Análisis de datos censurados: técnicas de estimación e inferencia no paramétricas y paramétricas

Jorge Muro Guerrero

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Directora del trabajo: Ana Carmen Cebrián Guajardo
10 de septiembre de 2019

Abstract

Failure times, denoted as T, are random variables that represent the time to a particular event and they are studied in different areas of research like medicine (survival analysis), engineering (reliability), economics or social sciences. When observing these variables, censored data arise, when the observed value of the variable is only partially known. These censored observations may appear when the event of interest hasn't happened by the end of the study or when the subject dropped out, moved away or died from other causes; then the only information we know is that the observed value of the variable is greater than a given value. Deleting censored data or taking them as true observations will lead to biased estimations, so we need to add this information properly to the analysis if we want to produce efficient estimations based on the obtained sample.

First, in Chapter 1 the existing types of censoring will be presented, followed by some definitions and properties about the probability distribution of failure times. The usual functions that uniquely characterize the distribution of a variable are usually the probability density function $f(t)$ (pdf) and the cumulative distribution function $F(t)$ (cdf), but in the study of failure times other functions are equally used: the survival function, defined as $S(t) = P(T > t)$, the hazard function $h(t)$, the cumulative hazard function $H(t)$ and the mean residual life function $m(t)$. Some relations between these functions will be proved and the main summary measures of location will be presented, like the mean survival time $E[T]$, the median survival time and the percentiles and also another function, the median residual lifetime.

In Chapter 2, some methods of estimation for censored samples are provided. Methods can be non-parametric, if they don't require any assumptions about the population distribution, or parametric, if they assume that sample data come from a given family of distributions and then the parameters of the distribution are estimated.

The main non-parametric statistic introduced is the Kaplan-Meier estimator, used to estimate the survival function. It is defined as,

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

with $t_{(i)}$ a time point at which at least one event occurred, d_i the number of events that happened at $t_{(i)}$ and n_i the number of individuals at risk at $t_{(i)}$, i.e., the individuals still in the study at time $t_{(i)}$. Some properties of this estimator are obtained and it is described how to calculate confidence intervals for $S(t)$ at any time t.

Using the Kaplan-Meier estimator of $S(t)$, one can obtain estimators of the cumulative hazard function $H(t)$, the mean survival time $E[T]$ and the percentiles together with their confidence intervals. Also there exists the Nelson-Aalen estimator for $H(t)$ and $S(t)$.

In order to make parametric estimations, a family of distributions must be selected. First, the most common distributions used to model failure times are described: the exponential, Weibull, Gamma, Log-normal and Loglogistic distributions. Then, a graphic procedure is described to see which distribution better fits the data.

Once a distribution is selected, the vector of parameters Θ must be estimated. Here the maximum likelihood estimation from complete samples is adapted to censored samples. The only change lies in the definition of the likelihood function, which is defined as,

$$L(\Theta) = \prod_{i=1}^{n} L_i(\Theta) = \prod_{t_i \in F} f(t_i;\Theta) \prod_{t_i^* \in C} S(t_i^*;\Theta),$$

where F if the set of instants of failure and C is the set of censoring times. The aim is to maximize this function. With the maximum likelihood estimator of the parameters one can obtain the estimator for any bijective function of them, such as the survival function or the mean of the distribution. Finally, confidence intervals for the parameters and functions of the parameters will be computed.

In Chapter 3, tests for comparing the distribution of T of two or more groups will be introduced. The first one is the non-parametric log-rank test family. The null hypothesis is that all the groups have the same distribution, and the alternative hypothesis is that at least one of the groups differs from the others at some time. A test statistic is constructed and when the null hypothesis is true it has a χ^2 distribution with $G-1$ degrees of freedom, with G the number of groups.

The second test is the likelihood ratio test for two groups and it is parametric. If the distribution of the groups is known, the null hypothesis is that the vectors of parameters are equal and the alternative is that they are different, i.e., at least one parameter is different in each group. The test statistic is

$$X_l = -2[l(\hat{\Theta}) - l(\hat{\Theta}_1, \hat{\Theta}_2)],$$

with $l(\hat{\Theta})$ the maximum loglikelihood under H_0 and $l(\hat{\Theta}_1, \hat{\Theta}_2)$ the maximum loglikelihood with no restriction on the vectors of parameters (Θ_1, Θ_2). This statistic, under the null, follows a χ^2 distribution with $k-1$ degrees of freedom, with k the number of parameters of the vectors. This test can be adapted to the situation in which the null hypothesis doesn't include all the parameters of the distribution.

Finally, in Chapter 4 a censored sample simulation and an application of different estimation methods to the simulated sample are carried out. The methods used are both the standard methods for uncensored samples and the methods proposed in this paper. Then, the different results are compared to show that the usual methods produce bad estimations, but the new ones work well.

Índice general

Capítulo 1

Las variables aleatorias tiempos de fallo y sus características

En este primer capítulo, se introducen las variables aleatorias tiempos de fallo y una característica muy habitual: la presencia de datos censurados en las muestras a analizar, es decir, datos cuyo valor solo se conoce parcialmente, así como los tipos de censura que se pueden dar. Además, se presentan las funciones que caracterizan la distribución de probabilidad de estas variables y sus medidas de posición.

1.1. El tiempo de fallo

Las variables aleatorias **tiempos de fallo o supervivencia**, también denominadas usualmente como tiempos de vida o tiempos de respuesta, representan el tiempo transcurrido desde un instante inicial hasta la ocurrencia de un determinado evento que es el fallo o respuesta. Estas variables se pueden analizar estadísticamente con el fin de caracterizar la frecuencia de fallo de un elemento. A partir de ahora denotaremos como T a este tipo de variables.

Los dos principales campos en los que aparecen problemas cuyo objetivo es estudiar el tiempo de funcionamiento de un elemento o individuo son:

- La industria y la tecnología: El estudio de estas variables recibe el nombre de Fiabilidad (de componentes). Un ejemplo de una variable tiempo de fallo es el tiempo hasta la primera avería de un determinado tipo de coche.
- La biomedicina: En este caso, el estudio de estas variables se denomina Análisis de supervivencia y se utiliza en ensayos clínicos. Por ejemplo, el tiempo hasta la reaparición de un tumor en un enfermo de cáncer o el tiempo hasta la curación de una enfermedad en un paciente desde que comienza su tratamiento.

En las últimas décadas se ha producido un gran desarrollo de los métodos estadísticos para analizar este tipo de datos y existen además aplicaciones en epidemiología, economía, finanzas, criminología y ciencias sociales, entre otros campos.

1.2. Obtención de los datos

Para definir una variable tiempo de fallo debe especificarse qué instante marca el comienzo y qué suceso es el fallo o respuesta. La medida utilizada en la mayor parte de los casos será el tiempo real transcurrido, pero también podría ser el tiempo operativo u otra cantidad no negativa adecuada.

Como en cualquier estudio estadístico, para garantizar la validez y poder extrapolar los resultados del estudio es necesario que la muestra sea representativa de la población de la que proviene. Además, para evitar la aparición de sesgo en los resultados se deben analizar muestras aleatorias simples (i.i.d.), en las que los tiempos de fallo de todos los elementos tengan la misma distribución.

1.3. Características de los tiempos de fallo: asimetría y datos censurados

La variable tiempo de fallo, T, es una variable no negativa que suele tener una distribución bastante asimétrica. Por este motivo, la distribución Normal no tendrá la relevancia que tiene en otros campos de la Estadística; su papel lo tomará la distribución Exponencial.

El rasgo diferencial que caracteriza el análisis de este tipo de variables es la presencia de observaciones incompletas o **censuradas** en las muestras sobre las que hay que realizar inferencia. El principal motivo es que la obtención de muestras completas suele requerir demasiado tiempo por lo que es habitual terminar los experimentos antes de haber observado todos los fallos y, además, puede haber alguna dificultad para observar la respuesta de un individuo como puede ser el abandono durante el ensayo por parte de este o su fallo por una causa distinta a la estudiada en el ensayo. Sin embargo, estas observaciones parciales sí aportan información: que el valor de T es mayor que el tiempo transcurrido hasta su censura, y debe incorporarse al análisis de forma adecuada, ya que si las observaciones censuradas se eliminan, o se toman por auténticas, las estimaciones sobre T pueden resultar sesgadas e ineficientes.

En el capítulo 4 se simula con R una muestra censurada de tamaño 300 de una distribución Exp(0.4) y se realizan distintas estimaciones. Eliminando los datos censurados, la media, por ejemplo, es 1.957, y considerándolos fallos, la media es 1.975. La media de una distribución Exp(0.4) es 2.5, por lo que se observa cierto sesgo en la estimación de la media si procedemos de alguna de estas maneras. El objetivo de las técnicas que se van a presentar en este trabajo es tratar adecuadamente los datos censurados.

1.4. Tipos de censura

La censura puede ser a derecha, a izquierda o en un intervalo. Una observación se dice censurada a derecha en L, si se desconoce el valor exacto de la observación y solo se sabe que es mayor que L. Por otro lado, una observación se dice censurada a izquierda en L, si solo se sabe que la observación es menor que el valor L. Además, en algunos experimentos, pueden aparecer datos censurados en un intervalo (t_I, t_D); es decir, que la información sobre ellos es que $t_I < T < t_D$. La censura a derecha es la más frecuente y es la única que consideraremos a partir de ahora.

1.4.1. Censura de tipo I y censura de tipo II

Según la manera en que se limita la duración del experimento que se realiza para obtener los datos, los dos esquemas de censura más frecuentes son:

- Censura de tipo I. En este esquema el experimento se programa con una duración, C, establecida a priori. El tiempo de fallo de un elemento se observará si es menor o igual que ese valor prefijado. En otro caso, la observación correspondiente tendrá un valor censurado C. En este esquema, el número de observaciones censuradas de la muestra es aleatorio y el valor de censura es fijo. Este es el diseño que se utiliza generalmente en los ensayos médicos.

- Censura de tipo II. Con este esquema de censura, un ensayo con n componentes idénticos finaliza en el momento en que se produce el r-ésimo fallo $(1 \leq r \leq n)$. Ese instante, $t(r)$, será el valor censurado de todos los elementos que en ese momento no hayan fallado todavía. Es decir, se observan los r tiempos de fallo más pequeños de la muestra y aparecen $n-r$ tiempos censurados en el valor $t(r)$. En este esquema, el número de observaciones censuradas de la muestra es fijo y el valor de censura aleatorio. Este diseño es más frecuente en los experimentos industriales.

