# UNIVERSIDAD NACIONAL

# AUTONOMA DE MEXICO

***FACULTAD DE CIENCIAS***

ANALISIS ESTADISTICO DE TIEMPOS
DE FALLA UNIVARIADOS

**T E S I S**

QUE PARA OBTENER EL TITULO DE

***A C T U A R I O***

P R E S E N T A

***JOSE SALVADOR ZAMORA MUÑOZ***



MEXICO, D.F. ENERO 1994


**UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso**

# AGRADECIMIENTOS

Haber terminado este trabajo me da la oportunidad de agradecer a todas las personas que colaboraron en en su realización, entre las que se encuentran de manera especial la Dra. Belem Trejo Valdivia por su paciente dirección, y sobre todo por compartir sus concimientos en el campo estadístico del Análsis de Supervivencia; agradezco asimismo a todas aquellas personas que hacen posible el proyecto UNAM, ya que con su colaboración contribuyen a que al final del milenio existan instituciones de esta naturaleza; con el mismo énfasis expreso mi más amplio agradecimiento al pueblo trabajador mexicano, que hacen que la escuela pública sea una realidad y una seria alternativa.

Quiero agradecer además el apoyo recibido por el Instituto de Investigación en Matemáticas Aplicadas y en Sistemas (IIMAS) por la utilización de su infraestructura física e intelectual durante el desarrollo de este trabajo.

# INDICE

página

# INTRODUCCION

Cuando trabajamos con la estadística tradicional, suponemos que un individuo en la muestra nos proporcionará información completa sobre alguna característica particular en que estemos interesados. Sin embargo, puede suceder que solamente nos proporcione información parcial al respecto de dicha característica, los métodos usuales de la estadística eliminarían a este individuo sin tomar en cuenta que la información que él proporciona puede ser de utilidad.

El objetivo del Análisis de Supervivencia es incorporar a los procesos estadísticos la información parcial que se recabe de un individuo. De uso común en estudios longitudinales, la información parcial en estos estudios la proporcionan aquellos individuos que no concluyen el proceso de observación, se dice que dichos individuos representan observaciones censuradas.

Como es lógico suponer, en el Análisis de Supervivencia se reconstruyen todos los procesos de la estadística clásica (estimación paramétrica, intervalos de confianza, pruebas de hipótesis, etc.) incorporando en sus desarrollos la presencia de observaciones que resulten censuradas.

El objetivo de este trabajo es presentar la metodología del Análisis de Supervivencia, además de dar a conocer esta herramienta para que el uso de ella se extienda a otros campos en los que sea pertinente. Con el fin de lograr lo anterior, primeramente se introducen los conceptos propios de esta rama estadística, así como las funciones básicas con las que se realiza el Análisis de Supervivencia.

En seguida se presentan las principales familias de modelos paramétricos para dichas funciones, para continuar con los métodos no paramétricos alternativos a los modelos anteriores. Proseguiremos desarrollando los procesos de inferencia para las familias paramétricas y finalmente se presentarán dos modelos que nos ayudarán a realizar el análisis en el caso de que las poblaciones no sean homogéneas (debido a la presencia de covariables).

# CAPITULO 1

# ANALISIS DE SUPERVIVENCIA

## 1.1 INTRODUCCION

El campo de estudio del Análisis de Supervivencia lo constituyen algunos procesos en los que existe interés por estudiar el tiempo entre la ocurrencia de dos sucesos determinados, el segundo de ellos conocido como falla. Sus principales fuentes de aplicación son en las ciencias biomédicas y en confiabilidad.

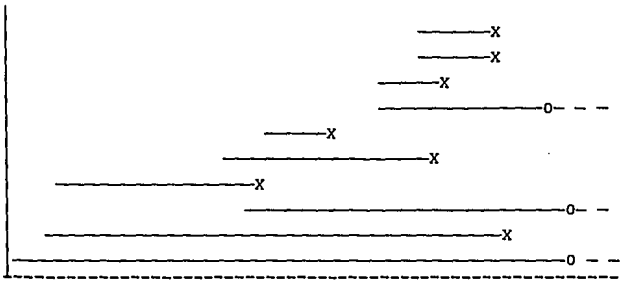
Para determinar este tiempo de falla necesitamos tres condiciones principales:

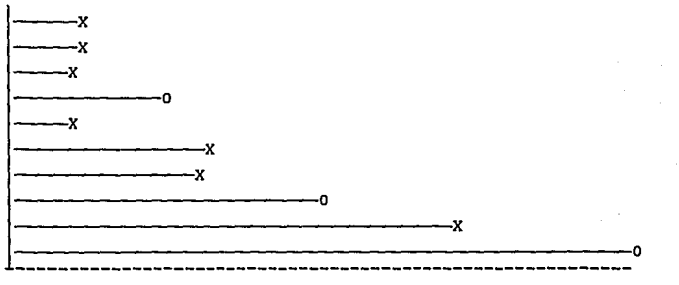
-Tiempo de inicio.- Determinar sin ambigüedad el tiempo en que se inicia el seguimiento de cada individuo en el estudio.
-Convenir una escala de medida para el tiempo dentro del estudio.
-Definir con precisión lo que se entiende por falla en el estudio.

**Ejemplos:**

**1.1.1 Estudio de una enfermedad mortal.** En un hospital se realiza el estudio de diez pacientes que tienen alguna enfermedad mortal. El ingreso de los pacientes se produce en cualquier momento a lo largo del estudio, por lo que el tiempo calendario para cada uno de ellos no es necesariamente el mismo. El tiempo de falla (muerte), será medido desde la fecha de ingreso de cada individuo en el estudio.

Las siguientes gráficas muestra este proceso




X muerte  0 salida del estudio

**1.1.2 Control de calidad:** En una fábrica de artículos electrónicos para evaluarla calidad, se ponen a trabajar en condiciones estándar, un cierto número de aparatos producidos y se registra el tiempo de la primera falla en cada uno. En este caso el tiempo de inicio es controlado por el investigador, por lo que el tiempo calendario y el tiempo de estudio coinciden. El evento de interés es la presencia de cualquier falla en cualquier aparato, la variable respuesta es el tiempo que transcurre desde el momento en que se pone a funcionar hasta que falla por primera vez.

**1.1.3 Reparación de artículos:** Después de la reparación de un artículo, el interés se centra en la ocurrencia de fallas consecutivas. Como en el caso anterior el tiempo inicial lo controla el investigador, así el tiempo calendario es igual al tiempo dentro del estudio. La variable respuesta es el tiempo transcurrido entre dos fallas consecutivas.

**1.1.4 Comparación entre dos grupos.** En un estudio de individuos que presentan alguna enfermedad (leucemia por ejemplo), una muestra aleatoria de ellos serán tratados con cierto medicamento y el resto se tomará como grupo control. En este estudio los pacientes ingresan en tiempos distintos por lo que tienen un tiempo calendario diferente. El tiempo dentro del estudio se medirá a partir de su ingreso. El interés es determinar si el tiempo de vida de los individuos tratados es mayor que el del grupo control. La variable respuesta es el tiempo transcurrido desde el inicio hasta su muerte (falla).

En todos estos ejemplos pueden presentarse salidas del estudio.

## 1.2 CENSURA

El Análisis de Supervivencia es particularmente útil cuando aparecen observaciones censuradas, que son aquellas en las que no es posible observar a los individuos hasta que se presente la falla. La censura es una de las razones más importantes para desarrollar modelos y procedimientos para analizar el tiempo de falla.

Algunas causas de censura pueden ser:

-El estudio termina y se tienen individuos en los que aún no ocurre el evento de interés.
-La falla se presenta en algunos individuos por causas que no son de interés para el estudio.
-Algunos individuos abandonan el estudio antes de ocurrir la falla.

Los individuos clasificados como salidas del estudio representan censuras en cada uno de los ejemplos anteriores.

### 1.2.1 TIPOS DE CENSURA

Lo que distingue al Análisis de Supervivencia de otros campos de la estadística es la posibilidad de censura. De manera vaga, una observación censurada contiene sólo información parcial acerca de la variable de interés.

En general se considera tres tipos de censura:

### 1.2.2 Censura Tipo I

Algunas veces no se tiene ni el tiempo ni el dinero suficiente para prolongar el estudio hasta que todos los individuos dentro de él presenten la falla, por lo que el investigador se ve forzado a determinar un tiempo límite para realizar sus observaciones, a este tiempo lo llamaremos *tiempo de censura fijo.* Su asignación puede ser con base a la experiencia del investigador, al presupuesto de que disponga, al tiempo para el estudio, etc..

Este tipo de censura lo podemos representar matemáticamente de la siguiente manera:

Sea $t_c$ algún número fijo ( previamente asignado ) que llamaremos tiempo de censura fijo y que corresponde al tiempo final de observación. En lugar de observar $T_1,\ldots,T_n$ ( las variables aleatorias de interés), se observa solamente $Y_1,\ldots, Y_n$ en donde

$$Y_i = \begin{cases} T_i \text{ si } T_i \leq t_c \\ t_c \text{ si } t_c < T_i \end{cases} \qquad i = 1,2,\ldots,n$$

Esto es, observamos la variable aleatoria $T_i$ si ésta no rebasa al tiempo fijo $t_c$ y observamos $t_c$ en lugar de $T_i$ si $T_i$ es mayor que $t_c$ .

### 1.2.3 Censura tipo II

En la realización de un estudio, el investigador decide observar sólo las r primeras fallas de n posibles ($r < n$) debido a diversas razones como el costo del estudio, estudios previos, entre otros, y decide tomar el valor de la última falla observada para el resto que no observó.

Técnicamente podemos representar lo anterior como sigue:

Sea $r<n$ fija y sea $T_{(1)}<T_{(2)}\ldots<T_{(n)}$ las estadísticas de orden[1] de $T_1, T_2, \ldots, T_n$. Realizamos las observaciones y suspendemos después de que ocurre la r-ésima falla, hemos observado $T_{(1)}, \ldots, T_{(r)}$ . La muestra está formada por:

$$
\begin{aligned}
Y_{(1)} &= T_{(1)} \\
&\cdot \\
&\cdot \\
&\cdot \\
Y_{(r)} &= T_{(r)} \\
Y_{(r+1)} &= T_{(r)} \\
&\cdot \\
&\cdot \\
&\cdot \\
Y_{(n)} &= T_{(r)}
\end{aligned}
$$

Es decir, observamos hasta la r-ésima falla y después asignamos ese valor ($T_{(r)}$) al resto de las variables aleatorias $Y_i$ para $i > r$ .

Estos dos tipos de censura aparecen algunas veces en aplicaciones de ingeniería.
Por ejemplo: ( Control de Calidad ). Tenemos un conjunto de transistores; los colocamos a todos a prueba en t=0, y anotaremos

---

1.- La estadística de orden se define como:

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n de una población con función de distribución $F(\cdot)$ . Entonces $Y_1 < Y_2 < \ldots < Y_n$ la muestra de las $X_i$'s dispuestas en orden de magnitud creciente son llamadas las estadísticas de orden de la muestra aleatoria $X_1, X_2, \ldots, X_n$

su tiempo de falla. Algunos transistores pueden tardar mucho tiempo en quemarse y no podemos esperar tanto para finalizar el experimento. Por lo tanto, podemos detener el experimento en un tiempo fijado de antemano $t_c$ , en este caso tenemos censura del tipo I, o podemos desconocer el valor del tiempo de censura fijo que debemos esperar para que una fracción pre-especificada de transistores se hayan quemado, en este caso tenemos censura del tipo II.

### 1.2.4 Censura Aleatoria

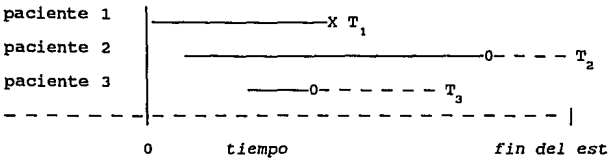
La censura aleatoria aparece en aplicaciones médicas, estudios de animales o tratamientos clínicos, y corresponden a aquellos casos donde el paciente se retira del estudio, muere por alguna causa ajena al evento de interés, etc. En este caso el tiempo de censura no es controlado por el investigador. En un tratamiento clínico, los pacientes pueden ingresar al estudio en diferentes tiempos, entonces cada uno se trata con una o varias terapias posibles. Es de interés observar su tiempo de vida, la censura puede ocurrir en alguna de las siguientes formas:

*Pérdida de secuencia*. El paciente decide mudarse a otra parte por lo que no lo volveremos a ver.

*Retiro*. La terapia puede tener efectos laterales muy malos, por lo que es necesario suspender el tratamiento, o bien el paciente rehusa continuar el tratamiento.

*Terminación del estudio*. El estudio concluye y tenemos pacientes que aún no presentan la falla de interés.

La siguiente gráfica ilustra un posible proceso.




Aquí, el paciente 1 entra al estudio en t=0 y muere, por lo tanto $T_1$ es una observación no censurada, el paciente 2 ingresa al estudio, y al final del mismo permanece aún vivo resultando una observación censurada $T_2$ y el paciente 3 ingresa al estudio y se pierde su secuencia antes de que termine el estudio, por lo que tenemos otra observación censurada $T_3$.

La manera de presentar matemáticamente esta censura es como sigue:

Sean $C_1, C_2, \ldots, C_n$ variables aleatorias idénticamente distribuidas con función de distribución G, $C_i$ es el tiempo de censura asociado al i-ésimo individuo. Observamos $(Y_1,\delta_1),\ldots,(Y_n,\delta_n)$ donde:

$$Y_i = \min (T_i, C_i) \quad y$$

$$\delta_i = I\,(T_i \leq C_i) = \begin{cases} 1 \text{ si } T_i \leq C_i \text{ , esto es, } T_i \text{ no es censurada.} \\ 0 \text{ si } T_i > C_i \text{, esto es, } T_i \text{ es censurada.} \end{cases}$$

Es común suponer que los tiempos de falla y de censura son independientes.

Este supuesto parece justificado ya que se tienen ingresos aleatorios y ocurren pérdidas de secuencia aleatoriamente dentro del estudio. Sin embargo, si la razón para desertar está

relacionada con el curso de la terapia, este hecho puede provocar dependencia entre $T_i$ y $C_i$, y los procedimientos aquí presentados no se aplicarían en tal situación.

### 1.2.5 Otros tipos de clasificacion de la censura

*Censura aleatoria por la derecha.* Se dice que una observación es censurada por la derecha cuando el valor exacto del tiempo de falla T se desconoce y sólo se sabe que es mayor que el tiempo en que se registró la censura. De manera similar, una observación se dice censurada por la izquierda si se sabe que el valor desconocido de T es menor que su tiempo de censura. Para esta última observaremos solamente

$$(Y_1, E_1), \ldots, (Y_n, E_n) \text{ donde}$$

$$Y_i = \text{máx}(T_i, C_i)$$

$$E_i = I(C_i \leq T_i)$$

El siguiente ejemplo ilustra estos dos tipos de censura:

**Niños africanos.** Un psiquiatra deseaba estudiar el tiempo en el que un grupo de niños africanos aprenden a ejecutar una tarea específica. Cuando llegó a la aldea, ésta tenía algunos niños que ya la sabían ejecutar, por lo que estos niños constituyen las observaciones con censura aleatoria por la izquierda. Otros niños aprendieron la tarea durante su estancia. El resto son niños que aún no aprenden a realizar la tarea, representando así las observaciones con censura aleatoria por la derecha.

Ambos tipos de censura aleatoria, izquierda y derecha, son casos especiales de censura por intervalos , en la que solamente podemos

realizar la observación de la variable aleatoria de interés en un intervalo. Si $T_i$ es una variable aleatoria con censura aleatoria por la derecha, realizaremos la observación de $T_i$ comprendida en el intervalo $[C_i,\infty)$, y si $T_i$ es una variable aleatoria con censura por la izquierda, realizaremos la observación en el intervalo $[0,C_i]$.

La posibilidad de censura hace necesario el desarrollo de métodos y procedimientos estadísticos para el análisis de la información procedente de estudios longitudinales en donde el mayor interés se centra en entender el comportamiento de un tiempo de falla.

# CAPITULO 2

## DISTRIBUCION DEL TIEMPO DEL TIEMPO DE FALLA

Ahora tenemos la necesidad de una herramienta matemática que nos sea útil para el Análisis de Supervivencia, esta herramienta la constituyen tres funciones que nos servirán para caracterizar el comportamiento de la variable aleatoria T que representa el tiempo de falla en una población homogénea.

Las funciones son:

-Función de Supervivencia
-Función de Densidad de Probabilidad de T
-Función de Riesgo.

Cada una de estas funciones tienen su utilidad dentro del Análisis de Supervivencia, pero guardan relación entre ellas.

En general una variable aleatoria de tiempo de falla se mide de manera discreta. Sin embargo, para fines operativos se analizará la variable de manera discreta si el periodo entre mediciones consecutivas modifica las condiciones del problema de estudio, de lo contrario se analizará dicha variable de forma continua.

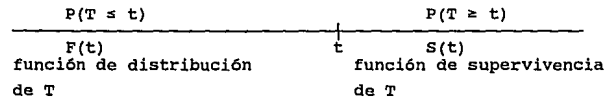
### 2.1 CASO CONTINUO

Analizaremos primero el caso en que T es una variable aleatoria no negativa absolutamente continua.

