Estimación de la función de supervivencia

Giancarlo Sal y Rosas, Ph.D.

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

31 de marzo de 2017

Outline

Estructura de datos

- Sean Y el tiempo de censura y T el tiempo a la ocurrencia del evento.
- Sean F y f la función acumulada de distribución y la función de densidad de T, respectivamente.
- Sea G y g la función acumulada de distribución y la función de densidad de Y, respectivamente.
- Los datos observados tienen la forma

$$(X, \Delta) = (T \wedge Y, I(T \leq Y))$$

- **Objetivo**: Estimar F

Estructura de los datos

- Estudio aleatorizado para estimar el tiempo de vida de pacientes con cancer de ovario bajo dos tratamientos
- Datos

```
> head(ovarian)
  futime fustat      age resid.ds rx ecog.ps
1     59      1 72.3315        2  1       1
2    115      1 74.4932        2  1       1
3    156      1 66.4658        2  1       2
4    421      0 53.3644        2  2       1
5    431      1 50.3397        2  1       1
6    448      0 56.4301        1  1       2
```

donde

  - futime: Tiempo de vida
  - fustat: El evento (muerte) ocurrio o no
  - rx: El tratamiento asignado

Estimación

- ¿Cuál es la probabilidad de vivir mas de 353 dias para estas personas ?
- **Opción A**: Ignoramos los datos censurados

```
> dat <- ovarian[ovarian$fustat==1,]
> dat <- dat[order(dat$futime,decreasing=F),]
> dat$futime
 [1]  59 115 156 268 329 353 365 431 464 475 563 638
> mean(dat$futime>353)
[1] 0.5
```

- **Opción B**: Usamos toda la data pero ignoramos la censura

```
> ovarian <- ovarian[order(ovarian$futime,decreasing=F),]
> ovarian$futime
 [1]   59  115  156  268  329  353  365  377  421  431  448  464  475  477  563
638  744  769
[19]  770  803  855 1040 1106 1129 1206 1227
> mean(ovarian$futime>353)
[1] 0.7692308
```

Estimación

```
> Surv(ovarian$futime,ovarian$fustat)
 [1]   59   115   156   268   329   353   365   377+  421+  431   448+  464   475   477+
563   638
[17]  744+  769+  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+
> tbl
   time n.risk n.event
1    59     26       1
2   115     25       1
3   156     24       1
4   268     23       1
5   329     22       1
6   353     21       1
7   365     20       1
8   431     17       1
9   464     15       1
10  475     14       1
11  563     12       1
12  638     11       1
```

- Para $t \in [0, 59)$, se tiene que

$$\hat{\lambda}_n(t) = 0 \quad , \quad \hat{S}_n(t) = 1$$

Estimación

- Para $t = 59$, se tiene que

$$\hat{\lambda}_n(t) = \tfrac{1}{26} \quad , \quad \hat{S}_n(t) = 1 \times \left(1 - \frac{1}{26}\right) = 0.962$$

- Para $t \in [115, 156)$, se tiene que

$$\hat{\lambda}_n(t) = \tfrac{1}{25} \quad , \quad \hat{S}_n(t) = 1 \times \left(1 - \frac{1}{26}\right)\left(1 - \frac{1}{25}\right) = 0.923$$

- Para $t \in [365, 431)$, se tiene que $\hat{\lambda}_n(t) = 1/20$ y

$$\hat{S}_n(t) = 1 \times \left(1 - \frac{1}{26}\right)\left(1 - \frac{1}{25}\right) \times \cdots \times \left(1 - \frac{1}{20}\right) = 0.731$$

Estimación

- Para $t \in [431, 464)$, se tiene que

$$\hat{\lambda}_n(t) = \frac{1}{17}$$

  - Entrando a este intervalo se **censuraron** (perdimos) 2 observaciones: 377 y 421 y ocurrio un evento (365)
  - El número de observaciones **en riesgo** es $19 - 2 = 17$

- La función de supervivencia para $t \in [431, 464)$, es

$$\hat{S}_n(t) = 1 \times \left(1 - \frac{1}{26}\right)\left(1 - \frac{1}{25}\right) \times \cdots \times \left(1 - \frac{1}{17}\right) = 0.688$$

