Introducción al Análisis de Supervivencia con R

Madrid, 15 de Octubre de 2015

Jesús Herranz Valera

(jesus.herranz@imdea.org)

Bioestadístico Senior

Instituto IMDEA Alimentación

Índice

- ✓ **Introducción al análisis de supervivencia**
- ✓ **Estimador de Kaplan-Meier**
- ✓ **Análisis descriptivo del tiempo de supervivencia**
- ✓ **Comparación de curvas de supervivencia**

Bibliografía

- D. Hosmer & S. Lemeshow. *Applied Survival Analysis*. Wiley, 2008
- D. Kleinbaum & M. Klein. *Survival Analysis*. Springer, 2012
- F. Harrell. *Regression Modeling Strategies*. Springer, 2001

Análisis de supervivencia

- El **análisis de supervivencia** reúne las técnicas estadísticas apropiadas para analizar estudios en los que los individuos son seguidos **a lo largo de un periodo de tiempo** hasta que ocurre un determinado **evento de interés**

- Ejemplos de **eventos clínicos**: recidiva, recaída, progresión, muerte, alta hospitalaria, curación, ….

- La **variable respuesta** a analizar es el **tiempo hasta que ocurre el evento**

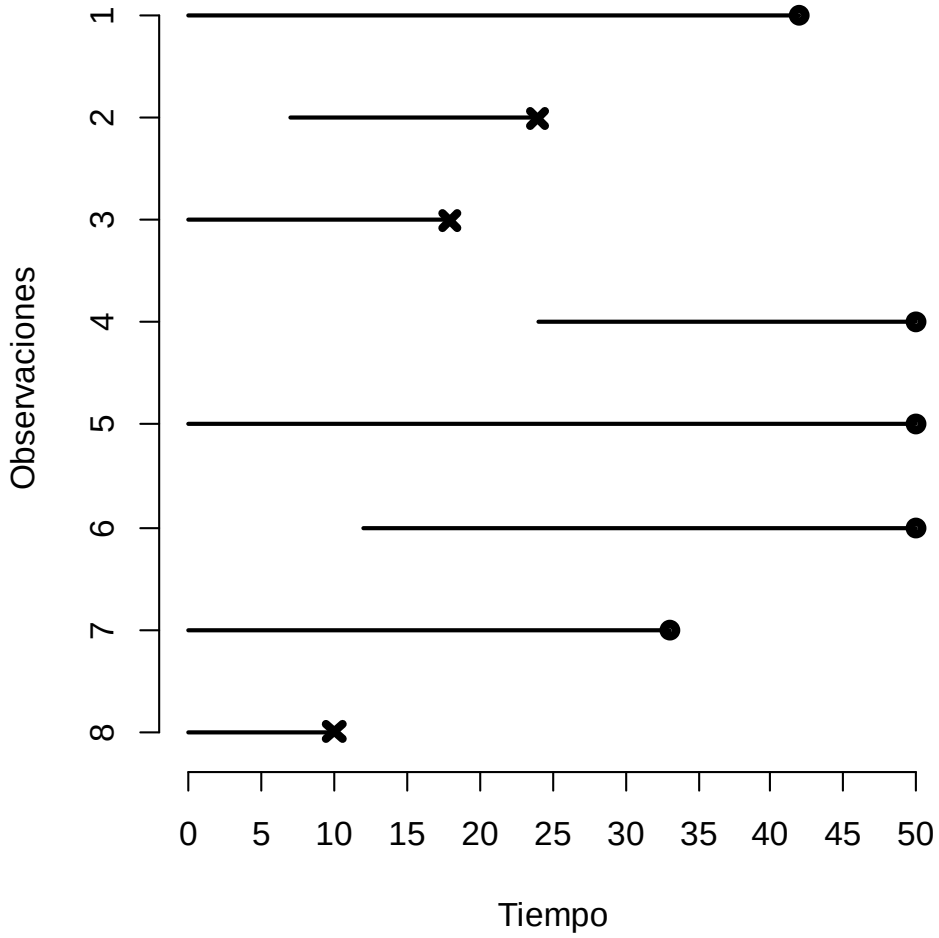
- **Estudios de seguimiento**
  - Fechas de inicio y final del estudio
  - Periodo de reclutamiento, en el que los individuos se incorporan al estudio