Definimos la función de supervivencia como la probabilidad de que T sobrepase un tiempo t . Esto es :

$$S(t) = P(T \geq t) \qquad \text{con} \quad t \in [0,\infty)$$

que es el complemento con respecto a 1 de la función de distribución de T , $F(t) = P(T \leq t)$.
Gráficamente tenemos




La función de densidad de probabilidad de T se define como

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

Por otro lado

$$\begin{aligned} P(\, t < T \leq t+\Delta t \,) &= P(T \leq t+\Delta t) - P(T \leq t) \\ &= 1-P(T \geq t+\Delta t) - [1-P(T \geq t)] \\ &= P(T \geq t) - P(T \geq t+\Delta t) \\ &= S(t) - S(t+\Delta t) \end{aligned}$$

es decir

$$f(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{\Delta t}$$

por lo tanto

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

de donde

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

De esta expresión se desprende

$$S(0) = \int_0^{\infty} f(x)dx = 1 \quad \text{ya que } f(t) \text{ es función de densidad}$$

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

$S(t)$ es monótona no creciente.

La última de nuestras funciones es la función de riesgo que es la tasa instantánea de falla al tiempo t y se define como:

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

como

$$P(t < T \leq t+\Delta t | T \geq t) = \frac{P(t < T \leq t+\Delta t)}{P(T \geq t)}$$

entonces

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

o de manera equivalente

$$h(t) = \left( \frac{-d}{dt} S(t) \right) / S(t) = \frac{-d}{dt} \left( \log S(t) \right)$$

Por lo que

$$S(t) = \exp\left\{-\int_0^t h(x)\,dx\right\} .$$

Como se había dicho en un principio se han establecido relaciones entre las tres funciones (S(t), f(t), h(t)) por lo que bastará conocer una de ellas para obtener las restantes. Estas relaciones son:

$$S(t) = \int_t^\infty f(x)\,dx$$

$$S(t) = \exp\left\{-\int_0^t h(x)\,dx\right\}$$

y como

$$f(t) = h(t)S(t)$$

entonces

$$f(t) = h(t)\exp\left\{-\int_0^t h(x)\,dx\right\} .$$

## 2.2 CASO DISCRETO

De manera similar al caso continuo trabajaremos el caso discreto. Supongamos T una variable aleatoria discreta que toma los valores $0 \leq t_1 < t_2 < \ldots$ .La función de densidad de probabilidad para T es:

$$f(t) = \begin{cases} P\,(T = t_j) & \text{si } t=t_j \\ 0 \text{ en otro caso} & \end{cases} \qquad j = 1,2.....$$

La función de supervivencia es:

$$S(t) = P(T \geq t) = \sum_{j:t_j \geq t} f(t_j)$$

donde, como en el caso continuo, tenemos:

$$S(0) = P(T \geq 0) = \sum_{j:t_j \geq 0} f(t_j) = 1$$

$$S(\infty) = 0$$

S(t) es una función monótona no creciente.

La función de riesgo se define para los valores $t_j$ j=1,2... y proporciona la probabilidad condicional de falla al tiempo $t=t_j$, dado que se estuvo vivo justo antes de $t_j$, por lo que:

$$h(t_j) = P(T = t_j \mid T \geq t_j)$$

$$= \frac{P(T = t_j)}{P(T \geq t_j)}$$

$$= \frac{f(t_j)}{S(t_j)} \qquad j = 1,2....$$

Observemos que

$$f(t_j) = S(t_j) - S(t_{j+1})$$

por lo tanto

$$h(t_j) = \frac{S(t_j) - S(t_{j+1})}{S(t_j)} = 1 - \frac{S(t_{j+1})}{S(t_j)}$$

despejando tenemos

$$S(t_{j+1}) = [1 - h(t_j)]S(t_j)$$

$$\text{Con } S(t_j) = [1 - h(t_{j-1})]S(t_{j-1})$$

$$\vdots$$

$$S(t_2) = [1 - h(t_1)]S(t_1) \quad \text{y } S(t_1) = 1 .$$

Por lo que

$$S(t_j) = [1-h(t_{j-1})][1-h(t_{j-2})]\ldots[1-h(t_1)]$$

es decir, la función de supervivencia queda expresada por

$$S(t) = \prod_{j:t_j<t} [1-h(t_j)]$$

Entonces la función de densidad de probabilidad en función de h(t) está dada por:

$$f(t_j) = h(t_j)S(t_j)$$

$$= h(t_j) \prod_{j:t_j<t} [1-h(t_j)]$$

## 2.3 CASO MIXTO

En este caso, la función de distribución de T posee una parte continua y una discreta por lo que la función de supervivencia consta de productos de términos para cada parte, es decir

$$S(t) = \exp\left[-\int_0^t h_c(u)du\right] \prod_{j:t_j<t} [1-h(t_j)]$$

con $h_c$ la función de riesgo para la parte continua y $t_1<t_2<...$ puntos con probabilidad positiva para la parte discreta.

En este caso la función de riesgo puede expresarse como

$$h(t)dt = h_c(t)dt + \sum_j h(t_j)\delta(t-t_j)$$

en donde $\delta(t-t_j)$ es la delta de Dirac, definida como

$$\delta(x)dx = \begin{cases} 1 & x = 0 \\ 0 & \text{en otro caso} \end{cases}$$

Otra función que también puede usarse en el análisis de la distribución de T, es la función de riesgo acumulada, definida de la siguiente manera

$$\Lambda(t) = \int_0^t h(u)du = \int_0^t h_c(u)du + \sum_{j|t_j<t} h(t_j)$$

Este caso no será abordado en el resto del presente trabajo.

## 2.4 CARACTERISTICAS DE LA FUNCION DE RIESGO

La función de riesgo describe la forma como cambia la tasa instantánea de muerte de un individuo al paso del tiempo (constante, lineal, exponencial, etc. ).

El conocimiento de la función de riesgo puede ayudar en la selección del modelo para la distribución del tiempo de vida. Por

ejemplo, puede ser útil al considerar restricciones para modelos con funciones de riesgo no decrecientes o modelos cuyas funciones de riesgo tienen alguna otra característica bien definida.

Las comparaciones entre grupos o individuos suelen hacerse con mayor precisión vía la función de riesgo.

Los modelos basados en el riesgo son a menudo convenientes cuando se tiene censura o varios tipos de falla.

Ahora ejemplificaremos las tres funciones anteriores en distribuciones de tiempo de falla particulares.

1). Familia Exponencial

$$h(t) = \lambda$$
$$f(t) = \lambda \exp(-\lambda t)$$
$$S(t) = \exp(-\lambda t)$$

2). Familia Weibull

$$h(t) = \lambda p(\lambda t)^{p-1}$$
$$f(t) = p\lambda(\lambda t)^{p-1}\exp\{-(\lambda t)^{p}\}$$
$$S(t) = \exp\{-(\lambda t)^{p}\}$$

3). Familia Gamma

$$f(t) = \frac{\lambda(\lambda t)^{k-1}\exp(-\lambda t)}{\Gamma(k)}$$

$$S(t) = 1-\int_0^t f(t) = 1-I_k G(\lambda t)$$

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

donde $IG_k(\cdot)$ es la gamma función incompleta dada por

$$\int_0^{\cdot} \frac{x^{k-1}\exp(-x)}{\Gamma(k)}\, dx$$

4). Familia Log-Normal

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

$$S(t) = 1-\Phi\left[\frac{\log t-\mu}{\sigma^2}\right]$$

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

5). Familia Log-logística

$$h(t) = \frac{kt^{k-1}p^k}{[1+(tp)^k]}$$

$$f(t) = \frac{kp^kt^{k-1}}{[1+(tp)^k]^2}$$

$$S(t) = \frac{kt^{k-1}p^k}{[1+(tp)^k]}$$

6). Familia de riesgos proporcionales con covariable $\underline{Z}$

$$h(t|\underline{Z}) = \psi(\underline{Z})h_0(t)$$

$$f(t|\underline{Z}) = \psi(\underline{Z})[S_0(t)]^{\psi(\underline{Z})-1} f_0(t)$$

$$S(t|\underline{Z}) = [S_0(t)]^{\psi(\underline{Z})}$$

7). Familia Poisson

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

$$S(t) = 1-\sum_{t=k}^{\infty} \frac{\lambda^t e^{-\lambda}}{t!}$$

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

Estos modelos son algunos de los más usuales, existen algunos otros como polinomios ortogonales, gaussiano inverso, translación, etc. que no se consideran en el presente trabajo.

## 2.5 MODELOS PARAMETRICOS

Muchos de los procesos de interés en el Análisis de Supervivencia pueden ser caracterizados por algún modelo matemático que dependa de uno o varios parámetros y que pertenezca a una familia específica de distribuciones, estos modelos son los modelos paramétricos. En general haremos el desarrollo de estos modelos a partir de su función de riesgo, e ilustraremos algunas situaciones donde su empleo es de utilidad. Es conveniente remarcar que los modelos paramétricos pueden ser propuestos a partir de la evidencia que, sobre un proceso particular, proporcione algún modelo no paramétrico.

Si tenemos una población homogénea cuyo tiempo de falla es

continuo, los modelos más comunes en el Análisis de Supervivencia son:

**-Modelo Exponencial**
**-Modelo Weibull**

Estas distribuciones admiten formas cerradas para las probabilidades y formas simples para las funciones de riesgo y de supervivencia, como fue visto en la sección anterior.

**-Modelo Log-Normal**
**-Modelo Gamma**

Estos modelos no tienen formas cerradas para las probabilidades pero son igualmente útiles.

### 2.5.1 MODELO EXPONENCIAL

Este modelo se genera al considerar una función de riesgo constante, $h(t) = \lambda$ , $\lambda > 0$. Es decir, la tasa instantánea de falla no depende de t por lo que el cambio condicional de falla es el mismo para cualquier intervalo de tiempo, esta propiedad es conocida como la pérdida de la memoria de la Exponencial.

Este modelo tiene su principal aplicación en los estudios para determinar el tiempo de vida útil de algunos artículos manufacturados. Pese a que el supuesto de una función de riesgo constante restringe mucho, el modelo Exponencial es aún útil en una amplia variedad de situaciones, y juega el papel de la distribución Normal en inferencia tradicional.

Su función de supervivencia es:

$$S(t) = \exp\left\{-\int_0^t h(x)\,dx\right\}$$

$$= \exp\left\{-\int_0^t \lambda dx\right\}$$

$$= \exp\left\{-\lambda t\right\}$$

$$f(t) = h(t)S(t) = \lambda \exp\left\{-\lambda t\right\}$$

$$E(T) = \int_0^\infty t\lambda \exp\{-\lambda t\}$$

$$= -t \exp\{-\lambda t\}\Big|_0^\infty + \int_0^\infty \exp\{-\lambda t\}dt = 1/\lambda$$

Como $E(T^2) = \int_0^\infty t^2\lambda \exp\{-\lambda t\}dt$

$$= 2\int_0^\infty 1/\lambda \exp\{-\lambda t\} = 2/\lambda^2$$

entonces

$$V(T) = E(T^2) - [E(T)]^2$$

$$= 2/\lambda^2 - (1/\lambda)^2 = 1/\lambda^2.$$

**2.5.2 MODELO WEIBULL**

Este modelo se genera a partir de una función de riesgo monótona,

esto es, se establece una relación creciente o decreciente entre el tiempo y la tasa de mortalidad.

Tiene aplicación en procesos para determinar el tiempo de vida útil de algunos artículos, así como en investigaciones biomédicas (ocurrencia de tumores en poblaciones humanas, experimentos de laboratorio con animales), y en muchas otras situaciones.

El modelo Weibull es una generalización del modelo Exponencial, su función de riesgo depende de dos parámetros y está dada por:

$$h(t) = \lambda p(\lambda t)^{p-1} \quad \text{con} \quad \lambda > 0,\ p > 0,\ t > 0$$

además

$$\frac{d\ h(t)}{dt} = \lambda(p-1)(\lambda p)(\lambda t)^{p-2}$$

por lo que h(t) es estrictamente creciente si p>1 y será estrictamente decreciente si p<1, cuando p=1 se reduce al modelo exponencial.

La función de supervivencia está dada por:

$$S(t) = \exp\left\{-\int_0^t h(x)dx\right\}$$

$$= \exp\left\{-\int_0^t \lambda p\ (\lambda x)^{p-1}dx\right\} = \exp\{-(\lambda t)^p\}$$

y la densidad por

$$f(t) = h(t)S(t) = p\lambda(\lambda t)^{p-1}\exp\{-(\lambda t)^p\}.$$

El r-ésimo momento $E(T^r)$ es

$$E(T^r) = \int_0^\infty t^r p\lambda(\lambda t)^{p-1}\exp\{-(\lambda t)^p\}$$

$$= p\lambda^{1-r}\int_0^\infty (\lambda t)^{p+r-1}\exp\{-(\lambda t)^p\}dt$$

Sean $u = (\lambda t)^p$ $\quad du = p(\lambda t)^{p-1}\lambda dt$

$$t = \lambda^{-1}u^{1/p} \quad dt = (\lambda p)^{-1}u^{1/p-1}\,du$$

entonces

$$E(T^r) = p\,\lambda^{1-r}\int_0^\infty u^{1+(r-1)/p}\exp\{-u\}(\lambda p)^{-1}u^{1/p-1}\,du$$

$$= \lambda^{-r}\int_0^\infty u^{r/p}\exp\{-u\}du$$

$$= \lambda^{-r}\Gamma(r/p + 1)\ .$$

Con $\Gamma(\cdot)$ la función gamma dada por

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

Por lo tanto

$$E(T) = \lambda^{-1}\Gamma(1/p + 1)$$

y

$$Var(T) = \lambda^{-2}\Gamma(2/p + 1) - \left(\lambda^{-1}\Gamma(1/p + 1)\right)^2$$

$$= \lambda^{-2}\left[\Gamma(2/p + 1) - \Gamma(1/p + 1)^2\right]$$

### 2.5.3 MODELO LOG-NORMAL

El modelo log-normal se ha utilizado como modelo de tiempo de vida, por ejemplo en el tiempo de falla de aislantes eléctricos y en los estudios del tiempo de aparición de cáncer pulmonar en los fumadores de cigarros.

Supongamos que $Y = \log T \sim N(\mu,\sigma^2)$, entonces la función de densidad de probabilidad de Y es :

$$f(y) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

de donde T tiene función de densidad de probabilidad dada por:

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

La función de supervivencia es:

$$S(t) = \int_t^\infty f(x)dx = 1-\int_0^t f(x)dx$$

$$= 1 - \int_0^t \frac{1}{\sqrt{2\pi}\ \sigma x} \exp\left\{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right\}dx$$

Haciendo $u = \dfrac{\log t-\mu}{\sigma}$, y tomando $\Phi(\cdot)$ la función de distribución de una normal estándar, se tiene que

$$S(t) = 1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$$

La función de riesgo es:

$$h(t) \quad \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right\}}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)}$$

Esta función toma el valor cero en t = 0, crece hasta un valor máximo luego tiende a cero cuando $t \to \infty$. Como la función de riesgo decrece para los valores grandes de T, el modelo puede ser inadecuado como un modelo de tiempo de vida en muchas situaciones.

Sin embargo, este modelo es apropiado para estudiar el tiempo de vida, en los casos en que no se está interesado en valores grandes de T.

La media y la varianza del modelo log-normal son:

$$E(T) = \exp\{\mu+\sigma^2/2\} \qquad V(T) = (\exp\{\sigma^2\}-1)(\exp\{2\mu+\sigma^2\})$$

### 2.5.4 MODELO GAMMA

El modelo Gamma se ajusta adecuadamente a una gran variedad de datos de tiempo de vida, por lo tanto existen varios procesos de falla que conducen a este modelo.

La función de densidad de probabilidad es:

$$f(t) = \frac{\lambda(\lambda t)^{k-1}\exp\{-\lambda t\}}{\Gamma(k)} \qquad \lambda,\ k,\ t > 0$$

Nótese que cuando k=1 se reduce al caso exponencial.

La función de supervivencia involucra la función gamma incompleta, ya que

$$S(t) = \int_t^{\infty} f(x)dx = 1-\int_0^t f(x)dx$$

y la función de riesgo es:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda(\lambda t)^{k-1}\exp\{-\lambda t\}}{\int_t^{\infty} \lambda(\lambda x)^{k-1}\exp\{-\lambda x\}dx}$$

La función h(t) es monótona creciente para k >1 con h(0)=0 y para $0 < k < 1$ h(t) es decreciente, con $\lim_{t\to 0^+} h(t) = \infty$ y $\lim_{t\to \infty} h(t) = \lambda$

Esta variable aleatoria tiene momentos dados por:

$$E(T) = k/\lambda \quad y \quad V(T) = k/\lambda^2.$$

# CAPITULO 3

# INFERENCIA NO PARAMETRICA DEL TIEMPO DE FALLA

Ahora discutiremos algunas técnicas no paramétricas para la estimación de la función de supervivencia, estas técnicas no requieren de supuestos de familias específicas de distribución y proporcionan alternativas a los métodos paramétricos. Los métodos no paramétricos están en conexión con procesos de bondad de ajuste para modelos complejos, e incluye la tabla de vida, la función de supervivencia empírica, el estimador Kaplan-Meier, entre otros. Es conveniente señalar que estos métodos pueden ser utilizados conjuntamente con algunos métodos paramétricos en el Análisis de Supervivencia.

La presencia de datos censurados provoca que los métodos comunes para datos sin censura sean modificados o se propongan métodos alternativos.

## 3.1 La Tabla de Vida

La tabla de vida es uno de los métodos más antiguos y más usual de presentar los datos de supervivencia, es empleada por lo menos desde el inicio del siglo XX, muy utilizada sobre todo por demógrafos y actuarios; aunque los métodos de la tabla de vida fueron usados por largo tiempo, la elaboración de sus propiedades estadísticas tiene un desarrollo más reciente. La introducción de la censura requiere una cuidadosa investigación de dichas propiedades estadísticas.