Estimación

- En general

	time	n.risk	n.event	surv
1	59	26	1	0.9615385
2	115	25	1	0.9230769
3	156	24	1	0.8846154
4	268	23	1	0.8461538
5	329	22	1	0.8076923
6	353	21	1	0.7692308
7	365	20	1	0.7307692
8	431	17	1	0.6877828
9	464	15	1	0.6419306
10	475	14	1	0.5960784
11	563	12	1	0.5464052
12	638	11	1	0.4967320

donde la columna **time** se refiere, tacitamente a un intervalo. Por ejemplo, la novena fila corresponde al intervalo $[464, 475)$

- Note que al final del estudio se tiene

$$\hat{S}_n(t) = 0.497 \ , \ t \in [638, \infty)$$

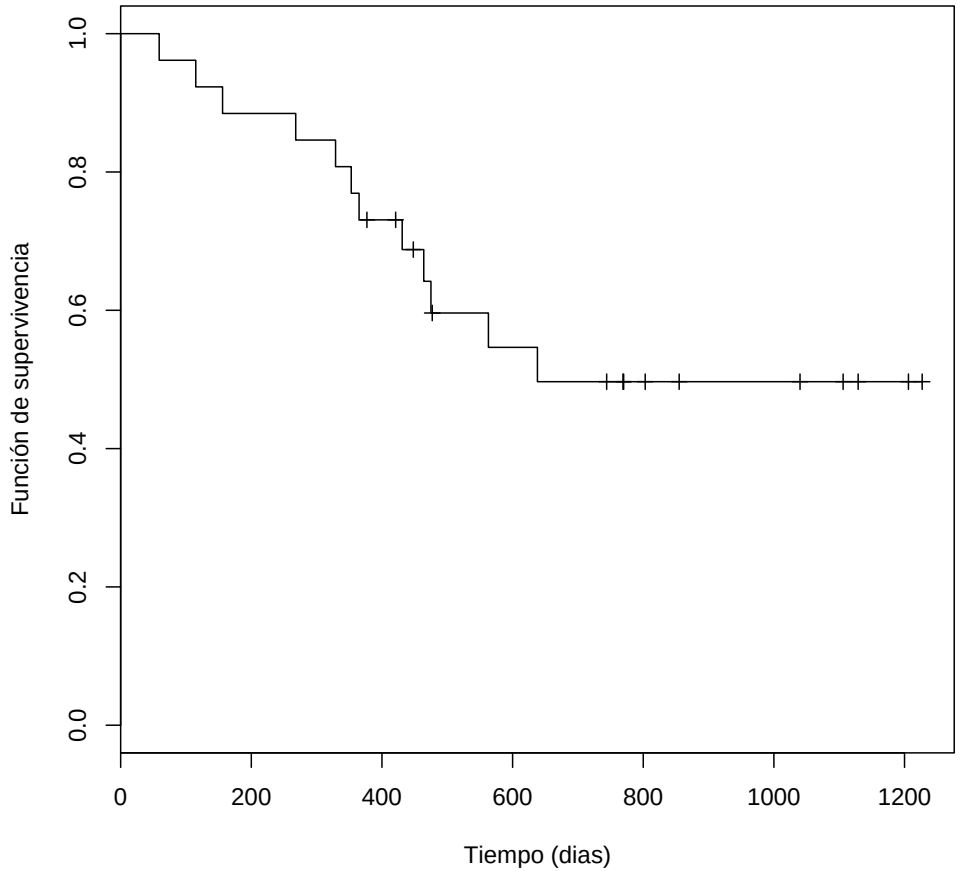

Figura 1 : Estimador de Kaplan-Meier: Cancer de ovario

Estimador de Kaplan-Meier

- Kaplan y Meier formalizaron esta técnica [1]
- Sea $t_1 < t_2 < \cdots < t_k$ los tiempos (dentro de la muestra) donde al menos un evento es observado.
- En base a estos tiempos, particionamos la muestra en $k + 1$ intervalos:

$$[t_0, t_1) \quad , \quad [t_1, t_2) \quad \ldots \quad [t_k, t_{k+1})$$

  donde $t_0 = 0$ y $t_{k+1} = \infty$.
- Dentro de cada intervalo $[t_j, t_{j+1})$, se observan los tiempos de censura $t_{j1}, t_{j2}, \ldots, t_{jm_j}$

