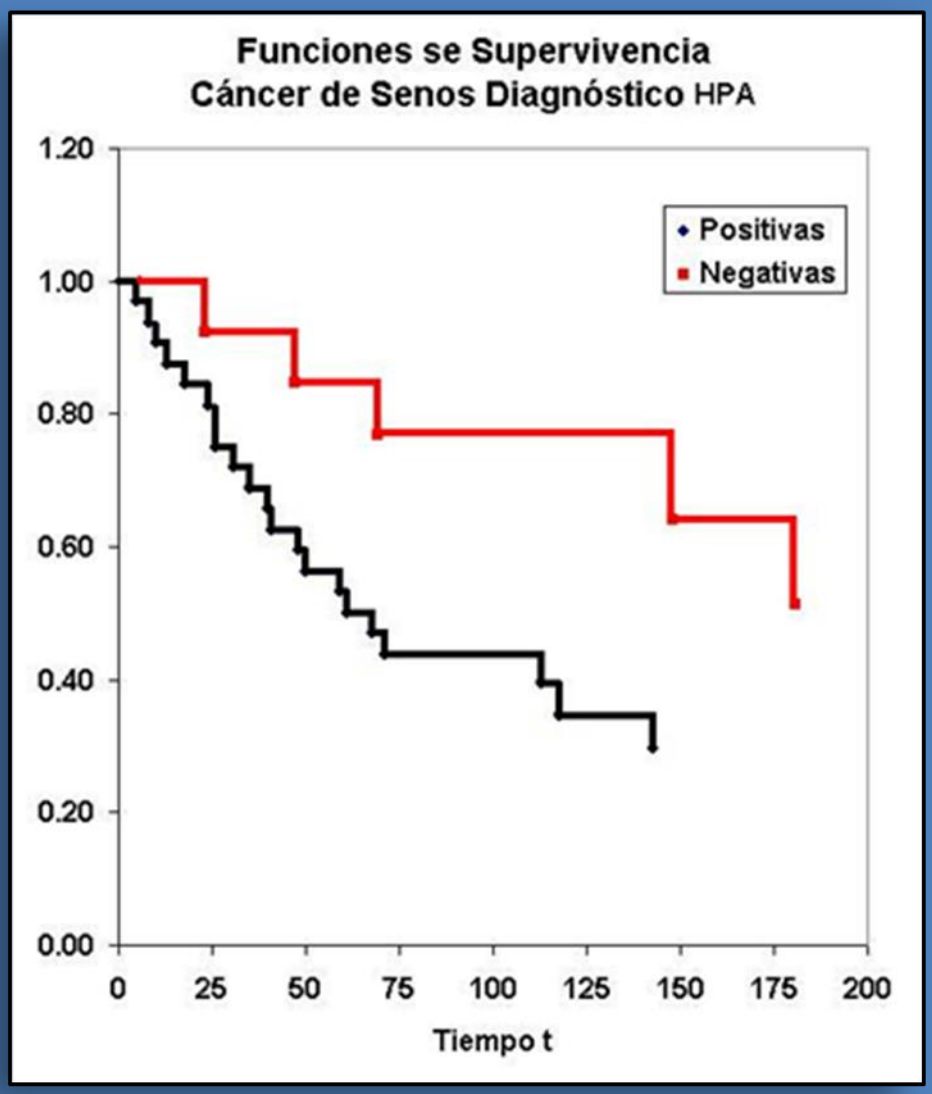
ANÁLISIS DE SUPERVIVENCIA



Profr. Francisco Sánchez Villarreal

2021

ANALISIS DE SUPERVIVENCIA

1. INTRODUCCION

El análisis de supervivencia es una rama de la estadística que estudia, entre otros, eventos vinculados con fallas de máquinas o muertes de organismos. Por ejemplo el tiempo que se espera dure un componente electrónico antes de fallar, el tiempo de duración de la hospitalización de pacientes, el tiempo esperado en la conclusión de la construcción de una presa, etc.

El análisis de supervivencia se interesa por resolver preguntas tales como ¿cuál es la proporción de la población que sobrevivirá pasado cierto tiempo?, ¿cómo afectan la supervivencia determinadas condiciones?. Por lo anterior, pareciera que el análisis de supervivencia se limita a eventos relacionados con la muerte o situaciones análogas tales como fallas de maquinarias, pero de hecho cualquier fenómeno distribuido en el tiempo puede considerarse bajo este enfoque. Sus orígenes se remontan al siglo XVII cuando se despierta el interés por la construcción de tablas de vida.

Los objetivos generales del análisis de supervivencia se pueden resumir de la forma siguiente:

- Estimar el tiempo para la realización de un evento para un determinado grupo de individuos.

- Comparar el tiempo de aparición de un evento para dos o más grupos de individuos que difieren en características controladas.

- Valorar las relaciones entre un conjunto de variables explicativas y la realización de un evento.

Es obvio que el elemento común de estos objetivos es el tiempo que pasa para que un evento se realice.

La definición del tiempo de ocurrencia de un evento debe hacerse en forma precisa y para ello se requieren tres condiciones:

El punto de origen para la medición del tiempo debe ser establecido sin ambigüedades.

La escala de medida del tiempo debe ser consistente. Inicialmente se debe decidir el tipo de unidad en que se midan los intervalos de tiempo o algo análogo.

El significado del evento debe ser precisado sin ambigüedades.

Profundicemos un poco en estos puntos. Se ha mencionado que el punto de origen para la medición del tiempo debe ser establecido sin ambigüedades. En un experimento controlado esto puede quedar definido en forma precisa, como el momento en que se pone a funcionar un equipo, pero existen circunstancias más sutiles, por ejemplo si se estudian pacientes con cáncer de mama, el punto de inicio que probablemente se adoptaría para la enfermedad sería el momento de la detección, la cual se puede dar en forma temprana o cuando la enfermedad está avanzada y ello trae aparejado un sesgo importante.

La escala de medición del tiempo no necesariamente corresponde a tiempo calendario, sino a unidades análogas, por ejemplo la duración de una llanta se mediría en función del kilometraje de rodamiento experimentado antes del primer reventón.

El momento en que se registra la ocurrencia de una falla también puede ser motivo de duda. La falla de un componente electrónico se puede determinar en el momento que deja de funcionar totalmente o cuando se detectan características o comportamientos que reducen su eficacia o eficiencia.

Es indudable que un análisis conceptual cuidadoso previo y la adopción de buenas definiciones operacionales es un requisito importante antes de la aplicación de modelos en éste y en cualquier terreno de investigación.

Los datos originados por un problema de análisis de supervivencia suelen **distribuirse en forma muy asimétrica**, como consecuencia el supuesto de normalidad frecuentemente no es sostenible. La asimetría eventualmente puede ser compensada y como consecuencia impulsar los datos hacia la normalidad mediante transformaciones, como tomar logaritmo natural, pero lo idóneo es buscar modelos más específicos.

2. CENSURA DE DATOS

Otro aspecto que se presenta en forma frecuente es el de **datos censurados**. Se dice que la medición de un individuo está censurada, cuando el sujeto no es observado hasta la presentación del evento objeto de estudio. Por ejemplo en un estudio clínico consistente en una terapia a la que se incorpora un grupo de sujetos y el evento asociado es la recuperación, puede ser que una parte de los sujetos abandonen la terapia antes de llegar a la recuperación o se les pierda el rastro y no se pueda tener conclusión asociada. Una alternativa obvia sería someter al análisis únicamente a los sujetos que completaron todo el estudio, pero esta alternativa, aunque lógica puede inducir sesgos en las conclusiones. Pues supone que los sujetos con información incompleta tienen el mismo comportamiento de aquellos que concluyeron. Otra forma de censura se puede dar de la siguiente forma. Un experimento consiste en registrar el

tiempo de falla de una muestra de 40 dispositivos y como plazo de observación se plantean 2 años, tiempo en el que se supone todos los dispositivos habrán fallado, pero al final de los 2 años quedan 5 dispositivos que continúan funcionando, pero no es posible esperar hasta su momento de falla y se les asigna como duración el máximo observado.

Estas forma de censura se dan al final, pero también puede haber censura al inicio. Por ejemplo, un grupo de pacientes es sometido a una operación y después de 6 meses se revisan para ver si hay indicios de recurrencia, pero puede haber pacientes que presenten la recurrencia en forma muy temprana y para cuando se hace la detección su estado es avanzado y por tanto incierto en inicio de esa recurrencia.

Los sujetos sometidos a estos estudios frecuentemente no se incorporan en forma sincrónica, sino que se van incorporando en forma irregular.

3. FUNDAMENTOS TEORICOS DE LA FUNCIÓN SUPERVIVENCIA.

Existen dos tipos de funciones que son de Interés fundamental en el análisis de supervivencia. La llamada Función de Supervivencia S(t) y la Función de Riesgo h(t).

El tiempo de supervivencia de un sujeto se puede definir como una variable aleatoria T. Los diferentes valores que T tienen una distribución acumulativa de probabilidad asociada *F(t)* con función de densidad *f(t)* definidas ambas en forma genérica como sigue:

$$F(t) = P(T \le t) = \int_0^t f(u)du$$

La función de supervivencia *S(t)* se define como la diferencia respecto de 1 de la función de distribución.

$$S(t) = P(T > t) = 1 - F(t)$$

La función de riesgo se refiere a la probabilidad de que un sujeto pueda morir en el intervalo $(t, t+\delta t)$ condicional a que haya vivido hasta el tiempo t, esto es $P(t \le T < t+\delta t / T \ge t)$. Por ejemplo, si un sujeto vive hasta el año 70 la función de riesgo nos da la probabilidad de que muera durante el año 71. Esta probabilidad expresa la probabilidad por unidad de tiempo al dividirse entre el intervalo δt da lugar a una razón, que al considerar al incremento δt muy pequeño, la función de riesgo se expresa formalmente como el límite siguiente:

$$h(t) = l\ \underset{\delta t \to 0}{im} \left\{ \frac{P(t \le T < t+\delta t / T > t}{\delta t} \right\}$$

La función de riesgo también se refiere como *tasa instantánea de mortalidad* o *fuerza de mortalidad*.

Recordando algunos resultados elementales del cálculo de probabilidades, la probabilidad de que suceda un evento A dado que ha sucedido un evento B se expresa como

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Se usará este resultado para obtener la función de riesgo. La probabilidad condicional se expresa:

$$P(t \le T < t + \delta t / T > t) = \frac{P(t \le T < t + \partial t)}{P(T > t)}$$

Esta fórmula expresada en términos de la función de distribución acumulativa F(t)b y la función de supervivencia S(t) nos da:

$$\frac{P(t < T < t + \delta t)}{P(T > t)} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

Ahora como la función de riesgo se define a través del límite:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t < T < t + \delta t / T > t}{\delta t} \right\}$$

$$= \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{S(t)\delta t} \right\} = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}$$

Pero el límite en el extremo izquierdo corresponde a la derivada de F(t) respecto a t y esta derivada es igual a la función de densidad f(t)

$$\lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} = \frac{dF(t)}{dt} = f(t)$$

De donde se concluye que la función de riesgo es el cociente de la función de densidad entre la función de supervivencia.

$$h(t) = \frac{f(t)}{S(t)}$$

Ahora considérese a S(t) y obtengamos la derivada de su logaritmo respecto a t

$$Ln(S(t)) = Ln(1 - F(t))$$

Por tanto

$$\frac{dLn(S(t))}{dt} = \frac{dLn(1 - F(t))}{dt}$$

$$= \frac{-f(t)}{1 - F(t)} = -\frac{f(t)}{S(t)}$$

De donde se concluye que la función de riesgo también se expresa como

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

Si se integra en ambos miembros se tendrá

$$\int_0^t h(u)du = -\int_0^t \frac{d}{dt} Ln(S(u))du = -Ln(S(t))$$

Si se define a la integral $H(t) = \int_0^t h(u)du$ y se la identifica como función acumulativa de riesgo, la conclusión es la siguiente.

$S(t) = \exp(-H(t))$, en forma equivalente $H(t) = -Ln(S(t)$

Supóngase que una función de supervivencia tiene como función de densidad la siguiente expresión, exponencial negativa con parámetro igual a 1.

$f(t) = e^{-t}$ para t ≥ o

La función de distribución acumulativa F(t) se obtiene por integración de f(t)

$$F(t) = \int_0^t f(x)dx = \int_0^t e^{-x}dx = -e^{x}\Big|_0^t = 1 - e^{-t}$$

La función de supervivencia S(t) se obtiene al restar F(t) de 1

$S(t) = 1 - F(t) = 1 - (1 - e^{-t}) = e^{-t}$

La función de riesgo h(t) resulta ser, en este caso, igual a 1, pues f(t) y S(t) coinciden.

$$h(t) = \frac{f(t)}{S(t)} = \frac{e^{-t}}{e^{-t}} = 1$$

Finalmente la función de riesgo acumulativa H(t) es igual a la idéntica

$$H(t) = -\ln(S(t)) = -\ln(e^{-t}) = t$$

Tanto la función de supervivencia, como la función de riesgo se estiman a partir de los datos observados. Los métodos de estimación que no requieren de un supuesto de distribución se denominan no paramétricos o de distribución libre y se designa como métodos paramétricos a los que suponen una función de distribución subyacente.

