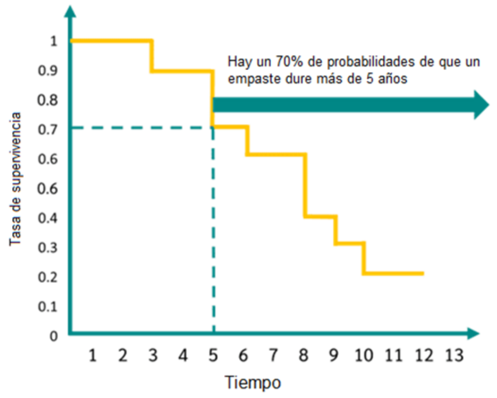
Análisis de Supervivencia con Stata

Luis Guillen Grados
Chief Data Science
lguillen@geoanalytics.pe
lguilleng@gmail.com



https://datatab.net/tutorial/survival-analysis

medium.com/@lguilleng

linkedin.com/in/datascientistlg

¿Qué es un análisis del tiempo de supervivencia?

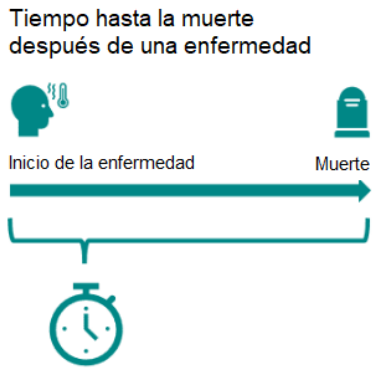
El análisis del tiempo de supervivencia es un grupo de métodos estadísticos en los que la variable objeto de estudio es el tiempo hasta que se produce un evento.

¿Qué significa "tiempo de ocurrencia de un evento"?

En el análisis del tiempo de supervivencia se considera una variable que tiene una hora de inicio y, cuando ocurre un determinado evento, una hora de finalización. El tiempo entre la hora de inicio y el evento es el foco del análisis del tiempo de supervivencia. Por ejemplo, el tiempo se puede medir en días, semanas o meses.

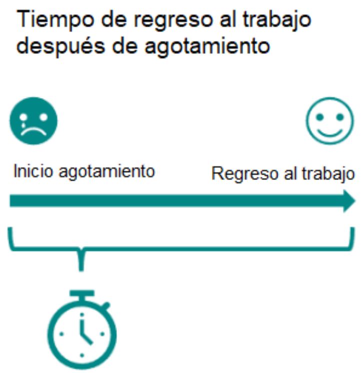
Casos de uso del análisis del tiempo de supervivencia

Un ejemplo sería mirar el tiempo que transcurre entre la abstinencia de una droga y la recaída de la persona en cuestión. La hora de inicio sería entonces el final de la abstinencia y el evento considerado sería entonces la recaída. Por ejemplo, podría estar interesado en saber si los diferentes tipos de tratamiento tienen un impacto en el tiempo de recaída.



Como supone el nombre "análisis del tiempo de supervivencia", también hay un ejemplo clásico: el tiempo hasta la muerte después de una enfermedad. Aquí, la hora de inicio es el reconocimiento de la enfermedad y la hora de finalización es la muerte. Entonces, a menudo es de gran interés si un determinado fármaco influye en el tiempo de supervivencia.

El evento no tiene por qué ser un evento negativo, también podrías fijarte en el momento de volver al trabajo después de un descanso por agotamiento, por ejemplo.

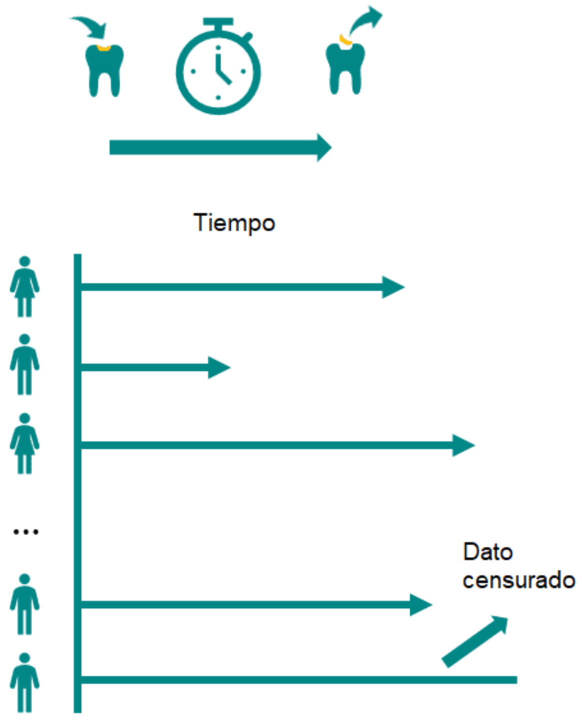


Además, el objeto investigado tampoco tiene por qué ser una persona. En ingeniería, por ejemplo, una pregunta común es cuánto durará un componente en una prueba sin fallar. Aquí, uno podría modificar varios parámetros y ver si tienen un efecto en el tiempo de supervivencia del objeto.

Ejemplo de análisis de tiempo de supervivencia

Supongamos que es un técnico dental y desea analizar el "tiempo de supervivencia" de un empaste en un diente.

Entonces, su hora de inicio es el momento en que una persona recibe un empaste en el diente. El tiempo final, es decir, el acontecimiento, es el momento en que se desprende el empaste. Ahora está interesado en el tiempo entre estos dos eventos.



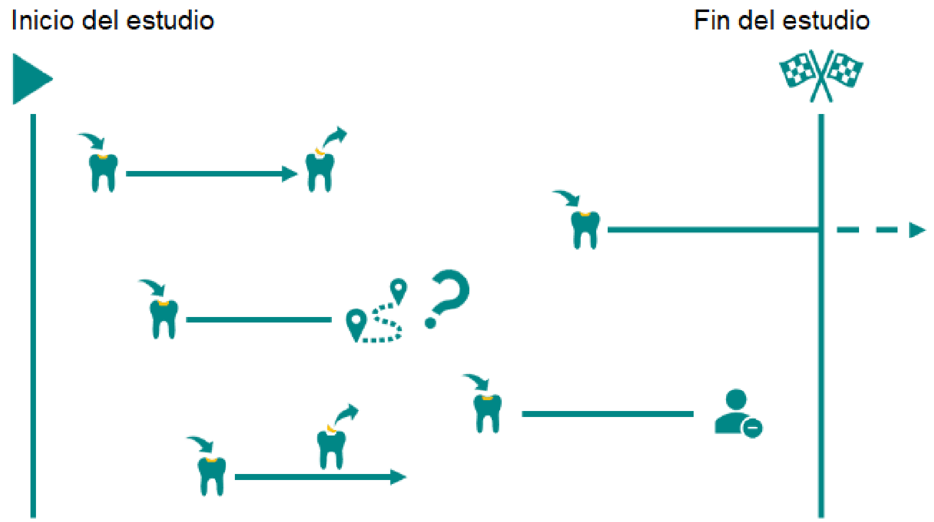
Primero, por supuesto, necesita personas para tener datos para evaluar. Para cada persona de prueba, anota el tiempo que transcurre hasta que se rompe el relleno.