Comenzaremos por presentar la experiencia de supervivencia de un grupo de individuos, algunas veces referido como una cohorte ; se considera este grupo como muestra aleatoria de alguna población, la tabla de vida proporciona también estimadores de probabilidades de supervivencia de la población. Esta tabla es esencialmente una extensión, para el caso de datos censurados, de la tabla de frecuencias relativas usual. En esta tabla de vida se hace énfasis en la estimación de la probabilidad condicional de muerte en un intervalo, dada la supervivencia al inicio del mismo, y la probabilidad de supervivencia después del final del intervalo. Si no hay censura, excepto tal vez en el último intervalo, es fácil estimar estas cantidades, pero como la censura altera estas estimaciones daremos una cuidadosa descripción de la construcción de la tabla:

Dividimos el tiempo de observación en k+1 intervalos $I_j = [a_{j-1}, a_j)$ $j=1,\ldots,$ k+1, con $0 \leq a_0 < a_1 < \ldots < a_k < a_{k+1} = \infty$. Asignemos a cada miembro de una muestra aleatoria de individuos un tiempo de vida T y un tiempo de censura L. Ya que los datos están agrupados, solamente se conoce en que intervalos mueren o son censurados los individuos, y no el tiempo de vida y de censura exacto. Los datos consisten por lo tanto en el número de tiempos de vida y tiempos de censura ocurridos en cada uno de dichos intervalos. Definimos las siguientes cantidades:

$N_j$ = Número de individuos en riesgo al inicio de $I_j$( i.e. vivos y no censurados al tiempo $a_{j-1}$).

$D_j$= Número de muertes (i. e. el número de fallas) en $I_j=[a_{j-1}, a_j)$

$W_j$ = Número de censuras en $[a_{j-1}, a_j)$.

De lo anterior tenemos que:

$$N_1 = n \quad y \quad N_j = N_{j-1} - D_{j-1} - W_{j-1} \qquad j = 2,\ldots,k+1$$

La distribución de tiempos de vida para la población en estudio tiene función de supervivencia $S(t)$, y definimos las siguientes probabilidades para $j=1,2,\ldots,k+1$

$P_j = S(a_j) = P$(un individuo sobreviva más allá del inicio de $I_j$).

$p_j = P$(un individuo sobreviva más allá del inicio de $I_j$, dado que sobrevivió más allá de $I_{j-1}$).

$$= \frac{P_j}{P_{j-1}} \qquad \ldots(1)$$

$q_j = 1-p_j = P$ (un individuo muera en $I_j$, dado que sobrevivió más allá de $I_{j-1}$).

Finalmente obsérvese que:

$$P_j = p_1 p_2,\ldots,p_j \qquad j=1,\ldots,k+1, \qquad \ldots(2)$$

ya que la probabilidad de que un individuo sobreviva más allá de $I_j$ es igual al producto de las probabilidades condicionales de sobrevivir en cada uno de los intervalos anteriores a $I_j$.

Una vez que definimos estas cantidades, nuestro primer trabajo será la estimación de las $P_j$'s. Si los datos no son censurados, el estimador obvio, es el de máxima verosimilitud dado por $N_{j+1}/n$ , la proporción de individuos en el muestreo vivos al tiempo $a_j$. Esto no es así, si los intervalos contienen censuras ( i.e. tiempos censurados ). Ya que $N_{j+1}$ no es necesariamente el número de individuos vivos al tiempo $a_j$, puesto que es muy probable que algún individuo censurado este vivo en $a_j$, $N_{j+1}/n$ en muchos casos

sobreestima a $P_j$ . El método de tabla de vida supera este problema.

La idea detrás de la tabla de vida es emplear (2) para estimar $P_j$ y está basada en la observación de que aún cuando haya censura, es posible generalmente dar estimadores adecuados de las $P_j$'s . El análisis de la tabla de vida involucra estimadores de los $q_j$'s. El proceso usual es como sigue: Si un intervalo particular $I_j$ no tiene censuras (i.e. $W_j=0$) entonces un estimador adecuado de $q_j$ es $\hat{q}_j = D_j/N_j$ , ya que $q_j$ es la probabilidad condicional de que un individuo muera en $I_j$, dado que está vivo al inicio de $I_j$, si $W_j>0$ en el intervalo $I_j$, $D_j/N_j$ puede subestimar $q_j$, ya que alguno de los individuos censurados en $I_j$ puede morir antes del final de $I_j$. Es deseable por lo tanto hacer algún ajuste por los individuos censurados. El procedimiento comunmente usado es estimar $q_j$ por:

$$\hat{q}_j = \frac{D_j}{N_j - W_j/2} \quad \text{Si } N_j > 0$$

Y definimos $\hat{q}_j=1$,cuando $N_j=0$ por razones de conveniencia que aparecerán después. El denominador $N'_j = N_j - W_j/2$ se puede pensar como el número efectivo de individuos en riesgo en el intevalo $I_j$; esto supone que, en algún sentido, un individuo censurado está en riesgo por la mitad del intervalo. Este ajuste es arbitrario pero adecuado en muchas situaciones. Algunas veces otros estimadores de $q_j$ son preferibles, por ejemplo, si todos las censuras en $I_j$ ocurren exactamente al final de $I_j$, el estimador $\hat{q}_j = D_j/N_j$ puede ser más apropiado, mientras que si todos las censuras ocurren al principio de $I_j$,$\hat{q}_j = D_j/(N_j - W_j)$ puede ser conveniente.

Una vez estimados $q_j$ y $p_j = 1 - q_j$ y estimado por (2) $P_j$, la tabla

de vida sintetiza la distribución de los datos y el valor de los estimadores antes mencionados. La tabla generalmente incluye columnas para cada intervalo con los valores de $N_j, D_j, W_j, \hat{q}_j, \hat{P}_j$. Algunas veces incluyen columnas adicionales con cantidades como $N'_j, \hat{P}_j$ y ocasionalmente estimadores de otras características.

Por lo que el formato general es:

Tabla de vida

| Intervalo | Num. de muertes | Num. de censuras | Ind. en riesgo | $N_j$ | $\hat{q}_j$ | $\hat{P}_j$ | $\hat{P}_j$ |
|---|---|---|---|---|---|---|---|

### 3.2 Propiedades de los estimadores

Las cantidades $\hat{q}_j$, $\hat{P}_j$ y $\hat{P}_j$ son estimadores sujetos a variación muestral, y por ello es deseable tener alguna idea de su precisión.

#### 3.2.1 El Caso sin Censura

Este caso es sencillo, pero es conveniente considerarlo en primer lugar, puesto que se pueden obtener resultados exactos y porque varias fórmulas usadas cuando hay censura están estrechamente relacionadas a estos resultados.

Si no hay censura el número de muertes $D_1 \ldots D_k$ en los intervalos $I_1, \ldots, I_k$ siguen una distribución multinomial con función de densidad de probabilidad

$$\Pr\ (D_1, \ldots, D_k) = \frac{n!}{D_1! \ldots D_k!\ D_{k+1}!} \prod_{j=1}^{k+1} \pi_j^{D_j}$$

en donde

$$D_1 + D_2 + \ldots + D_{k+1} = n$$

y $\pi_1 + \pi_2 + \ldots + \pi_{k+1} = 1$

Ya que $\pi_j$ es la probabilidad condicional de que un individuo muera en $I_j$ entonces

$$\pi_j = p_{j-1}q_j = p_1 \cdots p_{j-1}q_j$$

Realizaremos la estimación por máxima verosimilitud, por lo que la función de verosimilitud es proporcional a:

$$\prod_{j=1}^{k+1} \pi_j^{D_j} = \prod_{j=1}^{k+1} (p_1 \cdots p_{j-1}q_j)^{D_j}$$

Ahora dasarrollando tenemos

$$\prod_{j=1}^{k+1} (p_1 \cdots p_{j-1}q_j)^{D_j} = (p_0\, q_1)^{D_1} (p_0 p_1 q_2)^{D_2} \cdots (p_0 p_1 \cdots p_k q_{k+1})^{D_{k+1}}$$

$$p_0^{D_1+\ldots+D_{k+1}} \quad p_1^{D_2+\ldots+D_{k+1}} \quad p_2^{D_3+\ldots+D_{k+1}}$$

$$\cdots p_k^{D_{k+1}} \prod_{j=1}^{k+1} q_j^{D_j}$$

Recordando que:

$$n = N_1 = D_1 + D_2 + \ldots + D_{k+1}\,, \quad N_j = N_{j-1} - D_{j-1} \text{ y } \quad p_0 = 1$$

tenemos lo siguiente:

$$N_1 - D_1 = D_1 + D_2 + \ldots + D_{k+1} - D_1 = D_2 + \ldots + D_{k+1}$$

$$N_2 - D_2 = N_1 - D_1 - D_2 = D_1 + D_2 + \ldots + D_{k+1} - D_1 - D_2 = D_3 + \ldots + D_{k+1}$$

.

.

.

$$N_k - D_k = N_1 - D_1 - D_2 \ldots - D_k = D_1 + D_2 + \ldots + D_{k+1} - D_1 - D_2 - \ldots - D_k = D_{k+1}$$

por lo que:

$$\prod_{j=1}^{k+1} (p_1 \ldots p_{j-1} q_j)^{D_j} = \prod_{j=1}^{k+1} q_j^{D_j} \, p_j^{N_j - D_j}$$

de donde

$$\text{Log } L(p_j, q_j) = \sum_{j=1}^{k+1} D_j (\log q_j) + (N_j - D_j)(\log p_j) \qquad j=1,\ldots,k+1$$

La maximización de la función anterior se basa en el sistema de ecuaciones simultáneas dado por

$$\frac{\partial \log L}{\partial q_j} = \frac{D_j}{q_j} + \frac{D_j - N_j}{1 - q_j} = 0 \qquad j = 1,\ldots,k+1$$

de aquí $\hat{q}_j = D_j/N_j$ si $N_j > 0$. Si $N_j = 0$ se define $q_j = 1$.

Por invarianza de los estimadores máximo verosímiles, el estimador de $p_j$ está dado por:

$$\hat{p}_j = 1 - \hat{q}_j$$

por lo que el estimador máximo verosimil de

$$\hat{P}_j = \hat{P}_1 \cdots \hat{P}_j = \frac{N_{j+1}}{n} .$$

Bajo el modelo multinomial, la distribución marginal de cada $N_{j+1}$ es binomial con parámetro $P_j$, es decir:

$$P\ (N_{j+1} = x) = \binom{n}{x} P_j^x\ (1-P_j)^{n-x} \qquad x = 0,1,2,\ldots,n$$

por lo que

$$E(\hat{P}_j) = P_j \quad y \quad Var(\hat{P}_j) = \frac{P_j(1-P_j)}{n}$$

Supongamos $j < l \leq k$ entonces la covarianza entre $\hat{P}_j$ y $\hat{P}_l$ es

$$\begin{aligned} cov(\hat{P}_j, \hat{P}_l) &= cov\left(N_{j+1}/n\ ,\ N_{l+1}/n\right) \\ &= 1/n^2 cov(N_{j+1},\ N_{l+1}) \\ &= 1/n^2 cov(n - N_{j+1}, n - N_{l+1}) \end{aligned}$$

como

$$N_{j+1} = N_j - D_j$$

entonces

$$N_{j+1} = N_1 - D_1 - D_2 - \ldots - D_j \quad y \text{ ya que } N_1 = n$$

$$N_{j+1} = n - D_1 - D_2 - \ldots - D_j$$

de donde

$$\operatorname{cov}\left(\hat{P}_j,\hat{P}_l\right)=1/n^2\operatorname{cov}\left(n-(n-D_1-D_2-\ldots-D_j),n-(n-D_1-\ldots-D_j-D_{j+1}-\ldots-D_l)\right)$$

$$=1/n^2\operatorname{cov}(D_1+D_2+\ldots+D_j,D_1+D_2+\ldots+D_j+D_{j+1}\ldots+D_l)$$

$$=1/n^2\operatorname{var}(D_1+D_2+\ldots+D_j)+1/n^2\operatorname{cov}(D_1+D_2+\ldots+D_j,D_{j+1}+\ldots+D_l)$$

$$=1/n^2\ nP_j(1-P_j)+\ 1/n^2\sum_{i=1}^{j}\sum_{s=j+1}^{l}\operatorname{cov}(D_i,D_s)$$

ya que conjuntamente las $D_i$'s tienen distribución multinomial, entonces $\operatorname{Cov}(D_i,D_s) = -n\pi_i\pi_s$ $(i\neq s)$, por lo que la covarianza es:

$$= 1/n\ P_j(1-P_j)\ -1/n\sum_{i=1}^{j}\sum_{s=j+1}^{l}\pi_i\pi_s$$

$$= 1/n\ P_j(1-P_j)\ -1/n\sum_{i=1}^{j}\pi_i\sum_{s=j+1}^{l}\pi_s$$

Por otro lado se tiene

$$\sum_{i=1}^{j}\pi_i\ =\ \pi_1+\pi_2+\ldots+\pi_j$$

$$= p_0\ q_1\ +\ p_0\ p_1\ q_2\ +\ p_0\ p_1\ p_2\ q_3\ +\ \ldots+\ p_0\ldots p_{j-1}\ q_j$$

$$= p_0\ (1-\ p_1)\ +\ p_0\ p_1\ (1\ -p_2)\ +\ p_0\ p_1\ p_2\ (1\ -p_3)\ +\ \ldots$$

$$+p_0\ldots\ p_{j-1}\ (1-p_j)$$

$$= p_0 - p_0 p_1 + p_0 p_1 - p_0 p_1 p_2 + p_0 p_1 p_2 - p_0 p_1 p_2 p_3 + \cdots$$

$$p_0 \cdots p_{j-1} - p_0 \cdots p_{j-1} p_j \,, \qquad \text{ya que} \qquad p_0 = 1$$

$$= 1 - p_0 \cdots p_{j-1} p_j = 1 - P_j \qquad y$$

$$\sum_{s=j+1}^{l} \pi_s = \sum_{s=1}^{l} \pi_s - \sum_{s=1}^{j} \pi_s = 1 - P_l - (1 - P_j) = P_j - P_l$$

por lo que finalmente

$$\mathrm{Cov}(\hat{P}_j, \hat{P}_l) = 1/n \; P_j(1-P_j) - 1/n \; (1-P_j)(P_j - P_l) = \frac{(1-P_j)P_l}{n} \qquad j < l.$$

Supongamos ahora que $N_j > 0$ y notemos entonces que la distribución condicional de $D_j$ , dado $N_j$ , es binomial, con parámetros $N_j$ y $q_j$ .

Usaremos aquí los siguientes resultados de momentos condicionales

1) $E(Y) = E\left[E(Y|X)\right]$

2) $\mathrm{Var}(Y) = E\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[E(Y|X)\right]$

La esperanza de $\hat{q}_j = D_j/N_j$ es

$$E(\hat{q}_j) = E\left[E(\hat{q}_j|N_j)\right] = E(q_j) = q_j$$

$$\mathrm{Var}(\hat{q}_j) = E\left[\mathrm{Var}(\hat{q}_j|N_j) + \mathrm{Var}\left[E(\hat{q}_j|N_j)\right]\right.$$

$$= E\left(p_j q_j | N_j\right) + \mathrm{Var}(q_j) = p_j q_j E(1/N_j)$$

pues la $Var(q_j) = 0$

Si $N_j = 0$, las expresiones anteriores no son correctas ya que en dicho caso, $q_j$ se define por conveniencia como la unidad. El error que se comete es pequeño, ya que la probabilidad de que $N_j = 0$ es en general también pequeña. Si se condicionan los resultados a $N_j > 0$, entonces serán exactos.

Ahora estamos interesados en la covarianza de las $\hat{q}_j'$ s. Note que que para $i < j$ y $N_j > 0$ .

$$E\left[(\hat{q}_j - q_j)\,|\,\hat{q}_i\right] = E\left[\left(D_j/N_j - q_j\right)|D_i/N_i\right]$$

$$= E\left[E\left(D_j/N_j - q_j\right)|D_i/N_i, N_j\right]$$

$$= E(0|\hat{q}_i) = 0 .$$

Entonces para $i < j$

$$Cov(\hat{q}_i, \hat{q}_j) = E\left[(\hat{q}_i - q_i)(\hat{q}_j - q_j)\right]$$

$$= E\left[(\hat{q}_i - q_i)\; E(\hat{q}_j - q_j|\hat{q}_i)\right] = 0$$

Sin embargo, a pesar de que la covarianza entre $\hat{q}_i$ y $\hat{q}_j$ es cero, no son independientes.

Notemos que como $\hat{p}_j = 1 - \hat{q}_j$ la varianza y covarianza de $\hat{p}_j$ están dadas también por las dos expresiones anteriores, i.e.

$$V(\hat{p}_j) = p_j q_j E(1/N_j) \text{ y } cov(\hat{p}_i, \hat{p}_j) = 0 \qquad i < j .$$

El caso sin censura es simple. Se pueden hacer estimaciones de media, varianza y covarianza de los estimadores y se puede, en suma obtener intervalos de confianza para $P_j=S(a_j)$, usando teoría estándar para la distribución binomial.

ALGUNAS OBSERVACIONES SOBRE LAS TABLAS DE VIDA

Cuando los datos de tiempo de vida están agrupados, las tablas de vida dan una descripción concisa de la experiencia de supervivencia de los individuos en la muestra y también suministran estimadores no paramétricos de las probabilidades de supervivencia.

La tabla de vida ha sido tradicionalemente usada en áreas médicas, por ejemplo en estudios secuenciales de individuos que sufren enfermedades crónicas. Tales estudios frecuentemente involucran a un gran número de individuos, por lo que se llega a "perder la secuencia" de algunos de ellos u otros permanecen vivos aún al momento en que se colectan los datos. En el caso anterior su tiempo de vida es censurado.