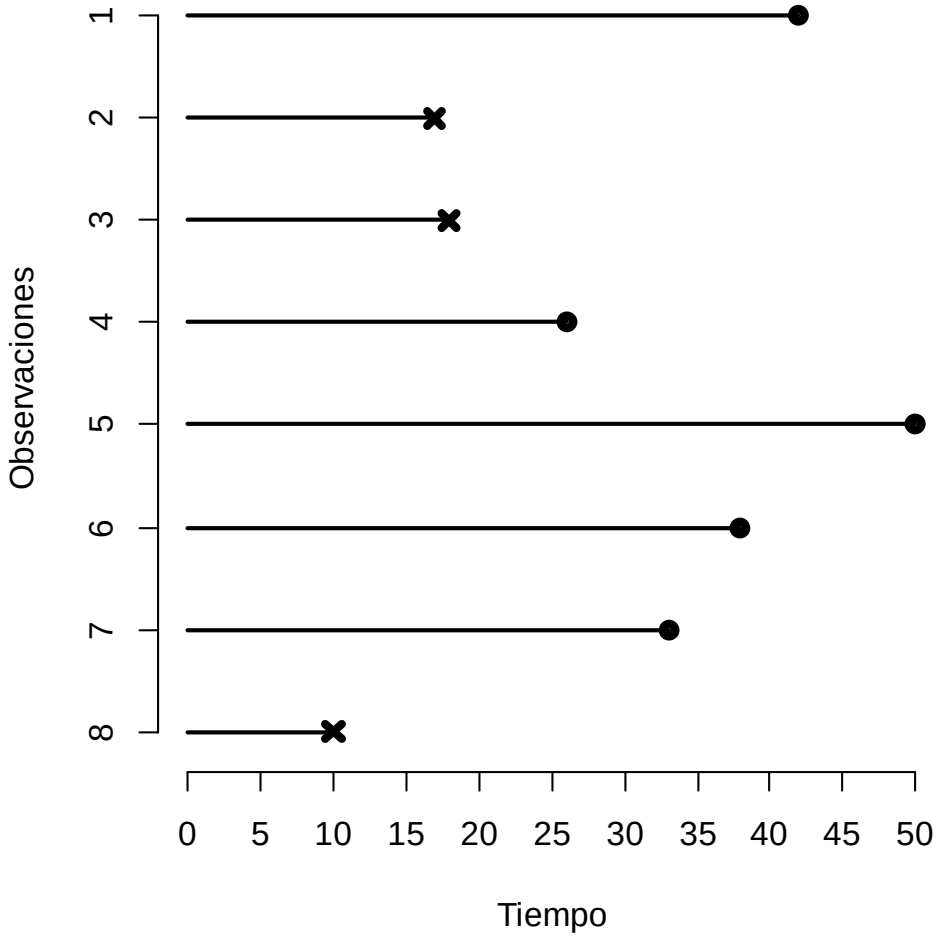
Tiempo de supervivencia

- **Tiempo de supervivencia**
  - Tiempo entre la **incorporación** al estudio y la fecha en la que ha **ocurrido el evento**

- **Observaciones censuradas**
  - Individuos para los que **no ha ocurrido el evento**
  - **Censuras a la derecha:** individuos en los que no ha ocurrido el evento al finalizar el estudio, o individuos perdidos en el seguimiento por otras causas
  - Se registra el **tiempo de seguimiento**: tiempo trascurrido entre la fecha de **incorporación** al estudio y la fecha de la **última observación**

Tiempo de supervivencia





x – evento
o – censurado

tiempo de supervivencia
tiempo de seguimiento

Análisis de supervivencia

- La **variable respuesta** en un análisis de supervivencia tiene **dos componentes**:
  - c_i es una variable **binaria**, llamada también **estado (status)**
    - 1 si ha ocurrido el evento
    - 0 si es una observación censurada
  - t_i es el **tiempo de seguimiento,** que coincide con **el tiempo de supervivencia** para las observaciones para las que ha ocurrido el evento

- El análisis de supervivencia incluye el análisis del “ritmo” o “velocidad” en la que se presenta el evento en el tiempo, es decir, la **tasa de incidencia del evento**
  - La variable tiempo no se puede analizar con regresión lineal, porque existen observaciones censuradas

Fichero de datos: anderson

- El fichero contiene datos **de tiempos de recaída** de **42 pacientes de leucemia**, para comparar **2 tratamientos** (variable “rx”)

Nombre	**Descripción**	**Categorías / Comentarios**
subj	Identificador	
survt	**Tiempo hasta la recaída**	Tiempo de seguimiento para los que no han recaído
status	**Recaída**	0 = observación censurada 1 = ha recaído
sex	Sexo	0 = mujer 1 = hombre
logwbc	Número glóbulos blancos	Escala logarítmica
rx	**Tratamiento**	0 = nuevo tratamiento 1 = tratamiento estándar

- Datos extraídos de *“Survival Analyis: a Self-Learning Text”* de David G. Kleinbaum

