Tema 4

Métodos de libre distribución para el análisis de datos de tiempo de vida

4.1. Introducción

En la realización de un estudio analítico de tiempos de vida hay procedimientos paramétricos, no paramétricos y semiparamétricos. En la sección anterior ya hemos visto algunas distribuciones paramétricas utilizadas en este campo. Los procedimientos no paramétricos no dependen de la presunción de una familia específica de distribuciones.

Las técnicas clásicas incluyen el uso de tablas de frecuencias relativas, representaciones y función de distribución empírica.

Las tablas de vida son de gran importancia en estudios demográficos, permitiendo conocer razones de muerte, esperanza de vida, tiempo esperado de vida que resta a una persona a una cierta edad en una población, tiempo esperado de vida si se elimina un riesgo de muerte, estudio del tiempo de permanencia de una persona en el paro, etc. Esto es de mucha utilidad en el planteamiento de distintos aspectos económicos: pensiones, seguridad social, paro, etc.

Cuando se va a realizar una tabla de vida se debe tener en cuenta si el muestreo es completo o los datos son censurados.

Un segundo procedimiento en el análisis de datos de tiempos de vida es el estudio de la función de supervivencia y de la razón de fallo. En un análisis sin datos censurados se haría mediante la función de supervivencia empírica, pero en el caso de tener datos censurados se utiliza el estimador Kaplan-Meier, también llamado estimador Producto-Límite (P-L).

4.2. Tablas de vida

En la elaboración de una tabla de vida debemos tener en cuenta la naturaleza del conjunto de datos, si se tratan de datos completos o datos censurados.

Supongamos en primer lugar que se tiene una muestra con datos completos de tamaño n. Para hacer una tabla de frecuencias relativas comenzamos dividiendo el tiempo $k+1$ intervalos $I_j = [a_{j-1}, a_j)$, para $j = 1, \ldots, k+1$ con $a_0 = 0$; $a_k = M$ y $a_{k+1} = \infty$, siendo M la mayor fecha observada.

Sea d_j es el número de tiempos de vida observados que hay en el intervalo I_j. A d_j se le denomina *frecuencia del intervalo j-ésimo* y es el número de individuos fallecidos en el intervalo I_j.

Una tabla de frecuencias es una lista de intervalos con sus respectivos d_j o d_j/n. Con estos datos se puede realizar un histograma relativo al problema donde la base de los rectángulos son $[a_{j-1}, a_j)$ y las alturas son los d_j/n si los intervalos tienen igual amplitud.

En el caso de que el muestreo sea censurado no se puede formar una tabla de frecuencias como tal. Si todos los datos censurados ocurren en el último intervalo $[M, \infty)$ no hay problema, pero si los tiempos censurados caen en otros intervalos no se puede calcular la frecuencia exacta.

4.2.1. Tabla de vida estándar

La tabla de vida es una extensión de la tabla de frecuencias relativas simples presentadas anteriormente. En este caso se estimará la probabilidad de muerte en un intervalo, condicionado a que sobrevive al comienzo del intervalo y la probabilidad de sobrevivir después del fin de intervalo. Es decir se estima la función de riesgo y la supervivencia. Para su calculo definimos los siguientes conceptos.

- N_j: número de individuos en riesgo (vivos y no censurados) en a_{j-1}
- D_j: número de individuos fallecidos (fallos o tiempos de vida) en $[a_{j-1}, a_j)$
- W_j: número de datos censurados en $[a_{j-1}, a_j)$

Claramente se tiene que $N_1 = n$ y que $N_j = N_{j-1} - W_{j-1} - D_{j-1}$ para $j = 1, \ldots, k+1$.

Las medidas que se van a estimar en una tabla de vida son las siguientes para $j = 1, \ldots, k+1$:

- $P_j = Pr\{\text{sobrevivir más allá de } I_j\}$
- $p_j = Pr\{\text{sobrevivir más allá de } I_j | \text{ sobrevive más allá de } I_{j-1}\} = P_j/P_{j-1}$
- $q_j = Pr\{\text{morir en } I_j | \text{ sobrevive más allá de } I_{j-1}\}$

Desde la propia definición se tiene que $q_{k+1} = 1$ y que $P_{k+1} = 0$.