En la formación de la tabla de vida, (Lawless 1982) menciona que es conveniente hacer los intervalos del mismo tamaño, pero no es necesario.

El número de intervalos usados dependerá de la cantidad de datos disponibles y de la finalidad del análisis. Ciertas propiedades estadísticas de los estimadores se tienen cuando el número de intervalos es muy grande, aunque generalmente es deseable tener al menos 8 ó 10 (Lawless 1982).

Finalmente lo apropiado de los estimadores de la tabla de vida estándar, en el caso con censura, dependerán de ciertos supuestos

acerca de los patrones de censura en la población.

### 3.2.2 El Caso con Censura

Para el caso general que involucra censura, las propiedades de los métodos de estimación en la tabla de vida dependen de algunos factores como el mecanismo particular de censura, el número de intervalos considerados para la tabla y ciertos resultados asintóticos para los estimadores, entre otros.

Primeramente haremos algunas consideraciones sobre los estimadores en el caso con observaciones censuradas para, posteriormente, desarrollar dichos estimadores bajo un mecanismo de censura aleatoria usando resultados de distribución asintótica.

El estimador más usual de la varianza es uno sugerido por Greenwood. En este caso $\hat{q}_j\hat{p}_j/N'_j$ se toma como una aproximación de la varianza de $\hat{q}_j$ o $\hat{p}_j$ y se deriva una aproximación a la varianza de $\hat{P}_j = \hat{p}_1 \ldots \hat{p}_j$ usando la siguiente aproximación asintótica estándar

$$\text{Var}\left(\hat{P}_j\right) \approx \sum_{i=1}^{j}\sum_{l=1}^{j} \frac{\partial P_j}{\partial p_i}\frac{\partial P_j}{\partial p_l}\,\text{Cov}(\hat{p}_i,\hat{p}_l) \qquad \ldots(3)$$

Suponiendo, como en el caso sin censura, que $\text{cov}(\hat{p}_i,\hat{p}_l) = 0$ y tomando la aproximación, ya escrita anteriormente, para la

$$\hat{\text{Var}}(\hat{p}_j) = \hat{q}_j\hat{p}_j/N'_j \qquad \ldots(4)$$

obtenemos:

$$\hat{\text{Var}}(\hat{P}_j) = \hat{P}_j^{\,2}\sum_{i=1}^{j}\frac{\hat{q}_i}{\hat{p}_i N'_i} \qquad \ldots(5)$$

que es el estimador propuesto por Greenwood. (su desarrollo se hará posteriormente ).

Este estimador es razonable a condición de que la E ($N'_j$) no sea demasiado pequeña, sin embargo, en tales circunstancias la distribución de $\hat{P}_j$ es muy sesgada por lo que su varianza no es una buena estimación de la precisión.

En la estimación de Var($\hat{P}_j$) hay algunas aproximaciones que analizaremos.

La aproximación (3) es razonable si n es suficientemente grande, pero la aproximación cov($\hat{P}_i$,$\hat{P}_i$) = 0 i ≠ 1 y (4), las cuales se escriben por analogía con el caso sin censura, son más cuestionables. Lo adecuado de ellas dependerá del mecanismo de censura y la distribución del tiempo de vida en el problema de estudio.

En el caso sin censura

$$\hat{V}(\hat{P}_j) = \hat{P}_j^{\,2} \sum_{j=1} \frac{\hat{q}_j}{\hat{p}_j N'_j}$$

Reproduce la estimación de la varianza de $\hat{P}_j$ para este caso.

Si no hay censura

$$N'_j = N_j = n\hat{P}_{j-1} \quad \text{ya que} \quad \hat{P}_{j-1} = N_j/n \ , \quad \text{de aquí}$$

$$N_j \hat{p}_j = n\hat{p}_j\hat{P}_{j-1} = n\hat{P}_j$$

entonces

$$\hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{P}_i N_i'} = \hat{P}_j^2/n \sum_{i=1}^{j} \frac{1-\hat{P}_i}{\hat{P}_i} = \hat{P}_j^2/n \sum_{i=1}^{j} \left(1/\hat{P}_i - 1/\hat{P}_{i-1}\right)$$

$$= \hat{P}_j^2/n \left(1/\hat{P}_1 - 1/\hat{P}_0 + 1/\hat{P}_2 - 1/\hat{P}_1 + 1/\hat{P}_3 - 1/\hat{P}_2 + - + \ldots + 1/\hat{P}_j\right)$$

$$= \hat{P}_j^2/n\left(-1 + 1/\hat{P}_j\right) = \hat{P}_j^2/n \left(\frac{1-\hat{P}_j}{\hat{P}_j}\right) = \frac{\hat{P}_j(1-\hat{P}_j)}{n}$$

En suma, detrás de las preguntas acerca de la estimación adecuada de la varianza (3) están las cuestiones de sesgo en $\hat{q}_j$ $\hat{P}_i$ y $\hat{P}_j$. En el caso sin censura hay estimadores insesgados de $p_j, q_j$ y $P_j$ respectivamente, pero esto es falso cuando hay censura.

Un problema al examinar las propiedades de los estimadores de la tabla de vida es especificar un modelo realista para la censura. Las propiedades de los estimadores en una situación dada dependen del mecanismo particular de censura en operación. Parece plausible que si el tiempo de censura está distribuido de manera uniforme sobre los intervalos y éstos no son muy anchos, y si hay independencia de los tiempos de vida en algún sentido, entonces los estimadores de la tabla de vida estándar pueden ser adecuados.

### 3.2.3 Propiedades Asintóticas de los Estimadores bajo un Modelo de Censura Aleatoria.

Desarrollaremos en esta parte los estimadores de las varianzas de $\hat{p}_j$ y $\hat{q}_j$ bajo el supuesto que las censuras dentro de la población siguen un modelo aleatorio.

Supongamos que los tiempos de vida $T_1,\ldots,T_n$ de n individuos bajo estudio son independientes e idénticamente distribuidos con función de supervivencia S(t). El primer objetivo del método de la tabla de vida es estimar $S(a_j)$, j=1,...,k donde los intervalos $I_j = [a_{j-1},a_j)$ están definidos como antes. Supongamos que cada individuo tiene asociado también un tiempo de censura $L_i$ y que los $L_i$'s son independientes e idénticamente distribuidos, independientes de los tiempos de vida, con función de supervivencia G(t). Los datos en la tabla de vida consisten del número de tiempos de vida (muertes) y los tiempos de censura en los k+1 intervalos. Esto se puede representar por el vector $D=(D_1,W_1,\ldots,D_k,W_k,N_{k+1})$, donde $D_j$ y $W_j$ denotan el número de muertes y censuras respectivamente, en $I_j$ y $N_{k+1}=n-D_1-W_1-\ldots-D_k-W_k$ es el número de individuos que sobreviven después del tiempo $a_k$. En notación usual, D tiene una distribución multinomial con parámetros n y $\underset{\sim}{\Pi} = (\Pi_1^D,\Pi_1^W,\ldots,\Pi_k^D,\Pi_k^W,\Pi_{k+1}^N)$ donde $\Pi_1^D+\Pi_1^W+\ldots+\Pi_{k+1}^N = 1$. Las probabilidades de $\underset{\sim}{\Pi}$ pueden ser determinadas en términos de S(t) y G(t). Por ejemplo:

$$\Pi_1^D = P(\text{un individuo muera en } I_1)$$

$$= P(T_1 \leq a_1, T_1 \leq L_1)$$

$$= \int_0^{a_1} G(x)\,|dS(x)|\,dx$$

de manera análoga,

$$\Pi_j^D = \int_{a_{j-1}}^{a_j} G(x)\,|dS(x)|$$

$$\Pi_j^W = \int_{a_{j-1}}^{a_j} S(x)\,|dG(x)|$$

Como D es multinomial, entonces $\sqrt{n}\,(D-n\Pi)$ converge a una normal multivariada con media $\underline{0}$ y matriz de covarianza

$$\Sigma = \text{diagonal}\,(\Pi_1^D,\Pi_1^W,\ldots,\Pi_{K+1}^N) - \underline{\Pi}\,\underline{\Pi}'.$$

En la tabla de vida estándar los estimadores $\hat{q}_j = D_j/(N_j - 1/2W_j)$ son funciones de $D_1,W_1,\ldots,N_{k+1}$ con k derivadas continuas. Ya que la distribución de $\sqrt{n}\,(\hat{q}-q^*)$ también converge a una normal multivariada con media $\underline{0}$ y matriz de covarianza $\Sigma$, donde

$$\hat{q} = (\hat{q}_1,\ldots,\hat{q}_k) \text{ y } q^* = (q_1^*,\ldots,q_k^*)$$

con $q^*$ un estimador consistente de $\hat{q}$ . Como

$$\hat{q}_j = \frac{D_j/n}{N_j/n - W_j/(2n)}$$

se sigue que

$$q_j^* = \frac{\Pi_j^D}{\Pi_j^N - \Pi_j^W/2}$$

donde

$$\Pi_j^N = E(N_j/n) = G(a_{j-1})S(a_{j-1})$$

por lo que

$$q_j^* = \frac{\left( \int_{a_{j-1}}^{a_j} G(x) \, |dS(x)| \right)}{\left( G(a_{j-1}) S(a_{j-1}) - 1/2 \int_{a_{j-1}}^{a_j} S(x) \, |dG(x)| \right)}$$

Como por lo general $q_j^* \neq q_j$ entonces los estimadores de la tabla de vida estándar no son consistentes, ya que tampoco $\hat{P}_j$ lo es de $P_j$.

Por lo tanto una cuestión práctica importante es si el sesgo asintótico en los estimadores es suficientemente pequeño para que esta inconsistencia sea importante. Esto parece ser un hecho en muchos casos.

$\Sigma$ es una matriz diagonal, y además $\hat{q}_i$ y $\hat{q}_j$ $(i \neq j)$ son asintóticamente no correlacionadas como en el caso sin censura.

La varianza asintótica de $\sqrt{n}(\hat{q}_j - q_j^*)$ es:

$$\text{Var asint} \left[ \sqrt{n} (\hat{q}_j - q_j^*) \right] = \frac{q_j^* - q_j^{*2} \left[ (\Pi_j^N - \Pi_j^W / 4) \right]}{\left( \Pi_j^N - \Pi_j^W / 2 \right)^2}$$

Hagamos la deducción de esta expresión, primero observemos lo siguiente:

$$\Pi_j^D + \Pi_j^W = \int_{a_{j-1}}^{a_j} G(x) \, |dS(x)| \, dx + \int_{a_{j-1}}^{a_j} S(x) \, |dG(x)| \, dx$$

$$= \int_{a_{j-1}}^{a_j} G(x)\,|dS(x)| + S(x)\,|dG(x)|\,dx = G(x)S(x)\Big|_{a_{j-1}}^{a_j}$$

$$= S(a_j)G(a_j) - S(a_{j-1})G(a_{j-1}) = \Pi_j^N - \Pi_{j-1}^N$$

y como

$$\Pi_{j+1}^N = \Pi_j^N - \Pi_j^D - \Pi_j^W \quad \text{y} \quad \Pi_1^N = 1$$

entonces

$$\Pi_{k+1}^N = 1 - \Pi_1^D - \Pi_1^W - \Pi_2^D - \Pi_2^W \cdots - \Pi_k^D - \Pi_k^W$$

como: $\Sigma^* = \text{diagonal}\,(\Pi_1^D, \Pi_1^W, \ldots, \Pi_{k+1}^N) - \underline{\Pi}\,\underline{\Pi}'$.

$$= \begin{bmatrix} \Pi_1^D(1-\Pi_1^D) & -\Pi_1^D\Pi_1^W & & -\Pi_1^D\Pi_2^D - \cdots - \Pi_1^D\Pi_{k+1}^N \\ -\Pi_1^D\Pi_1^W & \Pi_1^W(1-\Pi_1^W) & & -\Pi_1^W\Pi_2^D - \cdots - \Pi_1^W\Pi_{k+1}^N \\ \vdots & & \ddots & \\ -\Pi_1^D\Pi_{k+1}^N & \cdots & & \Pi_{k+1}^N(1-\Pi_{k+1}^N) \end{bmatrix}$$

Usando la aproximación (Lawless 1982)

$$\sqrt{n}\left[g(T_{1n},\ldots,T_{kn}) - g(\theta_1,\ldots,\theta_k)\right] \xrightarrow{D} N\left[0, \sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}\frac{\partial g}{\partial \theta_i}\frac{\partial g}{\partial \theta_j}\right]$$

en donde $T_{1n},\ldots,T_{kn}$ son estadísticas tales que, cuando $n \to \infty$

$$\sqrt{n}(T_{1n}-\theta_1,\ldots,T_{kn}-\theta_k) \xrightarrow{D} N(0,\Sigma)$$

con $\Sigma=(\sigma_{ij})$ y $g(x_1,\ldots,x_k)$ una función cuyas primeras derivadas existen.

En nuestro caso tenemos que $g(\cdot)=q^*$, $\theta = \underline{\Pi}$ y $\Sigma^* = (\sigma_{ij})$.

Para j=1 tenemos

$$q_1^* = \frac{\Pi_1^D}{\Pi_1^N - \Pi_1^{N}/2} = \frac{\Pi_1^D}{1-\Pi_1^N/2}$$

de donde

$$\frac{\partial q_1^*}{\partial \Pi_1^D} = \frac{1}{1-\Pi_1^N/2} \quad y \quad \frac{\partial q_1^*}{\partial \Pi_j^D} = 0 \qquad j= 2,\ldots,k$$

$$\frac{\partial q_1}{\partial \Pi_1^N} = \frac{1}{2(1-\Pi_1^N/2)} \quad y \quad \frac{\partial q_1^*}{\partial \Pi_j^D} = 0 \qquad j= 2,\ldots,k$$

por lo que

$$\sum_{i=1}\sum_{j=1} \sigma_{ij} \left(\frac{\partial q_1^*}{\partial \Pi_1}\right)\left(\frac{\partial q_1^*}{\partial \Pi_j}\right) = \frac{\Pi_1^D(1-\Pi_1^D)}{(1-\Pi_1^N/2)^2} + \frac{\Pi_1^N(1-\Pi_1^N)\Pi_1^{D^2}}{4(1-\Pi_1^N/2)^4} - \frac{2\Pi_1^{D^2}\Pi_1^N}{2(1-\Pi_1^N)^3}$$

$$= \frac{q_1^*(1-\Pi_1^D)}{(1-\Pi_1^N/2)} + \frac{q_1^{*2}\Pi_1^N(1-\Pi_1^N)}{4(1-\Pi_1^N/2)^2} - \frac{q_1^{*2}\Pi_1^N}{(1-\Pi_1^N/2)}$$

$$= \frac{q_1^*}{(1-\Pi_1^N/2} - q_1^{*2} + \frac{q_1^{*2}\Pi_1^N(1-\Pi_1^N)}{4(1-\Pi_1^N/2)^2} - \frac{q_1^{*2}\Pi_1^N}{(1-\Pi_1^N/2)}$$

$$= \frac{q_1^*}{(1-\Pi_1^N/2)} - \frac{q_1^{*2}}{(1-\Pi_1^N/2)^2}\left[(1-\Pi_1^N/2)^2 - \frac{\Pi_1^N(1-\Pi_1^N)}{4} + \Pi_1^N(1-\Pi_1^N/2)\right]$$

$$= \frac{q_1^*}{(\Pi_1^N - \Pi_1^W/2)} - \frac{q_1^{*2}}{(\Pi_1^N - \Pi_1^W/2)^2}\,(\Pi_1^N - \Pi_1^W/4)$$

ya que $\Pi_1^N = 1$

$$\text{para } j=2 \quad q_2^* = \frac{\Pi_2^D}{(\Pi_2^N - \Pi_2^W/2)}$$

entonces

$$\frac{\partial q_2^*}{\partial \Pi_1^D} = \frac{\Pi_2^D}{(\Pi_2^N - \Pi_2^W/2)^2} \qquad \frac{\partial q_2^*}{\partial \Pi_2^D} = \frac{1}{(\Pi_2^N - \Pi_2^W/2)}$$

$$\frac{\partial q_2^*}{\partial \Pi_1^W} = \frac{\Pi_2^D}{(\Pi_2^N - \Pi_2^W/2)^2} \qquad \frac{\partial q_2^*}{\partial \Pi_2^W} = \frac{\Pi_2^D}{2(\Pi_2^N - \Pi_2^W/2)^2}$$

por lo que

$$\sum_{i=1} \sum_{j=1} \sigma_{ij} \left(\frac{\partial q_2^*}{\partial \Pi_i}\right)\left(\frac{\partial q_2^*}{\partial \Pi_j}\right) = \frac{\Pi_2^{D\,2}\,\Pi_1^D\,(1-\Pi_1^D)}{(\Pi_2^N - \Pi_2^W/2)^4} + \frac{\Pi_2^{D\,2}\,\Pi_1^W\,(1-\Pi_1^W)}{(\Pi_2^N - \Pi_2^W/2)^4} + \frac{\Pi_2^D(1-\Pi_2^D)}{(\Pi_2^N - \Pi_2^W/2)^2}$$

$$+ \frac{\Pi_2^{D\,2}\,\Pi_2^W\,(1-\Pi_2^W)}{4(\Pi_2^N - \Pi_2^W/2)^4} - \frac{2\Pi_2^{D\,2}\,\Pi_1^D\,\Pi_1^W}{(\Pi_2^N - \Pi_2^W/2)^4} - \frac{2\Pi_2^{D\,2}\,\Pi_1^W}{(\Pi_2^N - \Pi_2^W/2)^3}$$

