UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS Y ESTIMACIÓN DE MODELOS DE SUPERVIVENCIA PARAMÉTRICOS

REPORTE DE SEMINARIO DE TITULACIÓN

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A:

JOSEL SÁNCHEZ CERVANTES

FACULTAD DE CIENCIAS
UNAM

TUTOR

ACT. JAIME VÁZQUEZ ALAMILLA

2007

Análisis y Estimación de Modelos de Supervivencia Paramétricos

Sánchez Cervantes Josel

2007

Hoja de Datos del Jurado

1.- Datos del alumno
Sánchez
Cervantes
Josel
50 44 21 65
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
300328600

2.- Datos del tutor
Act
Jaime
Vázquez
Alamilla

3.- Datos del sinodal 1
M en C
Salvador
Zamora
Muñoz

4.- Datos del sinodal 2
Mat
Margarita
Chávez
Cano

5.- Datos del sinodal 3
Dra
Guillermina
Eslava
Gómez

6.- Datos del sinodal 4
Act
Francisco
Sánchez
Villarreal

7.- Datos del trabajo escrito
Análisis y Estimación de Modelos de Supervivencia Paramétricos
63 p
2007

Índice general

1. Introducción **5**

1.1. Definición y formas de los modelos de supervivencia 6
1.2. Modelos de supervivencia actuariales 7
1.2.1. El Modelo Selecto 8
1.2.2. El Modelo Agregado 8
1.3. Estimación 8
1.4. Ejemplos 9

2. Las matemáticas de los modelos de supervivencia **12**

2.1. La distribución de T 12
2.1.1. Relación de la función de supervivencia con otras funciones 12
2.1.2. La función de tasa de riesgo 13
2.1.3. Modelos de supervivencia actuariales 15
2.2. Modelos de supervivencia paramétricos 16
2.2.1. Distribución uniforme 16
2.2.2. Distribución exponencial 17
2.2.3. Distribución Weibull 18
2.2.4. Distribución Gompertz 19
2.2.5. Distribución Makeham 20
2.3. Formas de medidas condicionales y distribuciones truncadas . 20
2.3.1. Probabilidades condicionales 21
2.3.2. Distribución de X truncada inferiormente 21
2.3.3. Distribución de X truncada superior e inferiormente . . 23
2.3.4. La tasa central 25

3. Estimación de modelos de supervivencia paramétricos **26**
3.1. Modelos univariados con datos no censurados 26
3.1.1. Método de momentos 27
3.1.2. Método de máxima verosimilitud 28
3.1.3. Método de mínimos cuadrados 29
3.1.4. Método de medianas 33
3.1.5. Método de percentiles 34
3.1.6. Tiempos de muerte agrupados 34
3.2. Modelos univariados con datos censurados 36
3.2.1. Método de máxima verosimilitud 36
3.3. Pruebas de hipótesis de modelos paramétricos 39
3.3.1. Prueba χ^2 (Ji-cuadrada) 39
3.3.2. Prueba *Kolmogorov-Smirnov* 40
3.4. Modelos multivariados 42
3.4.1. Modelo multiplicativo 44
3.5. Aplicaciones 46
3.5.1. Uso de variables concomitantes 48
3.5.2. Alternativa paramétrica de modelos exponenciales . . . 49

4. Modelos multiestados **52**
4.1. Modelo de progresión de una enfermedad 53
4.1.1. Propiedades del modelo 54
4.1.2. Estimación de los parámetros del modelo 57
4.2. Modelo de cuidado continuo de retiro en comunidades 59
4.2.1. Probabilidades de transición 60

Prefacio

El análisis de supervivencia se presenta en diversas áreas, incluyendo la medicina, biología, ingeniería y economía. Su importancia radica en el interés de poder medir probabilidades de supervivencia y predecir los tiempos de vida de diversos eventos en cada una de las áreas de investigación.

Un modelo de supervivencia es una fórmula matemática que sirve para representar el comportamiento del tiempo de falla de los individuos bajo estudio. Para esto es importante la obtención de datos que representen el fenómeno a estudiar, estos datos pueden presentarse principalmente en dos formas que son: completos y censurados, completos cuando se conocen todos los tiempos de falla de los individuos bajo estudio y censurados cuando por alguna causa no se conoce el tiempo de falla del individuo bajo estudio. Alrededor del análisis de supervivencia se presentan varias formas de distribuciones de probabilidad, las cuales son muy adecuadas para su aplicación, ya que representan de buena forma el tiempo de falla para los modelos en los cuales son usadas.

En este trabajo se pretende realizar un primer acercamiento hacia lo que son los modelos de supervivencia, desde por qué surgen y para qué sirven, hasta la obtención de los modelos y sus aplicaciones. Para ello se hace uso de herramientas básicas de estadística y probabilidad para un buen desarrollo y mejor entendimiento del análisis requerido para este tema. Asimismo, se trata de presentar un balance entre la teoría y la aplicación presentando algunos ejemplos para cada uno de los modelos que se mencionen.

En el primer capítulo se hace una introducción de lo que es un modelo de supervivencia. Se revisará brevemente en qué consiste un modelo de supervivencia, qué interpretaciones permite a partir de él, cómo se puede utilizar, cuales son las formas de presentarse y desarrollarse así como sus alcances y limitaciones. También se comentan las bases del análisis de supervivencia, sus fundamentos y cuales son sus aplicaciones prácticas mediante la descripción de algunos ejemplos.

En el segundo capítulo se explican los fundamentos matemáticos sobre los cuales se basan los modelos de supervivencia. Se explica cuál es la importancia de la función de tasa de riesgo y su utilización para el desarrollo de un

modelo. También se explican cuales son las formas de distribuciones probabilísticas más adoptadas para emplearse en un modelo (Weibull, Gompertz, Makeham y exponencial), sus distintas propiedades y su funcionalidad dentro del análisis de supervivencia. Se describen las formas de medidas condicionales y distribuciones truncadas y también se hace una breve mención acerca de la forma actuarial de representar dichas medidas.

En el tercer capítulo se explican los distintos métodos paramétricos de obtener un modelo de supervivencia a partir de una muestra (por ejemplo: máxima verosimilitud, mínimos cuadrados) y algunas propiedades de dichos métodos. También se explica lo que son las muestras censuradas y las adecuaciones que se realizan en los métodos de estimación para poder generar un modelo a partir de éstas. Una vez obtenido el modelo, se exponen las pruebas estadísticas formales (Kolmogorov-Smirnov y χ^2 (ji-cuadrada)) donde lo que se desea es verificar si se cumplen las hipótesis en las que se basa dicho modelo, y si verdaderamente éste se ajusta bien a los datos observados. A lo largo de todo este capítulo se desarrollan varios ejemplos prácticos para un mejor entendimiento del tema. Para finalizar, se hace una breve introducción hacia lo que son los modelos multivariados, explicando algunas de sus presentaciones y propiedades.

En el capitulo final se muestran algunas formas de los modelos multiestados, los cuales son una extensión más compleja de los modelos de supervivencia. Se describen las formas de los modelos multiestados presentados, así como sus bases y propiedades junto con algunas formas de estimar sus parámetros.

Capítulo 1

Introducción

El análisis estadístico de supervivencia el cual es referido como tiempo de vida, tiempo de supervivencia o tiempo de falla, es un importante tema para investigadores de diversas áreas como ingeniería y ciencias biomédicas.

Por conveniencia, cuando se habla de modelos de supervivencia se hace referencia a datos de tiempo de falla. En ocasiones los eventos de interés son muertes de individuos en un sentido real y el tiempo de vida es la edad actual de un individuo o quizás el tiempo de supervivencia medido desde algún punto inicial particular.

Los siguientes ejemplos ilustran algunos casos típicos en los cuales surge el planteamiento y uso de los modelos de supervivencia:

(a) Algunos tipos de artículos fabricados pueden ser reparados ante una falla. En este caso uno podría estar interesado en el periodo de tiempo entre fallas sucesivas de un artículo o pieza y referirse a éste como el tiempo de vida.

(b) En estudios médicos de enfermedades fatales el interés radica en el tiempo de supervivencia de individuos con esta enfermedad, medido desde la fecha del diagnóstico o desde algún otro punto inicial.

1.1. Definición y formas de los modelos de supervivencia

Un *Modelo de Supervivencia* es la función de distribución de probabilidad de un tipo especial de variable aleatoria, la cual se describe a continuación.

Sea T una variable aleatoria no negativa que representa el tiempo de vida (o de falla) de individuos en una población. Suponiendo que un individuo de la población se mantiene con vida en el tiempo $t = 0$, se está interesado en la probabilidad de que dicho individuo continúe con vida en el futuro (para cualquier t).

Así la variable aleatoria considerada, definida como el tiempo de falla de un individuo sabiendo que existe al tiempo $t = 0$, es frecuentemente llamada *variable aleatoria del tiempo de falla*. Si T denota el tiempo de falla, entonces la probabilidad de que el individuo continúe funcionando al tiempo t es la misma que la probabilidad de que el tiempo de falla exceda el valor de t. Simbólicamente esta probabilidad está denotada por $S(t)$. Formalmente:

$$S(t) = P(T \geq t). \tag{1.1}$$

(ver [7] pág. 8)

Por la definición de T se tienen las propiedades:

$$S(0) = 1 \tag{1.2}$$

$$\lim_{t \to \infty} S(t) = 0. \tag{1.3}$$

(ver [7] pág. 9)
y $S(t)$ es una función no creciente.

Ahora, si T es el tiempo de falla de un individuo que existe en $t = 0$, entonces T es también el *tiempo futuro de vida* del individuo medido desde $t = 0$.

Las formas en que se pueden presentar los modelos de supervivencia son: paramétrica y no paramétrica. Cuando las probabilidades de supervivencia son dadas mediante una fórmula matemática se dice que el modelo $S(t)$ está

en forma paramétrica, esto es porque los valores de $S(t)$ dependen de uno o más parámetros. Por ejemplo, si se supone un modelo de supervivencia $S(t) = e^{-\theta t}$, que corresponde a la función de supervivencia cuando T tiene una distribución de probabilidad exponencial, $S(t)$ depende del parámetro θ, del cual posteriormente se estima su valor. Por su parte, en los modelos de supervivencia no paramétricos los valores numéricos de $S(t)$ son presentados para ciertos valores seleccionados de t, más comúnmente valores enteros. El modelo de supervivencia no paramétrico más común es el conocido como *tabla de vida* o *tabla de mortalidad*.

Dado que el modelo de supervivencia no paramétrico $S(x)$ presenta valores sólo para números enteros $x = 0, 1, ...$ es claro que esta forma de modelo presenta una carencia de respuesta para valores de x fraccionales, por lo tanto es necesario hacer una suposición acerca de la forma de $S(x)$ entre enteros subsecuentes. Esta suposición es llamada distribución supuesta de mortalidad, la cual definirá los valores de $S(x)$ para toda $x \geq 0$. Para esta nueva distribución entre enteros subsecuentes las tres suposiciones más comunes son: lineal, exponencial e hiperbólica. Los modelos de supervivencia actuariales usualmente están dados en forma no paramétrica. En la siguiente sección se describen esos modelos.

1.2. Modelos de supervivencia actuariales

En el ámbito actuarial, los modelos estudiados se dividen principalmente en dos grandes ramas: vida y no vida. Como su nombre lo menciona, en el área de vida la variable aleatoria T representa el tiempo de vida de una persona. En los modelos de supervivencia actuariales de vida, se toma en cuenta la edad cronológica de los individuos en estudio, reconociendo que la supervivencia está en función de la edad.

Por su parte, en los modelos actuariales de no vida, la variable aleatoria T para este caso representa el tiempo de ocurrencia de un evento asociado con una persona distinto de la muerte, por ejemplo: tiempo de ocurrencia de una enfermedad o el tiempo de ocurrencia de un accidente automovilístico. Para este caso también se puede plantear un modelo de supervivencia mediante la función $S(t) = P(T \geq t)$, ya que la variable T sigue representando el tiempo de ocurrencia o falla de un evento.

En la rama de vida, existen dos tipos de modelos de supervivencia actuariales los cuales son: el modelo agregado y el modelo selecto.

1.2.1. El Modelo Selecto

Considerando un modelo de supervivencia que utilizado para el cálculo de primas de seguros con respecto a personas seleccionadas para una cierta cobertura a edad x, la función $S(t)$ puede tomar distintos valores dependiendo de la edad x, por lo cual necesitamos que $S(t)$ también dependa del valor x cuando $t = 0$. Para estos casos se hace uso del símbolo $S(t; x)$. En este nuevo contexto, la edad seleccionada x es llamada *variable concomitante*. En el ámbito actuarial la forma más usual de esta forma de modelo es simplemente tener por separado una función $S(t)$ para cada valor x.

La edad (x) no es la única variable concomitante que puede tener influencia sobre el modelo de supervivencia, otra variable importante podría ser el sexo, en este caso la variable concomitante probablemente reflejaría la necesidad de tener modelos por separado de hombres y mujeres.

1.2.2. El Modelo Agregado

Esta forma de modelo de supervivencia actuarial se distingue por la particularidad de que el tiempo de inicio $t = 0$ coincide con el nacimiento del individuo $(x = 0)$. Notando que ambas variables se mueven juntas podemos ocupar cualquiera de ellas para representarse en el modelo de interés, por conveniencia se usa x. Es claro que para este caso $S(x)$ y $S(t)$ son funciones idénticas pero ambas difieren de $S(t; x)$. En este caso, la variable aleatoria X es la edad de muerte o el tiempo futuro de vida, así como en su caso T representa una variable aleatoria del tiempo de muerte o tiempo de falla.

1.3. Estimación

Una vez especificada la forma del modelo $S(t)$ (o $S(x)$), es necesario establecer una aproximación o estimación al modelo real, la cual se denota

por $\widehat{S}(t)$. Se utilizarán varias aproximaciones para estimar $S(t)$ dependiendo de la naturaleza de los datos y del diseño del estudio.

Para el modelo no paramétrico usualmente se estiman las probabilidades condicionales de supervivencia sobre pequeños intervalos unitarios (generalmente periodos de un año) y se obtiene una estimación de $S(t)$ a partir de dichas probabilidades.

Respecto al modelo paramétrico, la estimación de los parámetros desconocidos de la supuesta forma de distribución adoptada para el modelo $S(t)$ produce el modelo estimado $\widehat{S}(t)$. En el modelo paramétrico se consigue la estimación por una primera aproximación de sucesiones de intervalos de probabilidad condicional y se adecua a la forma paramétrica elegida y estas aproximaciones llevan a una prueba de hipótesis para modelos paramétricos.

1.4. Ejemplos

Los siguientes ejemplos describen situaciones en donde pueden usarse los modelos de supervivencia.