Probablemente te hagas la pregunta: ¿Qué sucede si el empaste de una persona de prueba no se rompe en absoluto? ¿O qué sucede si una persona se muda, cambia de dentista y simplemente no se sabe cuándo se rompió o no el empaste?

Todos estos casos se resumen bajo el término "censura".

Datos censurados

En primer lugar, es importante recordar que un estudio no puede durar indefinidamente, sino que se extiende por un período de tiempo limitado. Por razones de recursos (tiempo, financieros, etc.) y simplemente porque desea publicar los resultados en algún momento, cada estudio tiene una fecha clara de inicio y finalización.



Si se inserta un empaste dentro de este período de tiempo y luego el empaste también se rompe nuevamente dentro de este período de tiempo y esto también está documentado, hay un caso válido. El evento ha ocurrido.

Sin embargo, también es posible que se inserte un relleno y luego se llegue al final del estudio antes de que ocurra el evento.

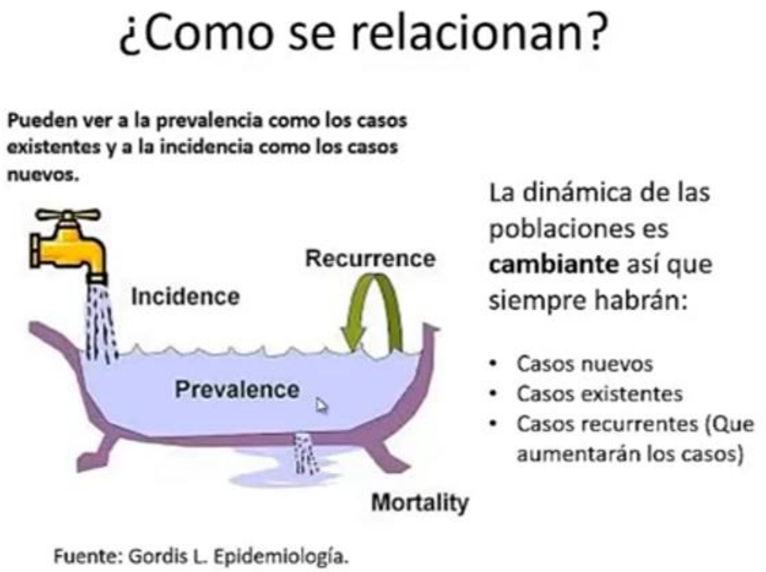
O puede ocurrir que un sujeto decida no continuar con el estudio. En ambos casos, no se sabe cuándo o si se ha producido el hecho que se está considerando.

También puede ocurrir que ocurra otro evento, que no esté considerado en el estudio. Por ejemplo, el paciente podría morir o incluso perder todo el diente. En ambos casos, el evento considerado, que el empaste se rompe, ya no puede ocurrir.

Por supuesto, también puede suceder que la persona no se dé cuenta de que el empaste se ha roto y esto solo se descubre durante el próximo examen de rutina.

En general, hay muchos casos en los que los datos no pueden estar completamente disponibles. Estos datos se denominan entonces "datos censurados".

Prevalencia e Incidencia



Prevalencia puntual

La prevalencia es una proporción (P = A/A+B) que mide la proporción de personas que se encuentran con alguna condición de interés al momento de ser observado,

$$\text{Prevalencia} = \frac{\text{Número de casos de una enfermedad presentes en una población en un tiempo especificado}}{\text{Número de personas en la población en ese tiempo especificado}} \times 10^n$$

Incidencia

Comúnmente denominamos solo como incidencia a la tasa de incidencia, dado que el concepto tasa va implícito. La principal propiedad de esta medida es determinar los casos nuevos que se presentan en una población en un tiempo determinado, de ahí que para su cálculo se requiere un periodo de seguimiento.

Incidencia acumulada

- Cuantifica la aparición de nuevos casos que se presentan en una población en un tiempo determinado.
- El denominador será solo las personas que se encuentren en riesgo de contraer la enfermedad (los sanos)

$$\text{Incidencia} = \frac{\text{Número de casos nuevos de una enfermedad que ocurren en una población durante un periodo de tiempo especificado}}{\text{Número de personas a riesgo de desarrollar la enfermedad durante ese periodo de tiempo}} \times 10^n$$

Tasa de incidencia

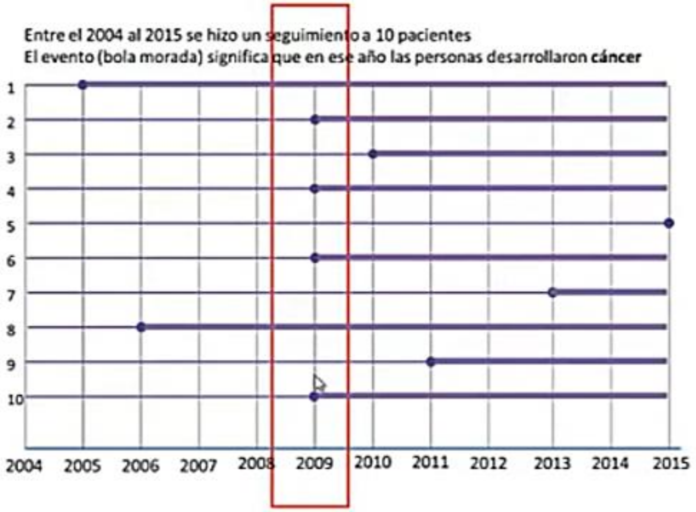
- También es conocida como Densidad de incidencia
- Considera el tiempo que estuvo cada persona con la condición

$$\text{Tasa de incidencia} = \frac{\text{Número de casos nuevos de una enfermedad que ocurren en una población durante un periodo de tiempo especificado}}{\text{Tiempo de exposición de cada persona, totalizado para la población (persona-tiempo de exposición)}} \times 10^n$$

Ejemplos

$$Prev. = \frac{6}{10} x100 = 60\%$$

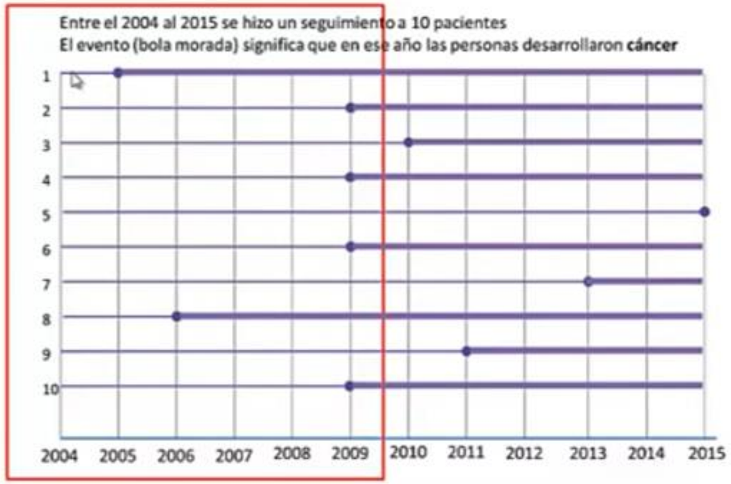
¿Incidencia de casos en el 2009?