$$- \frac{2\Pi_2^{D\,2}\,\Pi_1^D\,\Pi_2^W}{2(\Pi_2^N - \Pi_2^W/2)^4} - \frac{2\Pi_2^{D\,2}\,\Pi_1^D}{(\Pi_2^N - \Pi_2^W)^3} - \frac{2\Pi_2^{D\,2}\,\Pi_1^W\,\Pi_2^W}{2(\Pi_2^N - \Pi_2^W/2)^4}$$

$$- \frac{2\pi_2^{D^2}\pi_2^W}{2(\pi_2^N-\pi_2^W)^3}$$

$$= \frac{q_2^{*2}\pi_1^D(1-\pi_1^D)}{(\pi_2^N-\pi_2^W/2)^2} + \frac{q_2^{*2}\pi_1^W(1-\pi_1^W)}{(\pi_2^N-\pi_2^W/2)^2} + \frac{q_2^*}{(\pi_2^N-\pi_2^W/2)^2}$$

$$- \frac{q_2^{*2}(\pi_2^N-\pi_2^W/2)^2}{(\pi_2^N-\pi_2^W/2)^2} + \frac{q_2^{*2}\pi_2^W(1-\pi_2^W)}{4(\pi_2^N-\pi_2^W/2)^2} - \frac{2q_2^{*2}\pi_1^D\pi_1^W}{(\pi_2^N-\pi_2^W/2)^2}$$

$$- \frac{2q_2^{*2}\pi_1^D}{(\pi_2^N-\pi_2^W/2)} - \frac{q_2^{*2}\pi_1^D\pi_2^W}{(\pi_2^N-\pi_2^W/2)^2} - \frac{2q_2^{*2}\pi_1^W}{(\pi_2^N-\pi_2^W/2)}$$

$$- \frac{q_2^{*2}\pi_1^W\pi_2^W}{(\pi_2^N-\pi_2^W/2)^2} \qquad \frac{q_2^{*2}\pi_2^W}{(\pi_2^N-\pi_2^W/2)}$$

$$= \frac{q_2^*}{(\pi_2^N-\pi_2^W/2)^2} - \frac{q_2^*}{(\pi_2^N-\pi_2^W/2)^2}\Big(-\pi_1^D(1-\pi_1^D) \; -\pi_1^W(1-\pi_1^W)$$

$$+ (\pi_2^N-\pi_2^W/2)^2 - \frac{\pi_2^W(1-\pi_2^W)}{4} + 2\pi_1^D\pi_1^W + 2\pi_1^D(\pi_2^N-\pi_2^W/2)$$

$$+ \pi_1^D\pi_2^W + 2\pi_1^W(\pi_2^N-\pi_2^W/2) + \pi_1^W\pi_2^W + \pi_2^W(\pi_2^N-\pi_2^W/2)\Big)$$

$$= \frac{q_2^*}{(\pi_2^N-\pi_2^W/2)} - \frac{q_2^*}{(\pi_2^N-\pi_2^W/2)^2}\Big( -\pi_1^D + \pi_1^{D^2} - \pi_1^W + \pi_1^{W^2}$$

$$+\pi_2^{N^2} - \pi_2^W/4 + 2\pi_1^D \pi_1^W + 2\pi_1^D \pi_2^N + 2\pi_1^W \pi_2^N \Big)$$

$$= \frac{q_2^*}{(\pi_2^N - \pi_2^W/2)} - \frac{q_2^{*2}}{(\pi_2^N - \pi_2^W/2)^2} \left( 1 - \pi_1^D - \pi_1^W - \pi_2^W/4 \right)$$

$$= \frac{q_2^*}{(\pi_2^N - \pi_2^W/2)} - \frac{q_2^{*2}}{(\pi_2^N - \pi_2^W/2)^2} \left( \pi_2^N - \pi_2^W/4 \right)$$

por lo que de manera general tenemos

$$\text{Var asint.} \left[ \sqrt{n} \, (\hat{q}_j - q_j^*) \right] = \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma_{ij} \left( \frac{\partial q^*}{\partial \pi_i} \right) \left( \frac{\partial q^*}{\partial \pi_j} \right)$$

$$= \frac{q_j^*}{(\pi_j^N - \pi_j^W/2)} - \frac{q_j^{*2}}{(\pi_j^N - \pi_j^W/2)^2} \left( \pi_j^N - \pi_j^W/4 \right).$$

El estimador usual de la $\text{var}(\hat{q}_j)$ es

$$\hat{\text{Var}}(\hat{q}_j) = \frac{\hat{q}_j - \hat{q}_j^2}{N_j'}$$

Si $q_j$ y $q_j^*$ son parecidos, esto tiende a sobreestimar el valor verdadero de la varianza algunas veces, ya que $N_j'/n$ converge en probabilidad al denominador $(\pi_j^N - \pi_j^W/2)$ y el término entre paréntesis es menor que la unidad. Si $q_j$ es pequeña entonces $q_j^{*2}(\cdot)$ y $\hat{q}_j^2$ son pequeños comparados con $q_j^*$ y $\hat{q}_j$ y la igualdad entre las dos formulas es posible.

Para la var $(\hat{P}_j)$ tenemos que la distribución límite de $\sqrt{n}\,(\hat{P}_j - P_j)$'s es una normal multivariada, con media, varianza y covarianza que se determinan por métodos usuales. Usando nuevamente la aproximación anterior, ahora para $\hat{P}_j = \hat{p}_1 \ldots \hat{p}_j$ tenemos que:

$$\mathrm{Cov}\left(\sqrt{n}\,\hat{P}_j, \sqrt{n}\,\hat{P}_i\right) \cong \sum_{l=1}^{j}\sum_{s=1}^{i}\left[\frac{\partial \hat{P}_j}{\partial \hat{p}_l}\frac{\partial \hat{P}_i}{\partial \hat{p}_s}\bigg|_{p^*}\right] \mathrm{Cov}\left(\sqrt{n}\,\hat{p}_l, \sqrt{n}\,\hat{p}_s\right)$$

como

$$\frac{\partial \hat{P}_j}{\partial \hat{p}_i} = \frac{\partial\, \hat{p}_1 \ldots \hat{p}_j}{\partial\, \hat{p}_i} = \hat{p}_1 \ldots \hat{p}_{i-1}\hat{p}_{i+1}\ldots\hat{p}_j = \frac{\hat{P}_j}{\hat{p}_i}$$

tenemos que:

$$\mathrm{Cov}\left(\sqrt{n}\,\hat{P}_j, \sqrt{n}\,\hat{P}_i\right) = \sum_{l=1}^{j}\sum_{s=1}^{i}\frac{P_j^*}{p_l^*}\frac{P_i^*}{p_s^*}\,\mathrm{Cov}\left(\sqrt{n}\,\hat{p}_l, \sqrt{n}\,\hat{p}_s\right)$$

$$= P_j^* p_i^* \sum_{l=1}^{j}\frac{\mathrm{var}(\sqrt{n}\,\hat{p}_j)}{p_i^{*2}}$$

ya que

$\mathrm{cov}(\sqrt{n}\,\hat{P}_i, \sqrt{n}\,\hat{P}_s) = 0 \quad i \neq s$

Sean $\hat{P} = (\hat{P}_1 \ldots \hat{P}_k)$ y $P^* = (P_1^*, \ldots, P_k^*)$, donde $P_j^* = p_1^* \ldots p_j^*$.

Con $p_j^*$ un estimador consistente de $\hat{p}_j$. Entonces la distribución límite de $\sqrt{n}(\hat{P} - P^*)$ es normal multivariada con media $\underline{0}$ y matriz de

covarianza

$$p_j^* p_i^* \sum_{i=1}^{j} \frac{\text{var}\left[\sqrt{n}\,(\hat{q}_i - q_i^*)\right]}{(1-q_i^*)^2}$$

que se deduce del hecho que

$$\sqrt{n}\,(\hat{p}_j - p_j^*) = \sqrt{n}(1-q_j^* - (1-\hat{q}_j) = \sqrt{n}\,(\hat{q}_j - q_j^*)$$

y sustituyendo en la expresión

$$p_j^* p_i^* \sum_{i=1}^{j} \frac{\text{var}(\sqrt{n}\,\hat{p}_i)}{p_i^{*2}}$$

obtenemos el resultado anterior.

Ahora como la varianza se calcula con j = 1 , utilizando el estimador usual para la varianza de $\hat{q}_j$ y remplazando $p_i^*$ y $q_i^*$ por $\hat{p}_i$ y $\hat{q}_i$ tenemos que:

$$p_j^* p_i^* \sum_{i=1}^{j} \frac{\text{var}\sqrt{n}\,(\hat{q}_i - q_i^*)}{(1-q_i^*)^2} \cong \hat{p}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i - \hat{q}_i^2}{(1-\hat{q}_i)^2 N_i'} = \hat{p}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{N_i' \hat{p}_i}$$

que es la fórmula de Greenwood.

### 3.3 Función de Supervivencia Empírica

Una manera usual de representar los datos de supervivencia es calcular la función de supervivencia empírica o, equivalentemente, la función de distribución empírica. Esto provee un estimador no

paramétrico de la función de supervivencia para la distribución de tiempo de vida en estudio. Si en un muestreo de tamaño n no hay observaciones censuradas, la función de supervivencia empírica (F.S.E.) se define como:

$$\hat{S}(t) = \frac{\text{Número de observaciones} \geq t}{n} \qquad t \geq 0$$

$\hat{S}(t)$ es una función escalonada decreciente con saltos de tamaño 1/n después de cada tiempo de vida observado si todos son distintos. De manera general si tenemos d tiempos de vida iguales a t , la F.S.E. tendrá saltos de tamaño d/n después de cada tiempo t.

Cuando tratamos con datos censurados es necesario hacer algunas modificaciones a $\hat{S}(t)$ , ya que el número de tiempos de vida mayores o iguales a t, por lo general, no se conoce con exactitud. Esta modificación recibe comunmente el nombre del estimador "producto-límite" de la función de supervivencia, o también el estimador Kaplan-Meier (K-M).

### 3.4 El Estimador Kaplan-Meier

La motivación para el estimador (K-M) es esencialmente la misma que para los estimadores de las probabilidades de supervivencia $P_j=\hat{p}_1 \ldots \hat{p}_j$ en la tabla de vida. Esto es, el estimador K-M se forma a partir de un producto y cada término se puede pensar como un estimador de la probabilidad condicional de que un individuo sobreviva después del tiempo $t_j$, dado que sobrevivió inmediatamente antes de $t_j$. El estimador K-M es de hecho un caso límite del proceso de la tabla de vida estándar, que se obtiene cuando el número de intervalos en la tabla de vida tiende a infinito y el tamaño de los intervalos tiende a cero.

Este estimador puede construirse como un estimador de tipo máximo verosímil de la siguiente manera:

En la definición de S(t) supondremos que tenemos k tiempos de vida distintos $t_1 < \ldots < t_k$ con $d_j$ muertes en $t_j$ y $n_j$ individuos en riesgo en $t_j$.

En el intervalo $[t_{j-1}, t_j)$ supondremos que hay $\lambda_j$ tiempos de censura $L_i^j$ $(i=1,\ldots,\lambda_j)$ $j=1,\ldots,k+1$, con $t_0=0$ y $t_{k+1}=\infty$.

Por lo que

$$n_j = \sum_{j=1}^{k} (\lambda_i + d_i).$$

Hacemos la convención de ajustar los tiempos de censura infinitesimalmente a la derecha, de modo que las censuras en $t_j$ se considera que ocurren justo después de $t_j$. Esta convención es adecuada, ya que un individuo censurado al tiempo L casi seguramente sobrevive después de L.

En un caso general, la probabilidad de que un individuo muera en $t_j$ está dada por $S(t_j)-S(t_j+0)$ con

$$S(t_j+0) = \lim_{x \to 0^+} S(t_j+x)$$

La función de verosimilitud propuesta es de la forma:

$$L = \prod_{j=1}^{k} \left[ \prod_{i=1}^{\lambda_j} S(L_i^j) \right] \left[ S(t_j) - S(t_j+0) \right]^{d_j} \left\{ \prod_{i=1}^{\lambda_{k+1}} S(L_i^{k+1}) \right\}$$

Ya que $S(L_i^j)$ es la contribución de las censuras para $i = 1, \ldots, \lambda_j$

$\left[ S(t_j) - S(t_j+0) \right]^{d_j}$ es la contribución de las muertes por cada intervalo, y finalmente $S(L_i^{k+1})$ es con lo que contribuyen los individuos que permanecen vivos (censurados) después del tiempo final de observación.

Para maximizar L con respecto a S(t), primeramente observemos que $\hat{S}(t)$ debe ser discontinua en las $t_j$'s ya que de lo contrario $S(t_j+0)=S(t_j)$ y tendríamos L = 0 con lo que no habría máximo.

Además observemos que:

$$\hat{S}(t_1)=\hat{S}(L_j^1) = 1 \qquad i= 1,\ldots,\lambda_j \qquad \text{y también}$$

$$\hat{S}(L_1^{j+1})=\hat{S}(t_j+0) = \hat{S}(t_{j+1}) \quad j =1,..,k \quad i=1,\ldots\lambda_{j+1}$$

ya que $\hat{S}$ debe ser una función no creciente.

Con ésto la verosimilitud se reduce a

$$L = \prod_{j=1}^{k} \left[S(t_j)-S(t_j+0)\right]^{d_j} S(t_j)^{\lambda_j} S(t_{k+1})^{\lambda_{k+1}} .$$

Si escribimos $S(t_j+0) = P_j$ (j = 1,...,k) y definimos $P_0 = 1$, es necesario maximizar solamente:

$$L_1 = \prod_{j=1}^{k} (P_{j-1}-P_j)^{d_j} P_j^{\lambda_{j+1}}$$

la cual maximizaremos con respecto a $P_1,\ldots,P_k$ .

Dado que

$$p_j = P_j/P_{j-1} \text{ y } q_j = 1-p_j$$

tenemos

$$L_1 = \prod_{j=1}^{k} (p_1\cdots p_{j-1}q_j)^{d_j}(p_1\cdots p_j)^{\lambda_{j+1}}$$

Es conveniente, en este momento recordar que

$$n_j = \sum_{i=j}^{k} (\lambda_i + d_i)$$

y observar que si $\lambda_{k+1} > 0$ entonces la última observación, digamos $L^*$ es censurada, con lo que $\hat{S}(L^*) = \hat{S}(t_k+0)$ y $\hat{S}(t)$ queda indeterminada, por lo que consideraremos $\lambda_{k+1} = 0$. Ahora desarrollando el producto en $L_1$ tenemos :

$$\prod_{j=1}^{k} (p_1 \cdots p_{j-1} q_j)^{d_j} (p_1 \cdots p_j)^{\lambda_j} = \prod_{j=1}^{k} q_j^{d_j} (p_1 \cdots p_{j-1})^{d_j} (p_1 \cdots p_j)^{\lambda_j}$$

$$= \left( \prod_{j=1}^{k} q_j^{d_j} \right) p_0^{d_1} p_1^{\lambda_1} (p_0 p_1)^{d_2} (p_1 p_2)^{\lambda_2} (p_0 p_1 p_2)^{d_3} (p_1 p_2 p_3)^{\lambda_3} \cdots$$

$$(p_0 p_1 \cdots p_{k-1})^{d_k} (p_1 p_2 \cdots p_k)^{\lambda_k}$$

$$= \left( \prod_{j=1}^{k} q_j^{d_j} \right) p_1^{\lambda_1+d_2+\lambda_2+d_3+\lambda_3+\cdots+d_k+\lambda_k} p_2^{\lambda_2+d_3+\lambda_3+d_4+\lambda_4+\cdots+d_k+\lambda_k}$$

$$\cdots p_{k-1}^{\lambda_{k-1}+d_k+\lambda_k} p_k^{\lambda_k}$$

$$= \left( \prod_{j=1}^{k} q_j^{d_j} \right) p_1^{n_1-d_1} p_2^{n_2-d_2} \cdots p_{k-1}^{n_{k-1}-d_{k-1}} p_k^{n_k-d_k}$$

$$= \prod_{j=1}^{k} q_j^{d_j} p_j^{n_j-d_j}$$

Tomando log $L_1$ y maximizando para $p_j$ fijo y arbitrario obtenemos que el estimador máximo-verosímil es:

$$\hat{p}_j = \frac{n_j - d_j}{n_j}$$

y como

$$\hat{S}(t_j+0) = \hat{S}(t_{j-1}+0)\left(\frac{n_j - d_j}{n_j}\right)$$

con $n_1=n$ por lo que entonces el estimador Kaplan-Meier es:

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

Notemos que si no hay censura $n_1 = n$ y $n_j = n_{j-1} - d_{j-1}$ $j= 2 ,\ldots, k$ el estimador K-M se reduce a la F.S.E. . En ambos casos, con y sin censura, $\hat{S}(t)$ es una función escalonada que es igual a 1 en t=0 y decrece con saltos después de cada tiempo de falla $t_j$. y con una magnitud que depende del número de individuos aún vivos y el patrón de censura observado.

Deduciremos la expresión para $\hat{Var}(\hat{S}(t))$ bajo un modelo de censura aleatorio. Para un proceso con censura aleatoria, G(t) es la función de supervivencia asociada a los tiempos de censura.

