Fiabilidad y Análisis de Supervivencia

Jesús Abaurrea y Ana Carmen Cebrián
Dpto. Métodos Estadísticos. Universidad de Zaragoza

Índice General

Capítulo 1

Introducción a la Fiabilidad

1.1 Definición y objetivos

El término Fiabilidad (*reliability*) se usa generalmente para expresar la capacidad de un elemento para funcionar satisfactoriamente bajo ciertas condiciones ambientales durante un determinado periodo de tiempo. El elemento puede ser un componente o un sistema, formado por un conjunto de dos o más componentes organizados para realizar una determinada función. El concepto de Fiabilidad no hace referencia a la capacidad para realizar una función en un instante preciso, sino que está asociado al comportamiento que cabe esperar a lo largo del tiempo.

El objetivo de los modelos matemáticos desarrollados en Fiabilidad es describir la frecuencia de fallo de un elemento, así como resolver problemas de optimización relativos al funcionamiento y utilización de los mismos. La Fiabilidad se interesa también por el análisis y diseño de políticas de mantenimiento preventivo, de inspección y de reparación que permitan lograr un funcionamiento óptimo y minimizar costos.

El concepto de Fiabilidad comprende un gran número de facetas y problemas que corresponden a disciplinas distintas y requieren métodos de análisis diferentes; en particular, se puede distinguir una rama de carácter estadístico, cuyo objetivo es caracterizar el tiempo de fallo de una población, y otra, que suele incluirse dentro de la Investigación Operativa, que se interesa por la gestión óptima y el mantenimiento de sistemas.

Existen otras divisiones posibles, por ejemplo, se puede distinguir entre sistemas **reparables** y **no reparables**. Un sistema reparable es aquél que cuando falla puede repararse sustituyendo alguno de sus componentes. En el análisis de componentes no reparables se dispone de las observaciones del tiempo de fallo correspondientes a varios componentes del mismo tipo, y la hipótesis de independencia e idéntica distribución, i.i.d., suele ser habitual. En los sistemas reparables sin embargo, se utilizan normalmente medidas sucesivas realizadas sobre el mismo sistema, y la hipótesis de i.i.d. acerca de la duración de los intervalos de funcionamiento puede no ser plausible. En consecuencia, el análisis de sistemas reparables requiere técnicas diferentes a las utilizadas en sistemas no reparables, que se basan en muestras aleatorias.

En los sistemas complejos, el análisis individual de sus componentes -Fiabilidad de componentes- no suele ser suficiente para determinar la fiabilidad del sistema, y es necesario utilizar técnicas que analicen la configuración o estructura conjunta -Fiabilidad de sistemas-.

Es importante señalar la analogía de los problemas estadísticos de la Fiabilidad de componentes, con los de otras disciplinas, especialmente en el campo biomédico. En efecto, los datos correspondientes a la variable respuesta, son los tiempos transcurridos desde un instante inicial definido hasta la observación de un suceso específico que llamamos fallo. De este mismo tipo son los datos obtenidos en determinados estudios médicos -ensayos sobre tratamientos, pruebas con animales de laboratorio, etc.- de forma que los métodos que describiremos son también aplicables en este campo. El análisis estadístico de los datos obtenidos en este tipo de estudios biomédicos se denomina **Análisis de Supervivencia** (*survival analysis*), mientras que el término Fiabilidad se utiliza en el ámbito industrial y tecnológico. Aunque existen peculiaridades y problemas específicos de cada materia, la metodología estadística de ambas es, básicamente, la misma.

1.2 Contenido

Dada la amplitud del tema, nos limitaremos a estudiar los métodos estadísticos para analizar problemas relativos al tiempo de fallo de componentes simples no reparables y no trataremos temas relativos a la fiabilidad de configuraciones.

- Introducción. Tipos de datos y problemas en Fiabilidad. Similitud con los problemas de Análisis de Supervivencia. Mecanismos de censura.

- Revisión de conceptos probabilísticos básicos en Fiabilidad. Formas alternativas de caracterizar la distribución del tiempo de vida. Distribuciones más importantes: Exponencial, Weibull, Lognormal, Loglogística, Gumbel y Gamma. Características de estas distribuciones.

- Inferencia estadística no paramétrica a partir de una muestra. Estimación de las características de la distribución mediante tablas de vida clínicas. Estimador Kaplan-Meier de la función de supervivencia.

- Inferencia estadística paramétrica. Revisión del método de máxima verosimilitud y sus propiedades. Verosimilitud en muestras censuradas. Inferencia paramétrica con los modelos Exponencial y Weibull. Otros modelos paramétricos. Métodos gráficos para seleccionar un modelo adecuado a los datos.

- Métodos no paramétricos para comparar la supervivencia. Análisis de dos muestras: test de Mantel-Haenszel (*log-rank*) y test de Wilcoxon. Análisis estratificado. Comparación de tres o más muestras. Generalización del test de Mantel-Haenszel.

- Análisis de supervivencia con covariables. Modelo de riesgo proporcional de Cox. Modelo de tiempo de fallo acelerado.

1.3 Aspectos básicos de la Fiabilidad estadística

1.3.1 Obtención de los datos

El objetivo principal de la Fiabilidad estadística es el análisis del **tiempo de funcionamiento** de un elemento o individuo. Esta variable también se denomina, **tiempo hasta el fallo, tiempo de respuesta, tiempo de vida** o **tiempo de supervivencia**. En cada aplicación debe especificarse cómo se define y cuáles son los instantes que marcan sus límites. En muchos casos, el tiempo real transcurrido puede ser una buena medida, sin embargo, es frecuente utilizar el tiempo operativo u otra cantidad no negativa que resulte adecuada. Por ejemplo, para determinar

la fiabilidad de un vehículo es más útil analizar el kilometraje realizado entre dos averías que el tiempo transcurrido entre ambas y resulta más fácil de medir que el tiempo de utilización real del vehículo en ese intervalo.

Normalmente, las observaciones de la variable tiempo hasta el fallo se obtienen mediante la realización de ensayos. En los estudios de carácter industrial, para establecer las características del funcionamiento normal de un componente, se realizan ensayos en el laboratorio donde se miden los tiempos de fallo en muestras de esas piezas. En ocasiones, el ensayo se realiza en condiciones más exigentes que las de utilización normal, ensayos acelerados, a fin de abreviar su duración y disminuir costos.

La realización de ensayos clínicos, en los que el objetivo usual es comparar la experiencia de supervivencia de dos o más grupos de pacientes sometidos a tratamientos distintos, resulta más complicada que la de ensayos industriales, dada la mayor dificultad en definir con exactitud y seleccionar los tratamientos que se van a comparar, el tipo de pacientes y los métodos para evaluar la respuesta al tratamiento de cada uno de ellos.

Para garantizar la validez y poder extrapolar los resultados de un estudio clínico es necesario que los pacientes incluidos en el estudio sean representativos de la población objetivo. Además, para evitar la aparición de sesgos en los resultados se deben analizar grupos de pacientes que sean homogéneos, de forma que la única diferencia entre ellos sea el tratamiento, y la comparación no se vea afectada por otros factores. La forma más sencilla de evitar estos problemas es la asignación aleatoria del tratamiento a cada paciente. Además, si existe algún factor que pueda influir en la respuesta, debe efectuarse una **aleatorización estratificada**, es decir una asignación aleatoria por grupos o estratos.

Otra característica habitual de los ensayos clínicos es que, debido a la dificultad para disponer al comienzo del estudio de un número suficiente de pacientes que satisfagan todos los requisitos, la incorporación de éstos se produce de forma escalonada, en distintos instantes de tiempo, a lo largo de una primera fase del estudio. Los ensayos industriales, por lo general, no suelen presentar este problema.

1.3.2 Características diferenciales de los problemas

En los problemas de Fiabilidad, la variable respuesta, T, es una variable no negativa que, frecuentemente, tiene una distribución bastante asimétrica. Por este motivo, la distribución Normal no jugará aquí el papel preponderante que tiene en otros campos de aplicación de la Estadística; su papel de referencia pasarán a ocuparlo otras distribuciones como la Exponencial o la distribución Weibull. Este hecho, sin embargo, no es la característica diferencial de la Fiabilidad, ya que, una transformación adecuada de la variable T, por ejemplo, $ln(T)$ ó T^{-1}, o la utilización de herramientas no paramétricas, podrían resolver razonablemente esos problemas.

El rasgo específico del análisis estadístico en estos campos es la necesidad de realizar inferencia a partir de muestras en las que, junto con observaciones de la variable, aparecen observaciones incompletas, parciales o **censuradas**. La obtención de muestras completas suele requerir ensayos demasiado largos por lo que es habitual terminar los experimentos cuando se ha observado un determinado porcentaje de fallos o diseñarlos con un horizonte temporal prefijado, de forma que frecuentemente no se observa el fallo de un número importante de elementos de la muestra.

En los ensayos industriales la limitación de la duración del ensayo suele estar impuesta por razones económicas, mientras que en los ensayos clínicos la censura aparece debido al diseño de los mismos; en estos estudios los pacientes suelen incorporarse en diferentes instantes y es probable no observar el fallo de aquellos que se han incorporado más tarde. Además, es frecuente que durante el desarrollo de la prueba algunos individuos abandonen el estudio por lo que tampoco es posible conocer su tiempo de fallo exacto, sino sólo que éste no se había observado hasta el momento del abandono. Esta información parcial debe incorporarse al análisis, ya que, si las observaciones censuradas se desechan, o si se toman por observaciones auténticas, las estimaciones sobre T pueden resultar sesgadas e ineficientes.

1.4 Algunos ejemplos de problemas de Fiabilidad

En este apartado presentamos algunos ejemplos de estudios de Fiabilidad y Análisis de Supervivencia, con el fin de ilustrar los aspectos comentados

a 26 kV	a 30 kV	a 32 kV		a 34 kV		a 36 kV		a 38 kV
5.79	7.74	0.27	0.40	0.19	0.7	0.35	0.59	0.09
1579.52	17.05	0.69	0.79	0.96	1.31	0.96	0.99	0.39
2323.70	20.46	2.75	3.91	2.78	3.16	1.69	1.97	0.47
	21.02	9.88	13.95	4.15	4.67	2.07	2.58	0.73
a 28 kV	22.66	15.93	27.80	4.85	6.50	2.71	2.90	0.74
68.85	43.40	53.24	82.85	7.35	8.801	3.67	3.99	1.13
108.29	47.30	89.29	100.58	8.27	12.06	5.35	13.77	1.40
110.29	139.07	215.10		31.75	32.52	25.50		2.38
426.07	144.12			33.91	36.71			
1067.60	175.88			72.89				
	194.90							

Tabla 1.1: Datos de Nelson (ejemplo 1).

Ejemplo 1: Nelson presenta los resultados de un ensayo realizado para determinar el comportamiento de un fluido aislante. En el experimento se colocaron elementos fabricados con ese material entre dos electrodos. Los componentes aislantes fueron sometidos a un voltaje constante hasta que se estropearon. El material fue probado a siete voltajes distintos, entre 26 y 38 kilovoltios. En la tabla 1.1 se muestra el tiempo, en minutos, que cada elemento resistió hasta el fallo. El experimento anterior se prolongó el tiempo necesario para observar el fallo de todos los elementos aislantes sometidos a prueba, por lo que nos encontramos con una muestra aleatoria completa que se puede analizar con los métodos de inferencia habituales. Notemos cómo se alarga la duración del ensayo para observar el fallo de sólo tres componentes de la muestra.

El principal objetivo de este experimento era investigar la distribución del tiempo hasta el fallo del material aislante y la relación entre los parámetros de la distribución de T y el voltaje experimentado por el material. Una vez establecido un modelo para la relación tiempo de fallo-voltaje aplicado, interesará estimar la fiabilidad del material cuando funciona sometido a las condiciones de diseño. El percentil décimo de la distribución es una medida muy utilizada por los ingenieros para reflejar la fiabilidad de un elemento.

Ejemplo 2: Bartholomew plantea un ensayo en el que unidades de cierto componente de un equipo se van instalando cuando se produce un fallo de ese componente en alguno de los equipos disponibles. La información sobre los tiempos de vida y las sustituciones realizadas de 10 de esas piezas se muestra en la tabla 1.2. El

n^o de pieza	fecha instalación	fecha de fallo	tiempo de vida
1	11 junio	13 junio	2 días
2	21 junio	-	$>$ 72 días
3	22 junio	12 agosto	51 días
4	2 julio	-	$>$ 60 días
5	21 julio	23 agosto	33 días
6	31 julio	27 agosto	27 días
7	31 julio	14 agosto	14 días
8	1 agosto	25 agosto	24 días
9	2 agosto	6 agosto	4 días
10	10 agosto	-	$>$ 21 días

Tabla 1.2: Datos de Bartholomew (ejemplo 2).

experimento se dio por terminado el 31 de agosto y, en ese momento, tres de las piezas, las número 2, 4 y 10, no habían fallado todavía. Se trata, por consiguiente, de observaciones censuradas, incompletas, de las que no se conoce el tiempo de funcionamiento. Sí se sabe que éste fue superior a 72, 60 y 21 días, respectivamente.

El objetivo del estudio es analizar la distribución del tiempo de funcionamiento de estos componentes, así como estimar la proporción de equipos que fallarán dentro de un intervalo temporal determinado.

Ejemplo 3: Gehan estudió los resultados de un ensayo clínico sobre la eficacia de la droga 6-mercaptopurina (6-MP) para prolongar el estado de remisión, es decir, la ausencia de síntomas de la enfermedad, en enfermos que habían padecido leucemia aguda. Con el fin de contrastar su efecto, se administró esta droga a un grupo de pacientes, mientras a otro grupo, que servía de control, se le administró un placebo. La asignación de pacientes a cada uno de los dos grupos se hizo aleatoriamente. En el ensayo participaron 42 individuos cuyos tiempos de remisión, registrados en semanas, se muestran en la tabla 1.3. El periodo de observación fue de un año y, durante ese intervalo, los enfermos se fueron incorporando al ensayo más o menos regularmente. Las observaciones marcadas con asterisco son censuradas. En esos

6-MP	6, 6, 6 ,6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32* 34*, 35*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Tabla 1.3: Datos de Gehan (ejemplo 3).

AG positivos N=17		AG negativos N=16	
N^o Glob. B.	T. Superv.	N^o Glob. B.	T. Superv.
2.300	65	4.400	56
750	156	3.000	65
4.300	100	4.000	17
2.600	134	1.500	7
6.000	16	9.000	16
10.500	108	5.300	22
10.000	121	10.000	3
17.000	4	19.000	4
5.400	39	27.000	2
7.000	143	28.000	3
9.400	56	31.000	8
32.000	26	26.000	4
35.000	22	21.000	3
100.000	1	79.000	30
100.000	1	100.000	4
52.000	5	100.000	43
100.000	65		

Tabla 1.4: Datos de Feigl y Zelen (ejemplo 4).

pacientes, la enfermedad estaba aún en estado de remisión en el momento en que se les efectuó el último control.

Cabe destacar la gran dispersión que se aprecia en los datos, así como el hecho de que la censura sea frecuente en el grupo tratado y no exista en el grupo control.

Ejemplo 4: Feigl y Zelen, analizaron la supervivencia de dos grupos de enfermos de leucemia clasificados, en base a una característica celular, en AG positivos y AG negativos, tabla 1.4. Los tiempos de fallo corresponden al tiempo, medido en semanas, desde el instante del diagnóstico hasta el fallecimiento; también se registró la cantidad de glóbulos blancos de cada paciente en el momento del diagnóstico.

A diferencia del caso anterior, la clasificación en grupos viene determinada por una característica morfológica de los individuos y no está controlada por el investigador. Dado que el estado de los enfermos no es homogéneo, se ha medido una covariable, el n^o de glóbulos blancos, que se cree que puede ser indicativa del pronóstico de supervivencia del enfermo. Posibles objetivos del análisis serían establecer esa relación y analizar si la característica celular influye en la supervivencia.

Ejemplo 5: El objetivo principal de un estudio realizado por Prentice sobre la supervivencia de pacientes con cáncer de pulmón en estado avanzado era comparar el efecto de dos tratamientos de quimioterapia en la prolongación del tiempo de vida. Para ello, se separaron dos grupos, en uno de ellos se utilizó el tratamiento de quimioterapia usual y en el otro se ensayó el nuevo tratamiento. La asignación de los pacientes a cada uno de los grupos se realizó de forma aleatoria.

El principal problema de este estudio es la inhomogeneidad de la muestra debido a las diferentes características de los enfermos. Para realizar la comparación de los tratamientos en condiciones homogéneas era necesario tomar en consideración otros factores que también podían influir en el tiempo de supervivencia; con este objetivo se midieron variables que representaran estos factores:

- el tipo de tumor: escamoso, adeno, de célula pequeña o de célula grande
- la edad, en años, del enfermo
- si habían recibido, o no, terapia antes de entrar al estudio
- el número de meses transcurrido entre el diagnóstico del tumor en el enfermo y su incorporación al estudio
- el estado general del enfermo en el momento de incorporarse al estudio, codificado en una escala con valores 10, 20, 30, ..., 90. Un mayor valor de ese índice refleja un mejor estado del enfermo, correspondiendo los valores 10-30 a pacientes completamente hospitalizados, 40-60 a enfermos en hospitalización parcial y 70-90 a los que podían cuidarse por su cuenta.

En la tabla 1.5 se muestran los datos correspondientes a una parte de los 137 participantes en el ensayo, los 40 individuos que habían recibido tratamiento terapéutico contra el cáncer con anterioridad. Para esos pacientes, clasificados en tablas distintas según el tratamiento asignado y el tipo de tumor, se muestran las variables: tiempo de supervivencia -que corresponde al tiempo, en días, desde la incorporación al estudio hasta el fallecimiento- estado general del enfermo, edad e intervalo de tiempo transcurrido entre el momento del diagnóstico y su entrada al estudio. El conjunto completo de datos se puede encontrar en el fichero LUNG.DAT.

T.SUP.	ESTADO	EDAD	DIAG-EST	T.SUP	ESTADO	EDAD	DIAG-EST
Trat. habitual, Tumor escamoso				Trat. ensayo, Tumor escamoso			
411	70	64	5	999	90	54	12
126	60	63	9	231*	50	52	8
118	70	65	11	991	70	50	7
82	40	69	10	1	20	65	21
8	40	63	58	201	80	52	28
25*	70	48	9	44	60	70	13
11	70	48	11	15	50	40	13
Trat.habitual, Tumor cel. pequeña				Trat. ensayo, Tumor cel.pequeña			
54	80	63	4	103*	70	36	22
153	60	63	14	2	40	44	36
16	30	53	4	20	30	54	9
56	80	43	12	51	30	59	87
21	40	55	2				
287	60	66	25				
10	40	67	23				
Trat.habitual, Tumor adeno				Trat. ensayo, Tumor adeno			
8	20	61	19	18	40	69	5
12	50	63	4	90	60	50	22
				84	80	62	4
Trat. habitual, Tumor cel. grande				Trat. ensayo, Tumor cel. grande			
177	50	66	16	164	70	68	15
12	40	68	12	19	30	39	4
200	80	41	12	43	60	49	11
250	70	53	8	340	80	64	10
100	60	37	13	231	70	67	18

Tabla 1.5: Datos de Prentice (ejemplo 5).

1.5 Tipos de datos censurados. Esquemas de censura

Como hemos comentado los datos correspondientes a estudios de Fiabilidad y Análisis de Supervivencia presentan una particularidad que dificulta su análisis estadístico. Esta peculiaridad es la presencia de datos censurados: sólo se conoce el tiempo de fallo para una fracción, que puede ser pequeña, de los individuos de la muestra, mientras que del resto se dispone sólo de información parcial, habitualmente que el tiempo de vida es mayor que un valor dado.

Una observación se dice **censurada a derecha** en L, si se desconoce el valor exacto de la observación y sólo se sabe que ésta es mayor que L. Análogamente, una observación se dice **censurada a izquierda** en L, si sólo se sabe que la observación es menor que el valor L. La censura a derecha es mucho más frecuente que la censura a izquierda. En algunos experimentos, dependiendo del tipo de problema y el tipo de seguimiento, aparecen datos **censurados en un intervalo** (t_I, t_D); es decir, que sólo se sabe que $t_I < T < t_D$.

Generalmente la duración del tiempo de ensayo se debe limitar por razones prácticas y económicas. Existen dos esquemas básicos para establecer este límite.

Censura de tipo I. En este esquema el experimento se programa con una duración, C, establecida a priori. El tiempo de fallo de un individuo se observará, si es menor o igual que ese valor prefijado. En otro caso, la observación correspondiente será censurada, con valor C, y la denotaremos $C*$. En este esquema, el número de observaciones de la muestra es aleatorio.

Censura de tipo II. En los ensayos realizados bajo un esquema de tipo II, con n componentes idénticos, el ensayo finaliza en el momento en que se produce el r-ésimo fallo ($1 \leq r \leq n$). Ese instante, $t_{(r)}$, será el valor de los datos censurados correspondientes a los componentes que en ese momento sigan funcionando. De esta forma sólo se conocen las r observaciones más pequeñas de la muestra y aparecen $n - r$ tiempos censurados en el valor $t_{(r)}$. Este tipo de censura se usa con frecuencia en los experimentos industriales y es más fácil de analizar desde el punto de vista estadístico.

Es importante señalar que el valor de C en el esquema de tipo I y el valor de r (o la fracción r/n) que indica la tasa de censura en el esquema de tipo II deben fijarse

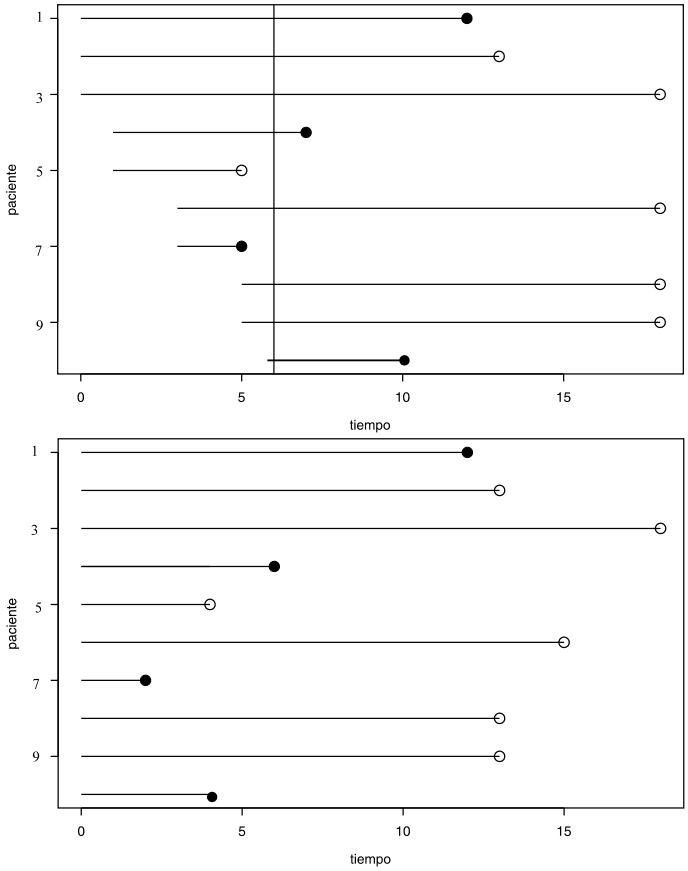

Figura 1.1: Esquema de los tiempos de fallo.

antes de iniciar el experimento y no durante el transcurso del mismo dependiendo de los resultados que se observen. La necesidad de que el mecanismo de censura sea independiente de la observación del fenómeno, es un requisito imprescindible para la validez de las conclusiones.