El valor de C en el esquema de tipo I y el valor de r (o de r/n) en el esquema de tipo II deben fijarse antes de iniciar el experimento, para garantizar la independencia entre el mecanismo de censura y la observación del fenómeno, que es una de las condiciones necesarias en el desarrollo de las herramientas estadísticas que presentaremos.

1.4.2. Muestras simplemente censuradas y muestras múltiplemente censuradas

Dependiendo de si existe un valor de censura único o no, las muestras generadas por los experimentos diseñados pueden ser:

- Muestras simplemente censuradas: Tienen un valor de censura común, $t(r)$ o C, para todas las observaciones censuradas. Esto requiere que todos los individuos comiencen el ensayo al mismo tiempo. En los ensayos industriales son más habituales este tipo de muestras.

- Muestras múltiplemente censuradas: Las observaciones censuradas de la muestra pueden tener valores de censura diferentes. Esto sucede en los ensayos médicos ya que, además de establecerse una limitación temporal C, es habitual que los individuos se incorporen al ensayo en distintos instantes de tiempo y que se produzcan abandonos durante el ensayo, que dan lugar a observaciones censuradas porque sólo se sabe que el fallo no se había observado hasta el momento del abandono.

En la Figura 1.1 se representa un ejemplo de muestra múltiplemente censurada que corresponde a un estudio de 18 meses de duración, en el que solo fueron admitidos pacientes durante los 6 primeros meses. En el gráfico superior se pueden observar los instantes de entrada y salida, por fallo (•) o censura (○), de cada paciente. Los pacientes 1, 4, 7 y 10 murieron durante el estudio, los pacientes 3, 6, 8 y 9 seguían vivos al terminar el estudio, y el 2 y el 5 lo abandonaron antes de que finalizara. El instante en el que se comenzó a medir cada observación no es de interés ya que la magnitud que queremos analizar es el tiempo de fallo T, es decir, el tiempo desde el instante de entrada hasta el instante de fallo; por esta razón, las observaciones suelen representarse con el mismo origen, como se muestra en el gráfico inferior.

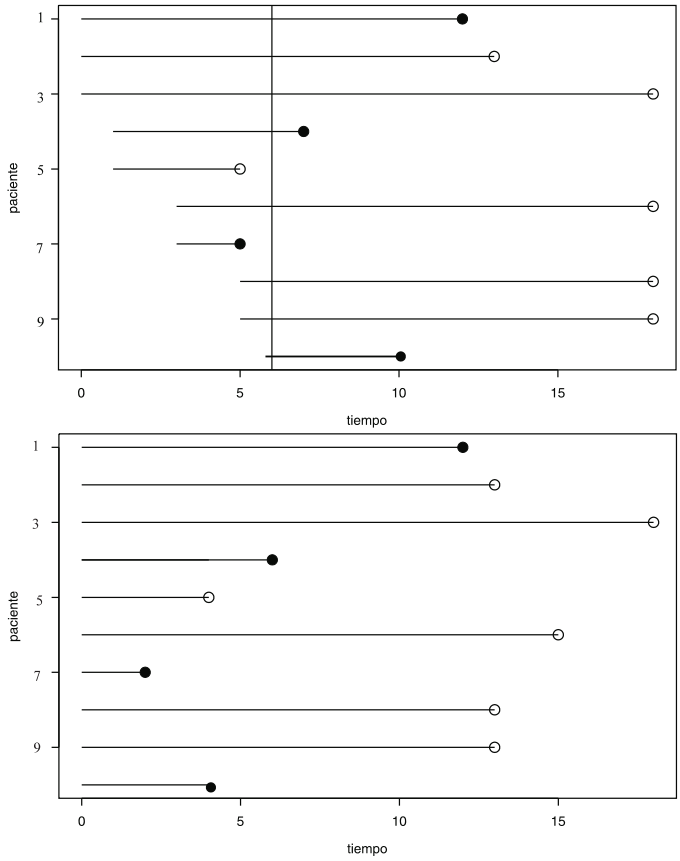

Figura 1.1: Esquema de los tiempos de fallo.

1.4.3. El modelo de censura aleatoria

En este modelo general de censura que engloba a los anteriores se supone que para cada individuo i de la muestra existen dos variables aleatorias, el tiempo de fallo T_i y el tiempo de censura C_i, y que son independientes. La muestra queda determinada por dos variables: el mínimo de cada par (T_i, C_i) y, además, una variable binaria para distinguir los fallos de las censuras, que vale 1 si $T_i \leq C_i$ y 0 en otro caso. Es necesario que la distribución del tiempo de fallo sea independiente de la censura para realizar estimaciones insesgadas a partir de la muestra.

1.5. La distribución de probabilidad de los tiempos de fallo y funciones que la caracterizan

Consideramos que los tiempos de supervivencia observados son observaciones independientes de una variable aleatoria T. El conjunto de valores que puede tomar la variable es no negativo, generalmente $\mathbb{R}^+ = [0, \infty)$ o un intervalo contenido en $\mathbb{R}^+$. En ocasiones, podría ser un conjunto discreto como $\mathbb{Z}^+$.

1.5.1. Variables continuas

Las funciones habituales que caracterizan una distribución de probabilidad son:

- La función de densidad: Se define como

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

  Es una función no negativa y $\int_0^\infty f(x)dx = 1$. Con intervalos suficientemente pequeños,

$$f(t) \approx \frac{P(t \leq T < t + \Delta t)}{\Delta t},$$

  es decir, la función de densidad en t se aproxima a la probabilidad de fallo en un intervalo de tiempo pequeño con inicio en t, expresada por unidad de tiempo. Equivalentemente, la probabilidad de fallo en un intervalo pequeño con inicio en t es aproximadamente la función de densidad en t por la longitud del intervalo: $P(t \leq T < t + \Delta t) \approx f(t)\Delta t$.

- La función de distribución: Se define como $F(t) = P(T \leq t)$ y para todo t,

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

Además de estas funciones, existen otras que se utilizan en el análisis de los tiempos de fallo:

- La función de supervivencia o de fiabilidad: Se define como $S(t) = P(T > t)$, es la probabilidad de que un individuo no haya fallado todavía en el instante t, es decir, que sobreviva más allá del instante t. Para todo t,

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

  Es una función monótona no creciente, $S(0) = 1$ y $\lim_{t \to \infty} S(t) = 0$.

- La función de riesgo: Se define como

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

  Cuando $\Delta t \to 0$, $h(t) \approx \dfrac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$, es decir, la función de riesgo en t se aproxima a la probabilidad por unidad de tiempo de que un individuo que ha sobrevivido hasta el instante t falle en un intervalo pequeño con inicio en t.

- La función de riesgo acumulado, definida como

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

representa el riesgo acumulado hasta el instante t.

Relaciones entre las funciones

Se presentan aquí las relaciones que se dan entre las funciones que se acaban de presentar y sus respectivas demostraciones.

- $f(t) = -\dfrac{dS(t)}{dt}$: $\quad f(t) = \dfrac{dF(t)}{dt} = \dfrac{d(1-S(t))}{dt} = -\dfrac{dS(t)}{dt}$.
- $h(t) = \dfrac{f(t)}{S(t)}$: $\quad h(t) = \lim\limits_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t} = \lim\limits_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{P(T \geq t)\Delta t} = \frac{1}{P(T \geq t)} \lim\limits_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t} = \dfrac{f(t)}{S(t)}$.
- $h(t) = -\dfrac{d \ln S(t)}{dt}$: $\quad h(t) = \dfrac{f(t)}{S(t)} = \dfrac{-\frac{dS(t)}{dt}}{S(t)} = -\dfrac{d \ln S(t)}{dt}$.
- $H(t) = -\ln(S(t))$: $\quad H(t) = \displaystyle\int_0^t h(x)dx = \int_0^t -\frac{d \ln S(x)}{dx}dx = -\ln(S(t))$.

1.5.2. Variables discretas

Suponemos que la variable toma valores en el conjunto de los enteros positivos, aunque este caso es poco frecuente. Las funciones que caracterizan la distribución en este caso son:

- La función de probabilidad: $p_k = P(T = k)$ para $k = 1, 2, \ldots$, con $p_k > 0$ y $\sum\limits_k p_k = 1$.
- La función de supervivencia: $S(t) = P(T > t) = \sum\limits_{k>t} p_k$.
- La función de riesgo: $h(t) = P(T = t | T \geq t) = \begin{cases} \frac{p_k}{S(k-1)} & \text{si } t = k \in \mathbb{Z}^+ \\ 0 & \text{si } t \neq k \in \mathbb{Z}^+ \end{cases}$
- La función de riesgo acumulado: $H(t) = \sum\limits_{k \leq t} h(k)$.

1.5.3. El tiempo de vida restante y la vida media residual

La variable aleatoria R_t, tiempo de vida restante en el instante t, se define como $T - t \mid T \geq t$, es decir, el tiempo que queda hasta el instante de fallo de un individuo dado que ha sobrevivido hasta el instante t. Dadas las funciones de la variable tiempo de fallo T que se han presentado, las funciones de R_t son:

- $S_{R_t}(x) = P(R_t > x) = P(T > x+t | T \geq t) = \dfrac{S(x+t)}{S(t)}$
- $f_{R_t}(x) = \dfrac{f(x+t)}{S(t)}$
- $h_{R_t}(x) = h(x+t)$
- $H_{R_t}(x) = H(x+t) - H(t)$

Otra función que también caracteriza la distribución de T es la función de vida media residual, $m(t)$, que representa el tiempo de vida restante esperado en cualquier instante $t \geq 0$ y se define como, $m(t) = E(R_t)$. La mediana de R_t también es muy utilizada en Análisis de Supervivencia.

1.6. Medidas resumen de posición

La medida de posición central más utilizada es la media, pero como la variable T suele ser asimétrica, la mediana es más adecuada en este caso.

- La media de T viene dada por la siguiente expresión:

$$E[T] = \int_0^\infty S(x)dx$$

  Aplicando esta relación al tiempo de vida restante en t R_t, la función de vida media residual se puede calcular como $m(t) = E(R_t) = \int_0^\infty S_{R_t}(x)dx$ y de aquí se obtiene la expresión

$$m(t) = \int_t^\infty \frac{S(x)}{S(t)}dx.$$

- El cuantil p se define como el valor t_p que verifica $S(t_p) = 1 - p$.