Ejemplo: Análisis de supervivencia

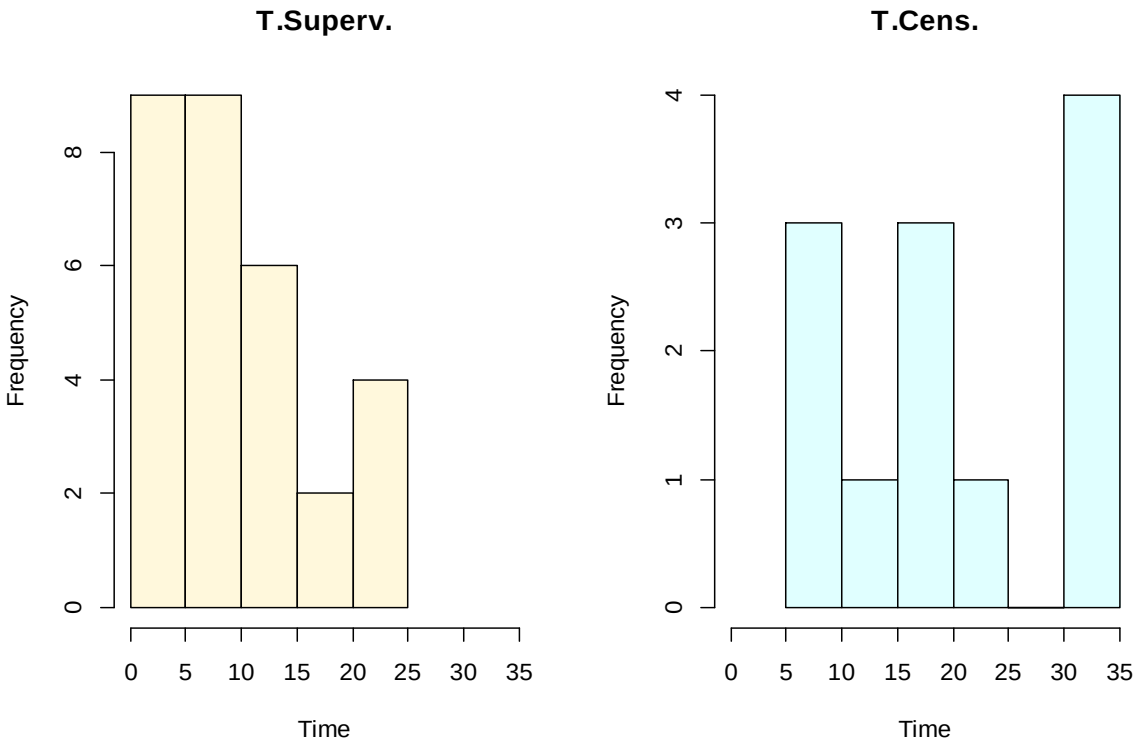
```
> library(survival)
> library(survMisc)
> ## Load the Data
> xx = read.csv ( " . . . /anderson.csv", sep=";", header=T )
> head(xx)
  subj survt status sex logwbc rx
1    1    35      0   1   1.45  0
2    2    34      0   1   1.47  0
3    3    32      0   1   2.20  0
4    4    32      0   1   2.53  0
5    5    25      0   1   1.78  0
6    6    23      1   1   2.57  0
> dim(xx)
[1] 42  6
> table(xx$rx)
 0  1
21 21
> table(xx$status)
 0  1
12 30
```

- Se cargan las **librerías *survival***, que contiene todas las técnicas básicas del análisis de supervivencia, y ***survMisc*** que contiene algunas funcionalidades adicionales
- El fichero contiene 42 observaciones, 21 en cada tratamiento (variable “rx”)
- De los 42 individuos, 30 han sufrido una recaída y 12 son observaciones censuradas

Ejemplo: Análisis de supervivencia

```
> sort ( xx$survt [ xx$status == 1 ] )     ## Tiempos de Seupervivencia
 [1]  1  1  2  2  3  4  4  5  5  6  6  6  7  8  8  8  8 10 11 11 12 12 13 15 16 17 22
[28] 22 23 23
> sort ( xx$survt [ xx$status == 0 ] )     ## Tiempos censurados
 [1]  6  9 10 11 17 19 20 25 32 32 34 35
>
> dev.new(); par ( mfrow = c (1,2) )
> hist( xx$survt [ xx$status == 1 ], xlim=c(0,35), xlab="Time", main="T.Superv." ,
col="cornsilk")
> hist( xx$survt [ xx$status == 0 ], xlim=c(0,35), xlab="Time", main="T.Cens.",
col="lightcyan" )
```

- Los tiempos de supervivencia, en los que se han producido los eventos, son muchos más cortos que los tiempos de las observaciones censuradas



Función de supervivencia

- T variable aleatoria **tiempo de supervivencia** (cuantitativa positiva, T>0)

- **Función de supervivencia:** probabilidad de que un individuo sobreviva durante un tiempo superior a t
  - Al inicio del estudio **S(0)=1** porque todos los individuos están vivos
  - **S(t) disminuye**

$$S(t) = Prob(T > t)$$

- El primer objetivo de un análisis de supervivencia es una **descripción univariante** de los datos observados mediante la **estimación de la función de supervivencia**