Estimador de Kaplan-Meier

```
> # Ordenar el tiempo de manera creciente
> ovarian <- ovarian[order(ovarian$futime,decreasing=F),]
>
> # Poner la estructura de censura
> Surv(ovarian$futime,ovarian$fustat)
 [1]   59   115   156   268   329   353   365   377+  421+  431   448+
[12]  464   475   477+  563   638   744+  769+  770+  803+  855+  1040+
[23] 1106+ 1129+ 1206+ 1227+
```

- Los tiempos en que se dieron los eventos son: 59, 115, 156, 268, 329, 353, 365, 431, 464, 475, 563 y 638
- Los intervalos a generarse son:

$$[0, 59), [59, 115), \ldots, [638, \infty)$$

- Dentro del primer intervalo no hay datos (tiempos) censurados
- Dentro del último intervalo hay 10 datos (tiempos) censurados

Estimador de Kaplan-Meier

- Definamos
  - $d_j = \#$ de personas que sufren el evento en el tiempo t_j
  - $m_j = \#$ de personas censuradas en el intervalo $[t_j, t_{j+1})$
  - $n_j = \#$ de persona en riesgo en el instante previo a t_j

$$n_j = (m_j + d_j) + (m_{j+1} + d_{j+1}) + \cdots + (m_k + d_k)$$

- En el intervalo $[t_j, t_{j+1})$ tenemos d_j eventos y m_j censuras que ocurrieron en los tiempos t_i y $t_{j1}, t_{j2}, \ldots, t_{jm_j}$, respectivamente.

Estimador de Kaplan-Meier

Podemos pensar que los datos han sido particionados en $k+1$ intervalos:

$[0, t_1)$	$[t_1, t_2)$	$[t_2, t_3)$	$\ldots$
0	d_1	d_2	$\ldots$
m_0	m_1	m_2	$\ldots$
n	$n_1 = n - m_0$	$n_2 = n_1 - d_1 - m_1$	$\ldots$

$\ldots$	$[t_{k-1}, t_k)$	$[t_k, \infty)$
$\ldots$	d_{k-1}	d_k
$\ldots$	m_{k-1}	m_k
$\ldots$	$n_{k-2} - d_{k-2} - m_{k-2}$	$n_{k-1} - d_{k-1} - m_{k-1}$

Estimador de Kaplan-Meier

- La verosimilitud asociada para el intervalo $[t_j, t_{j+1})$ es

$$[F(t_j) - F(t_j-)]^{d_j} \prod_{l=1}^{m_j} [1 - F(t_{jl})]$$

donde $f(t_j) = P(T = t_j) = F(t_j) - F(t_j-)$

- La verosimilitud de toda la muestra es

$$L_n = \prod_{j=0}^{k} \left\{ [F(t_j) - F(t_j-)]^{d_j} \prod_{l=1}^{m_j} [1 - F(t_{jl})] \right\}$$

Estimador de Kaplan-Meier

- Asumamos que $S(t_{jl}) = S(t_j)$. Es decir que las censuras ocurrieron al inicio del intervalo:

$$
\begin{aligned}
L_n &= \prod_{j=0}^{k} \left\{ [F(t_j) - F(t_j-)]^{d_j} \prod_{l=1}^{m_j} [1 - F(t_j)] \right\} \\
&= \prod_{j=0}^{k} \left\{ \left[\lambda_j \prod_{l=1}^{j-1} (1 - \lambda_l) \right]^{d_j} \prod_{l=1}^{m_j} \prod_{s=1}^{j} (1 - \lambda_s) \right\} \\
&= \prod_{j=0}^{k} \left\{ \lambda_j^{d_j} \prod_{l=1}^{j-1} (1 - \lambda_l)^{d_j} \prod_{s=1}^{j} (1 - \lambda_l)^{m_j} \right\}
\end{aligned}
$$

Estimador de Kaplan-Meier

- Podemos reordenar la verosimilitud en

$$L_n = \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j}$$

donde n_j es el número de personas en riesgo al inicio del intervalo

- El estimador de máxima verosimilitud de λ_j es

$$\hat{\lambda}_j = d_j / n_j$$

- El estimador de Kaplan-Meier es

$$\hat{S}^{KM}(t) = \prod_{j|t_j \leq t} \left(1 - \hat{\lambda}_i\right) = \prod_{j|t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Varianza