- La mediana de T es el valor en el que el 50 % de la población bajo estudio ha fallado, es decir, en el que la probabilidad de sobrevivir es la mitad de la que había al inicio, 1. La mediana es el valor $t_{0,5}$ tal que $S(t_{0,5}) = 0{,}5$ por lo que se calcula resolviendo esta ecuación. Aplicando esto al tiempo de vida restante en t_i R_{t_i}, la mediana de R_{t_i} es el valor $t_{R_{t_i};0,5}$ que verifica $S_{R_{t_i}}(t_{R_{t_i};0,5}) = 0{,}5$. En términos de la variable aleatoria T, verifica

$$\frac{S(t_i + t_{R_{t_i};0,5})}{S(t_i)} = \frac{1}{2}$$

  o, equivalentemente,

$$S(t_i + t_{R_{t_i};0,5}) = \frac{S(t_i)}{2}.$$

Capítulo 2

Estimación de la función de supervivencia y parámetros relacionados

En este capítulo, se presentan estimaciones sobre distintos parámetros y funciones relacionadas con las variables tiempos de fallo T. A la hora de estudiar el comportamiento de una variable aleatoria existen dos formas posibles de hacerlo: las estimaciones no paramétricas y las paramétricas. Las primeras no suponen ninguna hipótesis sobre la distribución de la variable y se emplean en las primeras fases del estudio cuando aún no se tiene información sobre el comportamiento de la variable. En cambio, las técnicas paramétricas requieren establecer una hipótesis sobre la familia de distribuciones a la que pertenece la variable y luego se calculan estimadores e intervalos de confianza de los parámetros de la distribución. A partir de estos se pueden obtener los estimadores de todas las funciones y medidas de la distribución de T.

2.1. Estimaciones no paramétricas

Dentro de las estimaciones no paramétricas, se introduce el estimador de Kaplan-Meier de la función de supervivencia. Además, a partir de este estimador se pueden obtener intervalos de confianza para $S(t)$ y otros estimadores de algunas funciones y medidas de la distribución de T.

El estimador natural de $S(t)$ es la función de supervivencia empírica,

$$\hat{S}(t) = \hat{P}(T > t) = \frac{\text{Nº de individuos con } T > t}{\text{Nº de individuos de la muestra}}.$$

Sin embargo, si la muestra contiene observaciones censuradas este no es un estimador adecuado porque trata las censuras como instantes de fallo y es probable que subestime $S(t)$. El estimador no paramétrico más utilizado para la función de supervivencia es el estimador producto-límite o estimador Kaplan-Meier (KM).

2.1.1. El estimador Kaplan-Meier de $S(t)$

Consideremos una muestra de n individuos y que esos individuos han fallado en $s \leq n$ instantes de fallo distintos $t_{(1)} < t_{(2)} < \ldots < t_{(s)}$ (estos instantes no incluyen los instantes de censura). Para cada instante de fallo $t_{(i)}$ se define:

- d_i, $(di \geq 1)$, es el numero de individuos de la muestra que han fallado en $t_{(i)}$.
- n_i, es el número de individuos en riesgo en ese instante, es decir, es el número de individuos que llegan al instante $t_{(i)}$ sin haber fallado ni haber sido censurados antes. Las observaciones con valor igual a $t_{(i)}$, fallos o censuras, se contabilizan como individuos en riesgo para calcular n_i, ya que han tenido la posibilidad de fallar en ese instante.

Con esta notación, el estimador Kaplan-Meier de $S(t)$ se define como,

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

En la Figura 2.1 se representa el estimador Kaplan-Meier de $S(t)$ con sus intervalos de confianza (más adelante se ve cómo se calculan) correspondientes a los datos del Ejemplo 2 de [7] sobre el tiempo, en meses, que tardan los clientes de una compañía en cancelar su suscripción con ella: 0.5, 1, 3*, 10, 10, 10*, 11, 13.5, 14, 19, 19.5, 30. Las observaciones censuradas (*) corresponden a clientes que todavía no se han marchado y solo se sabe que el instante en el que cancelarán la suscripción es mayor que el tiempo que llevan hasta el momento.

Propiedades del estimador Kaplan-Meier

- Es una función constante entre los tiempos de fallo consecutivos.
- Toma el valor 1 antes del menor tiempo de fallo, $t_{(1)}$.
- Su valor decrece, según el factor variable $\frac{n_i - d_i}{n_i}$, en cada instante de fallo $t_{(i)}$.
- No cambia de valor en los instantes donde ha ocurrido una censura, pero las observaciones censuradas influyen en el estimador a través de los valores n_i.

Notas:

1. Cuando el mayor de los tiempos observados en la muestra, t_M, es un fallo, la estimación KM toma el valor cero a partir de ese instante. Si t_M es una observación censurada, $\hat{S}(t)$ no está bien definido para $t > t_M$ puesto que $\hat{S}(t) = \hat{S}(t_M)$ es constante $\neq 0$ y produce un sesgo positivo ya que $\lim_{t \to \infty} S(t) = 0$. Tampoco se puede tomar $\hat{S}(t) = 0$ para $t > t_M$ porque se produce un sesgo negativo.
2. En una muestra sin observaciones censuradas, el estimador Kaplan-Meier coincide con el estimador natural de $S(t)$, la función de supervivencia empírica, $\hat{S}(t) = \frac{n - \sum_{k=1}^{j} d_k}{n}$ para $t_{(j)} \leq t < t_{(j+1)}$.
3. El estimador KM admite una formulación alternativa, menos intuitiva pero mas sencilla de calcular y programar: sean $t_1 \leq t_2 \leq \ldots \leq t_n$ los n valores de T observados en la muestra ordenada en orden creciente (considerando que las censuras que tienen un valor igual al tiempo de fallo de otro elemento son mayores que estos), el estimador KM se puede expresar como,

$$\hat{S}(t) = \prod_{\{r|t_r \text{ es tiempo de fallo} y t_r \leq t\}} \frac{n - r}{n - r + 1}.$$

Esta expresión es equivalente a la definición y proporciona exactamente las mismas estimaciones.

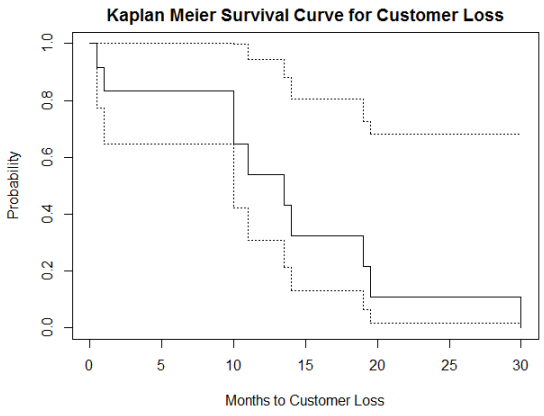

Figura 2.1: Estimación de $S(t)$ y los intervalos de confianza. [7]

Intervalos de confianza para $S(t)$

El mejor estimador de la varianza del estimador KM de $S(t)$ para un tiempo t fijo es el que proporciona la fórmula de Greenwood [1, pág. 23–25]:

$$\hat{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{\{i|t_{(i)}\leq t\}} \frac{d_i}{n_i(n_i-d_i)}.$$

Proposición 1. En ausencia de censura, la expresión anterior se reduce a $\hat{S}(t)(1-\hat{S}(t))/n$.

Demostración. En efecto, como $n_i - d_i = n_{i+1}$ y $\hat{S}(t) = \dfrac{n_{s+1}}{n}$ con $s = \text{máx}\{i \mid t_i \leq t\}$, se tiene

$$\sum \frac{d_i}{n_i(n_i-d_i)} = \sum \left(\frac{1}{n_i-d_i} - \frac{1}{n_i}\right) = \sum \left(\frac{1}{n_{i+1}} - \frac{1}{n_i}\right) = \frac{1}{n_{s+1}} - \frac{1}{n} = \frac{1}{n_{s+1}}(1-\hat{S}(t)) = \frac{1}{n\hat{S}(t)}(1-\hat{S}(t)).$$

Finalmente, $\hat{V}[\hat{S}(t)] = [\hat{S}(t)]^2 \sum \dfrac{d_i}{n_i(n_i-d_i)} = [\hat{S}(t)]^2 \dfrac{1}{n\hat{S}(t)}(1-\hat{S}(t)) = \hat{S}(t)(1-\hat{S}(t))/n.$ □

Bajo la hipótesis de que la censura de la muestra es aleatoria, el estimador $\hat{S}(t)$ es asintóticamente normal para un t fijo; utilizando esta propiedad se puede construir un intervalo de confianza aproximado para $S(t)$ en un tiempo t fijo; a un nivel del $100(1-\alpha)\,\%$ el intervalo es,

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

donde $z_{1-\alpha/2}$ es el cuantil correspondiente de la distribución $N(0,1)$ y el error estándar, $s.e.[\hat{S}(t)]$, se calcula utilizando la fórmula de Greenwood, $\sqrt{\hat{V}[\hat{S}(t)]}$.

Este intervalo aproximado no es demasiado satisfactorio en el caso de muestras pequeñas dada la normalidad asintótica de $\hat{S}(t)$; además, si t es un valor extremo, puede incluir valores fuera del rango $(0,1)$; la solución mas rápida es reemplazar los limites fuera del rango $(0,1)$ por 0 ó 1. Una alternativa más adecuada consiste en aplicar a $S(t)$ una transformación biyectiva $g : (0,1) \to \mathbb{R}$, calcular un intervalo de confianza para $g[S(t)]$,

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

y aplicar g^{-1} a los limites del intervalo calculado. La transformación más habitual es $g(x) = \ln[-\ln(x)]$, que transforma el intervalo $(0,1)$ en $(\infty, -\infty)$. El error estándar de $[g[\hat{S}(t)]]$ se calcula mediante el método delta: $V[g[X]] = (g'(E[X]))^2 V[X]$.

2.1.2. Estimación de otras funciones y parámetros

Función de riesgo acumulado

Existen dos posibles estimadores de $H(t)$:

- Estimador Kaplan-Meier de $H(t)$:

  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.