Sea $t^*<\infty$ tal que $S(t^*)>0$ y supongamos que S(x) y G(x) son funciones continuas, entonces la función aleatoria

$$\sqrt{n}\left(\hat{S}(x)-S(x)\right)$$

con $0 < x < t^*$ converge débilmente a un proceso Gaussiano con media cero y función de covarianza: (Lawless 1982)

$$S(x)S(y)\int_0^x \frac{|d\, S(u)|}{S^2(u)\, G(u)} \qquad 0 < x < y < t^*$$

Además para muestras grandes la varianza de $\hat{S}(t)$ se puede aproximar por:

$$\mathrm{Var}\left[\hat{S}(t)\right] \cong S(t)^2 \int_0^t \left( \frac{|dS(u)|}{S(u)\ S^{\circ}(u)} \right)$$

en donde $S^{\circ}(u) = S(u)G(u)$. Para estimar esta varianza debemos estimar $S(x)$ y $S^{\circ}(x)$ por sus correspondientes estimadores. Una manera adecuada puede ser remplazar $S(x)$ por el estimador K-M $\hat{S}(x)$, y $S^{\circ}(x)$ por $\hat{S}^{\circ}(x)=n_x/n$ donde $n_x$ es el número de individuos en riesgo al tiempo x, ya que $S^{\circ}(x)$ es la proporción esperada de individuos en riesgo al tiempo x.

Por lo que se tendrá el siguiente estimador de la varianza

$$\hat{\mathrm{Var}}\left[\ \hat{S}(t)\ \right] \cong \hat{S}(t)^2 \sum_{t_j<t} \frac{\hat{S}_{j-1}-\hat{S}_j}{\hat{S}_j(n_j/n)}$$

$$\text{con } \hat{S}_j = \prod_{i=1}^{j} \left( \frac{n_i-d_i}{n_i} \right)$$

entonces:

$$\hat{S}_{j-1}-\hat{S}_j = \prod_{i=1}^{j-1} \frac{n_i-d_i}{n_i} \left(1 - \frac{n_j-d_j}{n_j}\right) = \prod_{i=1}^{j-1} \frac{n_i-d_i}{n_i} \frac{d_j}{n_j}$$

$$\hat{S}_j(n_j/n) = \prod_{i=1}^{j} \frac{n_i-d_i}{n_i} (n_j/n) = \prod_{i=1}^{j-1} \frac{n_i-d_i}{n_i} \left( \frac{n_j-d_j}{n_j} \right) (n_j/n)$$

por lo que :

$$\frac{\hat{S}_{j-1}-\hat{S}_j}{\hat{S}_j\ (n_j/n)} = \frac{\prod_{i=1}^{j-1} (n_i-d_i)/n_i \quad (d_j/n_j)}{\prod_{i=1}^{j-1} (n_i-d_i)/n_i\,[(n_j-d_j)/n_j]n_j/n} = \frac{nd_j}{n_j(n_j-d_j)}$$

y por lo tanto

$$\hat{Var}\left[\hat{S}(t)\right] \cong \frac{\hat{S}(t)^2}{n} \sum_{j:t_j<t} \frac{nd_j}{n_j(n_j-d_j)} = \hat{S}(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}$$

que es el estimador usual para la $Var\left[\hat{S}(t)\right]$.

Observemos que en ausencia de censura, este estimador se reduce a

$$\frac{\hat{S}(t)[1-\hat{S}(t)]}{n}$$

que es el estimador común de la varianza, ya que en este caso

$$\frac{n_j-d_j}{n_j} = \frac{n_{j-1}-d_{j-1}-d_j}{n_j} = \ldots = \frac{n_1-d_1-d_2-\ldots-d_{j-1}-d_j}{n_j} = \frac{n_{j+1}}{n_j}$$

por lo que

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j-d_j}{n_j} = \prod_{j:t_j<t} \frac{n_{j+1}}{n_j} = \frac{n_{j+1}}{n} .$$

Por otro lado y como $n_{j+1}=n_j-d_j$ y $d_j = n_j-n_{j+1}$, entonces la expresión

$$\sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}$$

se transforma en:

$$\sum_{j:t_j<t} \frac{n_j-n_{j+1}}{n_j(n_{j+1})} = \sum_{j:t_j<t} \frac{1}{n_{j+1}} - \frac{1}{n_j} = \frac{1}{n_{j+1}} - \frac{1}{n_1}$$

como $n_1=n$, tenemos que

$$\hat{S}(t)^2 \sum_{j;\,t_j<t} \frac{d_j}{n_j(n_j-d_j)} = \left(n_{j+1}/n\right)^2 \left(1/n_{j+1}-1/n\right)$$

$$= n_{j+1}/n \left(\frac{1}{n} - \frac{n_{j+1}}{n^2}\right)$$

$$= \frac{\hat{S}(t)\,[1-\hat{S}(t)]}{n}$$

Finalmente el estimador K-M $\hat{S}(t)$ es un estimador consistente de $S(t)$, bajo supuestos adecuados sobre la censura y $\hat{Var}\ \hat{S}(t)$ es un estimador asintótico de la varianza de $S(t)$. (Lawless 1982)

Una característica adicional de este estimador es que puede servir como base para proponer una familia paramétrica para la variable T que nos permita un mejor análisis de su comportamiento. Esto se muestra para las siguientes familias.

Modelo exponencial. Para este modelo $S(t)= \exp\{-\lambda(t)\}$, entonces $\log S(t) = -\lambda t$. Así que empíricamente tendremos evidencia para ajustar una familia exponencial a los datos observados si al graficar el logaritmo natural del estimador Kaplan-Meier de $S(t)$ contra $t$ obtenemos aproximadamente una recta con pendiente negativa y que pasa por el origen.

Modelo Weibull. Si tenemos $\hat{S}(t)$ el estimador K-M de $S(t)$, entonces podemos comprobar de manera empírica que el modelo Weibull es adecuado graficando $\log(-\log S(t))$ contra $\log t$. Si el modelo es correcto, obtendremos aproximadamente una linea recta cuya pendiente es un estimador de $p$ y la ordenada al origen un estimador de $p \log \lambda$.

Modelo Log-normal. Para ajustar este modelo en algunos datos observados, como

$$1-S(t) = \Phi\left(\frac{\log t-\mu}{\sigma}\right)$$

debemos graficar $1-\hat{S}(t)$ contra logt en papel normal y si obtenemos una recta con pendiente positiva ( que estima a $1/\sigma$), y ordenada al origen (que estima a $-\mu/\sigma$) ajustaremos a los datos un modelo log-normal.

# CAPITULO 4

## INFERENCIA PARAMETRICA PARA LA DISTRIBUCION DE TIEMPO DE FALLA

Nos interesa conocer la forma de la función de verosimilitud para los diferentes casos de censura ( Tipo I, Tipo II, aleatoria) y posteriormente, con base en dicha función de verosimilitud, realizar inferencias sobre los parámetros asociados a las distribuciones

Iniciaremos con la censura de tipo II, ya que su desarrollo es relativamente fácil.

Supongamos que solamente podemos realizar las primeras r observaciones $T_{(1)} < T_{(2)} <...< T_{(r)}$ de una muestra de tamaño n. T tiene función de densidad de probabilidad f(t) y de supervivencia S(t). Entonces la función de densidad conjunta de $T_{(1)}$, $T_{(2)}$,..., $T_{(r)}$ es:

$$\frac{n!}{(n-r)!}\prod_{i=1}^{r} f(t_i)\left[S(t_r)\right]^{n-r}$$

en donde

$\prod_{i=1}^{r} f(t_i)$ es la parte correspondiente a las r fallas observadas, y

$\left[S(t_r)\right]^{n-r}$ constituye la aportación de las observaciones

censuradas después de la r-ésima falla observada

Si definimos la función indicadora

$$\delta_i = \begin{cases} 1 \text{ si } T_i \leq T_{(r)} \\ 0 \text{ si } T_i \geq T_{(r)} \end{cases}$$

entonces la función de verosimilitud para este tipo de muestra es proporcional a:

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} \left[S(t_i)\right]^{1-\delta_i}$$

Observese que en este caso $\Sigma\delta_i = r$ es el número de tiempos de vida observados.

Ahora consideramos la función de verosimilitud para muestras con censura de tipo I.

Supongamos que tenemos un muestreo aleatorio de n individuos con tiempo de vida $T_1, \ldots, T_n$ y que está asociado a cada individuo un tiempo fijo de censura $L_i > 0$ . Observaremos a $T_i$ solamente si $T_i \leq L_i$, por lo que los datos son parejas,

$(t_i, \delta_i)$ $i=1,\ldots,n$ donde $t_i = \min(T_i, L_i)$ y

$$\delta_i = \begin{cases} 1 \text{ si } t_i = T_i \\ 0 \text{ si } t_i = L_i \end{cases}$$

por lo que la forma de la verosimilitud para datos con este tipo de censura es:

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} \left[S(t_i)\right]^{1-\delta_i}$$

nuevamente $\Sigma\delta_i$ representa el total de los tiempos de vida observados.

Hay que notar que la forma de la verosimilitud es idéntica a la del caso con censura tipo II .

Por último, encontremos la función de verosimilitud para la censura aleatoria.

En este caso supondremos que para cada individuo tenemos $T_i$ y $L_i$ variables aleatorias independientes, con funciones de supervivencia S(t) y G(t) y funciones de densidad f(t) y g(t) respectivamente. Sean $(T_i,L_i)$, i=1,...,n independientes, la variable que observamos es $t_i=\min(T_i,L_i)$ y definimos

$$\delta_i = \begin{cases} 1 \text{ si } T_i \leq L_i \\ 0 \text{ si } T_i \geq L_i \end{cases}$$

por lo que las observaciones son las parejas

$$(t_i, \delta_i), \ i=1,\ldots, n.$$

Nótese que para i fijo

$$f(t_i = t,\delta_i = 0) = -\partial/\partial t \ P(T_i \leq t, T_i > L_i) = g(t)S(t)$$

$$f(t_i = t,\delta_i = 1) = -\partial/\partial t \ P(T_i \geq t, T_i \leq L_i) = f(t)G(t)$$

o equivalentemente,

$$f(t_i=t,\delta_i) = \left[f(t)G(t)\right]^{\delta_i} \left[g(t)S(t)\right]^{\delta_i}$$

La función de densidad de probabilidad para la muestra es entonces

$$\prod_{i=1}^{n} G(t_i)^{\delta_i} g(t_i)^{1-\delta_i} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Como L es independiente de T, G(t) y g(t) no involucran a los

parámetros de interés, entonces la función de verosimilitud es proporcional a:

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} \left[ S(t_i) \right]^{1-\delta_i}$$

Como antes, $\Sigma\delta_i$ representa el número total de tiempos de vida observados. Esta forma es la misma que la de los casos anteriores.

En suma, para cualquier mecanismo de censura muestral, la función de verosimilitud es esencialmente la misma, por lo que los procesos estadísticos basados en ella son válidos en cualquier caso.

### 4.1 Estimación para la Distribución Exponencial

La distribución exponencial ocupa un lugar muy importante en el trabajo con distribuciones de tiempo de vida.

La distribución exponencial fué el primer modelo de tiempo de vida desarrollado de manera extensa con métodos estadísticos. Muchos trabajos realizados arrojan un gran número de resultados y popularizan a la exponencial como modelo de distribución de tiempo de vida, especialmente en el área industrial de confiabilidad.

Como ya vimos anteriormente la función de verosimilitud para cualquier tipo de censura es de la forma:

$$\prod_{i=1}^{n} f(t_i)^{\delta_i} \left[ S(t_i) \right]^{1-\delta_i}$$

$$\text{con } \delta_i = \begin{cases} 1 \text{ si } T_i \leq L_i \\ 0 \text{ si } T_i > L_i \end{cases}$$

Para el caso particular de la distribución exponencial

$$f(t_i) = \lambda \exp\{-\lambda t_i\} \quad y \quad S(t) = \exp\{-\lambda t_i\}$$

por lo que la función de verosimilitud para $\lambda$ es:

$$L(\lambda) = \prod_{i=1}^{n} \left\{\lambda \exp\{-\lambda t_i\}\right\}^{\delta_i} \left\{\exp\{-\lambda t_i\}\right\}^{1-\delta_i}$$

Ahora para obtener el estimador de máxima verosimilitud para $\lambda$ tenemos

$$\log L(\lambda) = \Sigma \delta_i \log\lambda - \lambda \Sigma t_i \delta_i - \lambda \Sigma t_i (1-\delta_i)$$

en donde la suma corre sobre i=1,..,n

con primer derivada

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = \Sigma \delta_i / \lambda - \Sigma t_i \delta_i - \Sigma t_i (1-\delta_i)$$

Igualando a cero y resolviendo para $\lambda$ obtenemos:

$$\hat{\lambda} = \frac{\Sigma \delta_i}{\Sigma t_i}$$

Note que para el caso sin censura $\Sigma \delta_i = n$ y $\hat{\lambda}$ resulta ser el recíproco de la media muestral.

Para realizar pruebas de hipótesis y construir intervalos de confianza, usaremos la siguiente aproximación asintótica.

$$\frac{\hat{\lambda} - \lambda}{\sqrt{(I(\hat{\lambda}))}} \sim N(0,1)$$

con $I(\lambda)$ la información esperada de Fisher dada por

$$-E\left( \frac{\partial^2 \log L(\lambda)}{\partial \lambda^2} \right) .$$

Ejemplificaremos la construcción de $I(\lambda)$ con un esquema de censura aleatoria. Para ésto condicionaremos la censura a un valor observado, i.e., $L_i = c_i$.

La contribución del i-ésimo individuo a la verosimilitud es:

$$L_i(\lambda) = \left(\lambda \exp\{-\lambda t_i\}\right)^{\delta_i} \left(\exp\{-\lambda t_i\}\right)^{1-\delta_i}$$

cuya función de puntaje es:

$$U_i(\lambda) = \frac{\partial \log L_i(\lambda)}{\partial \lambda} = \frac{\delta_i}{\lambda} - t_i$$

además

$$E(U_i(\lambda)) = \frac{E(\delta_i)}{\lambda} - E(t_i) = 0, \text{ y}$$

$$\frac{-\partial^2 \log L_i(\lambda)}{\partial \lambda} = \frac{\delta_i}{\lambda^2}$$

entonces

$$E\left( \frac{-\partial^2 \log L_i(\lambda)}{\partial \lambda} \right) = \frac{E(\delta_i)}{\lambda^2}$$

pero

$$E(\delta_i) = 0 \; P_r(\delta_i = 0) + 1 \cdot P_r(\delta_i = 1)$$

$$= \int_0^{c_i} \lambda \exp\{-\lambda t_i\} dt_i = 1 - \exp\{-\lambda c_i\}$$

$$\text{entonces } E\left(\frac{-\partial^2 \log L_1(\lambda)}{\partial\lambda^2}\right) = \frac{1-\exp\{-\lambda c_1\}}{\lambda^2}$$

y finalmente

$$I(\lambda) = \sum_{i=1}^{n} \frac{1-\exp\{-\lambda c_1\}}{\lambda^2}$$

Como $I(\lambda)$ involucra todos los tiempos potenciales de censura para todos los individuos, tiempos que generalmente se desconocen, usaremos en su lugar:

$$I_0(\hat{\lambda}) = \left(\frac{-\partial^2 \log L_1(\lambda)}{\partial\lambda^2}\right)_{\hat{\lambda}}$$

que es la información observada.

## 4.2 Estimación para la Distribución Weibull

La distribución Weibull es particularmente importante como distribución de tiempo de vida. Una gran cantidad de métodos estadísticos han sido desarrollados para esta distribución. Una razón para este desarrollo la constituyen las propiedades estadísticas de la distribución Weibull. Como no hay, en general, dos estadísticas suficientes para los parámetros p y $\lambda$, existen grandes posibilidades de producir estimadores por diversos medios; así se ha llevado a cabo un trabajo extensivo en la producción de tablas para hacer inferencias y en el desarrollo de aproximaciones para distribuciones de cierto tipo de parámetros. Por fortuna, se cuenta con procesos estadísticos que son fáciles de realizar gracias a la existencia de computadoras muy rápidas.

Aquí desarrollaremos los principales procesos estadísticos de estimación para la distribución Weibull, cuya función de densidad de probabilidad es:

$$f(t) = p\lambda(\lambda t)^{p-1}\exp\{-(\lambda t)^{p}\}$$

Algunas veces en lugar de la distribución Weibull, es conveniente trabajar con la distribución equivalente del valor extremo, con función de densidad de probabilidad

$$f(x) = 1/b \exp\{(x-u)/b\} \exp\{e^{-(x-u)/b}\}$$

en la que $u\in R$ y $b>0$. Esta expresión se deriva de la anterior si hacemos el cambio de variable $x=\log t$, con $u= -\log\lambda$ y $b=p^{-1}$. La principal conveniencia de este cambio es que u y b son parámetros de localización y escala para esta nueva distribución. Todos los desarrollos se harán con la distribución del valor extremo.