4. METODOS NO PARAMETRICOS DE ESTIMACION.

4.1 Modelo de Kaplan Meier

Una forma sencilla de iniciar el análisis de supervivencia es mediante métodos descriptivos con la ayuda de tablas y gráficas que permiten presentar en forma resumida los datos sobre la supervivencia de uno o más grupos de sujetos. Los datos de supervivencia se resumen mediante la estimación empírica de funciones de supervivencia y riesgo; así como algunas estadísticas, entre las cuales son frecuentes estadísticas de orden como la mediana y otras medidas percentilares. Ello permite llegar a resultados muy útiles que no se apoyan en una distribución teórica de probabilidad, de ahí la designación genérica de métodos no paramétricos. Los métodos no paramétricos pueden utilizarse como paso previo de análisis con métodos paramétricos.

Uno de los métodos más frecuentemente utilizados en el análisis de datos censurados con un número moderado de observaciones es el de Kaplan Meier (1958) también conocido como *producto límite*.

Suponga como ejemplo que se tiene un grupo de 10 pacientes que ingresan a un estudio al inicio del año 2000, al final del año 6 pacientes han muerto y 4 sobreviven. Al inicio del año 2001, ingresan 20 nuevos pacientes y para el final de ese año 15 de los 20 adicionales han muerto y también han muerto 3 de los 4 que quedaban del año 2000.

Si se desea saber el valor de la función de supervivencia para t=2, la forma más simple consiste en ignorar a los 20 pacientes adicionals y fijarse solamente en los 10 iniciales, que han sido observados por 2 años, esto es S(2) = 1/10. El método propuesto por Kaplan y Meier (1958) considera que es aprovechable, al menos parcialmente, la información aportada por los pacientes que ingresaron en el año 2001.

Los pacientes que sobreviven 2 años se pueden considerar como la proporción de los sobrevivientes al segundo año, que previamente sobrevivieron el primer año.

$\hat{S}(2) = (\text{Proporción de pacientes que sobreviven 2 años, dado que sobrevieron al primer año})\text{X}$
$(\text{Proporción de pacientes que sobreviven 1 año})$

La probabilidad de supervivencia del primer año se puede obtener como la suma de los sobrevivientes al final del año 2000 del primer grupo (4) más los sobrevivientes del segundo grupo que sobrevivieron un año (5), esta suma dividida entre la suma de los pacientes que ingresaron el primer año(10), más los que ingresaron el segundo año (20)

$$\hat{S}(1) = \frac{4+5}{10+20} = 0.3$$

Ahora bien, solamente hay un sobreviviente después de 2 años de los 4 que previamente sobrevivieron al primer año. Esto es su probabilidad de sobrevivencia en el segundo año se estima como la proporción P(T=2) = ¼

De esta forma, la *S(2)* queda estimada como el producto:

S(2) = (0.3)x(0.25) = 0.075

Este sencillo razonamiento se puede generalizar de la siguiente forma:

Sea P_1 La probabilidad de sobrevivir el primer período

P_2 La probabilidad de sobrevivir el segundo período dado que se sobrevivió al primero

P_t La probabilidad de sobrevivir el período ***t*** dado que se sobrevivió al período ***t-1***

La probabilidad de sobrevivir hasta el período t esto es S(t) se obtiene como el producto de las probabilidades consecutivas.

$$S(t) = P_1 P_2 \ldots\ldots\ldots P_t = \prod_{i=1}^{t} P_i$$

Consideremos a n_t es el número de sobrevivientes al inicio del período t

d_t es el número de muertes en el período t

$n_t - d_t$ son los sobrevivientes al final del período t

La probabilidad de sobrevivir al periodo t se puede estimar como el cociente de los sobrevivientes al final del período entre los sobrevivientes al inicio del período.

$$\hat{P}_t = \frac{n_t - d_t}{n_t} = 1 - \frac{d_t}{n_t}$$

De donde la función de supervivencia resulta $\hat{S}(t) = \prod_{i=1}^{t}\left(1 - \frac{d_i}{n_i}\right)$

Este producto se conoce como el estimador *Producto Límite* de Kaplan Meier. Observe que se cumple una relación iterativa.

$$\hat{S}(t) = \left(1 - \frac{d_t}{n_t}\right)\hat{S}(t-1)$$

Si no se tienen observaciones censuradas se tiene el siguiente resultado, causado por la cancelación de los productos sucesivos.

$$\hat{S}(t) = \prod_{i=1}^{t}\left(1 - \frac{d_i}{n_i}\right)$$

$$= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right).....\left(1 - \frac{d_{t-1}}{n_{t-1}}\right)\left(1 - \frac{d_t}{n_t}\right)$$

$$= \left(\frac{n_1 - d_1}{n_1}\right)\left(\frac{n_2 - d_2}{n_2}\right).....\left(\frac{n_{t-1} - d_{t-1}}{n_{t-1}}\right)\left(\frac{n_t - d_t}{n_t}\right) \text{ como } n_k = n_{k-1} - d_{k-1}$$

$$= \left(\frac{n_2}{n_1}\right)\left(\frac{n_3}{n_2}\right).....\left(\frac{n_t}{n_{t-1}}\right)\left(\frac{n_t - d_t}{n_t}\right)$$

$$= \frac{n_t - d_t}{n_1}$$

Ejemplo 1. Considere una muestra de tiempos de supervivencia en meses de 10 pacientes. Con estos datos, ordenados de menor a mayor y calculadas las frecuencias de muertes a cada tiempo se construye la función de supervivencia.

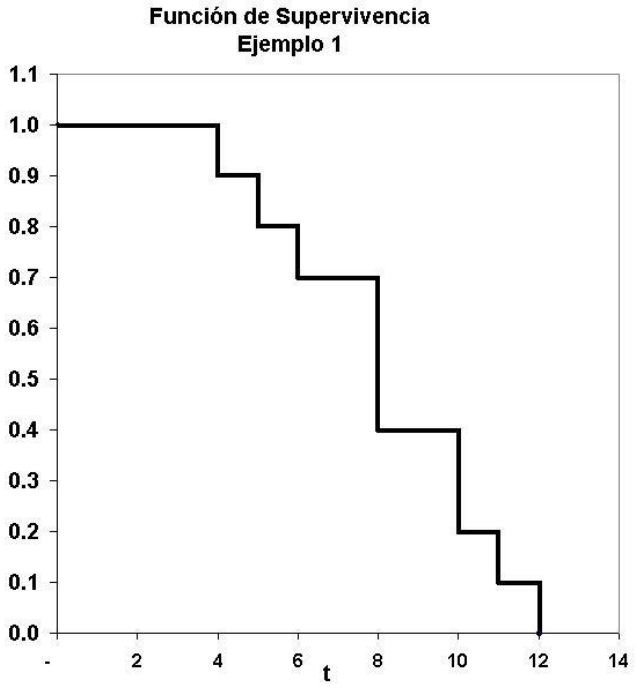
4 11 8 6 10 8 10 12 5 8

Figura 1. Función de Supervivencia del Ejemplo1

Tiempo de Supervivencia t	Sujetos vivos al inicio de t n_t	Sujetos muertos durante t d_t	$1-\frac{d_t}{n_t}$	$\hat{S}(t)$
4	10	1	0.900000	0.90000
5	9	1	0.888889	0.80000
6	8	1	0.875000	0.70000
8	7	3	0.571429	0.40000
10	4	2	0.500000	0.20000
11	2	1	0.500000	0.10000
12	1	1	0.000000	0.00000
		10		

La gráfica escalonada de la función de supervivencia se presenta a continuación:

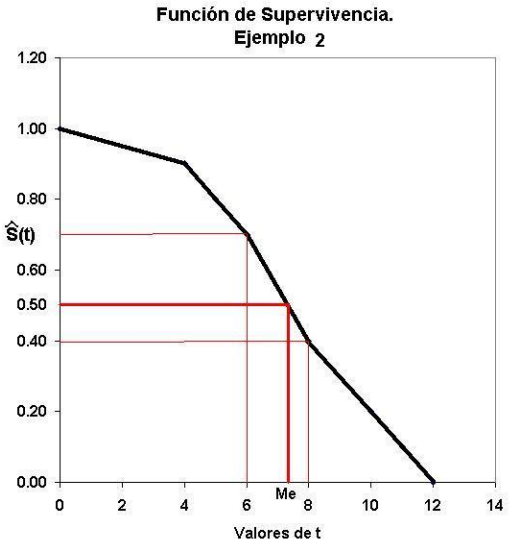
Figura 2 Gráfica de Función de Supervivencia



4.3 Estadísticas Básicas

Como el tiempo de supervivencia es una variable que en general se distribuye en forma muy asimétrica, la media aritmética resulta ser inconveniente como medida de tendencia central y es preferible usar la mediana cuando se quiere caracterizar en forma más específica a una función de supervivencia.

La mediana como se sabe es el valor que separa a la función de distribución acumulativa en dos regiones de valor 0.5. En una muestra de tamaño n la mediana se toma como el valor de la observación central si n es impar, o el promedio de las dos observaciones centrales si n es par. En el ejemplo 3, de acuerdo a esta definición, la mediana vale 8. Para datos agrupados se obtiene por interpolación lineal de los valores extremos del intervalo de contiene a la mediana. Para propósitos del cálculo de la mediana, la función empírica de supervivencia se suele adoptar en forma suavizada para evitar la discontinuidad que provocan los saltos abruptos de su forma escalonada. Así, en este ejemplo la gráfica suavizada de la función de supervivencia adopta la siguiente forma:



La gráfica muestra que el valor 0.5 se encuentra entre los valores S(6) y S(8) y en consecuencia la mediana se obtiene por interpolación entre *t=6 y t=8*.

Si se designa a *t=6* y *t'=8* y los respectivos valores de la función de supervivencia *S(t)=0.7 y S(t')=0.4.* La fórmula de interpolación queda:

$$Me = t + \frac{(t' - t)(\hat{S}(t) - 0.5)}{\hat{S}(t) - \hat{S}(t')}$$

Por lo tanto Me = 0.7333

4.4 Función de Riesgo y otras funciones relacionadas.

La función de densidad se aproxima mediante la proporción de muertes en el período t respecto del total de casos.

$$\hat{f}(t) = \frac{d_t}{n}$$

La función de riesgo *h(t)* es conocida también como tasa instantánea de mortalidad, fuerza de mortalidad, tasa condicional de mortalidad y tasa de mortalidad a edad específica. La *h(t)* proporciona el riesgo de mortalidad por unidad de tiempo.

Cuando no se tienen datos censurados, la función de riesgo es estimada como la proporción de pacientes que mueren en un intervalo por unidad de tiempo, dado que ellos han sobrevivido hasta el inicio del intervalo.

$$\hat{h}(t) = \frac{\text{Número de sujetos que mueren durante el tiempo t}}{\text{Número de sobrevivientes al inicio del tiempo t}}$$

De donde se tiene la aproximación de la *h(t)*

$$\hat{h}(t) = \frac{d_t}{n_t}$$

Función Acumulativa de riesgo se aproxima como el logaritmo natural del la función estimada de supervivencia.

$$\hat{H}(t) = -\ln(\hat{S}(t))$$

Ejemplo 2. Considere los siguientes datos de supervivencia de 22 sujetos.

Sujeto	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Sobrevivencia	10	11	12	13	13	14	14	14	14	14	15

Sujeto	12	13	14	15	16	17	18	19	20	21	22
Sobrevivencia	15	15	15	16	17	17	18	18	19	19	21

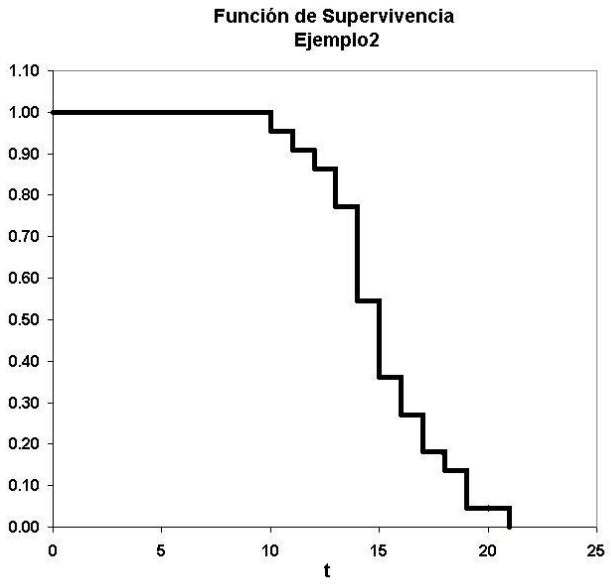
En la tabla siguiente se calcula la función de supervivencia junto con las otras funciones mencionadas en este apartado.