Dado que $p_j = P_j/P_{j-1}$ se tiene que $P_j = p_1 \cdot p_2 \cdots p_j$. De forma recursiva desde esta misma relación, se tiene que $P_j = p_j \cdot P_{j-1}$ siempre para

$j = 1, \ldots, k+1$. Estas medidas son la base de una tabla de vida, pero, 'cómo podemos estimarlas?

Si tuviésemos una muestra aleatoria de tamaño n, sin datos censurados, el estimador máximo verosímil de P_j es

$$\hat{P}_j = \frac{N_{j+1}}{n},$$

y de q_j es

$$\hat{q}_j = \frac{D_j}{N_j}.$$

Por contra, si los intervalos contienen datos censurados entonces N_{j+1} no es necesariamente el número de individuos vivos en a_j pues habría que descontar en número de datos censurados. En este caso el estimador propuesto anteriormente subestima a P_j. Veamos cómo se soluciona este problema en una tabla de vida con datos censurados. Un procedimiento es el siguiente. Si en un intervalo I_j hay W_j datos censurados, entonces se supone que los individuos censurados en dicho intervalo se distribuyen uniformemente a lo largo del intervalo, de modo que, el número de individuos en riesgo se estima considerando el valor esperado de individuos censurados en cualquier punto del intervalo que es $W_j/2$, quedando

$$\hat{q}_j = \frac{D_j}{N_j - \frac{W_j}{2}} = \frac{D_j}{N'_j}.$$

En el caso que $N_j = 0$ entonces $\hat{q}_j = 1$. El denominador N'_j se denomina *riesgo corregido para el intervalo* I_j.

Otros posibles estimadores de q_j son:

- Si los datos censurados están a la derecha del intervalo I_j se puede estimar q_j por $\hat{q}_j = D_j/N_j$
- Si los datos censurados están a la izquierda del intervalo I_j se puede estimar q_j por $\hat{q}_j = \frac{D_j}{N_j - W_j}$

El estimador de p_j y de P_j se puede hallar desde las relaciones de las probabilidades

$$\hat{p}_j = 1 - \hat{q}_j$$
$$\hat{P}_j = \hat{p}_1 \cdot \hat{p}_2 \cdots \hat{p}_j.$$

Por lo tanto una tabla de vida estándar queda como sigue

I	Fallecidos	Censura	Riesgo	N'_j	$\hat{q}_j$	$\hat{p}_j$	$\hat{P}_j$
I_j	D_j	W_j	N_j	$N_j - \frac{W_j}{2}$	$\frac{D_j}{N'_j}$	$1 - \hat{q}_j$	$\hat{p}_1 \cdot \hat{p}_2 \cdots \hat{p}_j$

Esta tabla se puede ampliar con los estimadores de las varianzas de los estimadores definidos anteriormente. Greenwood (1926) propuso el siguiente estimador para la varianza de $\hat{p}_j$, $\hat{q}_j$ y $\hat{P}_j$

$$\widehat{Var}(\hat{q}_j) = \widehat{Var}(\hat{p}_j) = \frac{\hat{q}_j \hat{p}_j}{N'_j}$$

$$\widehat{Var}(\hat{P}_j) = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{N_i' \hat{p}_i}.$$

En el caso de no censura se tiene

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

Una justificación detallada de estos estimadores se pueden ver en Lawless (1982).

4.2.2. Tabla de vida estándar de Chiang y Elveback

Se presenta una modificación al caso anterior. Se considera T la variable tiempo de vida y L la variable de censura. Para el individuo i se tiene el par (T_i, L_i). El método de Chiang requiere el conocimiento de los tiempos de censura de todos los individuos en estudio, incluso de aquéllos que mueren.