- El estimador de la varianza de $\hat{S}^{KM}(t)$ es

$$\hat{Var}(\hat{S}^{KM}(t)) = \hat{S}^{KM}(t)^2 \sum_{j|t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

- El error estandar es la raiz cuadrada de la varianza
- Note que si la ultima observación $(t_{(n)})$ es un evento, entonces

$$\hat{S}_n(t_{(n)}) = 0$$

y nuestra estimación de la varianza es **cero**.

- En general, la varianza tiende a aumentar en el tiempo

Error éstandar

- Resultados en R

	time	n.risk	n.event	surv	std.err
1	59	26	1	0.9615385	0.03771464
2	115	25	1	0.9230769	0.05225894
3	156	24	1	0.8846154	0.06265627
4	268	23	1	0.8461538	0.07075894
5	329	22	1	0.8076923	0.07729201
6	353	21	1	0.7692308	0.08262864
7	365	20	1	0.7307692	0.08698929
8	431	17	1	0.6877828	0.09188148
9	464	15	1	0.6419306	0.09652130
10	475	14	1	0.5960784	0.09992615
11	563	12	1	0.5464052	0.10320939
12	638	11	1	0.4967320	0.10510266

- Interpretación
  - La probabilidad estimada de vivir mas de 115 dias es 0.88 pero esta estimación tiene una variabilidad asociada de 0.06
  - La probabilidad estimada de vivir mas de 563 dias es 0.55 pero esta estimación tiene una variabilidad asociada de 0.10

Intervalo de confianza

- Un intervalo de confianza al 95 % para $S(t)$ es

$$\hat{S}^{KM}(t)) \pm 1.96 \times \sqrt{\hat{Var}(\hat{S}^{KM}(t))}$$

donde 1.96 es el percentil 97.5 de la distribución normal estandar.

- En general, un intervalo de confianza al $100(1-\alpha)$ % para $S(t)$ es

$$\hat{S}^{KM}(t)) \pm z_{1-\alpha/2} \times \sqrt{\hat{Var}(\hat{S}^{KM}(t))}$$

donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha)$ de la distribución normal estandar.

Intervalo de confianza

- Resultados en R

```
> model0 <- summary(survfit(Surv(futime,fustat)~1,data=ovarian))
> model0
Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian)

 time n.risk n.event survival std.err lower 95\% CI upper 95\% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

- **Interpretación**: La probabilidad estimada de vivir mas de 431, en esta población, es 0.69. Adicionalmente tenemos un 95 % de confianza que el valor real se encuentra entre 0.53 y 0.89

Intervalo de confianza

- **Precaución**:
  - Para el primer intervalo, su limite superior deberia ser (segun la formula!) de la forma

$$0.96 + 1.96 \times 0.04 = 1.04$$

    en estos casos, el programa lo acota a 1.
  - Este mismo fenomeno se presenta en el segundo y tercer intervalo

Intervalo de confianza

- Opción *conf.type* = "*plain*"

```
> summary(survfit(Surv(futime,fustat)~1,data=ovarian,conf.type="plain"))
Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian, conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.888        1.000
  115     25       1    0.923  0.0523        0.821        1.000
  156     24       1    0.885  0.0627        0.762        1.000
```

- Opción *conf.type* = "*log*" (por defecto)

```
> summary(survfit(Surv(futime,fustat)~1,data=ovarian,conf.type="log"))
Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian, conf.type = "log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
```

- Opción *conf.type* = "*log* − *log*"

```
> summary(survfit(Surv(futime,fustat)~1,data=ovarian,conf.type="log-log"))
Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian, conf.type = "log-log"