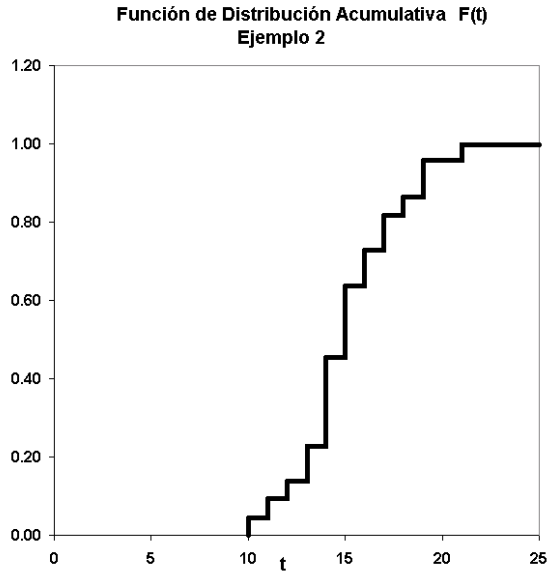
Figura 3. Función de Supervivencia y otras funciones relacionadas para el Ejemplo 2.

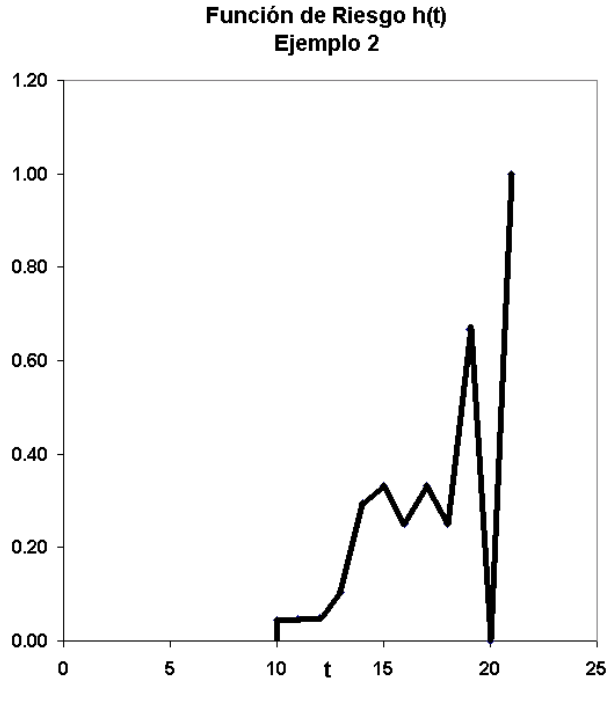
Tiempo de Supervivencia t	**Sujetos vivos al inicio de t** n_t	**Sujetos muertos durante t** d_t	$1-\frac{d_t}{n_t}$	$\hat{S}(t)$	$\hat{F}(t)$	$\hat{f}(t)$	$\hat{h}(t)$	$\hat{H}(t)$
10	**22**	**1**	**0.95455**	**0.95455**	**0.04545**	**0.04545**	**0.04545**	**0.04652**
11	**21**	**1**	**0.95238**	**0.90909**	**0.09091**	**0.04545**	**0.04762**	**0.09531**
12	**20**	**1**	**0.95000**	**0.86364**	**0.13636**	**0.04545**	**0.05000**	**0.14660**
13	**19**	**2**	**0.89474**	**0.77273**	**0.22727**	**0.09091**	**0.10526**	**0.25783**
14	**17**	**5**	**0.70588**	**0.54545**	**0.45455**	**0.22727**	**0.29412**	**0.60614**
15	**12**	**4**	**0.66667**	**0.36364**	**0.63636**	**0.18182**	**0.33333**	**1.01160**
16	**8**	**2**	**0.75000**	**0.27273**	**0.72727**	**0.09091**	**0.25000**	**1.29928**
17	**6**	**2**	**0.66667**	**0.18182**	**0.81818**	**0.09091**	**0.33333**	**1.70475**
18	**4**	**1**	**0.75000**	**0.13636**	**0.86364**	**0.04545**	**0.25000**	**1.99243**
19	**3**	**2**	**0.33333**	**0.04545**	**0.95455**	**0.09091**	**0.66667**	**3.09104**
20	**1**	**0**	**1.00000**	**0.04545**	**0.95455**	**0.00000**	**0.00000**	
21	**1**	**1**	**0.00000**	**0.00000**	**1.00000**	**0.04545**	**1.00000**	

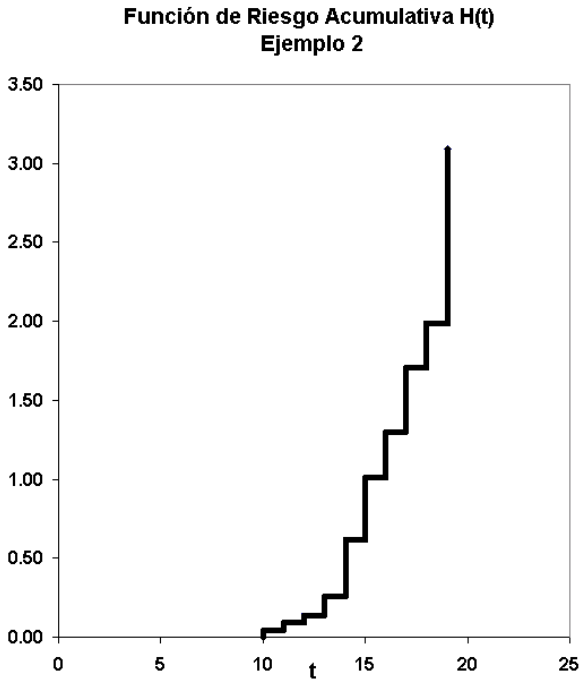
Gráficas de las funciones de supervivencia, de distribución acumulativa, de riesgo y función acumulativa de riesgo.

Figura 4. Gráficas de Funciones de Supervivencia, de Distribución, de Riesgo y Acumulativa de Riesgo correspondientes al ejemplo2.


Función de Supervivencia
Ejemplo2
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
5
10
15
20
25
t


Función de Distribución Acumulativa F(t)
Ejemplo 2
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
t


Función de Riesgo h(t)
Ejemplo 2
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
t


Función de Riesgo Acumulativa H(t)
Ejemplo 2
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0
5
10
15
20
25
t

4.5 Varianza y Error Estándar de la Función de Supervivencia.

La confiabilidad de los valores asociados a la función de supervivencia, se puede obtener si se dispone de intervalos de confianza. La forma usual de un intervalo de 100(1-α)% de confianza, bajo el supuesto de normalidad del estimador es la siguiente:

$$S(t) \pm Z_{1-\alpha/2} EE(S(t))$$

Una de las fórmulas más usuales para calcular la varianza y el error estándar de la función de supervivencia se debe a Greenwood (1926):

$$V(S(t)) = S^2(t) \sum_{i=0}^{t-1} \frac{d_j}{n_j (n_j - d_j)}$$

$$EE(S(t)) = S(t) \sqrt{\sum_{i=0}^{t-1} \frac{d_j}{n_j (n_j - d_j)}}$$

Demostración

$S(t) = \prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i}\right)$ En términos de Pj $S(t) = \prod_{i=1}^{t} P_j$

Se toma logaritmo

$Ln\left(S(t)\right) = Ln\left(\prod_{i=1}^{t} P_j\right)$ En forma equivalente $Ln\left(S(t)\right) = \sum_{j=1}^{t} Ln\left(P_j\right)$

Expresamos la varianza del logaritmo de la función de supervivencia.

$$V\left(Ln\left(S(t)\right)\right) = V\left(\sum_{j=1}^{t} Ln\left(P_j\right)\right)$$

Se consideran independientes la Pj y la varianza de la suma es igual a la suma de las varianzas.

$$V\left(Ln\left(S(t)\right)\right) = \sum_{j=1}^{t} V\left(Ln\left(P_j\right)\right)$$

Delta Método

Aplicando el Delta Método el cual se usa para obtener una aproximación de la distribución de probabilidad de un estimador estadístico asintóticamente normal y por lo tanto una estimación de la varianza del estimador.

Para su aplicación se requiere un estimador Tn cuya distribución depende de θ y una distribución asintótica.

Teorema. Sea Tn una sucesión de variables aleatorias (estimadores), cuya distribución depende de θ tal que tienda en distribución a una normal N(0,σ^2), cuando n →∞

$$\sqrt{n}\left(Tn-\theta\right)\xrightarrow{D} N\left(0,\sigma^2\right)$$

Entonces si g(θ) es una función diferenciable tal que g'(θ) ≠ 0 se verifica que cuando n →∞

$$\sqrt{n}\left(g\left(Tn\right)-g\left(\theta\right)\right)\xrightarrow{D} N\left(0,\sigma^2\left[g'(\theta)\right]^2\right)$$

Ahora consideremos una variable aleatoria X que se distribuye Binomial (n,P). Su esperanza y varianza se expresan:

$$E(X)=nP \quad V(X)=nP(1-P)$$

La siguiente transformación se aproxima a una normal estándar N(0.1)

$$Z=\frac{X-nP}{\sqrt{nP(1-P}}\rightarrow N(0.1)$$

En forma equivalente al dividir numerador y denominador entre n

$$Z=\frac{\frac{X}{n}-P}{\sqrt{\frac{P(1-P}{n}}}=\frac{\sqrt{n}\left(\frac{X}{n}-P\right)}{\sqrt{P(1-P}}\rightarrow N(0.1)$$

Entonces al multiplicar Z por el denominador se produce una cancelación y queda X/n y la varianza de la transformación que es ahora P(1-P)/n.

$$Z\sqrt{P(1-P}=\frac{X}{n}-P\rightarrow N(0.\frac{P(1-P)}{n})$$

Ahora se aplica el Método Delta a través de la función logarítmica. La derivada del logaritmo de P es 1/P, cuyo cuadrado se multiplica por la varianza anterior P(1-P)

$$\sqrt{n}\left[Ln\left(\frac{X}{n} \right) - Ln\left(P \right) \right] \to N\left(0.\frac{P(1-P)}{P^2} \right)$$

De donde se concluye.

$$V\left(\frac{X}{n} \right) = \frac{P(1-P)}{n}$$

Se vuelve al planteamiento de la varianza del logaritmo de S(t).

$$V\left(Ln\left(S(t) \right) \right) = \sum_{j=1}^{t} V\left(Ln\left(P_j \right) \right)$$

Al aplicar el Delta Método nuevamente.

$$V\left(Ln\left(P_j \right) \right) \approx \frac{1}{P_j^2} \frac{P_j\left(1-P_j \right)}{n_j}$$

$$V\left(Ln\left(S(t) \right) \right) = \sum_{j=1}^{t-1} \frac{1}{P_j^2} \frac{P_j(1-P_j)}{n_j}$$

Si nos apoyamos en la siguiente aproximación al aplicar nuevamente el Método Delta

$$V\left(\left(S(t) \right) \right) \approx S^2(t) V\left(Ln(S(t)) \right)$$

Se sustituye la varianza del Logaritmo natural de S(t) y se cancela una Pj

$$V\left(\left(S(t) \right) \right) = S^2(t) \sum_{j=1}^{t-1} \frac{1}{P_j} \frac{(1-P_j)}{n_j}$$

Considerando que la probabilidad de sobrevivir el periodo j es:

$$P_j = \frac{n_j - d_j}{n_j}$$

Al sustituir en la fórmula anterior, finalmente se tiene la fórmula de Greenwood

$$V\left(\left(S(t) \right) \right) = S^2(t) \sum_{j=1}^{t-1} \frac{d_j}{n_j\left(n_j - d_j \right)}$$

En la siguiente tabla se ilustra el cálculo del error estándar con la fórmula de Greenwood para la función de supervivencia del Ejemplo2 y los intervalos de 95% de confianza bajo el supuesto de normalidad.

Figura 5. Cálculo del Error estándar e intervalos de confianza para la función de supervivencia del Ejemplo 2.

Tiempo de Supervivencia t	Sujetos vivos al inicio de t n_t	Sujetos muertos durante t d_t	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j=0}^{t-1}\frac{d_j}{n_j(n_j-d_j)}$	EE S(t)	Límte Inferior	Límte Superior
10	22	1	0.00216	0.00216	0.04441	0.86750	1.04159
11	21	1	0.00238	0.00455	0.06129	0.78896	1.02922
12	20	1	0.00263	0.00718	0.07317	0.72023	1.00704
13	19	2	0.00619	0.01337	0.08935	0.59761	0.94785
14	17	5	0.02451	0.03788	0.10616	0.33738	0.75353
15	12	4	0.04167	0.07955	0.10256	0.16262	0.56465
16	8	2	0.04167	0.12121	0.09495	0.08662	0.45883
17	6	2	0.08333	0.20455	0.08223	0.02065	0.34299
18	4	1	0.08333	0.28788	0.07317	-0.00704	0.27977
19	3	2	0.66667	0.95455	0.04441	-0.04159	0.13250
20	1	0	0.00000	0.95455	0.04441	-0.04159	0.13250
21	1	1	0.00000	0.95455	0.00000	0.00000	0.00000
		22					

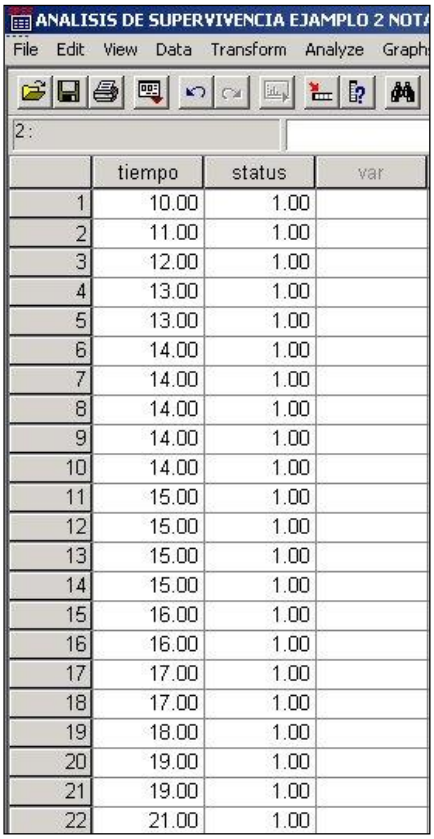
4.6 Cálculo de Kaplan Meier mediante SPSS

El paquete de software estadístico SPSS cuenta con un procedimiento para calcular la estimación de la función de supervivencia mediante el procedimiento Kaplan Meier.

