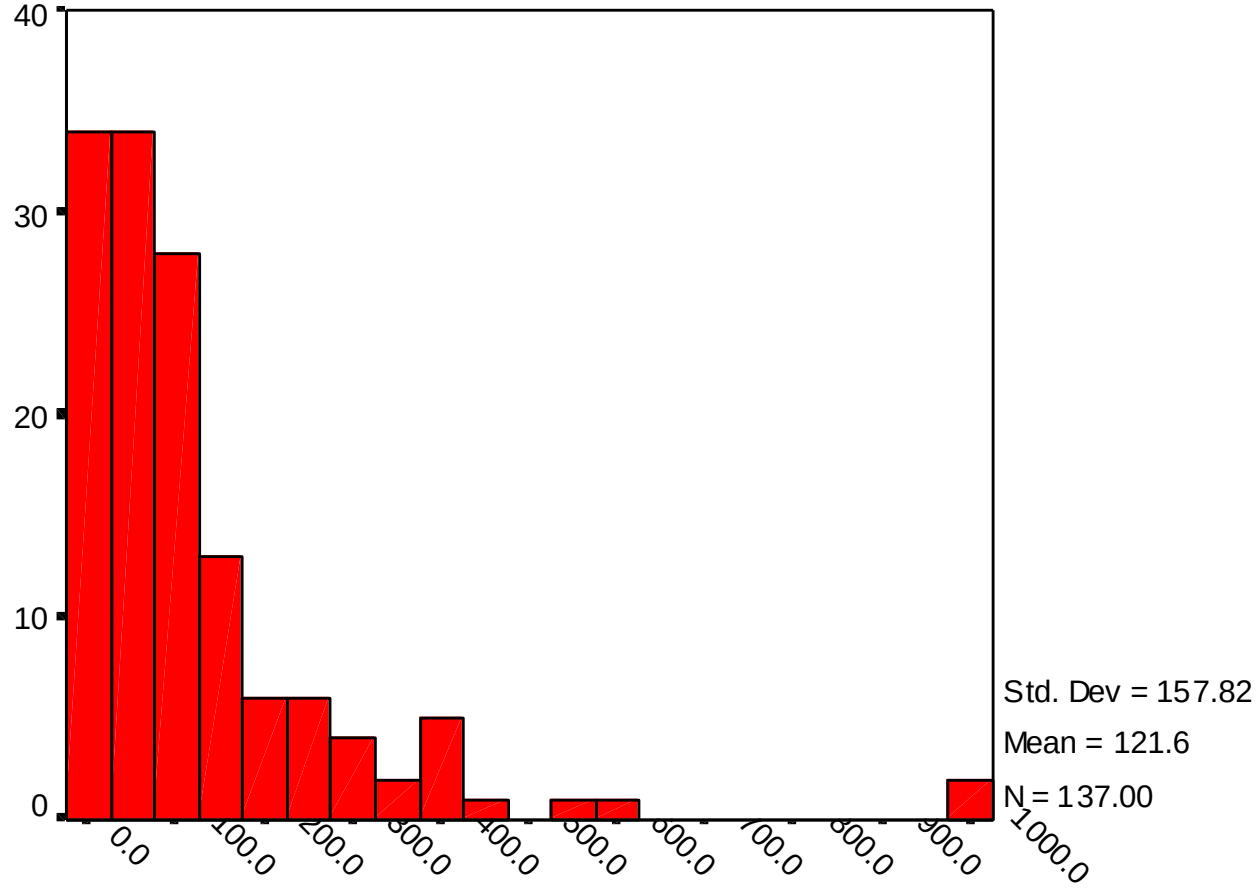


Variable de interés

- Tiempo hasta que ocurre un suceso

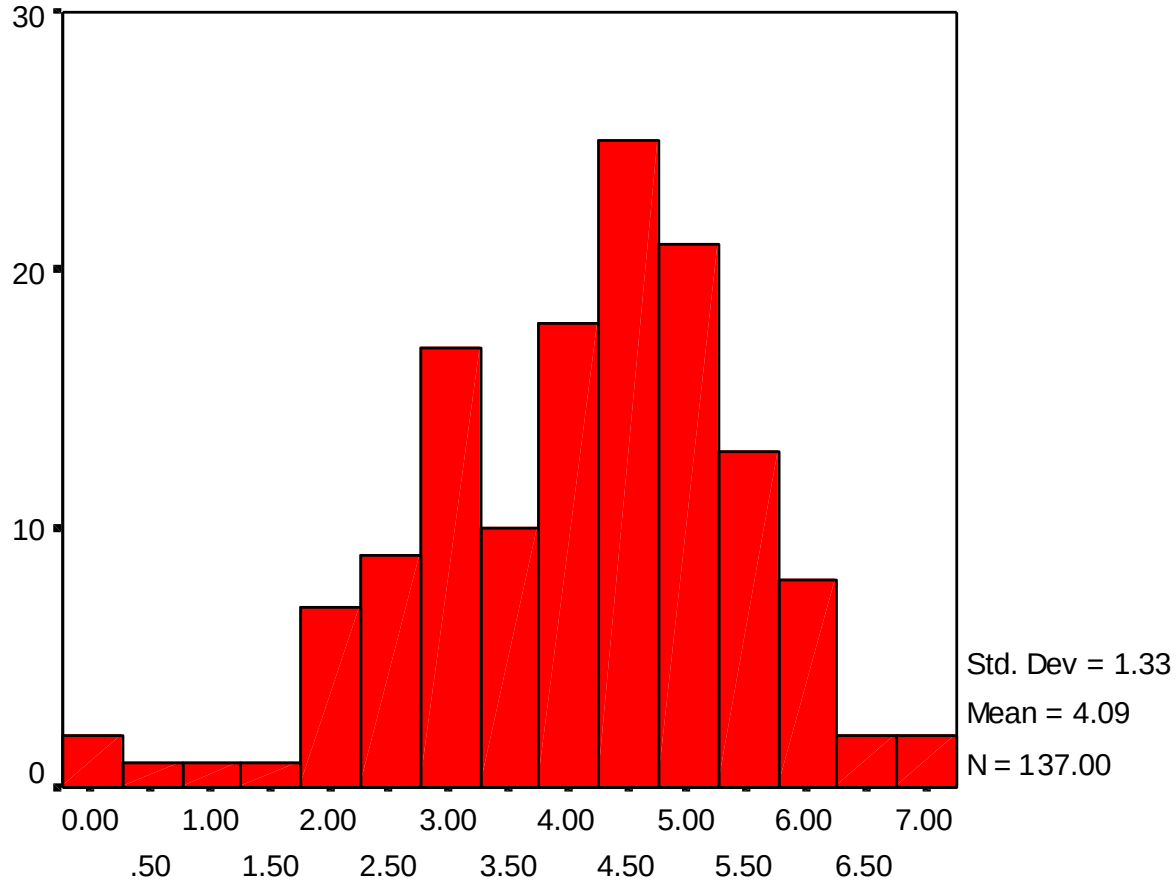
tiempo entrada → tiempo suceso

- Eventos:
  - muerte: **supervivencia**
  - recaída/metástasis: "tiempo libre de enfermedad"
  - curación
  - transplante



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
TIEMPO	137	1	999	121.63	157.82	3.127	.207	13.070	.411
Valid N (listwise)	137								



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
LOGT	137	.00	6.91	4.0934	1.3279	-.546	.207	.366	.411
Valid N (listwise)	137								

Datos Censurados

- Para algunos pacientes el evento de interés puede no haber ocurrido durante el tiempo de observación (t)
- Información incompleta: $T > t$

- Se necesitan dos variables para caracterizar los datos de supervivencia
  - T: tiempo de observación
  - δ: indicador del estado (binario)

Causas de censuras

- Final programado del estudio para el análisis
- Pérdidas de seguimiento
- Abandonos
- Muerte por otras causas diferentes de la de interés

Tipos de censura

- Tipo I. Todos los individuos se siguen hasta una fecha fin de estudio

  - Por la derecha: Pacientes vivos al finalizar el estudio

    Pacientes perdidos o abandonos
  - En intervalo: Las visitas de control son espaciadas
  - Por la izquierda: Se desconoce la fecha de inicio

- Tipo II. Los individuos se siguen hasta que han ocurrido r eventos

Truncamiento

- Los individuos entran en el estudio por un criterio determinado y los que no cumplen el criterio no son visibles al investigador.

Ejemplo:

- Interesa estudiar la edad al morir pero sólo se estudian ancianos de un asilo
  - Inicio: edad al ingresar al asilo (truncamiento)
  - Final: edad al morir
- Los muertos anteriores a la jubilación no pueden entrar en el asilo, por tanto los datos están truncados por la izquierda