- Estimador Nelson-Aalen de $H(t)$:

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

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

que va sumando la función de riesgo empírica d_j/n_j en los sucesivos instantes de fallo $t_{(j)}$. Una estimación de la varianza de este estimador es, $\hat{V}(\tilde{H}(t)) = \sum_{j,t_{(j)}\leq t} \frac{d_j}{n_j^2}$ [2, pág. 109]. A partir de este estimador de $H(t)$ es posible obtener un estimador alternativo de la función de supervivencia utilizando ahora la relación inversa, $\tilde{S}(t) = \exp(-\tilde{H}(t))$.

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 mas inestables, la diferencia entre ambos será, por lo general, pequeña.

Tiempo medio de vida

Dado que la esperanza del tiempo de vida se define como $\mu = E[T] = \int_0^\infty S(t)dt$, un estimador razonable de μ es,

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

El calculo de esta integral no resulta complicado ya que la integral es el área comprendida entre los ejes y $\hat{S}(t)$; como esta función es constante a trozos, la integral es la suma de las áreas de los rectángulos que tienen por base $t_{(i)} - t_{(i-1)}$ y por altura $\hat{S}(t_{(i-1)})$, es decir,

$$\hat{\mu} = \sum_{i=1}^{s} (t_{(i)} - t_{(i-1)})\hat{S}(t_{(i-1)}).$$

Este estimador es adecuado solo en el caso de que el máximo valor de la muestra, t_M, corresponda a un instante de fallo, $t_{(s)}$, ya que en ese caso $\hat{S}(t)$ es nula a partir de ese valor; si el máximo es una observación censurada, $\hat{S}(t)$ no está definido para $t > t_M$ y la integral anterior, hasta ∞, no puede calcularse. En esta situación, solo se puede estimar la media del tiempo de vida restringida a un intervalo $[0,L]$, $\mu_L = E[\text{mín}(T,L)] = \int_0^L S(t)dt$ y su estimador es $\hat{\mu_L} = \int_0^L \hat{S}(t)dt$. Si se toma $L = t_M$, μ_{t_M} será una buena aproximación de $E[T]$ si $P(T > t_M)$ es pequeña.

El estimador de la varianza de $\hat{\mu}$ más habitual es,

$$\hat{V}(\hat{\mu}) = \sum_{r\in F} A_r^2 \frac{d_r}{n_r(n_r - d_r)} = \sum_{r\in F} \frac{A_r^2}{(n-r)(n-r+1)},$$

donde $F = \{r \mid t_r \text{ es un tiempo de fallo de la muestra ordenada}\}$ y $A_r = \int_{t_r}^\infty \hat{S}(t)dt$ [6, pág. 118]. Este estimador es sesgado por lo que se suele corregir su sesgo multiplicándolo por $n_f/(n_f - 1)$, siendo n_f el número de instantes de fallo de la muestra, es decir, el número de sumandos del estimador [8, pág. , 332].

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 $\hat{S}(t_{(j)}) \leq 1 - p$.

Veamos cómo construir un intervalo de confianza aproximado para t_p. El procedimiento, propuesto por Brookmeyer, se basa en contrastar, para cada t, la hipótesis nula $H_0 : S(t) = 1 - p$, frente a la alternativa $H_1 : S(t) \neq 1 - p$, utilizando la normalidad asintótica del estimador KM de $S(t)$. El intervalo de confianza al $100(1-\alpha)\,\%$ para t_p está formado por los valores t para los que no se rechaza la hipótesis nula, es decir, que satisfacen,

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

donde el error estándar de $\hat{S}(t)$ se puede calcular a partir de la formula de Greenwood. En la práctica, para calcular el intervalo se comprueban si verifican la condición anterior solamente los valores de t que están en la muestra: el extremo inferior y superior del intervalo serán, respectivamente, el primer y el último valor de la muestra ordenada que la verifiquen.

2.2. Distribuciones de probabilidad para las variables tiempos de fallo T

Para las estimaciones paramétricas, hay que seleccionar una distribución conocida y entonces obtener estimadores de sus parámetros e intervalos de confianza. En esta sección se describen las distribuciones de probabilidad más frecuentes que presentan las variables tiempos de fallo T y un procedimiento para seleccionar la más adecuada.

2.2.1. Distribución exponencial

La distribución Exponencial de parámetro λ es la mas sencilla ya que supone que la función de riesgo es constante,

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

con λ una constante positiva.

Utilizando las relaciones entre las funciones vistas en el capítulo 1, las restantes funciones que caracterizan esta distribución son,

$$\begin{aligned} S(t) &= \exp(-\int h(t)dt) = \exp(-\int \lambda dt) = \exp(-\lambda t) \\ f(t) &= h(t)S(t) = \lambda exp(-\lambda t) \\ H(t) &= \int_0^t h(x)dx = \lambda t \qquad \text{para } 0 \leq t < \infty. \end{aligned}$$

Su media es $1/\lambda$ y su varianza $1/\lambda^2$. Otra propiedad importante que caracteriza esta distribución, es la ausencia de memoria, que implica que en cualquier instante t, la variable tiempo de vida restante, R_t, sigue también una distribución $Exp(\lambda)$. Otra parametrización frecuente de esta distribución utiliza un parámetro μ igual a su media, es decir, $\lambda = 1/\mu$.

Existen dos funciones que generalizan la Exponencial: Weibull y Gamma.

2.2.2. Distribución Weibull

La distribución Weibull generaliza la condición de riesgo constante con una función de riesgo de la forma,

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

donde el parámetro $\lambda > 0$ es el parámetro de escala y $\gamma > 0$ es el parámetro de forma. Esta función es siempre monótona: creciente si $\gamma > 1$ y decreciente si $\gamma < 1$. Si $\gamma = 1$, coincide con la distribución $Exp(\lambda)$. Las expresiones de las restantes funciones son,

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

En la Figura 2.2 se representan las funciones de densidad, riesgo y supervivencia de la distribución Weibull con $\lambda = 1$ y distintos valores de γ. Se puede observar la gran multitud de formas que puede tomar $f(t)$ dependiendo del parámetro de forma γ. Esta diversidad y la sencillez de sus funciones hace que sea una de las distribuciones más utilizadas para modelizar tiempos de fallo, sobre todo en Fiabilidad.

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

Otra parametrización frecuente de esta distribución es $h(t) = \lambda'\gamma' t^{\gamma'-1}$, que implica $S(t) = \exp[-\lambda' t^{\gamma'}]$, y la relación entre los parámetros es $\gamma' = \gamma$ y $\lambda' = \lambda^{\gamma}$.

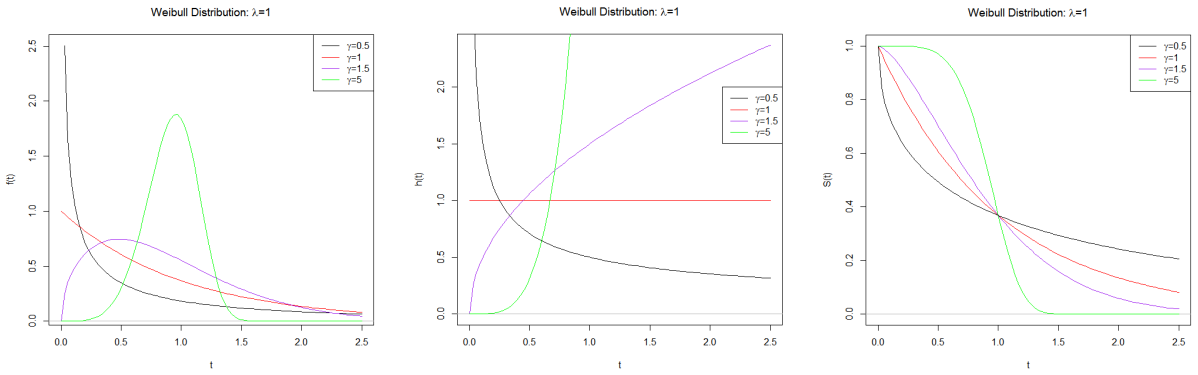

Figura 2.2: Gráfica de las funciones de densidad (izda.), riesgo (centro) y supervivencia (dcha.) de la distribución Weibull con $\lambda = 1$ y distintos valores de γ.

2.2.3. Distribución Gamma

La distribución Gamma está también caracterizada por dos parámetros positivos γ y λ. Su función de densidad es,

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

Su media y varianza son γ/λ y γ/λ^2.

La distribución Gamma, como la Weibull, es de riesgo creciente si $\gamma > 1$ y de riesgo decreciente si $\gamma < 1$. Si $\gamma = 1$ se obtiene la distribución $Exp(\lambda)$. Además $h(t)$ tiende a λ al crecer t.

Aunque en otros ámbitos la distribución Gamma es muy importante, en el análisis de los tiempos de fallo no es muy útil ya que para obtener expresiones de las funciones de riesgo y supervivencia se requiere calcular integrales y resulta complicado. Además, en general, produce estimaciones muy parecidas a las de la distribución Weibull, cuya inferencia es más sencilla, por lo que esta es más utilizada.

Además de estas distribuciones, existen otras dos que se construyen especificando la distribución de $\ln T$, para así asegurar que T solo toma valores positivos: las distribuciones Lognormal y Loglogística.

2.2.4. Distribución Lognormal

Diremos que T sigue una distribución Lognormal de parámetros μ y σ si $\ln(T)$ tiene una distribución $N(\mu,\sigma)$. La función de riesgo se caracteriza porque h(0) = 0, crece hasta alcanzar un máximo y a continuación decrece de forma que $\lim_{t\to\infty} h(t) = 0$.

La distribución Lognormal presenta el mismo inconveniente que la distribución Gamma: su función de supervivencia incluye una integral, lo que dificulta la inferencia, por lo que no es muy utilizada.