Ejemplo 1.1 *En la tabla 1.1 se muestra una tabla de vida para menores de 18 años como ejemplo de un modelo de supervivencia actuarial no paramétrico, en la cual se inicia con un radix de 100,000 personas vivas a la edad 0 y se tienen los datos de las muertes año con año. Apartir de las muertes d_x se calculan los valores para q_x y p_x. Para encontrar el valor estimado de $S(t)$ se realiza el cálculo $\widehat{S}(x) = p_0 \cdot p_1 \cdot ... \cdot p_{x-1}$.*

Ejemplo 1.2 *Unos científicos describieron la experiencia de supervivencia de un grupo de pacientes quienes estuvieron bajo tratamiento en conexión con un tipo de enfermedad fatal, dando como resultado la tabla 1.2.*

En este caso se calcularon las probabilidades de supervivencia en cada intervalo de unidad de tiempo, las cuales representan el modelo estimado.

Ejemplo 1.3 *En la tabla 1.3 se muestran 2 grupos de pacientes de leucemia, el tiempo de falla (tiempo hasta la muerte) en semanas y con 2 tipos de sangre y con glóbulos blancos incluidos. En el lado izquierdo se muestran los oh positivos en una muestra de N=17, mientras que del lado derecho los oh negativos con una muestra de N=16.*

En este caso se ilustra como es que se puede hacer uso de una variable concomitante al poder plantear modelos de supervivencia separados para cada tipo de Oh.

(ver [4] pág. 9)

Edad x	l_x	d_x	q_x	p_x	$\widehat{S}(x)$
0	100000	2042	0.02042	0.97958	1
1	97958	131	0.00134	0.99866	0,97958
2	97827	119	0.00122	0.99878	0,97827
3	97708	109	0.00112	0.99888	0,97708
4	97599	101	0.00103	0.99897	0,97599
5	97498	95	0.00097	0.99903	0,97498
6	97403	90	0.00092	0.99908	0,97403
7	97313	86	0.00088	0.99912	0,97313
8	97227	84	0.00086	0.99914	0,97227
9	97143	82	0.00084	0.99916	0,97143
10	97061	82	0.00084	0.99916	0,97061
11	96979	82	0.00085	0.99915	0,96979
12	96897	83	0.00086	0.99914	0,96897
13	96814	84	0.00087	0.99913	0,96814
14	96730	86	0.00089	0.99911	0,96730
15	96644	87	0.00090	0.99910	0,96644
16	96557	89	0.00092	0.99908	0,96557
17	96468	91	0.00094	0.99906	0,96468
18	96377	93	0.00096	0.999004	0,96377

Tabla 1.1

Intervalo (años)	# muertes	# retiros	# expuestos	Prob.sup.estimada
[0, 1)	90	0	374	0.759
[1, 2)	76	0	248	0.556
[2, 3)	51	0	208	0.420
[3, 4)	25	12	157	0.350
[4, 5)	20	5	120	0.291
[5, 6)	7	9	95	0.268
[6, 7)	4	9	79	0.254
[7, 8)	1	3	66	0.250
[8, 9)	3	5	62	0.237
[9, 10)	2	5	54	0.228
[10, ∞)	47	0	47	0

Tabla 1.2

GBI Oh+	Tiempo de muerte	GBI Oh-	Tiempo de muerte
2300	65	4400	56
750	156	3000	65
4300	100	4000	17
2600	134	1500	7
6000	16	9000	16
10500	108	5300	22
10000	121	10000	3
17000	4	19000	4
5400	39	27000	2
7000	143	28000	3
9400	56	31000	8
32000	26	26000	4
35000	22	21000	3
100000	1	79000	30
100000	1	100000	4
52000	5	100000	43
100000	65	-	-

Tabla 1.3

Capítulo 2

Las matemáticas de los modelos de supervivencia

2.1. La distribución de T

Hasta ahora se ha descrito el modelo de supervivencia en términos de la función $S(t)$ la cual representa la probabilidad $P(T \geq t)$ donde T es la variable aleatoria del tiempo de falla, esta función es llamada *función de supervivencia*. A continuación se desarrollan las matemáticas que llevan a entender la naturaleza de los modelos de supervivencia y los conceptos que dentro de ellos se manejan.

2.1.1. Relación de la función de supervivencia con otras funciones

La función de distribución $F(t)$ mide la probabilidad de que la falla ocurra antes del tiempo t.

La relación que existe entre la función de distribución de supervivencia y la función de distribución acumulativa es:

$$F(t) = 1 - S(t); \tag{2.1}$$

(ver [2] pág. 10)

por (1.2) y (1.3) se tiene que $F(0) = 0$ y que $\lim_{t\to\infty} F(t) = 1$

La *función de densidad de probabilidad* de T, se relaciona con $S(t)$ de la siguiente manera:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \tag{2.2}$$

como consecuencia se tiene que:

$$S(t) = \int_t^\infty f(y)dy. \tag{2.3}$$

Es importante notar que $f(t)$ es una función de densidad incondicional de falla al tiempo t, es por esto que esta función mide la densidad de probabilidad de falla al tiempo t dado que el individuo existió al tiempo $t = 0$.

2.1.2. La función de tasa de riesgo

Esta es una medida instantánea de falla al tiempo t dado que se ha sobrevivido al tiempo t, se conoce como *tasa de riesgo* y se denota por $\lambda(t)$. Esta función es el punto de partida para desarrollar un modelo de supervivencia, ya que a partir de la misma se obtiene las funciones antes mencionadas. Su importancia radica en que expresa el riesgo instantáneo de falla de un individuo a través del tiempo y es allí donde entran las suposiciones del modelo de supervivencia al establecer la forma posible del riesgo de falla instantáneo.

Dentro del ámbito de los modelos de supervivencia actuariales, la tasa de riesgo es comúnmente llamada *fuerza de mortalidad* ya que en el estudio de estos modelos la falla considerada es la muerte de los individuos.

Para una definición formal de la tasa de riesgo se considera la probabilidad de que la variable aleatoria asociada con el tiempo de supervivencia T, se encuentre entre t y $t + \delta t$, condicionado con que T sea mayor o igual a t, es decir $P\left(t \leq T < t + \delta t \mid T \geq t\right)$. Esta probabilidad condicional es expresada como una probabilidad por unidad de tiempo entre el intervalo de tiempo δt, para que resulte una tasa. La función de riesgo es el límite de esta cantidad cuando δt tiende a cero, teniendo la expresión:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P\left(t \leq T < t + \delta t \mid T \geq t\right)}{\delta t}. \tag{2.4}$$

(ver [2] pág. 12)

De la definición de la tasa de riesgo dada por (2.4), se puede obtener una relación entre la función de supervivencia y la tasa de riesgo. Usando el resultado de que la probabilidad de un evento A, condicionado a la probabilidad de ocurrencia de un evento B, está dada por $P\left(A \mid B\right) = P\left(AB\right)/P\left(B\right)$, donde $P(AB)$ es la probabilidad conjunta de ocurrencia del evento A y B, la probabilidad condicional dada en la definición de la tasa de riesgo en (2.4) es:

$$\frac{P\left(t \leq T < t + \delta t\right)}{P\left(T \geq t\right)}, \tag{2.5}$$

que es igual a

$$\frac{F\left(t + \delta t\right) - F\left(t\right)}{S(t)}, \tag{2.6}$$

donde $F(t)$ es la función de distribución de probabilidad de T. Sustituyendo en (2.4) se obtiene:

$$\lambda(t) = \lim_{\delta t \to 0} \left(\frac{F\left(t + \delta t\right) - F\left(t\right)}{\delta t} \right) \cdot \frac{1}{S(t)}, \tag{2.7}$$

Se observa que

$$\lim_{\delta t \to 0} \left(\frac{F\left(t + \delta t\right) - F\left(t\right)}{\delta t} \right) \tag{2.8}$$

es la definición de la derivada de $F(t)$ con respecto de t, la cual es $f(t)$. Con esto resulta la relación:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.9}$$

(ver [2] pág. 12)

dada la expresión de $\lambda(t)$ y la igualdad (2.2) se obtiene:

$$\lambda(t) = \frac{-\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\ln S(t) \tag{2.10}$$

e integrando ambos lados:

$$\int_0^t \lambda(y)dy = -\ln S(t), \tag{2.11}$$

despejando,

$$S(t) = \exp\left[-\int_0^t \lambda(y)dy\right]. \tag{2.12}$$

La función acumulativa de riesgo es definida como:

$$\Lambda(t) = \int_0^t \lambda(y)dy = -\ln S(t), \tag{2.13}$$

(ver [2] pág. 12)
por lo que:

$$S(t) = \exp[-\Lambda(t)] \tag{2.14}$$

2.1.3. Modelos de supervivencia actuariales

En el contexto de los modelos de supervivencia actuariales se hace uso de símbolos especiales para los conceptos anteriormente referidos. La función de tasa de riesgo, que es llamada fuerza de mortalidad, es denotada por el símbolo μ_x:

$$\mu_x = \frac{-\frac{d}{dx}S(x)}{S(x)} = -\frac{d}{dx}\ln S(x) \tag{2.15}$$

(ver [1] pág. 55)

Para referirse al primer momento de la variable aleatoria X , que para estos modelos representa el tiempo de vida de una persona, se hace uso del símbolo $\mathring{e}_0$, que es el valor esperado de X sabiendo que el individuo está vivo en $x = 0$; es también conocido como *esperanza de vida al nacer*.

$$\mathring{e}_0 = E[X] = \int_0^\infty x \cdot f(x)dx \tag{2.16}$$

(ver [1] pág. 68)

Para referirse al modelo selecto $S(t;x)$, se recuerda que t es el valor de la variable aleatoria T y x es la edad de la persona quien fue seleccionada para el modelo. Para este caso se hace uso de los símbolos $\mu_{[x]+t}$ para la función de tasa de riesgo y de $\mathring{e}_{[x]}$ para la esperanza de vida de una persona de edad x.

2.2. Modelos de supervivencia paramétricos

En esta sección se exploran varias distribuciones de probabilidad continuas no negativas que son consideradas para aplicarse como modelos de supervivencia paramétricos. En la práctica algunas de estas distribuciones ajustan mejor que otras dada la experiencia que se ha tenido en el estudio de la distribución del tiempo de falla.

2.2.1. Distribución uniforme

La distribución uniforme es una función de probabilidad con dos parámetros y una función de densidad de probabilidad constante. Los parámetros de la función son los límites del intervalo sobre el cual se define la variable aleatoria. De manera que para una variable aleatoria definida en el intervalo $[a, b]$ la función de densidad de probabilidad es la siguiente:

$$f(t) = \frac{1}{b-a}, a \leq t \leq b. \tag{2.17}$$

Para el caso de la variable aleatoria T se tiene que $a = 0$ y b es la longitud del intervalo. Cuando se hace uso de la distribución uniforme como un modelo de supervivencia se hace uso de la letra ω para el valor del parámetro, con lo cual la función de densidad de probabilidad para esta distribución queda definida como:

$$f(t) = \frac{1}{\omega}, 0 \leq t \leq \omega \tag{2.18}$$

de la cual se obtienen las siguentes expresiones:

$$F(t) = \int_0^t f(y)dy = \frac{t}{\omega} \tag{2.19}$$

$$S(t) = 1 - F(t) = \int_t^\omega f(y)dy = \frac{\omega - t}{\omega} \tag{2.20}$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{1}{\omega - t} \tag{2.21}$$

La distribución uniforme usada como un modelo de supervivencia no es apropiada para un amplio rango de tiempo, por lo menos en el caso de un modelo de supervivencia de personas. El mayor uso de esta distribución es sobre intervalos cortos de tiempo (o edad).

2.2.2. Distribución exponencial

Esta distribución queda definida por su función de riesgo constante dada por:

$$\lambda(t) = \theta \tag{2.22}$$

(ver [2] pág. 108)

su función de supervivencia se obtiene por:

$$S(t) = \exp\left(-\int_0^t \theta dy\right) = e^{-\theta t} \tag{2.23}$$

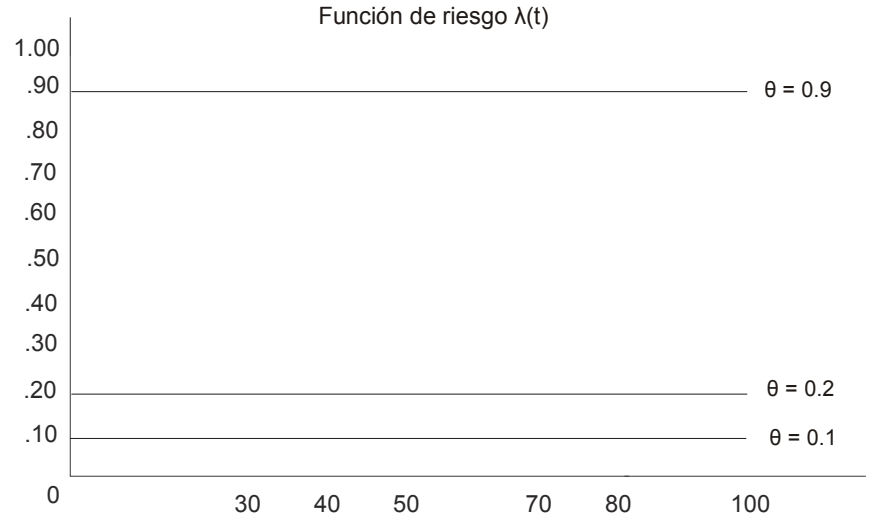
(ver [2] pág. 108)

y su función de densidad es:

$$f(t) = -\frac{d}{dt}S(t) = \theta e^{-\theta t}. \tag{2.24}$$

(ver [2] pág. 108)

Este es un modelo muy simple, dado que en él se supone que el riesgo de muerte para cualquier tiempo a partir del principio del estudio es el mismo. En la siguiente figura se ilustra la simplicidad de la función de riesgo para el modelo exponencial con distintos valores para θ.



La distribución exponencial ha sido utilizada ampliamente como modelo de supervivencia en diversas áreas desde estudios de tiempo de vida en artículos de manufactura hasta en investigaciones de supervivencia en enfermedades crónicas.

Históricamente la distribución exponencial fue la primera en usarse como un modelo para la distribución del tiempo de vida, esto es porque aparentemente esta distribucíon era muy adecuada para representar los tiempos de vida de diversas cosas, especialmente artículos de fabricación.