Estimador no paramétrico. Kaplan-Meier

- El **método de Kaplan-Meier** es un método **no paramétrico** estima las probabilidades de supervivencia $S(t_j)$ en los instantes en los que ha ocurrido el evento
  - para sobrevivir en un momento determinado, se ha tenido que haber **sobrevivido en todos los tiempos anteriores** hasta ese momento

- Se basa en una **probabilidad condicional** compuesta, es un producto de la supervivencia en el instante anterior y la tasa de supervivencia en ese instante

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \cdot \hat{S}(t_j / t_{j-1}) \qquad \hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$

  - cada probabilidad se obtiene **dividiendo** el número de individuos que estaban **en riesgo al final** del intervalo con los que lo estaban **al principio**
  - n_i número de individuos en riesgo en el tiempo $t_{(i)}$; d_i número observado de eventos

Medidas descriptivas del tiempo de supervivencia

- **Media** del tiempo de supervivencia
  - No se puede calcular porque no se dispone del tiempo para todos los individuos (observaciones censuradas)

- **Mediana** del tiempo de supervivencia
  - No se necesita conocer el tiempo de supervivencia de todos los individuos
  - No se puede calcular cuando hay muchas observaciones censuradas

$$S(t_{mediana}) = 0.50$$

- **Cuartiles** y percentiles son medidas también adecuadas

Ejemplo: Análisis descriptivo tiempo supervivencia

```
> ## Objeto Surv
> Surv ( xx$survt , xx$status )
 [1] 35+ 34+ 32+ 32+ 25+ 23  22  20+ 19+ 17+ 16  13  11+ 10+ 10   9+  7   6+  6   6   6
[22] 23  22  17  15  12  12  11  11   8   8   8   8   5   5   4   4   3   2   2   1   1
> ## Estimador KM
> kmfit1 = survfit ( Surv ( survt , status ) ~ 1 , data = xx )
> kmfit1
Call: survfit(formula = Surv(survt, status) ~ 1, data = xx)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     42      42      42      30      12       8      22
```

- La función ***Surv()*** define la **variable respuesta** en un análisis de supervivencia, y muestra **los tiempos de supervivencia**, indicando con un signo “+” las censuras, que significa que el tiempo de supervivencia es superior al dato registrado
- La función ***survfit()*** calcula el **estimador de Kaplan-Meier** de la función de supervivencia
- Si no se especifica en el modelo ninguna variable predictora (~ 1), se obtiene la **supervivencia global**, la de toda la muestra
- **La mediana de supervivencia es 12** (IC95%: 8 – 22), lo que indica que a los 12 meses la mitad de los individuos ha recaído

Ejemplo: Análisis descriptivo tiempo supervivencia

```
> summary ( kmfit1 )
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     42       2    0.952  0.0329       0.8901        1.000
    2     40       2    0.905  0.0453       0.8202        0.998
    3     38       1    0.881  0.0500       0.7883        0.985
    4     37       2    0.833  0.0575       0.7279        0.954
    5     35       2    0.786  0.0633       0.6709        0.920
    6     33       3    0.714  0.0697       0.5899        0.865
    7     29       1    0.690  0.0715       0.5628        0.845
    8     28       4    0.591  0.0764       0.4588        0.762
   10     23       1    0.565  0.0773       0.4325        0.739
   11     21       2    0.512  0.0788       0.3783        0.692
   12     18       2    0.455  0.0796       0.3227        0.641
   13     16       1    0.426  0.0795       0.2958        0.615
   15     15       1    0.398  0.0791       0.2694        0.588
   16     14       1    0.369  0.0784       0.2437        0.560
   17     13       1    0.341  0.0774       0.2186        0.532
   22      9       2    0.265  0.0765       0.1507        0.467
   23      7       2    0.189  0.0710       0.0909        0.395
```

- La función ***summary()*** muestra una tabla de todos los tiempos donde se ha producido algún evento, el número de individuos que estaba en riesgo en ese momento, el número de eventos y la estimación de la supervivencia con su IC95% y SE
- Por ejemplo, la supervivencia a los 6 meses es de 0.714 (IC95%: 0.590 – 0.865)
- La **Mediana** = 12 y **los cuartiles** Q3 (P_{75}) = 6 y Q1 (P_{25}) = 23

Ejemplo: Análisis descriptivo tiempo supervivencia

```
> summary ( kmfit1 )
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     42       2    0.952  0.0329        0.8901        1.000
    2     40       2    0.905  0.0453        0.8202        0.998
    3     38       1    0.881  0.0500        0.7883        0.985
    4     37       2    0.833  0.0575        0.7279        0.954
    5     35       2    0.786  0.0633        0.6709        0.920
    6     33       3    0.714  0.0697        0.5899        0.865
    7     29       1    0.690  0.0715        0.5628        0.845
    8     28       4    0.591  0.0764        0.4588        0.762
   10     23       1    0.565  0.0773        0.4325        0.739
   11     21       2    0.512  0.0788        0.3783        0.692
. . .
```