2.2.5. Distribución Loglogística

Se dice que T tiene una distribución Loglogística si la variable $\ln(T)$ tiene una distribución Logística. La distribución Logística, como la Normal, es una distribución con parámetros de localización μ y escala σ, es decir, $Y = \mu + \sigma W$, donde W es la distribución Logística estándar. W tiene una función de densidad simétrica muy parecida a la de la $N(0,1)$, excepto en las colas. Su función de densidad es

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

Las funciones que caracterizan la distribución Loglogística de parámetros $\theta = -\mu/\sigma, (-\infty < \theta < \infty)$ y $\kappa = 1/\sigma, (\kappa > 0)$ son,

$$S(t) = (1+e^{\theta}t^{\kappa})^{-1}, \quad f(t) = -\frac{dS(t)}{dt} = \frac{e^{\theta}\kappa t^{\kappa-1}}{(1+e^{\theta}t^{\kappa})^2}, \quad h(t) = \frac{f(t)}{S(t)} = \frac{e^{\theta}\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}}, \quad \text{para } t > 0.$$

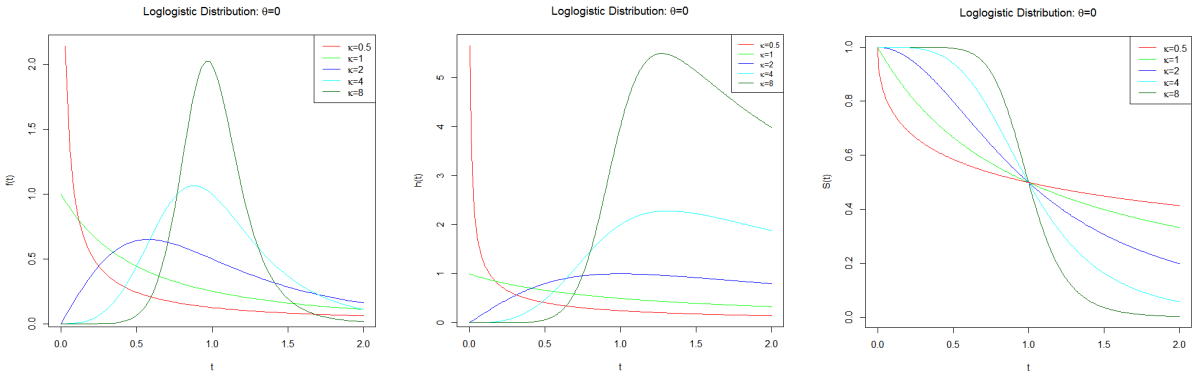

Figura 2.3: Gráficas de las funciones de densidad (izda.), riesgo (centro) y supervivencia (dcha.) de la distribución Loglogística con $\theta = 0$ y distintos valores de κ.

En la Figura 2.3 se representan las funciones de densidad, riesgo y supervivencia de la distribución Loglogística con $\theta = 0$ y distintos valores de κ. Como podemos observar, la función de riesgo es monótona decreciente si $\kappa \leq 1$. Si $\kappa > 1$, la función tiene el mismo comportamiento que el riesgo de la Lognormal, tiene un único máximo, lo que permite modelizar situaciones que presentan dos fases, una fase inicial con riesgo creciente a la que sigue otra con riesgo decreciente.

Dada la similitud existente entre las distribuciones Normal y Logística, la distribución Loglogística suele producir resultados similares a los que se obtienen con la Lognormal y, dada su mayor sencillez de cálculo, se utiliza más.

2.2.6. Selección de una distribución para T

Para escoger un modelo paramétrico para una variable T cuya distribución es desconocida, se deben comparar distintas distribuciones y elegir la más adecuada. Un criterio del que ya hemos hablado es el de la sencillez de las expresiones de las funciones de riesgo, supervivencia y densidad, según el cual elegiríamos el modelo Weibull o el Loglogístico antes que el Gamma o el Lognormal, cuando produzcan ajustes similares.

En este apartado se propone un procedimiento gráfico para analizar la bondad de ajuste de una distribución que se basa en comparar su función de supervivencia $S(t)$ con un estimador no paramétrico de $S(t)$ calculado a partir de la muestra obtenida. En primer lugar se debe encontrar una función de $S(t)$ que sea una función lineal de una función del tiempo, es decir, encontrar una expresión de la forma $g1(S(t)) = a + b \cdot g2(t)$, la cual se obtiene a partir de la ecuación de la función de supervivencia $S(t)$ de la distribución ensayada. A continuación, hay que representar $g1[\hat{S}(t_i)]$ frente a $g2(t_i)$, donde t_i son los instantes de fallo de la muestra, para así poder comparar la función de supervivencia $S(t)$ con el estimador no paramétrico $\hat{S}(t)$. Si la nube de puntos es aproximadamente una línea recta, la distribución ensayada es adecuada. Es importante señalar que para poder representar las funciones $g1$ y $g2$, estas no deben depender de ningún parámetro desconocido relacionado con la distribución cuya bondad de ajuste queremos analizar. A continuación se detallan los procedimientos para algunas distribuciones.

Exponencial

La función de supervivencia exponencial es $S(t) = \exp(-\lambda t)$. Tomando logaritmos en esta expresión se tiene,

$$\ln[S(t)] = -\lambda t.$$

De esta relación se deduce que, si la hipótesis de exponencialidad es cierta, al representar $\ln[\hat{S}(t_i)]$ frente a t_i debemos obtener una nube de puntos aproximadamente lineal de pendiente $-\lambda$. En este caso, $g1(x) = \ln(x)$ y $g2(x)$ es la función identidad. Además, la pendiente de la recta de regresión ajustada a la nube de puntos, $\hat{\beta}_1$, proporciona un estimador preliminar de λ, $\hat{\lambda} = -\hat{\beta}_1$.

Weibull

La función de supervivencia es $S(t) = \exp[-(\lambda t)^{\gamma}]$. Tomando logaritmos, $-\ln[S(t)] = (\lambda t)^{\gamma}$. Tomando logaritmos de nuevo y operando en el segundo miembro, $\ln[-\ln[S(t)]] = \ln[(\lambda t)^{\gamma}] = \gamma\ln(\lambda t) = \gamma\ln(\lambda) + \gamma\ln(t)$, y se obtiene la expresión

$$\ln[-\ln[S(t)]] = \gamma\ln(\lambda) + \gamma\ln(t).$$

De esta relación se deduce que, si la distribución Weibull es adecuada, al representar $\ln[-\ln[\hat{S}(t_i)]]$ frente a $\ln(t_i)$ debemos obtener una nube de puntos aproximadamente lineal de pendiente γ y ordenada en el origen $\gamma\ln(\lambda)$. En este caso, $g1(x) = \ln[-\ln[x]]$ y $g2(x) = \ln(x)$. Los estimadores de los parámetros de la recta de regresión ajustada a la nube de puntos, $\hat{\beta}_0$ y $\hat{\beta}_1$, proporcionan dos estimadores de γ y λ, $\hat{\gamma} = \hat{\beta}_1$ y $\hat{\lambda} = \exp(\hat{\beta}_0/\hat{\beta}_1)$.

Loglogística

La función de supervivencia es $S(t) = (1 + e^{\theta}t^{\kappa})^{-1}$. Por lo tanto, $\frac{1}{S(t)} - 1 = e^{\theta}t^{\kappa}$. Tomando logaritmos y operando en el segundo miembro, $\ln\left(\frac{1}{S(t)} - 1\right) = \ln(e^{\theta}t^{\kappa}) = \ln(e^{\theta}) + \ln(t^{\kappa}) = \theta + \kappa\ln(t)$, y se obtiene la expresión

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

De esta relación se deduce que, si la distribución Loglogística es adecuada, al representar $\ln\left(\frac{1-\hat{S}(t_i)}{\hat{S}(t_i)}\right)$ frente a $\ln(t_i)$ debemos obtener una nube de puntos aproximadamente lineal de pendiente κ y ordenada en el origen θ. En este caso, $g1(x) = \ln(\frac{1-x}{x})$ y $g2(x) = \ln(x)$. Los estimadores de los parámetros de la recta de regresión ajustada a la nube de puntos, $\hat{\beta}_0$ y $\hat{\beta}_1$, proporcionan dos estimadores de θ y κ, $\hat{\theta} = \hat{\beta}_0$ y $\hat{\kappa} = \hat{\beta}_1$.

En el caso de las distribuciones Lognormal y Gamma no existe una relación exacta pero se pueden utilizar relaciones basadas en aproximaciones a la distribución Normal:

Lognormal

Si T tiene una distribución Lognormal de parámetros μ y σ, se tiene que, $S(t) = 1 - P(T \leq t) = 1 - P(\ln(T) \leq \ln(t)) = 1 - P\left(\frac{\ln(T)-\mu}{\sigma} \leq \frac{\ln(t)-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)$, siendo Φ la función de distribución de una variable Normal estándar; de esta relación se deduce que,

$$\Phi^{-1}(1 - S(t)) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\ln(t).$$

En consecuencia, para verificar la adecuación de la distribución Lognormal se analiza la linealidad de la nube de puntos $\Phi^{-1}(1 - \hat{S}(t_i))$ frente a $\ln(t_i)$, y los estimadores de μ y σ son respectivamente, $-\hat{\beta}_0/\hat{\beta}_1$ y $1/\hat{\beta}_1$.

2.3. Estimación e inferencia paramétrica

Una vez seleccionado un modelo para la variable T hay que estimar sus parámetros a partir de la muestra. En esta última sección del capítulo, se emplea el método de máxima verosimilitud para realizar las estimaciones. En primer lugar, se presenta la función de verosimilitud generalizada para muestras con datos censurados; el resto del análisis es análogo al de muestras completas. Finalmente, se utiliza la distribución asintótica normal de los estimadores máximo verosímiles (EMV) para calcular intervalos de confianza de los parámetros y de funciones de los parámetros como podrían ser la función de supervivencia, la media de T y demás funciones y medidas de la distribución de T.

2.3.1. Estimación paramétrica en muestras con censura

En la inferencia paramétrica se asume que la distribución de la variable T es conocida excepto por un vector de parámetros Θ. El método de estimación más utilizado es el de máxima verosimilitud, que puede aplicarse en situaciones bastante generales y, en particular, puede adaptarse a la existencia de censura.

Una propiedad muy importante de los estimadores máximo verosímiles, EMV, es la **invarianza funcional**, que garantiza que si g es una función biyectiva que toma valores reales, el EMV de $g(\Theta)$ se puede calcular aplicando g al EMV $\hat{\Theta}$, $\widehat{g(\Theta)} = g(\hat{\Theta})$. Esta propiedad permite estimar $S(t), h(t)$, la esperanza, la mediana o cualquier otra medida relacionada con T que sea una función biyectiva de los parámetros, a partir de los EMV de los parámetros de la distribución.