2.2.3. Distribución Weibull

Esta distribución también se define mediante su función de riesgo:

$$\lambda(t) = \alpha\gamma t^{\gamma-1}, \quad \alpha, \gamma > 0, \tag{2.25}$$

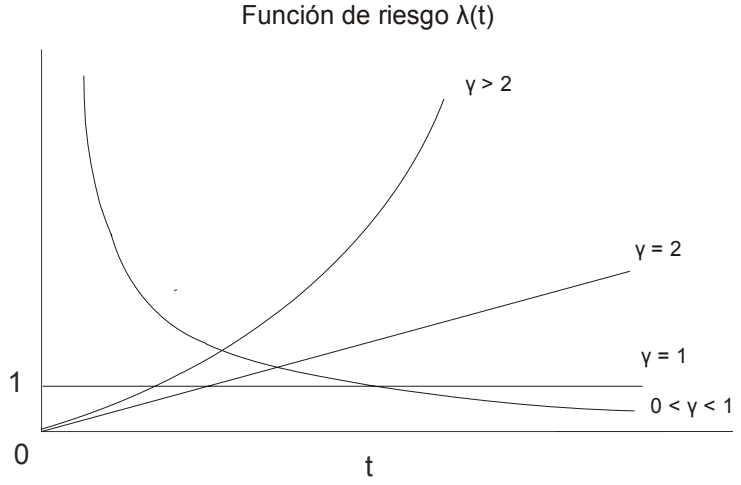
(ver [2] pág. 110)

y su función de distribución de supervivencia está dada por:

$$S(t) = \exp\left[-\int_0^t \alpha\gamma y^{\gamma-1} dy\right] = \exp\left[-\alpha t^{\gamma}\right] \tag{2.26}$$

(ver [2] pág. 110)

En el caso particular donde $\gamma = 1$, la función de riesgo toma un valor constante α y el tiempo de supervivencia tiene una distribución exponencial. Para otros valores de γ, la función de riesgo crece o decrece monótonamente, es decir, no cambia de dirección. La forma general para la función de riesgo de la distribución Weibull , para distintos valores de γ, se muestra en la siguiente figura.



Esta distribución es quizás la más utilizada en modelos de distribución del tiempo de vida. Al igual que la distribución exponencial, es muy usada para los modelos que representan el tiempo de vida de artículos de manufactura, así como en diversas áreas de física y también en estudios biomédicos, por ejemplo, en estudios del tiempo de ocurrencia de tumores en una población.

2.2.4. Distribución Gompertz

Esta distribución fue sugerida como un modelo de supervivencia por Gompertz en el año de 1825 y usualmente está definida por su función de riesgo:

$$\lambda(t) = Bc^t, t \geq 0, B > 0, c > 1 \tag{2.27}$$

(ver [1] pág. 78)

de aquí se obtiene la función de distribución de supervivencia dada por:

$$S(t) = \exp\left[-\int_0^t Bc^y dy\right] = \exp\left[\frac{B}{\ln c}(1 - c^t)\right] \tag{2.28}$$

La función de densidad de probabilidad está dada por $\lambda(t) \cdot S(t)$ y es claro que no tiene una expresión conveniente.

2.2.5. Distribución Makeham

Esta distribución fue propuesta por Makeham en 1860 como una modificación de la distribución de Gompertz, sugiriendo que la función de riesgo tiene una parte independiente de la edad en sí misma, así que fue agregada una constante al modelo de Gompertz:

$$\lambda(t) = A + Bc^t, t \geq 0, B > 0, c > 1, A > -B \tag{2.29}$$

(ver [1] pág. 78)

la función de supervivencia está dada por:

$$S(t) = \exp\left[-\int_0^t (A + Bc^y) dy\right] = \exp\left[\frac{B}{\ln c}(1 - c^t) - At\right] \tag{2.30}$$

Al igual que la distribución de Gompertz, la función de densidad de probabilidad está dada por $\lambda(t) \cdot S(t)$ y es claro que no tiene una forma matemática tratable y el cálculo de probabilidades, momentos y otras cantidades es realmente difícil.

2.3. Formas de medidas condicionales y distribuciones truncadas

Hasta ahora sólo se han considerado medidas de probabilidad desde una edad $x = 0$, denotando estas probabilidades como $S(x)$ o $F(x)$, estas prob-

abilidades son no condicionales. Ahora se considerarán casos en que se sabe que el individuo bajo estudio se encuentra funcionando (o vivo) para $x > 0$ y se buscarán las probabilidades de supervivencia medidas a partir de esa nueva edad $x > 0$.

2.3.1. Probabilidades condicionales

En esta sección el principal interés es encontrar cuál es la probabilidad de que un individuo sobreviva a la edad $x + n$, dado que se sabe que ha sobrevivido hasta la edad x. De manera general, si una probabilidad condicional se multliplica por la probabilidad del evento en el cual se condiciona se obtiene la probabilidad de la intersección de los eventos involucrados. Para este caso, la probabilidad del evento en el cual se condiciona es $S(x)$ y la probabilidad de la intersección de los eventos involucrados $S(x + n)$, tomando esto en cuenta se tiene que la probabilidad condicional, denotada por ${}_np_x$, está dada por:

$$ {}_np_x = \frac{S(x+n)}{S(x)} \tag{2.31} $$

(ver [1] pág. 54)

Por su parte, la probabilidad condicional de falla (o de muerte) antes de la edad $x + n$, denotada por ${}_nq_x$, está dada por:

$$ {}_nq_x = 1 - {}_n\, p_x = \frac{S(x) - S(x+n)}{S(x)} \tag{2.32} $$

(ver [1] pág. 54)

2.3.2. Distribución de X truncada inferiormente

Cuando se habla de probabilidades condicionales donde el evento sobre el cual se condiciona es que se sobreviva a la edad x, se hace referencia a la distribución en un subconjunto del espacio muestral que corresponde a los valores de X que exceden a x. Esta distribución es llamada la d*istribución de X truncada debajo de x.*

De esta forma la probabilidad de supervivencia condicional ${}_np_x$ queda establecida de la siguiente manera:

$$ {}_np_x = P(X > x + n \mid X > x) = S(x + n \mid X > x) \tag{2.33} $$

esta ecuación representa la probabilidad de que la edad de falla (o muerte) sea superior a $x+n$, dado que ésta excedió a x. Es claro que esta es la misma probabilidad mencionada en la sección anterior, por lo cual, de (2.31) y (2.33) se obtiene la igualdad:

$$ S(x + n \mid X > x) = \frac{S(x + n)}{S(x)} \tag{2.34} $$

De igual forma se tiene:

$$ \begin{aligned} {}_nq_x &= P(X \leq x + n \mid X > x) \\ &= P(x < X \leq x + n \mid X > x) = F(x + n \mid X > x) \end{aligned} \tag{2.35} $$

Comparando (2.32) y (2.35) se nota que:

$$ F(x + n \mid X > x) = \frac{S(x) - S(x + n)}{S(x)} = \frac{F(x + n) - F(x)}{1 - F(x)} \tag{2.36} $$

La *función de densidad de probabilidad condicional* para un tiempo de falla (o de muerte) y, dado que está vivo a la edad x $(y > x)$, es denotada por $f(y \mid X > x)$ y se obtiene por:

$$ f(y \mid X > x) = \frac{d}{dy} F(y \mid X > x) = \frac{d}{dy} \left[\frac{F(y) - F(x)}{1 - F(x)} \right] = \frac{f(y)}{1 - F(x)} \tag{2.37} $$

debido a que $\frac{d}{dy}F(x) = 0$. Como $F(x) = 1 - S(x)$ se tiene:

$$ f(y \mid X > x) = \frac{f(y)}{S(x)} \tag{2.38} $$

A este resultado también se puede llegar de una forma intuitiva dado que del lado izquierdo de (2.38) lo que se tiene es la probabilidad condicional

de falla al tiempo y dado que se ha sobrevivido al tiempo x $(y > x)$, si se multiplica esta probabilidad condicional por la probabilidad del evento sobre el cual se condiciona, que es $S(x)$, se obtiene la probabilidad de la intersección de los eventos que es estar vivo a la edad y.

Por su parte, la *funcion de tasa de riesgo condicional* se denota por $\lambda(y \mid X > x)$ y al calcular su expresión se obtiene un resultado particular:

$$\lambda(y \mid X > x) = \frac{f(y \mid X > x)}{S(y \mid X > x)} = \frac{f(y)}{S(x)} \div \frac{S(y)}{S(x)} = \frac{f(y)}{S(y)} = \lambda(y) \qquad (2.39)$$

es la misma expresión que para la distribución no truncada.

2.3.3. Distribución de X truncada superior e inferiormente

El caso general para el estudio de una distribucíon truncada es tomar en cuenta los valores de la variable aleatoria X sobre un intervalo determinado, de manera que la función de distribución de supervivencia queda establecida por:

$$S(x \mid y < X \leq z) = P(X > x \mid y < X \leq z) \qquad (2.40)$$

para $y < x \leq z$.

Esta expresión representa la probabilidad de que la falla (o muerte) ocurra después de la edad x, dado que ésta ocurre entre y y z (realmente se habla de una probabilidad de muerte entre y y z). De igual forma, si se multiplica esta probabilidad condicional por la probabilidad de la condición, que es $S(y) - S(z)$, se obtiene la probabilidad no condicional de falla (o muerte) entre x y z, que es $S(x) - S(z)$, con lo cual resulta:

$$S(x \mid y < X \leq z) = \frac{S(x) - S(z)}{S(y) - S(z)}. \qquad (2.41)$$

La correspondiente funcion de distribución de probabilidad doblemente truncada está dada por:

$$F(x \mid y < X \leq z) = P(y < X \leq x \mid y < X \leq z) \tag{2.42}$$

como $F(x \mid y < X \leq z) = 1 - S(x \mid y < X \leq z)$ se tiene:

$$F(x \mid y < X \leq z) = \frac{S(y) - S(x)}{S(y) - S(z)} = \frac{F(x) - F(y)}{F(z) - F(y)}. \tag{2.43}$$

La funcion de densidad de probabilidad doblemente truncada queda expresada como:

$$f(x \mid y < X \leq z) = -\frac{d}{dx} S(x \mid y < X \leq z) = -\frac{d}{dx}\left[\frac{S(x) - S(z)}{S(y) - S(z)}\right] \tag{2.44}$$

como $-\frac{d}{dx}S(x) = f(x)$ y $-\frac{d}{dx}S(z) = 0$ se tiene:

$$f(x \mid y < X \leq z) = \frac{f(x)}{S(y) - S(z)}. \tag{2.45}$$

Finalmente, la función tasa de riesgo doblemente truncada se obtiene como:

$$\lambda(x \mid y < X \leq z) = \frac{f(x \mid y < X \leq z)}{S(x \mid y < X \leq z)}, \tag{2.46}$$

de donde,

$$\lambda(x \mid y < X \leq z) = \frac{f(x)}{S(y) - S(z)} \div \frac{S(x) - S(z)}{S(y) - S(z)} = \frac{f(x)}{S(x) - S(z)}. \tag{2.47}$$

Como $f(x) = \lambda(x) \cdot S(x)$, se tiene que:

$$\lambda(x \mid y < X \leq z) = \frac{\lambda(x) \cdot S(x)}{S(x) - S(z)} \tag{2.48}$$

Es importante resaltar que (2.39) muestra cómo para el caso de una distribución truncada la función de tasa de riesgo no se ve afectada por un truncamiento inferior, lo cual es un resultado intuitivo, dado que la función de tasa de riesgo es una función condicional a la supervivencia de x, por lo cual, cualquier truncamiento antes de x no afecta a esta función. Por otra parte, en (2.48) se nota cómo un truncamiento superior de la función de distribución de X sí afecta a la función tasa de riesgo, y esto es porque el tiempo probable de falla restante se ve reducido.

2.3.4. La tasa central

Otro tipo de medida condicional sobre el intervalo de edad $[x, x+1]$ es llamada *tasa central de mortalidad* y es denotada por m_x. Ésta es definida como el valor de peso promedio de la función de tasa de riesgo sobre el intervalo $[x, x+1]$, usando como peso la función $\lambda(y)$ y la probabilidad de sobrevivir a la edad y.

$$m_x = \frac{\int_x^{x+1} S(y) \cdot \lambda(y) dy}{\int_x^{x+1} S(y) dy}. \tag{2.49}$$

(ver [1] pág. 70)

Más generalmente, ${}_nm_x$ es el riesgo promedio o tasa central de mortalidad, sobre el intervalo $[x, x+n]$ y está dada por:

$${}_nm_x = \frac{\int_x^{x+n} S(y) \cdot \lambda(y) dy}{\int_x^{x+n} S(y) dy}, \tag{2.50}$$

(ver [1] pág. 70)

aplicando el cambio de variable $y = x + s$, resulta:

$${}_nm_x = \frac{\int_0^n S(x+s) \cdot \lambda(x+s) ds}{\int_0^n S(x+s) ds} \tag{2.51}$$

dividiendo el numerador y el denominador por $S(x)$ se obtiene:

$${}_nm_x = \frac{\int_0^n \frac{S(x+s)}{S(x)} \cdot \lambda(x+s) ds}{\int_0^n \frac{S(x+s)}{S(x)} ds} = \frac{\int_0^n {}_sp_x \mu_{x+s} ds}{\int_0^n {}_sp_x ds} \tag{2.52}$$

donde ${}_sp_x$ es la probabilidad condicional $\frac{S(x+s)}{S(x)}$ y μ_{x+s} es el símbolo actuarial estándar para $\lambda(x+s)$.

En el siguiente capítulo se presentan los métodos de estimación paramétrica utlilizados para este tipo de modelos, así como algunas pruebas estadísticas que serviran para validar la aplicación de un modelo.

Capítulo 3

Estimación de modelos de supervivencia paramétricos

Este capítulo se enfoca a la estimación de los modelos de supervivencia en forma paramétrica. Se consideran dos grupos de modelos de supervivencia parámetricos. El primer grupo consiste en modelos univariados $S(t)$ (o $S(x)$) de los cuales se estiman sus parámetros con muestras de datos no censurados y con datos censurados. En el segundo grupo se considera el modelo de supervivencia paramétrico que incluye variables concomitantes como lo es el modelo selecto $S(t;x)$. Se consideran varias aproximaciones para la estimación paramétrica, haciendo énfasis en los métodos de mínimos cuadrados y máxima verosimilitud.

3.1. Modelos univariados con datos no censurados

En el estudio de los modelos de supervivencia paramétricos, se hace una distinción entre los modelos que involucran muestras en los cuales se conoce el tiempo exacto de falla (o muerte) y aquellos en los que el tiempo de falla (o muerte) ha sido agrupado. Esta sección se enfoca a la estimación de los modelos de supervivencia paramétricos en los cuales la muestra se presenta con datos no censurados y los tiempos de falla (o muerte) son totalmente conocidos.

3.1.1. Método de momentos

Supónga que se tiene una muestra de n individuos que se sabe que existen al tiempo $t = 0$ y en cada uno de ellos se observa su tiempo de falla. Suponiendo que estos tiempos de falla son independientes, se obtienen los datos $t_1, t_2, t_3 \ldots, t_n$ (cada uno como realizaciones de las variables aleatorias $T_1, T_2, T_3..., T_n$); para este caso se desea estimar $S(t)$ como un modelo de supervivencia paramétrico, es decir, se adopta una función particular de la variable T, dependiente de uno o más parámetros y se usan los datos observados para estimar esos parámetros desconocidos.