Como siempre, la función de verosimilitud para cualquier tipo de censura es de la forma:

$$\prod_{i=1}^{n} f(t_i)^{\delta_i}\left[S(t_i)\right]^{1-\delta_i} \qquad \text{con } \delta_i = \begin{cases} 1 \text{ si } T_i \le L_i \\ 0 \text{ si } T_i > L_i \end{cases}$$

En este caso la verosimilitud toma la forma:

$$L(u,b) = \prod_{i=1}^{n}\left(1/b \exp\{(x_i-u)/b - \exp(x_i-u)/b\}\right)^{\delta_i}\left(\exp\{(y_i-u)/b\}\right)^{1-\delta_i}$$

en donde $x_i= \log t_i$ y $y_i= \log L_i$, entonces

$$\log L(u,b) = \sum_{i=1}^{n}\delta_i \log b + \sum_{i\in D}(x_i-u)/b - \sum_{i\in D}\exp\{(x_i-u)/b\}$$

$$- \sum_{i\in C}\exp\{(x_i-u)/b\}$$

con D es el conjunto de individuos en los que ocurrió la falla y C es el conjunto de individuos censurados. Por lo que

$$\log L(u,b) = \sum_{i=1}^{n} \delta_i \log b + \sum_{i \in D} (x_i - u)/b - \sum_{i=1}^{n} \exp\{(x_i - u)/b\} .$$

Las primeras derivadas de log L(u,b) con respecto a u y b son:

$$\frac{\partial \log L}{\partial u} = \frac{-\Sigma\delta_i}{b} + 1/b \sum_{i=1}^{n} \exp\{(x_i - u)/b\}$$

$$\frac{\partial \log L}{\partial b} = \frac{\Sigma\delta_i}{b} \; 1/b \sum_{i=1}^{n} (x_i - u)/b\} + 1/b \sum_{i=1}^{n} \exp\{(x_i - u)/b\}(x_i - u)/b$$

para resolver las ecuaciones de verosimilitud, primero haremos

$$\frac{\partial \log L}{\partial u} = 0$$

con lo que obtenemos

$$\exp\{\hat{u}\} = \left[1/\Sigma\delta_i \sum_{i=1}^{n} \exp\{x_i/\hat{b}\}\right]^{\hat{b}}$$

y sustituyendo en $\partial \log/\partial b = 0$, tenemos:

$$\frac{\sum_{i=1}^{n} \exp\{x_i/\hat{b}\} x_i}{\sum_{i=1}^{n} \exp\{x_i/\hat{b}\} - \hat{b} - 1/\Sigma\delta_i \sum_{i \in D} x_i} = 0$$

esta última ecuación puede ser resuelta por métodos iterativos para encontrar $\hat{b}$ y después sustituir y encontrar $\hat{u}$. Las expresiones correspondientes para la Weibull se encuentran recordando que $u = -\log\lambda$ y $b = p^{-1}$.

Ahora ilustraremos el cálculo de la información observada de Fisher a partir de una muestra con esquema de censura aleatoria. Para ésto condicionaremos el desarrollo a un valor observado de la censura $L_i=c_i=(y_i-u)/b$.

Si definimos $(x_i-u)/b = z_i$, la contribución de una sola observación al logaritmo de la verosimilitud se puede escribir como:

$$\log L_i = \delta_i( z_i-\log b-\exp z_i) + (1-\delta_i)(-\exp c_i)$$

por lo que las primeras derivadas con respecto a u y b toman la forma

$$\frac{\partial \log L_i}{\partial u} = \delta_i(-1/b + 1/b \exp z_i) + (1-\delta_i)(1/b \exp c_i)$$

$$\frac{\partial \log L_i}{\partial b} = \delta_i(-1/b-z_i/b + z_i/b \exp z_i)+(1-\delta_i)(c_i/b \exp c_i)$$

$$\frac{\partial^2 \log L_i}{\partial u^2} = \delta_i(-1/b^2\exp z_i) - (1-\delta_i)(1/b^2\exp c_i)$$

$$\frac{\partial^2 \log L_i}{\partial b^2} = \delta_i(1/b^2 + 2z_i/b - 2z_i/b^2\exp z_i - z_i/b^2\exp z_i) + (1-\delta_i)(-1/b^2\exp c_i - c_i/b^2\exp c_i)$$

$$\frac{\partial^2 \log L_i}{\partial u\,\partial b} = \delta_i(+1/b^2 - 1/b^2\exp z_i - z_i/b^2\exp z_i) (1-\delta_i) (-1/b^2\exp c_i - c_i/b^2\exp c_i)$$

Observemos que dado $\delta_i=1$, $z_i$ tiene una distribución estándar del

valor extremo truncada en $c_i$, con función de densidad de probabilidad

$$f(z_i) = \frac{e^{z_i}\exp(-e^{z_i})}{1-\exp(e^{c_i})} \qquad -\infty < z_i \leq c_i$$

ya que $f(z_i|\delta_i=1) = \dfrac{f(z_i)}{1-S(L_i)}$ .

Además, notemos que $p(\delta_i=1) = 1-\exp(-e^{c_i}) = 1-p(\delta_i=0)$, con lo que obtenemos las siguientes esperanzas

$$I_{uu,i} = E\left(-\frac{\partial^2 \log L_i}{\partial u^2}\right) = 1/b^2\left(\int_0^{c_i} \exp z_i \frac{e^{z_i}\exp(-e^{z_i})}{1-\exp(-e^{c_i})}\, 1-\exp(-e^{c_i})dz + e^{c_i}\exp(-e^{c_i})\right) = 1/b^2(\exp(-e^{c_i}) + e^{c_i}\exp(-e^{c_i}))\ .$$

De manera análoga obtenemos

$$I_{bb,i} = E\left(-\frac{\partial^2 \log L_i}{\partial b^2}\right) = 1/b^2\left(\int_0^{c_i}(1+z_i^2\exp z_i)e^{z_i}\exp(-e^{z_i})dz + c_i^2 e^{c_i}\exp(-e^{c_i})\right)$$

$$I_{ub,i} = E\left(-\frac{\partial^2 \log L_i}{\partial u\, \partial b}\right) = 1/b^2\left(\int_0^{c_i} z_i \exp(2z_i)\exp(-e^{z_i})\, dz + c_i^2 e^{c_i}\exp(-e^{c_i})\right)$$

Las integrales no resueltas se pueden calcular numéricanente. Si sumamos estas expresiones sobre todos los individuos tendremos que la información esperada de Fisher es:

$$I(u,b) = \begin{pmatrix} \sum_{i=1}^{n} I_{uu,i} & \sum_{i=1}^{n} I_{ub,i} \\ \sum_{i=1}^{n} I_{ub,i} & \sum_{i=1}^{n} I_{bb,i} \end{pmatrix} .$$

En la práctica, usualmente estimamos $I(u,b)$ por $I(u,b)\big|_{(\hat{u},\hat{b})}$. Además, $(\hat{u},\hat{b})$ es asintóticamente normal bivariada con media $(u,b)$ y matriz de covarianza $\left[I(u,b)\right]^{-1}$. Intervalos de confianza y pruebas de hipótesis se pueden realizar a partir de este hecho.

### 4.3 Estimación para la Distribución Log-Normal

Para la distribución log-normal el logaritmo de los tiempos de falla se distribuye normal, por lo que en ausencia de censura, los resultados de estimación son ampliamente conocidos.

Cuando tenemos datos censurados la estimación por máxima verosimilitud es un poco más complicada. Denotamos por $\Phi(z)$ y $Q(z)$ las correspondientes funciones de densidad y supervivencia de la normal estándar, es decir

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \quad \text{y} \quad Q(z) = \int_z^{\infty} \Phi(x)dx .$$

Ahora , si $Y \sim N(\mu,\sigma^2)$ entonces las funciones de densidad de probabilidad y supervivencia de Y son:

$$f(y) = 1/\sigma \; \Phi\left[(y-\mu)/\sigma\right], \; S(y) = Q\left[(y-\mu)/\sigma\right]$$

y la función de verosimilitud toma la forma:

$$\prod_{i \in D} 1/\sigma \; \Phi((y_i-\mu)/\sigma) \; \prod_{i \in C} Q((y_i-\mu)/\sigma) \; .$$

D es el conjunto de individuos que registran la falla (cuyo número denotaremos como r) y C el de los individuos censurados. Entonces el logaritmo de verosimilitud es:

$$\log\ (\mu,\sigma) = -r \log \sigma - 1/2\sigma^2 \sum_{i \in D} (y_i-\mu)^2 + \sum_{i \in C} \log Q((y_i-\mu)/\sigma) \; .$$

Las primeras derivadas de log L son:

$$\frac{\partial \log L}{\partial \mu} = 1/\sigma^2 \sum_{i \in D} (y_i-\mu) + 1/\sigma \sum_{i \in C} \Phi((y_i-\mu)/\sigma)/Q((y_i-\mu)/\sigma)$$

$$\frac{\partial \log L}{\partial \sigma} = -r/\sigma + 1/\sigma^3 \sum_{i \in D} (y_i-\mu)^2 + 1/\sigma \sum_{i \in C} [(y_i-\mu)/\sigma \Phi((y_i-\mu)/\sigma)]/Q((y_i-\mu)/\sigma)$$

igualando a cero estas dos derivadas obtenemos las ecuaciones de máxima verosimilitud, que podemos resolver por métodos numéricos (por ejemplo: algoritmo EM, Newton Raphson). Para simplificar las expresiones de las segundas derivadas haremos el cambio de notación $z_i=(y_i-\mu)/\sigma$ , con este cambio dichas derivadas son:

$$\frac{\partial^2 \log L}{\partial \mu^2} = -r/\sigma^2 - 1/\sigma^2 \sum_{i \in C} \frac{\Phi(z_i)}{Q(z_i)} \left( \frac{\Phi(z_i)}{Q(z_i)} - z_i \right)$$

$$\frac{\partial^2 \log L}{\partial \sigma^2} = -r/\sigma^2 - 3/\sigma^2 \sum_{i \in D} z_i^2 - 2/\sigma^2 \sum_{i \in C} z_i \frac{\Phi(z_i)}{Q(z_i)}$$

$$-1/\sigma^2 \sum_{i \in C} z_i^2 \frac{\Phi(z_i)}{Q(z_i)} \left( \frac{\Phi(z_i)}{Q(z_i)} - z_i \right)$$

$$\frac{\partial^2 \log L}{\partial\mu\ \partial\sigma} = -2/\sigma^2 \sum_{i\in D} z_i - 1/\sigma^2 \sum_{i\in C} \frac{\Phi(z_i)}{Q(z_i)}$$

$$-1/\sigma^2 \sum_{i\in C} \frac{\Phi(z_i)}{Q(z_i)} \left( \frac{\Phi(z_i)}{Q(z_i)} - z_i \right) .$$

El cálculo de las esperanzas de estas expresiones es demasiado complicado por lo que, en lugar de trabajar con la matriz de información esperada de Fisher, se trabaja con la matriz de información observada, es decir

$$I_0 = \begin{bmatrix} -\partial^2 \log L/\partial\mu^2 & -\partial^2 \log L/\partial\mu\ \partial\sigma \\ -\partial^2 \log L/\partial\mu\ \partial\sigma & -\partial^2 \log L/\partial\sigma^2 \end{bmatrix}_{(\hat{\mu},\hat{\sigma})}$$

Las pruebas de hipótesis e intervalos de confianza se realizan basados en el hecho de que asintóticamente

$$(\hat{\mu},\hat{\sigma}) \sim N_2\left((\mu,\sigma), I_0^{-1}\right) .$$

### 4.4 Estimación para la Distribución Gamma

Con datos no censurados, algunos procesos de inferencia para la distribución gamma son simples y se discuten en un gran número de libros de estadística. Sin embargo, cuando los datos están censurados, o cuando es necesario obtener intervalos para cantidades estimables como cuantiles o la función de supervivencia, realizar estos procesos resulta más complicado. Esta es una razón por la que la distribución gamma tiene menor uso como distribución de tiempo de falla, que las distribuciones Weibull y Log-normal. No obstante, la gamma es un modelo usual, y se presentarán algunos procesos de inferencia para ella.

Suponga que en una muestra aleatoria de tamaño n se observan r tiempos de falla y n-r tiempos censurados. Ambos tiempos, de censura y falla, serán denotados por $t_i$, con D y C los conjuntos de individuos que presentaron la falla y que resultaron censurados, respectivamente.

Como ya hemos visto, la función gamma tiene función de densidad y supervivencia

$$f(t) = \frac{\lambda(\lambda t)^{k-1}\exp\{-\lambda t\}}{\Gamma(k)}$$

$$S(t) = Q(k,\lambda t) \quad \text{con}$$

$$Q(k,x) = 1-IG_k(x) \ .$$

Por lo que la función de verosimilitud para esta muestra es:

$$L = \left(\prod_{i\in D} f(t_i)\right)\left(\prod_{i\in C} S(t_i)\right)$$

$$= \left(\prod_{i\in D} \frac{\lambda(\lambda t_i)^{k-1}\exp\{-\lambda t_i\}}{\Gamma(k)}\right)\left(\prod_{i\in C} Q(k,\lambda t_i)\right) \ .$$

Si denotamos por

$$\bar{t} = \sum_{i\in D} \frac{t_i}{r} \quad \text{y} \quad \tilde{t} = \left(\prod_{i\in D} t_i\right)^{1/r}$$

las medias aritmética y geométrica de los tiempos de falla observados, podemos escribir el logaritmo de la función de verosimilitud como:

$$\log L = -r\log\Gamma(k) + r\log\lambda + (k-1)\sum_{i\in D}\log t_i$$

$$-\lambda \sum_{i \in D} t_i + \sum_{i \in C} \log\left[Q(k,\lambda t)\right]$$

$$= rk\log\lambda - r\log\Gamma(k) + r(k-1)\log\bar{t} - \lambda r\bar{t}$$

$$+ \sum_{i \in C} \log\left[Q(k,\lambda t)\right] .$$

Las primeras derivadas de log L con respecto a k y $\lambda$ son:

$$\frac{\partial \log L}{\partial k} = r\log\lambda - r\psi(k) + r\log\bar{t} + \sum_{i \in C} \frac{1}{Q(k,\lambda t)} \frac{\partial Q(k,\lambda t)}{\partial k}$$

$$\frac{\partial \log L}{\partial \lambda} = \frac{rk}{\lambda} - r\bar{t} + \sum_{i \in C} \frac{1}{Q(k,\lambda t)} \frac{\partial Q(k,\lambda t)}{\partial \lambda}$$

con $\psi(k) = \dfrac{d \log \Gamma(k)}{dk} = \dfrac{\Gamma'(K)}{\Gamma(k)}$ la función digamma, además

$$\frac{\partial Q(k,\lambda t)}{\partial k} = \frac{1}{\Gamma(k)} \int_{\lambda t}^{\infty} u^{k-1}(\log u) \exp\{-u\} du - \psi(k) Q(k,\lambda t)$$

y $$\frac{\partial Q(k,\lambda t)}{\partial \lambda} = \frac{\lambda(\lambda t)^{k-1}\exp\{-\lambda t\}}{\Gamma(k)} .$$

Las ecuaciones de máxima verosimilitud se obtienen igualando $\dfrac{\partial \log L}{\partial k}$ y $\dfrac{\partial \log L}{\partial \lambda}$ a cero, que se resuelven por métodos interativos. El cálculo de la segunda derivada de logL, que involucra la función trigamma, es aún más complicado, además de que su desarrollo resulta igualmente difícil. Para realizar pruebas de hipótesis sobre los paramétros de esta distribución se utiliza el cociente de verosimilitudes. Se podrían usar

aproximaciones asintóticas de las distribuciones de los estimadores máximo verosímiles, pero se tiene evidencia de que la distribución conjunta de $(\hat{k},\hat{\lambda})$ no es cercana a una normal bivariada.

# CAPITULO 5

# MODELOS CON COVARIABLES

## 5.1 INTRODUCCION

Hasta ahora hemos tratado con muestras de observaciones del tiempo de falla en poblaciones homogéneas. Sin embargo, en la práctica es común encontrar situaciones en las que se tienen poblaciones heterogéneas, esto es, observaciones no necesariamente idénticamente distribuidas.

Si la heterogeneidad de la población se debe a la presencia de variables explicativas como edad, sexo,tratamiento médico,etc., entonces es importante considerar la relación del tiempo de falla con estas variables explicativas llamadas comunmente covariables. Una manera de hacer ésto es através de modelos tipo regresión, en los que se especifica la dependencia entre el tiempo de falla y las covariables, por lo que supondremos que para cada individuo está definido un vector $\underline{Z}$ de p covariables. Los componentes de $\underline{Z}$ representan aspectos que pueden estar relacionados con el tiempo de falla, tales como:

1).- Tratamientos.
2).- Propiedades intrínsecas de los individuos.
3).- Variables exógenas.

Las covariables que miden las propiedades intrínsecas de los individuos incluyen por ejemplo, sexo, edad y variables que describen la historia clínica para la admisión del estudio.

Las variables exógenas definen, en particular, aspectos ambientales del problema y pueden ser necesarias para representar agrupaciones de los individuos.

Algunos ejemplos de situaciones en donde es necesario la presencia de covariables son:

-Para determinar los efectos de un tratamiento bajo estudio se divide a la población en dos grupos, el primero de ellos recibe el tratamiento, por lo que el valor de su covariable es $\underline{Z}=1$ y el segundo se tomará como grupo control $\underline{Z}=0$

-En un estudio con condensadores de vidrio, éstos fueron examinados operando bajo ciertos niveles de temperatura y voltaje. Se realizaron pruebas con ocho combinaciones diferentes de temperatura-voltaje; en cada combinación se pusieron a prueba varios condensadores y se observó su tiempo de falla. Las covariables en este caso son temperatura y voltaje.

-En un estudio clínico se recabaron datos sobre la supervivencia de 65 pacientes con mieloma múltiple, considerando su relación con otros factores. Se tomaron en cuenta 16 covariables en total incluyendo medidas fisiológicas como el aumento de hemoglobina en la sangre y el conteo de glóbulos blancos en diagnóstico, factores cualitativos como la presencia o ausencia de infecciones en diagnóstico y las características de cada individuo tales como edad y sexo. El estudio consistió en valorar cuáles de las covariables están fuertemente relacionadas con el tiempo de supervivencia.

-Prentice (1973), presenta algunos datos sobre las experiencias de supervivencia de un grupo de 40 pacientes con cáncer pulmonar avanzado, en este caso los tratamientos son un factor de importancia, un grupo de pacientes recibió un tratamiento y el resto otro. Otras covariables que se consideraron fueron, entre otras, la edad y la condición del paciente. El estudio consistió

en comparar el efecto de los tratamientos en el tiempo de supervivencia cuando se toma en cuenta las otras covariables.