Alta aparición de casos

$$Inc. = \frac{4}{8} x100 = 50\%$$

¿Tasa de incidencia al 2009?



En un grupo de 100 personas, en un año pueden aparecer 13,9 casos

$$\text{Tasa}.Inc. = \frac{6}{1+5+5+5+5+5+5+2+5+5} x100 = 13{,}9\%$$

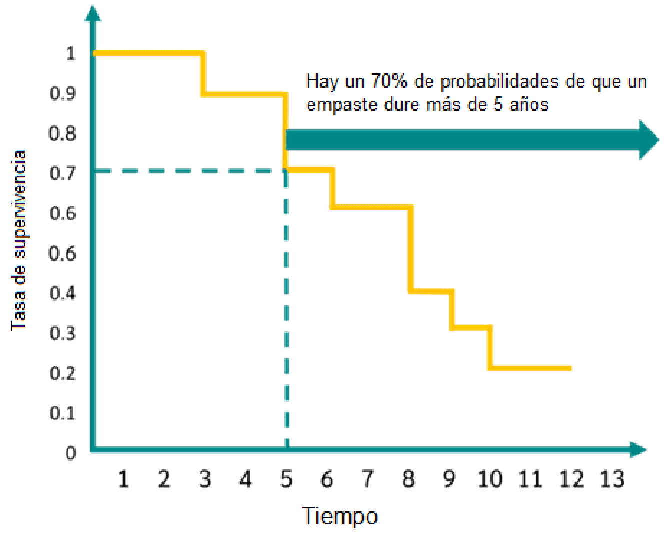
Métodos de análisis del tiempo de supervivencia

Los tres métodos más comunes de análisis del tiempo de supervivencia son (1) las curvas de tiempo de supervivencia de Kaplan Meier, (2) la prueba de rango logarítmico y (3) la regresión de Cox.

Curvas de tiempo de supervivencia de Kaplan Meier

La curva de Kaplan Meier se utiliza para representar gráficamente la tasa de supervivencia o función de supervivencia. Aquí, el tiempo se representa en el eje x y la tasa de supervivencia en el eje y.

¿Cuál es la tasa de supervivencia? En este punto, volvemos al ejemplo del empaste dental. Supongamos que hemos recopilado datos sobre cuánto tiempo tarda en romperse un relleno. En la curva de Kaplan Meier, ahora puede leer la probabilidad de que un empaste dure más de cierto tiempo.



En este contexto, puede interesarle, por ejemplo, la probabilidad de que su empaste dure más de 5 años. Para hacer esto, simplemente muévase a 5 años en el eje x del gráfico y vea cuál es la tasa de supervivencia (eje y). A los 5 años, la curva de Kaplan Maier te da un valor de 0,7.

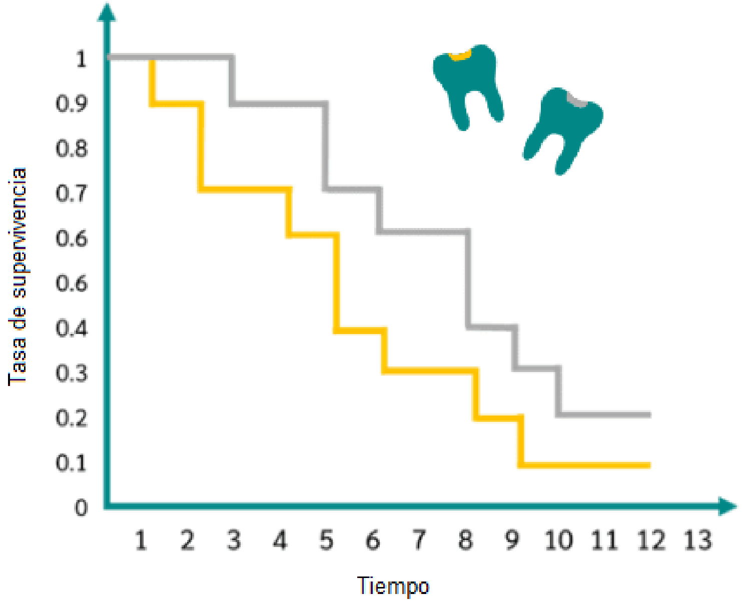
Por lo tanto, existe un 70 % de probabilidades de que un empaste dure más de 5 años. También puede interesarle saber si esta curva difiere para diferentes materiales de empaste, por ejemplo, si un material de empaste es mejor que otro. Para responder a esta pregunta, se usa la prueba de rango logarítmico.

Prueba de rango logarítmico

La prueba de rango logarítmico compara la distribución del tiempo hasta que ocurre un evento de dos o más muestras independientes. Por ejemplo, podría interesarle saber si existe una diferencia en el tiempo de supervivencia de dos materiales diferentes. En este ejemplo, utiliza el material A para la mitad de las materias y el material B para la otra mitad.

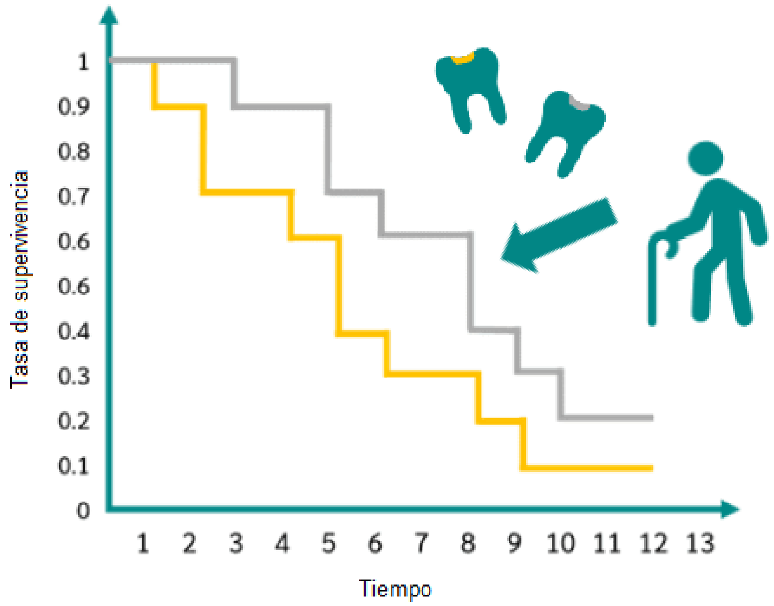
La prueba de rango logarítmico ahora le da una respuesta a la siguiente pregunta: ¿Existe una diferencia significativa entre las dos curvas?

O en otras palabras: ¿El material de empaste influye en el "tiempo de supervivencia" del empaste?



Regresión de Cox

¿Qué sucede si ahora desea verificar si hay otros parámetros que influyen en la curva? Digamos que desea saber no solo si el material influye en el tiempo de supervivencia, sino también si la edad de loas personas influye. Para responder a esta pregunta, la regresión de Cox es el método apropiado.