En los ensayos industriales, los experimentos diseñados con los procedimientos citados generan, habitualmente, **muestras simplemente censuradas**, es decir con un único valor común, $t_{(r)}$ o C, de las observaciones censuradas. En los ensayos médicos, aunque el experimento se diseñe con una limitación temporal como en el esquema I, es normal que los individuos se incorporen al ensayo en instantes aleatorios, cuando se dispone de los pacientes adecuados; además, es habitual que se produzcan abandonos durante la realización del ensayo. En consecuencia, las muestras resultantes son **múltiplemente censuradas**. Generalmente, este tipo de observaciones se presentan mediante un par de variables (T, δ), donde T es el tiempo

Paciente	T. de entrada	T. fallo o censura	Estado	T. de Supervivencia
1	0.0	11.8	F	11.8
2	0.0	12.5	C	12.5*
3	0.4	18.0	C	17.6*
4	1.2	6.6	F	5.4
5	1.2	4.4	C	3.2*
6	3.0	18.0	C	15.0*
7	3.4	4.9	F	1.5
8	4.7	18.0	C	13.3*
9	5.0	18.0	C	13.0*
10	5.8	10.1	F	4.3

Tabla 1.6: Datos correspondientes a la figura 1.1

transcurrido desde la entrada del individuo al ensayo hasta su salida del mismo y δ es una variable binaria indicadora del tipo de observación, que toma el valor 1 si se ha observado el fallo y el valor 0 si se trata de una observación censurada.

En la tabla 1.6 se muestran los datos recogidos en un estudio de 18 meses de duración, en el que sólo fueron admitidos pacientes durante los seis primeros meses. En la figura 1.1 se puede observar que cuatro pacientes habían muerto, cuatro permanecían vivos al finalizar el estudio y dos lo habían abandonado antes de que terminara. En el análisis de los tiempos de vida, el instante en el que se comenzó a medir cada observación no suele ser de interés, por lo que con frecuencia suelen representarse las observaciones con el mismo origen, como se muestra en la figura inferior.

1.6 Referencias de interés

Para terminar el capítulo indicamos algunas referencias que consideramos de interés. Es importante utilizar bibliografía reciente porque los métodos de análisis para tratar estos problemas y el software correspondiente han evolucionado con rapidez, especialmente los relativos a modelos de supervivencia con covariables.

En las referencias se indica, utilizando los símbolos (R) y (S), si el texto analiza el aspecto industrial o el biomédico respectivamente, y con un asterisco los libros de nivel más elemental. En los ejercicios propuestos al final de cada capítulo también se marcan con un asterisco aquellos que presentan mayor dificultad.

- Collet, D. (1994). Modelling Survival Data in Medical Research. Chapman and Hall. (S) *.
- Cox, D. R. y Oakes, D. (1984). Analysis of Survival Data. Chapman and Hall. (S).
- Crowder, M. J., Kimber, A. C., Smith, R. L. y Sweeting, T. J. (1991). Statistical Analysis of Reliability Data. Chapman and Hall. (R) *.
- Elandt-Johnson, R. C. y Johnson, N. L. (1980). Survival Models and Data Analysis. J. Wiley. (S).
- Gross, A. J. y Clark, V. A. (1975). Survival Distributions: Reliablility Applications in the Biomedical Sciences. J. Wiley. (S) *.
- Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. J. Wiley. (S, R)
- Lee, E. T. (1992). Statistical methods for survival data analysis. 2^a ed. J. Wiley. (S) *.
- Marubini, E. y Valsechi, M.G. (1995). Analysing Survival Data from Clinical Trials and Observational Studies. J. Wiley. (S) *.
- Nelson, W. (1982). Applied Life Data Analysis. J. Wiley. (R).

Capítulo 2

Conceptos probabilísticos básicos de Fiabilidad

2.1 Distribución del tiempo de supervivencia

En este capítulo se revisan e introducen algunos conceptos probabilísticos fundamentales para el estudio de la Fiabilidad.

En Fiabilidad la variable de mayor interés es la variable **tiempo hasta el fallo** o **tiempo de supervivencia**. Habitualmente, las observaciones de esta variable tienen gran variabilidad y una fuerte carga de indeterminación, por lo que cualquier modelo matemático para esos datos deberá contener una componente aleatoria que represente la variabilidad que no se pueda explicar.

En este capítulo consideraremos que los tiempos de supervivencia observados son observaciones independientes de una variable aleatoria (v.a.) T, no negativa. Asociado a esa variable T existe un espacio de probabilidad $(\xi, \omega, \wp_T)$ formado por,

ξ: el **espacio de resultados**, conjunto de todos los posibles valores que puede tomar la variable. Este espacio será, por lo general, $\mathbb{R}+ = [0, \infty)$ o un intervalo $I = [t_0, t_1]$ contenido en $\mathbb{R}+$. En algunas ocasiones, el espacio de resultados puede ser un conjunto discreto, por ejemplo, el de los enteros no negativos, $Z+ = \{0, 1, 2, 3, \ldots\}$.

ω: la **familia de sucesos** asociada al fenómeno. Un **suceso** es un hecho relativo a la variable T al que se puede asignar una probabilidad; algunos ejemplos son: $\{T \geq t_0\}$, $\{T < t_1\}$, $\{t_2 \leq T \leq t_3\}$ con t_0, t_1, t_2 y t_3, instantes de tiempo arbitrarios. Entre

los sucesos se definen las operaciones habituales -la ocurrencia simultánea de varios sucesos, la ocurrencia de alguno de ellos, la no ocurrencia de un suceso dado, etc.- y las operaciones conjuntistas correspondientes -intersección, unión, complementación, etc.

$\wp_T$: la **medida de probabilidad** asociada a la variable aleatoria, que llamaremos distribución de probabilidad de T, es una función que asigna un valor entre 0 y 1 a cada suceso de ω y queda caracterizada por la función de distribución F, que asigna a cada valor t de $\mathbb{R}$ la probabilidad de que se observe un tiempo de vida no superior a t,

$$F(t) = P(T \leq t).$$

La función F es monótona no decreciente y verifica $\lim_{t\to\infty} F(t) = 1$. En este caso, por ser T no negativa, $F(t) = 0$ para $t < 0$.

Cuestión: Con la ayuda de MINITAB, u otro paquete estadístico, genera muestras de distintos tamaños (n = 20, 50, 100, 250, 500), de:

1. Una distribución Uniforme en $[20, 200]$.
2. Una distribución Geométrica de parámetro $p = 1/3$.
3. Una distribución de Poisson de parámetro $\lambda = 1$, desplazada una unidad.
4. Una distribución Exponencial con $\lambda = 1$, desplazada 10 unidades.

Representa gráficamente los datos y calcula para las muestras obtenidas, las medidas descriptivas siguientes: media, mediana, moda, desviación típica, rango, rango intercuartílico, coeficiente de variación y coeficiente de asimetría. Calcula los correspondientes parámetros de dichas distribuciones de probabilidad y compáralos con las estimaciones obtenidas en las distintas muestras.

2.1.1 Modelos continuos

Diremos que una **variable** aleatoria es **continua** si su espacio de resultados es un subconjunto continuo de $\mathbb{R}$. En las definiciones siguientes supondremos que el espacio de resultados de T es $\xi = [0, \infty)$.

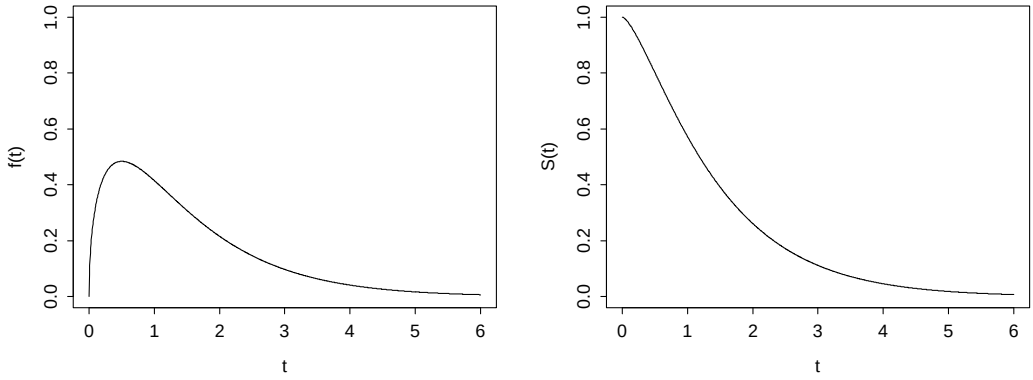

Figura 2.1: Función de densidad y función de supervivencia.

La función de densidad se define como una función f definida en $\mathbb{R}+$ que verifica,

$$\int_0^\infty f(x)dx = 1$$

y tal que, para todo t, su función de distribución puede expresarse como,

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

En la figura 2.1 se muestra la función de densidad de una variable tiempo de vida.

De la definición de densidad se deduce que, para todo t y cuando $t \to 0$,

$$f(t)\Delta t \approx P(t \leq T \leq t + \Delta t) = F(t + \Delta t) - F(t);$$

es decir, que el valor de la función de densidad en el instante t se puede interpretar como la probabilidad de fallo existente en torno a ese instante por unidad de tiempo.

En Fiabilidad es habitual utilizar otras funciones, además de las de distribución o densidad, para caracterizar la distribución de probabilidad de T. Todas ellas caracterizan unívocamente la distribución, pero proporcionan visiones diferentes relativas al tiempo de vida y sus características.

La **función de supervivencia** o **función de fiabilidad** $S(t)$, figura 2.1, asocia a cada valor t la probabilidad de que un individuo sobreviva a ese instante de tiempo:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx.$$

Esta función es monótona no creciente y verifica que $S(0) = 1$ y $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

Nota. Algunos autores definen la función de supervivencia como $S(t) = P(T \geq t)$. En el caso de distribuciones continuas ambas definiciones dan lugar a la misma función de supervivencia, pero no cuando la distribución es discreta o mixta.

La **función de riesgo** o **función de tasa de fallo**, $h(t)$ se define como,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Esta función representa la tasa instantánea de fallo en el instante t, dado que el individuo o pieza ha sobrevivido hasta esa edad, figura 2.2. Una visión intuitiva de la definición anterior nos dice que para todo t y cuando $\Delta t \to 0$:

$$h(t)\Delta t \approx P(t \leq T \leq t + Dt \mid T \geq t).$$

Las expresiones siguientes, muy sencillas de comprobar, muestran las relaciones existentes entre las funciones más utilizadas,

$$\begin{aligned} f(t) &= -\frac{d}{dt} S(t) \\ h(t) &= -\frac{d}{dt} \ln S(t) \\ S(t) &= \exp\left[-\int_0^t h(x)dx\right]. \end{aligned}$$

La **función de riesgo acumulado**, H(t), se define como,

$$H(t) = \int_0^t h(x)dx$$

y está relacionada con la función de supervivencia mediante la expresión,

$$S(t) = \exp[-H(t)].$$

Cuestión: Calcula las funciones de densidad, riesgo y supervivencia de una distribución uniforme en [20, 200] y de una Exponencial de parámetro $\lambda = 1$.

2.1.2 Modelos discretos

En algunos casos, en la práctica poco frecuentes, puede ser necesario tratar el tiempo de vida T como una variable aleatoria discreta; por ejemplo, cuando se mide el

tiempo de vida de forma tal que el número de valores distintos que puede tomar la variable es pequeño y son frecuentes los empates en las observaciones. Supondremos que el espacio de resultados asociado es ahora $\xi = \{t_{(1)}, t_{(2)}, t_{(3)}, \ldots\}$ donde $0 \leq t_{(1)} < t_{(2)} < \ldots$

La distribución de T viene determinada por su **función de probabilidad**,

$$p_j = P(T = t_{(j)}) j = 1, 2, \ldots$$

con $p_j > 0$ y $\sum_j p_j = 1$. La función de supervivencia,

$$S(t) = P(T > t) = \sum_{j:t_{(j)}>t} p_j$$

es en este caso una función monótona no creciente, que cambia en los instantes de fallo $t_{(j)}$, continua a derecha y tal que $S(\infty) = 0$. A diferencia del caso continuo, el valor de esta función de supervivencia difiere en los instantes de fallo del que se obtendría definiendo $S(t) = P(T \geq t)$.

La función de riesgo se define en los instantes $t_{(j)}$, como la probabilidad condicional de fallo en ese instante, dado que se ha llegado vivo a él:

$$\begin{aligned} h(t_{(j)}) &= h_j = P(T = t_{(j)} \mid T \geq t_{(j)}) = \frac{p_j}{S(t^-_{(j)})} = \frac{p_j}{S(t_{(j-1)})} \qquad j = 1, 2, \ldots \\ h(t) &= 0 \qquad \text{si } t \neq t_{(j)}. \end{aligned}$$

Como en el caso continuo, estas funciones determinan completamente la distribución de T. Las relaciones existentes entre ellas son:

$$\begin{aligned} h_j &= 1 - \frac{S(t_{(j)}}{S(t_{(j-1)})} \qquad j = 1, 2, \ldots \\ S(t) &= \prod_{j:t_{(j)} \leq t} (1 - h_j) \end{aligned}$$

La definición de función de riesgo acumulado es en este caso:

$$H(t) = \sum_{j:t_j \leq t} h_j$$

y no satisface la relación encontrada en el caso continuo pues, en general,

$$-\ln S(t) = -\sum_{j:t_{(j)} \leq t} \ln(1 - h_j) \neq \sum_{j:t_{(j)} \leq t} h_j;$$

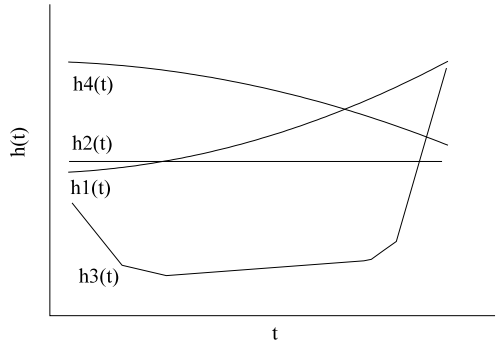

Figura 2.2: Distintas funciones de riesgo.

no obstante, como puede comprobarse desarrollando en serie $\ln(1-x)$, ambas expresiones producirán resultados próximos si los valores h_j son pequeños.

Cuestión: Calcula las funciones de probabilidad, riesgo y supervivencia de una distribución Geométrica de parámetro $p = 1/3$ y de una distribución de Poisson con $\lambda = 1$.

2.2 Interpretación de la función de riesgo y tiempo restante de vida

La función $h(t)$ representa la evolución de la probabilidad de fallo en relación con la edad de los individuos. Con el fin de ilustrar este concepto, comentaremos cuatro ejemplos de poblaciones cuyo perfil de riesgo se corresponde con cada uno de los tipos de curvas que se muestran en la figura 2.2:

- $h1(t)$: El conjunto de personas mayor de 65 años. Esta población presenta una función de riesgo creciente que nos indica que la tasa de fallo tiende a aumentar con el transcurso del tiempo. Por ejemplo, la probabilidad de que un individuo con 70 años viva más de 71, es mayor que la probabilidad de que un individuo con 80 viva más de 81.

- $h2(t)$: Una población de individuos sanos entre los 20 y 40 años de edad, para los que el único riesgo de muerte, en la práctica, viene dado por distintos tipos de accidentes (laborales, deportivos, de tráfico, etc.). En esta población, la función de riesgo es prácticamente constante.

- $h3(t)$: Una población, por ejemplo la de los españoles nacidos en la década

de los setenta, observada desde el nacimiento hasta la muerte. Esa población presentará una función de riesgo con forma de J, llamada también riesgo "bañera", típica de las tablas de vida poblacionales. Inicialmente se tiene un periodo con tasa de fallo alta, correspondiente a la etapa neonatal e infantil, que va decreciendo hasta estabilizarse. El riesgo permanece bajo y aproximadamente constante, hasta una cierta edad, en torno a los 40 años, a partir de la cual comienza a aumentar con el tiempo.

- $h4(t)$: Una población de personas jóvenes que padece cierto defecto congénito y que es sometida a un proceso quirúrgico complicado para corregirlo, analizada mientras dura el periodo de recuperación. Esta población presentará una tasa de riesgo decreciente ya que en estos casos, el principal riesgo de muerte aparece como consecuencia de la intervención o de sus complicaciones inmediatas.

El tiempo restante de vida. En muchos problemas de supervivencia tienen gran interés las cuestiones relativas al comportamiento esperado de T en los individuos que han alcanzado cierta edad. Formalmente, se trata de estudiar la distribución de la variable $R_t = T - t$, dado que ha ocurrido $T \geq t$ para un cierto valor t. Su función de supervivencia será:

$$S_{R_t}(x) = P(R_t > x) = P(T > t + x \mid T \geq t) = \frac{S(t+x)}{S(t^-)},$$

donde $S(t^-)$ denota el límite a izquierda en t de la función de supervivencia.

Si T es una variable continua que toma valores en $[0, \infty)$, con función de densidad $f(x)$, R_t será también una v.a. continua, con valores en $[0, \infty)$ y función de densidad:

$$f_{R_t}(x) = \frac{f(x+t)}{S(t)} \qquad\qquad x \geq 0.$$

En cuanto a la relación entre las funciones de riesgo y riesgo acumulado de T y R_t se verifica:

$$\begin{aligned} h_{R_t}(x) &= h(t+x) \\ H_{R_t}(x) &= H(t+x) - H(t) \qquad\qquad x \geq 0. \end{aligned}$$

Finalmente, una función que también caracteriza la distribución de T es la **función de vida media residual**, m(t). Para cada t positivo, se define como,

$$m(t) = E(R_t) = \int_0^\infty S_{R_t}(x)dx = \frac{\int_t^\infty S(x)dx)}{S(t^-)}.$$

Cuestión: Calcula la distribución de la variable R_t si T es $Exp(\lambda)$ y si T es Geométrica de parámetro p. Calcula también la función $m(t)$ en ambos casos.

2.3 Algunas distribuciones de probabilidad básicas

En este apartado se revisan algunas de las distribuciones de probabilidad más empleadas en Fiabilidad y Análisis de Supervivencia. En principio, cualquier distribución no negativa se puede utilizar para modelizar una variable tiempo de vida; el objetivo es disponer de un conjunto de distribuciones lo suficientemente flexibles para adaptarse a los distintos tipos de datos y lo más sencillas posibles para facilitar su análisis.

En general, la distribución más utilizada en Estadística es la distribución Normal. En Fiabilidad, sin embargo, la distribución de referencia es la Exponencial. Las buenas propiedades de esta distribución, consecuencia de su ausencia de memoria, permiten simplificar los problemas de inferencia; por el mismo motivo su aplicación práctica es limitada, siendo más utilizadas generalizaciones suyas como las distribuciones Weibull o Gamma.

Presentamos las distribuciones bajo la hipótesis de que el rango de valores de T es $\xi = [0, \infty)$. Todas las distribuciones de tiempo de fallo admiten una versión más general en la que aparece un nuevo parámetro G, llamado umbral o **tiempo de garantía**, que puede tomar cualquier valor no negativo. Esta versión generalizada, en la que el rango de T es $[G, \infty)$, se obtiene al considerar que $T - G$ tiene una distribución con rango $[0, \infty)$. Los modelos con parámetro de garantía tienen interés en situaciones en las que el conocimiento previo del fenómeno permite suponer que el riesgo de fallo en el intervalo $[0, G)$ es nulo. Generalmente el valor de G es desconocido y es necesario estimarlo junto con los demás parámetros.

En el tercer capítulo se desarrollan procedimientos de selección de la mejor distribución para modelizar la variable tiempo de vida. Si se encuentra una distribución de probabilidad que representa bien los datos, se pueden aplicar métodos de inferencia paramétricos basados en dicha distribución. Este tipo de análisis es más frecuente en el campo de la Fiabilidad que en el del Análisis de Supervivencia.

2.3.1 Distribución Exponencial

La distribución de probabilidad más sencilla es la que presenta una función de riesgo constante,

$$h(t) = \lambda \qquad \text{para } 0 \leq t < \infty,$$

con λ una constante positiva. Las restantes funciones que caracterizan este modelo son,

$$\begin{aligned} H(t) &= \lambda t \\ S(t) &= \exp(-\lambda t) \\ f(t) &= \lambda \exp(-\lambda t) \qquad \text{para} 0 \leq t < \infty. \end{aligned}$$

Esta distribución se denomina **Exponencial** de parámetro λ. Su media es $1/\lambda$, su varianza $1/\lambda^2$, y su coeficiente de variación la unidad. Una propiedad importante, característica de esta distribución, es la **ausencia de memoria**; en cualquier instante t, la variable tiempo de vida restante, $R_t = T - t \mid T \geq t$, sigue también una distribución $Exp(\lambda)$.

Cuestión: Formula la distribución Exponencial con dos parámetros λ y G. Calcula la expresión de las funciones que lo caracterizan y su media y varianza.

2.3.2 Distribución de Weibull

En la mayor parte de los fenómenos de interés la hipótesis de que la función de riesgo sea constante resulta demasiado restrictiva. La distribución de Weibull define un modelo más general cuya función de riesgo es,

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1} \qquad \text{para } 0 \leq t < \infty$$

donde los parámetros λ y γ, denominados parámetros de escala y forma respectivamente, toman valores positivos. Esta función es siempre monótona; es creciente (IFR) si $\gamma > 1$ y decreciente (DFR) si $\gamma < 1$. Si $\gamma = 1$, la función de riesgo es constante y corresponde al modelo $Exp(\lambda)$. Las restantes funciones características de la distribución son,

$$\begin{aligned} S(t) &= \exp[-(\lambda t)^{\gamma}] \\ f(t) &= \lambda\gamma(\lambda t)^{\gamma-1} \exp[-(\lambda t)^{\gamma}] \qquad \text{para } 0 \leq t < \infty. \end{aligned}$$

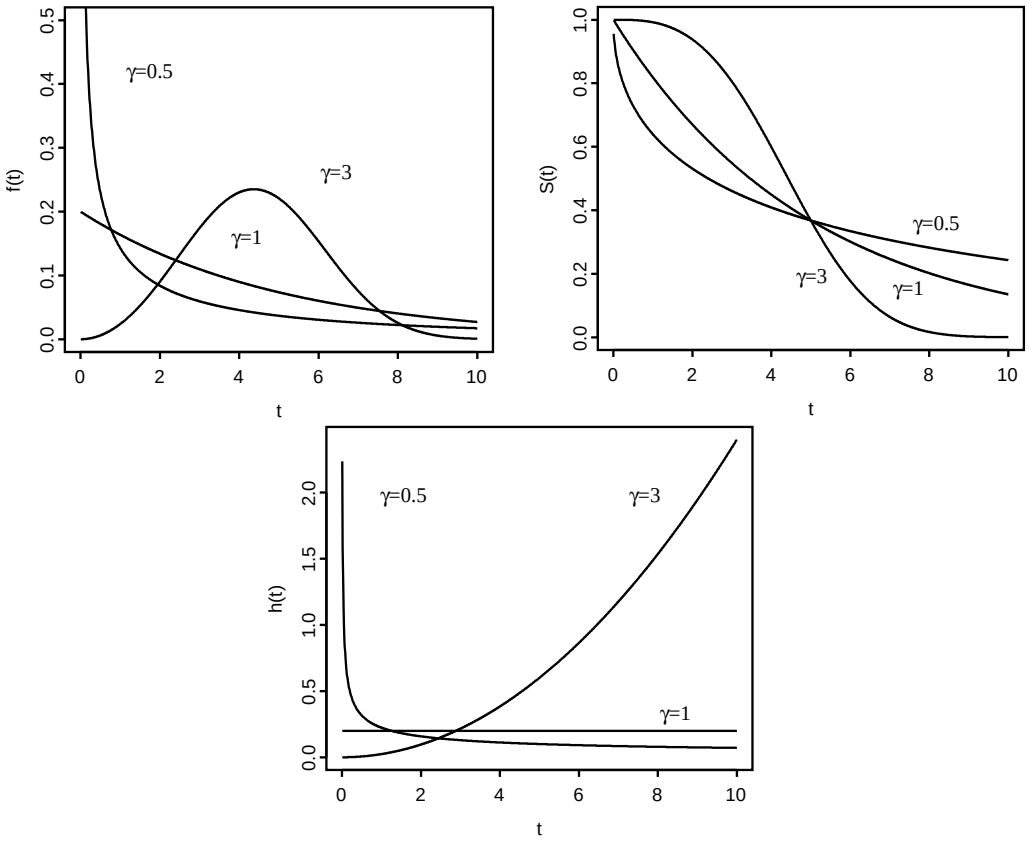

Figura 2.3: Funciones características de la distribución Weibull con $\lambda = 5$.