- El estimador de Kaplan-Meier se puede calcular manualmente:

$$\frac{42-2}{42} = \frac{40}{42} = 0.952$$

$$0.952 \cdot \frac{40-2}{40} = 0.905$$

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \cdot \hat{S}(t_j / t_{j-1})$$

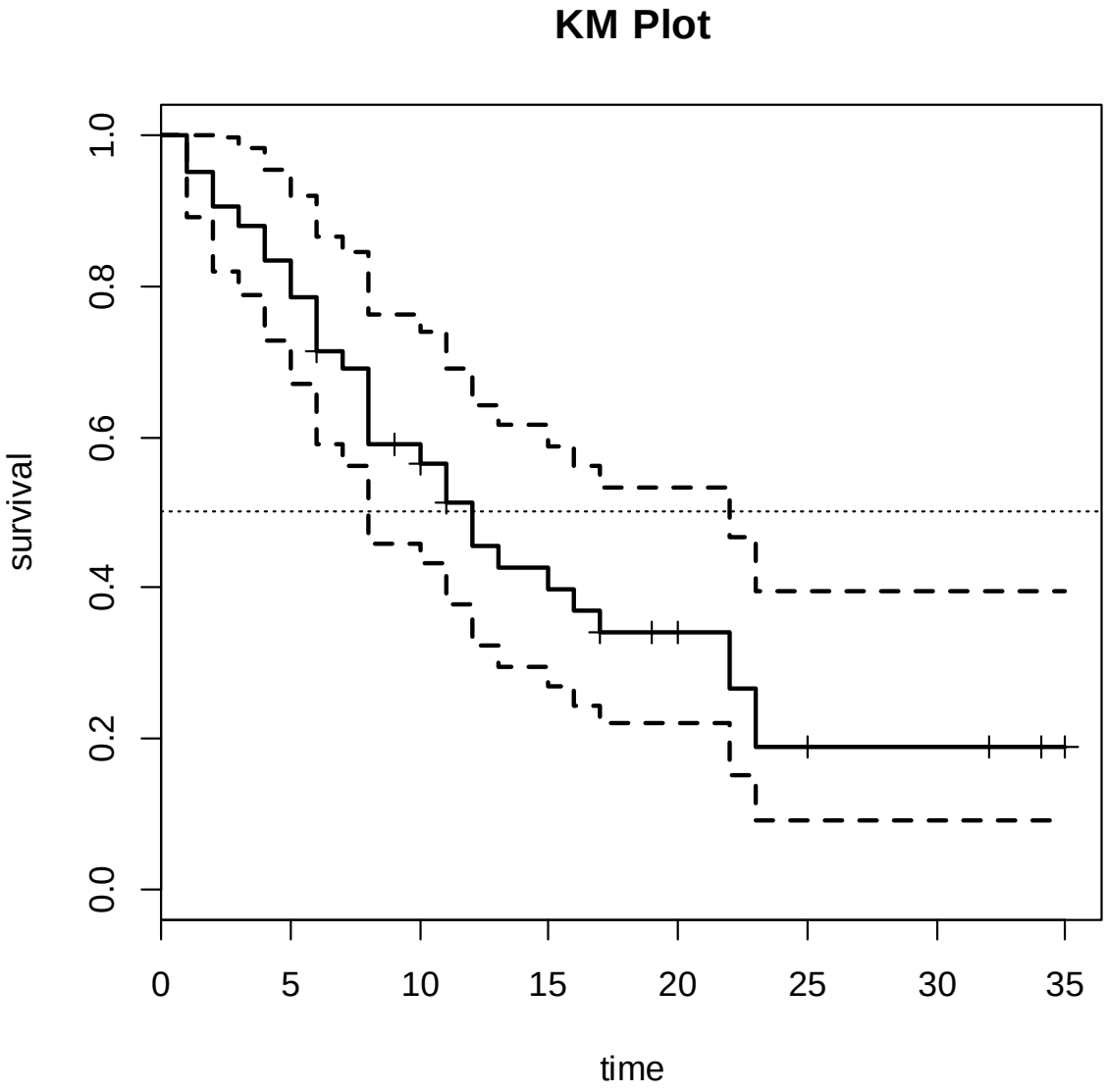
$$\hat{S}(t) = \prod_{t_{(i)} \le t} \frac{n_i - d_i}{n_i}$$

Ejemplo: Estimador de Kaplan-Meier

```
> ## Tabla de Vida
> summary ( kmfit1 , times = 6 )
Call: survfit(formula = Surv(survt, status) ~ 1, data = xx)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     33      12    0.714  0.0697        0.59        0.865
> summary ( kmfit1 , times = seq ( 6, 18, by=6 ) )
Call: survfit(formula = Surv(survt, status) ~ 1, data = xx)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     33      12    0.714  0.0697        0.590        0.865
   12     18      10    0.455  0.0796        0.323        0.641
   18     11       4    0.341  0.0774        0.219        0.532
>
> ## KM Plot
> dev.new()
> plot( kmfit1, xlab="time", ylab="survival", lwd=2, main="KM Plot" )
> abline ( h = 0.5 , lty = 3 )        # Mediana
```