Aplicaciones con Stata

Utilizaremos el archivo datos_supervivencia.dta, que contiene datos de pacientes con problemas cardiacos.

```
des

Contains data from datos_supervivencia.dta
  obs:           172                          Heart transplant data
 vars:            14                          20 Apr 2023 16:43
 size:         4,816
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
id              int     %8.0g                 Identificador del paciente
anio            byte    %8.0g                 Año de aceptación
edad            byte    %8.0g                 Edad del paciente
Fallecer        byte    %8.0g                 Estado de supervivencia
                                               (1=fallecido)
tiempo          int     %8.0g                 Tiempo de supervivencia (dias)
cirugia         byte    %8.0g                 Cirugia (ejm. CABG)
transplante     byte    %8.0g                 Transplante de corazón
tespera         int     %8.0g                 Tiempo de espera para recibir
                                              transplante
posttran        byte    %8.0g
t1              float   %9.0g
_st             byte    %8.0g                 1 si el registro debe ser utilizado;
                                              0 en caso contrario
_d              byte    %8.0g                 1 si falla; 0 si está censurado
_t              double  %10.0g                Tiempo de análisis cuando termina el
                                              registro
_t0             int     %10.0g                Tiempo de análisis cuando comienza
                                              el registro
--------------------------------------------------------------------------------
```

```
stset tiempo, failure(fallecer=1)

     failure event:  fallecer == 1
obs. time interval:  (0, tiempo]
 exit on or before:  failure

------------------------------------------------------------------------------
        172  total observations
          0  exclusions
------------------------------------------------------------------------------
        172  observations remaining, representing
         75  failures in single-record/single-failure data
     60,591  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =     1,799
```

Incidencia

```
stptime

         failure _d:  fallecer == 1
   analysis time _t:  tiempo
                 id:  id

Cohort     |  person-time   failures        rate   [95% Conf. Interval]
-----------+-----------------------------------------------------------
     total |         3285         30   .00913242   .0063853    .0130615
```

Para una mejor interpretación, pedimos a Stata que la incidencia la muestre en tantos por 1000 personas

```
stptime, per(1000)

         failure _d:  fallecer == 1
   analysis time _t:  tiempo
                 id:  id

Cohort     |  person-time   failures        rate   [95% Conf. Interval]
-----------+-----------------------------------------------------------
     total |         3285         30   9.1324202   6.385258    13.06151
```

El riesgo de fallecer es de 9 fallecimientos por cada 1000 personas en un día

El Método de la Tabla de Vida

Una tabla de vida es una extensión de la tabla de frecuencias tradicional que muestra funciones de supervivencia como la función de riesgo estimada y la función de supervivencia estimada para datos individuales o agregados. Aunque las tablas de vida se pueden construir con datos individuales, el método se utiliza típicamente con datos agregados.

La tabla siguiente muestra una típica tabla de vida con datos agregados. Esta tabla utiliza datos hipotéticos que representan la experiencia de los niños en acogimiento familiar al salir del cuidado temporal para lograr la reunificación familiar dentro de un horizonte de estudio de 12 meses. Un total de 310 niños conforman la muestra. La tabla de vida primero agrupa a los sujetos de la muestra utilizando un intervalo mensual. Por lo tanto, en la tabla aparecen un total de 13 intervalos de tiempo. El límite inferior del intervalo de tiempo es inclusivo y el límite superior es exclusivo. Por ejemplo, el primer intervalo de tiempo de la tabla es [0,1), lo que significa que este intervalo muestra la experiencia de supervivencia de los niños del estudio dentro de un intervalo de tiempo limitado exactamente a 0 meses y menos de 1 mes. Los 310 niños ingresan en el primer intervalo, ya que no se produce ningún evento de reunificación al comienzo de la ventana de estudio o al tiempo 0.

La columna etiquetada como "número de entradas en este intervalo" simplemente muestra el número total de sujetos de estudio que ingresan al intervalo, lo que forma la base, pero no exactamente el mismo número, del conjunto de riesgo para calcular la tasa de riesgo. La columna etiquetada como "número de fallas" indica el número de sujetos que experimentan el evento dentro de este intervalo. Tenga en cuenta que el término "falla" es sinónimo de "ocurrencia del evento"; en análisis de supervivencia, al igual que el término "fallecido", "falla" o "fracaso" es una terminología especial que indica la ocurrencia del evento. Hay 11 niños que se reunieron en el primer mes. La columna etiquetada como "número censurado" indica el número de sujetos que están censurados; para el primer intervalo, se censuran 14 niños. El método de la tabla de vida supone que todas las fallas y censuras ocurren exactamente en el punto medio del intervalo de tiempo. Con esta suposición, la tabla de vida calcula las funciones de supervivencia y de riesgo, junto con estadísticas adicionales como "tamaño de muestra efectivo" y "probabilidad condicional de falla". La fórmula exacta y la ilustración del cálculo de estas estadísticas se muestran en la tabla siguiente.

Construcción de una tabla de vida

Intervalo (LI, LS)	Número que ingresa en el intervalo	Número de fallas	Número de censuras	Tamaño de muestra efectivo	Probabilidad condicional de falla	Sobreviviente	Riesgo
	(a)	(b)	(c)	(d)=(a)-(c)/2	(e)=(b)/(d)	(f)=(1-(e_{i-1}))*h_{i-1}	(g)
[0,1[	310	11	14	303	0.0363	1.0000	0.0370
[1,2[	285	6	24	273	0.0220	0.9637	0.0222
[2,3[	255	10	24	243	0.0412	0.9425	0.0420
[3,4[	221	11	8	217	0.0507	0.9037	0.0520
…							
[12,+	126	0	126	63	0.0000	0.7022	0.0000

Tabla de vida con Stata

```
ltable tiempo fallecer
```

```
                        Beg.                                         Std.
   Interval            Total    Deaths     Lost      Survival       Error       [95% Conf. Int.]
-------------------------------------------------------------------------------------------------
    1      2             172         1        0        0.9942      0.0058       0.9595     0.9992
    2      3             171         3        0        0.9767      0.0115       0.9392     0.9912
    3      4             168         3        0        0.9593      0.0151       0.9165     0.9804
    5      6             165         2        1        0.9476      0.0170       0.9018     0.9724
    6      7             162         2        0        0.9359      0.0187       0.8873     0.9640
    8      9             160         1        0        0.9301      0.0195       0.8802     0.9597
    9     10             159         1        0        0.9242      0.0202       0.8731     0.9553
   11     12             158         0        1        0.9242      0.0202       0.8731     0.9553
   12     13             157         1        0        0.9184      0.0209       0.8660     0.9508
   16     17             156         3        2        0.9006      0.0229       0.8449     0.9370
   17     18             151         1        1        0.8946      0.0235       0.8379     0.9323
   18     19             149         1        0        0.8886      0.0241       0.8309     0.9275
   21     22             148         2        0        0.8766      0.0252       0.8170     0.9177
   28     29             146         1        1        0.8706      0.0258       0.8101     0.9128
   30     31             144         1        1        0.8645      0.0263       0.8031     0.9078
   31     32             142         0        1        0.8645      0.0263       0.8031     0.9078
   32     33             141         1        0        0.8584      0.0268       0.7961     0.9028
   35     36             140         1        0        0.8522      0.0273       0.7891     0.8977
   36     37             139         1        0        0.8461      0.0278       0.7822     0.8925
   37     38             138         1        0        0.8400      0.0283       0.7753     0.8874
   39     40             137         1        3        0.8338      0.0287       0.7683     0.8821
   40     41             133         2        0        0.8212      0.0296       0.7543     0.8715
   43     44             131         1        1        0.8149      0.0301       0.7472     0.8661
   45     46             129         1        1        0.8086      0.0305       0.7402     0.8607

                                            (omitido)

  915    916              25         0        2        0.4765      0.0488       0.3783     0.5681
  941    942              23         0        2        0.4765      0.0488       0.3783     0.5681
  979    980              21         1        1        0.4532      0.0517       0.3500     0.5507
  995    996              19         1        1        0.4287      0.0544       0.3211     0.5318
 1032   1033              17         1        1        0.4027      0.0570       0.2912     0.5114
 1141   1142              15         0        2        0.4027      0.0570       0.2912     0.5114
 1321   1322              13         0        2        0.4027      0.0570       0.2912     0.5114
 1386   1387              11         1        1        0.3644      0.0631       0.2433     0.4862
 1400   1401               9         0        1        0.3644      0.0631       0.2433     0.4862
 1407   1408               8         0        2        0.3644      0.0631       0.2433     0.4862
 1571   1572               6         0        2        0.3644      0.0631       0.2433     0.4862
 1586   1587               4         0        2        0.3644      0.0631       0.2433     0.4862
 1799   1800               2         0        2        0.3644      0.0631       0.2433     0.4862
-------------------------------------------------------------------------------------------------
```