Los EMV de los parámetros Θ son los valores $\hat{\Theta}$ que maximizan la función de verosimilitud $L(\Theta)$, que se define en el siguiente apartado. Para calcular los valores que maximizan esa función es más cómodo operar con la función de logverosimilitud, $l(\Theta) = \ln[L(\Theta)]$. Los valores que maximizan estas dos funciones son los mismos dada la monotonía de la función logaritmo y esta tiene la ventaja de que la estructura aditiva resulta más sencilla para derivar. Los valores que maximizan la función son aquellos que anulan las derivadas parciales de $l(\Theta)$ respecto a cada θ_i, por lo que para calcularlos hay que resolver el sistema formado por las m ecuaciones de verosimilitud,

$$\frac{\partial l(\Theta)}{\partial \theta_i} = 0 \quad i = 1, \ldots, m$$

El vector $\left(\frac{\partial l(\Theta)}{\partial \theta_1}, \ldots, \frac{\partial l(\Theta)}{\partial \theta_m}\right)'$ de las derivadas parciales se denomina vector de puntuaciones (*score*).

Función de verosimilitud

El procedimiento descrito es el mismo para el caso de muestras completas y para el de muestras con censura; el único cambio se da en la definición de la función de verosimilitud.

Se supone que la distribución de probabilidad es continua, que es conocida excepto por un vector de parámetros $\Theta' = (\theta_1, \ldots, \theta_m)$ y que $f(t;\Theta)$ es su función de densidad. Si la muestra no contiene observaciones censuradas, la función de verosimilitud de esa muestra, $L(\Theta)$, se define como

$$L(\Theta) = \prod_{i=1}^{n} L_i(\Theta) = \prod_{i=1}^{n} f(t_i;\Theta)$$

Para generalizar el caso anterior a muestras que contienen observaciones censuradas (a derecha), tengamos en cuenta que la información que aporta una de estas observaciones t_i^* es que $T > t_i^*$. Como el objetivo será encontrar los parámetros que hacen máxima la probabilidad conjunta de que sucedan todas las observaciones de la muestra, la verosimilitud en las observaciones censuradas se define como $P(T > t_i^*)$, esto es, la función de supervivencia en ese instante. Por lo tanto, la función de verosimilitud de la muestra se define como

$$L(\Theta) = \prod_{i=1}^{n} L_i(\Theta) = \prod_{t_i \in F} f(t_i;\Theta) \prod_{t_i^* \in C} S(t_i^*;\Theta)$$

donde F es el conjunto de observaciones de la muestra que son fallos y C es el conjunto de observaciones que son censuradas. Una formulación alternativa es

$$L(\Theta) = \prod_{i=1}^{n} f(t_i;\Theta)^{c_i} S(t_i;\Theta)^{1-c_i}$$

donde c_i es 1 si t_i es un tiempo de fallo y 0 si t_i es una censura.

La función de logverosimilitud con la que se opera queda,

$$l(\Theta) = \sum_{i=1}^{n} \ln[L_i(\Theta)] = \sum_{t_i \in F} \ln[f(t_i;\Theta)] + \sum_{t_i^* \in C} \ln[S(t_i^*;\Theta)] \tag{2.1}$$

En general, hay que resolver sistemas de ecuaciones no lineales cuya resolución requiere algoritmos numéricos. Sin embargo, para la distribución exponencial los cálculos son sencillos y pueden obtenerse expresiones analíticas de los estimadores.

Dada una muestra de tamaño n de una variable T con distribución $Exp(\lambda)$, en la que hay r observaciones completas t_i, y $n-r$ observaciones censuradas t_j^*, la verosimilitud de la muestra es,

$$L(\lambda) = \prod_{i=1}^{r} \lambda \exp(-\lambda t_i) \prod_{j=1}^{n-r} \exp(-\lambda t_j^*),$$

y la logverosimilitud,

$$l(\lambda) = \ln(L(\lambda)) = r\ln(\lambda) - \lambda \left(\sum_{i=1}^{r} t_i + \sum_{j=1}^{n-r} t_j^* \right) = r\ln(\lambda) - \lambda S,$$

con $S = \sum_{i=1}^{r} t_i + \sum_{j=1}^{n-r} t_j^*$ la suma de todas las observaciones de la muestra. Resolviendo $\frac{dl}{d\lambda}(\lambda) = \frac{r}{\lambda} - S = 0$, se deduce que el EMV de λ es $\hat{\lambda} = r/S$. Utilizando la propiedad de invarianza se pueden obtener los EMV de cualquier función o estadístico que sea una función biyectiva del parámetro λ. Así, se obtienen, por ejemplo, los estimadores $\hat{S}(t) = \exp(-\hat{\lambda}t)$, $\hat{h}(t) = \hat{\lambda}$ o $\widehat{E(T)} = 1/\hat{\lambda}$.

2.3.2. Inferencia paramétrica en muestras con censura

El procedimiento de inferencia que se detalla en este apartado se basa en la distribución asintótica Normal de $\hat{\Theta}$, el EMV de Θ. Para calcular intervalos de confianza y contrastes de hipótesis se necesitan los estimadores de los parámetros y los estimadores de sus varianzas, que se encuentran en la matriz de varianzas-covarianzas $\hat{V}(\hat{\Theta}) = I^{-1}$, donde $I = (I_{ij})_{m\times m}$ con $I_{ij} = -\frac{\partial^2}{\partial\theta_i\theta_j} l(\hat{\Theta})$ [5, pág. 34].

Los EMV son asintóticamente normales e insesgados ($E(\hat{\Theta}) = \Theta$). En consecuencia, para un parámetro θ_k, se tiene que $(\hat{\theta}_k - \theta_k)/s.e.[\hat{\theta}_k]$ con $s.e.[\hat{\theta}_k] = \sqrt{\hat{V}[\hat{\theta}_k]}$ es asintóticamente $N(0,1)$. Así, un intervalo de confianza para θ_k al nivel α de confianza se puede calcular de la forma habitual como

$$\hat{\theta}_k \pm z_{1-\alpha/2} s.e.[\hat{\theta}_k]$$

donde $z_{1-\alpha/2}$ es el correspondiente percentil de la distribución Normal estándar.

Intervalos de confianza para funciones de los parametros

Por la propiedad de invarianza funcional, se tiene que $g(\hat{\Theta})$ es el EMV de $g(\Theta)$ y de aquí se deduce que, asintóticamente,

$$g(\hat{\Theta}) \sim N(g(\Theta), \hat{V}[g(\hat{\Theta})]).$$

De esta afirmación, se obtiene inmediatamente el intervalo de confianza asintótico,

$$g(\hat{\Theta}) \pm z_{1-\alpha/2} s.e.[g(\hat{\Theta})],$$

donde $s.e.[g(\hat{\Theta})]$ se calcula con el método delta: $V[g[X]] = \sum_{i=1}\sum_{j=1} \frac{\partial g}{\partial X_i}(E[X_i]) \; \frac{\partial g}{\partial X_j}(E[X_j]) \; Cov[X_i, X_j]$.

Solo podemos garantizar que la aproximación es buena si la muestra es razonablemente grande; en otro caso, conviene utilizar otros procedimientos, como los intervalos de confianza basados en el estadístico de razón de verosimilitud.

Capítulo 3

Comparación de la supervivencia de dos o más grupos

En este capítulo se presentan dos herramientas para comparar la supervivencia de grupos definidos por un factor, por ejemplo el sexo, el tratamiento utilizado o el proceso de fabricación. La primera de ellas es la familia de contrastes log-rank, que es no paramétrica, por lo que se utiliza cuando no se tiene ninguna hipótesis sobre la distribución de T. La segunda es el test de razón de verosimilitud para comparar dos grupos de forma paramétrica, es decir, requiere establecer previamente una hipótesis sobre la familia a la que pertenece la distribución de T.

3.1. Comparación no paramétrica: familia de tests log-rank

Supongamos que se quiere comparar la supervivencia en G poblaciones de individuos; sean $t_{(1)} < \ldots < t_{(j)} < \ldots < t_{(J)}$ los J instantes de fallo distintos observados en la muestra conjunta de los G grupos, dispuestos en orden creciente.

Para cada $j = 1,\ldots,J$, $i = 1,\ldots,G$ se define d_{ij} como el número de fallos ocurridos en la muestra i en el instante $t_{(j)}$ y n_{ij} como el número de individuos en riesgo en la muestra i en el instante $t_{(j)}$. Además, para cada $j = 1,\ldots,J$, d_j es el número total de fallos en ese instante, es decir, $d_j = \sum_{i=1}^{G} d_{ij}$ y n_j es el número total de individuos en riesgo en ese instante, $n_j = \sum_{i=1}^{G} n_{ij}$.

La hipótesis que se quiere contrastar es que el tiempo de fallo de todos los grupos tiene la misma distribución. Como una distribución queda determinada unívocamente por cualquiera de las funciones que la caracterizan, la hipótesis se puede plantear en términos de cualquiera de ellas, aunque habitualmente se utiliza la función de riesgo. Por lo tanto, la hipótesis nula es,

$$H_0 : h_1(t) = h_2(t) = \ldots = h_G(t) \quad \forall t$$

frente a la alternativa de que el riesgo de al menos uno de los grupos difiera de los demás en algún instante t.

Para contrastar esta hipótesis se construye un estadístico basado en la discrepancia entre lo observado y lo esperado bajo H_0, en concreto, entre el número de fallos observado y esperado en cada grupo en cada uno de los instantes de fallo.

Para j fijo, supongamos conocido d_j y los valores n_{ij} $\forall i$. El vector aleatorio $(d_{1j}, d_{2j}, \ldots, d_{Gj})$ que representa el número de fallos ocurridos en cada grupo en el instante $t_{(j)}$, queda determinado por $G-1$ de las componentes, ya que una de ellas queda determinada por las $G-1$ restantes debido a la relación $d_j = \sum_{i=1}^{G} d_{ij}$ y podemos no tenerla en cuenta para construir el estadístico ya que no se pierde información.