El método de momentos para la estimación de los parámetros consiste en igualar el r-ésimo momento muestral con el r-ésimo momento poblacional, planteando así un sistema de ecuaciones y resolviendo para cada parámetro desconocido.

La forma más facil para estimar un modelo de supervivencia paramétrico es usando una función que dependa de una sola variable. Para este caso se toma de ejemplo la distribución exponencial con $S(t) = e^{-\theta t}$ para estimar su único parámetro θ.

Sea

$$\overline{t} = \frac{1}{n} \cdot \sum_{i=1}^{n} t_i \tag{3.1}$$

que es el tiempo promedio de falla de la muestra, también conocido como primer momento muestral. Por otro lado se sabe que $E[T] = \frac{1}{\theta}$ porque T se distribuye exponencial. Aplicando el método de momentos se tiene:

$$\widehat{\theta} = \frac{1}{\overline{t}} = \frac{n}{\sum_{i=1}^{n} t_i} \tag{3.2}$$

que es el valor estimado para el parámetro suponiendo una distribución exponencial.

Para un modelo paramétrico con dos parámetros, es necesario conocer el segundo momento muestral para aplicar el metódo de momentos.

3.1.2. Método de máxima verosimilitud

Para aplicar este método se obtiene la función de verosimilitud ($L(\theta)$) y el desarrollo consiste en maximizar la función de log-verosimulitud ($\ln L(\theta)$) con respecto a cada parámetro que se desea estimar.

Para este caso, nuevamente se toma de ejemplo la distribución exponencial con $S(t) = e^{-\theta t}$ y $f(t_i) = \theta e^{-\theta t_i}$.La función de verosimilitud está dada por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \theta e^{-\theta t_i} = \theta^n \cdot e^{-\theta \cdot \sum_{i=1}^{n} t_i} \tag{3.3}$$

y la función de log-verosimilitud es:

$$\ln L(\theta) = n \cdot \ln \theta - \theta \cdot \sum_{i=1}^{n} t_i \tag{3.4}$$

tomando la derivada de $\ln L(\theta)$ con respecto a θ e igualando a cero se tiene:

$$\frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} t_i = 0 \tag{3.5}$$

resolviendo se obtiene el estimador para θ el cual es:

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\bar{t}} \tag{3.6}$$

que es el mismo que se obtuvo por el método de momentos.

Para constatar que $\widehat{\theta}$ es un punto máximo, se encuentra la segunda derivada de $\ln L(\theta)$ con respecto a θ y se evalúa en el punto obtenido. si resulta un valor menor a cero, entonces $\widehat{\theta}$ es un valor máximo.

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{n}{\theta^2} < 0, \quad \forall \quad \theta \in \mathrm{R} \tag{3.7}$$

por lo tanto, $\widehat{\theta}$ si es un punto máximo.

Ejemplo 3.1 *En un experimento con 6 ratones recién nacidos se observan los siguientes tiempos exactos de muerte: 0.6, 2.2, 2.3, 3.1, 4.6 y 7.2. Se ajusta un modelo de supervivencia exponencial a los datos y se calcula el valor de* $\widehat{\theta}$ *mediante el método de máxima verosimilitud.*

Por los datos observados se tiene que $\bar{t} = \frac{1}{6} \cdot \sum_{i=1}^{6} t_i = \frac{20}{6}$, *y por el método de máxima verosimilitud se obtiene el valor* $\widehat{\theta} = \frac{1}{\bar{t}} = \frac{6}{20} = 0.3$

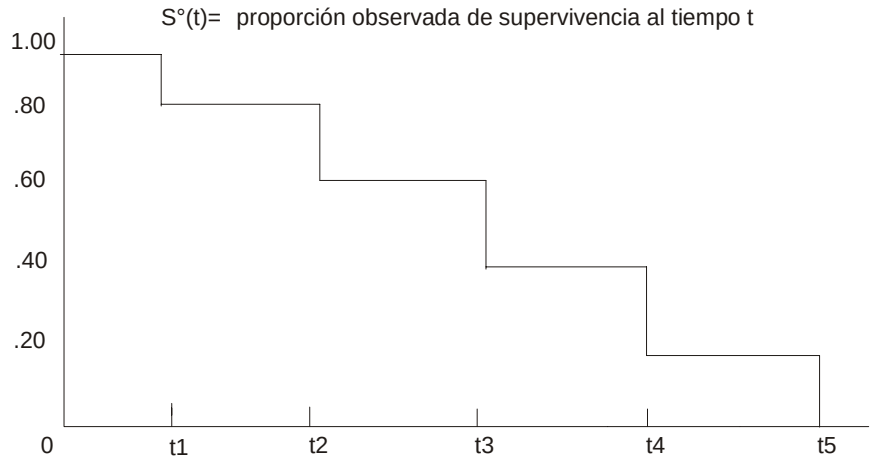
3.1.3. Método de mínimos cuadrados

Para aplicar este método se debe desarrollar primero un modelo que se llama *función de supervivencia empírica*, el cual se define de la siguiente manera: sean $t_1, t_2, t_3 ... t_n$ los tiempos de falla observados en los individuos bajo estudio, la función de supervivencia empírica está dada por:

$$S^{\circ}(t) = \begin{cases} 1 & si \quad t < t_1 \\ \frac{n-i}{n} & si \quad t_i \leq t < t_{i+1}; \quad i = 1, 2, 3, ..., n-1 \\ 0 & si \quad t \geq t_n \end{cases} \tag{3.8}$$

(ver [8] pág. 82)

Lo que se expresa en esta función es que se toma un modelo observando la propoción de supervivencia de acuerdo a los datos obtenidos en la muestra, de esta forma se construye una distribución de supervivencia empírica de acuerdo al remanente de individuos vivos bajo estudio. La siguiente gráfica muestra una posible forma de esta función para una muestra de tiempos de falla con $n = 5$:



Cabe aclarar que esta distribución es totalmente empírica y es sólo la interpretacion de los datos observados, es decir, no se debe mal interpretar pensando que, de acuerdo a esta función, la probabilidad de falla antes del tiempo t_1 es cero, ni tampoco que no se pueda sobrevivir más allá de un cierto tiempo t_5, es sólo que la muestra observada presenta este comportamiento.

Una vez teniendo esta función empírica, se adopta un modelo paramétrico para $S(t)$ dependiente de ciertos parámetros desconocidos y se considera la siguiente suma de cuadrados definida por:

$$SS = \sum_{i=1}^{n} [S(t_i) - S^{\circ}(t_i)]^2 \tag{3.9}$$

donde los únicos parámetros desconocidos son los de $S(t_i)$.

Hay que notar que (3.9) define las desviaciones entre $S(t)$ y $S^{\circ}(t)$ para la parte inferior izquierda de la función $S^{\circ}(t)$, este procedimiento ejerce una pequeña tendencia descendente al momento de calcular los valores estimados para los parámetros desconocidos de $S(t)$. De igual forma redefiniendo la función SS de la siguiente manera:

$$SS = \sum_{i=1}^{n} [S(t_i) - S^{\circ}(t_{i-1})]^2 \tag{3.10}$$

(ver [8] pág. 186)

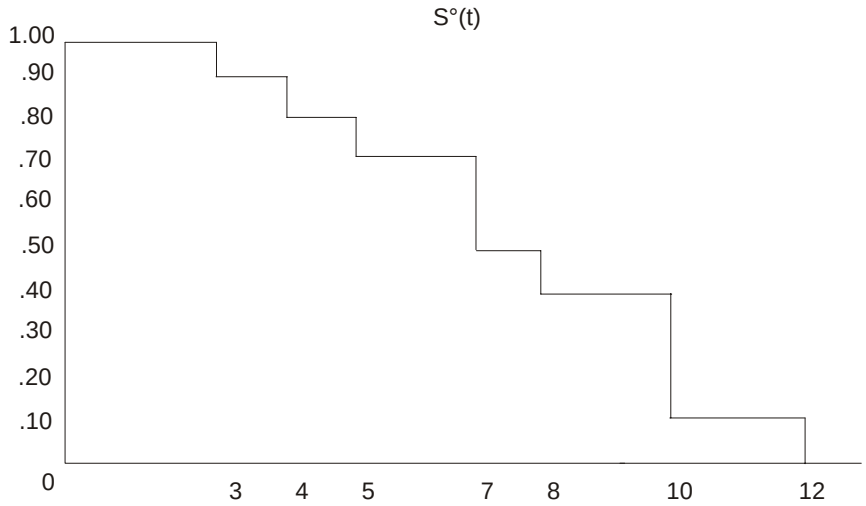
que ajusta los datos de $S(t)$ y $S^{\circ}(t)$ para la parte superior derecha de $S^{\circ}(t)$, se obtiene una cierta tendencia ascendente para el cálculo de los parámetros desconocidos de $S(t)$. Por lo anterior, es mejor considerar los puntos medios de los intervalos entre $S^{\circ}(t_i)$ y $S^{\circ}(t_{i-1})$ para realizar un mejor ajuste en el modelo, de lo cual se redefine la función SS como:

$$SS = \sum_{i=1}^{n} \left(S(t_i) - \frac{1}{2} \left[S^{\circ}(t_i) + S^{\circ}(t_{i-1}) \right] \right)^2 . \tag{3.11}$$

(ver [8] pág. 187)

El método de mínimos cuadrados consiste en derivar la función SS con respecto a cada parámetro, igualar cada derivada a cero y resolver para cada parámetro desconocido. Este método de estimación es fácil de aplicar cuando los parámetros a estimar son de forma lineal dentro de la función SS, o cuando las derivadas parciales son de forma lineal con respecto a los parámetros desconocidos. Como éste no es el caso más frecuente dentro de la estimación, en varias ocasiones es necesario aplicar una transformación, frecuentemente logarítmica, para poder desarrollar dicho método.

Ejemplo 3.2 *Supóngase que se realiza un estudio y se tiene la siguiente muestra que refleja los tiempos de muerte (en días) de ciertos individuos en estudio en un laboratorio: 3, 4, 5, 7, 7, 8, 10, 10, 10, 12, como se repiten algunos valores se considera* $n = 7$ *que son sólo los datos distintos dentro de la muestra. La correspondiente forma de* $S^{\circ}(t)$ *se muestra en la siguiente gráfica:*



Se trata de ajustar los datos a un modelo con distribución exponencial y se estima el parámetro θ.

Se tiene:

$$SS = \sum_{i=1}^{7} \left(e^{-\theta t_i} - \frac{1}{2} \left[S^{\circ}(t_i) + S^{\circ}(t_{i-1}) \right] \right)^2 \tag{3.12}$$

como $\frac{dSS}{d\theta} = 0$ *es una ecuación exponencial en* θ*, se aplica la transformación* $\ln S(t_i) = -\theta t_i$ *y* $\ln\left(\frac{1}{2}\left[S^{\circ}(t_i) + S^{\circ}(t_{i-1})\right]\right)$ *con lo cual se obtiene:*

$$SS = \sum_{i=1}^{7} \left(-\theta t_i - \ln \left\{ \frac{1}{2} \left[S^{\circ}(t_i) + S^{\circ}(t_{i-1}) \right] \right\} \right)^2, \tag{3.13}$$

derivando con respecto a θ

$$\frac{dSS}{d\theta} = 2 \cdot \sum_{i=1}^{7} \left(-\theta t_i - \ln \left\{ \frac{1}{2} \left[S^{\circ}(t_i) + S^{\circ}(t_{i-1}) \right] \right\} \right) (-t_i) \tag{3.14}$$

igualando a cero

$$\theta \cdot \sum_{i=1}^{7} t_i^2 + \sum_{i=1}^{7} t_i \cdot \ln \left\{ \frac{1}{2} \left[S^{\circ}(t_i) + S^{\circ}(t_{i-1}) \right] \right\} = 0 \tag{3.15}$$

con lo que se obtiene el estimador:

$$\widehat{\theta} = \frac{-\sum_{i=1}^{7} t_i \cdot \ln\left\{\frac{1}{2}\left[S^{\circ}(t_i) + S^{\circ}(t_{i-1})\right]\right\}}{\sum_{i=1}^{7} t_i^2} \tag{3.16}$$

La siguiente tabla muestra los valores necesarios para evaluar $\widehat{\theta}$.

i	t_i	$S^{\circ}(t_i)$	t_i^2	$t_i \cdot \ln\left\{\frac{1}{2}\left[S^{\circ}(t_i) + S^{\circ}(t_{i-1})\right]\right\}$
1	*3*	*0.90*	*9*	*-0.15387*
2	*4*	*0.80*	*16*	*-0.65007*
3	*5*	*0.70*	*25*	*-1.43841*
4	*7*	*0.50*	*49*	*-3.57578*
5	*8*	*0.40*	*64*	*-6.38806*
6	*10*	*0.10*	*100*	*-13.86294*
7	*12*	*0.00*	*144*	*-35.94879*
	49		*407*	*-62.01792*

con lo cual se obtiene el valor estimado $\widehat{\theta} = \frac{62.01792}{407} = 0.15238$.

(ver [8] pág. 188)

3.1.4. Método de medianas

Este método sólo se puede aplicar a modelos dependientes de un sólo parámetro. Sea $\widetilde{t}$ que representa la mediana de la muestra. La mediana de la distribución es el valor de t (denotado $t_{1/2}$) para el cual $S(t_{1/2}) = \frac{1}{2}$. El método consiste en igualar las medianas y resolver para el parámetro desconocido.

Suponiendo un modelo de supervivencia con distribucíon exponencial, se tiene que resolver $S(t_{1/2}) = \frac{1}{2}$ o $e^{-\theta t_{1/2}} = \frac{1}{2}$ para $t_{1/2}$ resultando:

$$t_{1/2} = -\frac{\ln \frac{1}{2}}{\theta} = \widetilde{t}, \tag{3.17}$$

(ver [8] pág. 183)

quedando el estimador para θ como:

$$\widehat{\theta} = -\frac{\ln \frac{1}{2}}{\widetilde{t}}. \tag{3.18}$$

3.1.5. Método de percentiles

Este método es una extensión del método de medianas para aplicarse a un modelo con más de un parámetro.

Si t_p es el valor de t tal que $S(t_p) = 1 - p$, entonces t_p es el $100p - ésimo$ percentil de la distribución. Si $p = \frac{1}{2}$, entonces t_p es la mediana. Para aplicar este método se calculan los correspondientes percentiles de la muestra y se igualan con los de la distribución supuesta para el modelo (que se encuentran en términos de los parámetros desconocidos) y se resuelve para cada parámetro.

3.1.6. Tiempos de muerte agrupados

Este diseño de estudio es más común en las técnicas de estimacíon para modelos no paramétricos. Al igual que en el caso anterior, se aplican los métodos de máxima verosimilitud y mínimos cuadrados para realizar la estimación de los modelos de supervivencia paramétricos.