El nombre de esta distribución proviene del físico sueco que la introdujo por primera vez en 1939 en relación con experimentos de resistencia de materiales. La media de la distribución es,

$$E(T) = \frac{\Gamma(1+\gamma^{-1})}{\lambda}$$

donde $\Gamma(x)$ es la función gamma, definida para todo $x > 0$ por la integral,

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du.$$

La función de riesgo para diferentes valores de γ y las correspondientes funciones de densidad y supervivencia se muestran en las gráficas de la figura 2.3. La gran variedad de formas que puede tomar esta distribución dependiendo del valor de γ, y la relativa sencillez de sus funciones, hacen que sea una de las distribuciones más utilizadas en el análisis paramétrico de tiempos de fallo.

Cuestión: Deduce la expresión de $S(t)$ y de $f(t)$ a partir de la expresión de $h(t)$.

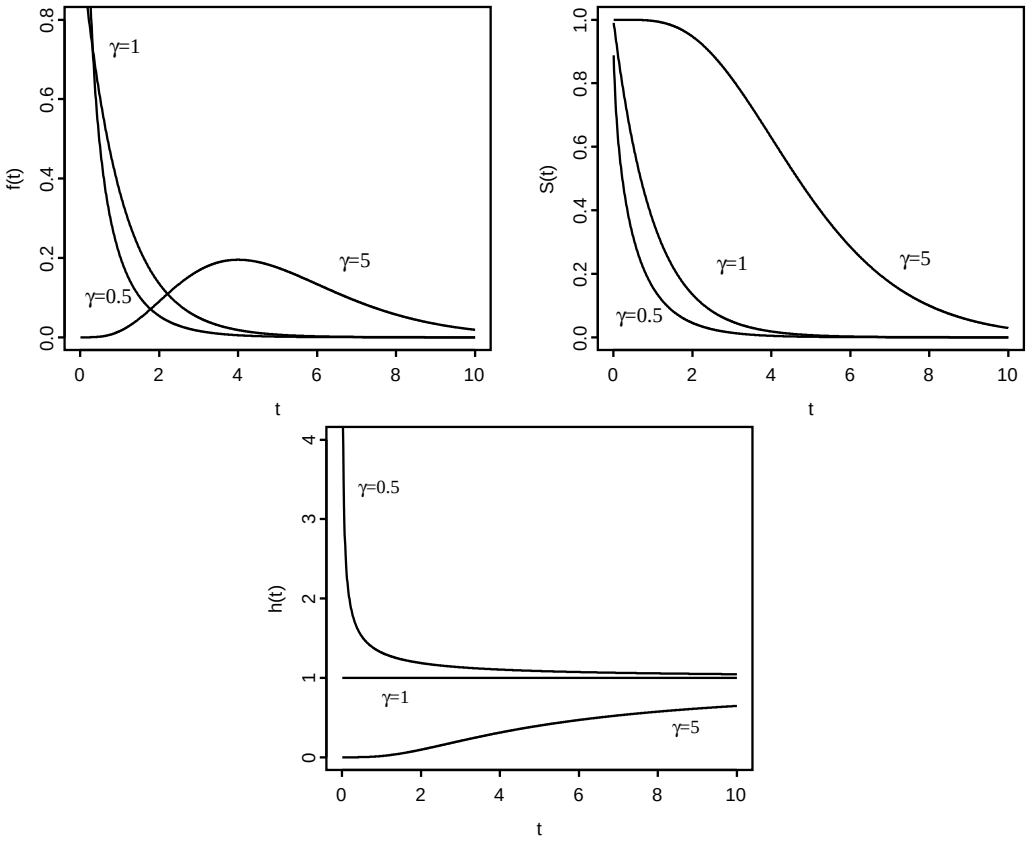

Figura 2.4: Funciones características de la distribución Gamma con $\lambda = 1$.

Comprueba también que la expresión de la esperanza de la distribución Weibull es la indicada.

Cuestión: Comprueba que si una variable T tiene una distribución Weibull de parámetros λ y γ, la variable T^γ tiene una distribución $Exp(\lambda^\gamma)$.

Nota. Conviene señalar que la forma de parametrizar las distribuciones no es única y que algunos textos y paquetes estadísticos utilizan expresiones distintas a las indicadas. En concreto, es frecuente definir como distribución $Exp(\lambda)$ la que tiene su media igual a λ; otra forma de presentar el modelo Weibull consiste en definir su función de riesgo como $h(t) = \lambda\gamma t^{\gamma-1}$, lo que implica $S(t) = \exp[-\lambda t^\gamma]$.

2.3.3 Otras distribuciones

Distribución Gamma. La distribución Gamma tiene dos parámetros positivos γ

y λ y su función de densidad es,

$$f(t) = \frac{\lambda}{\Gamma(\gamma)} (\lambda t)^{\gamma-1} \exp(-\lambda t) \qquad \text{para } t > 0.$$

Su media y varianza son γ/λ y γ/λ^2 respectivamente y su coeficiente de variación es $1/\sqrt{(}\gamma)$, independiente del valor de λ. En la figura 2.4 se muestran las gráficas de las funciones de densidad, distribución y riesgo para distintos valores de los parámetros.

Como ocurría en la distribución Weibull, el modelo gamma es IFR si $\gamma > 1$ y DFR si $\gamma < 1$; en ambos casos $h(t)$ tiende a λ al crecer t. Si $\gamma = 1$ se obtiene la distribución Exponencial. En el caso particular en que γ toma valores enteros, esta distribución suele denominarse de Erlang y aparece con frecuencia en los modelos de Teoría de Colas. El caso particular $\gamma = n/2$ y $\lambda = 1/2$, corresponde a la distribución χ^2 con n grados de libertad.

Aunque la distribución Gamma es una de las distribuciones continuas que toman valores positivos más importantes, en Fiabilidad no es muy utilizada, ya que las expresiones de las correspondientes funciones de riesgo y supervivencia son complicadas. La distribución de Weibull proporciona en muchos casos resultados similares a los que se obtienen con la distribución Gamma y la inferencia con ella es más sencilla.

Distribuciones Lognormal y Log-logística. Otra forma de formular un modelo para el tiempo de supervivencia, T, consiste en especificar una distribución para la variable $Y = \ln(T)$, que toma valores en toda la recta real. Una posibilidad es considerar que $\ln(T)$ tiene una distribución Normal de media μ y varianza σ^2; en este caso diremos que T sigue una distribución Lognormal de parámetros μ y σ. Desafortunadamente, esta distribución presenta el mismo inconveniente que la distribución Gamma, ya que su función de supervivencia no tiene una expresión explícita.

La función de riesgo verifica $h(0) = 0$, crece hasta alcanzar un máximo y posteriormente decrece de nuevo hacia 0; estas características lo hacen poco realista. En la figura 2.5 se muestran las funciones de densidad y riesgo para distintos valores de los parámetros.

Una distribución que se obtiene por el mismo procedimiento y que, a diferencia de la Lognormal, tiene expresiones de las funciones básicas relativamente sencillas es

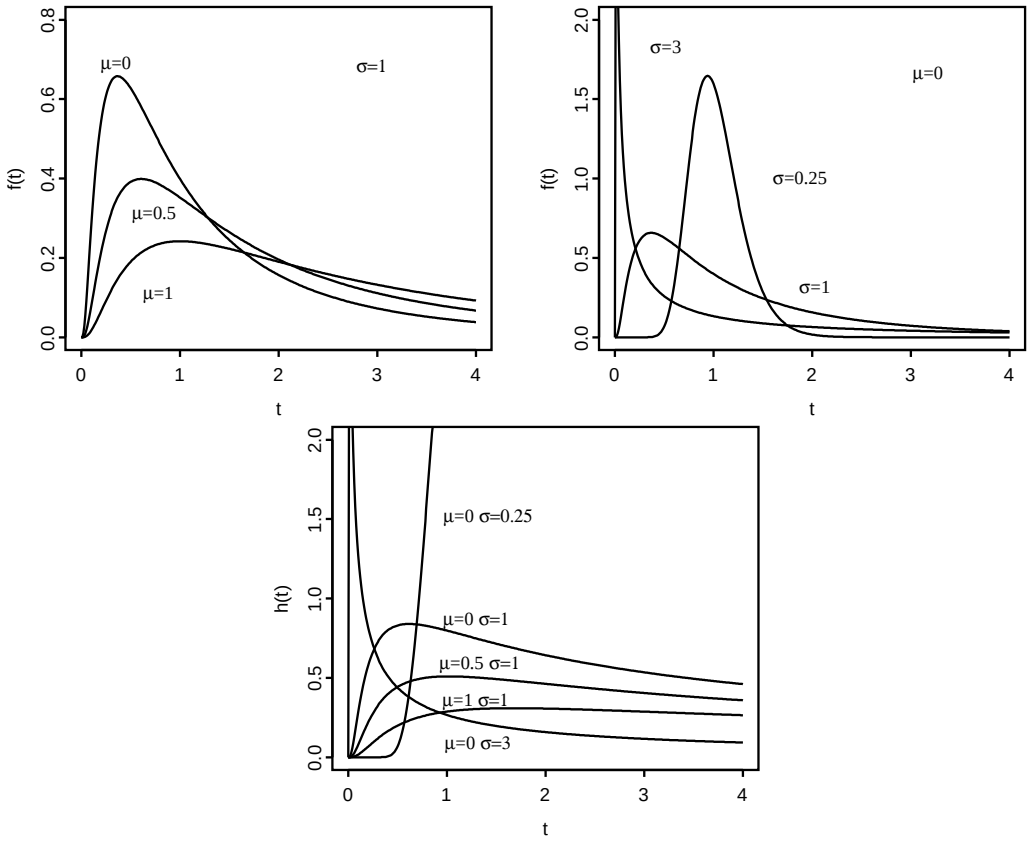

Figura 2.5: Funciones características de la distribución Lognormal.

la distribución Log-logística. Diremos que T tiene una distribución Log-logística si la variable $\ln(T)$ tiene una distribución Logística. Esta distribución, como la Normal, es una distribución de localización y escala,

$$Y = \ln(T) = \mu + \sigma W,$$

donde W es la distribución Logística estándar. W tiene una función de densidad simétrica parecida a la de la distribución $N(0,1)$, salvo en las colas.

Las expresiones de las funciones que caracterizan la distribución Log-logística de parámetros $\theta = -\mu/\sigma$ con $-\infty < \theta < \infty$ y $\kappa = 1/\sigma$, con $\kappa > 0$ son,

$$\begin{aligned} S(t) &= \left(1 + e^{\theta} t^{\kappa}\right)^{-1} \\ f(t) &= \frac{e^{\theta} \kappa t^{\kappa-1}}{\left(1 + e^{\theta} t^{\kappa}\right)^{2}} \\ h(t) &= \frac{e^{\theta} \kappa t^{\kappa-1}}{1 + e^{\theta} t^{\kappa}} \qquad \text{para } t > 0. \end{aligned}$$

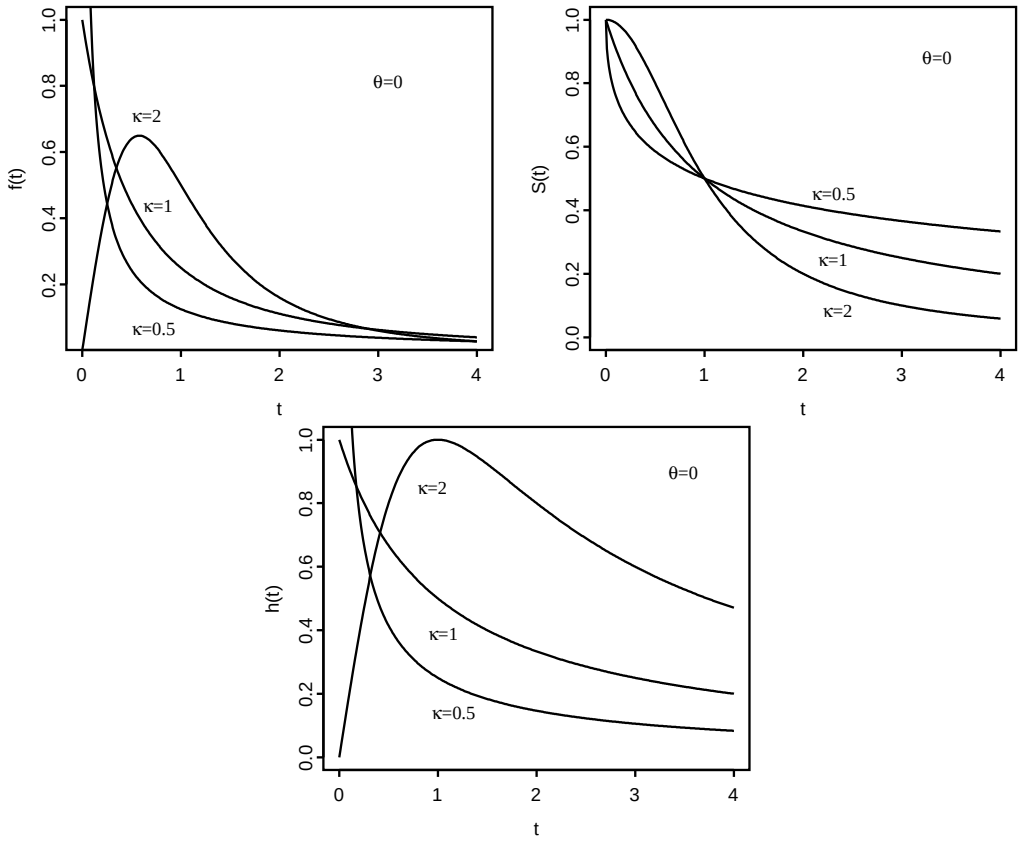

Figura 2.6: Funciones características de la distribución Log-logística con $\theta = 0$.

La función de riesgo es monótona decreciente si $\kappa \leq 1$. Si $\kappa > 1$, la función tiene un único máximo y permite modelizar situaciones que presentan dos fases diferentes, una inicial con riesgo creciente a la que sigue otra con riesgo decreciente. En la figura 2.6 se muestran las funciones de densidad, supervivencia y riesgo, para distintos valores del parámetro κ.

Dada la similitud existente entre las distribuciones Normal y Logística, el modelo Log-logístico produce resultados similares a los que se obtienen con el modelo Lognormal y, como hemos señalado, resulta más fácil de calcular.

Distribución Gumbel. Esta distribución tiene como rango de valores $(-\infty, \infty)$ y su función de supervivencia es,

$$S(t) = \exp\left(-\exp[(t-\alpha)/\beta]\right) \qquad \text{para } -\infty < t < \infty$$

donde α, con $-\infty < \alpha < \infty$, es un parámetro de localización y β ($\beta > 0$) un

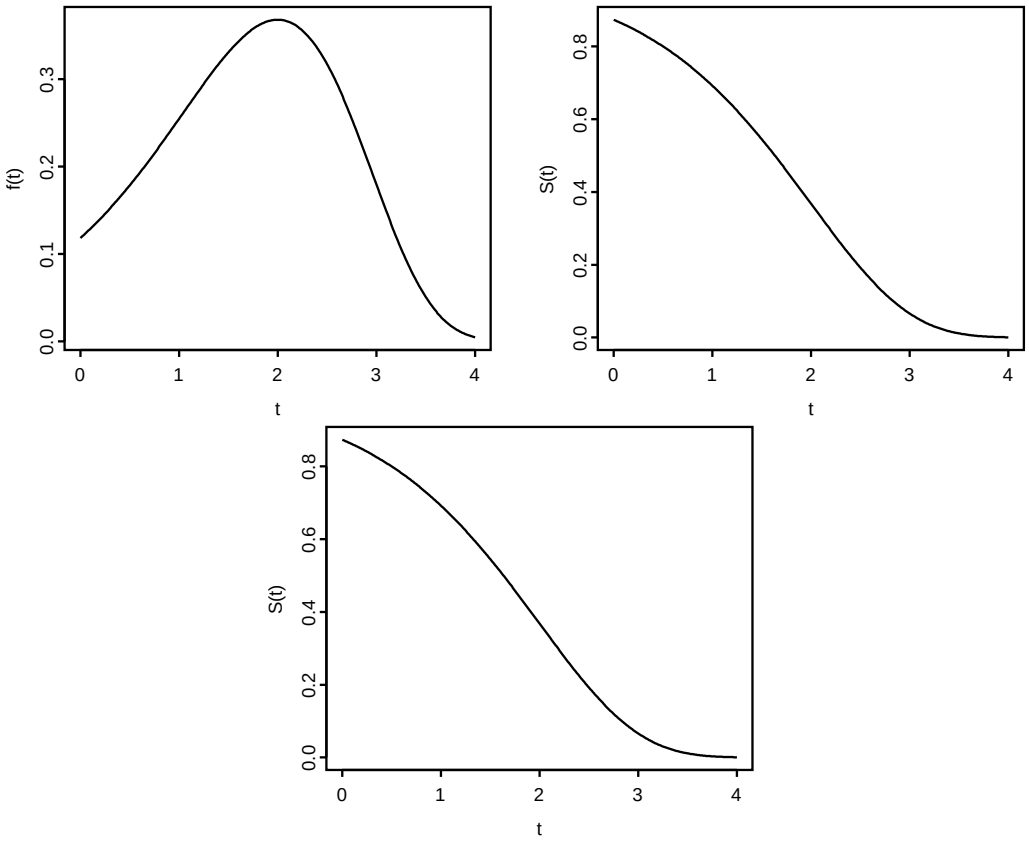

Figura 2.7: Funciones características de la distribución Gumbel con $\beta = 1$ y $\alpha = 2$.

parámetro de escala. Esta es una de las tres distribuciones de tipo Valor Extremo, VE, que corresponden a las distribuciones límite posibles del máximo de una muestra aleatoria. Su función de riesgo es,

$$h(t) = \frac{1}{\beta} \exp[(t - \alpha)/\beta].$$

El interés de esta distribución se debe en gran parte a su relación con la distribución Weibull: si el tiempo de supervivencia T es $Weibull(\gamma, \lambda)$, la variable $\ln(T)$ tiene una distribución Gumbel de parámetros $\alpha = -\ln(\lambda)$ y $\beta = 1/\gamma$. En las gráficas de la figura 2.7 se muestran las funciones de densidad, riesgo y supervivencia de esta distribución con $\beta = 1$ y $\alpha = 2$.

Aunque el rango de la distribución de valor extremo es $\mathbb{R}$, se puede utilizar como modelo para el tiempo de vida. Para ello es necesario truncarla en el valor cero; es decir, considerar la distribución condicionada a que tome valores no negativos. La distribución resultante se denomina distribución Gompertz y su función de riesgo

coincide con la de la distribución Gumbel (véase la cuestión final de este apartado). El modelo Gompertz puede a su vez generalizarse sumando un parámetro a su tasa de riesgo. La distribución así obtenida se denomina Gompertz-Makeham, y su función de riesgo es:

$$h(t) = \theta + \beta \exp(\alpha t) \qquad \text{para } t \geq 0.$$

Estas distribuciones resultan adecuadas en situaciones en las que el riesgo aumenta o decrece de forma rápida ya que la tasa de fallo es una función exponencial del tiempo.

Cuestión: Comprueba que si T tiene una distribución $Weibull(\gamma, \lambda)$, la variable $\ln(T)$ tiene una distribución Gumbel de parámetros $\alpha = -\ln(\lambda)$ y $\beta = 1/\gamma$.

Cuestión: Comprueba que al truncar a izquierda una distribución, la correspondiente función de riesgo no cambia; es decir, si $h_T(t)$ es la función de riesgo de T, y $t \geq \alpha$,

$$h_{T \mid T>\alpha}(t) = h_T(t).$$

Otras distribuciones. Otros modelos posibles son:

- La distribución de Rayleigh, o modelo lineal exponencial, es una distribución cuya función de riesgo es una función lineal del tiempo,

  $$h(t) = \lambda + \gamma t$$

  donde λ y γ pueden tomar cualquier valor siempre que $h(t)$ sea no negativa.

- Otra función de riesgo de interés es la denominada riesgo 'bañera'. Dos alternativas paramétricas que proporcionan funciones de riesgo con estas características son,

  $$\begin{aligned} h(t) &= \delta t + \frac{\beta}{t+\gamma} \\ h(t) &= \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left[\left(\frac{t}{\alpha}\right)^{\beta}\right]. \end{aligned}$$

  El inconveniente de estos modelos es su dificultad de tratamiento. Una alternativa más sencilla es descomponer la distribución original en dos fases: un periodo inicial con riesgo decreciente y otro, para quienes sobreviven a esa etapa inicial, con riesgo creciente.

- Un modelo sencillo y útil en algunas situaciones es el que presenta una función de riesgo constante a trozos:

$$h(t) = \begin{cases} \alpha_1 & \text{si } 0 \leq t < t_1 \\ \alpha_2 & \text{si } t_1 \leq t < t_2 \\ \dots & \dots \\ \alpha_{k-1} & \text{si } t_{k-2} \leq t < t_{k-1} \\ \alpha_k & \text{si } t \geq t_{k-1}. \end{cases}$$

  Este es el modelo probabilístico subyacente en el análisis de tablas de vida clínicas; en ellas las observaciones del tiempo de supervivencia se agrupan en intervalos en los que se supone el riesgo constante.

- Finalmente mencionaremos los modelos obtenidos mediante la mezcla de varias distribuciones. Estos modelos aparecen cuando en la población conviven individuos que provienen de k poblaciones distintas. Si suponemos que en la población global existe una proporción p_i de individuos de la subpoblación i, con $0 < p_i < 1$ y $\sum p_i = 1$ y que la distribución del tiempo de vida de cada una de esas subpoblaciones tiene una función de supervivencia $S_i(t)$, la función de supervivencia de un individuo de la población global es de la forma:

$$S(t) = p_1 S_1(t) + \ldots + p_k S_k(t).$$

  Estos modelos son adecuados en aquellas situaciones en las que la población no es homogénea y no es posible separar los individuos de cada tipo. Habitualmente las distribuciones correspondientes a cada subpoblación suelen pertenecer a la misma familia paramétrica, aunque no es necesario imponer esta restricción. Las propiedades de este tipo de modelos se pueden deducir a partir de las características de las k distribuciones que definen la mixtura. No es frecuente utilizar valores de k mayores que tres ya que en ese caso el número de parámetros desconocidos es demasiado grande y su estimación resulta complicada.

2.4 Ejercicios

1.- Si el tiempo de supervivencia T se mide en horas, ¿en qué unidades se deben expresar las funciones $f(t)$, $S(t)$, $h(t)$ y $H(t)$? ¿Qué relación guardan los valores de

estas funciones en cierto instante t, con sus valores en el mismo instante cuando T se mide en segundos?

2.- Comprueba que, tanto en el caso discreto como en el caso continuo, el valor esperado de la variable T puede calcularse como $E[T] = \int_0^\infty S(t)dt$.

3.- Sean $T_1, T_2, \ldots, T_n$ variables aleatorias independientes, continuas, no negativas y con funciones de riesgo $h_1(t), \ldots, h_n(t)$, respectivamente. Prueba que la función de riesgo de la variable $T = \min(T_1, \ldots, T_n)$ es $h_T(t) = \sum h_i(t)$.

4.- Se dispone de una muestra de 2000 unidades de un tipo de célula solar de la que se quieren estudiar sus características. Para ello, se realiza un ensayo acelerado con el fin de medir su vida útil (en un ensayo acelerado, la muestra se pone a prueba en condiciones más exigentes que las de uso habitual con el fin de acortar la duración del ensayo). Las frecuencias relativas correspondiente a dicho ensayo, con los tiempos de vida expresados en miles de horas, se muestran en la tabla 2.1.

t.vida (10^3h)	0-1	1-2	2-3	3-4	4-5	5-6	6-7
frec. rel.	0.15	0.25	0.25	0.10	0.10	0.08	0.07

Tabla 2.1: Tiempos de fallo de células solares. (ejercicio 4).