Primero se consideran los 22 datos del ejemplo 2, el cual carece de observaciones censuradas.

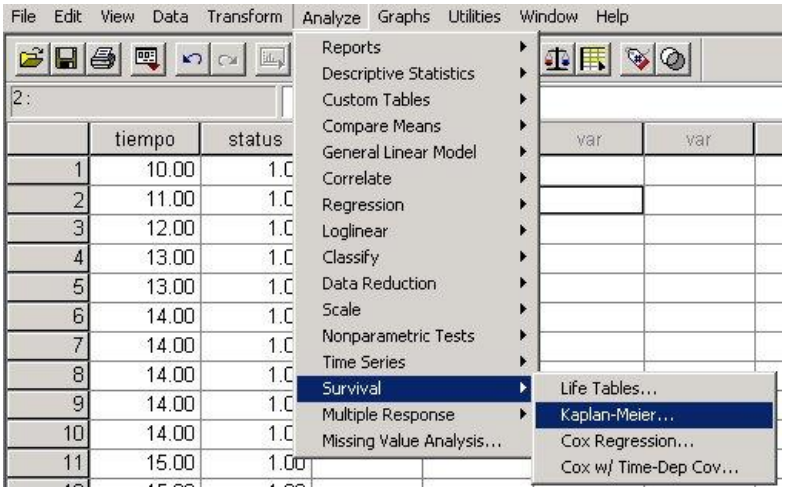
En primer término el usuario debe incorporar los datos de supervivencia en dos variables, la primera que se identificará como Tiempo contiene la supervivencia de cada caso, la segunda es una variable etiqueta que se denomina Statuts y que identifica los casos que llegaron a término con 1 y los casos censurados con cero. Como en este ejemplo no hay observaciones censuradas, todos los casos se marcaron con 1.

A continuación se selecciona la pestaña de procedimientos de análisis de datos (Analyze) y dentro de ésta la opción Kaplan Meier.

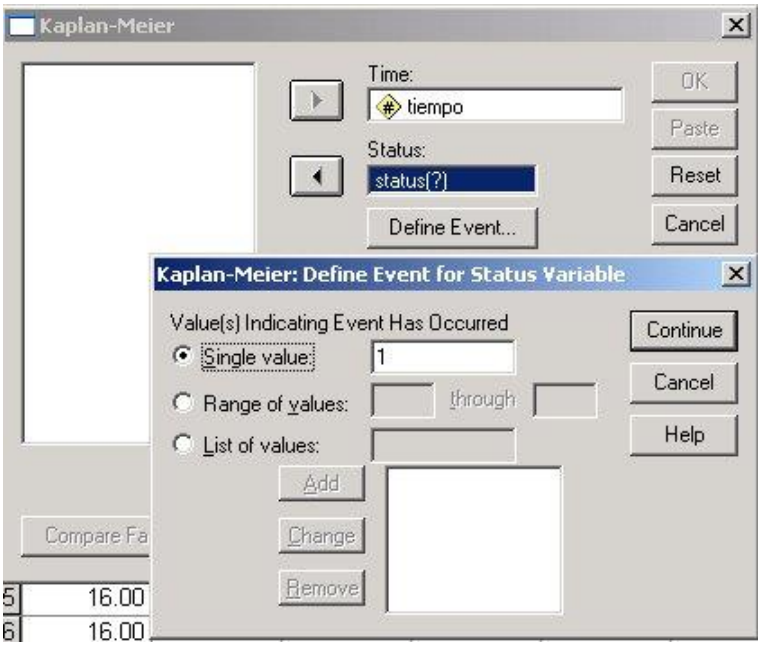


	tiempo	status	var
1	10.00	1.00	
2	11.00	1.00	
3	12.00	1.00	
4	13.00	1.00	
5	13.00	1.00	
6	14.00	1.00	
7	14.00	1.00	
8	14.00	1.00	
9	14.00	1.00	
10	14.00	1.00	
11	15.00	1.00	
12	15.00	1.00	
13	15.00	1.00	
14	15.00	1.00	
15	16.00	1.00	
16	16.00	1.00	
17	17.00	1.00	
18	17.00	1.00	
19	18.00	1.00	
20	19.00	1.00	
21	19.00	1.00	
22	21.00	1.00	

El procedimiento Kaplan Meier de SPSS abre una ventana de diálogo para que el usuario seleccione las variables asociadas al tiempo y al status del caso. En la variable Status se debe especificar el valor asociado a los casos completos, el 1 en nuestro caso. Este valor se debe especificar en la ventana que se abre al oprimir el botón **Define Event**



Una vez asignado este valor se oprime el botón Continue para cerrar la ventana que define el status y el botón OK de la ventana Kaplan Meier y así se pueden obtener resultados básicos en la forma siguiente:



Reporte de Resultqados de SPSS para Kaplan Meier

Kaplan-Meier

Survival Analysis for TIEMPO

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
10.00	1.00	.9545	.0444	1	21
11.00	1.00	.9091	.0613	2	20
12.00	1.00	.8636	.0732	3	19
13.00	1.00			4	18
13.00	1.00	.7727	.0893	5	17
14.00	1.00			6	16
14.00	1.00			7	15
14.00	1.00			8	14
14.00	1.00			9	13
14.00	1.00	.5455	.1062	10	12
15.00	1.00			11	11
15.00	1.00			12	10
15.00	1.00			13	9
15.00	1.00	.3636	.1026	14	8
16.00	1.00			15	7
16.00	1.00	.2727	.0950	16	6
17.00	1.00			17	5
17.00	1.00	.1818	.0822	18	4
18.00	1.00	.1364	.0732	19	3
19.00	1.00			20	2
19.00	1.00	.0455	.0444	21	1
21.00	1.00	.0000	.0000	22	0

Number of Cases: 22 Censored: 0 (.00%) Events: 22

	Survival Time	Standard Error	95% Confidence Interval	
Mean:	15.09	.57	(13.97,	16.21)
Median:	15.00	.50	(14.02,	15.98)

Compare la coincidencia de los valores de la columna *Cumulative Survival* del reporte SPSS con los de la columna S(t) de la tabla identificada como Figura 3 y la columna Standar Error con la columna EE(S(t)) de la tabla en la Figura 5.

4.7 Cálculo de Kaplan Meier con Casos Censurados.

El procedimiento de cálculo de Kaplan Meier con la presencia de casos censurados adopta ciertas particularidades. Como Ejemplo 3 considérense los siguientes datos correspondientes a pacientes de cáncer de colon que mencionan McIllmurray y Turkie publicados en Br. Med J. 294 (1987).

Paciente	Supervivencia t	Censurado Ct
1	3	+
2	6	
3	6	
4	6	
5	6	
6	8	
7	8	
8	12	
9	12	
10	12	+
11	15	+
12	16	+

Paciente	Supervivencia t	Censurado Ct
13	18	+
14	18	+
15	20	
16	22	+
17	24	
18	28	+
19	28	+
20	28	+
21	30	
22	30	+
23	33	+
24	42	

A continuación se presenta la tabla de cálculos, cuyos datos de base son el número de supervivientes, muertes y observaciones censuradas en cada tiempo. La n_t se ajusta en cada tiempo en función de las muertes y de los casos censurados. La ***S(t)*** se omite en renglones que incluyen únicamente observaciones censuradas.

$$n_t = n_{t-1} - d_{t-1} - C_{t-1}$$

$$P_t = 1 - \frac{d_t}{n_t} \qquad H(t) = -Ln(S(t))$$

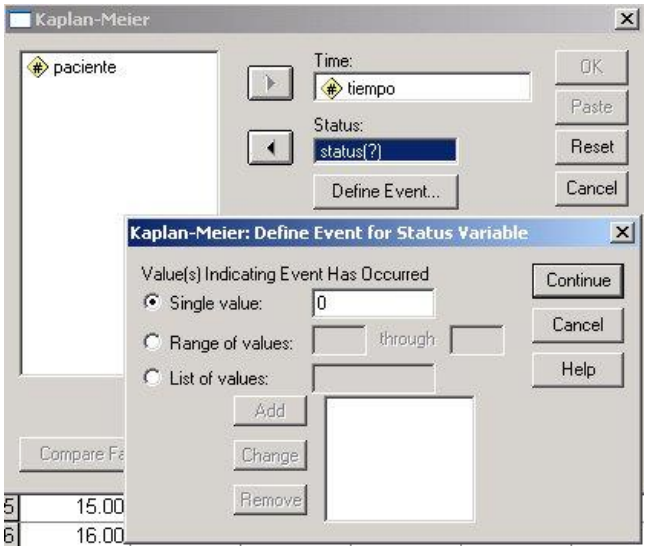
Figura 7. Calculo de la Función de Supervivencia. Ejemplo 3

t	n_t	d_t	C_t	P_t	$\prod_{i=1}^{t} P_i$	$S(t)$	$H(t)$
3	24	0	1	1.00000			
6	23	4	0	0.82609	0.82608696	0.826087	0.1910552
8	19	2	0	0.89474	0.73913043	0.739130	0.3022809
12	17	2	1	0.88235	0.65217391	0.652174	0.4274440
15	14	0	1	1.00000	0.65217391		
16	13	0	1	1.00000	0.65217391		
18	12	0	2	1.00000	0.65217391		
20	10	1	0	0.90000	0.58695652	0.586957	0.5328045
22	9	0	1	1.00000	0.58695652		
24	8	1	0	0.87500	0.51358696	0.513587	0.6663359
28	7	0	3	1.00000	0.51358696		
30	4	1	1	0.75000	0.38519022	0.385190	0.9540180
33	2	0	1	1.00000	0.38519022		
42	1	1	0	0.00000	0	0.000000	

4.8 SPSS Cálculo de Kaplan Meier con Casos Censurados.

	paciente	tiempo	status
1	1.00	3.00	1.00
2	2.00	6.00	.00
3	3.00	6.00	.00
4	4.00	6.00	.00
5	5.00	6.00	.00
6	6.00	8.00	.00
7	7.00	8.00	.00
8	8.00	12.00	.00
9	9.00	12.00	.00
10	10.00	12.00	1.00
11	11.00	15.00	1.00
12	12.00	16.00	1.00
13	13.00	18.00	1.00
14	14.00	18.00	1.00
15	15.00	20.00	.00
16	16.00	22.00	1.00
17	17.00	24.00	.00
18	18.00	28.00	1.00
19	19.00	28.00	1.00
20	20.00	28.00	1.00
21	21.00	30.00	.00
22	22.00	30.00	1.00
23	23.00	33.00	1.00
24	24.00	42.00	.00

Si se utiliza el SPSS para calcular la función de supervivencia del Ejemplo 3, en lo único que debe poner atención el usuario es en disponer su archivo de datos con 3 variables: la identificación, la variable correspondiente al tiempo de supervivencia y la variable de status para identificar si el caso es o no censurado. Esta última variable es una variable dicotómica (0,1). El usuario deberá determinar que valor está asociado a la realización del evento (muerte). En el ejemplo, los datos censurados se etiquetan con 1 y los concluyentes con 0. El 0 es el que se incorpora al SPSS.



El reporte emitido por SPSS adopta la siguiente forma:

Survival Analysis for TIEMPO

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
3.00	.00			0	23
6.00	1.00			1	22
6.00	1.00			2	21
6.00	1.00			3	20
6.00	1.00	.8261	.0790	4	19
8.00	1.00			5	18
8.00	1.00	.7391	.0916	6	17
12.00	1.00			7	16
12.00	1.00	.6522	.0993	8	15
12.00	.00			8	14
15.00	.00			8	13
16.00	.00			8	12
18.00	.00			8	11
18.00	.00			8	10
20.00	1.00	.5870	.1087	9	9
22.00	.00			9	8
24.00	1.00	.5136	.1173	10	7
28.00	.00			10	6
28.00	.00			10	5
28.00	.00			10	4
30.00	1.00	.3852	.1418	11	3
30.00	.00			11	2
33.00	.00			11	1
42.00	1.00	.0000	.0000	12	0

Number of Cases: 24 Censored: 12 (50.00%) Events: 12

	Survival Time	Standard Error	95% Confidence Interval	
Mean:	25.88	3.54	(18.95,	32.81)
Median:	30.00	7.03	(16.23,	43.77)

4.9 Método Actuarial para Cálculo de Función de supervivencia

Es frecuente que en los datos para cálculo de la función de sobrevivencia no se disponga de valores individuales, sino que se presenten en estadísticas agrupadas por edad o intervalos de tiempo. Los intervalos usualmente son del mismo tamaño, pero es factible la utilización de intervalos de diferente tamaño.