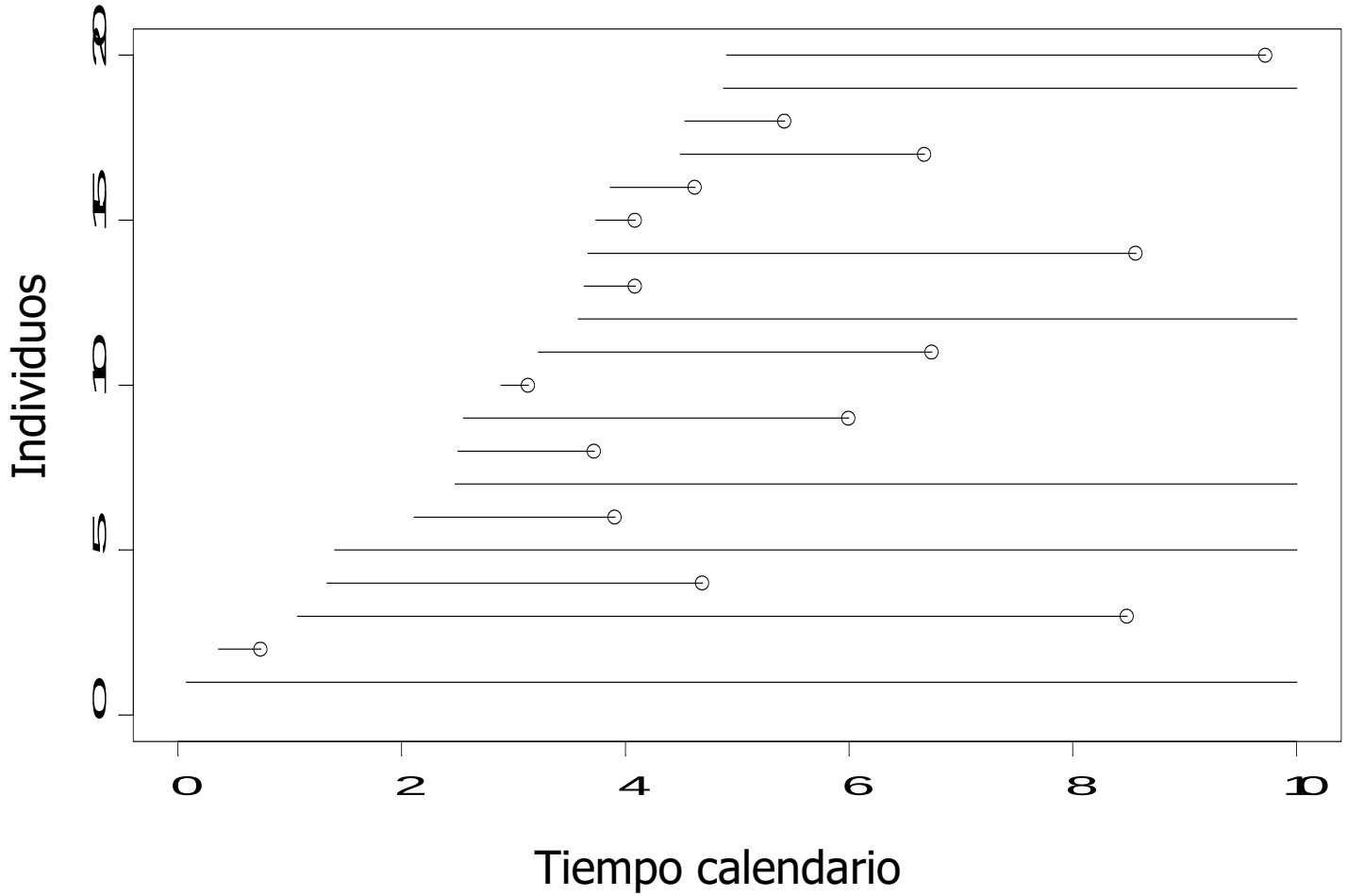
Escalas de tiempo

- Calendario: Inicio a fin del estudio

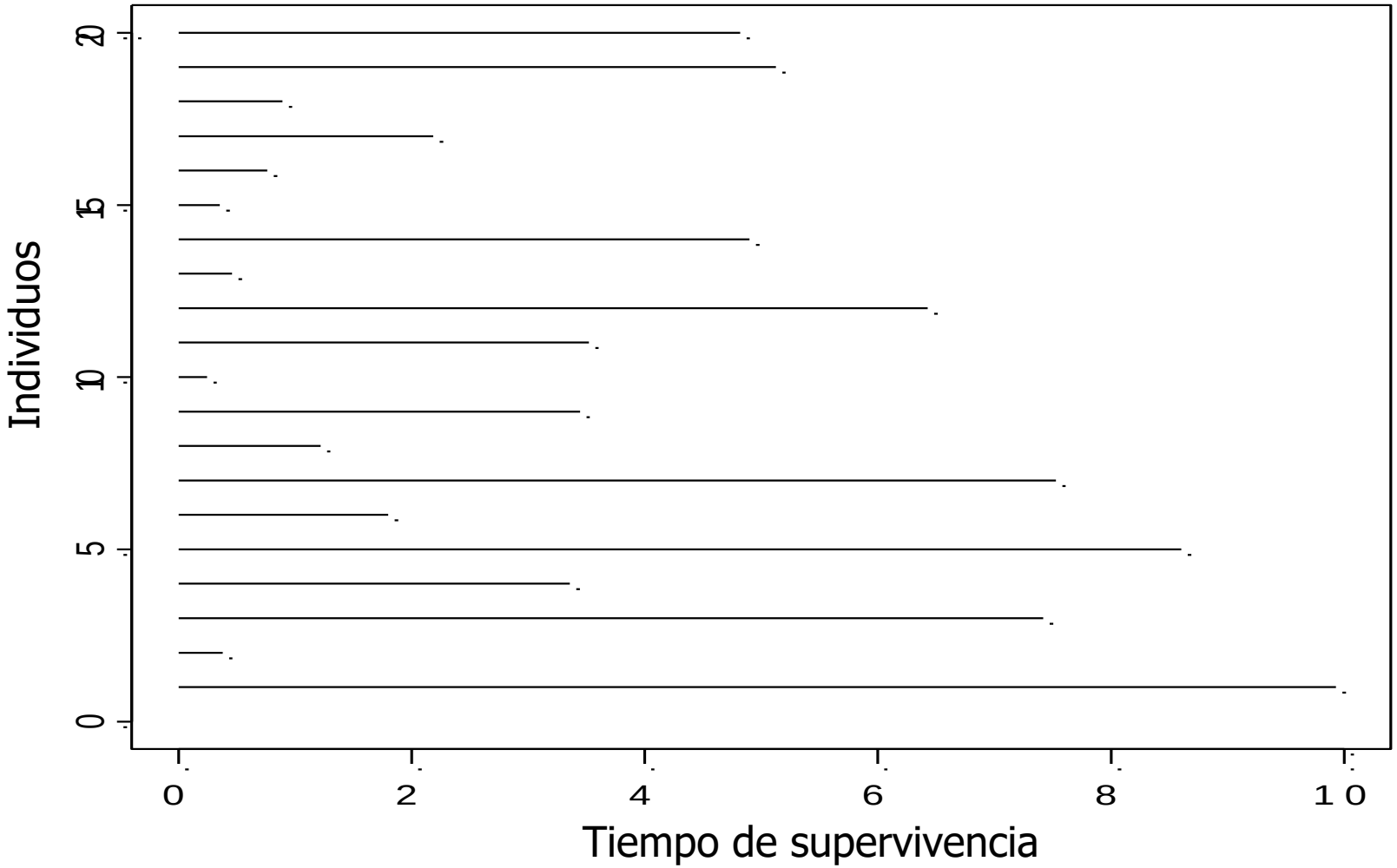
- Tiempo del paciente en el estudio:
  entrada a salida (por muerte o censura)

- Otras escalas pueden ser de interés:
  - edad "en el momento actual"
  - duración de una exposición

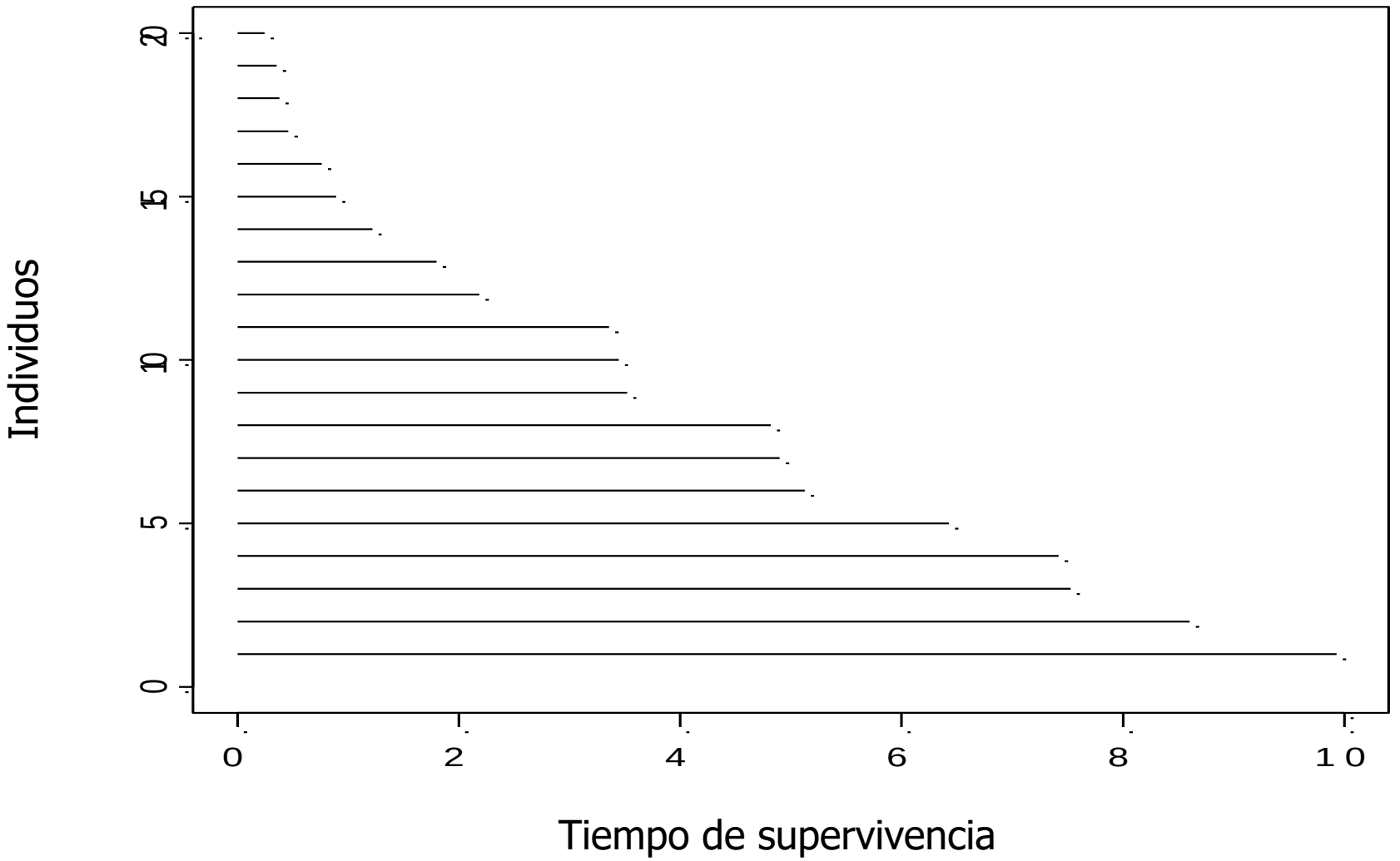
Escalas de tiempo



Escalas de tiempo



Escalas de tiempo



Descripción del tiempo de seguimiento

- Describir el tiempo de seguimiento.
  - ¿En eventos o en censuras o ambos?
  - ¿Qué estadístico/s de resumen es/son útil/es?