Máxima verosimilitud

Supóngase que se tiene n individuos vivos en $t = 0$ y se observan las muertes de éstos en k intervalos independientes de igual longitud. Sea d_i el número de muertes observadas en el intervalo $(i, i+1]$. La probabilidad de muerte en $(i, i+1]$ de un individuo vivo en $t = 0$ es $S(i) - S(i+1)$, con lo cual, la aportación a la función de verosimilitud del i-ésimo intervalo es:

$$L_i(\theta) = [S(i) - S(i+1)]^{d_i}, \tag{3.19}$$

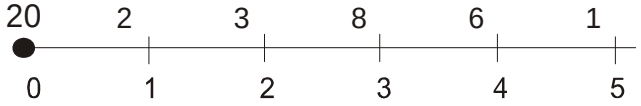
(ver [8] pág. 190)

por lo tanto, la función de verosimilitud total es:

$$L(\theta) = \prod_{i=0}^{k-1} [S(i) - S(i+1)]^{d_i}. \tag{3.20}$$

(ver [8] pág. 190)

Ejemplo 3.3 *Se considera una muestra de 20 individuos que existen al tiempo $t = 0$. Todos ellos mueren dentro de un periodo de 5 semanas y sólo se conoce la semana en que murieron: 2 en la primera semana, 3 en la segunda, 8 en la tercera, 6 en la cuarta y 1 en la quinta. Se tiene que ajustar un modelo de supervivencia exponencial para este grupo de datos. El siguiente diagrama muestra la forma del estudio.*



En general, $S(i) - S(i+1) = e^{-\theta i} - e^{-\theta(i+1)} = e^{-\theta i}(1 - e^{-\theta})$, con lo cual se tiene:

$$L(\theta) = \prod_{i=0}^{4} \left[e^{-\theta i}(1 - e^{-\theta})\right]^{d_i} = e^{-\theta \sum_{i=0}^{4} i \cdot d_i} \cdot (1 - e^{-\theta})^{\sum_{i=0}^{4} d_i} \tag{3.21}$$

$$\ln L(\theta) = -\theta \cdot \sum_{i=0}^{4} i \cdot d_i + \ln(1 - e^{-\theta}) \cdot \sum_{i=0}^{4} d_i, \tag{3.22}$$

haciendo $\frac{d \ln L(\theta)}{d\theta} = 0$ resulta:

$$e^{-\widehat{\theta}} = \frac{\sum_{i=0}^{4} i \cdot d_i}{\sum_{i=0}^{4} d_i + \sum_{i=0}^{4} i \cdot d_i}. \tag{3.23}$$

De los datos se obtiene $\sum_{i=0}^{4} d_i = 20$ y $\sum_{i=0}^{4} i \cdot d_i = 41$, lo cual produce el estimador $\widehat{\theta} = 0.3973$

Mínimos cuadrados

Se reconoce que $S(i) - S(i+1) = S(i) \cdot q_i$ es la probabilidad multinomial de que un individuo de la muestra original de n muera en el intervalo $(i, i+1]$, entonces $n\,[S(i) - S(i+1)]$ corresponde al número esperado de muertes en

el intervalo $(i, i+1]$, como d_i es el número observado de muertes en dicho intervalo, el método de estimación por mínimos cuadrados se realiza minimizando:

$$SS = \sum_{i=0}^{k-1} \{n\,[S(i) - S(i+1)] - d_i\}^2 . \tag{3.24}$$

(ver [8] pág. 191)

3.2. Modelos univariados con datos censurados

Esta sección se enfoca a la estimación de modelos supervivencia paramétricos con datos censurados, es decir, se consideran muestras que han sido truncadas de acuerdo al diseño del experimento y a partir de dichas muestras se estiman los parámetros adecuados para cada modelo. Se considera el método de máxima verosimilitud, con adecuaciones respecto al presentado en la sección anterior debido a la presencia de censura.

3.2.1. Método de máxima verosimilitud

Supóngase un estudio en el cual se tienen n individuos vivos en $t = 0$, en él se observa el tiempo exacto de cada muerte para un tiempo mayor a $t = r$ y se suspende la observación en alguna t donde todavia continúan individuos vivos. Como no todos los individuos en estudio han muerto, se tiene una situación con datos censurados. En el contexto médico se dice que el estudio ha sido truncado.

Para el caso de este diseño de estudio se aplica la aproximación por el método de máxima verosimilitud, sólo que la función de verosimilitud tiene una modificación dada por la alteración que en ésta se produce debido al truncamiento de los datos. La aportación de cada muerte a la función de verosimilitud corresponde a su respectiva probabilidad de muerte hasta $t = r$ y la contribución de cada sobreviviente hasta el tiempo $t = r$ es simplemente la probabilidad de sobrevivir hasta $t = r$, la cual es $S(r)$. Si de la muestra

de n individuos se observan d muertes en total, la función de verosimilitud queda de la siguiente manera:

$$L(\theta) = [S(r)]^{n-d} \cdot \prod_{i=1}^{d} f(t_i). \tag{3.25}$$

(ver [2] pág. 116)

Ejemplo 3.4 *Considere los datos descritos en el estudio del ejemplo 3.2 y que éstos han sido truncados en* $t = 9$*. Al estimar el parámetro* θ *para este caso se observa que en* $t = 9$ *han ocurrido* 6 *muertes y* 4 *continúan con vida, se tiene:*

$$L(\theta) = [S(9)]^4 \cdot \prod_{i=1}^{6} f(t_i), \tag{3.26}$$

en este caso $S(9) = e^{-9\theta}$ *y* $f(t_i) = \theta e^{-\theta t_i}$*, con* $t_i = 3, 4, 5, 7, 7, 8$*, sustituyendo:*

$$L(\theta) = \left[e^{-9\theta}\right]^4 \cdot \prod_{i=1}^{6} \theta e^{-\theta t_i} = \theta^6 \cdot e^{-36\theta} \cdot e^{-34\theta}, \tag{3.27}$$

derivando $\ln L(\theta)$ *y haciendo* $\frac{d \ln L(\theta)}{d\theta} = 0$*, se obtiene* $\widehat{\theta} = \frac{6}{70} = 0.08571$*.*

De manera más general, supóngase que el i-ésimo individuo en estudio se mantiene en observación al tiempo r_i y sale de dicha observación al tiempo t_i, un individuo puede salir de la observación por muerte o por el fin de el periodo en el cual sería observado. Se supone que la muerte es la única causa aleatoria de salida. La probabilidad de sobrevivir del tiempo r_i a t_i es ${}_{t_i - r_i}p_{r_i} = \frac{S(t_i)}{S(r_i)}$ que es la aportación total a la función de verosimilitud de los individuos vivos. Para los individuos muertos se usa la función de densidad de probabilidad, la cual se obtiene multiplicando ${}_{t_i - r_i}p_{r_i}$ por $\lambda(t_i)$, se tiene:

$$L(\theta) = \prod_{\overline{D}} \frac{S(t_i)}{S(r_i)} \cdot \prod_{D} \frac{S(t_i) \cdot \lambda(t_i)}{S(r_i)}, \tag{3.28}$$

(ver [8] pág. 193)

donde el primer producto denota a todos los individuos con vida y el segundo a las muertes que se presentaron durante el periodo de observación.

Definiendo la siguiente variable indicadora:

$$\delta_i = \begin{cases} 1 & \textit{si el individuo } i \textit{ muere en } (t, t+1] \\ 0 & \textit{en otro caso,} \end{cases} \tag{3.29}$$

como (3.28) contiene $\frac{S(t_i)}{S(r_i)}$ para todos los n individuos en estudio y $\lambda(t_i)$ solo para las muertes, se puede escribir la función de verosimilitud de la siguiente forma:

$$L(\theta) = \prod_{i=1}^{n} \frac{S(t_i) \cdot [\lambda(t_i)]^{\delta_i}}{S(r_i)}. \tag{3.30}$$

(ver [8] pág. 194)

Ejemplo 3.5 *Considere el siguiente cuadro que presenta a seis pacientes que recibieron un transplante de corazón y se mantuvieron en observación hasta el 31 de Diciembre de 2005.*

Paciente	*Fecha del transplante*	*Fecha de muerte*
1	*1 Enero 2004*	*1 Abril 2005*
2	*1 Abril 2004*	*1 Abril 2005*
3	*1 Julio 2004*	—
4	*1 Octubre 2004*	*1 Julio 2005*
5	*1 Enero 2005*	—
6	*1 Abril 2005*	*1 Octubre 2005*

Se estima el parámetro del modelo suponiendo una distribución exponencial y tomando el año calendario de 2005 como el periodo de observación.

Al momento de entrar en el estudio, los seis pacientes tienen los siguientes valores de r_i*:* $r_1 = 1, r_2 = 0.75, r_3 = 0.50, r_4 = 0.25, r_5 = 0$ *y* $r_6 = 0$*. Al tiempo de salida del estudio, los valores de* t_i *son los siguientes:* $t_1 = 1.25, t_2 = 1, t_3 = 1.50, t_4 = 0.75, t_5 = 1$ *y* $t_6 = 0.50$*. Hay que notar que los tiempos* t_3 *y* t_5 *son tiempos en los cuales la observación ha finalizado, mientras que los demás* t_i *son tiempos de muerte.*

De (3.30) se tiene $\ln L(\theta) = \sum_{i=1}^{6} [-\theta \cdot t_i + \theta \cdot r_i + \ln \theta \cdot \delta_i]$*. Derivando y haciendo* $\frac{d \ln L(\theta)}{d\theta} = 0$*, se obtiene* $\widehat{\theta} = \frac{\sum_{i=1}^{6} \delta_i}{\sum_{i=1}^{6} (t_i - r_i)} = \frac{4}{3.5} = 1.1429$*.*

(ver [8] pág. 194)

3.3. Pruebas de hipótesis de modelos paramétricos

En esta sección se exploran brevemente las pruebas estadísticas formales que son usadas para determinar la aceptación de un modelo paramétrico como una adecuada representación del modelo real $S(t)$ en cuestión. Se realizan las pruebas de hipótesis para los diferentes diseños de estudio anteriormente tratados, es decir, tanto para el caso de tiempos de muerte agrupados como para el caso donde los tiempos de muerte son totalmente conocidos. En ambos casos se supone que los datos son no censurados.

3.3.1. Prueba χ^2 (Ji-cuadrada)

Esta prueba estadística está diseñada para emplearse en un modelo estimado para muestras de tiempos de muerte agrupados. Supóngase que mediante algún procedimiento anteriormente descrito se ha obtenido el modelo estimado $\widehat{S}(t)$ y ahora lo que se desea es saber qué tan buena aproximación es este modelo estimado del modelo real $S(t)$.

Sea

$$\widehat{E}_i = n\left[\widehat{S}(i) - \widehat{S}(i+1)\right], \quad i = 0, 1, 2, ..., k-1 \tag{3.31}$$

(ver [3] pág. 186)

el número estimado de individuos muertos en el intervalo $(i, i+1]$ de acuerdo con el modelo estimado. Se recuerda que para la obtención del modelo estimado para muestras de tiempos agrupados, se subdivide el tiempo de observación en k intervalos de igual longitud. Tomando d_i, que representa el número de individuos muertos observados en $(i, i+1]$, se usa la estadística de prueba

$$\chi^2 = \sum_{i=0}^{k-1} \frac{(\widehat{E}_i - d_i)^2}{\widehat{E}_i}, \tag{3.32}$$

(ver [3] pág. 186)

para el desarrollo de la prueba de hipótesis. Dicha cantidad aproximadamente se distribuye como una variable aleatoria χ^2 con $k-1-r$ grados

de libertad, donde r es el número de parámetros desconocidos estimados en $S(t)$.

La prueba de consiste en plantear la hipótesis nula H_0 = el modelo ajusta a la distribucuón supuesta estimada *vs* H_a = el modelo no ajusta correctamente a la distribución supuesta estimada. La regla de desición es rechazar H_0 con grado de significancia α si χ^2 excede el valor en tablas de $\chi^{2(1-\alpha)}_{k-1-r}$, en otro caso aceptar H_0. (para mayores referencias, ver [3] pág. 186)

Ejemplo 3.6 *Se realiza la prueba de hipótesis χ^2 para ver si el modelo $\widehat{S}(t)$ determinado en el ejemplo 3.3 es una buena representación de $S(t)$. Como $\widehat{\lambda} = 0.3973$ se calculan los valores $\widehat{E}_i$ y se obtiene la siguiente tabla para $k = 6$:*

i	$\widehat{E}_i$	d_i	$(\widehat{E}_i - d_i)^2$	$\frac{(\widehat{E}_i - d_i)^2}{\widehat{E}_i}$
0	*6.5574*	*2*	*20.7698*	*3.1673*
1	*4.4074*	*3*	*1.9807*	*0.4494*
2	*2.9624*	*8*	*25.3774*	*8.5665*
3	*1.9911*	*6*	*26.0712*	*8.0715*
4	*1.3383*	*1*	*0.1144*	*0.0855*
5	*2.7435*	*0*	*7.5267*	*2.7435*
				23.0839

Como sólo hay un parámetro estimado se compara con una estadística χ^2 con 4 grados de libertad. Una tabla de valores para χ^2 muestra que la probabilidad de obtener una medida de 23.0839 o mayor es menos de 0.001, con esto se encuentra una evidencia significativa de que el modelo $\widehat{S}(t)$ no es una buena representación de $S(t)$.

3.3.2. Prueba *Kolmogorov-Smirnov*

Esta prueba se realiza para modelos estimados mediante datos completamente conocidos, en este caso se desea comparar el modelo estimado $\widehat{S}(t)$ como una representación de $S(t)$ a partir de la distribución de supervivencia empírica $S^{\circ}(t)$. Para su aplicación se define la siguiente estadística de prueba:

$$D_n = \sup_t \left| \widehat{S}(t) - S^{\circ}(t) \right|, \tag{3.33}$$

conocida como *estadística de Kolmogorov-Smirnov.*

Esta medida representa la mayor desviación absoluta entre $\widehat{S}(t)$ y $S^{\circ}(t)$ sobre todo el dominio de t. El subíndice de D_n significa que esta medida depende del tamaño de la muestra n. Dado esto, la medida que será utilizada para efectos de la prueba es:

$$Y = \sqrt{n} \cdot D_n \tag{3.34}$$

Para esta prueba, normalmente la estadística *Kolmogorov-Smirnov* se plantea de la forma $D_n = \sup_t \left|\widehat{F}(t) - F^{\circ}(t)\right|$, pero sustituyendo $S(t)$ en (3.33) se tiene

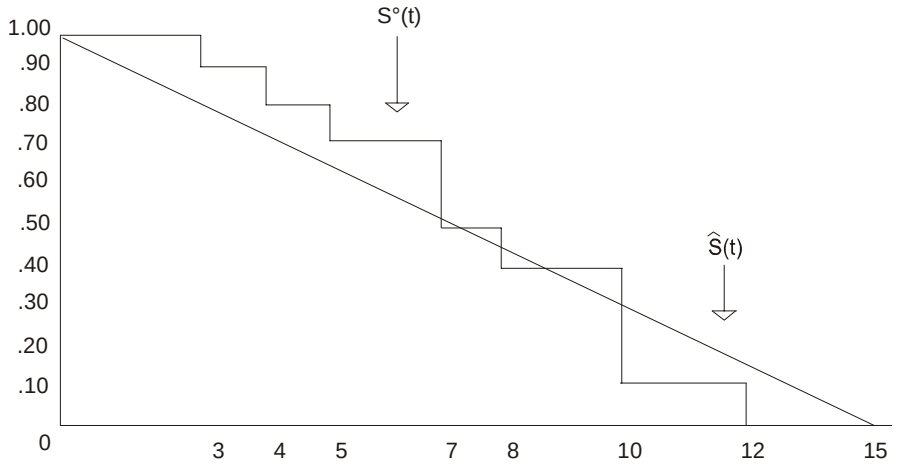
$$\begin{aligned} \left|1 - \widehat{F}(t) - (1 - F^{\circ}(t))\right| &= \left|-\widehat{F}(t) + F^{\circ}(t)\right| \\ &= \left|-\left(-\widehat{F}(t) + F^{\circ}(t)\right)\right| \\ &= \left|\widehat{F}(t) - F^{\circ}(t)\right| \end{aligned} \tag{3.35}$$

quedando la estadistica comúnmente planteada para esta prueba, y de esta manera las regiones de rechazo son las mismas.