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.757        0.994
  115     25       1    0.923  0.0523        0.726        0.980
  156     24       1    0.885  0.0627        0.684        0.961
```

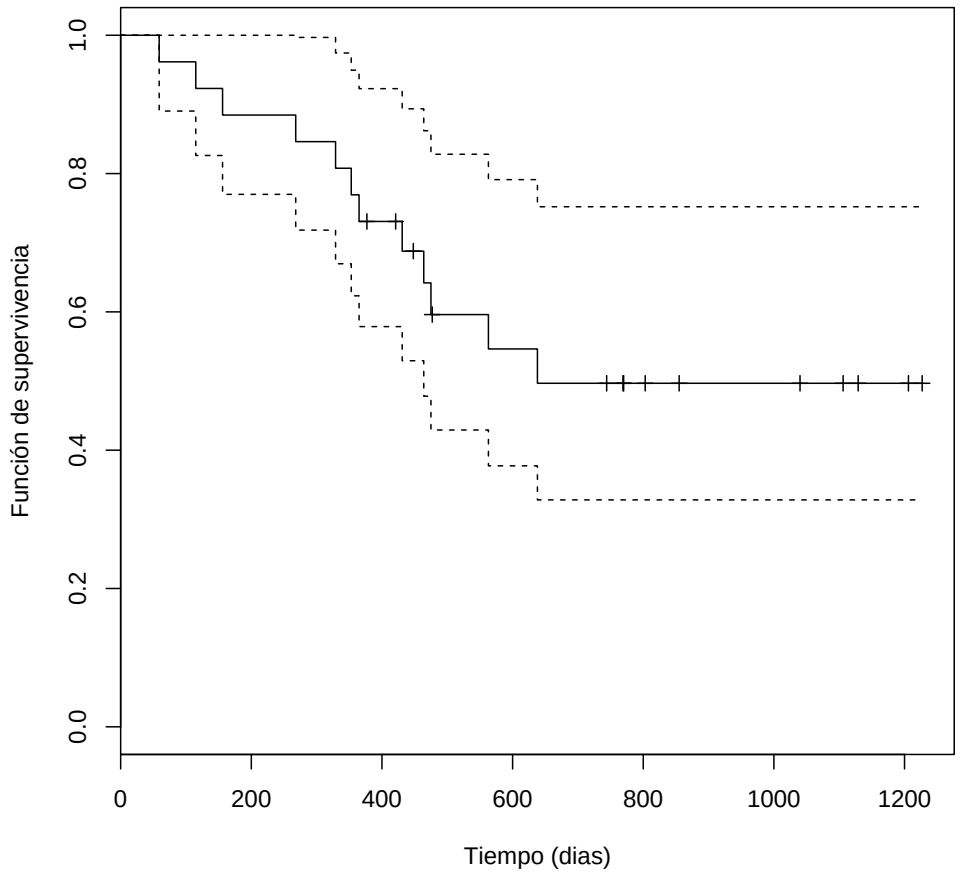

Figura 2 : Estimador de Kaplan-Meier e intervalos de confianza: Cancer de ovario

Transplante de corazón

- La base de datos **lung** esta disponible en la libreria survival
- Estudia el tiempo de vida de pacientes con cancer avanzado de pulmon
- Datos

```
> 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
```

Funciones **Surv** y **survfit**

```
> Surv(lung$time,lung$status)[1:40]
 [1]   5  11  11  11  12  13  13  15  26  30  31  53  53  54  59  60
60
[18]  61  62  65  65  71  79  81  81  88  88  92  92+ 93  95  95  105 105+
[35] 107 107 110 116 118 122