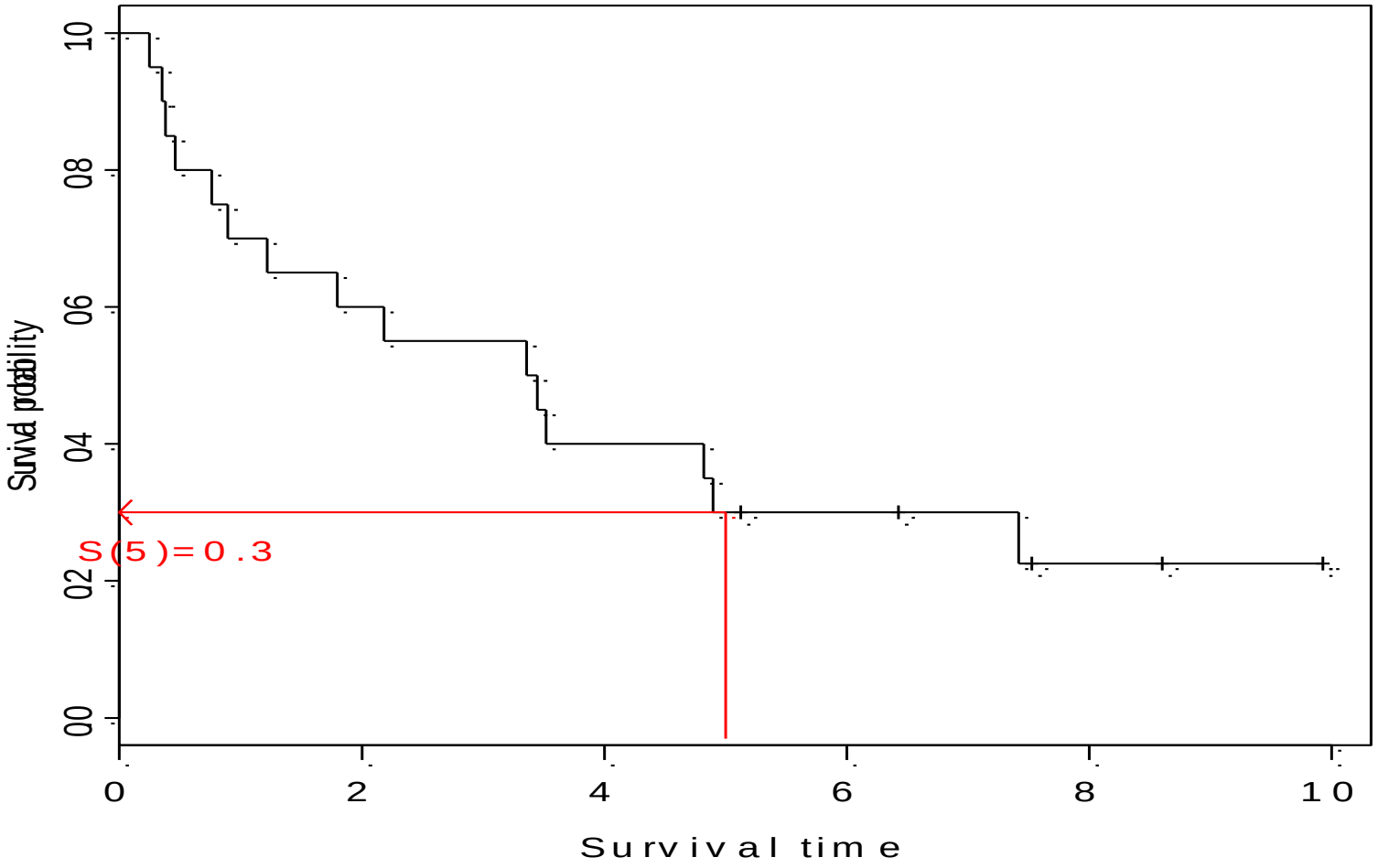
Descripción de la supervivencia

- T es cuantitativa continua
- Descripción (supervivencia, probabilidad, densidad, riesgo, riesgo acumulado…) :
  - **Supervivencia**: Probabilidad de sobrevivir *t* o más:

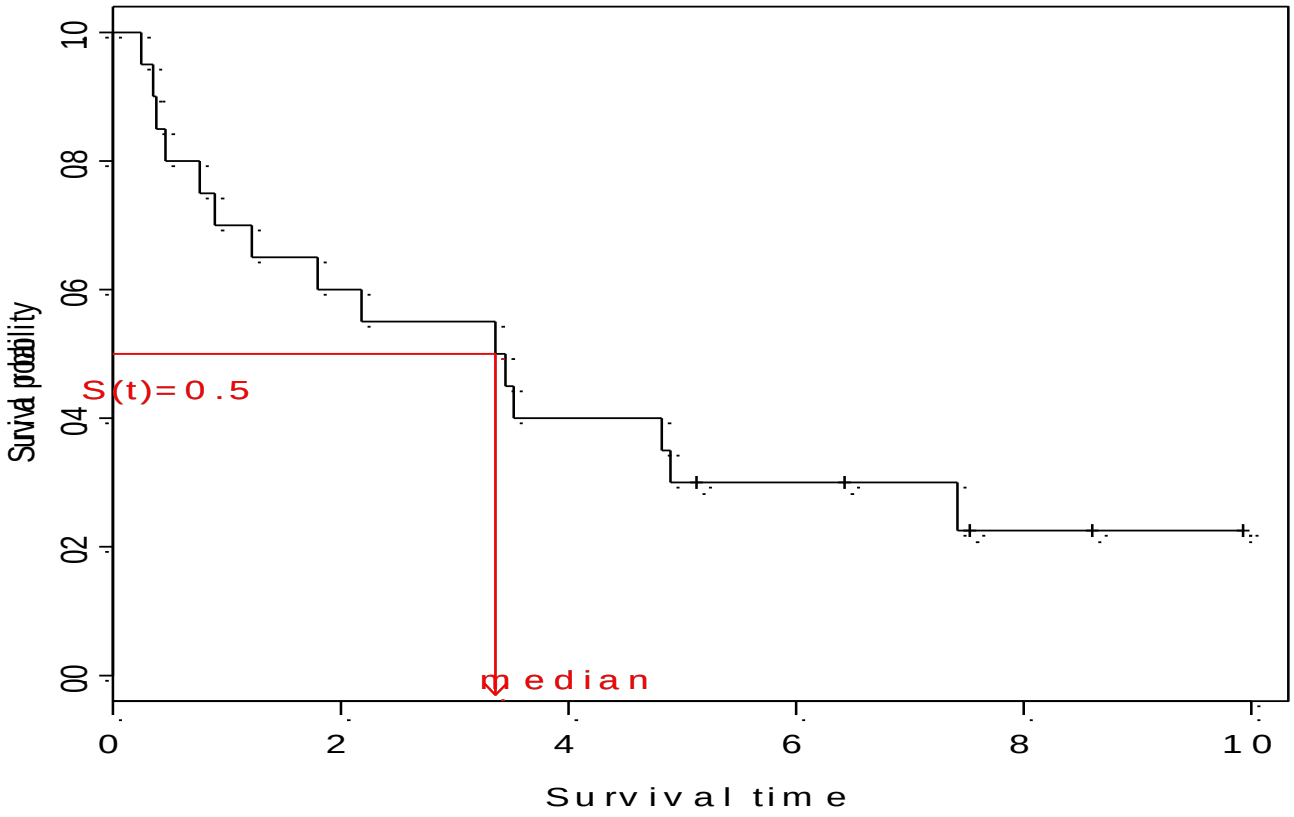
$$S(t) = Pr\ (T \geq t)$$

    - Acumulativa
  - **Percentiles**: tiempo que sobrevive una proporción de la población