Recordemos, en primer lugar, la definición de la distribución hipergeométrica multivariante. Consideremos una población de N elementos de lo cuales N_1 son de tipo 1, N_2 son de tipo 2,..., y N_G son de tipo G y en la que la probabilidad de cada elemento de ser extraído es la misma; si extraemos de dicha población una muestra de tamaño n (sin reemplazo), el vector aleatorio $(X_1,\ldots,X_{G-1})$ dado por

X_i = *número de elementos de tipo i en la muestra de tamaño n*, tiene una distribución hipergeométrica de dimensión $G-1$ de parámetros $N, n, N_1, \ldots, N_{G-1}$.

Esta situación es análoga al caso en el que tenemos $N = n_j$ individuos en riesgo en el instante $t_{(j)}$. De esos individuos, $N_1 = n_{1j}$ son del grupo 1, $N_2 = n_{2j}$ son del grupo 2,..., y $N_G = n_{Gj}$ son del grupo G. De esos individuos, hay $n = d_j$ que fallan simultáneamente. Si la probabilidad de cada individuo de fallar es la misma (es decir, si la distribución del tiempo de fallo de los individuos de todos los grupos es la misma, esto es, si H_0 es cierta), el vector aleatorio $(X_1, \ldots, X_{G-1})$ dado por $X_i = d_{ij}$ *número de individuos del grupo i que han fallado entre los* d_j sigue una distribución hipergeométrica de dimensión $G-1$ de parámetros $n_j, d_j, n_{1j}, \ldots, n_{(G-1)j}$; luego cada d_{ij} tiene una distribución hipergeométrica de parámetros n_j, n_{ij}, d_j.

Por lo tanto, bajo la hipótesis nula, el número esperado de fallos en el grupo k 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,

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

con $k = 1, \ldots, G-1$ y $l = 1, \ldots, G-1$. [4, pág. 173]

Para obtener una medida global de la discrepancia entre lo observado y lo esperado en cada grupo, se suman las diferencias observadas en los distintos instantes de fallo para cada grupo y se obtiene un vector $U = (U_1, \ldots, U_{G-1})$, cuyas componentes son,

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

donde w_j son los elementos de un vector de pesos que se utiliza para dar más valor a las discrepancias $d_{kj} - e_{kj}$ observadas en los instantes iniciales ya que están calculadas con muestras más grandes (de tamaño n_j) que las de los instantes finales que tienen más variabilidad. Así, con $w_j = 1$ se obtiene el test Mantel-Cox, con pesos $w_j = n_j$ el test de Gehan-Breslow, y con pesos $w_j = \sqrt{n_j}$ el test de Tarone-Ware.

Suponiendo que los J vectores $(d_{1j}, d_{2j}, \ldots, d_{(G-1)j})$ son independientes, $Cov[d_{ij}, d_{kl}] = 0\ \forall i,k$, si $j \neq l$, y por lo tanto $Var[U_i] = \sum_{j=1}^{J} w_j^2 Var[d_{ij}]$ y $Cov[U_i, U_k] = \sum_{j=1}^{J} w_j^2 Cov[d_{ij}, d_{kj}]$, luego la matriz de covarianzas del vector U, V, se calcula a partir de las matrices de covarianzas correspondientes a los J instantes de fallo,

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

Aplicando el teorema central del límite se tiene que la distribución de U es asintóticamente (si el número de tiempos de fallo, J, es suficientemente alto) Normal multivariante de vector de medias nulo bajo H_0. Por ello, se define el estadístico $Q = U'V^{-1}U$ que, bajo H_0, tiene una distribución aproximada χ^2 con $G-1$ grados de libertad.

La hipótesis nula se rechaza si el valor del estadístico Q es muy extremo, es decir, si el p-valor $P(X > \mid q \mid)$, donde $X \sim \chi^2_{G-1}$ y q es el valor observado del estadístico Q, es muy pequeño.

3.2. Comparación paramétrica: test de razón de verosimilitud

Si los tiempos de fallo siguen una distribución conocida, los tests paramétricos son más eficientes que los no paramétricos. Vamos a ver el test de razón de verosimilitud para comparar dos grupos.

Supongamos que la distribución de los tiempos de fallo en los dos grupos es del mismo tipo y que todos los parámetros son desconocidos y queremos contrastar la hipótesis de que los vectores de parámetros son iguales (es decir, que la distribución del tiempo de fallo de los dos grupos es la misma),

$$H_0 : \Theta_1 = \Theta_2 (\equiv \Theta),$$

con Θ desconocido, frente a la alternativa $H_1 : \Theta_1 \neq \Theta_2$. El estadístico del test de razón de verosimilitud asociado es,

$$X_l = -2[l(\hat{\Theta}) - l(\hat{\Theta}_1, \hat{\Theta}_2)]$$

donde:

- $\hat{\Theta}$ representa el EMV del vector de parámetros común Θ obtenido a partir de la muestra conjunta obtenida al unir los dos grupos y $l(\hat{\Theta})$ la función de logverosimilitud asociada a esa muestra evaluada en $\hat{\Theta}$.
- $\hat{\Theta}_1$ es el EMV de Θ_1 calculado solo con la muestra del primer grupo y $l_1(\hat{\Theta}_1)$ su logverosimilitud asociada.
- $\hat{\Theta}_2$ es el EMV de Θ_2 calculado solo con la muestra del segundo grupo y $l_2(\hat{\Theta}_2)$ su logverosimilitud asociada.
- $l(\hat{\Theta}_1, \hat{\Theta}_2)$ es la logverosimilitud conjunta calculada como $l(\hat{\Theta}_1, \hat{\Theta}_2) = l_1(\hat{\Theta}_1) + l_2(\hat{\Theta}_2)$.

Bajo H_0, X_l tiene una distribución asintótica χ^2 con k grados de libertad, siendo k el número de parámetros de Θ.

La hipótesis nula se rechaza a un nivel de significación α si el p-valor $P(\chi^2_k > x_l)$, donde x_l es el valor observado en la muestra del estadístico X_l, es menor que α.

Este procedimiento se puede adaptar al caso en que la hipótesis nula no incluya a todos los parámetros de Θ. Distingamos dos casos:

- Caso 1: Si sabemos que una parte de los parámetros son iguales y conocemos su valor, podemos contrastar si el resto de parámetros que no conocemos son iguales de la misma manera: calculando los EMV de esos parámetros a partir de la muestra conjunta y por separado, el valor de las funciones de logverosimilitud y el estadístico X_l, que tendrá una distribución χ^2 con los mismos grados de libertad que parámetros desconocidos.
- Caso 2: Si solo sabemos que una parte de los parámetros son iguales pero desconocemos su valor, primero se calcula el EMV de todos los parámetros con la muestra conjunta. Los parámetros que ya se sabe que son iguales intervienen con el mismo valor (el EMV obtenido) en la logverosimilitud bajo H_0, l, y bajo H_1, $l_1 + l_2$. A continuación se procede como en el caso anterior.

Capítulo 4

Análisis de supervivencia con R y aplicación a una muestra simulada

En este capítulo se presentan algunas herramientas que tiene el software R para realizar análisis de supervivencia y después se realiza un ejercicio de simulación para comparar el funcionamiento de los métodos de estimación específicos para muestras con censura propuestos en temas anteriores, con los métodos estándar para muestras sin censura, cuando se aplican a muestras censuradas. En primer lugar, se genera una muestra con una distribución conocida con la cual se pueden obtener de forma teórica valores como la media, percentiles de la distribución y su función de supervivencia. A continuación se genera un mecanismo de censura aleatoria independiente y se obtiene la muestra censurada correspondiente. Posteriormente, a partir de esa muestra se estiman los mismos parámetros mediante los métodos paramétricos y no paramétricos para muestras censuradas y mediante los métodos habituales para muestras completas, eliminando los datos censurados o considerándolos fallos. Por último, se comparan los resultados obtenidos con los distintos métodos con los valores teóricos y la función de supervivencia real de la distribución simulada.

4.1. Análisis de supervivencia con R: el paquete *survival*

Para estimar la función de supervivencia y sus intervalos de confianza se utiliza el paquete *survival* [9] de R, que contiene las principales tareas de análisis de supervivencia, como la definición de objetos de supervivencia, curvas de Kaplan-Meier y Aalen-Johansen, modelos Cox, etc. y se carga mediante la orden `library(survival)`. En el tutorial [3] se pueden consultar estas y otras herramientas de análisis de supervivencia con R, como la comparación de la supervivencia de dos o más grupos.

En primer lugar, se deben disponer los datos de la forma adecuada, creando lo que se denomina un objeto de supervivencia, que establezca claramente qué observaciones son fallos y cuales son censuradas. Esto se realiza mediante la función **Surv()** y necesita dos argumentos para muestras censuradas solo a derecha: el primero es un vector que contiene todas las observaciones y el segundo es el vector que indica para cada observación si es un fallo (1) o está censurada (0).

Para obtener el estimador Kaplan-Meier de la función de supervivencia y sus intervalos de confianza se utiliza la función **survfit(formula, conf.int, conf.type)**, donde *formula* es un objeto de supervivencia, *conf.int* es el nivel de confianza de los intervalos (por defecto 0.95) y *conf.type* es la transformación utilizada para construir los intervalos, por ejemplo, para $g(t) = \log(-\log(t))$ es "log-log". Con la orden `plot()` podemos obtener la gráfica del estimador y con la orden `summary()` podemos ver cada tiempo de fallo con sus valores n_i y d_i y la estimación de $S(t)$ en ese instante con su error estándar y los límites de su intervalo de confianza.

4.2. Simulación de una muestra censurada y aplicación de los métodos de estimación con R

En esta sección estimaremos de distintas maneras la función de supervivencia, la media y los cuantiles 0.25, 0.5, 0.75 y 0.95 a partir de una muestra simulada con R.

Para obtener una muestra simulada que contenga datos censurados utilizaremos el modelo de censura aleatoria. En este modelo la censura es independiente de los tiempos de fallo, por lo que se generan los fallos y las censuras por separado. Por ejemplo, para una muestra Exp(0.4) de tamaño 300:

```
y <- rexp(300, rate = 0.4)
cen <- rexp(300, rate = 0.1)
```

La muestra estará formada por el mínimo de cada par:

```
ycen <- pmin(y, cen)
```

Además se genera una variable binaria que indique si la observación es un fallo (1) o una censura (0):

```
di <- as.numeric(y <= cen)
```

La muestra generada contiene un 20 % de censura.

4.2.1. Estimación mediante métodos para muestras completas

Para utilizar los métodos de estimación habituales podemos eliminar los datos censurados o considerarlos fallos.