Se supone que la relación existente entre los parámetros de escala de la distribución del tiempo de funcionamiento en condiciones normales y de la distribución en condiciones aceleradas, es 10:1. Contesta las siguientes cuestiones relativas al tiempo de funcionamiento de las células en condiciones operativas normales.

i.- Estima la función de supervivencia de esas células en el instante $t = 3.5$ años.

ii.- ¿Cuál es la tasa de fallo de una célula cuando lleva un año funcionando?

iii.- Entre las células que han estado funcionando durante más de 20.000 horas, ¿qué porcentaje se espera que funcione más de 40.000 horas?

iv.- De las células que han alcanzado la edad de 10.000 horas, ¿cuál es el porcentaje esperado de células que fallarán cuando lleven funcionando entre 20.000 y 40.000 horas?

5.- La función de supervivencia del tiempo de vida de cierta pieza medido en horas es:

$$S(t) = \frac{1}{2}\exp(-t/2) + \frac{1}{2}\exp(-t/3).$$

i.- Calcula el valor del tiempo medio hasta el fallo de esta pieza.

ii.- Comprueba que su función de riesgo es:

$$h(t) = \frac{1/2 + (1/3)\exp(t/6)}{1 + \exp(t/6)}$$

y calcula el riesgo en el instante $t = 1$.

iii.- Prueba que la función de tasa de fallo es una función decreciente.

iv.- Si se sabe que la pieza ha funcionado más de dos horas, calcula la probabilidad de que falle en el intervalo $(2, 3)$.

6.- En ocasiones, las tasas de fallo se expresan en una unidad denominada FIT, que equivale al número esperado de fallos del componente por cada 10^9 horas de funcionamiento efectivo. Supongamos que cierta pieza tiene una tasa constante de fallo correspondiente a 325.000 FITs.

i.- ¿Cuál es la probabilidad de que falle por primera vez entre el sexto y el duodécimo mes de funcionamiento, sabiendo que no ha fallado durante los seis primeros meses? Considera que 1 mes equivale a 160 horas de funcionamiento efectivo.

ii.- ¿Cuál es el número esperado de fallos del componente durante un periodo de funcionamiento de 10^4 horas?

7.- La tasa de fallo de cierta pieza que opera de manera continuada puede describirse mediante la siguiente función,

$$h(t) = \left| \begin{array}{ll} 0.1 \text{ min}^{-1} & 0 \leq t \leq 10 \text{ min.} \\ 0.001 \text{ min}^{-1} & 10 < t \leq 1010 \text{ min.} \\ 0.01 \text{ min}^{-1} & t > 1010 \text{ min.} \end{array} \right.$$

i.- Especifica la función de supervivencia correspondiente a esta función de riesgo y represéntala gráficamente.

ii.- ¿Qué proporción de piezas, si empleamos un gran número de ellas, puede esperarse que funcionen entre 80 y 100 horas?

iii.- ¿Qué vida media tiene una pieza de estas características? Calcula la mediana del tiempo de vida.

iv.- Cuando una pieza lleva funcionando una semana ininterrumpidamente, ¿cuál es la distribución de R_t en ese instante? Calcula su valor esperado.

8.- La tabla 2.2 resume una tabla de vida poblacional correspondiente a los Estados Unidos de América, con una población base de 100.000 individuos. Para su construcción se ha utilizado la información estadística sobre mortalidad correspondiente al periodo 1959-1961.

i.- Calcula el porcentaje de personas que mueren en cada intervalo. Dibuja el histograma correspondiente a esta tabla de vida.

ii.- Calcula la vida media de las personas que viven más de 5 años y menos de 85. Para calcular este valor considera que la muerte se produce en el punto medio de cada intervalo).

iii.- Estima la función de supervivencia en el extremo inferior de cada intervalo para las personas que viven al menos 5 años y calcula los percentiles 10, 50 y 90 de la distribución de su tiempo de vida. ¿Qué proporción de esos individuos alcanzará los 65 años?

iv.- Estima la función de densidad en el punto medio de cada intervalo y represéntala.

v.- Estima la función de riesgo en el punto medio de cada intervalo y dibújala.

vi.- Estima la función de riesgo para las personas con 10 años de edad y para las de 50. ¿Qué conclusiones obtienes al comparar ambas funciones?

9(*).- Dada una distribución de probabilidad continua, considera las cuatro propiedades siguientes:

i.- $h(t)$ es no decreciente para todo $t \geq 0$. Las distribuciones con esta propiedad suelen llamarse distribuciones IFR, o con tasa de fallo creciente.

Intervalo edad	0-1	1-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
N^o muertes	2593	409	233	214	440	594	612	761	1080	1686
Intervalo edad	45-50	50-55	55-60	60-65	65-70	70-75	75-80	80-85	> 85	
N^o muertes	2622	4045	5644	7920	10290	12687	14594	15034	18542	

Tabla 2.2: Tabla poblacional de EEUU. (ejercicio 8).

ii.- $H(t)/t$ es no decreciente para todo $t > 0$. Las distribuciones con esta propiedad suelen llamarse distribuciones IFRA, o con tasa de fallo creciente en promedio.

iii.- $m(t) \leq m(0)$ para todo $t \geq 0$. Si esto ocurre, suele decirse que las distribuciones tienen la propiedad NBU o "nueva mejor que usada".

iv.- $m(t)$ es una función no creciente para todo $t \geq 0$. En este caso se habla de distribuciones con vida residual media decreciente.

Definidas las propiedades anteriores, comprueba que:

a.- i $\Longrightarrow$ ii $\Longrightarrow$ iii

b.- i $\Longrightarrow$ iv $\Longrightarrow$ iii.

10(*).- Sean $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$ los estadísticos ordenados correspondientes a una muestra aleatoria de tamaño n de una variable con distribución exponencial de parámetro λ. Definimos las variables aleatorias:

$$\begin{aligned} W_1 &= nT_{(1)} \\ W_i &= (n-i+1)\left[T_{(i)} - T_{(i-1)}\right] \quad \text{para } i = 2, \ldots, n \end{aligned}$$

i.- Comprueba que cada variable W_i sigue una distribución exponencial de parámetro λ.

ii.- Comprueba que las variables W_i son independientes.

11.- Considera la distribución Gamma de parámetros λ y γ,

i.- Comprueba que la esperanza y la varianza de esta distribución son,

$$\begin{aligned} E[T] &= \gamma/\lambda \\ Var[T] &= \gamma/\lambda^2 \end{aligned}$$

ii.- Comprueba que la vida media residual de esta distribución verifica,

$$\lim_{t\to\infty} m(t) = \lambda^{-1}.$$

iii(*).- Comprueba que la suma de n variables exponenciales independientes de parámetro λ sigue una distribución Gamma de parámetros n y λ. Verifica que si T sigue un distribución $Erlang(n, \lambda)$ se cumple,

$$P(T \geq t) = P(Y_{\lambda t} < n)$$

donde $Y_{\lambda t}$ sigue una distribución Poisson de media λt. Para demostrarlo, comprueba que,

$$\int_t^\infty f(x)dx = \sum_{i=0}^{n-1} \frac{(\lambda t)^i \exp(-\lambda t)}{i!}$$

donde $f(x)$ es la función de densidad de la variable $Erlang(n, l)$.

iv(*).- Verifica que una distribución $Gamma(n/2, 1/2)$ coincide con una distribución χ^2_n.

12.- Dada una variable con distribución Lognormal de parámetros μ y σ,

i.- Comprueba que la función de densidad es,

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right].$$

ii.- Comprueba que la media y la varianza de esta distribución son,

$$\begin{aligned} E(T) &= \exp\left(\mu + \sigma^2/2\right) \\ Var(T) &= [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2). \end{aligned}$$

13.- Sea T un tiempo de vida con una distribución Log-logística. La variable $Y = \ln(T)$ tendrá una distribución Logística cuya función de densidad es,

$$f(y) = \frac{\exp[(y-\mu)/\sigma]}{(1 + \exp[(y-\mu)/\sigma])^2}\sigma^{-1} \qquad -\infty < y < \infty.$$

i(*).- Comprueba que la función generatriz de momentos de $W = (Y - \mu)/\sigma$ es,

$$M(\theta) = E[\exp(\theta W)] = \Gamma(1+\theta)\Gamma(1-\theta)$$

y que su media y su varianza son 0 y $\pi^2/3$, respectivamente. Calcula a partir de este resultado la media y la varianza de Y.

ii.- Comprueba que la función de supervivencia de $T = \exp(Y)$ es,

$$S(t) = \left(1 + e^\theta t^\kappa\right)^{-1}$$

donde $\theta = -\mu/\sigma$ y $\kappa = \sigma^{-1}$.

iii.- Comprueba que si $\kappa \leq 1$, la función de riesgo de la variable T es monótona decreciente y que si $\kappa > 1$, tiene un comportamiento como el de la de la distribución Lognormal; es decir, $h(0) = 0$, $h(t)$ aumenta hasta un valor máximo y decrece después y que $\lim_{t\to\infty} h(t) = 0$.

Capítulo 3

Estimación no paramétrica de la supervivencia: análisis de una muestra

3.1 Introducción

El problema principal en los problemas de Fiabilidad y Análisis de Supervivencia es la estimación de la función de supervivencia $S(t)$. Esta función es la base para estimar la mayor parte de las funciones y parámetros de interés en el análisis del tiempo de vida.

Si la muestra no contiene observaciones censuradas, la función de supervivencia se estima mediante la **función de supervivencia empírica**, fse, definida como,

$$\hat{S}(t) = \frac{N^o \text{ de individuos con tiempo de supervivencia mayor que t}}{N^o \text{ de individuos de la muestra}}$$

Este estimador es una función no creciente, toma el valor 1 en todo instante anterior al tiempo de fallo más pequeño y 0 a partir del máximo tiempo de fallo observado; la función permanece constante entre dos instantes de fallo consecutivos y presenta un salto descendente en cada tiempo de fallo observado. Si no hay empates en la muestra, todos los saltos de la función son de altura $1/n$, mientras que si se observan d tiempos de vida iguales a t_i, el salto de $\hat{S}(t)$ en ese instante será de altura d/n.

Cuando en la muestra existen observaciones censuradas la fse no es un estimador

adecuado porque tiende a subestimar la función de supervivencia. En efecto, el utilizar este estimador es equivalente a considerar que todos los individuos censurados fallan en el instante de censura. Dado que es posible que alguno de los individuos con tiempo de censura menor que t esté vivo en el instante t, será necesario introducir alguna modificación en el estimador para evitar ese sesgo.

Los métodos estadísticos para estimar los parámetros y funciones de la distribución del tiempo de vida se clasifican en **paramétricos** y **no paramétricos** según se basen, o no, en hipótesis específicas sobre la familia a la que pertenece la distribución de T. Las técnicas no paramétricas, que utilizan menos hipótesis, se emplean preferentemente en las primeras fases del estudio cuando se tiene poca información sobre el comportamiento del fenómeno; los resultados obtenidos en estos análisis ayudan a determinar qué distribución de probabilidad representa mejor los datos observados.

En este capítulo se estudian dos procedimientos no paramétricos de estimación a partir de una muestra homogénea: la estimación mediante **tablas de vida**, que se basa en datos agrupados en intervalos, y el **estimador producto-límite** o estimador de **Kaplan-Meier** de la función de supervivencia, que requiere observaciones individuales. En el último apartado se presenta un procedimiento gráfico para seleccionar un modelo paramétrico a partir de una estimación no paramétrica de $S(t)$.

3.2 Tablas de vida

Las tablas de vida son un procedimiento clásico para describir la mortalidad que experimenta una población. Este método, cuyo origen se atribuye a Halley (1693), sigue siendo una herramienta muy utilizada en campos como la demografía o los seguros de vida. El objetivo de una tabla de vida es expresar el patrón de mortalidad que experimenta un colectivo de individuos en unas condiciones dadas. Distinguiremos dos tipos de tablas: las **tablas poblacionales**, que son una herramienta de carácter fundamentalmente descriptivo, y las **tablas de vida clínicas**, que tienen una estructura análoga a la de las anteriores y sirven para estimar la supervivencia de una población a partir de una muestra.

3.2.1 Tablas de vida poblacionales

El objetivo de este tipo de tablas es describir y establecer previsiones sobre la mortalidad de una población. Una tabla de vida poblacional se puede construir utilizando dos procedimientos; el más intuitivo consiste en considerar una **cohorte**, esto es, un conjunto de personas representativo de la población de interés al que se hace un seguimiento temporal. Por ejemplo, para estudiar el patrón de mortalidad de los españoles nacidos en la década de los 70, una posible cohorte podría estar constituida por 10.000 niños nacidos durante el año 1975. Fijada la cohorte, se hace un seguimiento de cada uno de sus miembros, comprobando si vive o ha fallecido. El seguimiento se hace a intervalos de tiempo prefijados, año a año por ejemplo, hasta que se registra el fallecimiento del último superviviente. Las tablas de vida poblacionales así construidas se denominan **tablas de cohorte**.

El inconveniente del procedimiento anterior estriba, obviamente, en que requiere un periodo de seguimiento muy largo; por este motivo, las tablas de vida más habituales son las denominadas **tablas actuales**. En las tablas de vida actuales la cohorte es ficticia y su evolución se obtiene aplicando a un grupo hipotético de individuos las tasas de mortalidad específicas de cada edad. Estas tasas se calculan con los datos estadísticos registrados y a partir de ellas se calculan las restantes funciones de las tablas de vida poblacionales. Las tablas así construidas muestran un patrón de mortalidad ficticio, que no ha sido experimentado por ninguna cohorte real sino que es el resultado de combinar lo observado en cohortes diferentes. Este procedimiento tiene la ventaja de ser rápido y poco costoso y es adecuado si las fuentes estadísticas disponibles son fiables. Los resultados serán válidos mientras no se produzcan cambios significativos en las condiciones de vida de la población respecto a la situación a la que corresponden los datos utilizados.

Las fuentes estadísticas necesarias para calcular las tasas de mortalidad por edades son:

- Los datos censales del número de personas vivas de cada edad en un cierto periodo, por ejemplo un año, medidos en el punto medio del intervalo de tiempo considerado.

- Las estadísticas del número de muertes, por edades, en el mismo periodo.

Los procedimientos para estimar las tasas de mortalidad a partir de estos datos

se pueden consultar en el capítulo cuarto del libro de Elandt-Johnson y Johnson (1980).

Funciones básicas de una tabla de vida poblacional

${}_tq_x$: **probabilidad condicional de muerte** en el intervalo $(x, x+t)$ dado que el individuo alcanza la edad x. Es la función básica de la tabla y se estima a partir de las fuentes estadísticas. Habitualmente se utiliza como periodo t un año, en cuyo caso suele suprimirse el índice t de la notación, o cinco años.

l_x: **número esperado de supervivientes** a la edad x de los l_0 considerados inicialmente. Esta función, que expresa la supervivencia de la población, es una función continua de x, pero en la práctica suele tabularse sólo para los valores enteros de x y algunos valores fraccionarios menores que uno. El valor l_0 se denomina **base de la tabla de vida** y suele tomarse igual a 100.000 ó 1.000.000.

De acuerdo con su definición, los sucesivos valores de l_x pueden calcularse de forma recursiva como el producto del número esperado de supervivientes al comienzo del intervalo anterior por la tasa de supervivencia en dicho intervalo,

$$l_x = l_{x-t}\,(1 -_t q_{x-t})\,.$$

Reiterando este cálculo, se puede establecer la relación de l_x con la base l_0,

$$l_{nt} = l_0\,(1 -_t q_0)\,(1 -_t q_t)\,(1 -_t q_{2t}) \ldots \left(1 -_t q_{(n-1)t}\right).$$

P_x: **proporción esperada de supervivientes** con edad x,

$$P_x = \frac{l_x}{l_0}.$$

${}_td_x$: **número esperado de fallecimientos** con edad comprendida en el intervalo $(x, x+t)$,

$${}_td_x = l_x - l_{x+t} = l_{xt}q_x.$$

${}_tL_x$: **número total esperado de años vividos** entre las edades x y $x+t$. Este valor es la suma total de años que las l_x personas que alcanzan la edad x esperan vivir durante el intervalo $(x, x+t)$. Formalmente se define como,

$${}_tL_x = \int_x^{x+t} l_y dy.$$

Los individuos de la cohorte que alcanzan la edad $x+t$, contribuyen a ${}_tL_x$ con la longitud total del intervalo t, aquéllos que no alcanzan la edad x no aportan nada y el resto, los que mueren con una edad comprendida entre x y $x+t$, contribuyen sólo con una fracción de t, cuyo valor no se conoce exactamente. En la práctica, con la excepción del primer intervalo, se suele suponer que las muertes se producen de modo uniforme en $(x, x+t)$, y se estima que la contribución a ${}_tL_x$ de cada una de esas personas es $t/2$. Así se tiene,

$$ {}_tL_x = t\left(l_{x+t} + \frac{1}{2}\ {}_td_x\right) = \frac{t}{2}\left(l_x + l_{x+t}\right). $$

En el primer intervalo esta hipótesis no es plausible ya que los fallecimientos suelen concentrarse en la etapa inicial. Las soluciones habituales a este problema son dos; una subdividir el periodo en intervalos de tiempo más pequeños, y otra considerar que los individuos que mueren en dicho intervalo aportan una contribución menor a ${}_tL_x$.

T_x: **número total esperado de años vividos con edad superior a** x por las personas vivas a esa edad. Este valor se obtiene aplicando la idea que lleva a definir la función ${}_tL_x$ al intervalo (x, ∞). ${}_tL_x$ y T_{x+t} verifican las relaciones siguientes,

$$ \begin{aligned} T_x &= {}_tL_x + {}_tL_{x+t} + {}_tL_{x+2t} + \ldots \\ T_x &= {}_tL_x + T_{x+t}. \end{aligned} $$

Se llama **población estacionaria** a una población hipotética que satisface las tres condiciones siguientes:

- Experimenta el patrón de mortalidad que se refleja en la tabla de vida.
- En cada periodo considerado nacen, uniformemente distribuidos, l_0 individuos.
- Ha transcurrido el tiempo necesario para que alcance el equilibrio.

En estas condiciones, en cada periodo se incorporan y desaparecen l_0 individuos de la población, por lo que el tamaño de la población y su composición de edades será constante. En la población estacionaria, l_x es el número de individuos con edad mayor que x, ${}_tL_x$ representa el número de personas vivas con edades en el intervalo $(x, x+t)$ y T_x es el número de personas con más de x años de edad.

e_x: **esperanza del tiempo restante de vida** de un individuo a la edad x. Esta función se calcula como,

$$e_x = \frac{T_x}{l_x}$$

y es utilizada para calcular las primas y los beneficios de los seguros de vida, así como para comparar la supervivencia de poblaciones.

Notemos para finalizar que, en una tabla de vida, las funciones,

- l_x hace referencia a la edad exacta x.

- ${}_tq_x$, ${}_td_x$, ${}_tL_x$ corresponden a intervalos de tiempo de la forma $(x, x+t)$

- T_x y e_x corresponden a intervalos del tipo (x, ∞).

Todas estas funciones se tabulan en un formato normalizado que se muestra en la tabla 3.1 Se trata de una tabla correspondiente al conjunto de la población de los EE.UU. de América para el periodo 1979-81 construida por el *National center for Health Statistics*. Las funciones de esa tabla poblacional actual se han calculado a partir de los datos del censo del año 1980 y de las estadísticas de defunciones de los años 1979-81. Como se ha señalado anteriormente, se describe con más detalle la mortalidad del primer año de vida

En la figura 3.1 se representa la evolución temporal de las funciones l_x y ${}_tq_x$, calculadas en la tabla anterior. Se observa cómo l_x es una curva que decrece muy rápidamente en el intervalo inmediatamente posterior al nacimiento, que se mantiene casi constante hasta la edad de 50 años y que posteriormente decrece cada vez más rápidamente. La gráfica de ${}_tq_x$ muestra también el comportamiento correspondiente a esa evolución.

(t_x, t_{x+1})	${}_tq_x$	l_x	${}_td_x$	${}_tL_x$	T_x	e_x
0-1 días	0.00463	100000	463	273	7387758	73.88
1-7	0.00246	99537	245	1635	7387485	74.22
7-28	0.00139	99292	138	5708	7385850	74.38
28-365	0.00418	99154	414	91357	7380142	74.43
0-1 años	0.01260	100000	1260	98973	7387758	73.88
1-2	0.00093	98740	92	98694	7288785	73.82
2-3	0.00065	98648	64	98617	7190091	72.89
3-4	0.00050	98584	49	98560	7091474	71.93
4-5	0.00040	98535	40	98515	6992914	70.97
5-6	0.00037	98495	36	98477	6894399	70.00
6-7	0.00033	98459	33	98442	6795922	69.02
7-8	0.00030	98426	30	98412	6697480	68.05
8-9	0.00027	98396	26	98383	6599068	67.07
9-10	0.00023	98370	23	98358	6500685	66.08
10-11	0.00020	98347	19	98338	6402327	65.10
11-12	0.00019	98328	19	98319	6303989	64.11
12-13	0.00025	98309	24	98297	6205670	63.12
13-14	0.00037	98285	37	98266	6107373	62.14
14-15	0.00053	98248	52	98222	6009107	61.16
15-16	0.00069	98196	67	98163	5910885	60.19
16-17	0.00083	98129	82	98087	5812722	59.24
17-18	0.00095	98047	94	98000	5714635	58.28
18-19	0.00105	97953	102	97902	5616635	57.34
19-20	0.00112	97851	110	97796	5518733	56.40
20-21	0.00120	97741	118	97682	5420937	55.46
21-22	0.00127	97623	124	97561	5323255	54.53
22-23	0.00132	97499	129	97435	5225694	53.60
23-24	0.00134	97370	130	97306	5128259	52.67
24-25	0.00133	97240	130	97175	5030953	51.74
25-26	0.00132	97110	128	97046	4933778	50.81
26-27	0.00131	96982	126	96919	4836732	49.67
27-28	0.00130	96856	126	96793	4739813	48.94
28-29	0.00130	96730	126	96667	4643020	48.00
29-30	0.00131	96604	127	96541	4546353	47.06
30-31	0.00133	96477	127	96414	4449812	46.12
31-32	0.00134	96350	130	96284	4353398	45.18
32-33	0.00137	96220	132	96155	4257114	44.24
33-34	0.00142	96088	137	96019	4160959	43.30
34-35	0.00150	95951	143	95880	4064940	42.36

(t_x, t_{x+1})	${}_tq_x$	l_x	${}_td_x$	${}_tL_x$	T_x	e_x
35-36	0.00159	95808	153	95731	3969060	41.43
36-37	0.00170	95655	163	95574	3873329	40.49
37-38	0.00183	95492	175	95404	3777755	39.56
38-39	0.00197	95317	188	95224	3682351	38.63
39-40	0.00213	95129	203	95027	3587127	37.71
40-41	0.00232	94926	220	94817	3492100	36.79
41-42	0.00254	94706	241	94585	3397283	35.87
42-43	0.00274	94465	264	94334	3302698	34.96
43-44	0.00306	94201	288	94057	3208364	34.06
44-45	0.00335	93913	314	93756	3114307	33.16
45-46	0.00356	93599	343	93427	3020551	32.27
46-47	0.00401	93256	374	93069	2927124	31.39
47-48	0.00442	92882	410	92677	2834055	30.51
48-49	0.00488	92472	451	92246	2741378	29.65
49-50	0.00538	92021	495	91773	2649132	28.79
50-51	0.00589	91526	540	91256	2557359	27.94
51-52	0.00642	90986	584	90695	2466103	27.10
52-53	0.00699	90402	631	90086	2375408	26.28
53-54	0.00761	89771	684	89430	2285322	25.46
54-55	0.00830	99087	739	88717	2195892	24.65
55-56	0.00902	88348	797	87950	2107175	23.85
56-57	0.00978	87551	856	87122	2019225	23.06
57-58	0.01059	86695	919	86236	1932103	22.29
58-59	0.01151	85776	987	85283	1845867	21.52
59-60	0.01254	84789	1063	84258	1760584	20.76
60-61	0.01368	83726	1145	83153	1676326	20.02
61-62	0.01493	82581	1233	81965	1593173	19.29
62-63	0.01628	81348	1324	80686	1511208	18.58
63-64	0.01767	80024	1415	79316	1430522	17.88
64-65	0.01911	78609	1502	77859	l351206	17.19
65-66	0.02059	77107	1587	76314	1273347	16.51
66-67	0.02216	75520	1674	74683	1197033	15.85
67-68	0.02389	73846	1764	72964	1122350	15.20
68-69	0.02585	72082	1864	71150	1049386	14.56
69-70	0.02806	70218	1970	69233	978236	13.93
70-71	0.03052	68248	2083	67206	909003	13.32
71-72	0.03315	66165	2193	65069	841797	12.72
72-73	0.03593	63972	2299	62823	776728	12.14
73-74	0.03882	61673	2394	60476	713905	11.58
74-75	0.04184	59279	2480	58039	653429	11.02

(t_x, t_{x+1})	${}_tq_x$	l_x	${}_td_x$	${}_tL_x$	T_x	e_x
75-76	0.04507	56799	2560	55520	595390	10.48
76-77	0.04867	54239	2640	52919	539870	9.95
77-78	0.05274	51599	2721	50238	486951	9.44
78-79	0.05742	48878	2807	47475	436713	8.93
79-80	0.06277	46071	2891	44626	389238	8.45
80-81	0.06882	43180	2972	41694	344612	7.98
81-82	0.07552	40208	3036	38689	302918	7.53
82-83	0.08278	37172	3077	35634	264229	7.11
83-84	0.09041	34095	3083	32553	228595	6.70
84-85	0.09842	31012	3052	29486	196042	6.32
85-86	0.10725	27960	2999	26461	166556	5.96
86-87	0.11712	24961	2923	23500	140095	5.61
87-88	0.12717	22038	2803	20636	116595	5.29
88-89	0.13708	19235	2637	17917	95959	4.99
89-90	0.14728	16598	2444	15376	78042	4.70
90-91	0.15868	14154	2246	13031	62666	4.43
91-92	0.17169	11908	2045	10886	49635	4.17
92-93	0.18570	9863	1831	8948	38749	3.93
93-94	0.20023	8032	1608	7228	29801	3.71
94-95	0.21495	6424	1381	5733	22573	3.51
95-96	0.22976	5043	1159	4463	16840	3.34
96-97	0.24338	3884	945	3412	12377	3.15
97-98	0.25637	2939	754	2562	8965	3.05
98-99	0.26868	2185	587	1892	6403	2.93
99-100	0.28030	1598	448	1374	4511	2.82
100-101	0.29120	1150	335	983	3137	2.73
101-102	0.30139	815	245	692	2154	2.64
102-103	0.31089	570	177	481	1462	2.57
103-104	0.31970	393	126	330	981	2.50
104-105	0.32786	267	88	223	651	2.44
105-106	0.33539	179	60	150	428	2.38
106-107	0.34233	119	41	99	378	2.33
107-108	0.34870	78	27	64	179	2.29
108-109	0.35453	51	18	42	115	2.25
109-110	0.35988	33	12	27	73	2.20

Tabla 3.1: Tabla poblacional de EEUU correspondiente a 1979-81.