Se supone que se tienen m intervalos 1,2,…,m y se dispone del número de sujetos expuestos al inicio de cada intervalo n_t, también de los decesos d_t y de los casos censurados Ct. Se adopta el supuesto de que los casos censurados se distribuyen en forma uniforme en el intervalo correspondiente, así el número de individuos que realmente se encuentran en riesgo es:

$$n_t^* = n_t - \frac{C_t}{2}$$

El valor de n^*_t se conoce como la hipótesis o supuesto actuarial, para tener un valor puntual de referencia del intervalo se puede adoptar la marca de clase, esto es el valor medio del intervalo. Al limitarnos al intervalo j-ésimo, la probabilidad de muerte ht se puede estimar mediante:

$$h(t) = \frac{d_t}{n_t^*}$$

La probabilidad de sobrevivir en el intervalo j-ésimo es simplemente la diferencia:

$$P_t = 1 - h(t)$$

La función de sobrevivencia y las otras estadísticas se estiman en forma análoga mediante la aplicación de n*t

$$\hat{f}(t) = \frac{d_t}{n} \qquad \hat{S}(t) = \prod_{i=1}^{t}\left(1 - \frac{d_i}{n_t^*}\right) \qquad H(t) = -Ln(S(t))$$

Los cálculos de la varianza y error estándar de la función de sobrevivencia quedan expresados como sigue:

$$\sum_{j=0}^{t-1} \frac{d_j}{n_j^*\left(n_j^* - d_j\right)} \qquad V(S(t)) = S^2(t)\sum_{i=0}^{t-1} \frac{d_j}{n_j^*(n_j^* - d_j)} \qquad EE(S(t)) = S(t)\sqrt{\sum_{i=0}^{t-1} \frac{d_j}{n_j^*(n_j^* - d_j)}}$$

Los intervalos de 95% de confianza para la S(t) se calculan de la forma acostumbrada con el supuesto de normalidad.

$$Li95\% = S(t) - 1.96 * EE(S(t))$$
$$Ls95\% = S(t) + 1.96 * EE(S(t))$$

Para ilustrar el cálculo mediante el método actuarial se presenta el siguiente ejemplo:

Intervalo de Tiempo Li<x<= Ls	Casos nt	Muertes dt	Censurados Abandono Ct	$n_t^* = n_t - \frac{C_t}{2}$	Función de Riesgo h t	$\hat{f}(t)$	Probabilidad de Sobrevivir Pt
0 a 5	949	731	18	940	0.777660	0.770285	0.222340
5 a 10	200	52	16	192	0.270833	0.054795	0.729167
10 a 15	132	14	75	94.5	0.148148	0.014752	0.851852
15 a 20	43	10	33	26.5	0.377358	0.010537	0.622642

Función de Sobrevivir St	Riesgo Acumulado Ht	$\frac{d_j}{n_j^*\left(n_j^* - d_j\right)}$	$\sum_{j=00}^{t-1} \frac{d_j}{n_j^*\left(n_j^* - d_j\right)}$	V(S(t))	EE(S(t))	Lim Inf 95%	Lim Sup 95%
0.222340	1.503546	0.003721	0.003721	0.000184	0.013563	0.195758	0.248923
0.162123	1.819399	0.001935	0.005655	0.000149	0.012192	0.138227	0.186020
0.138105	1.979741	0.001840	0.007496	0.000143	0.011957	0.114670	0.161540
0.085990	2.453526	0.022870	0.030366	0.000225	0.014984	0.056620	0.115359

4.10 Comparación de dos funciones de supervivencia.

Es frecuente la necesidad de comparar dos funciones de supervivencia originadas por dos grupos de sujetos sometidos a diferentes circunstancias, por ejemplo dos diferentes tipos de tratamientos terapéuticos.

El siguiente ejemplo corresponde a dos grupos de pacientes de cáncer de seno, uno de los cuales ha sido diagnosticado positivo y el otro negativo para un reactivo conocido como HPA. Se considera que la supervivencia está asociada a esa prueba clínica. El grupo de pacientes positivas a la prueba HPA está integrado por 32 pacientes de las cuales 11 son casos censurados.

PERSONA	TIEMPO	CENSURADO	HPA
1	5	0	1
2	8	0	1
3	10	0	1
4	13	0	1
5	18	0	1
6	24	0	1
7	26	0	1
8	26	0	1
9	31	0	1
10	35	0	1
11	40	0	1
12	41	0	1
13	48	0	1
14	50	0	1
15	59	0	1
16	61	0	1

PERSONA	TIEMPO	CENSURADO	HPA
17	68	0	1
18	71	0	1
19	76	1	1
20	105	1	1
21	107	1	1
22	109	1	1
23	113	0	1
24	116	1	1
25	118	0	1
26	143	0	1
27	154	1	1
28	162	1	1
29	188	1	1
30	212	1	1
31	217	1	1
32	225	1	1

El grupo de pacientes negativas está integrado por 13 casos, de los cuales 8 son censurados.

PERSONA	TIEMPO	CENSURADO	HPA
33	23	0	0
34	47	0	0
35	69	0	0
36	70	1	0
37	71	1	0
38	100	1	0
39	101	1	0
40	148	0	0
41	181	0	0
42	198	1	0
43	208	1	0
44	212	1	0
45	224	1	0

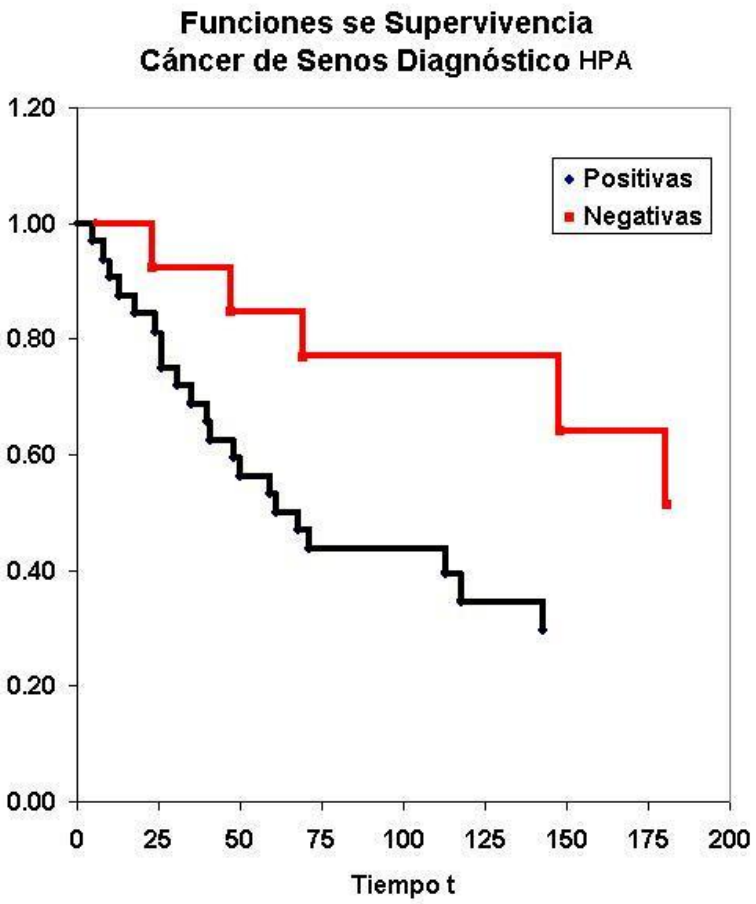
A continuación se calculan las tablas de supervivencia por separado
Tabla de cálculos de función de supervivencia para pacientes positivas de HPA

t	n_t	d_t	C_t	P_t	$\prod_{i=1}^{t} P_i$	$S(t)$	$H(t)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j=0}^{t-1}\frac{d_j}{n_j(n_j-d_j)}$	EE(S(t))
5	32	1	0	**0.96875**	0.968750	**0.96875**	0.0317487	0.00100806	0.00100806	0.0308
8	31	1	0	**0.96774**	0.937500	**0.93750**	0.0645385	0.00107527	0.00208333	0.0428
10	30	1	0	**0.96667**	0.906250	**0.90625**	0.0984401	0.00114943	0.00323276	0.0515
13	29	1	0	**0.96552**	0.875000	**0.87500**	0.1335314	0.00123153	0.00446429	0.0585
18	28	1	0	**0.96429**	0.843750	**0.84375**	0.1698990	0.00132275	0.00578704	0.0642
24	27	1	0	**0.96296**	0.812500	**0.81250**	0.2076394	0.0014245	0.00721154	0.0690
26	26	2	0	**0.92308**	0.750000	**0.75000**	0.2876821	0.00320513	0.01041667	0.0765
31	24	1	0	**0.95833**	0.718750	**0.71875**	0.3302417	0.00181159	0.01222826	0.0795
35	23	1	0	**0.95652**	0.687500	**0.68750**	0.3746934	0.00197628	0.01420455	0.0819
40	22	1	0	**0.95455**	0.656250	**0.65625**	0.4212135	0.0021645	0.01636905	0.0840
41	21	1	0	**0.95238**	0.625000	**0.62500**	0.4700036	0.00238095	0.01875	0.0856
48	20	1	0	**0.95000**	0.593750	**0.59375**	0.5212969	0.00263158	0.02138158	0.0868
50	19	1	0	**0.94737**	0.562500	**0.56250**	0.5753641	0.00292398	0.02430556	0.0877
59	18	1	0	**0.94444**	0.531250	**0.53125**	0.6325226	0.00326797	0.02757353	0.0882
61	17	1	0	**0.94118**	0.500000	**0.50000**	0.6931472	0.00367647	0.03125	0.0884
68	16	1	0	**0.93750**	0.468750	**0.46875**	0.7576857	0.00416667	0.03541667	0.0882
71	15	1	0	**0.93333**	0.437500	**0.43750**	0.8266786	0.0047619	0.04017857	0.0877
76	14	0	1	**1.00000**	0.437500					
105	13	0	1	**1.00000**	0.437500					
107	12	0	1	**1.00000**	0.437500					
109	11	0	1	**1.00000**	0.437500					
113	10	1	0	**0.90000**	0.393750	**0.39375**	0.9320391	0.01111111	0.05128968	0.0892
116	9	0	1	**1.00000**	0.393750					
118	8	1	0	**0.87500**	0.344531	**0.34453**	1.0655705	0.01785714	0.06914683	0.0906
143	7	1	0	**0.85714**	0.295313	**0.29531**	1.2197212	0.02380952	0.09295635	0.0900
154	6	0	1	**1.00000**	0.295313					
162	5	0	1	**1.00000**	0.295313					
188	4	0	1	**1.00000**	0.295313					
212	3	0	1	**1.00000**	0.295313					
217	2	0	1	**1.00000**	0.295313					
225	1	0	1	**1.00000**	0.295313					

Tabla de cálculos de función de supervivencia para pacientes negativas de HPA

t	n_t	d_t	C_t	P_t	$\prod_{i=1}^{t} P_i$	$S(t)$	$H(t)$	$\frac{d_j}{n_j(n_j-d_j)}$	$\sum_{j=0}^{t-1}\frac{d_j}{n_j(n_j-d_j)}$	EE(S(t))
23	13	1	0	**0.92308**	0.923077	**0.92308**	0.0800427	0.0064103	0.0064103	**0.0739**
47	12	1	0	**0.91667**	0.846154	**0.84615**	0.1670541	0.0075758	0.0139860	**0.1001**
69	11	1	0	**0.90909**	0.769231	**0.76923**	0.2623643	0.0090909	0.0230769	**0.1169**
70	10	0	1	**1.00000**	0.769231					
71	9	0	1	**1.00000**	0.769231					
100	8	0	1	**1.00000**	0.769231					
101	7	0	1	**1.00000**	0.769231					
148	6	1	0	**0.83333**	0.641026	**0.64103**	0.4446858	0.0333333	0.0333333	**0.1170**
181	5	1	0	**0.80000**	0.512821	**0.51282**	0.6678294	0.0500000	0.0833333	**0.1480**
198	4	0	1	**1.00000**	0.512821					
208	3	0	1	**1.00000**	0.512821					
212	2	0	1	**1.00000**	0.512821					
224	1	0	1	**1.00000**	0.512821					

La observación de las funciones de supervivencia permiten concluir en una primera revisión que tienen mayor supervivencia las pacientes diagnosticadas negativas para la prueba HPA. La gráfica siguiente permite apreciar en forma más inmediata ese resultado.