La prueba de consiste en plantear la hipótesis nula $H_0 =$ el modelo ajusta a la distribucuón supuesta estimada vs $H_a =$ el modelo no ajusta correctamente a la distribución supuesta estimada. La regla de desición es rechazar H_0 con grado de significancia α si Y excede el valor en tablas de $Y^{(1-\alpha)}$, en otro caso aceptar H_0. (para mayores referencias, ver [3] pág. 293)

Ejemplo 3.7 *Se realiza una prueba de hipótesis con la estadística Kolmogorov-Smirnov suponiendo que a los datos del ejemplo 3.2 se ajusta un modelo estimado* $\widehat{S}(t) = 1 - \frac{t}{15}$. *La siguiente figura muestra el modelo de distribución*

empírica $S^{\circ}(t)$ comparado con el modelo supuesto para los datos $\widehat{S}(t)$.



La siguiente tabla compara los valores de $\widehat{S}(t)$ con los de $S^{\circ}(t)$, donde $S^{\circ}(t^{-})$ denota los valores de $S^{\circ}(t)$ justo antes de t y $S^{\circ}(t^{+})$ denota los valores de $S^{\circ}(t)$ justo después de t.

t	$\widehat{S}(t)$	$S^{\circ}(t^{-})$	$S^{\circ}(t^{+})$	$\mid \widehat{S}(t) - S^{\circ}(t) \mid$
3	*0.80000*	*1.00*	*0.90*	*0.20000*
4	*0.73333*	*0.90*	*0.80*	*0.16667*
5	*0.66667*	*0.80*	*0.70*	*0.13333*
7	*0.53333*	*0.70*	*0.50*	*0.16667*
8	*0.46667*	*0.50*	*0.40*	*0.06667*
10	*0.33333*	*0.40*	*0.10*	*0.23333*
12	*0.20000*	*0.10*	*0.00*	*0.20000*
15	*0.00000*	*0.00*	—	*0.00000*

De la tabla se observa que la máxima desviación es 0.23333, con lo cual se tiene el valor $y = (0.23333) \cdot \sqrt{10} = 0.73786$ y se busca la probabilidad $P(Y > 0.73786)$. Buscando este valor en tablas no se encuentra evidencia significativa en contra de la hipótesis del modelo estimado supuesto $\widehat{S}(t)$.

3.4. Modelos multivariados

Ahora se considera el caso general donde hay s variables concomitantes $z_1, z_2, z_3..., z_s$, todas ellas denotadas por el vector columna $\mathbf{z}$ y la función

de supervivencia denotada por $S(t;\mathbf{z})$. Al igual que como se definieron las funciones para el caso univariado, se tienen expresiones semejantes en el caso multivariado para cada una de ellas:

$$\lambda(t;\mathbf{z}) = -\frac{d}{dt}\ln S(t;\mathbf{z}), \tag{3.36}$$

(ver [8] pág. 210)

$$S(t;\mathbf{z}) = \exp\left(-\int_0^t \lambda(u;\mathbf{z})du\right), \tag{3.37}$$

(ver [8] pág. 210)
y

$$f(t;\mathbf{z}) = S(t;\mathbf{z})\cdot\lambda(t;\mathbf{z}). \tag{3.38}$$

(ver [8] pág. 210)

Cada individuo del grupo al cual se aplica $S(t;\mathbf{z})$ tiene su propia función de tasa de riesgo, donde los valores de $\mathbf{z}$ se aplican específicamente a cada individuo. Por ejemplo, supóngase que z_1 es la edad del individuo y que z_2 y z_3 son variables indicadoras definidas por:

$$z_2 = \begin{cases} 0 & \text{si es hombre} \\ 1 & \text{si es mujer} \end{cases}, \tag{3.39}$$

$$z_3 = \begin{cases} 0 & \text{si es alcohólico} \\ 1 & \text{si no es alcohólico} \end{cases}, \tag{3.40}$$

para este caso un individuo i de 30 años, no alcohólico, mujer tendría una función de tasa de riesgo $\lambda(t;\mathbf{z}_i)$ donde $\mathbf{z}_i' = [30,1,1]$.

Como en casos anteriores, $S(t;\mathbf{z})$ está en forma paramétrica y las variables t y de $\mathbf{z}$ dependen de ciertos parámetros. Al igual que se hizo para el caso univariado, se escoge una forma paramétrica para $S(t;\mathbf{z})$ y se estiman los valores para los parámetros a partir de la muestra dada usando el método de máxima verosimilutud.

Para la construcción de la función de verosimilitud se supone que la persona i entra en observación al tiempo r_i y sale de dicha observación al tiempo t_i ya sea vivo o muerto. La contribución a la función de verosimilitud del i-ésimo individuo vivo es:

$$L_i = \frac{f(t_i; \mathbf{z}_i)}{S(r_i; \mathbf{z}_i)}, \tag{3.41}$$

(ver [8] pág. 211)

si la observación termina por muerte, o es:

$$L_i = \frac{S(t_i; \mathbf{z}_i)}{S(r_i; \mathbf{z}_i)}, \tag{3.42}$$

(ver [8] pág. 211)

si la observación termina al tiempo t_i y el individuo continúa vivo. Ambos casos pueden ser incorporados escribiendo

$$L_i = \frac{S(t_i; \mathbf{z}_i) \cdot [\lambda(t_i; \mathbf{z}_i)]^{\delta_i}}{S(r_i; \mathbf{z}_i)}, \tag{3.43}$$

(ver [8] pág. 211)

donde δ_i es la variable indicadora definida en (3.29). La función de verosimilitud total queda dada por:

$$L = \prod_{i=1}^{n} [\lambda(t_i; \mathbf{z}_i)]^{\delta_i} \cdot \frac{S(t_i; \mathbf{z}_i)}{S(r_i; \mathbf{z}_i)}. \tag{3.44}$$

(ver [8] pág. 211)

En teoría, siempre se pueden estimar los parámetros en (3.44) maximizando la función L, pero la dificultad de la maximización dependerá de la complejidad de $\lambda(t; \mathbf{z})$ así como de $S(t; \mathbf{z})$.

3.4.1. Modelo multiplicativo

En esta forma de modelo el riesgo total $\lambda(t; \mathbf{z})$ está compuesto por un riesgo subyacente como función del tiempo $\lambda(t)$, y el riesgo adicional surge de las variables concomitantes multiplicadas juntas. El modelo más conocido de este grupo es el *Modelo de Cox* donde

$$\lambda(t; z_i) = \exp(a_j z_j), \tag{3.45}$$

(ver [7] pág. 251)

de donde el riesgo total es:

$$\lambda(t; z_i) = \lambda(t) \cdot \prod_{j=1}^{s} \lambda(t; z_i) = \lambda(t) \cdot \exp\left(\sum_{j=1}^{s} a_j z_j\right). \tag{3.46}$$

(ver [7] pág. 251)

Además, si el riesgo subyacente es constante durante todo el tiempo, se puede hacer $\lambda(t) = e^{a_0}$ obteniendo:

$$\lambda(t; z_i) = \exp\left(\sum_{j=0}^{s} a_j z_j\right) = e^{\mathbf{a'z}} \tag{3.47}$$

donde $\mathbf{a'} = [a_0, a_1, ..., a_s]$ y $\mathbf{z} = [z_0, z_1, ..., z_s]$. (ver [8] pág. 214)

Para estimar los parámetros a_i de (3.47) se construye la función de verosimilitud dada por (3.44). Como $\lambda(t; z_i) = e^{\mathbf{a'z}}$ es constante, se tiene:

$$\frac{S(t_i; \mathbf{z})}{S(r_i; \mathbf{z})} = \exp\left[-(t_i - r_i) \cdot e^{\mathbf{a'z}}\right], \tag{3.48}$$

(ver [8] pág. 215)
sustituyendo

$$L = \prod_{i=1}^{n} \left[e^{\mathbf{a'z}}\right]^{\delta_i} \cdot \exp\left[-(t_i - r_i) \cdot e^{\mathbf{a'z}}\right], \tag{3.49}$$

(ver [8] pág. 215)
resultando

$$\ln L = \sum_{i=1}^{n} \delta_i \cdot \mathbf{a'z} - \sum_{i=1}^{n} (t_i - r_i) \cdot e^{\mathbf{a'z}}, \tag{3.50}$$

haciendo $\frac{\partial \ln L}{\partial a_i} = 0$ se obtiene:

$$\sum_{i=1}^{n} \delta_i \cdot z_{ji} - \sum_{i=1}^{n} (t_i - r_i) \cdot e^{\mathbf{a'z}} = 0 \tag{3.51}$$

donde z_{ji} es la j-ésima variable concomitante del individuo i. Las $(s+1)$ ecuaciones resultantes $(j = 0, 1, ..., s)$ son resueltas simultáneamente para $\widehat{a}_0$, $\widehat{a}_1$, ..., $\widehat{a}_s$, que usualmente se hace por métodos numéricos.

En este trabajo no se adentra demasiado en estas formas de modelos, ya que esta fuera de los fines del mismo y sólo se hace referencia a estos para dar a conocer las diferentes maneras que existen para su planteamiento de acuerdo a los requerimientos del estudio.

3.5. Aplicaciones

En esta sección se describe un ejemplo de la aplicación de los modelos de supervivencia junto con la forma de estimar sus parámetros a partir de la muestra dada. Existen varias situaciones de importancia dentro del análisis económico y financiero para el desarrollo de los modelos de supervivencia, un caso de importancia económica será descrito mediante un ejemplo el cual es la situación de una huelga laboral.

La siguiente tabla muestra la duración de las huelgas reportadas en E. U. por las industrias de manufactura sobre el periodo de 1968 a 1976. Para lograr un grado de homogeneidad, sólo se cuentan como huelgas oficiales aquellas en las que participan 1000 o más trabajadores. Hay un total de 62 huelgas en la muestra y 12 fueron censuradas debido a que excedían el periodo de observación de 80 días. También hay varios empates en la muestra, con 4 huelgas que duraron tres días, otras cuatro que duraron también 4 días y así sucesivamente. (para mayores referencias, ver [5])

Para cada orden de duración j, sea t_j periodo de duración y d_j denota el número de huelgas que duraron exactamente t_j días.

Orden j	Duración t_j	d_j	Orden j	Duración t_j	d_j
1	1	1	21	26	1
2	2	4	22	27	1
3	3	4	23	28	1
4	4	1	24	29	1
5	5	1	25	32	1
6	7	1	26	33	1
7	8	1	27	36	1
8	9	2	28	37	1
9	10	1	29	38	1
10	11	1	30	41	1
11	12	2	31	42	1
12	13	1	32	43	2
13	14	1	33	44	1
14	15	1	34	49	2
15	17	1	35	52	2
16	19	1	36	61	1
17	21	2	37	72	1
18	22	1			
19	23	1			
20	25	1			

Tabla 3.1

Como se ha visto en secciones anteriores, los parámetros de la distribución pueden ser estimados por el método de máxima verosimilitud. Si se tiene una muestra completa, la función de verosimilitud es:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(t_j; \boldsymbol{\theta}), \tag{3.52}$$

donde $\boldsymbol{\theta}$ es un vector de uno o más parámetros y $f(t_j; \boldsymbol{\theta})$ denota la función de densidad de probabilidad del j-ésimo periodo de duración.

Como se está en una situación con datos censurados (12 huelgas censuradas que rebasaron el periodo de observación de 80 días) se hace uso de la función de verosimilitud modificada con la variable indicadora δ_j anteri-

ormente definida, como:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \left[f(t_j;\boldsymbol{\theta})\right]^{\delta_j} \cdot \left[S(t_j;\boldsymbol{\theta})\right]^{1-\delta_j}, \tag{3.53}$$

con lo cual la función log-verosimilitud es:

$$\ln L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \delta_j \cdot \ln\left[f(t_j;\boldsymbol{\theta})\right] + \sum_{j=1}^{n} \left[1-\delta_j\right] \cdot \ln\left[S(t_j;\boldsymbol{\theta})\right], \tag{3.54}$$

substituyendo $f(t_j;\boldsymbol{\theta}) = S(t_j;\boldsymbol{\theta}) \cdot \lambda(t_j;\boldsymbol{\theta})$ se obtiene:

$$\ln L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln\left[S(t_j;\boldsymbol{\theta})\right] + \sum_{j=1}^{n} \delta_j \cdot \ln\left[\lambda(t_j;\boldsymbol{\theta})\right]. \tag{3.55}$$

Si se selecciona una distribución exponencial para el modelo, se tiene el resultado obtenido en la sección 3.2.1 (ver Ejemplo 3.5) que es:

$$\widehat{\boldsymbol{\theta}} = \widehat{\lambda} = \frac{\sum_{j=1}^{n} \delta_j}{\sum_{j=1}^{n} t_j}, \tag{3.56}$$

donde $\sum_{j=1}^{n} \delta_j$ es el número observado de huelgas finalizadas y $\sum_{j=1}^{n} t_j$ es la exposición exacta de tiempo de las huelgas.

De los datos se calcula $\sum_{j=1}^{62} \delta_j = 50$ y $\sum_{j=1}^{62} t_j = 2118$, resultando $\widehat{\lambda} = \frac{50}{2118} = 0.0236$

(para mayores referencias acerca de todo el ejemplo, ver [5])

3.5.1. Uso de variables concomitantes

La distribución del periodo de duración de las huelgas puede estar sujeta a varias caracteristicas del grupo laboral en estudio, como pueden ser: (1) tamaño del grupo, (2) tipo de industria, (3) tiempo del año, (4) propuestas

en disputa y (5) estado actual de la economía. Todas estas variables concomitantes pueden ser representadas en un modelo parámetrico para ajustar mejor los datos. En el caso de los datos para las huelgas, las primeras cuatro influencias antes listadas pueden ser consideradas.