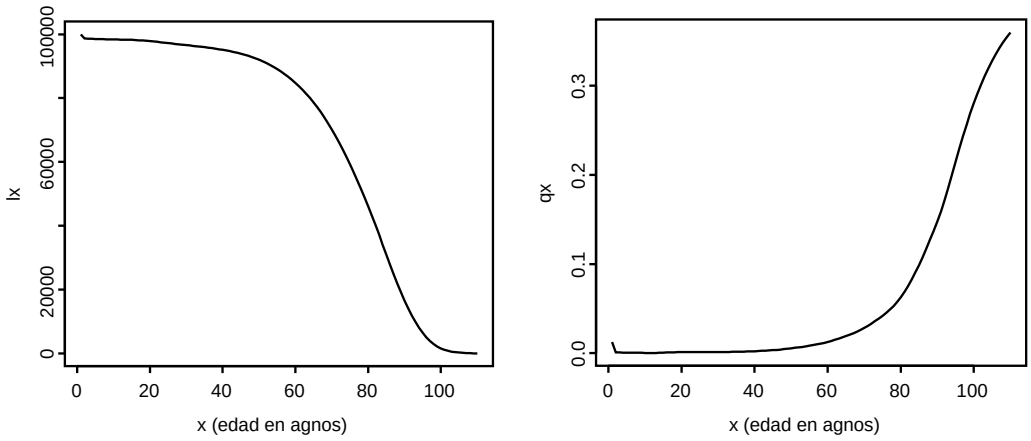

Figura 3.1: Representación de l_x y ${}_tq_x$ de la tabla 3.1.

Cuestión: Si una persona, cuya supervivencia se corresponde con el patrón expresado en la tabla 3.1, tiene ahora 30 años de edad,

i.- ¿Cuál es su edad esperada de fallecimiento?

ii.- ¿Cuál es la probabilidad de morir a la edad de 44 años en esa población? ¿ y la de las personas que tienen 30 años?

3.2.2 Tablas de vida clínicas

Las tablas de vida clínicas son un procedimiento de estimación de la supervivencia de una población desarrollado a partir de las tablas de vida poblacionales. En este caso no se analiza una cohorte sino un conjunto de datos procedente de ensayos clínicos en los que se realiza el seguimiento de una muestra homogénea de individuos. El procedimiento de estimación de estas tablas es similar al utilizado en las tablas de frecuencias, con la única diferencia de que en este caso pueden existir datos censurados en la muestra. La estimación se realiza a partir de datos agrupados; esta agrupación puede estar originada por:

- La forma de seguimiento de de los individuos: a partir de un mismo suceso inicial se realizan controles de la evolución de los individuos -si han fallado, sobreviven o no se tiene información sobre ellos- en intervalos de tiempo preestablecidos, iguales para todos. El protocolo, la frecuencia y forma de seguimiento de los individuos, determina los intervalos, $[t_0, t_1)$, $[t_1, t_2), \ldots, [t_s, t_{s+1})$, que se utilizarán en el análisis y que no tienen por qué ser de la misma longitud.

- La agrupación en intervalos de las observaciones individuales. En la actualidad, dada la facilidad de medios de cálculo, no suele estar justificado agrupar las observaciones individuales, porque implica una pérdida de información y el establecimiento de los intervalos introduce un cierto grado de arbitrariedad.

El objetivo principal de una tabla de vida clínica es la estimación de las funciones de supervivencia, riesgo y densidad del tiempo de vida T. Para facilitar esta estimación las tablas presentan una estructura normalizada que incluye, en distintas columnas, la siguiente información:

$[t_{i-1}, t_i)$: identifica los extremos del intervalo de tiempo i-ésimo. El extremo inferior del primer intervalo, t_0, suele ser 0; si el ensayo no termina hasta que se observe el fallo de todos los individuos, el extremo superior del último intervalo es, $t_{s+1} = \infty$.

t_{mi}: punto medio del intervalo i-ésimo. Esta columna se incluye debido a que la estimación de las funciones de riesgo y de densidad se realiza en esos puntos.

b_i: anchura del intervalo i-ésimo. Este valor se utiliza en la estimación de la función de densidad. Las funciones de densidad y riesgo no se pueden estimar en el último intervalo cuando éste es de longitud infinita.

l_i: *(lost)* número de individuos que abandonan el experimento y cuyo tiempo de censura, pertenece al intervalo i-ésimo.

w_i: *(withdrawn)* número de individuos cuya respuesta o fallo no se ha observado en el momento de finalizar el ensayo y que en ese instante llevan en el estudio un tiempo comprendido entre t_{i-1} y t_i. La información sobre el tiempo de vida de estos individuos es también censurada.

El tratamiento en el proceso de estimación de ambos tipos de observaciones censuradas, l_i e w_i, es idéntico; por ello, en algunas tablas no se hace la distinción anterior y se define una única columna,

c_i: *(censored)* número de individuos cuyo tiempo de observación, que es censurado, toma un valor comprendido entre t_{i-1} y t_i . Evidentemente,

$$c_i = l_i + w_i.$$

d_i: *(dead)* número de individuos cuyo tiempo de fallo observado pertenece al intervalo $[t_{i-1}, t_i)$.

n'_i: número de individuos que permanecen vivos y en el estudio al comienzo del intervalo i-ésimo. El valor n'_1 es el número total de individuos que comenzaron el ensayo; los restantes valores se calculan mediante la relación,

$$n'_i = n'_{i-1} - d_{i-1} - c_{i-1}.$$

n_i: número estimado de individuos expuestos a riesgo durante el intervalo i-ésimo. Si en el intervalo no hay observaciones censuradas, $n_i = n'_i$; en otro caso, la hipótesis usual es suponer que las observaciones censuradas ocurridas durante el intervalo se distribuyen en él uniformemente. Por ello, en media, los individuos con observación censurada están expuestos a riesgo durante la mitad de la duración del intervalo y, en consecuencia,

$$n_i = n'_i - \frac{1}{2}c_i.$$

$\hat{q}_i$: proporción de fallo en el intervalo i-ésimo. Este valor estima la probabilidad de fallo en ese intervalo, condicionada a que el individuo no había fallado al comienzo del mismo. Si en el intervalo no hay observaciones censuradas, el estimador natural de esta probabilidad es d_i/n'_i; si las hay, este estimador tiende a subestimar dicha probabilidad, ya que es posible que alguno de los individuos censurados en el intervalo muera antes de finalizar el mismo. Por esta causa, es necesario realizar algún tipo de ajuste; la alternativa habitual consiste en sustituir n'_i por el número estimado de individuos expuestos a riesgo, n_i,

$$\hat{q}_i = \frac{d_i}{n_i}.$$

Si el último intervalo es infinito, $\hat{q}_{s+1} = 1$.

La validez de este ajuste, en cierta medida arbitrario, depende de las características de los procesos de censura y de fallo. Breslow y Crowley (1974) establecieron que, bajo el esquema de censura aleatoria, este estimador era inconsistente y sesgado. Sin embargo, cuando la proporción de censura no es excesiva y se distribuye de forma homogénea, si los intervalos no son muy amplios y los valores de n_i no son demasiado pequeños, el comportamiento de este estimador es aceptable.

$\hat{p}_i$: proporción de supervivencia en el intervalo i-ésimo. Este valor estima la probabilidad de que un individuo sobreviva al instante t_i, dado que estaba vivo al inicio del intervalo i-ésimo. Se define,

$$\hat{p}_i = 1 - \hat{q}_i.$$

$\hat{P}(t_i)$: proporción de supervivencia en el instante t_i del ensayo. Este valor estima $P(T \geq t_i)$. Si un individuo sobrevive hasta el inicio del intervalo (i+1)-ésimo implica que dado que ha sobrevivido hasta el comienzo del intervalo i-ésimo, no falla durante el intervalo i-ésimo; así,

$$\hat{P}(t_i) = \hat{p}_i \hat{P}(t_{i-1}).$$

Aplicando reiteradamente esta relación y que $\hat{P}(t_0)$ es igual a 1, se obtiene,

$$\hat{P}(t_i) = \hat{p}_i \hat{p}_{i-1} \dots \hat{p}_1.$$

Este estimador de la función de supervivencia se llama estimador actuarial. En los instantes que no son extremo de un intervalo, es habitual calcular su valor mediante interpolación lineal entre los valores de los correspondientes extremos.

$\hat{f}(t_{mi})$: tasa o proporción de fallo en el intervalo i-ésimo por unidad de tiempo. Este valor estima la función de densidad de T en el punto medio del intervalo.

$$\hat{f}(t_{mi}) = \frac{\hat{P}(t_{i-1}) - \hat{P}(t_i)}{b_i} = \frac{\hat{P}(t_{i-1})\hat{q}_i}{b_i}.$$

$\hat{h}(t_{mi})$: tasa instantánea condicional de fallo correspondiente al punto medio del intervalo i-ésimo. Con este valor se estima la función de riesgo en t_{mi},

$$\hat{h}(t_{mi}) = \frac{\hat{f}(t_{mi})}{\hat{P}(t_{mi})}.$$

Cuestión: Comprueba que la estimación de la función de riesgo también puede expresarse como,

$$\hat{h}(t_{mi}) = \frac{2\hat{q}_i}{b_i(1+\hat{p}_i)} = \frac{d_i}{b_i(n_i - d_i/2)}.$$

Notas:

i.- Los datos imprescindibles para la estimación de las funciones definidas en una tabla de vida clínica son, d_i, c_i, y n'_1. Notemos que los resultados obtenidos

con este método de estimación son sensibles a la elección de los intervalos de tiempo con los que se construye la tabla. En general, el número de intervalos debe depender de la cantidad de datos disponibles y del objetivo del análisis. Como regla general, es conveniente utilizar al menos 8 ó 10 intervalos.

ii.- Como se comentó en el primer capítulo, la validez de éste y otros métodos de estimación exige que la distribución del tiempo de fallo de todos los individuos, censurados y no censurados, sea la misma.

En la tabla 3.3 se muestra una tabla de vida clínica, Parker *et al.* (1946), correspondiente a 2418 hombres enfermos de angina de pecho. El tiempo de supervivencia mide, en años, el tiempo transcurrido desde el instante del diagnóstico hasta el fallecimiento del enfermo. En la figura 3.2 se han representado las estimaciones de las funciones de supervivencia, densidad y riesgo. En la gráfica de la función de riesgo se observa que una vez superado el primer año, que presenta la tasa de fallo más alta, ésta permanece prácticamente constante hasta el décimo año; en ese instante vuelve a aumentar durante un intervalo de tres años. El comportamiento final podría no reflejar una característica real de la población, ya que en la cola de la distribución, al disminuir el número de datos, la estimación es menos fiable. Por esta razón es conveniente disponer de una medida de la precisión de las estimaciones. Algunas de estas medidas se presentan en el siguiente apartado.

Cuestión: Completa las casillas de la tabla de vida clínica 3.2:

i	t_{m_i}	b_i	n_i'	l_i	n_i	d_i	$\hat{p}_i$	$\hat{P}_i$	$\hat{f}(t_{mi})$	$\hat{h}(t_{mi})$
[0,1)			1000	10		22				
[1,2)				8		19				
[2,3)				21		26				
[3,4)				26		31				
[4,5)				37		27				
[5,∞)				-		773				

Tabla 3.2: Tabla de vida clínica.

3.2.3 Precisión de las estimaciones

Los valores de $\hat{q}_i$, $\hat{p}_i$ y $\hat{P}(t_i)$ son estimaciones sujetas a la variabilidad inherente al proceso de muestreo, por lo que deben completarse con información relativa a su precisión. Bajo determinadas hipótesis sobre los mecanismos de censura es posible,

Año	t_{m_i}	b_i	l_i	w_i	d_i	n'_i	n_i	$\hat{q}_i$	$\hat{p}_i$
0	0.5	1.0	0	0	456	2418	2418.0	0.1886	0.8114
1	1.5	1.0	39	0	226	1962	1942.5	0.1163	0.8837
2	2.5	1.0	22	0	152	1697	1686.0	0.0902	0.9098
3	3.5	1.0	23	0	171	1523	1511.5	0.1131	0.8869
4	4.5	1.0	24	0	135	1329	1317.0	0.1025	0.8975
5	5.5	1.0	107	0	125	1170	1116.5	0.1120	0.8880
6	6.5	1.0	133	0	83	938	871.5	0.0952	0.9048
7	7.5	1.0	102	0	74	722	671.0	0.1103	0.8897
8	8.5	1.0	68	0	51	546	512.0	0.0996	0.9004
9	9.5	1.0	64	0	42	427	395.0	0.1063	0.8937
10	10.5	1.0	45	0	43	321	298.5	0.1441	0.8559
11	11.5	1.0	53	0	34	233	206.5	0.1646	0.8354
12	12.5	1.0	33	0	18	146	129.5	0.1390	0.8610
13	13.5	1.0	27	0	9	95	81.5	0.1104	0.8896
14	14.5	1.0	23	0	6	59	47.5	0.1263	0.8737
15	-	-	0	0	0	30	30.0	1.0000	0.0000

Año	$\hat{P}(t_i)$	$\hat{f}(t_{m_i})$	$\hat{h}(t_{m_i})$	s.e.$[\hat{S}(t_i)]$	s.e.$[\hat{f}(t_{m_i})]$	s.e.$[\hat{h}(t_{m_i})]$	$\hat{t}_{mr_i}$	s.e.$[\hat{t}_{mr_i}]$
0	1.0000	0.1886	0.2082	-	0.0080	0.0097	5.33	0.17
1	0.8114	0.0944	0.1235	0.0080	0.0060	0.0082	6.25	0.20
2	0.7170	0.0646	0.0944	0.0092	0.0051	0.0076	6.34	0.24
3	0.6524	0.0738	0.1199	0.0097	0.0054	0.0092	6.23	0.24
4	0.5786	0.0593	0.1080	0.0101	0.0049	0.0093	6.22	0.19
5	0.5193	0.0581	0.1186	0.0103	0.0050	0.0106	5.91	0.18
6	0.4611	0.0439	0.1000	0.0104	0.0047	0.0110	5.60	0.19
7	0.4172	0.0460	0.1167	0.0105	0.0052	0.0135	5.17	0.27
8	0.3712	0.0370	0.1048	0.0106	0.0050	0.0147	4.94	0.28
9	0.3342	0.0355	0.1123	0.0107	0.0053	0.0173	4.83	0.41
10	0.2987	0.0430	0.1552	0.0109	0.0063	0.0236	4.69	0.42
11	0.2557	0.0421	0.1794	0.0111	0.0068	0.0306	4.00+	-
12	0.2136	0.0297	0.1494	0.0114	0.0067	0.0351	3.00+	-
13	0.1839	0.0203	0.1169	0.0118	0.0065	0.0389	2.00+	-
14	0.1636	0.0207	0.1348	0.0123	0.0080	0.0549	1.00+	-
15	0.1429	-	-	0.0133	-	-	-	-

Tabla 3.3: Tabla de vida clínica de los pacientes de angina de pecho.

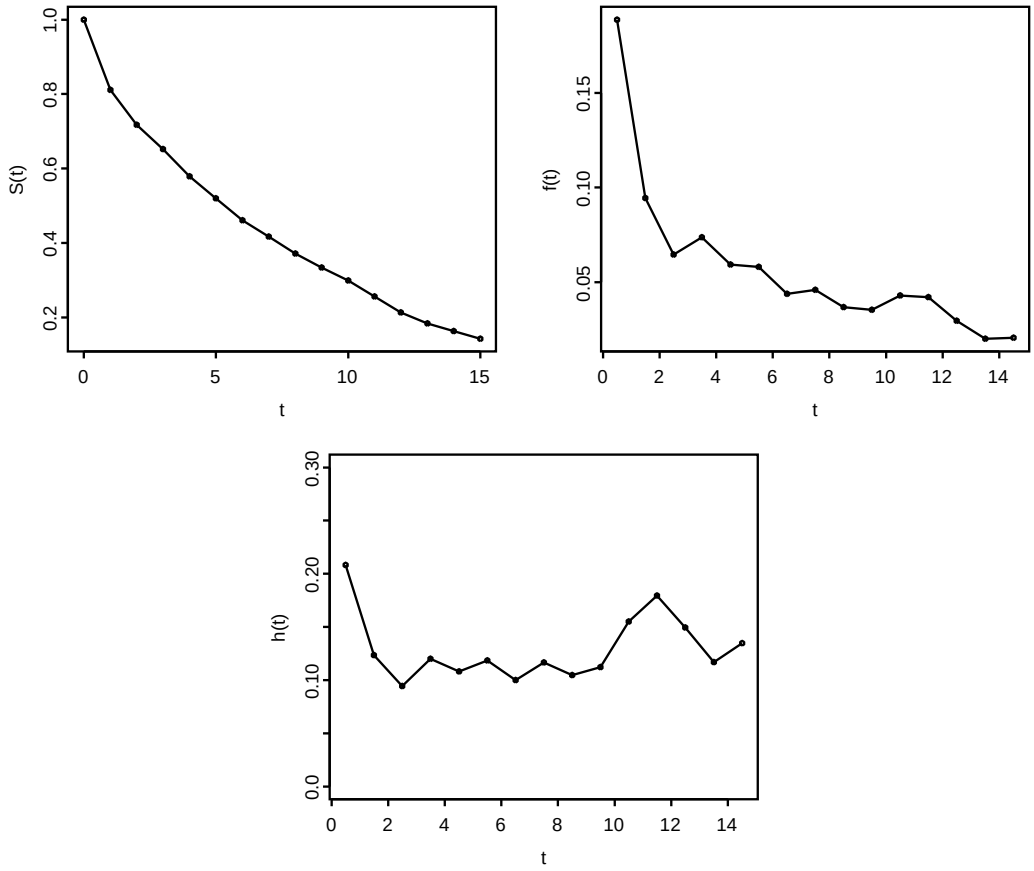

Figura 3.2: Estimación de las funciones características de los pacientes de angina de pecho.

aunque complicado, deducir estimaciones de sus varianzas. Por esta razón, aunque la metodología de las tablas de vida clínicas es antigua, el estudio teórico de las propiedades estadísticas de sus estimadores es reciente y está aún por completar. En este capítulo se presentan algunas de las propiedades y resultados más utilizados. La mayor parte de estos resultados se han obtenido para el caso de muestras completas, pero en se suelen generalizar y aplicar también al caso de muestras censuradas.

Fórmula de Greenwood La estimación más empleada de la varianza de $\hat{P}(t_j)$ es la propuesta por Greenwood en 1926,

$$Var[\hat{P}(t_j)] \approx [\hat{P}(t_j)]^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{p}_i n_i} = \left[\hat{P}(t_j)\right]^2 \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}.$$

Esta estimación, resultado de una aproximación asintótica que comentaremos más adelante, es razonable cuando el valor esperado de n_j no es demasiado pequeño y

requiere, si la proporción de censura en la muestra es importante, que el número de intervalos considerados no sea muy pequeño. La fórmula de Greenwood tiende a subestimar la varianza de $\hat{P}(t_j)$, especialmente en los intervalos de la cola derecha de la distribución donde el valor esperado de n_j suele ser pequeño. No obstante, en esos casos su cálculo no es adecuado ya que la distribución de $\hat{P}(t_j)$ suele ser muy sesgada y, en consecuencia, la varianza no es una buena medida de precisión de la estimación.

Justificación. La estimación actuarial de la función de supervivencia es, $\hat{P}(t_j) = \hat{p}_j\hat{p}_{j-1}\ldots\hat{p}_1$. Para obtener su varianza aproximada se aplica un procedimiento, denominado **método delta**, que consiste en calcular la varianza de la aproximación de primer orden de la función obtenida a partir de su desarrollo en serie de Taylor. En el caso de una función de una sola variable, se tiene,

$$g(X) \approx g(\theta) + g'(\theta)(X-\theta)$$

y la varianza aproximada es,

$$Var[g(X)] \approx \left[g'(\theta)\right]^2 Var(X)$$

donde θ es un parámetro tal que, asintóticamente, $E(X-\theta) = 0$. En el caso de una función de varias variables, aplicando un procedimiento análogo, se tiene,

$$Var[g(\mathbf{X})] \approx \sum_{i=1}^{j}\sum_{k=1}^{j} \frac{\partial g}{\partial \theta_i}\frac{\partial g}{\partial \theta_k} Cov(X_i, X_k) \tag{3.1}$$

donde $\frac{\partial g}{\partial \theta_i}$ denota $\frac{\partial g(x)}{\partial x_i}\,|_{x=\theta}$ y θ es un vector de parámetros tal que, asintóticamente, $E[X-\theta] = 0$. Generalmente, el vector θ no es conocido, por lo que las derivadas se evalúan en su estimador, $\hat{\theta}$.