> summary(survfit(Surv(time,status)~1,data=lung))
Call: survfit(formula = Surv(time, status) ~ 1, data = lung)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    228       1   0.9956 0.00438        0.9871        1.000
   11    227       3   0.9825 0.00869        0.9656        1.000
   12    224       1   0.9781 0.00970        0.9592        0.997
   13    223       2   0.9693 0.01142        0.9472        0.992
   15    221       1   0.9649 0.01219        0.9413        0.989
   26    220       1   0.9605 0.01290        0.9356        0.986
   30    219       1   0.9561 0.01356        0.9299        0.983
   31    218       1   0.9518 0.01419        0.9243        0.980
   53    217       2   0.9430 0.01536        0.9134        0.974
   54    215       1   0.9386 0.01590        0.9079        0.970
   59    214       1   0.9342 0.01642        0.9026        0.967
   60    213       2   0.9254 0.01740        0.8920        0.960
   61    211       1   0.9211 0.01786        0.8867        0.957
   62    210       1   0.9167 0.01830        0.8815        0.953
   65    209       2   0.9079 0.01915        0.8711        0.946
```

Interpretación

- Dos cientos veinte y ocho pacientes iniciaron el estudio
- La probabilidad estimada de sobrevivir 65 dias es 0.91
- El valor estimado del error éstandar de nuestra estimación es 0.02
- Tengo un 95 % de confianza que el valor real de la supervivencia, a los 65 dias, esta entre 0.87 y 0.95

Estimador de Kaplan-Meier

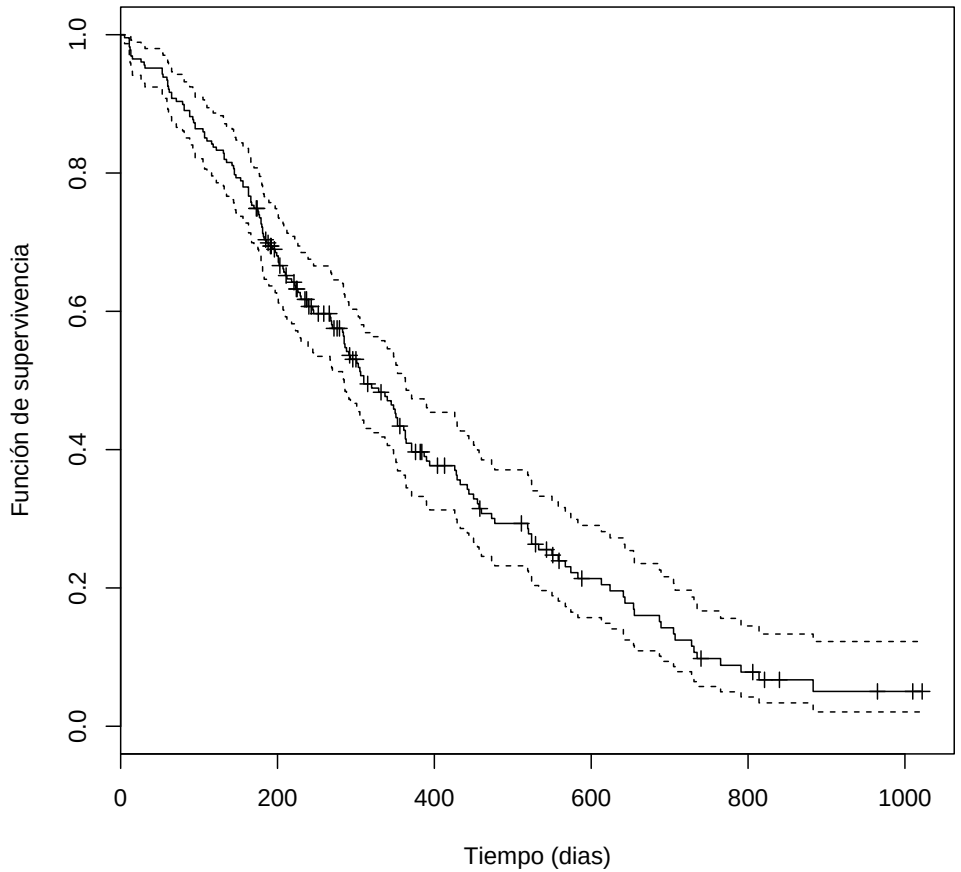

Figura 3 : Estimador de Kaplan-Meier

Mediana e intervalo de confianza

- La mediana del tiempode sobrevida es

$$t_{(50)} = inf\{m : S(m) \leq 0.5\}$$

- Los percentiles 25 y 75 del tiempo de sobrevida son

$$t_{(25)} = inf\{m : S(m) \leq 0.75\}$$

$$t_{(75)} = inf\{m : S(m) \leq 0.25\}$$

- En general, puedo calcular cualquier percentil

$$t_{(q)} = inf\left\{m : S(m) \leq 1 - \frac{q}{100}\right\}$$

Pacientes cáncer de ovario

- Datos

```
> summary(survfit(Surv(futime,fustat)~1,data=ovarian))
Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