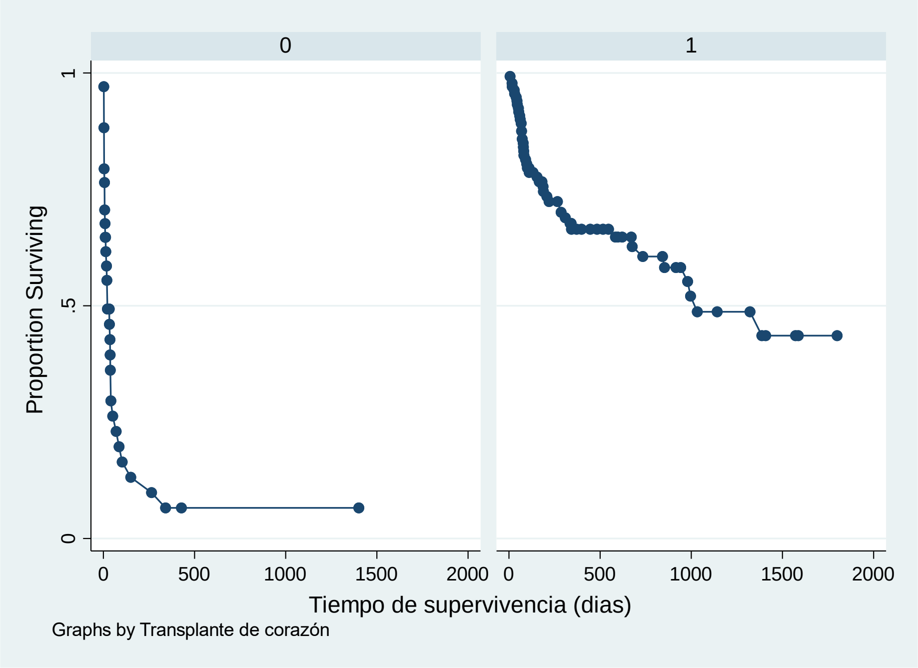
```
ltable tiempo fallecer, by(transplante)
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
transplante = 0								
1	2	34	1	0	0.9706	0.0290	0.8090	0.9958
2	3	33	3	0	0.8824	0.0553	0.7163	0.9541
3	4	30	3	0	0.7941	0.0693	0.6161	0.8961
5	6	27	1	0	0.7647	0.0727	0.5842	0.8747
6	7	26	2	0	0.7059	0.0781	0.5224	0.8296
8	9	24	1	0	0.6765	0.0802	0.4924	0.8060
9	10	23	1	0	0.6471	0.0820	0.4630	0.7818
			(omitido)					
50	51	9	1	0	0.2629	0.0786	0.1261	0.4224
69	70	8	1	0	0.2301	0.0753	0.1029	0.3869
85	86	7	1	0	0.1972	0.0714	0.0810	0.3503
102	103	6	1	0	0.1643	0.0666	0.0606	0.3125
149	150	5	1	0	0.1315	0.0609	0.0419	0.2732
263	264	4	1	0	0.0986	0.0538	0.0253	0.2322
340	341	3	1	0	0.0657	0.0448	0.0117	0.1889
427	428	2	0	1	0.0657	0.0448	0.0117	0.1889
1400	1401	1	0	1	0.0657	0.0448	0.0117	0.1889
transplante = 1								
5	6	138	1	1	0.9927	0.0072	0.9495	0.9990
16	17	136	2	2	0.9780	0.0126	0.9334	0.9929
17	18	132	1	1	0.9706	0.0145	0.9235	0.9889
28	29	130	1	1	0.9631	0.0162	0.9136	0.9845
30	31	128	1	1	0.9555	0.0177	0.9037	0.9798
39	40	126	1	3	0.9479	0.0192	0.8937	0.9748
43	44	122	1	1	0.9401	0.0206	0.8837	0.9696
45	46	120	1	1	0.9322	0.0218	0.8737	0.9641
51	52	118	1	1	0.9243	0.0231	0.8637	0.9585
53	54	116	1	1	0.9163	0.0242	0.8538	0.9528
			(omitido)					
995	996	18	1	1	0.5206	0.0651	0.3868	0.6385
1032	1033	16	1	1	0.4870	0.0690	0.3471	0.6132
1141	1142	14	0	2	0.4870	0.0690	0.3471	0.6132
1321	1322	12	0	2	0.4870	0.0690	0.3471	0.6132
1386	1387	10	1	1	0.4357	0.0785	0.2807	0.5810
1407	1408	8	0	2	0.4357	0.0785	0.2807	0.5810
1571	1572	6	0	2	0.4357	0.0785	0.2807	0.5810
1586	1587	4	0	2	0.4357	0.0785	0.2807	0.5810
1799	1800	2	0	2	0.4357	0.0785	0.2807	0.5810

```
ltable tiempo fallecer, by(transplante) interval(365)
```

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
transplante = 0								
0	365	34	30	2	0.0909	0.0500	0.0233	0.2167
365	730	2	0	1	0.0909	0.0500	0.0233	0.2167
1095	1460	1	0	1	0.0909	0.0500	0.0233	0.2167
transplante = 1								
0	365	138	37	49	0.6740	0.0440	0.5795	0.7518
365	730	52	2	20	0.6419	0.0474	0.5409	0.7263
730	1095	30	5	11	0.5109	0.0645	0.3789	0.6283
1095	1460	14	1	7	0.4623	0.0745	0.3130	0.5990
1460	1825	6	0	6	0.4623	0.0745	0.3130	0.5990