- Con *summary()* se pueden obtener estimaciones para un tiempo determinado o una **tabla de vida** que es similar a KM, pero con tiempos agrupados que se especifican en el parámetro *times*=
- Con la función ***plot()*** se muestra la curva de supervivencia obtenida con el **estimador Kaplan-Meier**

Ejemplo: Estimador de Kaplan-Meier

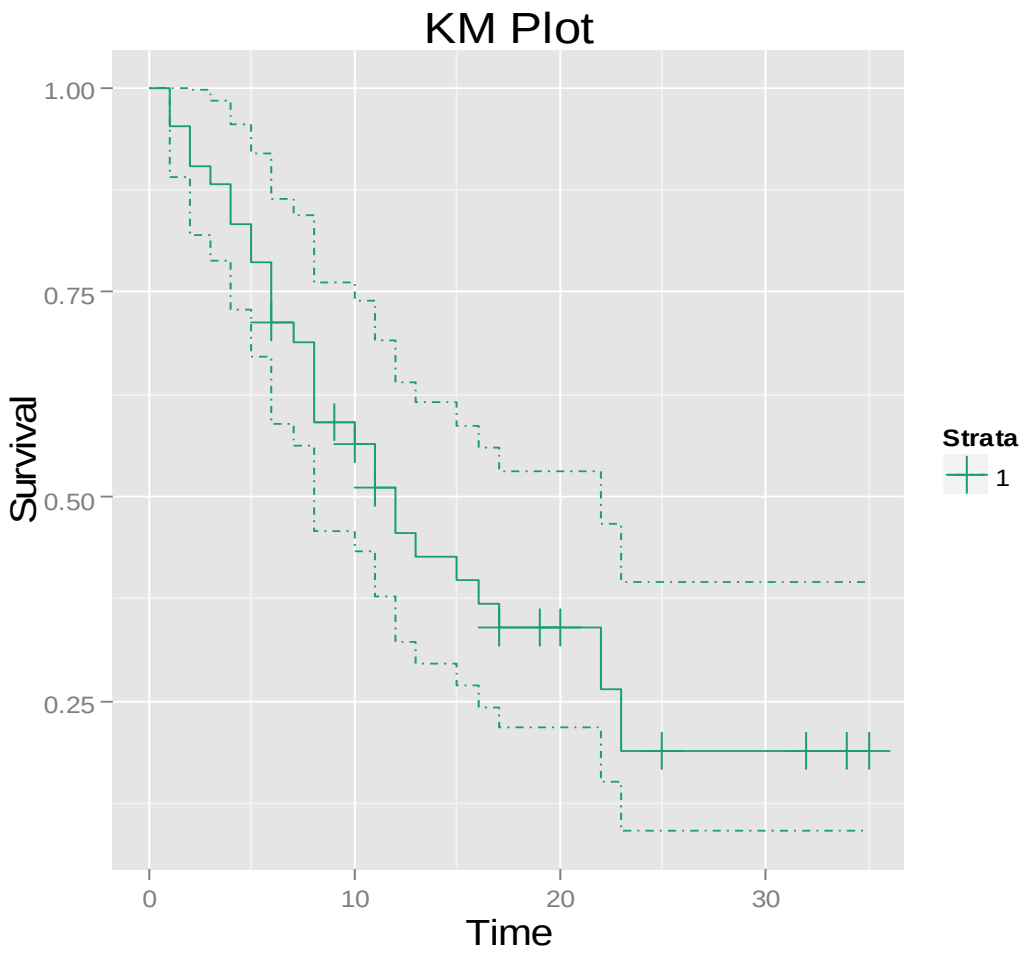


- La **curva KM** estima la función de supervivencia, probabilidad acumulada de que no haya ocurrido el evento
- Empieza en **1** y es una curva **decreciente escalonada** (donde ha ocurrido algún evento)
- La **mediana es 12**, es el tiempo en el que S(12)=0.50
- Las cruces representan a las observaciones censuradas

Ejemplo: Estimador de Kaplan-Meier

```
> ## KM Plot con survMisc (ggplot2)
> dev.new()
> autoplot ( kmfit1 , title= "KM Plot", type="CI", alpha=1 )
```

- Con la función ***autoplot()*** del paquete ***survMisc*** se genera un gráfico KM con ***ggplot2***



Comparación de curvas de supervivencia

- **Analizar si la supervivencia** en 2 o más grupos es igual o si hay diferencias estadísticamente significativas entre los grupos