4.11 Prueba Log Rank para comparar dos funciones de supervivencia.

Para probar la significancia estadística de la diferencia entre dos funciones, una de las pruebas más usadas es la prueba Log Rank . La prueba parte de la aplicación de tablas de contingencia 2x2 aplicadas a cada tiempo t en la siguiente estructura:

Grupo	Muertes t	Sobreviven t	Expuestos t
Grupo 1	d_{1t}	$n_{1t} - d_{1t}$	n_{1t}
Grupo 2	d_{2t}	$n_{2t} - d_{2t}$	n_{2t}
Total	d_{t}	$n_t - d_t$	n_t

Al suponer igualdad de proporciones, se calculan los valores esperados de muertes para ambos grupos. La distribución de probabilidad asociada a una tabla de contingencia es una hipergeométrica cuyo valor esperado es el siguiente:

$$e_{it} = \frac{n_{it} d_t}{n_t}$$

La varianza en términos de la tabla se expresa de la siguiente forma:

$$V\left(d_{it} - e_{it}\right) = \frac{n_{1t} n_{2t} d_t (n_t - d_t)}{n_t^2 (n_t - 1)}$$

Las hipótesis nula y alternativa que se plantean son las siguientes:

Ho: $S_1(t)=S_2(t)$ vs Ha: $S_1(t) \neq S_2(t)$

La estadística de prueba L se obtiene mediante el conciente de las sumas de cuadrados de diferencias entre valores observados y valores esperados, entre la suma de varianzas asociadas a cada tabla.

$$\text{Log Rank } L_i = \frac{\left[\sum_t d_{it} - \sum_t e_{it}\right]^2}{\sum_t V\left(d_{it} - e_{it}\right)}$$

La estadística L se distribuye aproximadamente como una Ji cuadrada con 1 grado de libertad.

La siguiente tabla muestra los cálculos para obtener la estadística L para dos grupos. En SPSS hay que solicitar la prueba en la opción COMPARE FACTOR.

$$e_{1t} = \frac{n_{1t} d_t}{n_t} \qquad e_{2t} = \frac{n_{2t} d_t}{n_t} \qquad V_{1t} = \frac{n_{1t} n_{2t} d_t (n_t - d_t)}{n_t^2 (n_t - 1)}$$

$$\text{Log Rank } L_i = \frac{\left[\sum_t d_{it} - \sum_t e_{it}\right]^2}{\sum_t V_{it}} \qquad L \rightarrow \chi_1^2$$

CALCULO DE LA ESTADISTICA L PARA LA PRUEBBA LOG RANK DE DIFERENCIA DE FUNCIONES DE SUPERVIVENCIA

Complemente los cálculos y obtenga la estadística de prueba LOG RANK

t	n1t	n2t	nt	d1t	d2t	dt	c1t	c2t	e1t	e2t	Vt
1	21	21	42		2	2			1.000	1.000	0.48780
2	21	19	40		2	2			1.050	0.950	0.48596
3	21	17	38		1	1			0.553	0.447	0.24723
4	21	16	37		2	2			1.135	0.865	0.47723
5	21	14	35		2	2			1.200	0.800	0.46588
6	21	12	33	3		3	1		1.909	1.091	0.65083
7	17	12	29	1		1			0.586	0.414	0.24257
8	16	12	28		4	4			2.286	1.714	0.87075
9	16	8	24			0	1		0.000	0.000	0.00000
10	15	8	23	1		1	1		0.652	0.348	0.22684
11	13	8	21		2	2	1		1.238	0.762	0.44807
12	12	6	18		2	2			1.333	0.667	0.41830
13	12	4	16	1		1			0.750	0.250	0.18750
15	11	4	15		1	1			0.733	0.267	0.19556
16	11	3	14	1		1			0.786	0.214	0.16837
17	10	3	13		1	1	1		0.769	0.231	0.17751
19	9	2	11			0	1		0.000	0.000	0.00000
20	8	2	10			0	1		0.000	0.000	0.00000
22	7	2	9	1	1	2			1.556	0.444	0.30247
23	6	1	7	1	1	2			1.714	0.286	0.20408
25	5	0	5			0	1		0.000	0.000	0.00000
32	4	0	4			0	2		0.000	0.000	0.00000
34	2	0	2			0	1		0.000	0.000	0.00000
35	1	0	1			0	1		0.000	0.000	
				9	21	30	12		19.251	10.749	6.257

El cálculo de la Estadística de Prueba de Log Rang arroja el valor 16.792941 cuya probabilidad asociada es 0.000042 y por tanto entre ambos grupos hay diferencia estadísticamente significativa.

Solución	
Grupo 1	
L =	16.792941
P value	0.000042

5. METODOS PARAMETRICOS DE ESTIMACION.
5.1 La distribución Exponencial Negativa

Una de las funciones más frecuentes para modelar la supervivencia es la distribución exponencial negativa.

La función exponencial negativa se caracteriza de la forma siguiente:

Función de densidad:

$$f_T(t) = \lambda e^{-\lambda t} \quad \text{para } t>0$$

Cuya media y varianza son: $E(T) = \frac{1}{\lambda}$ $V(T) = \frac{1}{\lambda^2}$

El estimador máximo verosímil de **λ** es el recíproco de la media aritmética

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

Su función de distribución acumulativa es:

$$F_T(t) = 1 - e^{-\lambda t}$$

Por tanto su función de supervivencia es la diferencia:

$$S(t) = 1 - F(t) = e^{-\lambda t}$$

De donde se obtiene fácilmente la función de riesgo h(t) como una constante igual al parámetro **λ**

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

La función acumulativa de riesgo H(t) es igual al negativo del logaritmo natural de la función de supervivencia que da lugar a una recta con pendiente que pasa por el origen:

$$H(t) = -Ln(S(t)) = -Ln(e^{-\lambda t}) = \lambda t$$

Ejemplo. Considere a 12 sujetos cuyos tiempos de supervivencia se presentan a continuación:

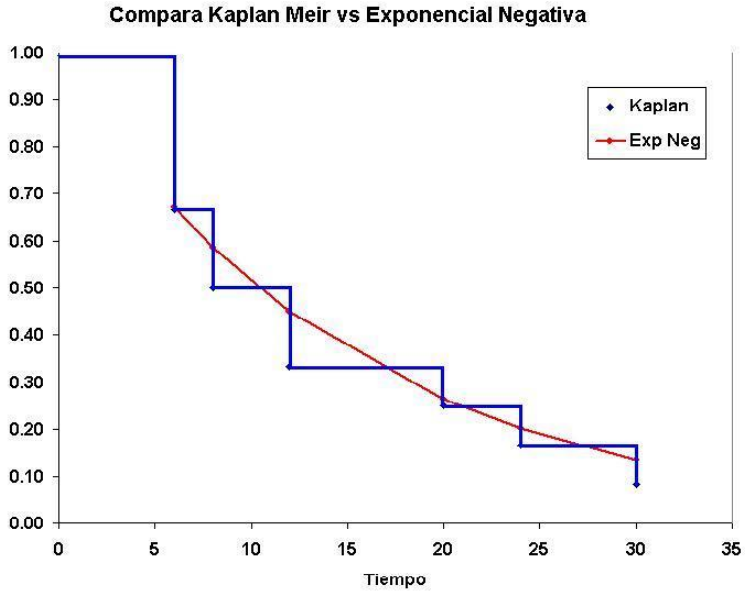
Sujeto	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Tiempo	**6**	**6**	**6**	**6**	**8**	**8**	**12**	**12**	**20**	**24**	**30**	**42**

A continuación se efectúan los cálculos de la función de supervivencia mediante el método Kaplan Meier

Tiempo de Supervivencia t	**Sujetos vivos al inicio de t** n_t	**Sujetos muertos durante t** d_t	$1-\frac{d_t}{n_t}$	$\hat{S}(t)$	$\hat{F}(t)$	$\hat{f}(t)$	$\hat{h}(t)$	$\hat{H}(t)$
6	**12**	**4**	**0.66667**	**0.66667**	**0.33333**	**0.33333**	**0.33333**	**0.40547**
8	**8**	**2**	**0.75000**	**0.50000**	**0.50000**	**0.16667**	**0.25000**	**0.69315**
12	**6**	**2**	**0.66667**	**0.33333**	**0.66667**	**0.16667**	**0.33333**	**1.09861**
20	**4**	**1**	**0.75000**	**0.25000**	**0.75000**	**0.08333**	**0.25000**	**1.38629**
24	**3**	**1**	**0.66667**	**0.16667**	**0.83333**	**0.08333**	**0.33333**	**1.79176**
30	**2**	**1**	**0.50000**	**0.08333**	**0.91667**	**0.08333**	**0.50000**	**2.48491**
42	**1**	**1**	**0.00000**	**0.00000**	**1.00000**	**0.08333**	**1.00000**	
		12				**1.00000**		

Ahora se procede a calcular el promedio de los tiempos de supervivencia de los 12 sujetos, lo cual da el valor 15. El recíproco se toma como estimador de y es entonces 0.06666667. En la siguiente tabla y gráfica se presentan los valores estimados mediante la función de supervivencia por Kaplan Meier y mediante ajuste a la exponencial negativa.

Tiempo de Supervivencia t	**Kaplan Meier** $\hat{S}(t)$	**Exp Neg.** $\hat{S}(t)$
6	**0.66667**	**0.67032**
8	**0.50000**	**0.58665**
12	**0.33333**	**0.44933**
20	**0.25000**	**0.26360**
24	**0.16667**	**0.20190**
30	**0.08333**	**0.13534**



$$S(t) = 1 - F(t)$$

$H(t)$

5.2 LA FUNCIÓN DE WEIBULL

Su función de densidad se caracteriza por 3 parámetros: α parámetro de forma, β parámetro de escala y γ parámetro de posición. El parámetro de posición marca el punto de arranque de la distribución.

$$f(t,\alpha,\beta,\gamma)=\begin{cases}\frac{\alpha}{\beta}\left(\frac{t-\gamma}{\beta}\right)^{\alpha-1} e^{-\left(\frac{t-\gamma}{\beta}\right)^{\alpha}} & t\geq\gamma\\ 0 & t<\gamma\end{cases}$$

Es ev**id**ente que si α= 1 y γ= 0 la función se reduce a la conocida exponencial negativa con un parámetro β. La incorporación de dos parámetros adicionales a la exponencial negativa, lo cual le proporciona a la distribución de Weibull mayor flexibilidad y adaptabilidad en algunas aplicaciones muy particulares.

Si se considera nulo el parámetro de posición γ= 0, se tiene la distribución de Weibull estándar, sobre la cual enfocaremos nuestra atención. Su caracterización reduce entonces a dos parámetros: α, parámetro de forma y β parámetro de escala. Las fórmulas de las funciones de densidad y de distribución acumulativa se presentan a continuación:

$$f(t,\alpha,\beta)=\begin{cases}\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} e^{-(t/\beta)^{\alpha}} & t\geq 0\\ 0 & t<0\end{cases} \qquad F(t,\alpha,\beta)=\begin{cases}1-e^{-(t/\beta)^{\alpha}} & t\geq 0\\ 0 & t<0\end{cases}$$

La función de supervivencia asociada a la distribución de Weibulla, usual en el ámbito actuarial S(t). También conocida como función de confiabilidad en ingeniería R(t), es el complemento de la función de distribución acumulativa.

$$S(t)=1-F(t,\alpha,\beta) \qquad S(t)=e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$

La tasa de riesgo instantánea h(t) y la función de riesgo acumulado H(t)

$$h(t)=\frac{\alpha t^{\alpha-1}}{\beta^{\alpha}} \qquad H(t)=\left(\frac{t}{\beta}\right)^{\alpha}$$

La estimación de sus parámetros se puede resolver por varios métodos. El de linealización de la función de distribución acumulativa empírica **(*)**

$$\mathrm{Ln}\left[\mathrm{Ln}\left(\frac{1}{1-\hat{F}(t)}\right)\right]=\alpha\mathrm{Ln}(t)-\alpha\mathrm{Ln}(\beta)$$

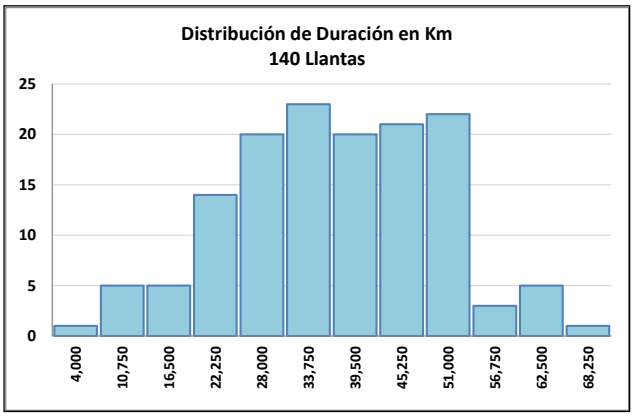
EJEMPLO DE AJUSTE DE WEIBULL

Se han recolectado datos de la duración en kilómetros de 140 llantas sometidas a diferentes superficies de rodamiento y hasta su primer reventón. Ajuste la función de supervivencia mediante una distribución de Weibull.