- En este conjunto solo un tiempo tiene una supervivencia menor a 0.5 ($t = 638$) y en consecuencia este es la mediana

$$t_{(50)} = 638$$

Percentiles

- Percentil 25: Los tiempos con supervivencia menor o igual a 0.75 son 365, 431, 464, 475, 563 y 638. Entonces

$$t_{(25)} = 365$$

- Percentil 75: No tenemos tienes con supervivencia menor a 0.25. Entonces

$$t_{(75)} = \infty$$

- El rango interquartil del tiempo de vida es (365,∞)

Mediana: Intervalo de confianza

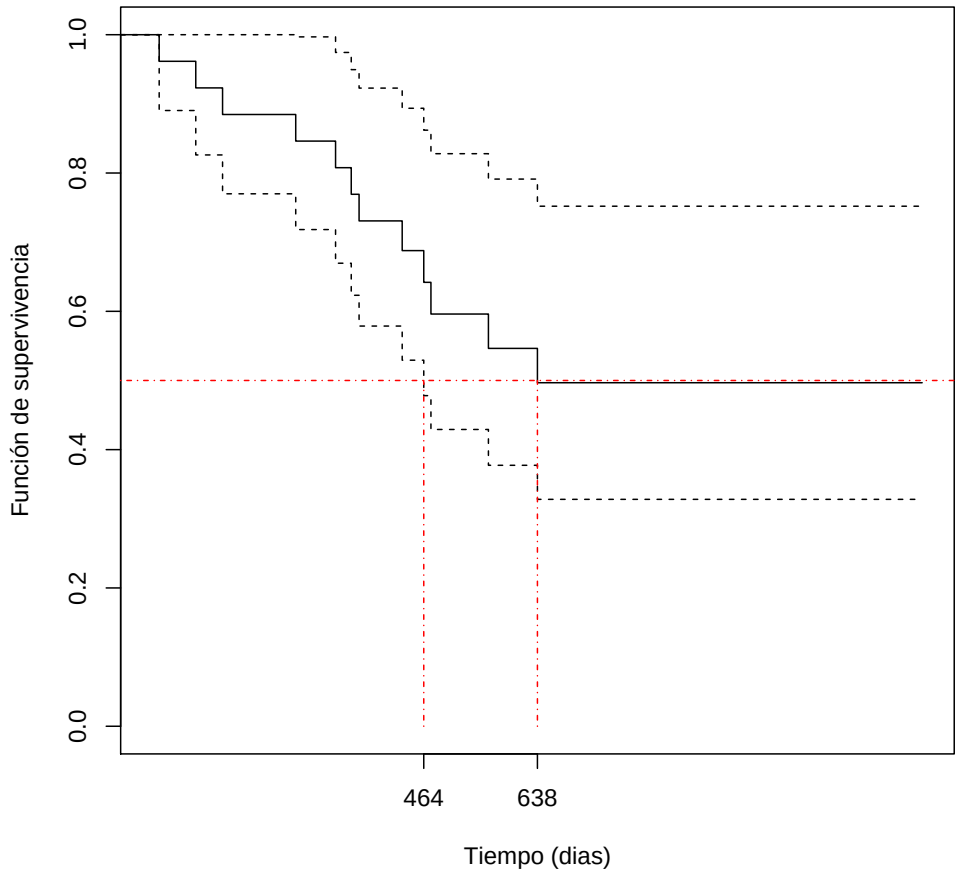

Figura 4 : Mediana e intervalo de confianza: 638 [95 %IC: $(464, \infty)$]

Intervalo de confianza

Codigo de R para calcular un percentil *p* y su intervalo de confianza:

```
> per_surv <- function(mod,p=0.5)
+ {
+   qp  <- ifelse(min(mod$surv)   > 1-p,Inf,min(mod$time[mod$surv <= 1-p]))
+   lqp <- ifelse(min(mod$lower) > 1-p,Inf,min(mod$time[mod$lower <= 1-p]))
+   uqp <- ifelse(min(mod$upper) > 1-p,Inf,min(mod$time[mod$upper <= 1-p]))
+   return(list(qp=qp,ic=c(lqp,uqp)))
+ }
>
> km <- survfit(Surv(futime,fustat)~1,data=ovarian)
> per_surv(km,p=0.5)
$qp
[1] 638

$ic
[1] 464 Inf

> per_surv(km,p=0.25)
$qp
[1] 365

$ic
[1] 268 Inf
```