- Representar las **gráficas KM** de los grupos

- Se compara el **número de eventos observados** en cada uno de los k grupos con el **número de eventos esperados**, en los tiempos en los que ha ocurrido algún evento
  - El número de eventos esperados se calculará suponiendo que la supervivencia es igual en todos los grupos

Prueba log-rank (Mantel-Haenszel)

- El **test log-rank**, es la suma de las diferencias entre los eventos observados y los eventos esperados para todos los tiempos de supervivencia observados, dividido por una estimación de la varianza

$$Q = \frac{\left[\sum_{i=1}^{m}(d_{1i} - \hat{e}_{1i)}\right]^2}{\sum_{i=1}^{m}\hat{v}_{1i}}$$

Bajo la hipótesis nula (las 2 curvas de supervivencia son iguales), Q sigue una distribución chi-cuadrado con 1 gl

- Generalización a **k grupos** (k>2), Q sigue una chi-cuadrado con k-1 gl

- **Test de Wilcoxon** es un versión ponderada, usando como pesos el número de individuos en riesgo. Da más importancia a los **tiempos iniciales** donde hay más individuos

Ejemplo: Comparación de curvas

```
> kmfit2 = survfit ( Surv ( survt , status ) ~ rx , data = xx )
> summary ( kmfit2 )
                rx=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807
                rx=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000     NaN           NA           NA
```

- Se usa la función ***survfit()*** para calcular el estimador KM para los grupos

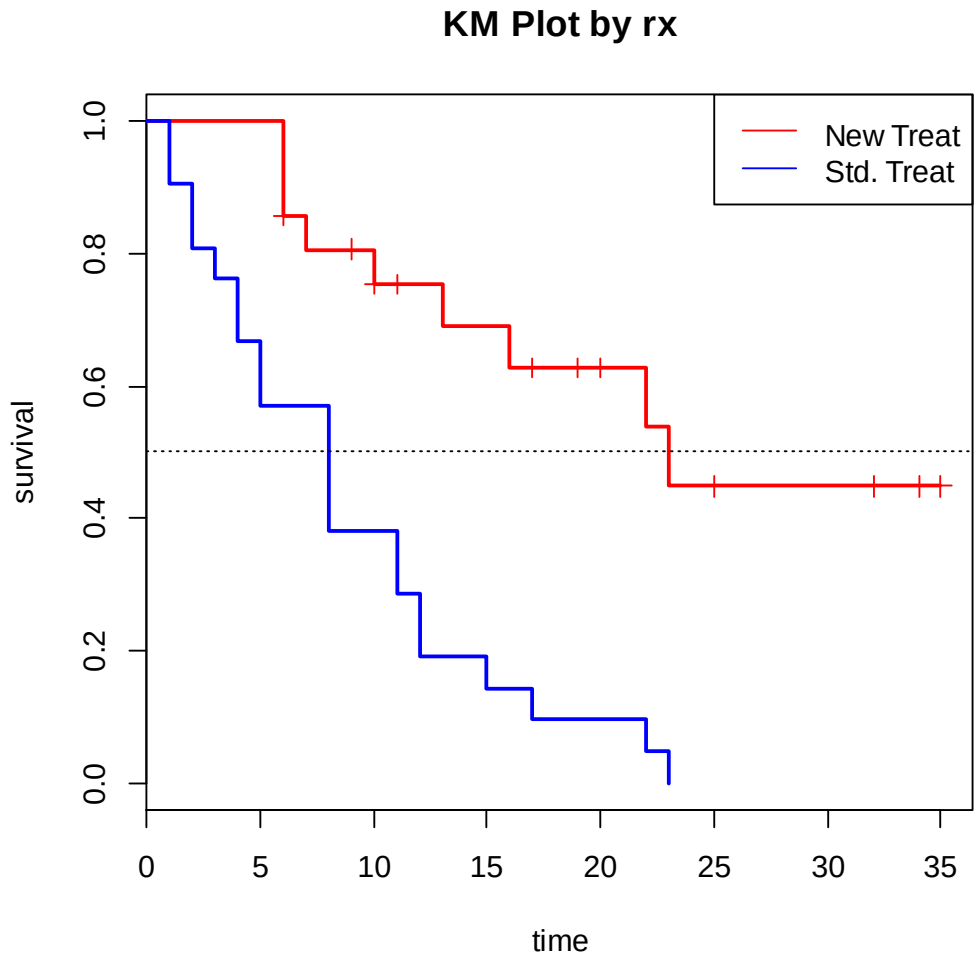
Ejemplo: Comparación de curvas

```
> kmfit2
     records n.max n.start events median 0.95LCL 0.95UCL
rx=0      21    21      21      9     23      16      NA
rx=1      21    21      21     21      8       4      12
> ## Log-rank test
> survdiff ( Surv ( survt , status ) ~ rx , data = xx )

      N Observed Expected (O-E)^2/E (O-E)^2/V
rx=0 21        9     19.3      5.46      16.8
rx=1 21       21     10.7      9.77      16.8
 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
> ## Wilcoxon test - Peto-Peto test
> survdiff ( Surv ( survt , status ) ~ rx , data = xx , rho=1 )

      N Observed Expected (O-E)^2/E (O-E)^2/V
rx=0 21     5.12    12.00      3.94      14.5
rx=1 21    14.55     7.68      6.16      14.5
 Chisq= 14.5  on 1 degrees of freedom, p= 0.000143
```