14,448.4	26,581.4	39,337.6	44,455.9	21,713.2	36,349.2	47,724.5	22,330.2	51,752.6	13,245.7
28,416.8	51,009.2	38,772.9	32,878.7	36,069.5	36,806.5	45,290.4	51,503.6	69,657.6	26,775.2
33,410.7	34,096.5	45,730.2	46,667.8	48,014.1	33,427.1	33,487.9	36,533.5	47,262.0	62,552.4
55,832.0	48,774.3	45,087.4	34,836.9	51,817.9	52,862.5	40,554.6	50,619.5	35,059.2	30,537.4
20,174.8	46,613.7	13,268.9	53,842.9	35,872.6	28,111.8	65,915.8	34,775.0	42,611.8	49,961.0
24,380.9	38,348.9	9,892.1	30,540.9	41,720.8	30,194.3	29,572.4	12,956.6	62,331.8	67,579.3
64,872.5	29,480.0	51,488.0	46,720.4	56,185.7	22,916.2	31,445.8	48,482.8	58,848.8	20,933.0
33,809.1	26,266.8	9,851.0	49,194.8	53,920.5	38,028.6	17,919.8	54,693.5	60,372.2	30,627.8
32,419.3	50,889.3	34,063.0	52,490.4	45,555.4	45,080.8	22,677.9	49,017.5	22,314.6	43,056.9
24,023.4	29,422.1	43,529.0	51,051.3	53,781.3	29,843.6	47,282.6	53,227.2	22,705.3	41,838.2
38,537.1	20,757.9	37,595.3	40,404.4	50,292.8	39,218.3	47,771.9	26,635.2	43,611.6	52,456.6
29,927.9	25,973.3	39,078.1	29,903.3	40,739.4	41,314.5	33,050.0	25,714.9	32,440.5	34,199.6
27,512.4	48,180.1	31,142.2	52,743.3	42,005.9	46,533.0	36,910.0	7,314.5	15,327.5	53,941.9
30,009.6	36,637.3	43,851.5	52,425.5	17,213.3	18,928.0	20,074.7	36,765.1	53,256.4	41,258.4

La tabla de frecuencias y el histograma correspondiente se presentan a continuación.

Límite		Frecuencia			
Inferior	Superior	Absoluta	Acumulada	Relativa	Rel. Acum
0	8,000	1	1	0.0071429	0.0071429
7,500	14,000	5	6	0.0357143	0.0428571
13,000	20,000	5	11	0.0357143	0.0785714
18,500	26,000	14	25	0.1000000	0.1785714
24,000	32,000	20	45	0.1428571	0.3214286
29,500	38,000	23	68	0.1642857	0.4857143
35,000	44,000	20	88	0.1428571	0.6285714
40,500	50,000	21	109	0.1500000	0.7785714
46,000	56,000	22	131	0.1571429	0.9357143
51,500	62,000	3	134	0.0214286	0.9571429
57,000	68,000	5	139	0.0357143	0.9928571
62,500	74,000	1	140	0.0071429	1.0000000
		140		1.0000000	



Estadísticas básicas

Media	38,449.5
Mediana	38,443.0
Desv, Est	13,259.05
Mínimo	7,314.5
Máximo	69,657.6

Para la estimación de los parámetros mediante la función linealizada se construye una tabla con los 140 datos, de la cual se presentan los primeros 10 renglones.

			X	Y
No	t	S(t)	Ln(t)	Ln(Ln(1/(1-S(t))
1	7,314.5	0.0071429	8.897614	-4.93806
2	9,851.0	0.0142857	9.195328	-4.24131
3	9,892.1	0.0214286	9.199492	-3.83222
4	12,956.6	0.0285714	9.469361	-3.54089
5	13,245.7	0.0357143	9.491428	-3.31408
6	13,268.9	0.0428571	9.493178	-3.12806
7	14,448.4	0.0500000	9.578339	-2.97020
8	15,327.5	0.0571429	9.637404	-2.83292
9	17,213.3	0.0642857	9.753438	-2.71138
10	17,919.8	0.0714286	9.793662	-2.60223

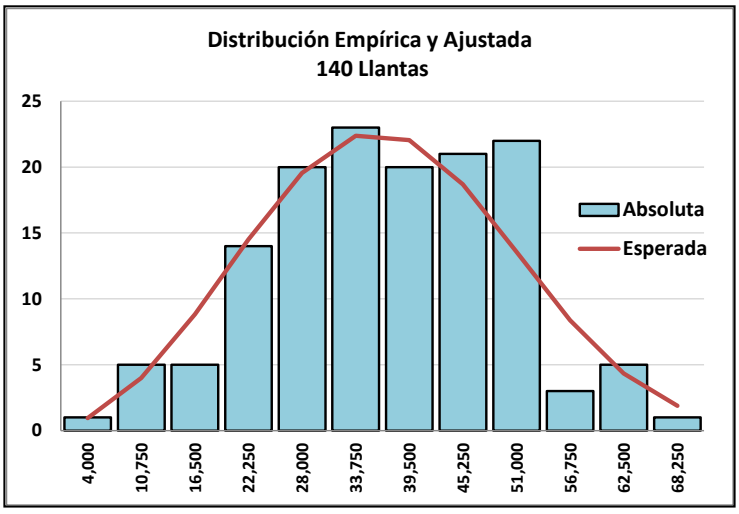
Los parámetros estimados mediante este método se obtienen los siguientes valores:

Beta	42,929.4392
Alfa	2.9685

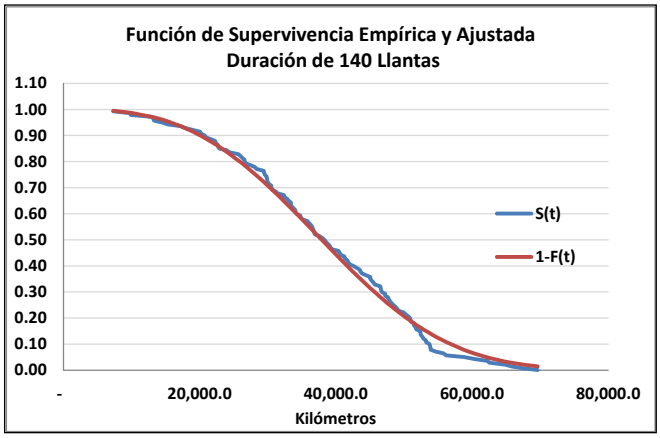
La función de distribución de Weibull ajustada permite el cálculo de frecuencias esperadas que se pueden comparar con las observadas:

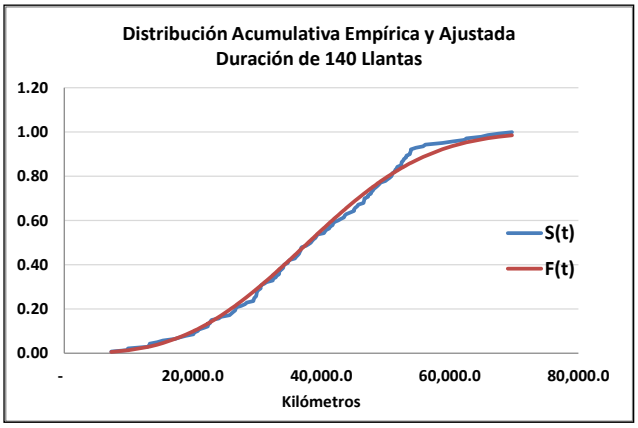
Límite		Frecuencia				Probabilidad		Frecuencia
Inferior	Superior	Absoluta	Acumulada	Relativa	Rel. Acum	Intervalo	Acumulada	Esperada
0	8,000	1	1	0.0071429	0.0071429	0.0067995	0.0067995	1.0
7,500	14,000	5	6	0.0357143	0.0428571	0.0284902	0.0352897	4.0
13,000	20,000	5	11	0.0357143	0.0785714	0.0631028	0.0983925	8.8
18,500	26,000	14	25	0.1000000	0.1785714	0.1036390	0.2020316	14.5
24,000	32,000	20	45	0.1428571	0.3214286	0.1396202	0.3416517	19.5
29,500	38,000	23	68	0.1642857	0.4857143	0.1598863	0.5015380	22.4
35,000	44,000	20	88	0.1428571	0.6285714	0.1574577	0.6589957	22.0
40,500	50,000	21	109	0.1500000	0.7785714	0.1334566	0.7924523	18.7
46,000	56,000	22	131	0.1571429	0.9357143	0.0968817	0.8893341	13.6
51,500	62,000	3	134	0.0214286	0.9571429	0.0597585	0.9490926	8.4
57,000	68,000	5	139	0.0357143	0.9928571	0.0310110	0.9801036	4.3
62,500	74,000	1	140	0.0071429	1.0000000	0.0133896	0.9934931	1.9
		140		1.0000000				

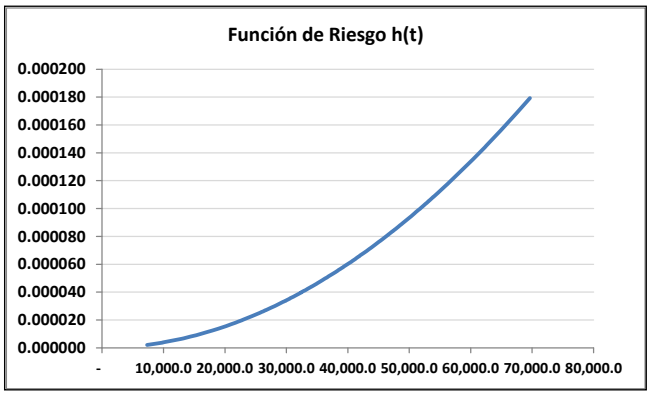
Las siguientes gráficas muestran las funciones de distribución acumulativa, función de supervivencia, de riesgo y de riesgo acumulado, empíricas y ajustadas. Mediante la función de supervivencia se puede afirmar

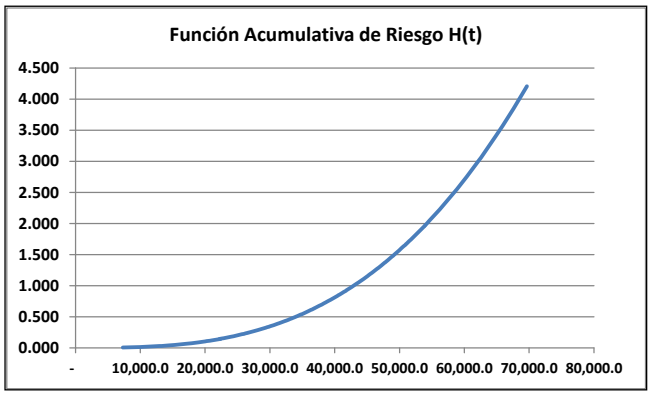


que la probabilidad de que una llanta dure 38,384 Km. es de 0.50. Si se emite una garantía de reemplazo para una duración de 30,000 Km., se puede calcular la probabilidad de que una llanta al azar dure 30,000 km. o más y es 0.72, esto es se esperaría un 28% de reclamaciones y en consecuencia la garantía debe reducirse, por ejemplo a 20,000 Km. con lo cual solamente 10% de las llantas no cumplirían con la garantía.









6. MODELO DE REGRESION DE COX DE RIESGOS PROPORCIONALES.

El modelo de Kaplan Meier para calcular funciones de supervivencia y la prueba Log Rank que permite la comparación de dos o más funciones de distribución se complementa con la incorporación de la influencia en el incremento o decremento del riesgo que ejercen una serie de variables explicativas con la misma intencionalidad. El modelo de Cox para la función de riesgo adopta la siguiente forma:

$$h(t) = h_o(t)e^{\beta_o + \beta_1 x_1 + \ldots + \beta_k x_k}$$

El modelo tiene dos componentes ho(t) que depende del tiempo y la parte exponencial que depende de las variables explicativas.

Si se toma logaritmo de ambos miembros de la expresión anterior se tiene:

$$\frac{h(t)}{h_o(t)} = e^{\beta_o + \beta_1 x_1 + \ldots + \beta_k x_k}$$

Ahora la combinación lineal se despeja en el miembro derecho y del lado izquierdo queda el logaritmo de la razón o **proporción de riesgo.**

$$Ln\left(\frac{h(t)}{h_o(t)}\right) = \beta_o + \beta_1 x_1 + \ldots + \beta_k x_k$$

La estimación de los coeficientes se apoya en un algoritmo denominado de verosimilitud parcial planteado por el propio Cox en 1972 en un artículo publicado en JRSS "Regression Models and Life Tables. La aproximación a los valores se realiza mediante el método Newton Raspón.

Los coeficientes se deben interpretar en el sentido de que si su signo es positivo, se incrementa el riesgo y si es negativo el riesgo disminuye.

Ejemplo.