Aplicando este método a la expresión de $\hat{P}(t_j)$, se obtiene como varianza aproximada,

$$Var\left[(P(t_j)\right] \approx \sum_{i=1}^{j}\sum_{k=1}^{j} \frac{\partial \hat{P}(t_j)}{\partial p_i}\frac{\partial \hat{P}(t_j)}{\partial p_k}\; Cov(\hat{p}_i, \hat{p}_k).$$

Para calcular esta expresión se necesita,

1. Estimadores de p_j y q_j. Los más utilizados son

$$\begin{aligned} \hat{p}_j &= (n_j - d_j)/n_j \\ \hat{q}_j &= d_j/n_j. \end{aligned}$$

En el caso sin censura, éstos son los estimadores máximo verosímiles; este resultado se obtiene utilizando que el vector $(d_1, d_2, \ldots, d_{s+1})$ sigue una distribución Multinomial de parámetros n_1, el número de individuos que inician el estudio, y $\pi_1, \pi_2, \ldots, \pi_s$, donde,

$$\begin{aligned} n_1 &= \sum_{i=1}^{s+1} d_i \\ \pi_j &= P(t_{j-1}) - P(t_j) = p_1 \ldots p_{j-1}(1-p_j) \qquad\qquad j = 1, 2, \ldots, s+1. \end{aligned}$$

2. Calculo de las derivadas parciales de la función de supervivencia respecto a cada componente del vector $p = (p_1, p_2, \ldots, p_s)$

$$\frac{\partial \hat{P}(t_j)}{\partial p_i} = \frac{P(t_j)}{p_i}.$$

Como los valores $P(t_j)$ y p_i no son conocidos, se sustituyen por sus estimaciones.

3. Cálculo de las varianzas y covarianzas de los estimadores $\hat{p}_i$. Presentamos los resultados en el caso de muestras sin observaciones censuradas ya que su justificación en el caso general resulta complicada.

Proposición 1: Bajo la hipótesis $n_j > 0$, d_j sigue una distribución Binomial de parámetros n_j y q_j; en consecuencia,

$$\begin{aligned} E[\hat{q}_j] &= q_j \\ E[\hat{p}_j] &= p_j \\ Var[\hat{q}_j] &= Var[\hat{p}_j] = p_j q_j E\left(\frac{1}{n_j}\right) \\ Cov(\hat{q}_i, \hat{q}_j) &= Cov(\hat{p}_i, \hat{p}_j) = 0 \qquad\qquad \text{con } i < j. \end{aligned}$$

En muestras con observaciones censuradas, los resultados anteriores no son ciertos aunque, asintóticamente, se verifica que $Cov[\hat{p}_i, \hat{p}_j] = 0$ y la estimación de $Var(\hat{p}_j)$ puede aproximarse mediante $\hat{p}_j \hat{q}_j / n_j$.

Sustituyendo los valores obtenidos en los puntos 1 y 2 en la expresión 3.1 de $\widehat{Var}[\hat{P}(t_j)]$ se obtiene la fórmula de Greenwood.

$$Var\left[(P(t_j)\right] \approx [\hat{P}(t_j)]^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{p}_i n_i} = \left[\hat{P}(t_j)\right]^2 \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}.$$

Notas:

i.- En el proceso de estimación de la varianza de $\hat{P}(t_j)$ se han utilizado varias aproximaciones e hipótesis. La aproximación de la varianza dada por el método delta, produce resultados razonables si n es suficientemente grande. La expresión de la varianza de $\hat{p}_i$ y la hipótesis de que $Cov(\hat{p}_i, \hat{p}_k] = 0$, adoptadas por analogía con el caso de las muestras sin censura, son más cuestionables, dependiendo su validez del mecanismo de censura así como de la distribución del tiempo de vida en el problema bajo estudio.

ii.- En el caso de una muestra sin datos censurados, es fácil comprobar que la fórmula de Greenwood produce la estimación

$$\widehat{Var}\left[\hat{P}(t_j)\right] = \frac{\hat{P}(t_j)\left[(1-\hat{P}(t_j)\right]}{n}.$$

Esta expresión es razonable ya que en este caso $\hat{P}(t_j) = \hat{p}_j\hat{p}_{j-1}\ldots\hat{p}_1 = \frac{n_{j+1}}{n}$ y la distribución de n_{j+1} es Binomial de parámetros n, $P(t_j)$.

iii.- Cuando la muestra es completa, los estimadores de p_j, q_j y $P(t_j)$ son insesgados; esta propiedad se pierde cuando en la muestra hay observaciones censuradas.

iv.- Desarrollando procesos de aproximación similares al anterior se pueden obtener estimaciones de la varianza para las funciones de densidad y riesgo,

$$\begin{aligned} Var\left[\hat{f}(t_{mi})\right] &\approx \left(\frac{\hat{P}(t_{i-1})\hat{q}_i}{b_i}\right)^2 \sum_{j=1}^{i-1}\left(\frac{\hat{q}_j}{n_j\hat{p}_j} + \frac{\hat{p}_i}{n_i\hat{q}_i}\right) \\ Var[\hat{h}(t_{mi})] &\approx \frac{\left[\hat{h}(t_{mi})\right]^2}{n_i\hat{q}_i}\left[1 - \left(\frac{1}{2}\hat{h}(t_{mi})b_i\right)^2\right]. \end{aligned}$$

Cuestión: Estima las varianzas de las estimaciones de la función de supervivencia, de densidad y de riesgo, para el segundo y el cuarto intervalo de la tabla propuesta en la cuestión del apartado anterior.

3.2.4 Estimación de la mediana y otras medidas relacionadas

La distribución del tiempo de supervivencia es habitualmente asimétrica y sesgada positivamente; por este motivo la mediana resulta más adecuada que la media como medida de posición central de la distribución.

La estimación de la mediana se puede obtener a partir de la función de supervivencia estimada. Recordemos que la mediana es el valor en el que el 50% de la población bajo estudio ha fallado, esto es, el valor $t_{0.5}$ tal que $S(t_{0.5}) = 0.5$. Utilizando la estimación de $S(t)$ obtenida en los instantes t_i de la muestra y haciendo la hipótesis de que, entre esos instantes, el estimador decrece linealmente, la mediana se obtiene mediante una simple interpolación lineal. El mismo procedimiento se aplica a la estimación del percentil t_p.

Denotaremos por $med(R_{t_i})$ a la mediana de la variable tiempo de vida restante en el instante t_i. Este parámetro, asociado a la supervivencia, se puede estimar por el procedimiento anterior ya que se caracteriza porque $S\left[med(R_{t_i})\right] = P(t_i)/2$. Una estimación de la varianza de este estimador es,

$$\widehat{Var}\left[\widehat{med}(R_{t_i})\right] \approx \frac{\hat{P}^2(t_i)}{4n_{i+1}\hat{f}^2(t_{mj})},$$

donde t_{mj} es el punto medio del intervalo $[t_{j-1}, t_j)$ para el que se verifica $\hat{P}(t_{j-1}) \geq \hat{P}(t_i)/2$ y $\hat{P}(t_j) < \hat{P}(t_i)/2$. Esta misma expresión sirve para estimar la varianza aproximada de $\hat{t}_{0.5}$ sin más que tener en cuenta que $\hat{t}_{0.5} = \widehat{med}(R_{t_0})$. En la tabla 3.3 de los pacientes de angina de pecho se muestra un ejemplo de la estimación de estos valores.

3.3 Estimador Kaplan-Meier de la función de supervivencia

El estimador no paramétrico más utilizado de la función de supervivencia, si se dispone de los tiempos de supervivencia individuales, es el **estimador producto-límite**, denominado también **estimador Kaplan-Meier**, KM, por ser estos autores los primeros en estudiar sus propiedades en 1958.

3.3.1 Definición del estimador KM

Consideremos una muestra de n individuos de los que se conoce su tiempo de fallo o el instante de censura. Supondremos que se han observado s, $s \leq n$, tiempos de fallo distintos que denotamos, una vez ordenados, $t_{(1)}$, $t_{(2)}$, $\ldots$, $t_{(s)}$. Es posible que en la muestra se produzcan **empates**, es decir, observaciones cuyo tiempo de

fallo es el mismo y por eso se define d_i, $(d_i \geq 1)$, como el número de fallos que se producen en el instante $t_{(i)}$. Las restantes observaciones, $n - \sum d_i$, son los tiempos de seguimiento de los individuos cuyo fallo no ha sido observado.

El estimador Kaplan-Meier de S(t) se define como,

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

donde n_i es el número de individuos en riesgo en el instante $t_{(i)}$; es decir, el número de individuos vivos y no censurados, justo antes de $t_{(i)}$. Si existe alguna observación censurada cuyo valor coincide con un tiempo de fallo, se hace la hipótesis de que la observación censurada ocurre inmediatamente después del tiempo de fallo y, en consecuencia, los individuos censurados en ese instante se contabilizan como individuos en riesgo.

El estimador KM es una función constante entre los tiempos de fallo consecutivos, que vale 1 antes del menor tiempo de fallo, $t_{(1)}$, y cuyo valor decrece según un factor variable en cada instante de fallo. El estimador no cambia en los tiempos de observación correspondientes a los individuos censurados, aunque esas observaciones influyen en el estimador a través de los valores n_i. Cuando el mayor de los tiempos observados en la muestra, t_M, es un tiempo de fallo, la estimación KM toma el valor cero a partir de ese instante; si t_M corresponde a una observación censurada, es habitual considerar que $\hat{S}(t)$ no está definido para $t > t_M$.

Justificación. La estructura de este estimador es similar a la del estimador actuarial de las tablas de vida clínicas, ya que se construye como un producto en el que cada término estima la probabilidad condicional de sobrevivir al tiempo de fallo $t_{(i)}$, dado que se ha sobrevivido hasta ese instante,

$$\hat{S}(t) = \prod_{i, t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \prod_{i, t_{(i)} \leq t} \hat{p}_i.$$

De hecho, el estimador KM se puede considerar como un caso límite del estimador actuarial, cuando el número de intervalos considerado tiende a infinito y su longitud, excepto la del último, se aproxima a cero. En efecto, sean t'_0 y t'_f los puntos inicial y final del periodo de seguimiento de un ensayo. Dividamos el intervalo $(t'_0, t'_f]$ en M intervalos, $(t'_0,\ t'_1],\ (t'_1,\ t'_2], \ldots,\ (t'_{M-1},\ t'_f]$, lo suficientemente

pequeños para que la probabilidad de que en alguno de ellos se produzca más de un tiempo observado distinto, se pueda considerar despreciable. El estimador de la probabilidad condicional de fallo en el intervalo $(t'_{j-1}, t'_j]$ es,

$$\hat{q}_j = \begin{cases} \frac{d_i}{n_i} & \text{si existe un tiempo de fallo } t_{(i)} \text{ en el intervalo } (t'_{j-1}, t'_j] \\ 0 & \text{en otro caso.} \end{cases}$$

Como $(t'_{j-1}, t'_j]$ puede ser arbitrariamente pequeño, se tiene que $\hat{q}_j = \frac{d_i}{n_i}$ sólo si t'_j es uno de los tiempos de fallo $t_{(i)}$. La probabilidad de supervivencia en cada intervalo es,

$$\hat{p}_j = 1 - \hat{q}_j = \begin{cases} \frac{n_i - d_i}{n_i} & \text{si } t'_{(j)} \text{ es un tiempo de fallo } t_{(i)} \\ 1 & \text{en otro caso.} \end{cases}$$

Al sustituir estas estimaciones en la expresión del estimador actuarial, dado que los intervalos en los que no se produce fallo no modifican el valor de la estimación, se obtiene el estimador producto límite,

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

Notas:

i.- El estimador KM de la función de supervivencia puede también deducirse como un estimador máximo verosímil generalizado. Posee buenas propiedades en el caso de muestras grandes y, bajo condiciones de censura bastante generales, es un estimador consistente de $S(t)$. Al igual que ocurría con los estimadores de las tablas de vida clínicas, el estudio de sus propiedades es una tarea compleja.

ii.- En una muestra sin observaciones censuradas, el estimador Kaplan-Meier coincide con el estimador natural de $S(t)$, la función de supervivencia empírica,

$$\hat{S}(t) = \frac{n - i}{n} \qquad \text{si } t_{(i)} \leq t < t_{(i+1)}.$$

iii.- El estimador KM admite una formulación alternativa, menos intuitiva pero más sencilla de calcular. Si $\tilde{t}_{(1)} \leq \tilde{t}_{(1)} \leq \ldots \leq \tilde{t}_{(n)}$, son los n tiempos de supervivencia observados en la muestra, ya sean tiempos de fallo u observaciones censuradas, dispuestos en orden creciente, el estimador KM se puede expresar como,

$$\hat{S}(t) = \prod_r \frac{n - r}{n - r + 1}$$

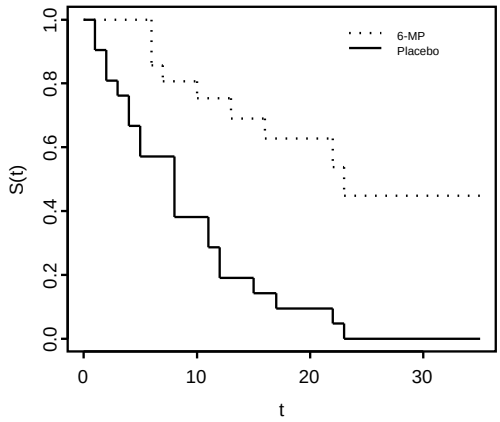

Figura 3.3: Estimación de $S(t)$ de los dos grupos del Ejemplo 3.

donde r recorre los enteros positivos tales que $\tilde{t}_{(r)} \leq t$, siendo $\tilde{t}_{(r)}$ un tiempo de fallo observado.

Cuestión: En un experimento clínico se han observado los siguientes tiempos de vida, en meses, de 10 pacientes: 1, 2, 3, 3*, 4, 4*, 5, 5*, 8, 9*. Calcula el estimador Kaplan-Meier de su función de supervivencia.

Cuestión: Comprueba la nota ii, es decir que en una muestra sin censura el estimador KM coincide con la función de supervivencia empírica.

Cuestión: Comprueba que la expresión del estimador KM dada en la nota iii es equivalente a la propuesta en la definición.

Ejemplo: En la figura 3.3 se han representado las estimaciones de la función de supervivencia correspondientes a los dos grupos del ensayo clínico de la droga 6-MP (Ejemplo 3 del capítulo 1). En la gráfica se aprecia claramente que la droga en ensayo parece ser eficaz en la prolongación del tiempo hasta el fallo de los pacientes.

3.3.2 Varianza del estimador Kaplan-Meier. Intervalos de confianza

Para establecer la fiabilidad de las estimaciones obtenidas con el estimador KM, se requiere tener una medida de su precisión. Se ha estudiado el comportamiento de distintos estimadores de la varianza de este estimador y se concluye que la mejor es

la que proporciona la fórmula de Greenwood,

$$\widehat{Var}\left[\hat{S}(t)\right] \approx \left[\hat{S}(t)\right]^2 \sum_{i,t_{(i)}\leq t} \frac{d_i}{n_i(n_i - di)} = \left[\hat{S}(t)\right]^2 \sum_r \frac{1}{(n-r)(n-r+1)},$$

expresión análoga a la correspondiente al estimador actuarial, que puede también obtenerse utilizando la teoría asintótica para estimadores máximo verosímiles.

Cuestión: Comprueba la validez de la segunda expresión, obtenida en base a la formulación alternativa del estimador KM hecha en la nota iii del apartado anterior.

Cuestión: Comprueba que en ausencia de censura, la expresión anterior se reduce a $\hat{S}(t)\left[1-\hat{S}(t)\right]/n$.

Intervalo de confianza basado en la normalidad Para analizar el comportamiento del estimador Kaplan-Meier es necesario especificar hipótesis sobre el mecanismo de censura. Breslow y Crowley (1974) establecieron, bajo la hipótesis de censura aleatoria, el carácter asintóticamente gaussiano del proceso estocástico

$$\left\{\sqrt{n}\left[\hat{S}(x) - S(x)\right]\right\}_{0<x<T} \qquad \text{con } T < \infty \text{ t.q. } S(T) > 0.$$

Utilizando la normalidad asintótica de $\hat{S}(x)$, se puede construir el siguiente intervalo de confianza aproximado para S(t), a un nivel del $100(1-\alpha)\%$,

$$\hat{S}(t) \pm z_{1-\alpha/2} \text{ s.e.}\left[\hat{S}(t)\right]$$

donde $z_{1-\alpha/2}$ es el cuantil correspondiente de la distribución Normal tipificada y el error estándar, s.e.$\left[\hat{S}(t)\right]$, se calcula mediante la fórmula de Greenwood.

Este intervalo de confianza no es demasiado satisfactorio en el caso de muestras pequeñas; además, si t es un valor extremo, puede incluir valores fuera del rango $(0,1)$. Una alternativa para evitar estas dificultades consiste en considerar una transformación biyectiva, g, de $S(t)$, que mejore la aproximación normal en muestras pequeñas y que evite las restricciones de rango. Tras calcular un intervalo de confianza para $g[S(t)]$, el intervalo para $S(t)$ se determina aplicando la transformación inversa, g^{-1}, a los límites del intervalo de confianza obtenido; la estimación de la varianza de $g\left[\hat{S}(x)\right]$ se obtiene aplicando el método delta. Algunas transformaciones adecuadas son $g(x) = \ln[-\ln(x)]$ y $g(x) = \arcsin\sqrt{x}$.

t_i	$\hat{S}\left(t_{(i)}\right)$	s.e.$\left[\hat{S}\left(t_{(i)}\right)\right]$	I.C. 95%
0	1.000	0.000	-
10	0.944	0.054	(0.839,1.000)
19	0.881	0.079	(0.727,1.000)
30	0.814	0.098	(0.622,1.000)
36	0.746	0.111	(0.529,0.963)
59	0.653	0.130	(0.397,0.908)
75	0.559	0.141	(0.283,0.836)
93	0.466	0.145	(0.182,0.751)
97	0.373	0.143	(0.093,0.653)
107	0.249	0.139	(0.000,0.522)

Tabla 3.4: Estimación de $S(t)$ y un intervalo de confianza al 95% de los tiempos de fallo del uso del DIU.

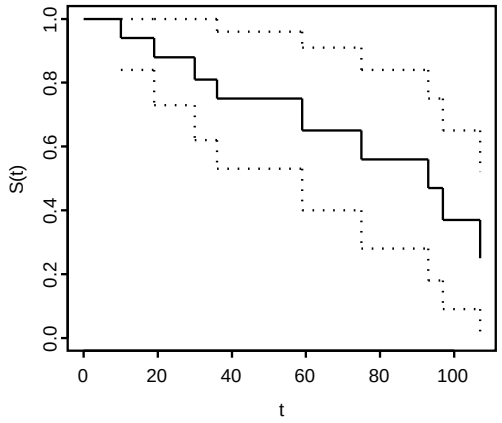

Figura 3.4: Estimación de $S(t)$ y un intervalo de confianza al 95% de los tiempos de fallo del uso del DIU.

Ejemplo: Los datos siguientes corresponden a una muestra de 18 mujeres y representan al número de semanas transcurrido entre la implantación de un dispositivo intrauterino (DIU) y su abandono debido a la aparición de alteraciones en la menstruación: 10, 13*, 18*, 19, 23*, 30, 36, 38*, 54*, 56*, 59, 75, 93, 97, 104*, 107, 107*, 107*. Las observaciones con asterisco corresponden a mujeres que dejaron de usar el DIU por otras razones.

En la tabla 3.4 se muestra la estimación de la función de supervivencia, el error estándar asociado y los extremos del intervalo de confianza al 95% para $S(t)$, con t en el intervalo $[t_{(i-1)}, t_{(i)})$. Los intervalos se han calculado utilizando la normalidad asintótica de $\hat{S}(t)$, reemplazando los límites fuera del rango $(0, 1)$ por 0 ó 1, en caso necesario. En la figura 3.4 se representa la estimación de la función de supervivencia

y el intervalo de confianza correspondiente. Se puede observar que la amplitud de los intervalos aumenta con el tiempo como consecuencia de la disminución del tamaño de muestra en que se basan las estimaciones.

Intervalo de confianza alternativo En el caso de una muestra sin observaciones censuradas de tamaño n, se puede construir un intervalo de confianza aproximado para $S(t)$ calculando las dos soluciones de la ecuación,

$$\frac{\hat{S}(t) - S(x)}{\sqrt{\frac{S(x)[1-S(x)]}{n}}} = z_{1-\alpha/2},$$

que tiene $S(x)$ como incógnita. Si en la muestra hay observaciones censuradas, Rothman (1978) propuso que se podría sustituir en la ecuación anterior n por n^*, el tamaño efectivo de muestra, valor que se calcula mediante la expresión,

$$n^* = \frac{1 - \hat{S}(x)}{\hat{S}(t) \sum_{j,t_{(j)} \leq x} \frac{d_j}{n_j(n_j - d_j)}},$$

que se obtiene al igualar la fórmula de Greenwood a la estimación de $Var[\hat{S}(t)]$ en el caso sin censura.

Algunos autores opinan que la definición anterior de tamaño muestral efectivo tiende a producir, en particular en la cola de la distribución, intervalos de confianza cuya probabilidad de cubrir S(t) es menor que el $100(1-\alpha)\%$ previsto. Para evitar este riesgo se suele utilizar otra definición del tamaño efectivo de muestra, más conservadora, propuesta por Peto (1977),

$$n^{**} = \frac{n_j}{\hat{S}\left(t_{(j)}^{-0}\right)} = \frac{n_j}{\hat{S}\left(t_{(j-1)}\right)},$$

siendo n_j el número de individuos en riesgo en $t_{(j)}$, el mayor tiempo de fallo menor o igual que el instante t considerado.

Cuestión: Con los datos del experimento clínico que aparecen en la primera cuestión del apartado anterior, calcular la estimación de la varianza de la función de supervivencia en el instante $t = 5$ meses.

3.3.3 Estimación de otras funciones y parámetros de interés

A partir del estimador KM se pueden obtener estimaciones de otras funciones de interés asociadas a la distribución, así como de parámetros de la misma.

Función de riesgo acumulado La función de riesgo acumulado se relaciona con $S(t)$ mediante la expresión $H(t) = -\ln S(t)$; en consecuencia, un estimador de dicha función es,

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

donde $\hat{S}(t)$ es el estimador KM de la función de supervivencia.

Otro estimador posible de $H(t)$ propuesto por Nelson-Aalen, NA, es la función de riesgo acumulado empírica,

$$\tilde{H}(t) = \sum_{j, t_{(j)} \leq t} \frac{d_j}{n_j},$$

que acumula las contribuciones d_j/n_j de la función de riesgo en los sucesivos instantes de fallo $t_{(j)}$. La mejor estimación de la varianza de este estimador es,

$$\widehat{Var}\tilde{H}(t) = \sum_{j, t_{(j)} \leq t} \frac{d_j}{n_j^2}.$$

A partir de este estimador de H(t) es posible obtener un estimador alternativo de la función de supervivencia, denominado estimador de Nelson-Aalen, utilizando la relación,

$$\tilde{S}(t) = \exp\left[-\tilde{H}(t)\right].$$

Si T es una variable continua, $\hat{H}(t)$ y $\tilde{H}(t)$ son dos estimadores asintóticamente equivalentes y, salvo para valores altos de t, donde las estimaciones son más inestables, la diferencia entre ambos será, por lo general, pequeña. En realidad, $\tilde{H}(t)$ es la aproximación lineal de primer orden de la función $\hat{H}(t)$. Desde el punto de vista teórico no hay argumentos para preferir un estimador al otro; el estimador $\tilde{H}(t)$ tiene la ventaja de su sencillez de cálculo.

Interesa señalar la utilidad de la función de riesgo acumulado para caracterizar la distribución del tiempo de vida. Resulta más fácil analizar el comportamiento de la tasa de fallo de la distribución, representando gráficamente $\hat{H}(t)$ o $\tilde{H}(t)$ frente

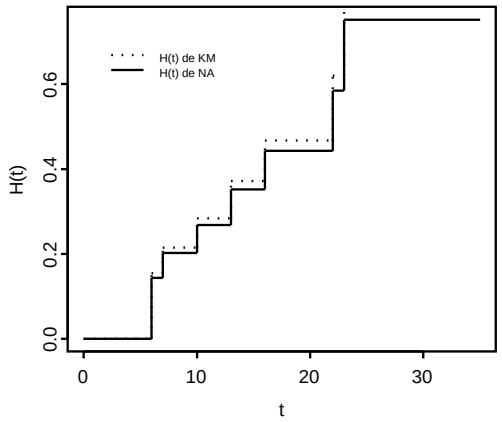

Figura 3.5: Estimación del riesgo acumulado por el procedimiento de Kaplan-Meier y el de Nelson-Aalen del grupo tratado con la droga 6-MP del Ejemplo 3.

al tiempo, que representando $\hat{S}(t)$, ya que al ser la función de riesgo la derivada de $H(t)$, se tiene que:

- Si la distribución tiene tasa de fallo constante, $H(t)$ es una función lineal en t.