Cáncer de pulmon

```
> km1 <- survfit(Surv(time,status)~1,data=lung)
> per_surv(km1,p=0.5)
$qp
[1] 310

$ic
[1] 285 363

> per_surv(km1,p=0.25)
$qp
[1] 170

$ic
[1] 145 197
```

- La mediana del tiempo de vida de estos pacientes es 310 dias [95 %IC: 285-363]
- El percentil 25 del tiempo de vida de estos pacientes es 170 [95 %IC: 145-197]

Mediana: Intervalo de confianza

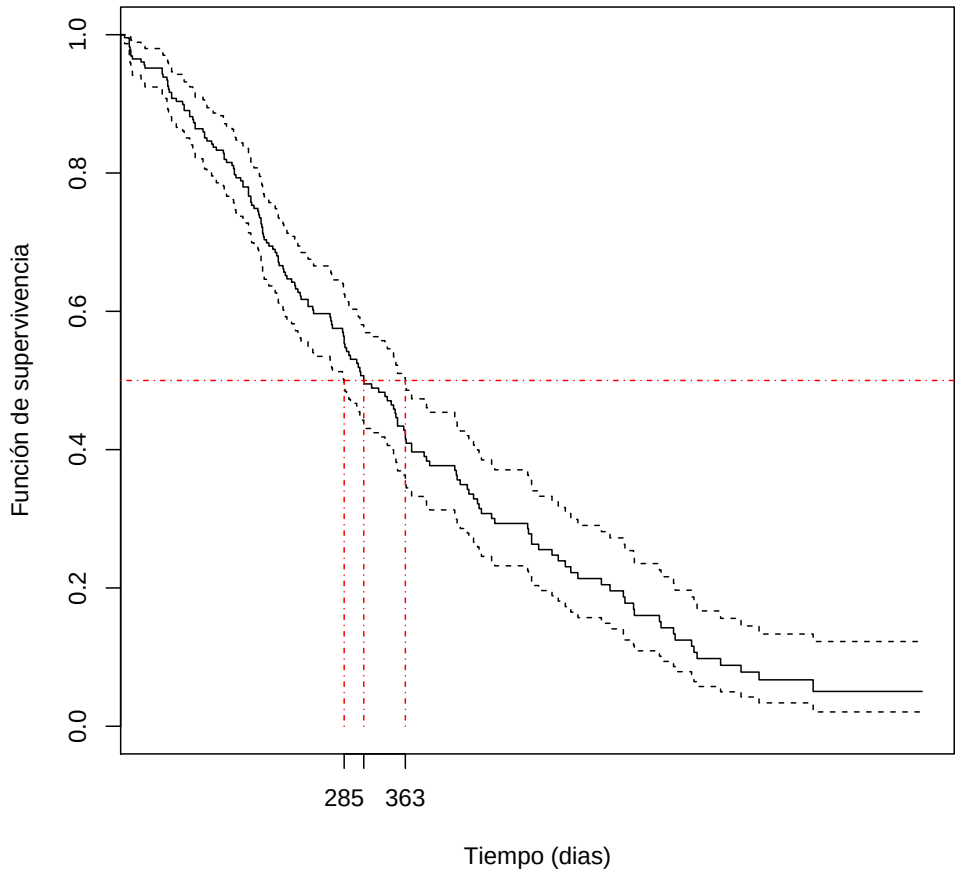

Figura 5 : Mediana e intervalo de confianza: 310 dias [95 %IC: 285-363]

Censura no informativa

Supuesto:

- Individuos que son censurados estan en el mismo riesgo de sufrir el evento que aquellos que se mantienen en seguimiento
- Las personas que se mantienen en seguimiento, en cualquier momento, deben representar a la población en riesgo en ese momento.

Ejemplos de censura

- **No Informativa**
  - Se estudia la vida útil de un catéter, sin embargo fallece el paciente que lo llevaba.
  - Se estudia el tiempo hasta graduarse de la universidad, sin embargo el estudiante recibe una beca para estudiar en el extranjero y abandona la universidad por esta.
- **Informativa**
  - Un paciente con cáncer abandona el estudio por fuerte dolor, lo cual puede ser indicio de un pronto fallecimiento.

Referencias I

[1] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *American Statistical Association*, 53(282):457–481, 1958.