En esta sección se desarrolla el modelo paramétrico para incluir una variable concomitante que es un índice de la producción industrial. Se utiliza el modelo de Cox caracterizado por tener una tasa de riesgo constante $\lambda = e^{a_0}$ y la variable concomitante z_1 representa un valor fijo del índice de la producción industrial asociada con cada huelga. Con esto la función de riesgo es:

$$\lambda(t_j; a_0, a_1, z_1) = e^{a_0 + a_1 \cdot z_1}, \tag{3.57}$$

(ver [8] pág. 280)

con función de supervivencia

$$S(t_j; a_0, a_1, z_1) = e\left[-t_j \cdot e^{a_0 + a_1 \cdot z_1}\right], \tag{3.58}$$

(ver [8] pág. 280)

y de (3.55) se obtiene la función de log-verosimilitud que es:

$$\ln L(a_0, a_1) = -\sum -t_j \cdot e^{a_0 + a_1 \cdot z_1} + \sum \delta_j \left(a_0 + a_1 \cdot z_1\right) \tag{3.59}$$

(para mayores referencias acerca de todo el ejemplo, ver [5])

3.5.2. Alternativa paramétrica de modelos exponenciales

Como se ha visto en varios casos anteriores, la distribución exponencial es frecuentemente seleccionada para modelar el tiempo de falla de una variable aleatoria, en parte se debe a la propiedad de tener una función de riesgo constante.

Una extensión simple para el modelo exponencial es suponer que la tasa de riesgo se mantiene constante al nivel λ_1 hasta el tiempo y y que ésta cambia al valor λ_2 para un tiempo mayor de y. Esta forma se conoce como modelo exponencial por casos y se define por su función de tasa de riesgo que es:

$$\lambda(t) = \begin{cases} \lambda_1 & 0 \leq t < y \\ \lambda_2 & t \geq y \end{cases}, \tag{3.60}$$

la función de supervivencia queda como:

$$S(t) = \begin{cases} e^{-t\cdot\lambda_1} & 0 \leq t < y \\ e^{-y\cdot\lambda_1} \cdot e^{-(t-y)\lambda_2} & t \geq y \end{cases}, \tag{3.61}$$

y la función de densidad de probabilidad es:

$$f(t) = \begin{cases} \lambda_1 \cdot e^{-t\cdot\lambda_1} & 0 \leq t < y \\ \lambda_2 \cdot e^{-y\cdot\lambda_1} \cdot e^{-(t-y)\lambda_2} & t \geq y \end{cases}. \tag{3.62}$$

Usando esta nueva forma de modelo se puede estimar los valores para λ_1 y λ_2 por el método de máxima verosimilitud para los datos de la siguiente tabla,

Unidad	Resultado	Tiempo del resultado
A	Falla	$t_A < y$
B	Retirado	$t_B < y$
C	Falla	$t_C > y$
D	Retirado	$t_D > y$

De la sección 3.2.1 se sabe que la contribución a la función de verosimilitud para la i-ésima falla es $f(t_i)$ y la contribución para la j-ésima retirada es $S(t_j)$, con esto se obtiene:

$$\begin{aligned} L(\lambda_1, \lambda_2) &= (\lambda_1 \cdot e^{-t_A\cdot\lambda_1})(e^{-t_B\lambda_1}) \cdot \\ &\quad (\lambda_2 \cdot e^{-y\cdot\lambda_1} \cdot e^{-(t_C-y)\lambda_2})(e^{-y\cdot\lambda_1} \cdot e^{-(t_D-y)\lambda_2}) \end{aligned} \tag{3.63}$$

(ver [8] pág. 296)

la función de log-verosimilitud es:

$$\begin{aligned} \ln L(\lambda_1, \lambda_2) &= \ln\lambda_1 + \ln\lambda_2 - \lambda_1(t_A + t_B + 2y) - \lambda_2\left[(t_C - y) + (t_D - y)\right] \\ &= \ln\lambda_1 + \ln\lambda_2 - \lambda_1 \cdot E_1 - \lambda_1 \cdot E_2, \end{aligned} \tag{3.64}$$

(ver [8] pág. 296)

donde E_i denota la exposición exacta al riesgo λ_i, derivando e igualando a cero

$$\frac{\partial \ln L(\lambda_1, \lambda_2)}{\partial \lambda_i} = \frac{1}{\lambda_i} - E_i = 0 \tag{3.65}$$

resultando los valores estimados como

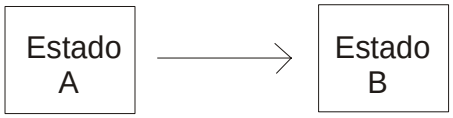
$$\widehat{\lambda}_i = \frac{1}{E_i}. \tag{3.66}$$

En el siguiente capítulo se presentan algunos modelos multiestados junto con sus propiedades y sus aplicaciiones dentro del ámbito del análisis de supervivencia.

Capítulo 4

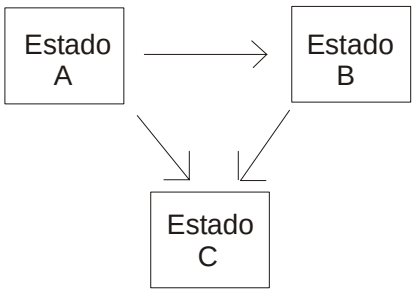
Modelos multiestados

Dentro del estudio de los modelos de supervivencia algunos de estos pueden ser vistos como modelos de dos o más estados, esto es en el sentido de que un individuo puede estar en cualquiera de los estados vivo o muerto. Suponiendo un modelo en el que el estado A denota estar vivo y el estado B denota estar muerto, este modelo sólo permite pasar del estado A al estado B. Con esto, la transferencia desde el estado A es sólo la muerte, entonces la tasa de riesgo para cualquier transferencia desde el estado A se llamada fuerza de mortalidad. En un caso más general, donde la transferencia desde un estado A hacia cualquier otro puede ser por diversas causas además de la muerte, se conoce como *fuerza de transferencia*.



En esta nueva forma de modelo existen varios ejemplos dentro del ámbito actuarial que van más allá del caso de solamente dos estados. Por ejemplo, sea A el estado en que un individuo se mantiene activo y puede pasar al estado B que es estar inhabilitado de acuerdo con una fueza de inhabilitación (fuerza de transferencia). Los individuos del estado B pueden morir y pasar al estado C de acuerdo con una fuerza de mortalidad, por supuesto también los individuos del estado A pueden pasar al estado C por una fuerza de mortalidad aplicable a los individuos del estado A. Con esto se plantea un modelo donde un individuo tiene la posibilidad de pasar del estado A a cualquiera de los dos estados B o C, y también de ser tranferido del estado

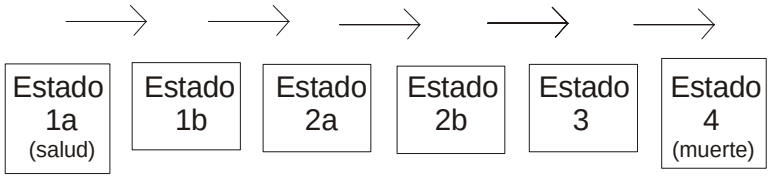
B al estado C. Se supone que todas las inhabilitaciones son permanentes, asi que no es posible pasar del estado A a B y regresar al estado A, y lógicamente una vez llegando al estado C no es posible regresar a ningún otro estado.



4.1. Modelo de progresión de una enfermedad

El primer modelo multiestados que se considera es un modelo progresivo de distintas etapas de una enfermedad como lo es el modelo del SIDA propuesto por Panjer.

En el modelo de Panjer se definen un total de seis estados. Primero se considera el estado en que el individuo no está infectado por el virus denotado por $1a$, si el individuo está infectado por el virus pero no presenta sintomas de la enfermedad se considera que está en el estado $1b$. Los siguientes tres estados denotados por $2a, 2b$ y 3 representan las etapas progresivas de la enfermedad, el estado final es la muerte denotado por 4. Se supone que los individuos progresan a través de las etapas de la enfermedad sin poder regresar a ninguna de éstas, también, por simplicidad, se supone que la muerte antes del estado 3 es un riesgo lo suficientemente menor como para ser ignorado. Lo principal es que un individuo en cualquier estado sólo puede pasar hacia el siguiente estado en el modelo.



Entonces el único decremento está dado por la progresión de los individuos hacia el siguiente estado, con esto el modelo queda completamente definido

por las tasas de riesgo (fuerza de progresión) operando en cada caso. Un supuesto de simplificación es que estas fuerzas de progresión son constantes dentro de cada estado, denotadas por μ_j para $j = 1a, 1b, 2a, 2b$,y 3, con lo cual no dependen de la edad o sexo del individuo.

4.1.1. Propiedades del modelo

Si la fuerza de progresión es constante, entonces la variable aleatoria T_j que mide el tiempo de espera en el estado j tiene distribución exponencial con:

$$S_{T_j}(t) = P(T_j > t) = e^{-t \cdot \mu_j} \tag{4.1}$$

$$F_{T_j}(t) = P(T_j \leq t) = 1 - e^{-t \cdot \mu_j} \tag{4.2}$$

$$f_{T_j}(t) = \frac{d}{dt} F_{T_j}(t) = \mu_j \cdot e^{-t \cdot \mu_j} \tag{4.3}$$

La variable aleatoria T_j es independiente de las demás variables aleatorias asociadas con los otros estados, es decir, la progresión de un estado hacia el siguiente no depende del tiempo que el individuo haya estado en el estado previo. En el modelo se supone que un individuo sólo puede dejar el estado 3 por la muerte, con lo cual μ_3 es la fuerza de mortalidad para un individuo con SIDA y T_3 es el tiempo de vida de dicho individuo. Entonces la variable aleatoria definida por:

$$R = T_{2b} + T_3 \tag{4.4}$$

denota el tiempo de vida de un individuo en el estado $2b$, con un valor esperado

$$E(R) = E(T_{2b}) + E(T_3) = \frac{1}{\mu_{2b}} + \frac{1}{\mu_3}. \tag{4.5}$$

De igual forma se pueden definir otras variables aleatorias como suma de las variables aleatorias básicas.

Ejemplo 4.1 *Dadas las suposiciones del modelo, se busca la probabilidad de que un individuo que en este momento está en el estado i, se encuentre en el siguiente estado en t años a partir de ahora.*

Considere un individuo que se mueve del estado i al estado $i+1$ en un tiempo r, donde $0 < r < t$ y éste permanece en el estado $i+1$ hasta el tiempo t. La probabilidad (diferencial) de este evento es $e^{-r\cdot\mu_i} \cdot \mu_i \cdot e^{-(t-r)\cdot\mu_{i+1}} dr$, entonces la probabilidad total es:

$$\begin{aligned} P_{i,i+1}(t) &= \int_0^t e^{-r\cdot\mu_i} \cdot \mu_i \cdot e^{-(t-r)\cdot\mu_{i+1}} dr \\ &= \mu_i \cdot e^{-t\cdot\mu_{i+1}} \int_0^t e^{r\cdot(\mu_{i+1}-\mu_i)} dr \\ &= \mu_i \cdot e^{-t\cdot\mu_{i+1}} [\frac{e^{-t\cdot(\mu_{i+1}-\mu_i)} - 1}{\mu_{i+1} - \mu_i}] \\ &= \frac{\mu_i(e^{-t\cdot\mu_i} - e^{-t\cdot\mu_{i+1}})}{\mu_{i+1} - \mu_i} \\ &= \mu_i(\frac{e^{-t\cdot\mu_i}}{\mu_{i+1} - \mu_i} + \frac{e^{-t\cdot\mu_{i+1}}}{\mu_i - \mu_{i+1}}) \end{aligned} \tag{4.6}$$

(ver [8] pág. 251)

Para el caso especial donde el estado $i+1$ es la muerte, desde este no existen más posibles progresiones, entonces $\mu_{i+1} = 0$ y el resultado del ejemplo anterior se reduce a $1 - e^{-t\cdot\mu_i}$, que es la probabilidad de morir dentro de t años cuando se está en el estado i.

El caso general para el ejemplo anterior puede ser extendido para encontrar las probabilidades de progresión desde cualquier estado hacia cualquier otro posterior y permanecer en ese estado en el tiempo t. Por ejemplo, considérese la probabilidad de moverse desde el estado i, a través del estado j, hacia el estado k y estar en éste en el tiempo t. Si la transferencia del estado i al estado j ocurre en el tiempo $r, 0 < r < t$, y la transferencia del estado j al estado k ocurre al tiempo $s, r < s < t$, entonces la probabilidad deseada está dada por:

$$
\begin{aligned}
P_{i,j,k}(t) &= \int_0^t \int_r^t e^{-r\cdot\mu_i}\cdot\mu_i\cdot e^{-(s-r)\cdot\mu_j}\cdot\mu_j\cdot e^{-(t-s)\cdot\mu_k} ds dr \\
&= \mu_i\cdot\mu_j\cdot e^{-t\cdot\mu_k}\int_0^t e^{r\cdot(\mu_j-\mu_i)}\int_r^t e^{s(\mu_k-\mu_j)} ds dr \\
&= \mu_i\cdot\mu_j\cdot e^{-t\cdot\mu_k}\int_0^t e^{r\cdot(\mu_j-\mu_i)}\left(\frac{e^{t(\mu_k-\mu_j)}-e^{r(\mu_k-\mu_j)}}{\mu_k-\mu_j}\right) dr \\
&= \frac{\mu_i\cdot\mu_j\cdot e^{-t\cdot\mu_k}}{\mu_k-\mu_j}\int_0^t \left(e^{t\cdot(\mu_k-\mu_j)}\cdot e^{r(\mu_j-\mu_i)}-e^{r(\mu_k-\mu_i)}\right) dr \\
&= \frac{\mu_i\cdot\mu_j\cdot e^{-t\cdot\mu_k}}{\mu_k-\mu_j}\left[e^{t\cdot(\mu_k-\mu_j)}\left[\frac{e^{t\cdot(\mu_j-\mu_i)}-1}{\mu_j-\mu_i}\right]-\frac{e^{t\cdot(\mu_k-\mu_i)}-1}{\mu_k-\mu_i}\right] \\
&= \frac{\mu_i\cdot\mu_j}{\mu_k-\mu_j}\left[\frac{e^{-t\cdot\mu_i}-e^{-t\cdot\mu j}}{\mu_j-\mu_i}-\frac{e^{-t\cdot\mu_i}-e^{-t\cdot\mu_k}}{\mu_k-\mu_i}\right] \\
&= \frac{\mu_i\cdot\mu_j}{\mu_k-\mu_j}\left[\frac{e^{-t\cdot\mu_i}(\mu_k-\mu_j)}{(\mu_j-\mu_i)(\mu_k-\mu_i)}-\frac{e^{-t\cdot\mu_j}}{\mu_j-\mu_i}+\frac{e^{-t\cdot\mu_k}}{\mu_k-\mu_i}\right] \\
&= \mu_i\cdot\mu_j\left[\frac{e^{-t\cdot\mu_i}}{(\mu_j-\mu_i)(\mu_k-\mu_i)}+\frac{e^{-t\cdot\mu_j}}{(\mu_i-\mu_j)(\mu_k-\mu_j)}+\right. \\
&\quad \left.\frac{e^{-t\cdot\mu_k}}{(\mu_i-\mu_k)(\mu_j-\mu_k)}\right]. \qquad (4.7)
\end{aligned}
$$