- Si la distribución es IFR, $H(t)$ es una función convexa.

- Si la distribución es DFR, $H(t)$ es una función cóncava.

En la figura 3.3 se muestra la gráfica de los estimadores $\hat{H}(t)$ y $\tilde{H}(t)$ para el grupo de control del ensayo de la droga 6-MP. Como se ve, las dos estimaciones son bastante próximas, y sólo se separan apreciablemente al crecer t. Ambos estimadores sugieren el carácter exponencial de la distribución de T, dado el aspecto aproximadamente lineal de las dos gráficas.

Tiempo medio de vida Dado que la esperanza del tiempo de vida coincide con el área comprendida entre los ejes y la curva $S(t)$, un posible estimador de $E[T]$ es,

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt.$$

Su cálculo no resulta complicado dado que $\hat{S}(t)$ es una función constante a trozos.

Cuando el máximo tiempo de supervivencia observado en la muestra, t_M, corresponde a una observación censurada, la integral anterior no puede calcularse, al no estar definido $\hat{S}(t)$ para $t > t_M$. En este caso resulta más conveniente estimar la media del tiempo de vida limitado, o restringido, al instante L, $\mu_L = E[\min(T, L)]$.

Tomando $L = t_M$, μ_{t_M} puede ser una buena aproximación del valor medio de T si $P(T > t_M)$ es pequeña. Se puede comprobar que,

$$\hat{\mu}_{t_M} = \int_0^{t_M} \hat{S}(t)dt.$$

Cuestión: Comprueba que cuando no hay observaciones censuradas, el estimador $\hat{\mu}$ es la media muestral.

Un estimador de la varianza de $\hat{\mu}$ es,

$$\widehat{Var}(\hat{\mu}) = \sum_r \frac{A_r^2}{(n-r)(n-r+1)}$$

donde el índice r recorre los enteros positivos tales que la observación $t_{(r)}$, de la muestra ordenada, no es censurada, y A_r es el área bajo la curva $\hat{S}(t)$ a la derecha de $t_(r)$. Este es un estimador sesgado por lo que es habitual -así lo hacen muchos paquetes estadísticos- corregir su sesgo multiplicándolo por un factor $n_f/(n_f - 1)$, siendo n_f el número de observaciones no censuradas de la muestra.

Percentiles y su varianza La estimación de la mediana o cualquier percentil, t_p, de la distribución es el menor tiempo de fallo observado $t_{(j)}$ tal que, $\hat{S}\left[t_{(j)}\right] \leq 1 - p$. Se puede obtener una estimación de la varianza de ese estimador aplicando el método delta al cálculo de $\widehat{Var}\left[\hat{S}(\hat{t}_p)\right]$. Utilizando que $S'(t) = -f(t)$ y reordenando términos se tiene,

$$\widehat{Var}\left[\hat{t}_p\right] = \left(\frac{1}{\hat{f}(\hat{t}_p)}\right)^2 \widehat{Var}\left[\hat{S}(\hat{t}_p)\right].$$

Utilizando la normalidad asintótica de $\hat{t}_p$ y calculando con la fórmula anterior s.e.$\left(\hat{t}_p\right)$ se obtiene un intervalo de confianza para t_p.

Intervalo de Brookmeyer Brookmeyer y Crowley propusieron otro procedimiento para construir un intervalo de confianza para el percentil de orden p, t_p, que no requiere estimar la función de densidad. Se define un intervalo de confianza aproximado para t_p, al nivel $100(1-\alpha)\%$, como el conjunto de valores, t, que satisfacen,

$$\frac{\left|\hat{S}(t) - (1-p)\right|}{\text{s.e.}\left[\hat{S}(t)\right]} \leq z_{1-\alpha/2},$$

donde el error estándar de $\hat{S}(t)$ puede calcularse a partir de la fórmula de Greenwood y $z_{1-\alpha/2}$ es el cuantil correspondiente de la distribución normal tipificada. Este procedimiento de construcción se basa en contrastar, para cada t, a un nivel α, la hipótesis nula $H_0 : t_p = t$, frente a la alternativa $H1 : t_p \neq t$, utilizando la normalidad asintótica del estimador Kaplan-Meier. Los valores t para los que no se rechaza la hipótesis nula son los que forman el intervalo de confianza para t_p.

3.4 Selección de un modelo paramétrico

Si se quiere proponer un modelo paramétrico para una muestra de tiempos de supervivencia, cuando no se tiene conocimiento a priori sobre la distribución del tiempo de fallo, se deberán estudiar posibles distribuciones y elegir la que parezca más adecuada. El gran número de distribuciones existentes dificulta esta tarea, por lo que conviene disponer de algunos criterios que permitan realizar una primera selección de forma rápida y simple; estos modelos se someterán posteriormente a un análisis más detallado.

Se debe tener en cuenta que en muchas ocasiones no se dispone de un tamaño de muestra suficiente para realizar la selección a partir de un análisis puramente empírico. En esta situación se pueden utilizar otros criterios prácticos, por ejemplo,

1. La forma de la función de riesgo. Hemos visto que un método simple de mostrar la forma genérica de esa función, consiste en dibujar la estimación de la función de riesgo acumulado frente a los valores observados, y analizar si la gráfica corresponde a la de una distribución IFR (gráfica convexa), DFR (gráfica cóncava) o al modelo Exponencial (crecimiento lineal).

2. La adecuación del modelo a los datos en el rango de valores de mayor interés. El comportamiento de la función de supervivencia en los tiempos pequeños, por ejemplo, suele ser de especial interés en las aplicaciones industriales; mientras que en la mayor parte de los estudios de carácter médico, suele tener más importancia su comportamiento en los tiempos grandes.

3. La disponibilidad de expresiones sencillas de las funciones de riesgo, supervivencia y densidad. Este criterio llevaría por ejemplo a preferir, en condiciones similares, el modelo Weibull al Gamma o el Log-logístico al Lognormal.

3.4.1 Análisis gráfico de la adecuación de una distribución

Antes de hacer inferencia basada en una hipótesis sobre la forma de la distribución, es necesario contrastar si esa hipótesis es adecuada a los datos observados. Para ello proponemos un procedimiento basado en la comparación de una estimación no paramétrica de $S(t)$ y la función de supervivencia del modelo elegido. Esta comparación puede hacerse gráficamente dibujando en un diagrama bidimensional $g\left[\hat{S}(t)\right]$ frente a $h(t)$. Las funciones g y h se eligen de acuerdo con el modelo que se quiere analizar, de forma que, si éste es adecuado, la nube de puntos se disponga, aproximadamente, según una línea recta.

Supongamos que se dispone de una muestra de tiempos de supervivencia y que se quiere analizar la hipótesis de que siguen una distribución Exponencial. Si efectivamente los datos satisfacen esta hipótesis, su función de supervivencia es,

$$S(t) = \exp(-\lambda t).$$

Tomando logaritmos en esa expresión y sustituyendo S(t) por su estimación Kaplan-Meier, se tiene,

$$\ln\left[\hat{S}(t)\right] = -\lambda t.$$

De esta relación se deduce que, si la hipótesis sobre el carácter Exponencial es aceptable, el gráfico de $\ln\left[\hat{S}(t)\right]$ frente a t deberá ser, aproximadamente, lineal. Este análisis exploratorio nos permite además obtener una estimación preliminar del parámetro λ: la pendiente de la recta ajustada por mínimos cuadrados a la nube de puntos representada en el gráfico.

Cuestión: Utilizando un procedimiento similar, encontrar el procedimiento para analizar la adecuación de un modelo Weibull a un conjunto de datos.

Para verificar la adecuación de la distribución Lognormal se puede analizar la relación lineal entre $\Phi^{-1}\left[1-\hat{S}(t)\right]$ y $\ln(t)$, siendo Φ la función de distribución de una variable Normal tipificada.

En el caso de la distribución Log-logística, de la expresión de su función de supervivencia, se deduce que,

$$\ln\left[\frac{1-S(t)}{S(t)}\right] = \theta + \kappa \ln(t),$$

por lo que bastará analizar la relación lineal entre $\frac{\ln([1-\hat{S}(t)]}{\hat{S}(t))}$ y $\ln(t)$.

En el caso de la distribución Gamma no existe una relación exacta pero se pueden utilizar relaciones basadas en aproximaciones a la distribución Normal; por ejemplo, si la distribución de T es Gamma, $\Phi^{-1}\left[1-\hat{S}(t)\right]$ debe ser, aproximadamente, una función lineal de $\sqrt{t}$, y también de $t^{1/3}$.

Otros procedimientos gráficos que ayudan en la elección de una distribución adecuada, desarrollados en el segundo capítulo de Cox y Oakes (1984), se basan en el estudio del coeficiente de variación o en la relación entre el coeficiente estandarizado de tercer orden y el coeficiente de variación.

3.5 Ejercicios

1.- En cierta clínica fueron tratados, entre los años 1944 y 1953, 388 pacientes de melanoma maligno. La tabla 3.5 muestra la información relativa a los tiempos de supervivencia de estos enfermos, agrupados en intervalos de un año de longitud. La tabla proporciona para cada intervalo: el número de personas vivas y en tratamiento al comienzo del mismo n_i', el número de personas que murieron a causa de la enfermedad d_i y el de las que abandonaron el tratamiento por alguna causa, l_i, durante ese periodo

n_i'	388	219	173	127	108	91	79	71	66	59
d_i	167	45	45	19	17	11	8	5	6	7
l_i	2	1	1	0	0	1	0	0	1	0

Tabla 3.5: Tabla de la supervivencia de pacientes de melanoma maligno (Ejercicio 1).

i.- Construye una tabla de vida clínica a partir de los datos anteriores, calculando en cada intervalo: el n^o de personas expuestas a riesgo, la proporción de muertes y las estimaciones de las funciones de supervivencia, densidad y riesgo.

ii.- Interpreta el significado de la función de riesgo estimada. ¿En qué momento el pronóstico de un paciente de esas características es mejor? ¿En cuál es peor?

iii.- Calcula una estimación de la mediana del tiempo restante de vida para un enfermo en el momento de iniciar el tratamiento, para uno que sigue vivo

tras recibirlo durante dos años y para otro que ha sobrevivido cuatro años. Interpreta los resultados obtenidos.

2.- El mieloma múltiple es una enfermedad mortal que se caracteriza por la acumulación de células enfermas en la médula. Los datos de la tabla 3.6 proceden de un estudio realizado con el fin de establecer la posible asociación entre la supervivencia y el sexo de los pacientes.

La variable tiempo de supervivencia mide, en meses, el tiempo transcurrido desde que se diagnosticó la enfermedad al paciente hasta su fallecimiento, la variable Estado indica el tipo de observación (0 observación censurada, 1 muerte causada por la enfermedad) y la variable Sexo toma el valor 1 en los hombres y 2 en las mujeres.

i.- Tomando como límite inferior de los intervalos los puntos: 0, 6.5, 11.5, 19.5 y 60, construye dos tablas de vida, una para los hombres y otra para las mujeres. En cada una de ellas, calcula para cada intervalo, el n^o de personas expuestas a riesgo, la proporción de muertes y las estimaciones de las funciones de supervivencia, densidad y riesgo en los puntos medios de los intervalos.

ii.- Estima el percentil 75 del tiempo de vida correspondiente a los hombres. Estima la mediana del tiempo restante de vida para una mujer a la que se acaba de detectar la enfermedad y para otra que lleva diagnosticada y en tratamiento 11.5 meses.

iii.- ¿A partir de qué característica se puede determinar fácilmente el pronóstico de supervivencia que tiene un paciente en cada momento? Compara los pronósticos obtenidos para los hombres y para las mujeres y comenta los resultados.

T. Sup.	Estado	Sexo	T. Sup.	Estado	Sexo
1.25	1	1	25	1	1
1.25	1	1	26	1	2
2	1	1	32	1	1
2	1	1	35	1	1
2	1	1	37	1	1
3	1	2	41	1	1
5	1	2	41	1	2
5	1	1	51	1	1
6	1	1	52	1	2
6	1	2	54	1	1
6	1	1	58	1	2
6	1	2	66	1	1
7	1	1	67	1	1
7	1	2	88	1	2
7	1	2	89	1	1
9	1	1	92	1	2
11	1	1	4	0	1
11	1	1	4	0	2
11	1	1	7	0	2
11	1	1	7	0	1
11	1	2	8	0	2
13	1	2	12	0	2
14	1	1	11	0	1
15	1	1	12	0	2
16	1	1	13	0	2
16	1	2	16	0	1
17	1	1	19	0	2
17	1	1	19	0	2
18	1	2	41	0	1
19	1	1	53	0	1
19	1	2	57	0	1
24	1	2	77	0	1

Tabla 3.6: Tabla de la supervivencia de pacientes de mieloma múltiple (Ejercicio 2).

3.- En un experimento realizado para comparar dos tratamientos se observó la supervivencia de 20 animales de laboratorio sometidos a un tratamiento estándar, grupo A, y 20 sometidos a un tratamiento experimental, grupo B, obteniéndose los siguientes datos:

A: 1, 1, 3*, 4, 5*, 8, 8 10, 13, 14, 15*, 18, 20, 23, 25, 25*, 39, 45, 49*, 56*.

B: 3, 4, 4, 6, 6, 6*, 9, 10*, 11, 13, 20, 21, 22, 22, 24, 31*, 36, 42*, 55, 68.

El tiempo de supervivencia se ha medido en días. Las observaciones marcadas con un asterisco indican el número de días que estuvieron sometidos a tratamiento los animales que no habían muerto cuando el experimento finalizó. Responde, para cada uno de los grupos, a las siguientes cuestiones:

i.- Calcula el estimador Kaplan-Meier de la función de supervivencia y represéntalo gráficamente.

ii.- Estima la mediana de la distribución y el tiempo medio de vida. Obtén un intervalo de confianza al 95% para la mediana por el procedimiento estándar y por el método de Brookmeyer.

iii.- Estima las probabilidades de que un animal sobreviva más de 8, 50 y 60 días, respectivamente. Construye un intervalo de confianza al 95% para la probabilidad de sobrevivir más de 8 días.

4.- Se define el tiempo medio de vida restringido a un tiempo especificado L, como,

$$\mu_L = \int_0^L S(x)d(x).$$

i.- Comprueba que $\mu_L = E[\min(T, L)]$, donde T representa el tiempo de vida.

ii.- Comprueba que si $L \to \infty$, se obtiene el tiempo medio de vida $E[T]$.

5.- La tabla siguiente muestra los tiempos de supervivencia, en meses, de un conjunto de pacientes con la enfermedad de Hodgkin que participaron en el ensayo de un nuevo medicamento. Los enfermos están clasificados en dos grupos, A y B, dependiendo de si habían recibido terapia, o no, con anterioridad a su entrada al estudio.

A: 1.25, 1.41, 4.98, 5.25, 5.38, 6.92, 8.89, 10.98, 11.18, 13.11, 13.21, 16.33, 19.77, 21.08, 21.84*, 22.07, 31.38*, 32.62*, 37.18*, 42.92.

B: 1.05, 2.92, 3.61, 4.20, 4.49, 6.72, 7.31, 9.08, 9.11, 14.49*, 16.85, 18.82*, 26.59*, 30.26*, 41.34*.

i.- Calcula el estimador de Kaplan-Meier de la función de supervivencia para cada uno de los grupos anteriores y compáralos. ¿Se aprecia alguna diferencia en la probabilidad de sobrevivir un año entre los dos tipos de enfermos?

ii.- Estima la media del tiempo de vida en el grupo A y la mediana en el B.

iii.- Estima la probabilidad de que un individuo del grupo A sobreviva más de 5 meses y calcula un intervalo de confianza para ese valor.

iv.- Representa las estimaciones de la función de riesgo acumulado $H(t)$ en cada grupo. Utiliza esos gráficos para examinar y comparar las dos distribuciones de vida.

6.- Con los datos del Ejemplo 3 del primer capítulo, relativos al ensayo de la droga 6-MP, contesta a las siguientes cuestiones:

i.- Calcula y representa gráficamente el estimador de Kaplan-Meier de $S(t)$ en el grupo tratado con la droga 6-MP y en el grupo de control.

ii.- Calcula la varianza de la estimación de la función de supervivencia en el instante $t = 10$ en el grupo tratado con 6-MP y en el instante $t = 3$ en el grupo de control.

iii.- Compara las estimaciones de la mediana del tiempo de remisión en los dos grupos.

7.- Los datos de la tabla 3.7 son los tiempos de vida, en meses, de un grupo de 121 pacientes con cáncer de pecho, que fueron tratadas durante el periodo 1929-38. Utilizando un paquete estadístico, responde a las siguientes cuestiones,

i.- Calcula el estimador KM de la función de supervivencia. Estima las probabilidades de sobrevivir un año y cinco años y calcula la varianza aproximada de estas estimaciones.

ii.- Agrupa los datos en intervalos de un año de longitud y construye una tabla de vida clínica. Compara las estimaciones de la probabilidad de sobrevivir un año y cinco años calculadas con los datos agrupados, con las obtenidas en el apartado anterior.

0.3	0.3*	4.0*	5.0	5.6	6.2	6.3	6.6	6.8	7.4*	7.5
8.4	8.4	10.3	11.0	11.8	12.2	12.3	13.5	14.4	14.4	14.8
15.5*	15.7	16.2	16.3	16.5	16.8	17.2	17.3	17.5	17.9	19.8
20.4	20.9	21.0	21.0	21.1	23.0	23.4*	23.6	24.0	24.0	27.9
28.2	29.1	30	31	31	32	35	35	37*	37*	37*
38	38*	38*	39*	39*	40	40*	40*	41	41	41*
42	43*	43*	43*	44	45*	45*	46*	46*	47*	48
49*	51	51	51*	52	54	55*	56	57*	58*	59*
60	60*	60*	61*	62*	65*	65*	67*	67*	68*	69*
78	80	83*	88*	89	90	93*	96*	103*	105*	109*
109*	111*	115*	117*	125*	126	127*	129*	129*	139*	154*

Tabla 3.7: Tiempos de vida de enfermos de cáncer de pecho (Ejercicio 7).

8.- Un empleado es encargado durante 5 días de transcribir los datos médicos correspondientes a un conjunto de pacientes. El número de datos correctamente transcritos entre cada dos fallos cometidos es el siguiente: 73, 12, 40, 65, 100, 15, 70, 40, 110, 64, 200, 6, 90, 102, 20, 102, 90, 34*. Si se supone que la tasa de fallo del empleado no varía de forma apreciable durante los 5 días de trabajo, ¿qué conclusiones obtienes al analizar gráficamente esta información?

9.- En un experimento sobre la duración de un modelo de reloj despertador realizado con una muestra de 12 unidades, se obtuvieron 11 tiempos de respuesta, medidos en meses: 30.5, 33, 33, 36, 42, 55, 55.5, 76, 76, 106, 106, y una observación censurada igual a 107.5 meses. Comprueba gráficamente la hipótesis de que la duración de los despertadores sigue una distribución Weibull.

i.- ¿Proporciona esta distribución un ajuste adecuado a los datos?

ii.- Ajusta por mínimos cuadrados una recta a las observaciones del gráfico y obtén estimaciones de los parámetros de forma y escala de la distribución.

iii.- Comenta la naturaleza, creciente o decreciente en el tiempo, de la tasa de fallo de este producto.

iv.- Estima la mediana del tiempo de vida de este modelo de despertador mediante un procedimiento paramétrico basado en que T sigue un modelo Weibull.

v.- Analiza gráficamente si es adecuado suponer que la variable tiempo hasta el fallo sigue una distribución Gamma. ¿Qué hipótesis, Weibull o Gamma, se ajusta mejor a los datos?

10.- Con los datos del Ejemplo 1 del primer capítulo sobre el comportamiento de un elemento aislante sometido a distintos voltajes, dibuja los gráficos de adecuación de una distribución Weibull a los tiempos de fallo correspondientes a dos de los voltajes.

i.- ¿Es plausible la hipótesis de que los tiempos de fallo en ambos casos siguen una distribución Weibull con el mismo parámetro de forma γ?

ii.- Calcula mediante mínimos cuadrados los estimadores de los parámetros del modelo Weibull y compara la estimación de la función de supervivencia obtenida a partir de ellos, con la función de supervivencia empírica.

11.- En la siguiente lista se muestra la presión necesaria, en kg/cm^3, para romper 20 cubos de $1dm.$ de lado de cierto material: 94.9, 106.9, 229.7, 275.7, 144.5, 112.8, 159.3, 153.1, 270.6, 322.0, 216.4, 544.6, 266.2, 263.6, 138.5, 79.0, 114.6, 66.1, 131.2, 91.1.

i.- Dibuja un gráfico para analizar la hipótesis de que los datos provienen de una distribución Lognormal. Haz una estimación preliminar de la media y la desviación típica de dicha distribución.

ii.- Dibuja un gráfico para comprobar la adecuación de la distribución Log-logística a los mismos datos. ¿Qué distribución proporciona, a tu juicio, un mejor ajuste?

12.- Se ponen en funcionamiento, simultáneamente, 58 ventiladores. Suponemos que su tiempo hasta el fallo sigue una distribución exponencial de media 28.700 horas.

i.- Predice el número de ventiladores que fallarán durante las próximas 2000 horas si, cuando un ventilador falla, éste es sustituido inmediatamente por otro ventilador de un nuevo modelo que, se supone, no falla nunca.

ii.- Calcula una cota superior para el número de fallos que pueden observarse con probabilidad 0.90.

iii.- Responde a los apartados i y ii suponiendo ahora que cuando se produce un fallo, el ventilador es sustituido inmediatamente por otro del mismo modelo.

Capítulo 4

Análisis comparativo de la supervivencia: Métodos no paramétricos

4.1 Introducción

Un problema frecuente en los estudios de Fiabilidad y Análisis de Supervivencia es la comparación de la supervivencia de dos o más poblaciones que se diferencian en alguna característica, por ejemplo, el proceso de fabricación, las condiciones de uso o, en el campo biomédico, el tratamiento aplicado. En los problemas de Fiabilidad industrial es frecuente realizar este análisis comparativo utilizando procedimientos paramétricos; en los ensayos de tipo médico sin embargo, no se suele disponer de la información suficiente para formular hipótesis sobre la forma de la función de supervivencia, por lo que es más habitual la aplicación de métodos no paramétricos. En este capítulo se presentan algunos de los tests no paramétricos más utilizados.

4.2 Test log-rank para dos muestras

Supongamos que se quiere comparar la supervivencia en dos poblaciones de individuos, A y B, y que se dispone de una muestra de cada población de tamaño n_A y n_B respectivamente; sea $n = n_A + n_B$ el tamaño de la muestra combinada y $t_{(1)} < \ldots < t_{(j)} < \ldots < t_{(J)}$, los J tiempos de fallo distintos observados en la muestra conjunta ordenados en forma creciente. Denotaremos por d_{Aj} y d_{Bj} el número de

Grupo	N^o fallos en $t_{(j)}$	N^o individuos vivos en $t_{(j)}$	N^o individuos en riesgo en $t_{(j)}$
A	d_{Aj}	$n_{Aj} - d_{Aj}$	n_{Aj}
B	d_{Bj}	$n_{Bj} - d_{Bj}$	n_{Bj}
Total	d_j	$n_j - d_j$	n_j

Tabla 4.1: Tabla correspondiente al instante $t_{(j)}$ para comparar dos grupos.

fallos ocurridos en el instante $t_{(j)}$ en cada muestra y por d_j el número total de fallos observados en ese instante, es decir, $d_j = d_{Aj} + d_{Bj}$. Análogamente, denotaremos por n_{Aj} y n_{Bj} el número de individuos en riesgo en cada muestra justo antes del instante $t_{(j)}$, y por n_j la suma $n_{Aj} + n_{Bj}$. Toda esta información se puede disponer en un conjunto de J tablas de contingencia 2×2, una para cada tiempo de fallo. La tabla correspondiente al instante $t_{(j)}$ se muestra en la tabla 4.1.