- Los pacientes con **el nuevo tratamiento** (rx=0) presentan una **mejor supervivencia**: la mediana de la supervivencia es 23 frente a 8 con el tratamiento estándar
- La función ***survdiff()*** se utiliza para calcular el **test log-rank**, y con el parámetro *rho=1* se puede calcular el **test de Wilcoxon**
- Hay **diferencias significativas** en la supervivencia por tratamiento (P<0.001)

Ejemplo: Comparación de curvas

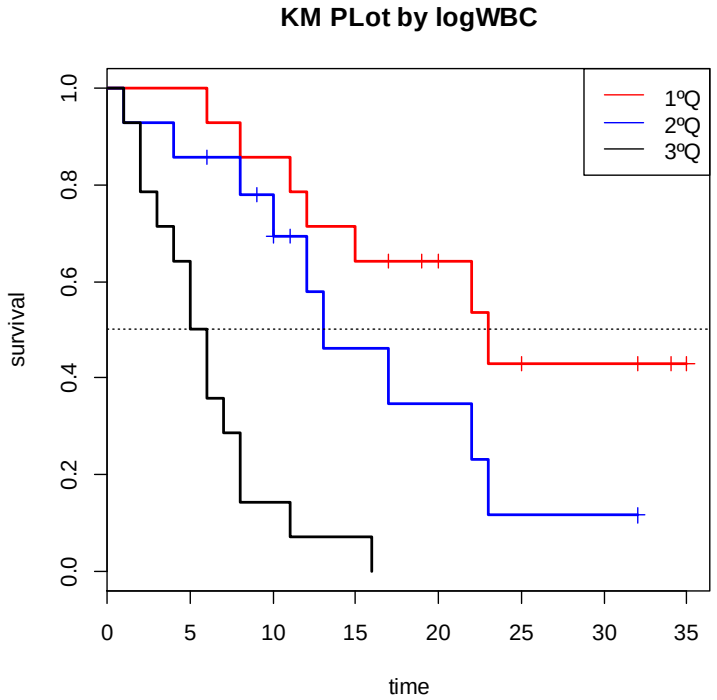
```
> ## KM Plot
> dev.new()
> plot( kmfit2, xlab="time", ylab="survival", col=c("red", "blue"),
+       lwd=2, main="KM Plot by rx" )
> legend ( "topright", c("New Treat", "Std. Treat" ), col=c("red", "blue"), lty=1)
> abline ( h = 0.5 , lty = 3 )        # Mediana
```



Ejemplo: Comparación de curvas

```
> ## Variable continua
> xx$logwbc.gr = cut( xx$logwbc ,
+                  breaks=c( 0, stats::quantile(xx$logwbc, c(0.33,0.66)), 100), lab=1:3)
> kmfit3 = survfit ( Surv ( survt , status ) ~ logwbc.gr , data = xx )
> ## Log-rank test
> survdiff ( Surv ( survt , status ) ~ logwbc.gr , data = xx )
. . .
 Chisq= 27.5  on 2 degrees of freedom, p= 1.08e-06
> ## KM Plot
> dev.new()
> plot( kmfit3, xlab="time", ylab="survival", col=c("red", "blue", "black"),
+       lwd=2, main="KM PLot by logWBC" )
> legend ( "topright", c("1ºQ", "2ºQ", "3ºQ" ), col=c("red", "blue", "black"), lty=1)
```

- Las curvas KM solo tienen sentido para subgrupos de la muestra
- Si se quieren usar para explorar la relación de la supervivencia con una **variable continua**, se puede categorizar esa variable con los cuartiles o terciles



Curso de Formación Continua

Estadística Aplicada con 

Módulos	Fechas 2015
1. Introducción a R	24, 25 Septiembre
2. Métodos de Regresión con R	15, 16 Octubre
3. Métodos de Regresión Avanzados para la Investigación en Ciencias Naturales con R	19, 20, 21 Octubre
4. Estadística Aplicada a la Investigación Biomédica con R	11, 12, 13 Noviembre
5. Modelos Mixtos / Jerárquicos / Multinivel con R	18, 19, 20 Noviembre
6. Estadística Multivariante con R	26, 27 Noviembre
7. Técnicas Estadísticas de Data Mining con R	14, 15, 16, 17 Diciembre

Información: http://goo.gl/whB1MM y en http://www.alimentacion.imdea.org/unidad-de-formacion