Proporción que sobrevive t o más



Tiempo mediano de supervivencia

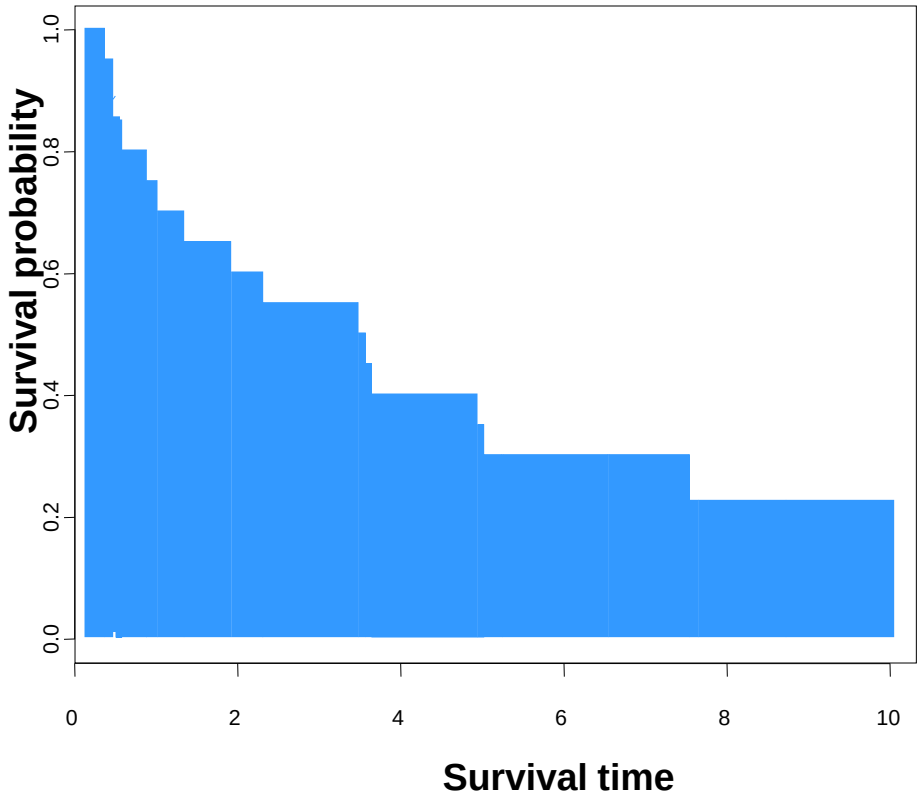


Nota: NO tiempo medio

Tiempo medio de supervivencia

- Media = área bajo $S(t)$
- No estimable si $S(t)$ no llega a 0
- Sesgado

  (T asimétrico)
- No es un buen resumen



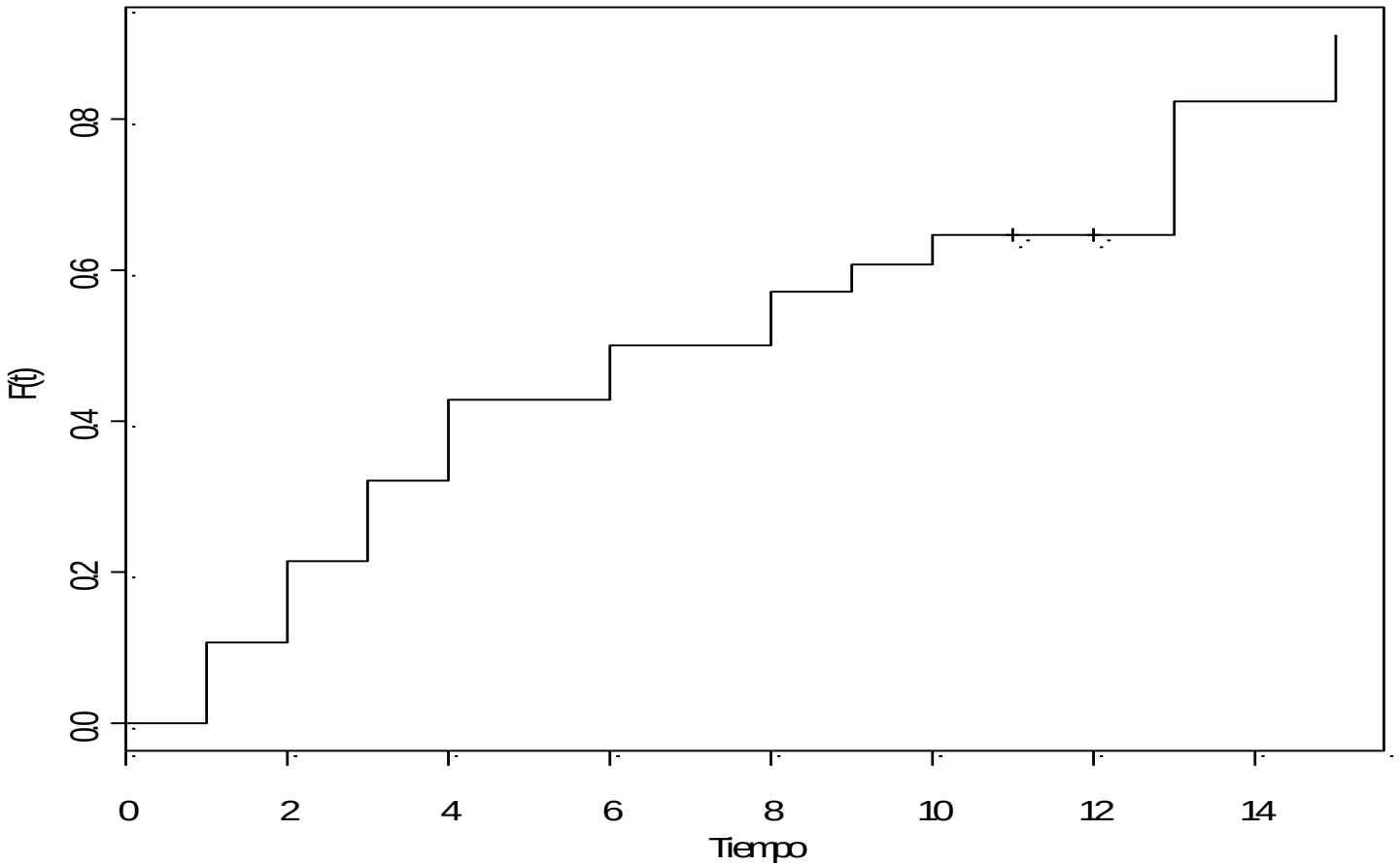
Función de distribución (de la mortalidad): F(t)

- Pr(morir en t o antes) : acumulada
- Ejemplos:
  - Pr(morir a los 65 años o antes)
  - Pr(recidivar a los 3 años o antes)

$$F(t) = Pr(T \leq t)$$

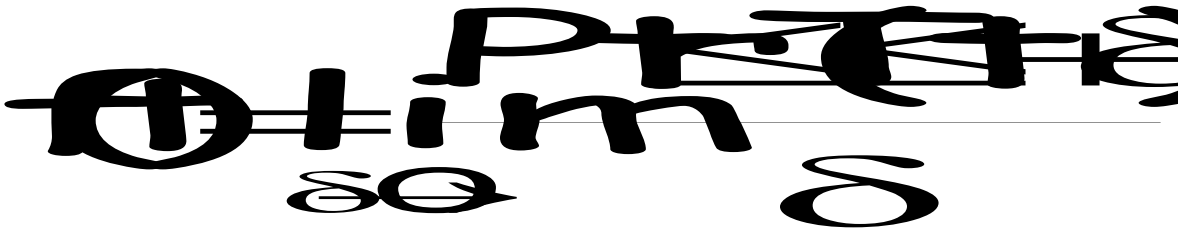
- Es equivalente a S(t): eventos acumulados

Función de distribución: F(t)



Función densidad: f(t)

- Tasa de mortalidad instantánea en t
  - Tiempo en el denominador (δ)
  - $f(t) \times \delta = Pr(\text{morir entre } t \text{ y } t+\delta)$
- Ejemplos:
  - Pr(morir a los 65 años)
  - Pr(tener un reinfarto a los 2 meses del 1º)
- Estimación:



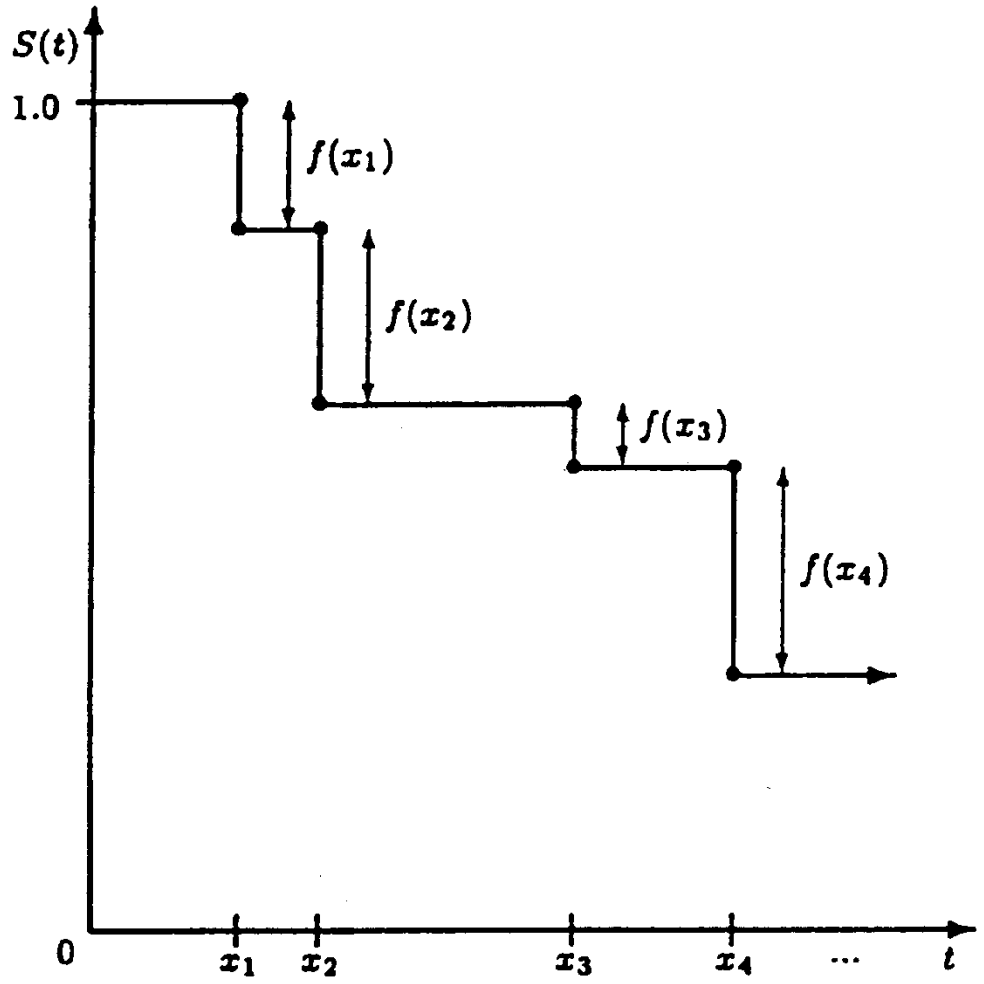
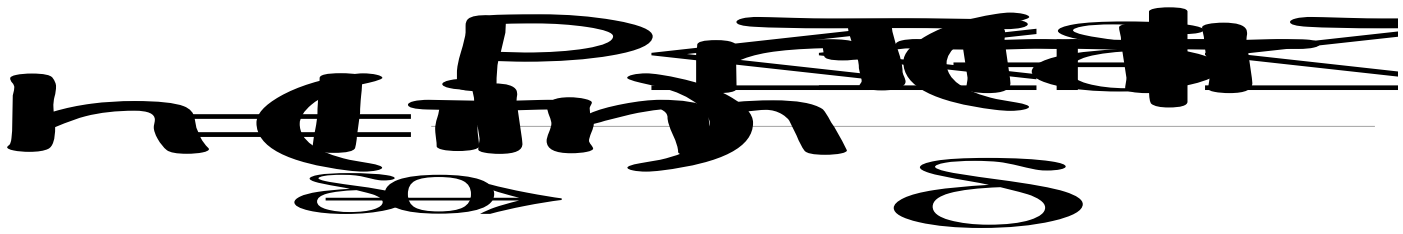


FIGURE 1.7. $S(t)$ is a step function for a discrete T.