La hipótesis a contrastar es que las funciones de supervivencia en ambas poblaciones coinciden en el intervalo de tiempo observado; esto es,

$$H_0 : S_A(t) = S_B(t) \qquad \forall t \leq \tau$$

o, equivalentemente,

$$H_0 : h_A(t) = h_B(t) \qquad \forall t \leq \tau.$$

donde, en general, τ se toma igual al mayor tiempo de supervivencia observado. Para contrastar esta hipótesis consideraremos la discrepancia que se observa entre los individuos que fallan en cada uno de los grupos en cada instante de fallo, y el correspondiente número esperado de fallos bajo H_0. Dada la historia de fallo y censura hasta el instante $t_{(j)}$, es decir conocidos los valores marginales n_{Aj}, n_{Bj} y d_j, los cuatro frecuencias observadas en la tabla correspondiente al instante $t_{(j)}$, quedan completamente determinadas dada una de ellas, por ejemplo d_{Aj}. En esas condiciones, si H_0 es cierta, d_{Aj} sigue una ley hipergeométrica de parámetros n_j, n_{Aj} y d_j, cuya esperanza y varianza son,

$$\begin{aligned} E[d_{Aj}] &= e_{Aj} = n_{Aj}\frac{d_j}{n_j} \\ Var[d_{Aj}] &= v_{Aj} = \frac{n_{Aj}n_{Bj}d_j(n_j - d_j)}{n_j^2(n_j - 1)}. \end{aligned}$$

El siguiente paso es construir un estadístico que combine la información de las J tablas 2×2 y proporcione una medida global de la desviación existente entre los

valores observados y los esperados bajo H_0. Sumando las desviaciones observadas en los distintos instantes de fallo se obtiene:

$$U_L = \sum_{j=1}^{J} (d_{Aj} - e_{Aj}) = \sum_{j=1}^{J} d_{Aj} - \sum_{j=1}^{J} e_{Aj},$$

estadístico que tiene media cero y, bajo la hipótesis de que las J tablas son independientes, varianza la suma de las varianzas de los sumandos. Aplicando el teorema central del límite se puede justificar, si el número de tiempos de fallo no es demasiado pequeño, que U_L tiene una distribución aproximadamente Normal, es decir,

$$\frac{U_L}{(Var[U_L])^{1/2}} = \frac{\sum_{j=1}^{J} (d_{Aj} - e_{Aj})}{\left(\sum_{j=1}^{J} Var[d_{Aj}]\right)^{1/2}} \sim N(0,1).$$

El cuadrado de una variable Normal estándar tiene una distribución χ^2_1, por lo que una forma equivalente de evaluar las diferencias es construir un test basado en $Q_L = U_L^2 / Var(U_L)$.

El test basado en U_L se denomina test log-rank o test de Mantel y Haenszel (1959), ya que fueron estos autores quienes lo plantearon en 1959. Aunque el razonamiento empleado para justificar la distribución de U_L no es completamente correcto ya que las tablas de contingencia, que corresponden a tiempos de fallo sucesivos, no son independientes, la validez del resultado puede justificarse utilizando argumentos de verosimilitud parcial, Kalbfleish y Prentice, capítulos 4 y 5.

Cuestión: Contrasta, utilizando el test log-rank, si la supervivencia de los dos grupos de enfermos del ensayo clínico sobre el tratamiento 6-MP, ejemplo 1.2.3 del capítulo 1, es la misma.

4.3 Una familia de tests para comparar dos poblaciones

El test log-rank da la misma importancia a las discrepancias observadas en todos los instantes de fallo, independientemente del número de individuos en riesgo en cada uno de esos instantes. Sin embargo, resulta razonable considerar que las discrepancias observadas en los primeros instantes de fallo deben influir en el estadístico con un peso mayor que las observadas en los últimos instantes, ya que están basadas en

lo ocurrido a un mayor número de individuos y, por consiguiente, son menos sensibles a fluctuaciones aleatorias que las diferencias observadas en los últimos tiempos de fallo.

En estas condiciones se define una familia de tests cuyo estadístico tiene la estructura siguiente,

$$U = \frac{\sum_{j=1}^{J} w_j (d_{Aj} - e_{Aj})}{\left(\sum_{j=1}^{J} w_j^2 Var[d_{Aj}]\right)^{1/2}},$$

donde w_j es la componente j-ésima de un vector de pesos $w = (w_1, w_2, \ldots, w_J)'$, que pondera las diferencias $(d_{Aj} - e_{Aj})$ correspondientes a los instantes $t_{(j)}$. Este estadístico U tiene una distribución $N(0,1)$ bajo la hipótesis nula. Los diferentes tests se obtienen al considerar distintos vectores de ponderación.

- Test log-rank, de Mantel-Haenszel, de Mantel- Cox o generalizado de Savage, U_L: : $w_j = 1$.
- Test de Gehan, Breslow o de Wilcoxon generalizado, U_G: $w_j = n_j$.
- Test de Tarone, Tarone-Ware, U_T: $w_j = \sqrt{n_j}$.
- Test de Prentice, Peto-Prentice, U_P: $w_j = \prod_{i=1}^{j} \frac{n_i - d_i + 1}{n_i + 1}$

Los diferentes vectores de peso proporcionan a cada test propiedades que los hacen más adecuados en determinadas condiciones:

i.- El test log-rank, es el más adecuado y potente cuando la hipótesis alternativa es la de **riesgo proporcional**:

$$H_1 : h_A(t) = \theta h_B(t) \qquad \forall t \leq \tau,$$

donde θ es una constante distinta de 1. Esta hipótesis puede formularse también en términos de la función de supervivencia mediante:

$$H_1 : S_A(t) = S_B^{\theta}(t) \qquad \forall t \leq \tau.$$

La propiedad de riesgo proporcional implica que las correspondientes funciones de supervivencia no se cruzan, ya que $S_A(t)$ será siempre mayor -o menor- que $S_B(t)$ según el valor de θ sea menor -o mayor- que 1. Un método gráfico para contrastar la hipótesis de riesgo proporcional, consiste en representar frente al tiempo el logaritmo del estimador de la función de riesgo acumulado, $\hat{H}(t)$

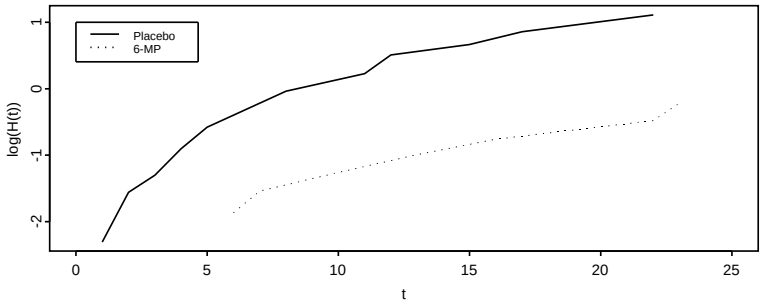

Figura 4.1: Gráfico para comprobar la hipótesis de riesgo proporcional en los dos grupos del ensayo 6-MP.

o $\tilde{H}(t)$, para cada uno de los grupos. Si la hipótesis de riesgo proporcional es cierta, deberán obtenerse dos líneas aproximadamente paralelas. En la figura 4.1 se muestra el gráfico correspondiente a los dos grupos del ensayo del tratamiento 6-MP.

Si la hipótesis de riesgo proporcional no se verifica, el test log-rank resulta demasiado sensible a la variabilidad muestral.

ii.- El test de Gehan fue deducido originalmente por ese autor como generalización del test de Mann-Whitney, o test de suma de rangos de Wilcoxon, al caso de muestras con observaciones censuradas. Breslow generalizó el test de Kruskal-Wallis para k muestras. En este estadístico se utiliza como peso el número total de individuos en riesgo en cada instante. En situaciones de riesgo no proporcional, este test resulta más potente que el log-rank.

iii.- Tarone y Ware (1977) comprobaron que el test de Gehan resulta muy sensible a los esquemas de censura, y que puede ser poco fiable cuando la distribución de los tiempos de censura en los grupos que se comparan es muy diferente. Tras estudiar distintos esquemas de pesos de la forma $w_j = f(n_j)$, propusieron $w_j = \sqrt{n_j}$, que representa un compromiso entre el test de Gehan y el log-rank.

iv.- El test de Prentice (1978) es también una generalización del test de Wilcoxon a muestras censuradas. Los pesos utilizados son valores próximos a la estimación Kaplan-Meier de la función de supervivencia en los instantes $t_{(j)}$, calculada a partir de la muestra combinada.

La obtención de estos tests como tests de rangos está desarrollada en Lawless(1982), pp. 412-427, y la deducción del test de Gehan como generalización

del test de Mann-Whitney en Gross y Harris, pp. 243-249.

Se debe observar que para rechazar la hipótesis nula, el estadístico U debe acumular discrepancias del mismo signo. En efecto, el valor de U será significativamente grande, o pequeño, si la mayor parte de los sumandos de su numerador tienen el mismo signo, lo que ocurre cuando la diferencia en la supervivencia entre los dos grupos es consistente a lo largo del tiempo; es decir, si $S_A(t) < S_B(t)$ o $S_A(t) > S_B(t)$ para todo t. En otras condiciones, el estadístico U no siempre detecta diferencias en la supervivencia de los grupos, pues al sumar términos de diferente signo, se pueden obtener valores de U muy próximos a 0. En general, el comportamiento de la supervivencia en la comparación de tratamientos suele ser consistente; si no es así, es preferible utilizar otro tipo de tests, como las generalizaciones del test de Kolmogorov-Smirnov a muestras con datos censurados.

4.4 Generalización al caso de tres o más muestras

Los tests citados en el apartado 4.3 para comparar la supervivencia en dos grupos de población pueden generalizarse al caso de G grupos, con $G \geq 3$. La hipótesis nula sigue siendo,

$$H_0 : h_1(t) = h_2(t) = \ldots = h_G(t) \qquad \forall t \leq \tau,$$

frente a la alternativa de que las tasas de fallo de al menos dos de los grupos difieran para algún instante t.

La información correspondiente al conjunto de muestras se puede resumir en J tablas de contingencia, ahora de dimensión $G \times 2$, una para cada instante de fallo, en las que $t_{(1)} < \ldots < t_{(j)} < \ldots < t_{(J)}$ representan los J tiempos de fallo distintos que se han observado en el conjunto de los G grupos, dispuestos en orden creciente. La tabla correspondiente al instante $t_{(j)}$ se muestra en la tabla 4.2.

Condicionalmente a los totales marginales observados, el vector de variables aleatorias $(d_{1j}, d_{2j}, \ldots, d_{Gj})$ que indica el número de fallos ocurridos en el instante $t_{(j)}$ en cada una de las muestras, sigue una distribución hipergeométrica $G-1$ dimensional, pues existe una relación que determina uno de ellos conocidos los $G-1$ restantes, $d_{Gj} = d_j - \sum_{j=1}^{J-1} d_{Gi}$. Bajo la hipótesis nula, el número esperado de fallos en la

Grupo	N^o fallos en $t_{(j)}$	N^o individuos vivos en $t_{(j)}$	N^o individuos en riesgo en $t_{(j)}$
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
...	...	...	...
k	d_{kj}	$n_{kj} - d_{kj}$	n_{kj}
...	...	...	...
G	d_{Gj}	$n_{Gj} - d_{Gj}$	n_{Gj}
Total	d_j	$n_j - d_j$	n_j

Tabla 4.2: Tabla correspondiente al instante $t_{(j)}$ para comparar G grupos.

muestra k-ésima es,

$$E[d_{kj}] = e_{kj} = n_{kj}\frac{d_j}{n_j}$$

y la matriz de covarianzas correspondiente, $V[t_{(j)}]$, de dimensión $(G-1)\times(G-1)$ está formada por elementos de la forma,

$$V_{kl} = Cov[d_{kj}, d_{lj}] = \begin{cases} \frac{n_{kj}(n_j - n_{kj})d_j(n_j - d_j)}{n_j^2(n_j-1)} & \text{para } k = l \\ \frac{n_{kj}n_{lj}d_j(n_j-d_j)}{n_j^2(n_j-1)} & \text{para } k \neq l \end{cases}$$

con $k = 1, \ldots, G-1$ y $l = 1, \ldots, G-1$.

Acumulando las discrepancias correspondientes a los distintos instantes de fallo se obtiene un vector U, de dimensión $G-1$, cuyas componentes son,

$$U_k = \sum_{j=1}^{J} w_j(d_{kj} - E[d_{kj}]),$$

donde w_j es la componente j-ésima de un vector de pesos. Suponiendo de nuevo que las J tablas son independientes, la matriz de covarianzas del vector U, V, es la suma de las matrices de covarianzas correspondientes a los J tiempos de fallo, ponderadas con el cuadrado del peso correspondiente,

$$V = \sum_{j=1}^{J} w_j^2 V[t_{(j)}].$$

Los vectores de pesos que se utilizan para construir los diferentes tests son los mismos que en el caso de dos grupos.

El test para contrastar la igualdad de la supervivencia en los G grupos se basa en el estadístico,

$$Q = U'V^{-1}U$$

que, bajo la hipótesis nula, dado al carácter asintóticamente Normal de la distribución de U, tiene una distribución aproximada χ^2 con $G-1$ grados de libertad.

4.5 Análisis estratificado

Para poder establecer comparaciones válidas entre dos o más grupos es fundamental que los individuos de la muestra sean lo más homogéneos posible, excepto en el factor que define los grupos, a fin de garantizar que las diferencias que se observan en las distintas submuestras estén asociadas únicamente a ese factor. Si existen variables conocidas y controlables en el ensayo que se cree que pueden influir en la supervivencia -como el sexo, la edad, la fase de la enfermedad, etc.- es necesario controlar el efecto de esas covariables realizando una asignación estratificada. En este tipo de asignación, a partir de los diferentes valores de esas covariables, se definen categorías, denominadas **estratos**, de forma que los individuos de un mismo estrato puedan considerarse homogéneos y en cada uno de ellos se realiza una asignación aleatoria de los individuos a los grupos. Si no se procede de esta forma, se podrían atribuir al efecto del factor que define los grupos, diferencias debidas a las distintas características de los individuos.

Los tests citados en la sección 4.3 son también aplicables en esta situación con ligeras modificaciones. Supongamos que se quiere comparar la supervivencia de G grupos y que a partir de un conjunto de covariables se han definido M estratos. La hipótesis que se desea contrastar ahora es,

$$H_0 : h_{1s}(t) = h_{2s}(t) = \ldots = h_{Gs}(t) \qquad \text{para } s = 1, \ldots, M, \text{ y } \forall t \leq \tau,$$

frente a la alternativa de que la supervivencia de algún grupo, en alguno de los estratos, sea diferente.

Utilizando los datos del estrato s-ésimo se calculan las cantidades ${}_sd_{Aj}$, ${}_se_{Aj}$, $\mathrm{Var}[{}_sd_{Aj}]$. A partir de ellas y del correspondiente vector de pesos se construye, de forma análoga a la indicada en el apartado 4.4, el vector U_s y la matriz V_s, que resumen la situación dentro de ese estrato. Sumando los vectores obtenidos en los diferentes estratos se definen $U = \sum_{s=1}^{M} U_s$ y $V = \sum_{s=1}^{M} V_s$, con los que se construye el estadístico global,

$$Q = U' V^{-1} U.$$

Si el tamaño de muestra es grande, y bajo la hipótesis nula, este estadístico tiene una distribución χ^2 con $G-1$ grados de libertad.

4.6 Tests para tendencia

La variable categórica que define los grupos que se comparan puede no ser una variable nominal, como el sexo o el grupo sanguíneo, sino una variable ordinal cuyas categorías tienen un orden intrínseco, por ejemplo la gravedad de un tumor, la cantidad de dosis administrada, etc. En estos casos, en lugar de plantear una hipótesis alternativa genérica -que la función de riesgo de alguno de los grupos es diferente- es preferible plantear una hipótesis alternativa más específica, como la existencia de un orden o tendencia en la supervivencia ligada a la variable ordinal que define los grupos. Supongamos que hay G, con $G > 2$, grupos en estudio, enumerados de acuerdo con el orden que se piensa puede existir entre las funciones de riesgo. La alternativa a la hipótesis de igualdad es ahora,

$$H_1 : h_1(t) \leq h_2(t) \leq \ldots \leq h_G(t) \qquad \forall t \leq \tau,$$

o, equivalentemente,

$$H_1 : S_1(t) \geq S_2(t) \geq \ldots \geq S_G(t) \qquad \forall t \leq \tau.$$

Al comparar con un test general grupos en los que existe una estructura de orden como la indicada con un test general, es posible que no se lleguen a apreciar diferencias significativas en su supervivencia. Sin embargo, con un test que incorpore información específica sobre la estructura de orden, es más probable que se detecten las diferencias existentes.

Para construir un test de tendencia se asocia a cada grupo una puntuación a_k, con $a_1 < a_2, < \ldots, < a_G$, de modo que los valores mayores correspondan a los grupos de mayor riesgo. La elección habitual de ese vector $a = (a_1, a_2, \ldots, a_G)$ consiste en tomar $a_k = k$, aunque dependiendo de la información que aporta la variable clasificatoria pueden elegirse cantidades que caractericen mejor los distintos grupos, como la edad media, la dosis recibida, etc. El estadístico del test se construye a partir del vector U y la matriz V calculados como se indicó en el apartado 4.4, pero

que ahora son de dimensión G y $G \times G$ respectivamente. La expresión del estadístico es,

$$Q_{tend} = \frac{(a'U)^2}{a'Va},$$

que, bajo la hipótesis nula, y si los tamaños de muestra son suficientemente grandes, tiene una distribución χ^2 con un grado de libertad.

4.7 Ejercicios

1.- En la tabla 4.3 se muestran los datos obtenidos en un estudio comparativo de supervivencia de pacientes con cáncer de recto tratados entre 1935 y 1944, periodo P1, y entre 1945 y 1954, periodo P2. En dicha tabla, d_i denota el número de muertes que se produjeron cada año y $n_i - d_i$ el de individuos en riesgo durante el mismo periodo. Contrasta utilizando el test log-rank si la supervivencia de los pacientes diagnosticados y tratados durante el primer periodo es la misma que la de los tratados durante el segundo.

2.- Utilizando los datos sobre enfermos de mieloma múltiple del ejercicio 2 del capítulo 3, analiza gráficamente si la hipótesis de riesgo proporcional en los grupos de hombres y mujeres es plausible y, según el resultado, contrasta la igualdad de supervivencia en los dos grupos utilizando el test que consideres adecuado.

3.- Con los datos de los pacientes con enfermedad de Hodking del ejercicio 5 del capítulo 3, contrasta si la supervivencia del grupo que había recibido terapia anteriormente y la del que no la había recibido son significativamente distintas, utilizando el test log-rank y el de Gehan. Analiza los resultados obtenidos con los dos tests.

		1	2	3	4	5	6	7	8	9	10
P1	d_i	167	45	45	19	17	11	8	5	6	7
	$n_i - d_i$	220.0	173.5	127.5	108.0	91.0	79.5	71.0	66.0	59.5	52.0
P2	d_i	185	88	55	43	32	31	20	7	6	6
	$n_i - d_i$	559.0	461.0	396.0	343.0	299.0	235.0	170.0	132.0	101.5	71.0

Tabla 4.3: Tabla ejercicio 1.

Nueva Terapia	Control
30, 67, 79*, 82*, 95, 148, 170, 171, 176, 193, 200, 221, 243, 261, 262, 263, 399,414, 446, 446*, 464, 777	57, 58, 74, 79, 89, 98, 101, 104, 110, 118, 125, 132, 154, 159, 188, 203, 257, 257, 431, 461, 497, 723,747, 1313, 2636

Tabla 4.4: Tabla ejercicio 4.

4.- Los datos de la tabla 4.4 son tiempos de supervivencia correspondientes a una muestra de pacientes con cáncer en el conducto biliar, que participaron en un estudio que pretendía determinar si un tratamiento, que combinaba la radioterapia con la administración de un nuevo medicamento 5-FU, prolongaba la supervivencia, Fleming et al. (1980). En la tabla se muestra los tiempos de supervivencia, en días, del grupo de pacientes que recibió la nueva terapia y los del grupo de control.

Representa el estimador producto límite de la función de supervivencia en los dos grupos y observa su forma. Contrasta la hipótesis de que las dos curvas de supervivencia son iguales utilizando distintos tests de comparación y analiza sus conclusiones.

5.- Se ha realizado un estudio para comparar dos tratamientos quirúrgicos aplicados en pacientes con cáncer de pecho. De cada paciente se conoce el número de nodos linfáticos afectados que se detectaron en la intervención quirúrgica. Con esa información se definen cuatro estratos: ningún nodo afectado, un nodo afectado, dos o tres y, por último, cuatro o más nodos afectados. En la tabla 4.5 se muestra, para cada estrato, m_n: el número de pacientes, ${}_md_A$: el número de pacientes que mueren en el grupo A, $E[{}_md_A]$: el número esperado de fallos en ese grupo bajo la hipótesis nula de igualdad de los dos tratamientos y, por último, la varianza de $U_m =_m d_A - E[{}_md_A]$.

A partir de esta información plantea un test para compara los dos tratamientos.

N^o nodos	m_n	${}_md_A$	$E[{}_md_A]$	$Var[U_m]$
0	520	64	65.35	32.99
1	100	11	15.02	6.96
2-3	43	10	11.90	4.03
≥ 4	38	9	9.32	4.73

Tabla 4.5: Tabla ejercicio 5.

Tratamiento 1	Tratamiento 2	Tratamiento 3
4, 5, 9, 10, 12, 13, 10, 23, 28, 28, 28, 29, 31, 32, 37, 41, 41, 57, 62, 74, 100, 139, 20*, 258*,269*	8, 10, 10, 12, 14, 20, 48, 70, 75, 99, 103, 162, 169, 195, 220, 161*,199*, 245*	8, 10, 11, 23, 25, 25, 28, 28, 31, 31, 40, 48, 89, 124, 143, 12*, 159*, 190*,217* 196*, 197*, 205*, 219*

Tabla 4.6: Tabla ejercicio 7.

6.- Utilizando los datos del ensayo realizado para comparar dos tratamientos de quimioterapia en enfermos de cáncer de pulmón descrito en el ejemplo 5 del capítulo 1, compara la eficacia de los dos tratamientos, tomando en consideración la posible influencia del tipo de tumor. El conjunto de datos se encuentra en el fichero LUNG.DAT.

7.- Se han registrado los tiempos de remisión, medidos en días, de 66 pacientes de leucemia. Los pacientes habían sido asignados aleatoriamente a tres tratamientos distintos y los tiempos observados se muestran en la tabla 4.6. Compara la supervivencia de los tratamientos.

8.- Los datos de la tabla 4.7 corresponden a un estudio de carcinogénesis, Thomas et al. (1977), en el que se administraron diferentes dosis de cierto agente cancerígeno a individuos asignados aleatoriamente a tres grupos de tratamiento. Se define como tiempo hasta el fallo, el intervalo, en días, hasta que se detecta la aparición de un tumor.

Analiza si el riesgo en los tres grupos es el mismo, o si existen evidencias de que el riesgo varía de forma proporcional a la dosis a la que los individuos han estado expuestos.

9.- En un estudio de carcinogénesis realizado con el fin determinar el efecto cancerígeno de un aditivo utilizado en la fabricación de productos alimenticios, se realizó un ensayo con 398 ratones machos y hembras. Se establecieron cuatro grupos, uno de control y tres de tratamiento, a los que se suministró diferentes dosis del aditivo. El tiempo de supervivencia es el número de semanas transcurridos hasta la detec-

Dosis: 2.0	Dosis:1.5	Dosis: 0.0
41*, 41*, 47, 47*, 47*, 58, 58, 58, 100*, 117	43*, 44*, 45*, 67, 68*, 136, 136, 150, 150, 150	73*, 74*, 75*, 76, 76, 76*, 99, 166, 246*

Tabla 4.7: Tabla ejercicio 8.

ción de un tumor. Suponiendo que los niveles de dosis utilizados fueron linealmente crecientes, analiza si el aditivo tiene un efecto cancerígeno en los ratones. Realiza ese mismo contraste tomando en consideración la posible influencia del sexo. Los datos se encuentran en el fichero TUMOR.DAT y las variables son: dosis (codificada de 1 a 4), sexo, tiempo de supervivencia y estado de la observación.