Supongamos que se dispone de esta información, descomponiendo N_j y D_j como

$$N_j = N_{j1} + N_{j2} \quad ; \quad D_j = D_{j1} + D_{j2},$$

siendo N_{j2} y N_{j1} el número de individuos en riesgo en I_j siendo o no censurados en dicho intervalo respectivamente. Por otro lado D_{j1} y D_{j2} son el número de muertos en I_j de entre los N_{j1} y N_{j2} respectivamente.

Al conocer los tiempos de censura para todos los individuos, podemos determinar qué individuos están en el grupo *se espera que sean censurados* y aquéllos que están en el grupo contrario.

Cuando todos lo individuos tienen la misma distribución de tiempo de vida, se sigue que dado N_{j1}, D_{j1} es binomial con parámetros N_{j1} y q_j.

La probabilidad q_j se puede estimar mediante D_{j1}/N_{j1}, el cual es un estimador insesgado de la probabilidad. A menos que la censura sea muy débil este estimador pierde gran información por lo que no es usado. A este estimador se le denomina *estimador de q_j con muestra reducida.*

Para utilizar la información de la censura se debe considerar la distribución de D_{j2}.

Chiang introduce el supuesto adicional siguiente, *el tiempo de censura para individuos que se espera que sean censurados en I_j, dado que sobreviven al comienzo de dicho intervalo, está uniformemente distribuido sobre I_j y los tiempos de vida están distribuidos exponencialmente.*

Bajo estas hipótesis se puede ver que, dado N_{j2}, entonces D_{j2} tiene una distribución binomial con parámetros N_{j2} y q_j^*, siendo

$$q_j^* = 1 + \frac{q_j}{\lg\left(1 - q_j\right)}.$$

Combinando las verosimilitudes binomiales de D_{j1} y D_{j2} se obtiene la verosimilitud para q_j, siendo

$$L(q_j) = \left(\begin{array}{c} N_{j1} \\ D_{j1} \end{array} \right) q_j^{D_{j1}} \left(1 - q_j\right)^{N_{j1} - D_{j1}} \left(\begin{array}{c} N_{j2} \\ D_{j2} \end{array} \right) q_j^{*D_{j2}} \left(1 - q_j^*\right)^{N_{j2} - D_{j2}}.$$

Chiang aproxima q_j^* si $q_j \leq 0{,}3$ mediante

$$q_j^* = 1 + \frac{q_j}{\lg(1-q_j)} \doteq 1 - (1-q_j)^{1/2}\,.$$

Sustituyendo esta aproximación se puede obtener el estimador máximo verosímil.

Por otra parte, Elveback (1958) da una estimación de q_j de una manera similar a la anterior, suponiendo que, *dado un individuo que muere en I_j, el tiempo de muerte está uniformemente distribuido, como los tiempo de censura, en el intervalo*. Esto conduce a una verosimilitud como sigue

$$L(q_j) = \left(\begin{array}{c} N_{j1} \\ D_{j1} \end{array} \right) q_j^{D_{j1}} (1-q_j)^{N_{j1}-D_{j1}} \left(\begin{array}{c} N_{j2} \\ D_{j2} \end{array} \right) \left(\frac{q_j}{2}\right)^{D_{j2}} \left(1-\frac{q_j}{2}\right)^{N_{j2}-D_{j2}},$$

ya que $q_j^* = q_j/2$, quedando

$$L(q_j) = \left(\begin{array}{c} N_{j1} \\ D_{j1} \end{array} \right) q_j^{D_{j1}} (1-q_j)^{N_{j1}-D_{j1}} \left(\begin{array}{c} N_{j2} \\ D_{j2} \end{array} \right) \left(\frac{1}{2}\right)^{N_{j2}} q_j^{D_{j2}} (2-q_j)^{N_{j2}-D_{j2}}\,.$$

4.3. Estimador de la función de supervivencia

En esta sección se presenta el estimador de la función de supervivencia en el caso de muestreo completo y censurado.