```
ltable tiempo fallecer, by(transplante) graph
```



La estimación de Kaplan-Meier

También conocido como el estimador de límite de producto, el estimador Kaplan-Meier de la función de supervivencia es más popular que el método de tabla de vida y es el enfoque dominante para estimar la función de supervivencia entre los investigadores biomédicos. Este estimador incorpora información de todas las observaciones disponibles, tanto censuradas como no censuradas, considerando la supervivencia a cualquier punto en el tiempo como una serie de pasos definidos por los tiempos de supervivencia y censura observados. La fórmula de Greenwood para estimar la varianza (y por lo tanto el error estándar) de la función de supervivencia permite a los analistas comparar las distribuciones de supervivencia entre grupos y realizar una prueba de significancia. El enfoque de Greenwood se basa en una expansión de la serie de Taylor de primer orden (Hosmer y Lemeshow, 1999).

La tabla siguiente muestra el método de estimación del enfoque de Kaplan-Meier con un conjunto de datos ilustrativos. Supongamos que en un estudio de bienestar infantil, un investigador observó las duraciones de tiempo que permanecieron en hogares de acogida para una muestra hipotética de 14 niños dentro de un período de 12 meses. De los 14 niños, 7 están censurados y sus duraciones de tiempo se etiquetan con *. Al igual que en el método de tabla de vida, el estimador de Kaplan-Meier crea primero una serie de intervalos de tiempo, pero los intervalos se forman de tal manera que cada intervalo es solo para los sujetos que tienen el evento dentro del intervalo de tiempo. Según esta definición, los sujetos censurados no contribuyen a la creación de intervalos de tiempo en la tabla. Este es el caso de los sujetos que están censurados, por ejemplo, en el mes 3 y en el mes 7.

El método de Kaplan-Meier comienza ordenando los tiempos de evento en orden ascendente y creando intervalos de tiempo de tal manera que cada intervalo j contenga solo sujetos que tienen tiempos de evento en el intervalo. Sea nj el número de sujetos que ingresan al intervalo j y dj el número de eventos que ocurren dentro de j. La tabla muestra que, inicialmente, los 14 sujetos ingresan al intervalo de tiempo [0,1); como no ocurre ningún evento en este intervalo, dj = 0. Luego, los 14 sujetos ingresan al segundo intervalo [1, 5) y nj sigue siendo 14. Dado que un sujeto salió del cuidado de crianza dentro de este intervalo, dj para [1, 5) = 1.

Construcción del estimador Kaplan-Meier de la función de supervivencia

Intervalo de tiempo j	n_j	dj	$[1-(d_j/n_j)]$	$\hat{S}(t)$	$S.E\{\hat{S}(t)\}$	95% C.I	
[0,1[	14	0	1.0000	1.0000	0.0000	1.0000	1.0000
[1,5[	14	1	0.9286	0.9286	0.0688	0.7937	1.0635
[5,6[	11	1	0.9091	0.8442	0.1019	0.6443	1.0440
[6,8[	10	1	0.9000	0.7597	0.1218	0.5210	0.9984
[8,9[	7	2	0.7143	0.5427	0.1562	0.2365	0.8488
[9,12[	5	1	0.8000	0.4341	0.1582	0.1240	0.7443
[12,+[	3	1	0.6667	0.2894	0.1584	-0.0210	0.5999

* Este valor estimado es reemplazado por el máximo del límite superior 1.0000 o el mínimo del límite inferior 0.0000 porque el valor estimado es mayor que 1.0000 o menor que 0.0000.

$$\hat{S}(t) = \prod_{j:t_j \leq t}\left[1 - \frac{d_j}{n_j}\right]$$

$$S.E.\{\hat{S}(t)\} \approx [\hat{S}(t)]\left\{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right\}^{1/2}$$

Estimación de la probabilidad de supervivencia con Kaplan Meier

Tiempo cuando ocurre un evento

Eventos ocurridos

Censuras

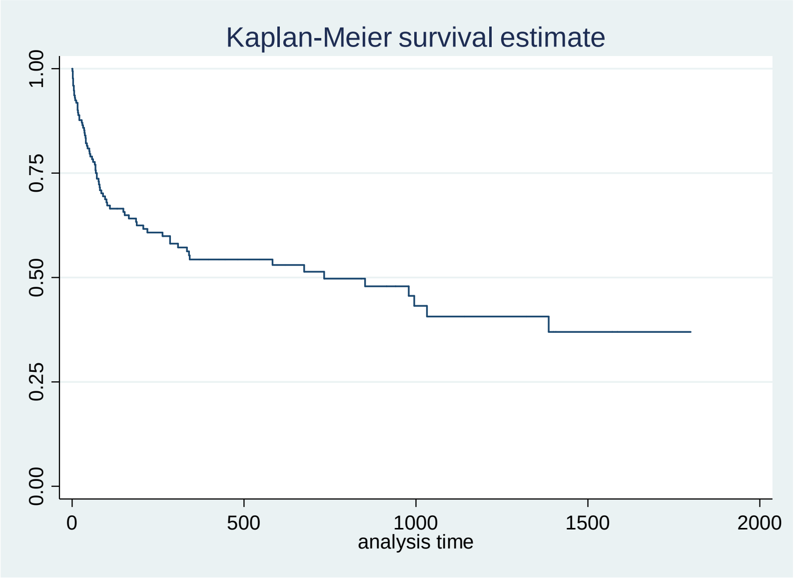
Probabilidad de supervivencia

`sts list`

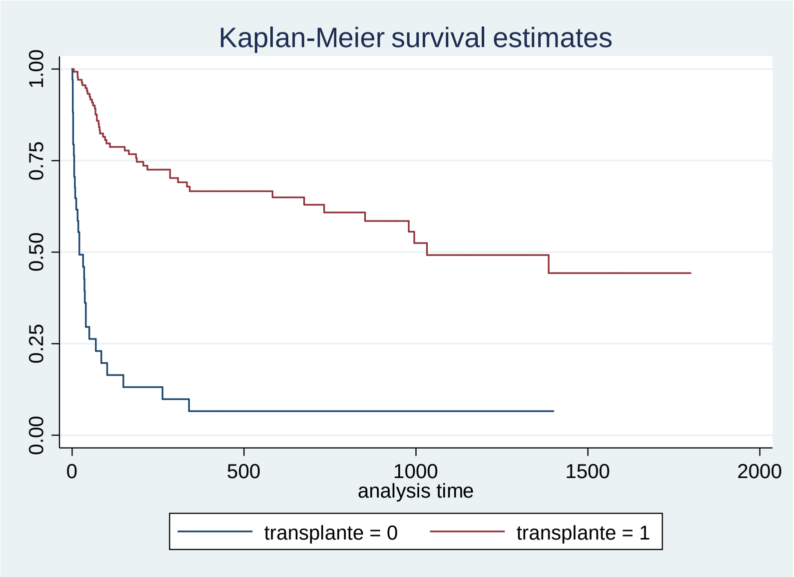
(omitido)

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
1	172	1	0	0.9942	0.0058	0.9595	0.9992
2	171	3	0	0.9767	0.0115	0.9392	0.9912
3	168	3	0	0.9593	0.0151	0.9165	0.9804
5	165	2	1	0.9477	0.0170	0.9019	0.9724
6	162	2	0	0.9360	0.0187	0.8874	0.9640
8	160	1	0	0.9301	0.0195	0.8802	0.9597
9	159	1	0	0.9243	0.0202	0.8732	0.9553
11	158	0	1	0.9243	0.0202	0.8732	0.9553
12	157	1	0	0.9184	0.0209	0.8661	0.9508
16	156	3	2	0.9007	0.0229	0.8452	0.9371
17	151	1	1	0.8948	0.0235	0.8382	0.9324
18	149	1	0	0.8888	0.0241	0.8311	0.9276
21	148	2	0	0.8767	0.0252	0.8172	0.9179
28	146	1	1	0.8707	0.0257	0.8103	0.9129
30	144	1	1	0.8647	0.0263	0.8034	0.9080
31	142	0	1	0.8647	0.0263	0.8034	0.9080
32	141	1	0	0.8586	0.0268	0.7964	0.9029
35	140	1	0	0.8524	0.0273	0.7894	0.8978
36	139	1	0	0.8463	0.0278	0.7825	0.8927
37	138	1	0	0.8402	0.0282	0.7755	0.8875
				(omitido)			
979	21	1	1	0.4561	0.0513	0.3535	0.5528
995	19	1	1	0.4321	0.0540	0.3252	0.5343
1032	17	1	1	0.4066	0.0565	0.2960	0.5143
1141	15	0	2	0.4066	0.0565	0.2960	0.5143
1321	13	0	2	0.4066	0.0565	0.2960	0.5143
1386	11	1	1	0.3697	0.0623	0.2499	0.4896
1400	9	0	1	0.3697	0.0623	0.2499	0.4896
1407	8	0	2	0.3697	0.0623	0.2499	0.4896
1571	6	0	2	0.3697	0.0623	0.2499	0.4896
1586	4	0	2	0.3697	0.0623	0.2499	0.4896
1799	2	0	2	0.3697	0.0623	0.2499	0.4896

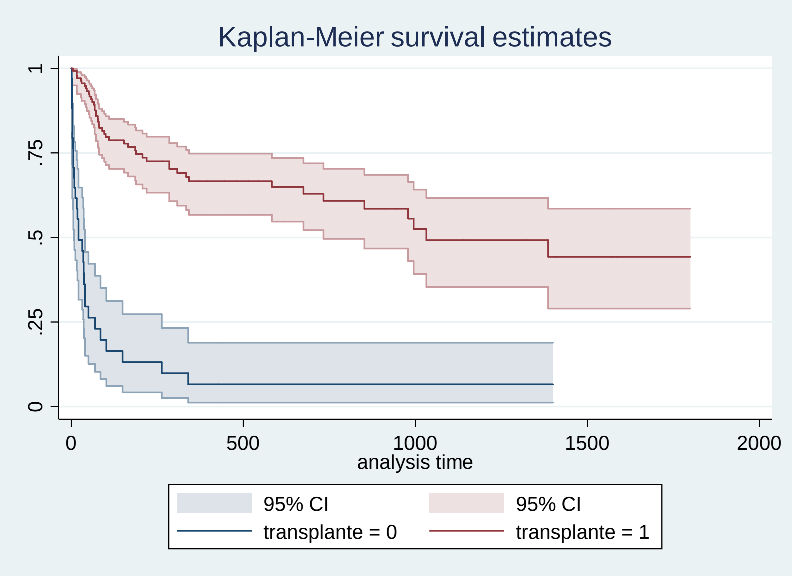
`sts graph`



`sts graph, by(transplante)`

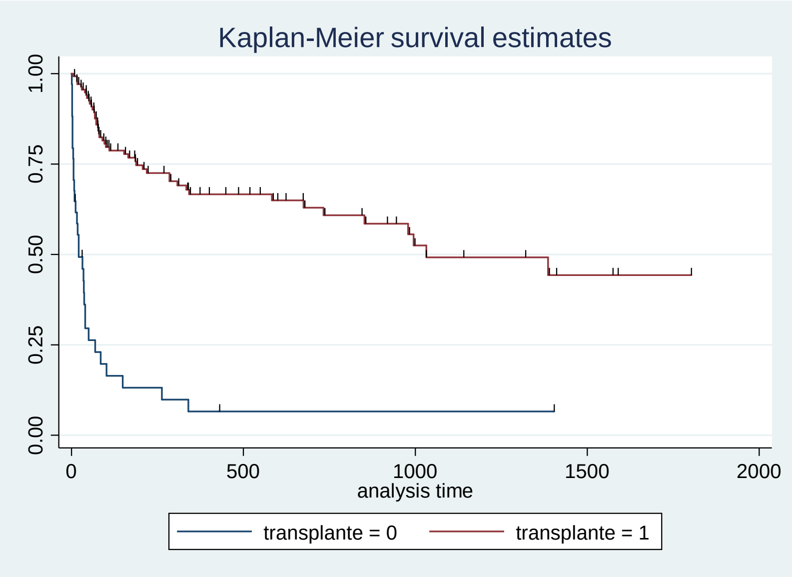