Función de riesgo (Hazard)

- Tasa de mortalidad en el momento (t a t+δ) condicional a estar vivo en t



  - Es una <u>tasa</u> de mortalidad instantánea:
    - Tiempo en el denominador (δ)
    - $h(t)\times\delta$ = Pr(morir entre t y t+δ |vivo en t)
- Util para modelar la supervivencia

Riesgo acumulado

- Tasa de mortalidad acumulada

$$H(t) = \int_0^t h(u)\,du$$

- Relacionada directamente con la función de supervivencia:

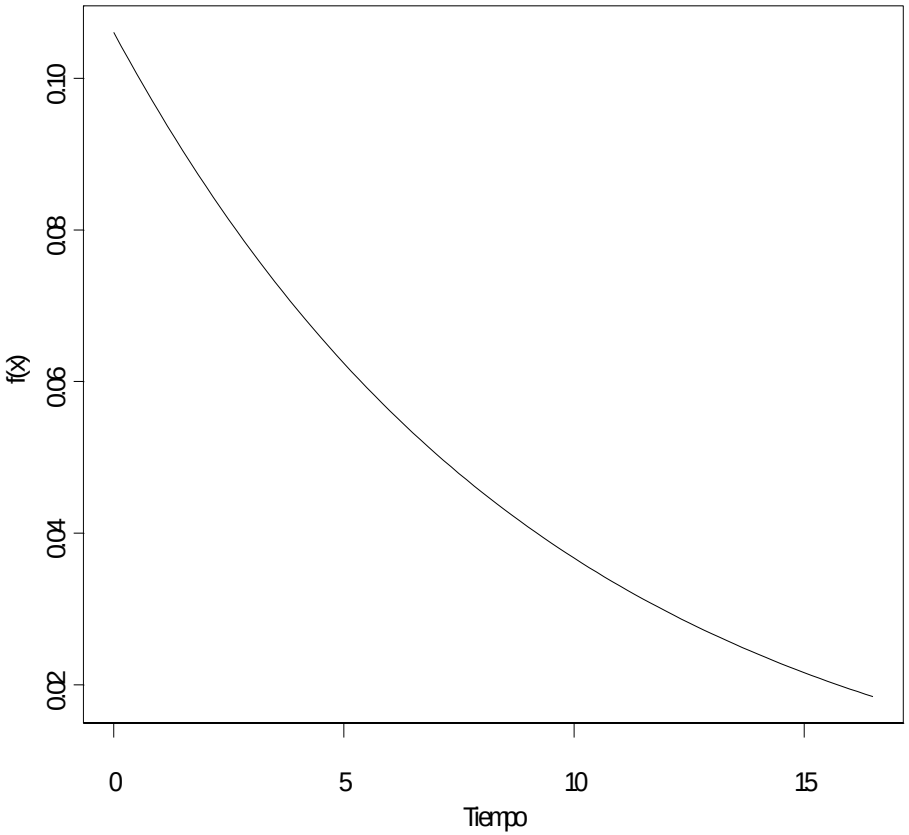
$$S(t) = \exp(-H(t)) \qquad H(t) = -\log(S(t))$$

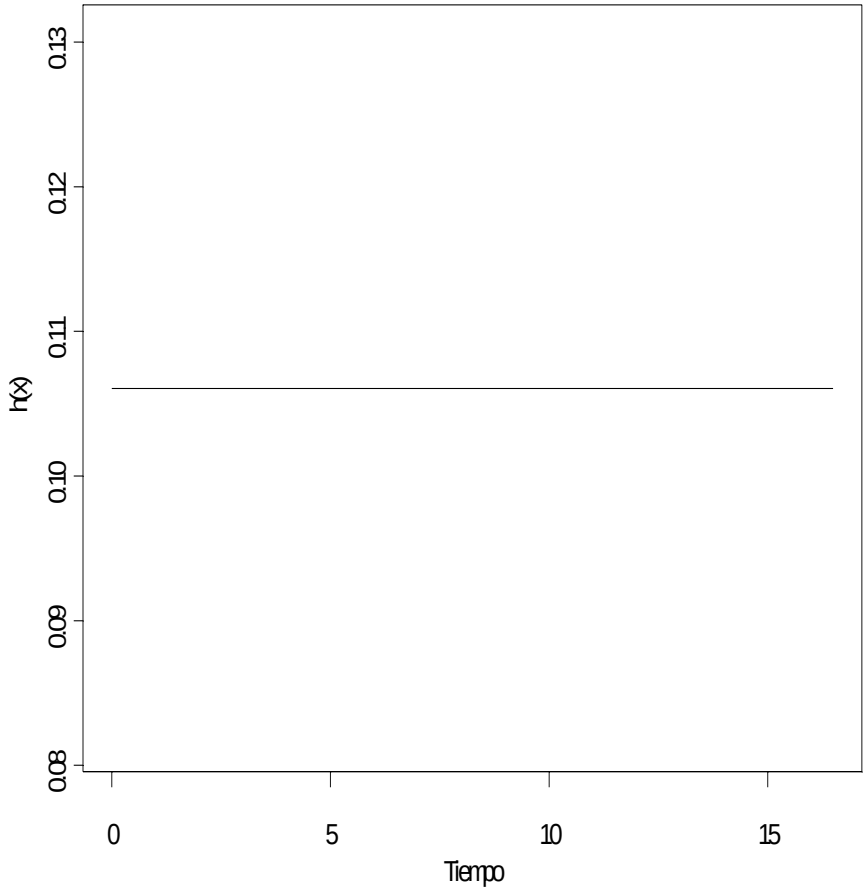
Densidad: f(t) vs riesgo: h(t)

- Las dos son probabilidades instantáneas
- La diferencia es el denominador:
  - densidad: toda la población
  - riesgo: la población viva antes de t

- Ejemplo:
  - f(65): Pr(morir a los 65,00-65,99 años)
  - h(65): Pr(morir a los 65,00-65,99 años | vivo a los 65)
    - es mayor pues el denominador es menor

densidad: f(t) y riesgo: h(t)





Relaciones entre funciones

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

$$= -\frac{d \ln S(t)}{dt}$$

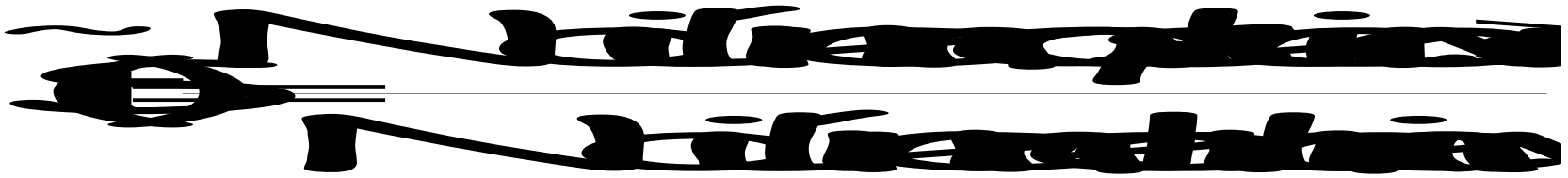
$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$= \exp[-H(t)]$$

Estimación de S(*t*)

- Si no hay censuras

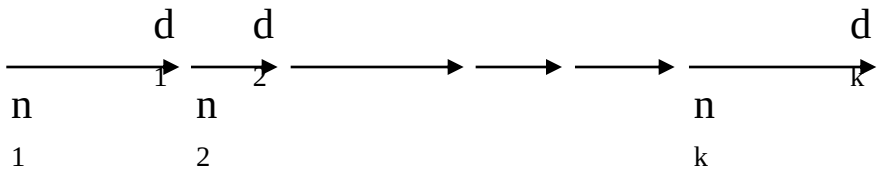
  función de supervivencia empírica:



S(*t*) es una función escalonada. Se mantiene coñstante entre los tiempos de dos muertes adyacentes

Estimador de Kaplan-Meier de S(t)

- Se divide el tiempo en 'k' intervalos de manera que cada intervalo acaba justo cuando un paciente (o varios si hay empates) muere o queda censurado

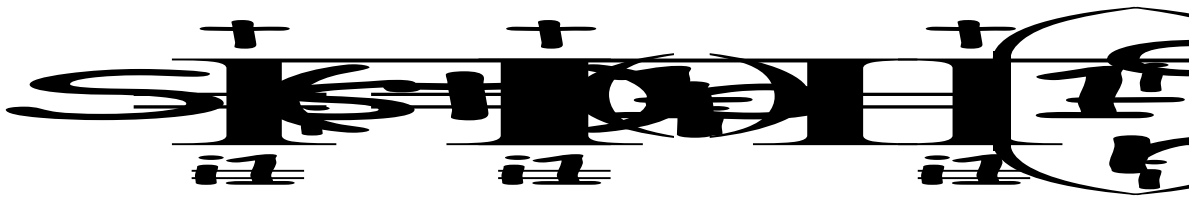


- Para cada intervalo 'i = 1 ··· k':
  - n_i están vivos al inicio
  - d_i mueren al final. d_i suele ser 1, pero varios eventos pueden registrarse en el mismo tiempo por problemas de redondeo o es 0 si censura.

- Probabilidad de morir en el intervalo, $T \in (t_{i-1}, t_i]$, condicional a estar vivo al inicio $p_i = d_i / n_i$
- Probabilidad de sobrevivir al final de intervalo, $T > t_i$, condicional a estar vivo al inicio

$$s_i = 1 - p_i = 1 - d_i / n_i$$

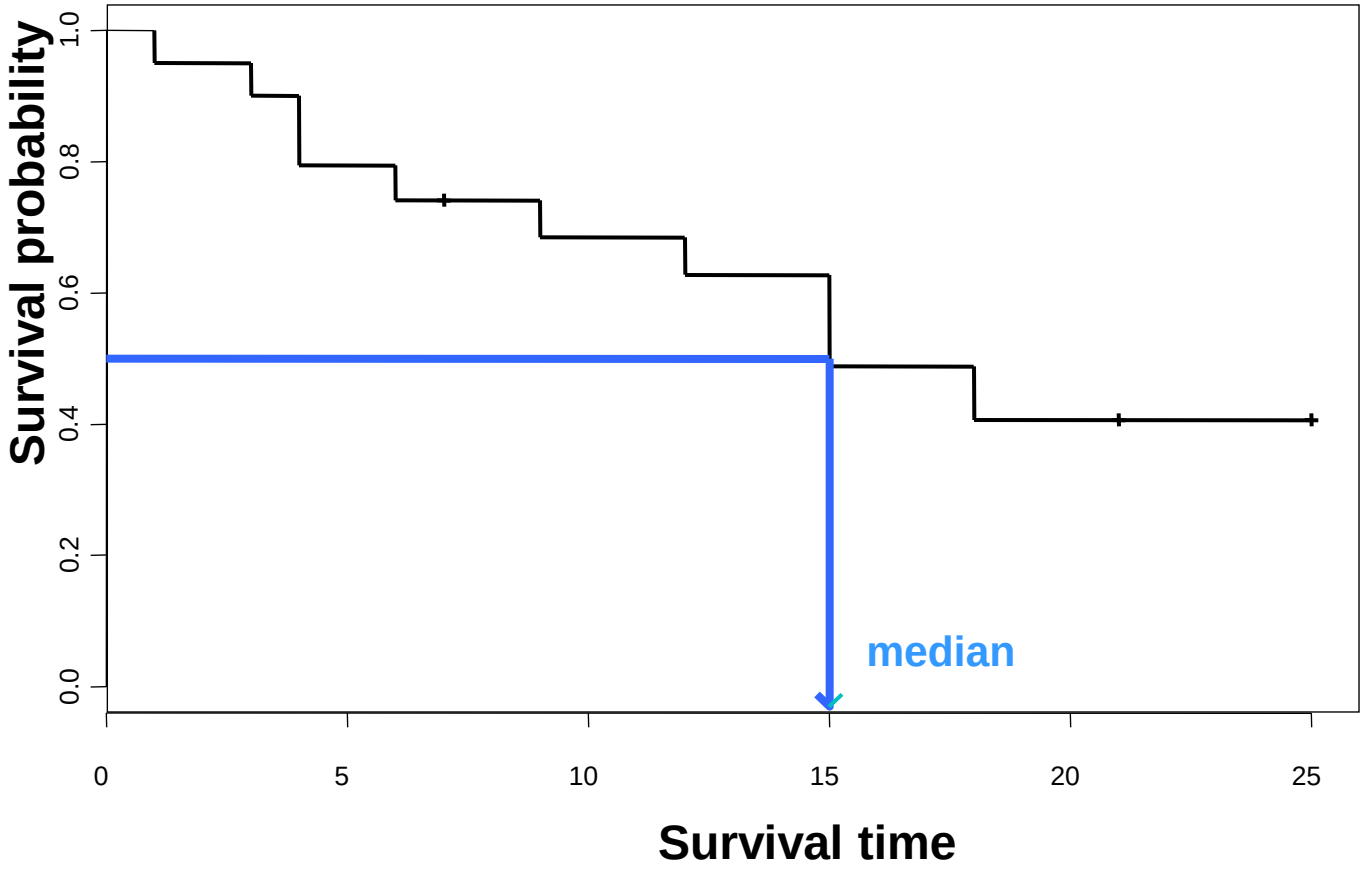
- Como los intervalos son independientes, la probabilidad acumulada de sobrevivir t desde el tiempo 0