4.3.1. Estimador de la función de supervivencia: Muestreo Completo

Si se tiene una muestra de tamaño n de tiempos de vida no censurados entonces la estimación no parmétrica de la función de supervivencia es conocida como

$$\hat{S}(t) = \frac{n^o observaciones \geq t}{n} \quad ; \quad t \geq 0,$$

ya que $S(t) = P[T \geq t]$. Se trata de una función escalonada que decrece $1/n$ después de cada tiempo de vida observado, si son todas las observaciones distintas. Si hay d tiempos iguales de tiempos de vida, el escalón en dicho tiempo es de longitud d/n justo después de t. Según esta definición, se considera continua a la derecha la función de supervivencia. Este estimador es un estimador consistente de $S(t)$ ya que converge uniformemente a la función de supervivencia según el teorema de Glivenko-Cantelli.

4.3.2. Estimador de Kaplan-Meier (PL)

Cuando hay datos censurados y alguno lo es antes de un determinado tiempo, entonces no es conocido el número de sobrevivientes después de dicho tiempo.

Para solventar el problema se modifica el estimador de la función de supervivencia descrito en la sección anterior. Dichas modificaciones dan lugar al estimador *producto-límite (PL) o estimador de Kaplan-Meier* gracias a Kaplan y Meier, 1958. El estimador se define como sigue.

Sean n individuos en estudio de los cuales hay r distintos tiempos de vida (de fallo o de muerte). Sean estos tiempos $t_1 < t_2 < t_3 \dots t_r$ y sea d_j el número de muertos en el tiempo t_j. El estimador Kaplan-Meier se define como

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j},$$

siendo n_j el número de individuos en riesgo en t_j, vivos y no censurados en t_{j-} (al principio de t_j).

Si el tiempo de censura de un individuo coincide con un tiempo de fallo, entonces se considera que en este tiempo dicho individuo est en riesgo. Es decir, dicho individuo está en n_i.

La motivación del estimador es la misma que en la tabla de vida para la probabilidad de supervivencia $\hat{P}_j = \hat{p}_1\hat{p}_2 \dots \hat{p}_j$. Es decir, $S(t)$ se puede expresar como un producto donde cada término es la función de riesgo: probabilidad de sobrevivir al tiempo t_j dado que vive al comienzo de t_j. El estimador de la función de riesgo es

$$\hat{h}_j = \hat{P}[T > t_j \| T > t_{j-1}] = \frac{n_j - d_j}{n_j}.$$

Este estimador resulta de considerar en la tabla de vida estándar una partición de la longitud del tiempo de vida con infinitos intervalos y amplitud de los mismos tendiendo a cero.

Cuando no hay censura $n_1 = n$ y $n_j = n_{j-1} - d_{j-1}$ para $j = 2, \dots k$, y el estimador es la función de supervivencia empírica descrita en la sección anterior.

En el caso censurado y no censurado la función $\hat{S}(t)$ es escalonada no creciente, con $\hat{S}(0) = 1$ y cada salto es de longitud $\frac{n_j - d_j}{n_j}$ después de cada tiempo t_j. En los tiempos de censura el estimador no cambia, esta tiene efecto en el tamaño de los saltos. Por otro lado decir que el estimador permanece constante a partir del último tiempo de vida.

El estimador de Kaplan-Meier tiene buenas propiedades. Entre ellas numeramos las siguientes

- Es un estimador consistente bajo condiciones generales.
- Es un estimador con varianza asintótica válida.
- Es un estimador máximo verosímil.

Finalmente se puede estimar la varianza de $\hat{S}(t)$como sigue

$$\widehat{Var}\hat{S}(t) = \hat{S}(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}.$$

4.3.3. Estimador de la función de riesgo acumulada: Nelson-Aalen

Es conocido que la función de supervivencia se puede expresar en función de la función de riesgo. Por lo que $H(t) = -\lg S(t)$. Si se deseas estimar la función de riesgo acumulada entonces claramente se tiene que $\hat{H}(t) = -\lg \hat{S}(t)$, siendo $\hat{S}(t)$el estimador Kaplan-Meier de lafunción de supervivencia.