```
sts graph, by(transplante) gw
```



Intervalos de confianza

```
sts graph, by(transplante) cens(s)
```



Muestra las censuras

Comparar analítica de curvas de supervivencia

Log Rank Test

Ho: En la población as curvas de supervivencia son iguales

```
sts test transplante

         failure _d:  fallecer == 1
   analysis time _t:  tiempo


Log-rank test for equality of survivor functions

             |   Events         Events
transplante  |  observed       expected
-------------+-------------------------
0            |        30           6.72
1            |        45          68.28
-------------+-------------------------
Total        |        75          75.00

                  chi2(1) =      91.40
                  Pr>chi2 =     0.0000   --> Rechaza Ho
```

El modelo de riesgos proporcionales de Cox

El objetivo ahora es plantear un modelo de regresión para el riesgo, o la supervivencia, en función de variables "explicatorias", que permita comparar dichas estimaciones, teniendo en cuenta el efecto de otras variables distintas de la que se utiliza para definir los grupos.

Por ejemplo, la supervivencia a dos tratamientos alternativos puede depender no sólo del tratamiento, sino también de otras variables como la edad, el sexo, o la gravedad de la afección de cada paciente. En los métodos previos se asume que el muestreo aleatorio hace que los distintos grupos sean homogéneos con respecto a todas las demás variables, sin embargo no siempre es así (el muestreo aleatorio sólo garantiza que las muestras homogéneas sean las más probables) y, por otro lado, a veces interesa estimar la supervivencia para distintos valores de las otras variables. Los modelos de regresión permiten hacer ambas cosas.

$$h(t,X) = h_o(t)e^{\alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots \alpha_k X_k}$$

Es decir, $h_0(t)$ es el riesgo cuando todas las variables X_i son 0, o riesgo basal, que es variable con el tiempo. Otra forma de expresarlo es:

$$ln \frac{h(t,X)}{h_o(t)} = \alpha_1 X_1 + \alpha_2 X_2 + \ldots\ldots \alpha_k X_k$$

El modelo plantea el logaritmo del riesgo relativo como una función lineal de las variables independientes. Se supone, por lo tanto, que el riesgo relativo, a diferencia del riesgo propiamente dicho, no depende del tiempo o, dicho de otra manera, que es constante a lo largo del tiempo (de ahí el nombre de modelo de riesgo proporcional).

La forma anterior hace explícita la interpretación de los coeficientes: α_i es el logaritmo del riesgo relativo cuando X_i aumenta una unidad, manteniéndose constantes las demás variables, y por tanto, $\exp(\alpha_i)$ es el riesgo relativo cuando X_i aumenta una unidad, manteniéndose constantes las demás.