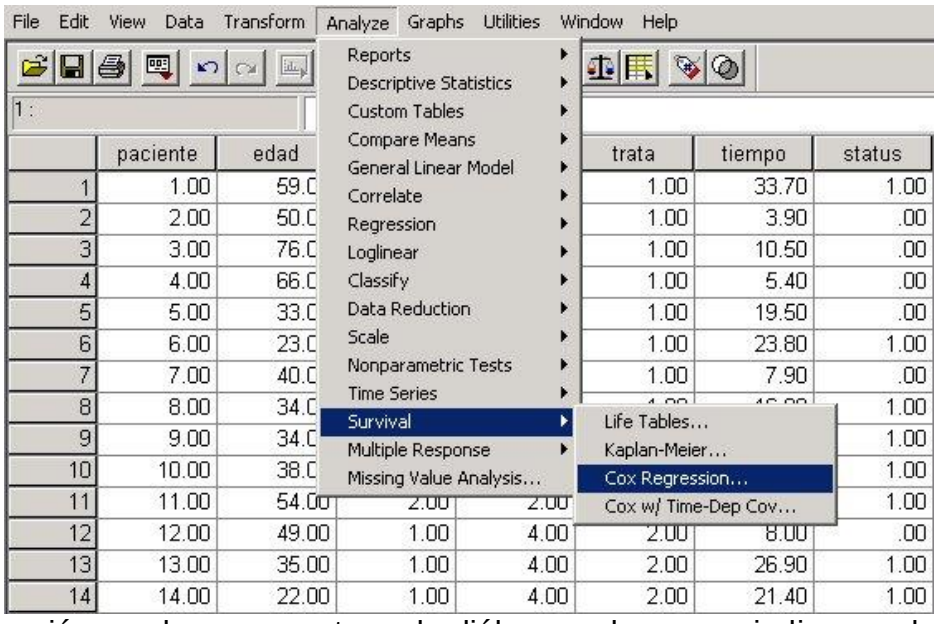
Se tienen datos de 30 pacientes de tumores cancerosos de Melanoma. Los datos disponibles son la edad de los pacientes, el sexo (1 Hombre 2 Mujer) el grado inicial del tumor, el tipo de tratamiento (1 BCG 2 Parvum) y el estado censurado (1 censurado 0 no censurado). Se pretende ajustar un modelo de regresión de Cox y

en particular identificar el efecto del tratamiento en presencia de las otras variables.

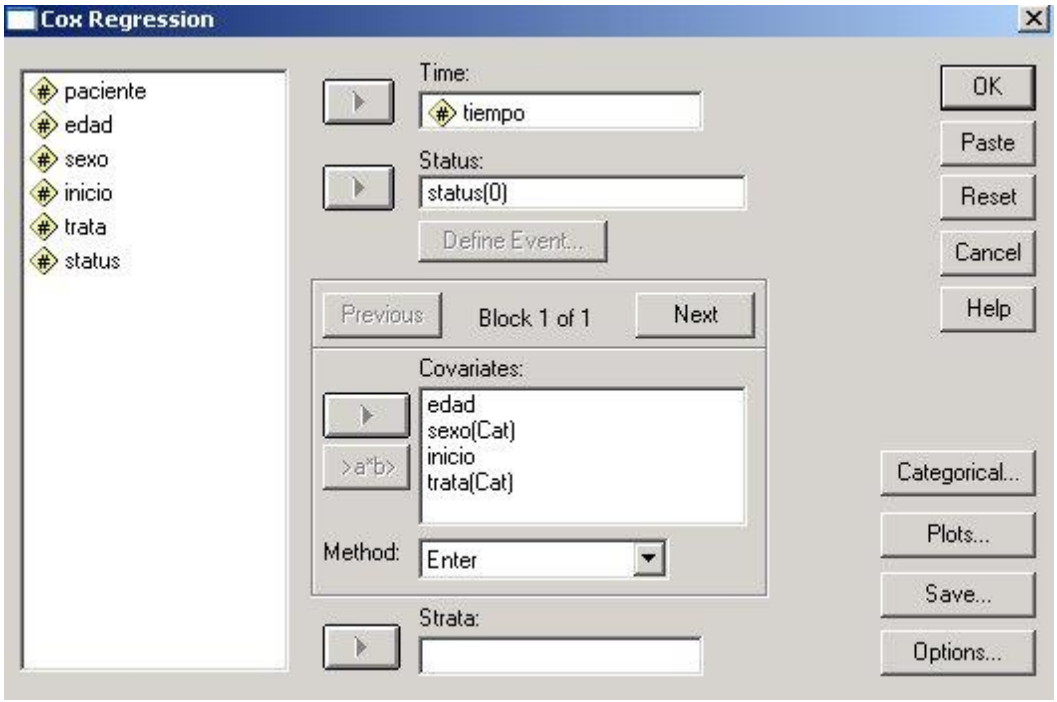
En la tabla siguiente se presentan los datos de los 30 pacientes.

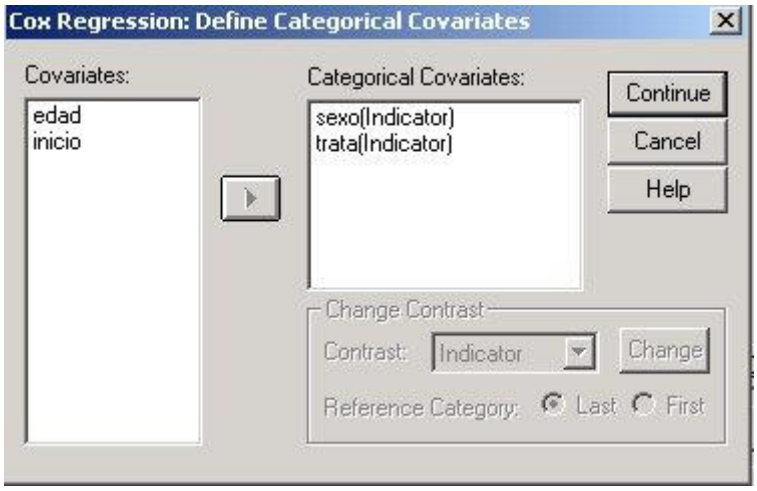
PACIENTE	EDAD	SEXO	INICIAL	TRATA	TIEMPO	STATUS
1	59	2	4	1	33.7	1
2	50	2	4	1	3.9	0
3	76	1	4	1	10.5	0
4	66	2	4	1	5.4	0
5	33	1	4	1	19.5	0
6	23	2	4	1	23.8	1
7	40	2	4	1	7.9	0
8	34	1	4	1	16.9	1
9	34	1	4	1	16.6	1
10	38	2	2	1	33.7	1
11	54	2	2	1	17.1	1
12	49	1	4	2	8.0	0
13	35	1	4	2	26.9	1
14	22	1	4	2	21.4	1
15	30	1	4	2	18.1	1
16	26	2	4	2	16.0	1
17	27	1	4	2	6.9	0
18	45	2	4	2	11.0	0
19	76	2	3	2	24.8	1
20	48	1	3	2	23.0	1
21	91	1	5	2	8.3	0
22	82	2	5	2	10.8	1
23	50	2	5	2	12.2	1
24	40	1	5	2	12.5	1
25	34	1	3	2	24.4	0
26	38	1	5	2	7.7	0
27	50	1	2	2	14.8	1
28	53	2	2	2	8.2	1
29	48	2	2	2	8.2	1
30	40	2	2	2	7.8	1

Se procede a seleccionar la opción de Regresión de Cox.

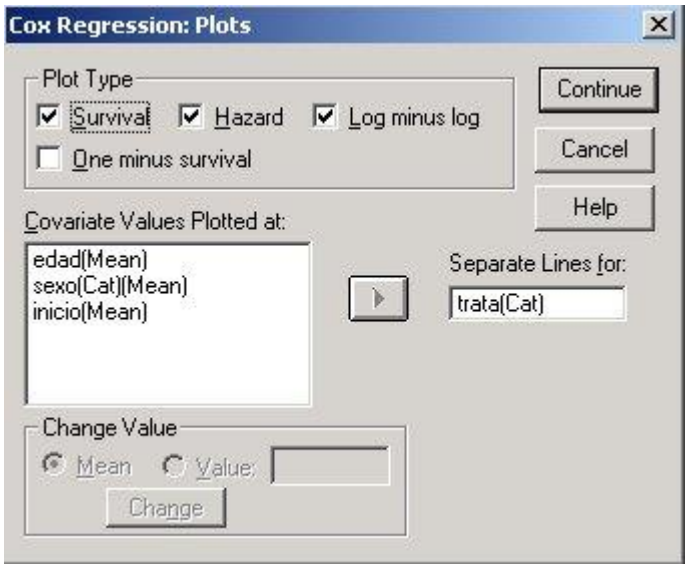


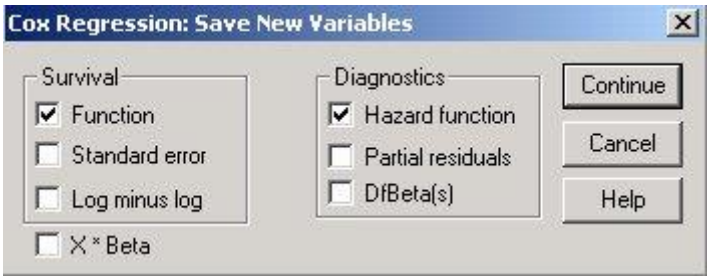
A continuación se abre una ventana de diálogo en la que se indica que la variable **tiempo** como dependiente. La variable **status** se codifica con "0", como valor válido para la realización del evento. Las variables **edad, sexo, inicio y trata** se trasladan como covariables. A continuación se le indica al SPSS que las variables **sexo** y **trata** son categóricas para que el paquete tome sus valores como etiquetas.





Se procede a seleccionar los diferentes gráficos de la función de supervivencia, acumulativa de riesgo y de distribución acumulativa. Se piden gráficas diferenciadas en función del tratamiento (trata). Finalmente se solicita que al archivo de datos se le graben las funciones de supervivencia y de riesgo acumulado.



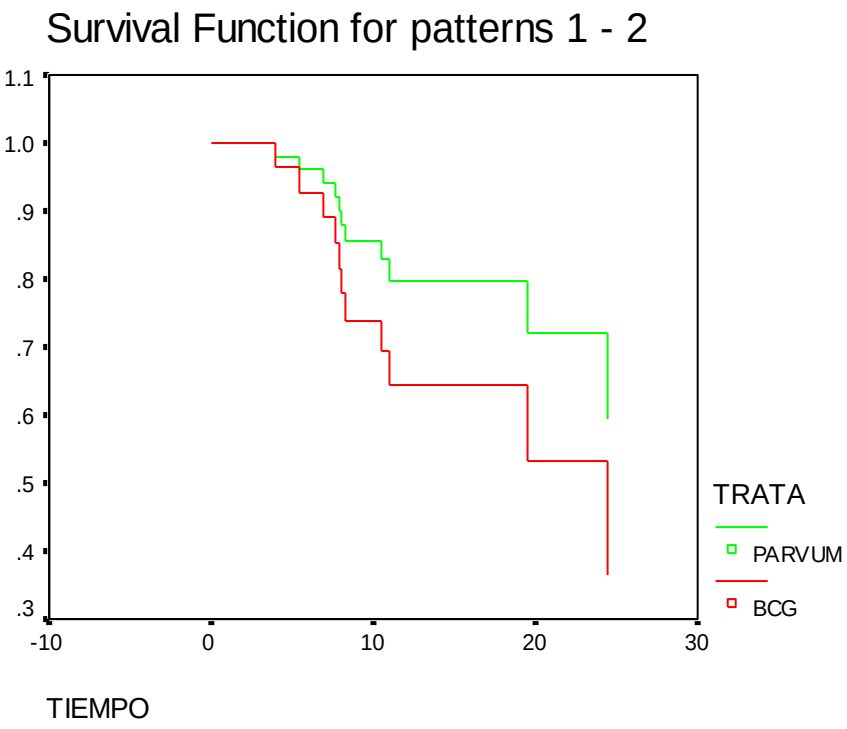


Se ejecutan las opciones y el paquete reporta los valores de los coeficientes de la combinación lineal del modelo (betas) con sus estimaciones puntuales, los errores estándares, la estadística de Wald, que cumple una función análoga a la prueba de t para verificar la significancia del coeficiente. Los grados de libertad y la significancia de los coeficientes de presentan en las siguientes columnas. Observe que ninguno de los coeficientes presenta probabilidad menor a 0.05 y por tanto no son significativamente diferentes de cero, sin embargo se los considera útiles.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
EDAD	.012	.016	.624	1	.430	1.013
SEXO	.507	.658	.593	1	.441	1.660
INICIO	.723	.474	2.326	1	.127	2.061
TRATA	.659	.636	1.071	1	.301	1.932

A medida que el valor del coeficiente aumenta en forma positiva, aumenta el riesgo, pero una mejor forma de apreciar este incremento se tiene al aplicar la exponencial del valor del coeficiente, pues da un valor del incremento de los momios de riesgo. Estos cálculos se tienen en la última columna. Así, el coeficiente para la edad es 0.012, si se calcula EXP(0.012)=1.013 ello indica que por cada año el momio del riesgo aumenta en un factor de 1,013. El sexo parece ser más importante. El paquete registró 1 como hombre y 2 como mujer, pero recodifica 1 a hombre y 0 a mujer. Entonces que el paciente sea hombre da lugar a un incremento en el momio del riesgo en un facto de 1.66 o en otros términos 66% mayor. El grado inicial del tumor también es muy importante. A medida que se aumenta un grado el riesgo se incrementa en un momio de 2.061 y finalmente el paquete interpreta al tratamiento BCG como 1 y 0 al Parvum. Entonces el BCG tiene un incremento del momio de riesgo de 1.932. Entonces resulta mejor el parvum. Esto se puede verificar al observar la gráfica de supervivencia de ambos tratamientos.


Survival Function for patterns 1 - 2
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
-10
0
10
20
30
TIEMPO
TRATA
PARVUM
BCG