Una segunda alternativa para estimar la razón de fallo acumulada sería la siguiente

$$\tilde{H}(t) = \sum_{j:t_j<t} \frac{d_j}{n_j},$$

que es la llamada función acumulada empírica de la función de riesgo. En el caso continuo se puede considerar un estimador de la función de riesgo en t_j a d_j/n_j.

En el caso de datos censurados tipo II, se puede ver que

$$E[H(t_j)] = \tilde{H}(t_{j+}).$$

En modelos continuos ambos estimadores son asintóticamente equivalentes pero difieren para grandes valores de t. Realmente se está tomado una aproximación de primer orden del estimador de Kaplan-Meier, pues

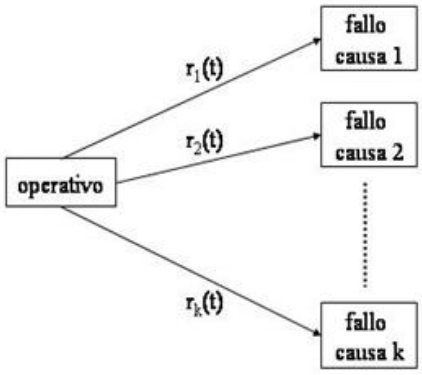
$$\begin{aligned} \hat{H}(t) &= -\lg \hat{S}(t) = -\lg \prod_{j:t_j<t} \frac{n_j - d_j}{n_j} = -\sum_{j:t_j<t} \lg\left(1 - \frac{d_j}{n_j}\right) \\ &= \sum_{j:t_j<t} \left(\frac{d_j}{n_j} + \frac{d_j^2}{2n_j^2} + \cdots\right) \approx \sum_{j:t_j<t} \frac{d_j}{n_j} = \tilde{H}(t) \end{aligned}$$

El estimador de la varianza de este estos estimadores es igual a

$$\widehat{Var}\left(\hat{H}(t)\right) = \widehat{Var}\left(\tilde{H}(t)\right) = \frac{\widehat{Var}\left(\hat{S}(t)\right)}{\hat{S}(t)^2}.$$

4.4. Análisis de datos de fenómenos con respuesta múltiple: El método de la incidencia acumulada

Hasta ahora, en el análisis de supervivencia, siempre hemos hablado de la observación del tiempo que transcurre hasta la aparición de un suceso (fallo, muerte, recaída en una enfermedad, etc.), observación que puede ser completa o censurada. Sin embargo, en multitud de casos reales, la situación puede ser más complicada, en el sentido de que el suceso a observar puede no ser dicotómico, es decir fallo SI/NO. El modelo de riesgos competitivos plantea que existe un estado inicial (operativo o vivo) y k posibles estados finales, gráficamente



donde $r_i(t)$ es la función de riesgo relativa a la causa de fallo i-ésima.

Presentamos a continuación un método para estimar la probabilidad de supervivencia en una situación de riesgos competitivos, como alternativa al método de Kaplan Meier, ya que este procedimiento no es adecuado cuando se presentan sucesos alternativos. Por ejemplo, supongamos que en un determinado modelo de coche se instala una nueva pieza en el sistema de arranque, que se quiere probar. Se trataría en este caso de estudiar el tiempo transcurrido hasta que se detecta un fallo en dicha pieza. Si queremos calcular esta probabilidad de fallo en un coche que ha sufrido un accidente de tráfico y ha sido calificado como siniestro total, antes de haber observado el fallo de la pieza; no es lo más adecuado que este ítem (coche) se considere como una observación censurada en el momento en que se produce el accidente, puesto que esto implicaría que suponemos que no sabemos si va a presentarse el fallo de la pieza a partir de ese instante, cosa que no es cierta, dado que el coche no volverá a rodar nunca más después del accidente, y por tanto es imposible que se presente el fallo de la pieza en el futuro.