```
. stcox transplante, hr

         failure _d:  fallecer == 1
   analysis time _t:  tiempo

Iteration 0:   log likelihood = -343.94384
Iteration 1:   log likelihood = -316.68117
Iteration 2:   log likelihood = -316.65424
Iteration 3:   log likelihood = -316.65424
Refining estimates:
Iteration 0:   log likelihood = -316.65424

Cox regression -- Breslow method for ties

No. of subjects =          172                     Number of obs    =         172
No. of failures =           75
Time at risk    =        60591
                                                   LR chi2(1)       =       54.58
Log likelihood  =   -316.65424                     Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 transplante |   .1337891   .0326471    -8.24   0.000     .0829299    .2158391
------------------------------------------------------------------------------
```

En la población de personas con problemas cardiacos, el riesgo de morir es (1-0.1337891)*100% = 86,6% menor en la población que recibe transplante de la que no recibe dicho tratamiento.

Significancia estadística

Añadiendo variables explicativas

```
stcox transplante edad cirugia tespera

         failure _d:  fallecer == 1
   analysis time _t:  tiempo

Iteration 0:   log likelihood = -343.94384
Iteration 1:   log likelihood =  -332.9251
Iteration 2:   log likelihood = -308.58276
Iteration 3:   log likelihood = -304.02902
Iteration 4:   log likelihood = -304.00593
Iteration 5:   log likelihood = -304.00593
Refining estimates:
Iteration 0:   log likelihood = -304.00593

Cox regression -- Breslow method for ties

No. of subjects =          172                  Number of obs    =         172
No. of failures =           75
Time at risk    =        60591
                                                LR chi2(4)       =       79.88
Log likelihood  =   -304.00593                  Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 transplante |   .1259027   .0392659    -6.64   0.000      .068323    .2320082
        edad |   1.059112   .0155365     3.92   0.000     1.029095    1.090005
     cirugia |   .4985788   .2213326    -1.57   0.117     .2088635    1.190159
     tespera |    .992293   .0048643    -1.58   0.115     .9828048    1.001873
------------------------------------------------------------------------------
```

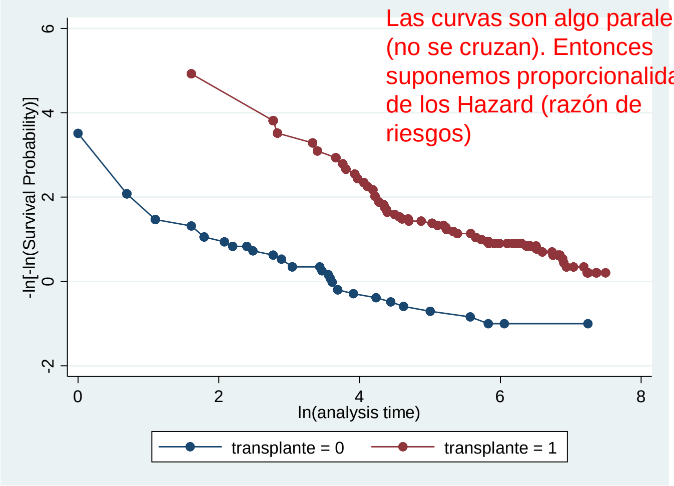
Por cada aumento de unidad en la edad, el riesgo de fallecer aumenta en 6% (1.06), manteniendo constante (en media) las demás variables.

El riesgo de morir se reduce en los pacientes con cirugía en un (1-0.49)x100=50%, respecto a los que no recibieron cirugía, manteniendo constante (en media) las demás variables.

Supuestos del modelo:

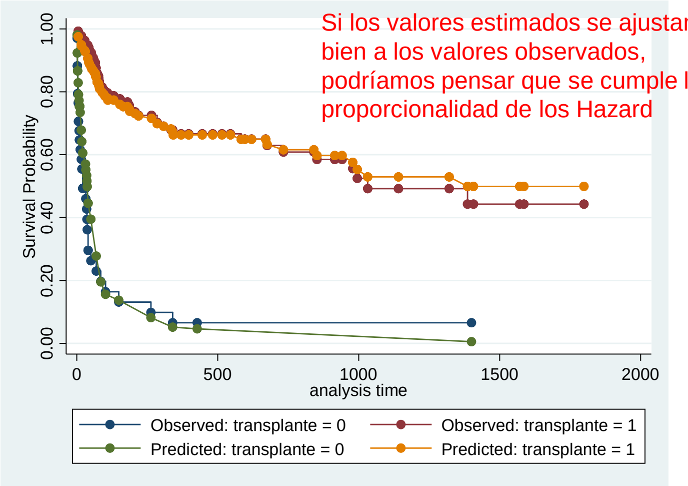
- Observaciones independientes
- Censuras independientes
- Proporcionalidad de los Hazard

```
stphplot, by(transplante)
```



Las curvas son algo paralelas (no se cruzan). Entonces suponemos proporcionalidad de los Hazard (razón de riesgos)

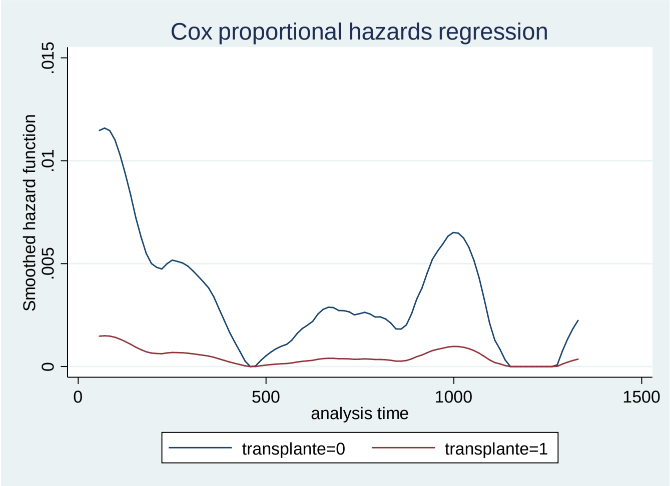
```
stcoxkm, by(transplante)
```



Si los valores estimados se ajustan bien a los valores observados, podríamos pensar que se cumple la proporcionalidad de los Hazard

```
stcurve, hazard at1(transplante=0) at2(transplante=1)
```

No se aprecia las curvas paralelas. No se puede decidir si se cumple o no la condición de proporcionalidad de hazard




```
stphtest
```

```
Test of proportional-hazards assumption

Time:  Time
-----------------------------------------------------------------
            |                     chi2       df        Prob>chi2
------------+----------------------------------------------------
global test |                     8.43        4          0.0769
-----------------------------------------------------------------
```

Ho: Los hazard son proporcionales

Con un riesgo de 0,05, no se puede rechazar Ho, por o tanto, se espera que los hazard sean proporcionales