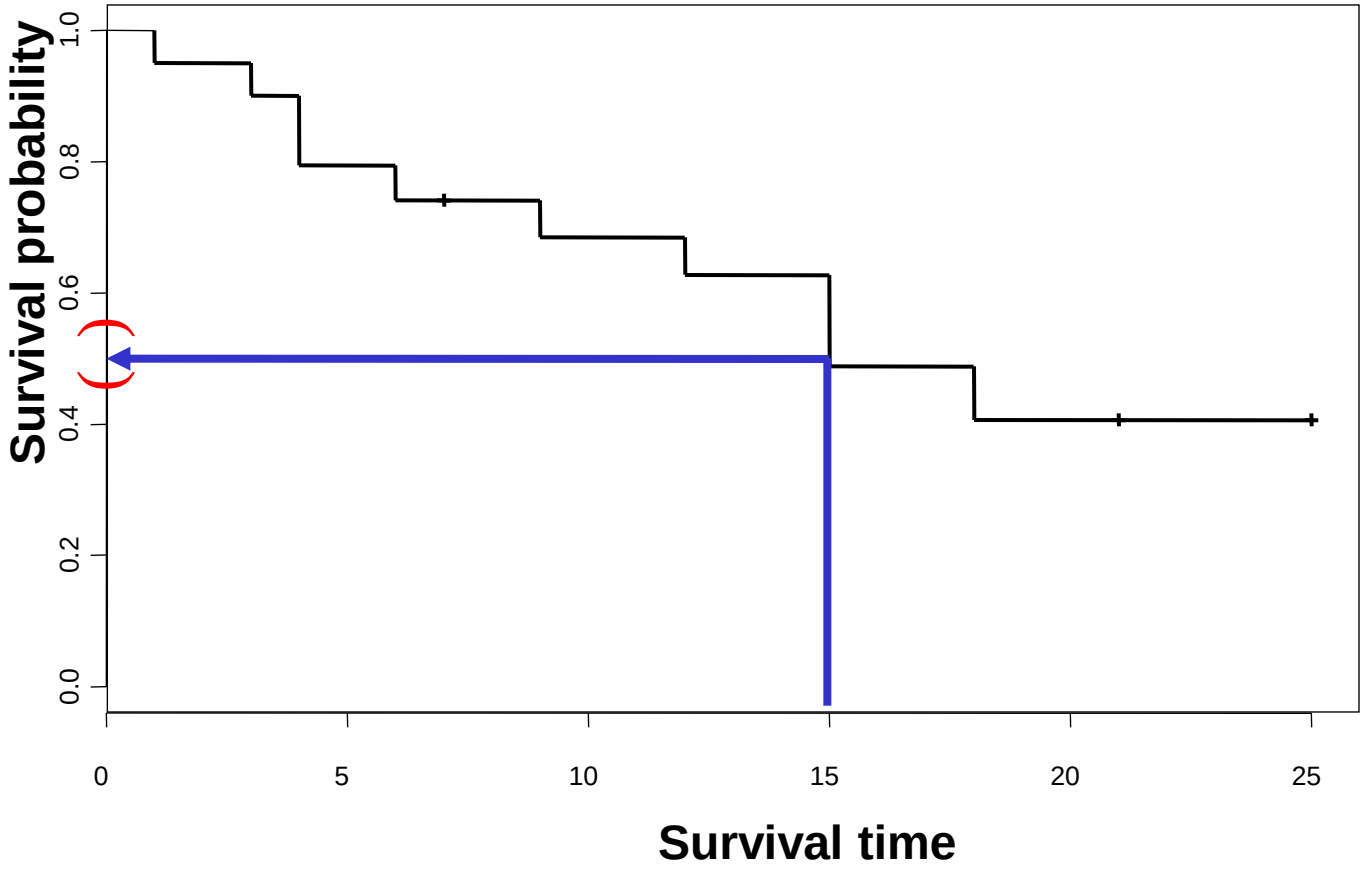
Cálculos de Kaplan-Meier

Time	n_i	d_i	c_i	$1-d_i/n_i = s_i$	$S(t)$
0	20	-	-	-	1.00
1	20	1		$1-1/20 = 0.95$	0.95
3	19	1	1	$1-1/19 = 0.95$	0.90
4	17	2		$1-2/17 = 0.88$	0.79
6	15	1		$1-1/15 = 0.93$	0.74
7	14		1	$1-0/14 = 1.00$	0.74
9	13	1		$1-1/13 = 0.92$	0.68
12	12	1	2	$1-1/12 = 0.92$	0.63
15	9	2	1	$1-2/9 = 0.78$	0.49
18	6	1	2	$1-1/6 = 0.83$	0.41
21	3		1	$1-0/3 = 1.00$	0.41
25	2		2	$1-0/2 = 1.00$	0.41

Tiempo mediano de supervivencia

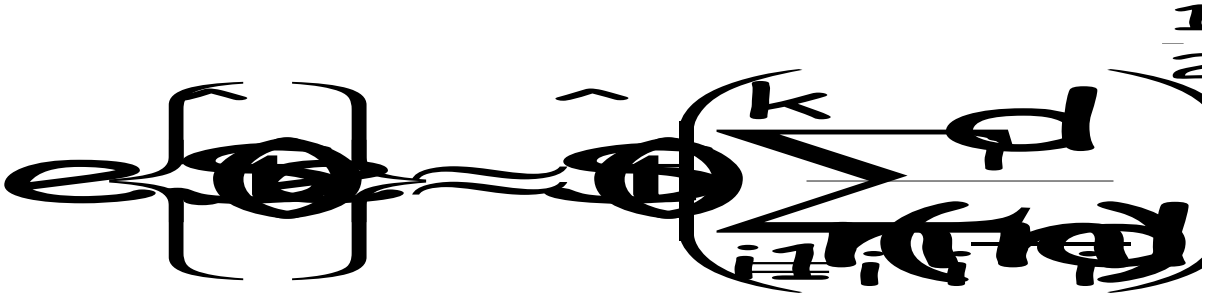


Precisión de S(t)



Precisión de S(*t*)

- El error estándar de S(t) se puede calcular para cada tiempo mediante la fórmula de Greenwood:



- El intervalo de confianza al 95% se calcula de la manera usual:

$$S(t) \pm 1.96 \text{ e.e.}\{S(t)\}$$

IC 95% para S(t)

- Para valores de S(t) cercanos a 1 y 0 el IC podría contener valores no válidos para una probabilidad (<0 ó >1)
- Se debe calcular el e.e. De una transformación de S(t)
  - logaritmo: log(S) ←
  - logit: log{S/(1-S)}
  - log-log: log{-log(S)}

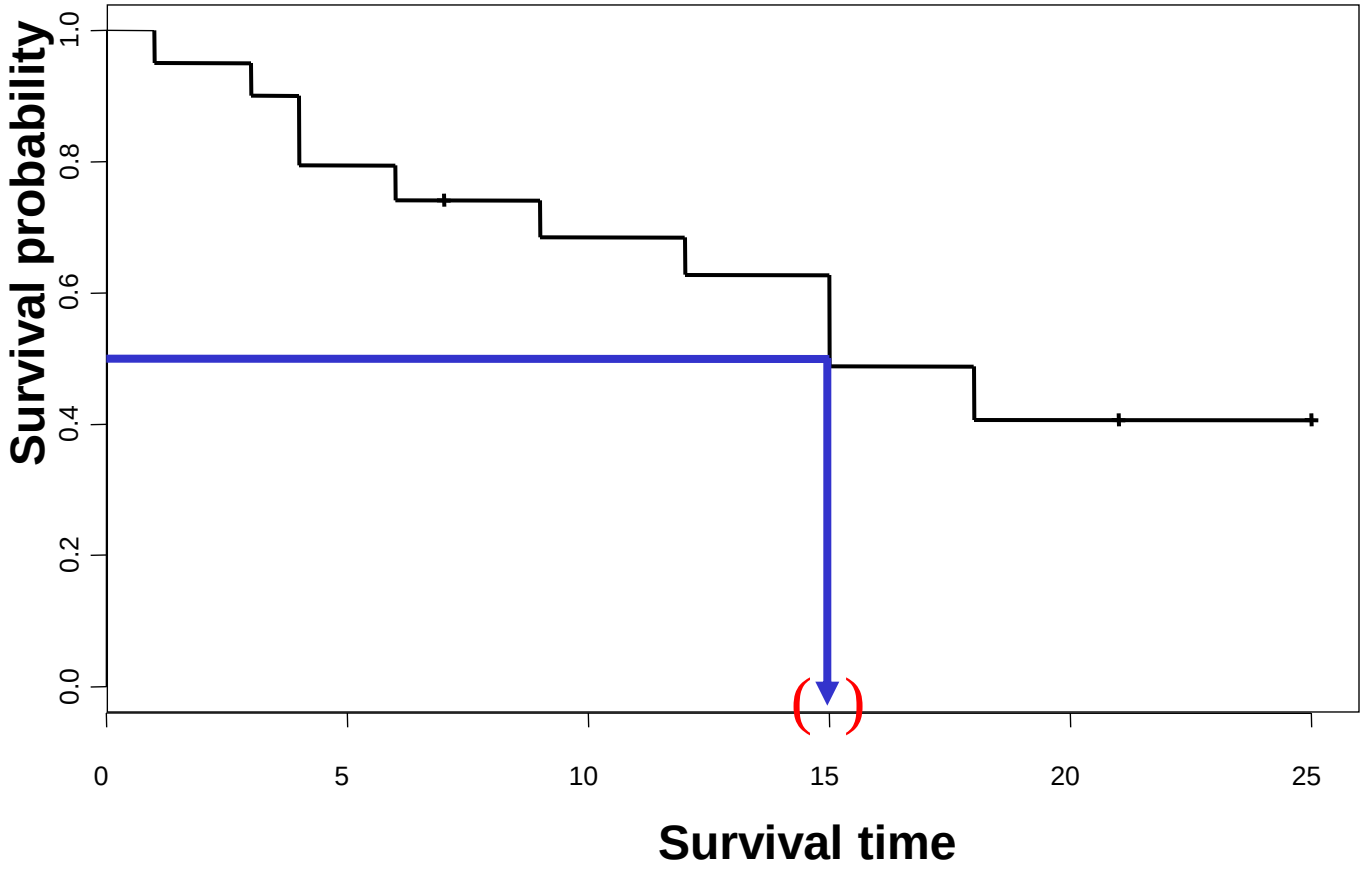
log(S)