Un estimador de probabilidad de fallo, alternativo al método de Kaplan Meier, es el que se conoce como *incidencia acumulada o probabilidad de riesgo o fallo específico.*

Aplicamos el procedimiento al caso en que tengamos únicamente dos posibles sucesos alternativos. Se observan n ítems a los que se asocia un tiempo t_i en el que puede haber ocurrido una de las siguientes situaciones:

- Ocurre el suceso principal en el instante t_i (el fallo de la pieza);
- Ocurre el suceso alternativo o competitivo en t_i (hay un accidente fatal);
- Finalizado el periodo de observación en t_i, no ha ocurrido ninguno de los dos sucesos anteriores (el coche continúan rodando con la pieza en funcionamiento).

El último caso corresponde a observaciones censuradas o incompletas. Definimos las siguientes cantidades:

- n: número total de observaciones;
- e_i: número de ítems que experimentan el suceso principal en el instante t_i;
- r_i: número de ítems que experimentan el suceso alternativo en el instante t_i;
- c_i: número de observaciones censuradas en t_i;
- n_i: número de ítems que continúan expuestos al riesgo después de t_i. Esta cantidad será igual al número de ítems que entraron en estudio menos los que se han ido perdiendo por diferentes causas, es decir

$$n_i = n - \sum_{j=1}^{i} (e_j + r_j + c_j) .$$

Si no se tiene en cuenta la presencia del riesgo alternativo y los ítems que experimentan el suceso alternativo se consideran observaciones incompletas para ese momento, estaremos efectuando la estimación mediante el método de Kaplan Meier,

$$KM_1(t) = \prod_{i=1}^{s} \left(1 - \frac{e_i}{n_{i-1}}\right),$$

para s tal que $t_1 < t_2 < \ < t_s \leq t$.

Del mismo modo, para estimar la probabilidad de supervivencia en el caso del suceso alternativo, podemos considerar como observaciones incompletas aquellas que corresponden a los ítems que en ti presentan el suceso principal, de modo que

$$KM_2(t) = \prod_{i=1}^{s} \left(1 - \frac{r_i}{n_{i-1}}\right),$$

para s tal que $t_1 < t_2 < \ < t_s \leq t$.

Por lo tanto, la probabilidad de no experimentar el suceso principal y no experimentar el suceso alternativo, al menos hasta el tiempo t, que llamaremos $KM_{12}(t)$, se calcula como el producto de las dos probabilidades anteriores, como la probabilidad de ocurrencia de dos sucesos independientes

$$KM_{12}(t) = KM_1(t)KM_2(t).$$

Estimamos entonces la probabilidad de que ocurra el suceso de interés en el instante t_i sabiendo que está condicionada a que para ello hay que haber llegado hasta ese instante sin que haya ocurrido ninguno de los sucesos (cuya probabilidad viene dada por la última expresión), y así llegamos a la fórmula para la estimación de la incidencia acumulada

$$IC(t) = \sum_{i=1}^{s} \frac{e_i}{n_{i-1}} KM_{12}(t),$$

para s tal que $t_1 < t_2 < \ < t_s \leq t$; donde el primer factor de cada sumando corresponde a la probabilidad de que ocurra un suceso principal en el instante t_i, y el segundo, es la probabilidad de haber llegado hasta ese instante. El sumatorio nos indica que la probabilidad de haber experimentado un suceso principal se calcula como la probabilidad de que el suceso haya ocurrido en t_1 o en t_2 o en $t_i \leq t_s$.

Notemos que la probabilidad de fallo estimada directamente mediante la función de KM_1 subestima la supervivencia y por tanto sobre estima el riesgo, que vendría dado por $1-KM_1$, ya que estaríamos considerando que los ítems en los que ha ocurrido el fallo alternativo son susceptibles de sufrir un fallo principal en el futuro, lo cual no es cierto, por tanto siempre ocurre que $IC \leq 1 - KM_1$.

Si queremos calcular la probabilidad de supervivencia, en lugar de la probabilidad de fallo, tendríamos

$$\hat{R}_{IC}(t) = 1 - IC(t).$$