Opción 1: Eliminar los datos censurados

Para empezar, definimos un subconjunto de `ycen` que contenga solo los fallos, mediante la orden `s1 <- subset(ycen, subset= di == "1")`. La media y los cuantiles de este subconjunto son:

```
> mean(s1)
[1] 1.957189

> quantile(s1, probs=c(.25,.5,.75,.95))
      25%       50%       75%       95%
0.6380547 1.4234583 2.7288136 5.5139521
```

Para estimar la función de supervivencia, buscamos el estimador Kaplan-Meier ya que para muestras no censuradas este coincide con la función de supervivencia empírica. Más adelante obtendremos su gráfica.

```
> my.surv1 <- Surv(s1, di[di==1])

> my.fit1 <- survfit(formula = my.surv1~1, conf.int = 0.95, conf.type = "log-log")
```

Para obtener estimaciones paramétricas, consideramos la distribución Exp(1/1.957189) y calculamos sus cuantiles. Más adelante se obtendrá la gráfica de su función de supervivencia.

```
> qexp(c(.25,.5,.75,.95), rate=1/1.957189, lower.tail=TRUE)
[1] 0.5630482 1.3566200 2.7132401 5.8632143
```

Opción 2: Considerar los datos censurados como fallos

En este caso, calculamos la media y los cuantiles con el vector *ycen* completo:

```
> mean(ycen)
[1] 1.974848

> quantile(ycen, probs=c(.25,.5,.75,.95))
      25%       50%       75%       95%
0.6369582 1.4666644 2.7935258 5.5129800
```

Para estimar $S(t)$ lo hacemos del siguiente modo:

```
> my.surv2 <- Surv(ycen, rep(1,length(ycen)))

> my.fit2 <- survfit(formula = my.surv2~1, conf.int = 0.95, conf.type = "log-log")
```

Para obtener estimaciones paramétricas, consideramos la distribución Exp(1/1.974848) y calculamos sus cuantiles. Más adelante se obtendrá la gráfica de su función de supervivencia.

```
> qexp(c(.25,.5,.75,.95), rate=1/1.974848, lower.tail=TRUE)
[1] 0.5681284 1.3688603 2.7377206 5.9161159
```

4.2.2. Estimación mediante métodos para muestras censuradas

Vamos a realizar una estimación no paramétrica, con el estimador Kaplan-Meier, y una estimación paramétrica, con el método de máxima verosimilitud.

Estimación no paramétrica

En primer lugar, obtenemos el estimador Kaplan-Meier de $S(t)$:

```
> my.surv3 <- Surv(ycen, di)

> my.fit3 <- survfit(formula = my.surv3~1, conf.int = 0.95, conf.type = "log-log")
```

A partir del estimador Kaplan-Meier, obtenemos la media restringida al máximo fallo y los cuantiles:

```
> print(my.fit3, print.rmean=TRUE)
Call: survfit(formula = my.surv3 ~ 1, conf.int = 0.95, conf.type = "log-log")

        n     events     *rmean *se(rmean)     median    0.95LCL    0.95UCL
  300.000    239.000      2.437      0.146      1.779      1.497      2.030
    * restricted mean with upper limit =  10.6

> quantile(my.fit3, probs=c(.25,.5,.75,.95))
$quantile
       25        50        75        95
0.7562912 1.7786439 3.4671224 7.1670871

$lower
       25        50        75        95
0.5734806 1.4974955 2.9798724 5.5878471

$upper
       25        50        75        95
0.8896492 2.0304145 4.0609495 9.7319758
```

Estimación paramétrica

Para la estimación paramétrica, suponemos que la distribución es Exp(λ) y obtendremos el EMV de λ utilizando la función **mle** del paquete *stats4*.

En primer lugar, calculamos la función de logverosimilitud (2.1) negativa. Como la verosimilitud es distinta para fallos y para censuras, se define, además del subconjunto `s1` de los fallos, el subconjunto `s0` que contiene las censuras con la orden `s0 <- subset(ycen, subset= di == "0")`. Para los fallos consideramos la función de densidad y para las censuras la función de supervivencia:

```
> llh_exp <- function(lambda){
+   llh <- -sum(dexp(s1, rate=lambda, log=TRUE))-sum(pexp(s0, rate=lambda, lower = FALSE, log=TRUE))
+   return(llh)}
```

La función **mle** necesita la logverosimilitud negativa como primer argumento y, además, un valor inicial para el parámetro:

```
> fit_exp <- mle(llh_exp, start = list(lambda = 0.1))

> summary(fit_exp)
Maximum likelihood estimation

Call:
mle(minuslogl = llh_exp, start = list(lambda = 0.1))

Coefficients:
        Estimate Std. Error
lambda 0.4034074  0.0260941

-2 log L: 911.9333
```

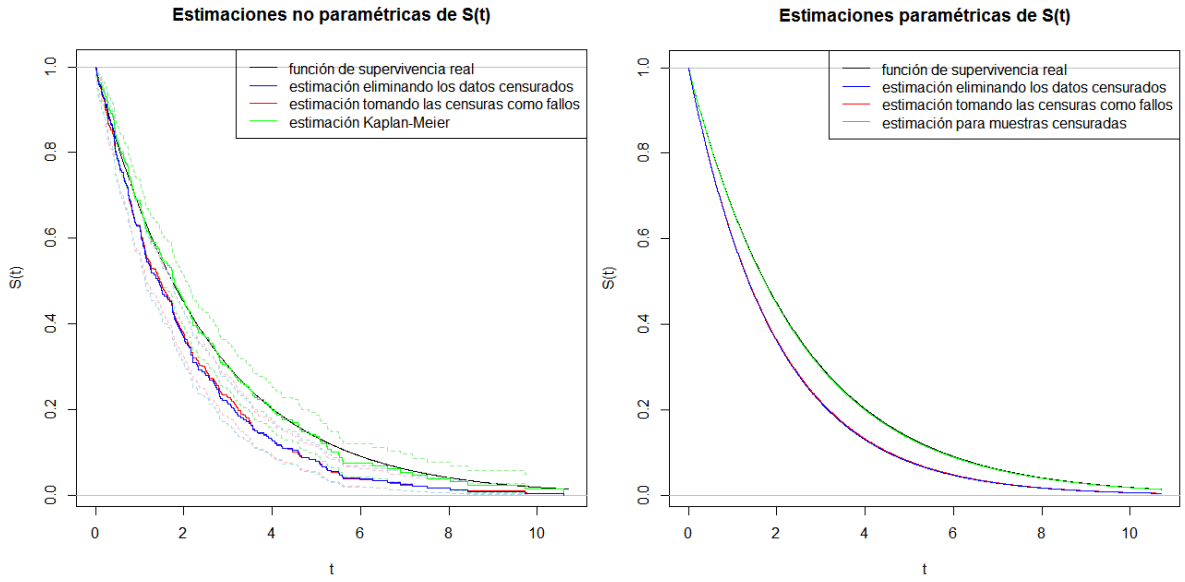

Figura 4.1: Gráficas de la función de supervivencia real y de sus estimaciones no paramétricas (izda.) y paramétricas (dcha.).

El valor estimado de λ es 0.4034074. A partir de este valor, podemos obtener los estimadores de $S(t)$, de la media y de los percentiles:

```
> 1/0.4034074
[1] 2.478884

> qexp(c(.25,.5,.75,.95), rate=0.4034074, lower.tail=TRUE)
[1] 0.7131304 1.7182312 3.4364624 7.4260717
```

4.3. Comparación de los resultados con la distribución real

Por último, comprobemos con qué métodos se han obtenido mejores estimaciones. En la siguiente tabla se recogen la media y los cuantiles obtenidos en las distintas estimaciones, además de los valores teóricos de la distribución Exp(0.4). Además, en la figura 4.1 se representan las gráficas de las distintas estimaciones de la función de supervivencia.

	Media	25 %	50 %	75 %	95 %
Estimación no param. eliminando los datos censurados	1.957	0.638	1.423	2.729	5.514
Estimación no param. tomando las censuras como fallos	1.975	0.637	1.467	2.794	5.513
Estimación Kaplan-Meier	2.437	0.756	1.779	3.467	7.167
Estimación param. eliminando los datos censurados	1.957	0.563	1.357	2.713	5.863
Estimación param. tomando las censuras como fallos	1.975	0.568	1.369	2.738	5.916
EMV para muestras censuradas	2.479	0.713	1.718	3.436	7.426
Valores reales	2.5	0.719	1.733	3.466	7.489

Comparando los distintos resultados, podemos observar que aplicar los métodos habituales a una muestra censurada no produce buenas estimaciones, ya que subestiman todas las medidas y la función de supervivencia. Sin embargo, los métodos para muestras censuradas que se han presentado en este trabajo funcionan bien.

Bibliografía

[1] D. COLLETT, *Modelling Survival Data in Medical Research*, 2ª ed., Chapman & Hall/CRC, 2003.

[2] MARIO CLEVES, WILLIAM GOULD, YULIA MARCHENKO, *An Introduction to Survival Analysis Using Stata*, 2ª ed., Stata Press, 2008.

[3] DAVID M. DIEZ, *Survival Analysis in R*, https://drive.google.com/file/d/1iaovmIcHKnAPlxFaIBOXHyLagV1oXM0e/edit, 2013.

[4] NORMAN L. JOHNSON, S. KOTZ, N. BALAKRISHNAN, *Discrete multivariate distributions* (A Wiley-Interscience Publication), John Wiley & Sons, Inc., 1997.

[5] MARC KERY, J. ANDREW ROYLE, *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS*, Vol.1, Academic Press, 2015.

[6] JOHN P. KLEIN, MELVIN L. MOESCHBERGER, *Survival analysis : techniques for censored and truncated data* (Statistics for biology and health), 2ª ed., Springer, 2003.

[7] P. ROSENMAI, *Calculating Kaplan Meier Survival Curves and Their Confidence Intervals in SQL Server*, https://eurekastatistics.com/calculating-kaplan-meier-survival-curves-and-their-confidence-intervals-in-sql-server/.

[8] STEVE SELVIN, *A Biostatistics Toolbox for Data Analysis*, Cambridge University Press, 2015.

[9] TERRY M THERNEAU, *survival: Survival Analysis*, https://CRAN.R-project.org/package=survival.