- $Var\{\log(S)\} \sim Var(S)/\{S \log(S)\}^2$

$\phi = ee\{\log(S)\} = Var\{\log S\}^{1/2} = ee(S)/\{S\log(S)\}$

$IC\ 95\% = S^{\exp(\pm 1.96\phi)}$

IC para un percentil



Estimación de S(*t*) para datos agrupados. Método de la tabla de vida

- El tiempo se divide en bandas amplias, usualmente de tamaño fijo (3m, 6m, 1a)
- Para cada banda '$i = 1 \cdots k$':
  - n_i están vivos al inicio
  - d_i mueren en la banda
  - c_i son censurados en la banda
- Las observaciones censuradas se supone que se distribuyen de manera uniforme a lo largo de la banda

- Las personas a riesgo se ajustan para tener en cuenta las observaciones censuradas

$$n_i' = n_i - c_i / 2$$

- Probabilidad de morir en la banda, condicional a estar vivo al inicio

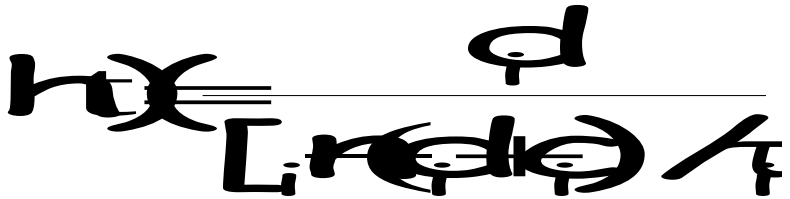
$$p_i = d_i / n_i'$$

- Probabilidad de sobrevivir la banda, condicional a estar vivo al inicio

$$s_i = 1 - p_i = 1 - d_i / n_i'$$

- Como las bandas son independientes, la probabilidad acumulada de sobrevivir t desde el tiempo 0

- El método de tabla de vida permite estimar la función de riesgo $h(t)$, suponiendo que la tasa de mortalidad es constante en la banda