(ver [8] pág. 253)

Ahora, se trata de encontrar la distribución exacta de la variable aleatoria $R = T_{2b} + T_3$, que representa el tiempo de vida de un individuo que se encuentra en el estado $2b$. La función de distribución de probabilidad se denota por $F_R = F_{T_{2b}} + F_{T_3}$ y está dada por (4.7), con $i = 2b$, $j = 3$ y $k = 4$, como el estado 4 es la muerte, se tiene que $\mu_4 = 0$.

$$
\begin{aligned}
F_R &= \frac{\mu_3 \cdot e^{-t\cdot\mu_{2b}}}{-(\mu_3-\mu_{2b})} + \frac{\mu_{2b}\cdot e^{-t\cdot\mu_3}}{-(\mu_{2b}-\mu_3)} + 1 \\
&= \frac{\mu_3 - \mu_{2b} + \mu_{2b}\cdot e^{-t\cdot\mu_3} - \mu_3 \cdot e^{-t\cdot\mu_{2b}}}{(\mu_3-\mu_{2b})} \\
&= \frac{\mu_3(1-e^{-t\cdot\mu_{2b}})}{(\mu_3-\mu_{2b})} + \frac{\mu_{2b}(1-e^{-t\cdot\mu_3})}{(\mu_{2b}-\mu_3)} \\
&= \frac{\mu_3}{\mu_3-\mu_{2b}} \cdot F_{T_{2b}}(t) + \frac{\mu_{2b}}{\mu_{2b}-\mu_3}\cdot F_{T_3}(t) \qquad (4.8)
\end{aligned}
$$

Con esto, la función de supervivencia de R es:

$$
\begin{aligned}
S_R(t) &= 1 - F_R \\
&= \frac{\mu_3}{\mu_3-\mu_{2b}} \cdot S_{T_{2b}}(t) + \frac{\mu_{2b}}{\mu_{2b}-\mu_3}\cdot S_{T_3}(t) \qquad (4.9)
\end{aligned}
$$

y la función de densidad de probabilidad es:

$$
\begin{aligned}
f_R(t) &= \frac{d}{dt}F_R \\
&= \frac{\mu_3}{\mu_3-\mu_{2b}} \cdot f_{T_{2b}}(t) + \frac{\mu_{2b}}{\mu_{2b}-\mu_3}\cdot f_{T_3}(t) \qquad (4.10)
\end{aligned}
$$

4.1.2. Estimación de los parámetros del modelo

Dada la suposición de una fuerza constante de progresión, el modelo de Panjer es viable para estimar los valores de μ_j , $j = 1a, 1b, 2a, 2b, 3$, mediante algunos métodos como se ha visto en temas anteriores, teniendo en cuenta que se tiene una muestra de los tiempos exactos en los que los individuos bajo estudio progresan a través de cada estado planteado en el modelo. Con esto, se puede usar el método de máxima verosimilitud para estimar cada fuerza de progresión constante, así como el número de individuos que progresan hacia el siguiente estado, donde los parámetros μ_j son estimados de forma separada dependiendo de la muestra. A pesar de esto, en la práctica por lo regular no se tiene una información tan completa y detallada de los tiempos de muerte de todos los individuos en estudio, así que muchas veces por conveniencia se

subdividió el periodo de observación en k grupos, de donde se tendrán los datos a partir de la muestra original para observar la forma en que progresan los individuos a través de los distintos estados del modelo.

Para cada j-ésimo estado inicial se toma en cuenta el número total de individuos en cada etapa denotado por n_i, y el número de los que progresan hacia el siguiente estado durante la observación, denotado por d_i, donde n_i y d_i se refieren al i-ésimo grupo en tiempo bajo observación. Entonces, $n_i - d_i$ denota el número de individuos que permanecieron en el grupo inicial durante el periodo de observación.

Sea p_i la probabilidad de permanecer en la etapa inicial y $q_i = 1 - p_i$ la probabilidad de progresar hacia el siguiente estado del modelo. De esta forma se tiene un modelo binomial y la verosimilitud para estimar μ_j es:

$$L_i(\mu_j) = (1 - p_i)^{d_i} \cdot (p_i)^{n_i - d_i} \tag{4.11}$$

(ver [8] pág. 256)

para el i-ésimo grupo en tiempo bajo observación, y

$$L(\mu_j) = \prod_{i=1}^{k} (1 - p_i)^{d_i} \cdot (p_i)^{n_i - d_i} \tag{4.12}$$

(ver [8] pág. 256)

dado que hay k grupos totales bajo observación.

La función log-verosimilitud es

$$\ln(L) = \sum_{i=1}^{k} d_i \cdot \ln(1 - p_i) + (n_i - d_i) \cdot \ln p_i \tag{4.13}$$

(ver [8] pág. 256)

resultando la ecuación

$$\frac{\partial \ln(L)}{\partial \mu_j} = \sum_{i=1}^{k} - \frac{\partial p_i}{\partial \mu_j} \left[\frac{d_i}{1 - p_i} - \frac{n_i - d_i}{p_i} \right] = 0 \tag{4.14}$$

que puede resolverse para $\widehat{\mu_j}$ por iteración.

Par realizar las estimaciones de $\widehat{\mu_j}$, es claro que se debe de expresar p_i en téminos de μ_j. Si μ_j es supuesta como constante, entonces $p_i = e^{-r_i \cdot \mu_j}$,

donde r_i denota el periodo de tiempo que permanece en la j-ésima etapa el individuo del i-ésimo grupo bajo observación.

4.2. Modelo de cuidado continuo de retiro en comunidades

Este es un modelo multiestados más complejo que el presentado en la sección anterior, donde la conveniencia del modelo radicaba en que los individuos solamente podían progresar de un estado a otro sin poder regresar a través de ellos. En este modelo se analiza el patrón de supervivencia de los residentes de una comunidad de cuidado continuo en retiro (CCCR). Este modelo fue propuesto por B. L. Jones.

Este modelo es representado por cuatro estados dados por:

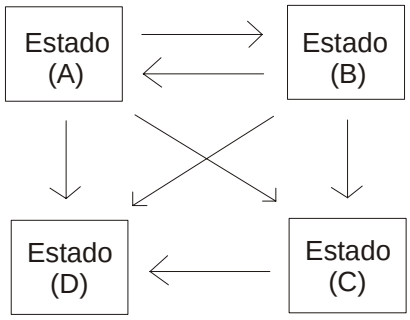
1 Estado (A): Residente como unidad de vida independiente.

2 Estado (B): Residencia temporal con servicio de enfermería.

3 Estado (C): Residencia permanente con servicio de enfermería.

4 Estado (D): Salida de CCCR, por muerte o salida voluntaria.

Las personas en el estado (A) pueden ser transferidas a cualquiera de los estados (B), (C) o (D). Las personas en el estado (B) pueden ser transferidas de vuelta al estado (A) o hacia los estados (C) o (D). Las personas en el estado (C) sólo pueden transferirse al estado (D), sin poder regresar a ningún estado anterior. Finalmente, las personas en el estado (D) han salido del CCCR y por lo tanto no se transfieren hacia ningún otro estado. La siguiente figura ilustra la forma del modelo.



4.2.1. Probabilidades de transición

Al igual que como se vio en el modelo de Panjer, se supone que existe una fuerza de transición constante del estado (i) al estado (j), denotada por μ_{ij}. Por la forma del modelo, se puede ver que algunas de las probabilidades de transición son cero (como μ_{ca} y μ_{cb}) y que las personas de algunos estados están sujetas a más de una fuerza de transición (como en el estado (A) y (B)). En el caso de una persona en el estado (A) está expuesta a un decremento triple, en particular, la fuerza de transición total para una persona en el estado (A) está dada por:

$$\mu_a^* = \mu_{ab} + \mu_{ac} + \mu_{ad} \tag{4.15}$$

Como la fuerza de transición total es constante, la variable aleatoria T_a que denota el periodo de tiempo de estancia en el estado (A) tiene una distribución exponencial. Entonces la función de supervivencia para una persona en el estado (A) al tiempo $t = 0$ está dada por:

$$S_a(t) = e^{-t \cdot \mu_a^*}. \tag{4.16}$$

El valor esperado del periodo de tiempo de estancia en el estado (A) está dado por:

$$E(T_a) = \frac{1}{\mu_a^*}, \tag{4.17}$$

y la varianza es:

$$Var(T_a) = \left(\frac{1}{\mu_a^*}\right)^2. \tag{4.18}$$

Se aplican las mismas observaciones para la variable aleatoria T_b, que denota el periodo de tiempo de estancia en el estado (B), la cual tiene una fuerza de transición total dada por:

$$\mu_b^* = \mu_{ba} + \mu_{bc} + \mu_{bd} \tag{4.19}$$

Las personas en el estado (C) están sujetas a una sola fuerza de transición hacia el estado (D), con lo cual, es un decremento simple. La variable aleatoria T_c tiene distribución exponencial con función de supervivencia

$$S_c(t) = e^{-t \cdot \mu_c^*} \tag{4.20}$$

Debido a la propiedad de pérdidad de memoria de la distribución exponencial, las probabilidades de supervivencia en cada estado son independientes del tiempo que haya permanecido la persona en ese estado. La suposición de una fuerza de transición constante con su consecuente distribición exponencial es claramente cuestionable, estas suposiciones se han hecho para la simplificación del análisis.

Ejemplo 4.2 *Para una persona en el estado (A) se busca la expresiones para:*

a) *La probabilidad de que no deje el estado (A) durante el siguiente año.*

b) *La probabilidad de que deje el estado (A) sólo por la tranferencia hacia el estado (C) durante el siguiente año.*

c) *La probabilidad de que deje el estado (A) por la tranferencia hacia el estado (B) durante el siguiente año y continúe en ese estado hasta el término del año.*

a) *La probabilidad está dada directamente por* $S_a(1) = e^{-\mu_a^*}$

b) *La fuerza de transferencia hacia el estado (C) se denota por* μ_{ac} *y la probabilidad deseada está dada por:*

$$\int_0^1 e^{-t\cdot\mu_a^*}\mu_{ac}dt = \frac{\mu_{ac}}{\mu_a^*}(1-e^{-\mu_a^*}). \tag{4.21}$$

c) *Aquí el evento es transferirse al estado (B) y permanecer en él hasta el tiempo* $t = 1$*, la probabilidad del evento está dada por:*

$$\begin{aligned}\int_0^1 e^{-t\cdot\mu_a^*}\mu_{ab}\cdot e^{-(1-t)\mu_b^*}dt &= \mu_{ab}\cdot e^{-\mu_b^*}\int_0^1 e^{t(\mu_b^*-\mu_a^*)}dt \\ &= \mu_{ab}\cdot e^{-\mu_b^*}\left(\frac{e^{t(\mu_b^*-\mu_a^*)}-1}{\mu_b^*-\mu_a^*}\right) \\ &= \mu_{ab}\left(\frac{e^{-\mu_a^*}-e^{-\mu_b^*}}{\mu_b^*-\mu_a^*}\right) \\ &= \mu_{ab}\left(\frac{e^{-\mu_a^*}}{\mu_b^*-\mu_a^*}+\frac{e^{-\mu_b^*}}{\mu_a^*-\mu_b^*}\right). \end{aligned} \tag{4.22}$$

Comentarios finales

En este trabajo se dio una perspectiva inicial del análisis y modelos de supervivencia, comentando sus formas, estimaciones y aplicaciones junto con las perspectivas actuariales en cuanto a este tipo de modelos.

En lo que concierne a las matemáticas de los modelos de supervivencia, dio una introducción sobre las bases en las cuales se desarrollan los modelos de supervivencia, incluyendo las formas actuariales para su aplicación. Se mostraron cuales son las principales distribuciones de probabilidad que se usan para el desarrollo del análisis junto con una breve explicación de cual es su utilidad. Esto es importante, ya que a partir de estas distribuciones se obtienen las propiedades posteriores de cada modelo, lo cual determina lo práctico de su uso y aplicación.

En cuanto a la estimación paramétrica, se dio un énfasis en los métodos más comúnmente usados que son máxima verosimilitud y mínimos cuadrados, también se hizo mención de otros métodos menos comunes pero útiles dependiendo el tipo de análisis que se requiera. Dada la complejidad de los modelos multivariados, no se abundó demasiado en lo relacionado a éstos y solo se consideró hacer una pequeña mención de sus formas para tener un panorama más amplio del alcance y entorno de los modelos de supervivencia.

Finalizando, en el capítulo de los modelos multiestados se mostraron solo unos modelos de forma muy particular sin adentrarse de manera general en modelos de este tipo. Para tener una perspectiva más amplía de este tipo de modelos se requeriría un trabajo mucho más extenso para poder desarrollarse de forma más minuciosa, pero esto queda fuera de los objetivos del presente trabajo.

En este documento se desarrollaron los temas previstos de una forma básica y sencilla, en los ejemplos se muestra como con las herramientas básicas de estadística y probabilidad, se pueden obtener y aplicar los modelos de supervivencia paramétricos.

Este trabajo cumplió con sus objetivos, ya que lo principal no era abundar demasiado en los temas sino dar una introducción a las formas y aplicaciones generales de los modelos de supervivencia.

Bibliografía

[1] Bowers Newton, L. (1997). *Actuarial Mathematics*. The Society of Actuaries, Schaumburg Illinois.

[2] Collet, David (1994). *Modelling Survival Data in Medical Research*. Chapman and Hall. Text in Statistical Sciences.

[3] Conover, W. J. (1971). *Practical Nonparametric Statistics*. John Wiley &Sons.

[4] Cox, D. R. and Oakes D. (1984). *Analysis of Survival Data*. Chapman and Hall. London, New York.

[5] Kennan, J. (1985). "The Duration of Contract Strikes in U.S. Manufacturing", *Journal of Econometrics*. 28(1).

[6] Lawless J. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley &Sons.

[7] Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. Second Edition. John Wiley

[8] London, Dick (1997). *Survival Models and their Estimation*. Third Edition. ACTEX Publications.