$\forall \tau_i$ es la amplitud de la banda en unidades de tiempo

Método de la tabla de vida

```
. ltable tiempo
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
5	6	23	2	0	0.9130	0.0588	0.6949	0.9775
8	9	21	2	0	0.8261	0.0790	0.6006	0.9309
9	10	19	1	0	0.7826	0.0860	0.5542	0.9032
12	13	18	1	0	0.7391	0.0916	0.5092	0.8734
13	14	17	2	0	0.6522	0.0993	0.4235	0.8084
16	17	15	1	0	0.6087	0.1018	0.3827	0.7737
18	19	14	1	0	0.5652	0.1034	0.3432	0.7376
23	24	13	2	0	0.4783	0.1042	0.2683	0.6613
27	28	11	1	0	0.4348	0.1034	0.2329	0.6212
28	29	10	1	0	0.3913	0.1018	0.1988	0.5798
30	31	9	1	0	0.3478	0.0993	0.1663	0.5371
31	32	8	1	0	0.3043	0.0959	0.1354	0.4928
33	34	7	1	0	0.2609	0.0916	0.1062	0.4469
34	35	6	1	0	0.2174	0.0860	0.0791	0.3993
43	44	5	1	0	0.1739	0.0790	0.0544	0.3495
45	46	4	2	0	0.0870	0.0588	0.0150	0.2417
48	49	2	1	0	0.0435	0.0425	0.0031	0.1824
161	162	1	1	0	0.0000	.	.	.

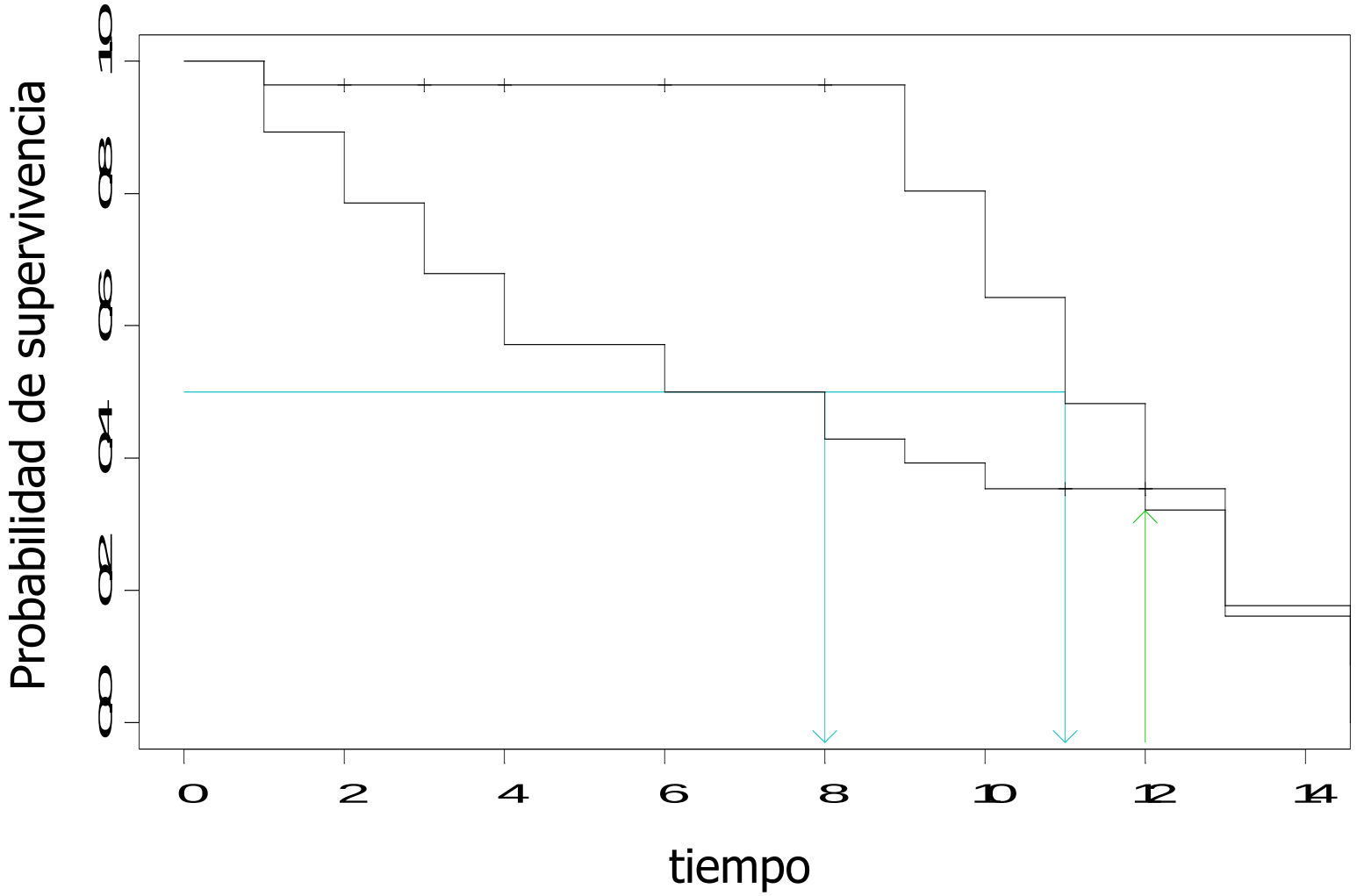
Comparación de grupos

Comparación de grupos

- Comparaciones puntuales
  - Probabilidad de sobrevivir cierto tiempo (supervivencia a 1 ó 3 ó 5 años)
  - Tiempo Mediano de supervivencia u otros percentiles

- Comparación global de la curva
  - Tests no paramétricos
  - Modelos paramétricos o semi-paramétricos

Comparación de grupos



Comparación de 2 grupos

- El tiempo se divide en intervalos de acuerdo con los tiempos de los eventos
- Para cada intervalo se crea una tabla de 2x2

Grupo	Muerto	Vivo	
A	d_{ai}		n_{ai}
B			n_{bi}
	d_i	$n_i - d_i$	n_i

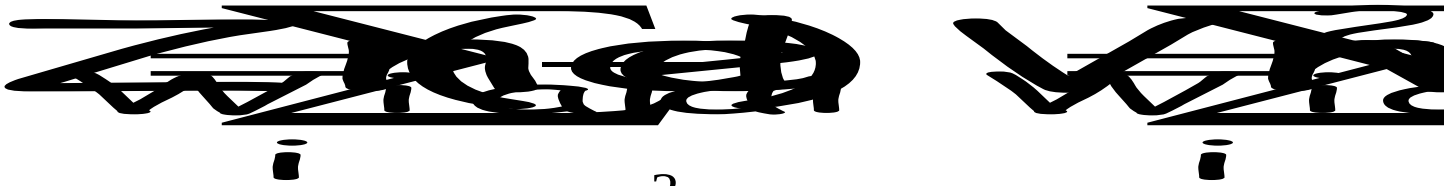
d_{ai} sigue una distribución hipergeométrica

Bajo la hipótesis de independencia, el número esperado de muertes es

Test de Log-rank

$$\frac{Z_L^2}{V_L} \sim \chi_1^2$$

Test de Wilcoxon

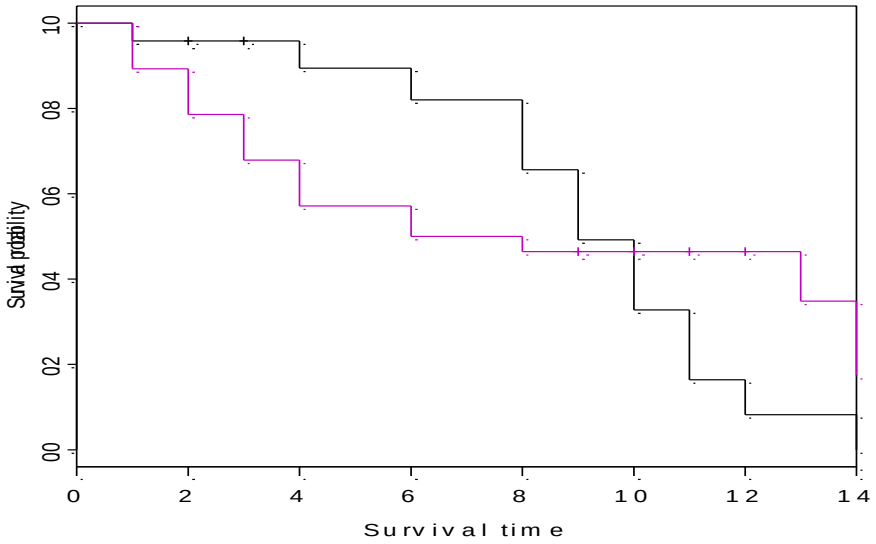


$$\frac{Z_w^2}{V_w} \sim \chi_1^2$$

En general

- Log-rank
  - $w_i = 1$
  - más poder para detectar diferencias al final de la curva
  - Más poder si los riesgos son proporcionales: $\lambda_b = \psi \lambda_a$
- Wilcoxon
  - $w_i = n_i$
  - Más poder para detectar diferencias al inicio de la curva

- Se pueden usar otros pesos
  - Tarone-Ware: $w_i = \sqrt{n_i}$
  - Peto: $w_i = S_i$
- Como todos los tests usan (O-E), ninguno es bueno cuando las curvas se cruzan



Más de 2 grupos

- Z y V se pueden generalizar para la comparación de más de 2 grupos (g)

Grupo	Muerto	Vivo	
A	d_{ai}		n_{ai}
B	d_{bi}		n_{bi}
C	d_{bi}		n_{ci}
D			n_{di}
	d_i	n_i-d_i	n_i

$$Z_k = \sum_i w_i (d_{ki} - e_{ki})$$

$$k = 1 + g - 1$$

V: matriz de varianza-covarianza

$$Z V^{-1} Z \sim \chi^2_{g-1}$$

Test de tendencia

- Cuando los grupos están definidos por una variable ordinal: - Categorías de edad
  - Grupos de dosis
  - Estadío tumoral

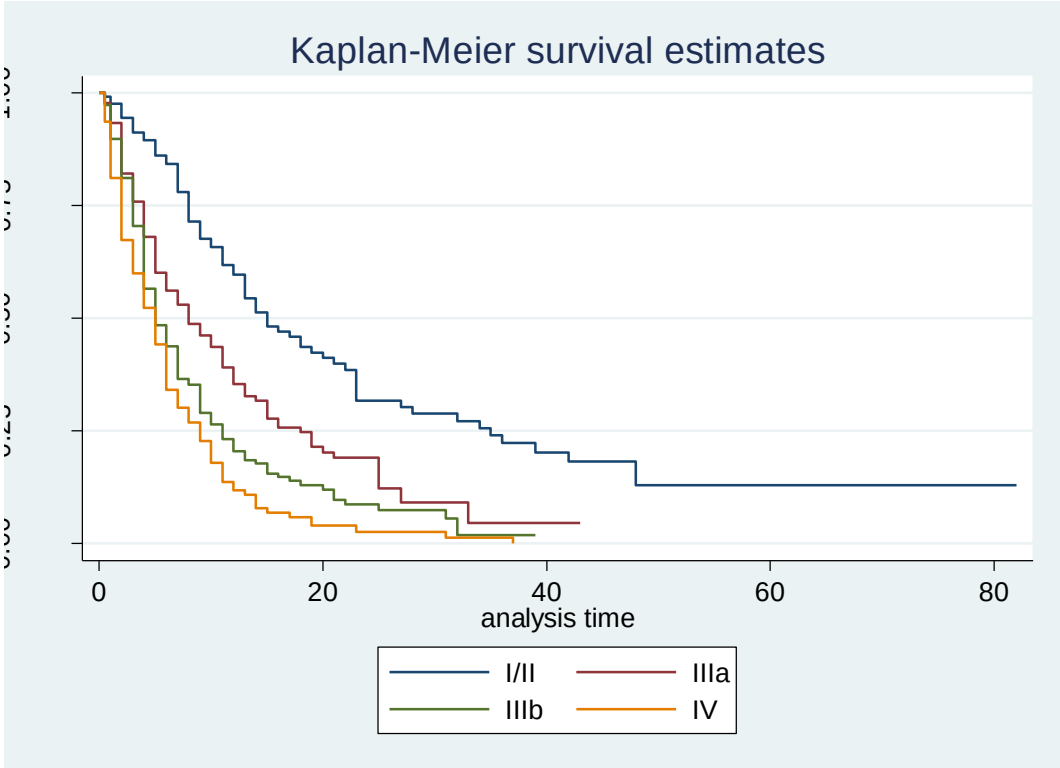
- Similar al test de Mantel-Haenszel para tendencias en proporciones

w_k codifica la métrica de la tendencia:

1 2 3 4 para tendencia lineal

$Z^2/V \sim \chi^2_1$ Solo 1 grado de libertad

Test de tendencia



Test de tendencia

Tendencia

Log-rank test for equality of survivor functions

estclin_num	Events observed	Events expected
I/II	79	154.30
IIIa	106	115.42
IIIb	155	123.12
IV	134	81.15
Total	474	474.00

chi2(3) = 94.80
Pr>chi2 = 0.0000

Test for trend of survivor functions

chi2(1) = 94.65
Pr>chi2 = 0.0000

Asociación

Log-rank test for equality of survivor functions

estclin	Events observed	Events expected
I/II	79	154.30
IIIa	106	115.42
IIIb	155	123.12
IV	134	81.15
Total	474	474.00

chi2(3) = 94.80
Pr>chi2 = 0.0000

Test estratificado

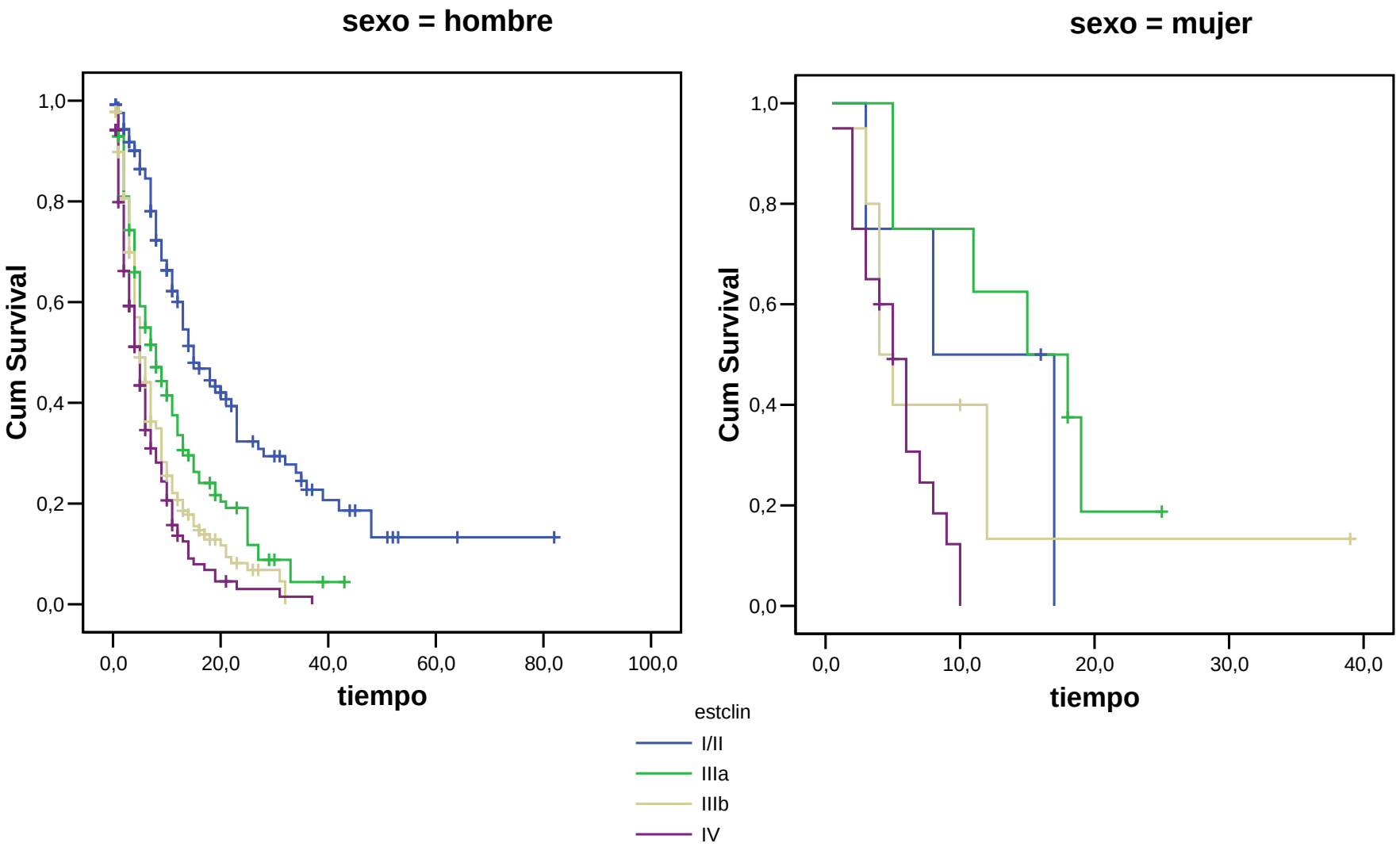
- Ajuste de factores de confusión mediante un test no paramétrico
- Se comparan grupos controlando el efecto de un atercera variable. La comparación entre grupos se realiza dentro de cada categoría (estrato) de la variable confusora.
- Z_k y V_k se calculan para cada estrato y después se combinan.

  Para 2 grupos, k estratos:

$$\frac{\sum_k Z_k}{\sum_k V_k} \sim \chi^2_1$$

- Se puede generalizar a más de 2 grupos

Test estratificado



Test estratificado

Stratified log-rank test for equality of survivor functions

estclin	Events observed	Events expected(*)
I/II	79	153.62
IIIa	106	116.48
IIIb	155	122.46
IV	134	81.43
Total	474	474.00

(*) sum over calculations within sexo

chi2(3) = 95.87
Pr>chi2 = 0.0000