Un problema inicial al que nos enfrentamos es modelar el cambio que ocurre en la función de riesgo básica al cambiar las condiciones estándar. Por este motivo, frecuentemente los modelos se pueden desarrollar en dos partes

1).- Un modelo para la población base, que es aquella en la que los individuos tienen asociado un vector de covariables $\underline{Z}=\underline{0}$, cuya función de riesgo $h_0(t)$ represente el riesgo básico. Este problema está resuelto, ya que los modelos para $h_0(t)$ son aquellos que se presentaron en los primeros capítulos.

2).- Una representación del cambio introducido en la función de riesgo básica por una covariable $\underline{Z} \neq \underline{0}$, $h(t|\underline{Z})$, frecuentemente en términos de alguna forma paramétrica. A este respecto se han propuesto dos grandes familias de modelos para $h(t|\underline{Z})$, el de riesgos proporcionales y el de vida acelerada.

## 5.2 Modelo de Riesgo Proporcionales

La primera familia de modelos que consideraremos para explicar la dependencia entre T (tiempo de falla) y $\underline{Z}$ (covariables) la constituyen los llamados modelos de riesgos proporcionales.

Para un vector constante de covariables $\underline{Z}$, supondremos que la función de riesgo básica, $h_0(t)$, cambia de manera proporcional al valor de cierta función que depende solamente de $\underline{Z}$, es decir:

$$h(t|\underline{Z}) = h_0(t)g(\underline{Z}) \ .$$

Con sus correspondientes funciones de supervivencia y densidad

$$S(t|\underline{Z}) = [S_0(t)]^{g(\underline{Z})}$$

$$f(t|\underline{Z}) = g(\underline{Z})[S_0(t)]^{g(\underline{Z})-1} f_0(t)$$

con $h_0$, $S_0$ y $f_0$ las funciones de riesgo, supervivencia y densidad básicas, respectivamente.

Ya que tanto $h_0(t)$ como $g(\underline{Z})$ son funciones arbitrarias, el modelo resulta demasiado general lo que complica el análisis, por esta razón supondremos que g es una función paramétrica lineal, esto es

$$g(\underline{Z}) = g(\underline{Z}'\underline{\beta})$$

con $\underline{\beta}$ un vector de parámetros de dimensión p. El modelo es entonces

$$h(t|\underline{Z}) = h_0(t)g(\underline{Z}'\underline{\beta}) \ .$$

La elección de $g(\cdot)$ se puede realizar a partir de los datos. Sin embargo, la elección de uso más común es $g(x) = \exp\{x\}$ ya que al evaluar la función de riesgo en $\underline{Z}=0$, regresamos a la función básica y es estrictamente positiva. Por lo que el modelo es:

$$h(t|\underline{Z}) = h_0(t)\exp(\underline{Z}'\underline{\beta}) \ .$$

Algunas razones para considerar el uso de este modelo

-El modelo tiene una interpretación simple en el sentido de que el efecto de una covariable (ejemplo, tratamiento) multiplica el riesgo por un factor constante.

-Se tiene evidencia empírica que sustenta el supuesto de riesgos proporcionales en distintos grupos de tratamientos.

-La censura y la ocurrencia de varios tipos de falla son relativamente fáciles de introducir en el modelo y, en particular, los problemas técnicos de inferencia estadística cuando $h_0(t)$ es

arbitraria, tienen una solución simple.

## 5.3 Modelo de Vida Acelerada

La segunda familia que usaremos para explicar la dependencia entre T y $\underline{Z}$ está formada por los llamados modelos de vida acelerada. La idea para estos modelos se introducirá a través de un ejemplo sencillo, con $\underline{Z}$ una covariable binaria.

Supongamos que en un proceso industrial dos máquinas serán sometidas a diferentes condiciones de trabajo, dichas condiciones estarán representadas por valores de una covariable Z . La primera máquina trabajará en condiciones estándar (Z=0), y la segunda lo hará a un ritmo diferente (Z=1). Pese a que el tiempo observado es el mismo en ambas máquinas, mientras que la primera tiene una escala de mediciones, las mediciones de la segunda están rescaladas por un valor constante g. Por lo que sus funciones de supervivencia están relacionadas de la siguiente manera

$$S(t) = S_0(gt)$$

y además se tiene

$$h(t) = gh_0(gt)$$

$$f(t) = gf_0(gt) \ .$$

Con $S_0$, $f_0$ y $h_0$ las correspondientes funciones de supervivencia, densidad y riesgo en condiciones estándar. El efecto de g es acelerar o desacelerar el proceso de vida de las máquinas. Una razón para ésto es el hecho de que un individuo que haya sobrevivido al tiempo t bajo Z = 0 debe sobrevivir al tiempo t/g, bajo Z = 1 i.e. las correspondientes variables están relacionadas por

$$T = \frac{T_0}{g} .$$

De manera más general, supongamos que existe una función $g(\underline{Z})$, de tal manera que ahora la relación entre las funciones de supervivencia, densidad y riesgo son

$$h(t|\underline{Z}) = h_0(g(\underline{Z})t)g(\underline{Z})$$

$$S(t|\underline{Z}) = S_0(g(\underline{Z})t)$$

$$f(t|\underline{Z}) = f_0(g(\underline{Z})t)g(\underline{Z})$$

$S_0$, $f_0$ y $h_0$ definidas como antes.

Ahora haremos la deducción del modelo. Sabemos que

$$f(t|\underline{Z}) = f_0(t\ g(\underline{Z}))g(\underline{Z}),$$

entonces

sea $Y = \log T$ ,por lo tanto $T = \exp\{Y\}$ y $\left|\frac{dT}{dY}\right| = \exp\{Y\}$

por lo que

$$f(y|\underline{Z}) = f(t(Y)|\underline{Z})\left|\frac{dT}{dY}\right| = f_0(e^y g(\underline{Z}))g(\underline{Z})e^y.$$

Ahora sea

$$Y_1 = Y - \mu_0 + \log g(\underline{Z}),$$

es decir

$$Y_1 = \log T - \mu_0 + \log g(\underline{Z}) \qquad \ldots(5)$$

de donde

$$\exp\{Y_1\} = \frac{T}{\exp\{\mu_0\}}\, g(\underline{Z}).$$

Despejando T tenemos

$$T = \frac{\exp\{Y_1+\mu_0\}}{g(\underline{Z})} \quad y \quad \left|\frac{dT}{dY_1}\right| = \frac{\exp\{Y_1+\mu_0\}}{g(\underline{Z})}$$

con lo que tenemos

$$f(Y_1 | \underline{Z}) = f(t(Y_1) | \underline{Z}) \left|\frac{dT}{dY_1}\right| = f_0\left( \frac{\exp\{Y_1+\mu_0\}}{g(\underline{Z})}\, g(\underline{Z})\right) g(\underline{Z})\, \frac{\exp\{Y_1+\mu_0\}}{g(\underline{Z})}$$

$$= f_0(\exp\{Y_1+ \mu_0\})\, \exp\{Y_1+ \mu_0\}$$

Observemos que $f(Y_1|\underline{Z})$ no depende de $\underline{Z}$ .Por semejanza con los modelos de regresión usuales, y dada la ecuación (5) desearíamos que $E(Y_1) = 0$ ya que toma el papel del error. Pero por otro lado

$$E(Y_1) = \int_{-\infty}^{\infty} Y_1 f_0(\exp\{Y_1+ \mu_0\}) \exp\{ Y_1+ \mu_0\} dY_1$$

Consideremos el cambio de variable $u = \exp\{Y_1+ \mu_0\}$, entonces

$$E(Y_1) = \int_{-\infty}^{\infty} (\log u - \mu_0) f_0(u) du$$

$$= \int_{-\infty}^{\infty} \log u\, f_0(u) du - \mu_0 \int_{-\infty}^{\infty} f_0(u) du$$

$$= E(\log T | \underline{Z}=0) - \mu_0$$

Para que $E(Y_1)=0$ debemos tomar $\mu_0$ tal que

$$E(\log T \mid \underline{Z} = 0) = \mu_0 .$$

Si definimos $\varepsilon = Y_1$, obtenemos

$$\varepsilon = \log T - \mu_0 + \log g(\underline{Z})$$

o equivalentemente,

$$\log T = \mu_0 - \log g(\underline{Z}) + \varepsilon$$

con $\varepsilon$ independiente de $\underline{Z}$ y $E(\varepsilon) = 0$. La contribución de $\underline{Z}$ es como en el caso anterior. Es decir $g(\underline{Z}\underline{\beta}) \geq 0$ y $g(\underline{0}) = 1$ , por lo que un candidato natural es $g(\underline{Z}\underline{\beta}) = \exp\{\underline{Z}'\underline{\beta}\}$ con $\underline{\beta}$ vector de p parámetros. Por lo que finalmente tenemos como el modelo de vida acelerada más usual el dado por la relación

$$\log T = \mu_0 - \underline{\beta}'\underline{Z} + \varepsilon$$

## 5.4 Estimación y Pruebas de Hipótesis para los Modelos de Riesgos Proporcionales

Si $h_0(t)$ se define de manera totalmente paramétrica, podemos construir una función de verosimilitud a partir de $f(t|\underline{Z})$, $S(t|\underline{Z})$ y utilizar los métodos descritos en el capítulo anterior.

El caso general supone $h_0(t)$ desconocida por lo cual es necesario considerar otro tipo de funciones que nos permitan dar estimadores y hacer pruebas de hipótesis para analizar el comportamiento del tiempo de falla. Cox (1975) propuso una extensión de la función de verosimilitud a la función de verosimilitud parcial, la cual está dada por:

Sea Y un vector con f.d.p. $f(Y;\theta)$. Suponga que Y se puede transformar en $(X_1,S_1,X_2,S_2,\ldots,X_m,S_m)$ donde los componentes pueden ser vectores.

Sean $X^{(j)}=(X_1,\ldots,X_j)$ y $S^{(j)}=(S_{(1)},\ldots,S_{(j)})$, $j=1,..,m$. Denotamos $f_1^{(j)}$ la densidad condicional de $X_j$ dado $X^{(j-1)}$ y $S^{(j-1)}$ y $f_2^{(j)}$ la densidad condicional de $S_j$ dado $X^{(j)}$ y $S^{(j-1)}$entonces, la verosimilitud total de esta sucesión se puede escribir como

$$\prod_{j=1}^{m} f_1^{(j)}\,(X_j|X^{(j-1)},S^{(j-1)};\theta)\;\prod_{j=1}^{m} f_2^{(j)}(S_j|X^{(j)},S^{(j-1)};\theta)\;.$$

LLamaremos al segundo producto la verosimilitud parcial para $\theta$ basada en S de la sucesión $\{X_j,S_j\}$.

En nuestro caso, supongamos que en una muestra aleatoria de n individuos, se observaron k tiempos de falla distintos y n-k tiempos de censura. Los k tiempos de falla observados, podemos denotarlos como $t_{(1)}< t_{(2)}<\ldots<t_{(k)}$ y definimos $R_i$ el conjunto de individuos en riesgo al tiempo $t_{(i)}$, es decir los individuos vivos y no censurados justo antes de la i-ésima falla. Cox sugiere la siguiente "función de verosimilitud" para estimar $\underline{\beta}$ sin necesidad de conocer $h_0(t)$

$$L(\underline{\beta}) = \prod_{i=1}^{k} \frac{\exp(Z_i'\underline{\beta})}{\sum_{j\in R_i}\exp(Z_j'\underline{\beta})}$$

donde $Z_i$ es el vector de covariables asociado con el individuo muerto en $t_{(i)}$. En este contexto, la motivación para trabajar con $L(\underline{\beta})$ es que, dado $R_i$, la probabilidad de que el individuo i-ésimo muera al tiempo $t_{(i)}$ es:

$$\frac{h(t_{(i)}|Z_i)}{\sum_{j\in R_i} h(t_{(j)}|Z_j)} = \frac{\exp(Z_i'\beta)}{\sum_{j\in R_i} \exp(Z_j'\beta)}$$

La razón intuitiva para que $h_0(t)$ no aparezca en esta expresión es el hecho de que el conocimiento de $h_0(t)$ no aporta más información sobre $\beta$ , ya que no depende de las covariables.

La verosimilitud parcial no es una verosimilitud en el sentido usual ya que su forma depende de los conjuntos de individuos en riesgo,y la llegada o salida de un nuevo individuo puede cambiar la forma de la verosimilitud de acuerdo a su posición en el tiempo y según resulte censurado o no. Pese a ésto, para propósitos de inferencias sobre $\beta$, $L(\beta)$ puede ser tratada como cualquier función de verosimilitud, pues se ha probado que asintóticamente se comporta como tal.

Tomando $L(\beta)$ en sentido usual, tenemos que la aportación del i-ésimo individuo a la verosimilitud es:

$$L_i = \frac{\exp(Z_i'\beta)}{\sum_{j\in R_i} \exp(Z_j'\beta)} .$$

Entonces

$$\frac{\partial \log L_i}{\partial \beta_r} = Z_{ir} - \frac{\sum_{j\in R_i} Z_{jr} \exp(Z_j'\beta)}{\sum_{j\in R_i} \exp(Z_j'\beta)} \qquad r = 1,2,\ldots,p$$

$$\frac{\partial \log L_i}{\partial \beta_r \partial \beta_s} = \frac{\sum_{j \in R_i} z_{jr} z_{js} \exp(z_j'\underline{\beta})}{\sum_{j \in R_i} \exp(z_j'\underline{\beta})}$$

$$+ \frac{\sum_{j \in R_i} z_{jr} \exp(z_j'\underline{\beta}) \sum_{j \in R_i} z_{js} \exp(z_j'\underline{\beta})}{\left( \sum_{j \in R_i} \exp(z_j'\underline{\beta}) \right)^2}$$

Las ecuaciones de máxima verosimilitud están dadas por

$$\frac{\partial \log L_i}{\partial \beta_r} = 0 \qquad r = 1, \ldots, p$$

y son usualmente resueltas por métodos iterativos, de donde se obtiene el estimador máximo verosimil $\hat{\underline{\beta}}$ . La matriz de información está dada por

$$I = -E\left( \frac{\partial^2 \log L}{\partial \beta_r \partial \beta_s} \right) \qquad r,s = 1, \ldots, p$$

En este caso, como es usual, $E\left( \frac{\partial \log L_i}{\partial \beta_r} \right) = 0$ ya que

$$E\left( \frac{\partial \log L_i}{\partial \beta_r} \right) = E\left[ E\left( \frac{\partial \log L_i}{\partial \beta_r} \middle| R_i \right) \right]$$

$$= E\left[ E\left( z_{ir} - \sum_{j \in R} z_{jr} \exp(z_j \underline{\beta}) \middle| R_i \right) \right]$$

$$= E\left[ \sum_{j \in R_i} z_{jr} \left( \frac{\exp(z_j \beta)}{\sum_{j \in R_i} \exp(z_j \beta)} \right) - \frac{\sum_{j \in R_i} z_{jr} \exp\{z_j \beta\}}{\sum_{j \in R_i} \exp(z_j \beta)} \right]$$

$$= E\left[ \frac{\sum_{j \in R_i} z_{jr} \exp\{z_j \beta\}}{\sum_{j \in R_i} \exp(z_j \beta)} - \frac{\sum_{j \in R_i} z_{jr} \exp\{z_j \beta\}}{\sum_{j \in R_i} \exp(z_j \beta)} \right] = E(0) = 0$$

por otro lado

$$-E\left( \frac{\partial^2 \log L}{\partial \beta_r \partial \beta_s} \right) = -E\left( E\left( \frac{\partial^2 \log L}{\partial \beta_r \partial \beta_s} \mid R_i \right) \right)$$

$$-E\left( E\left[ \frac{\sum_{j \in R_i} z_{js} \exp(z_j \beta) \sum_{j \in R_i} z_{js} \exp(z_j \beta)}{\left( \sum_{j \in R_i} \exp(z_j \beta) \right)^2} \mid R_i \right] \right)$$

Para obtener de manera explícita esta esperanza necesitamos conocer el mecanismo de censura, lo que normalmente no se da en la práctica, por lo que usualmente se utiliza en el análisis la matriz de información observada:

$$I_0 = \left( \frac{-\partial^2 \log L}{\partial \beta_r \partial \beta_s} \right)_{\hat{\beta}}$$

Bajo condiciones muy generales $L(\beta)$ proporciona un estimador $\hat{\beta}$ que es asintóticamente normal con media $\beta$ y matriz de covarianzas $I^{-1}$, la cual puede ser estimada consistentemente por $I_0^{-1}$. Basado en lo anterior se construyen regiones de confianza para $\beta$ o bien pruebas de hipótesis para $\beta = \beta_0$, con $\beta_0$ un punto de interés para el

estudio.

### 5.4.1 Estimación de la Función de Riesgo Básica

Ahora centraremos nuestra atención en realizar estimaciones sobre la función de riesgo básica $h_0(t)$.

Si $h_0(t)$ es paramétrica, digamos $h_0(t,\phi)$, una manera conveniente de resolver el problema es tomar $\underline{\beta}_0$ fijo y maximizar $L(\underline{\beta}_0, \phi(\underline{\beta}_0))$ que nos lleva a lo que se conoce como función de verosimilitud perfil.

En el caso de $h_0(t)$ desconocida, la estimaremos por métodos no paramétricos.

Definimos

$$H_0(t) = \int_0^t h_0(u)\, du \quad \text{la función de riesgo acumulada.}$$

Breslow (1974), desarrolló el estimador no paramétrico de la función de riesgo básica, el cual está dado por

$$\hat{H}_0(t) = \sum_{j \mid t_j < t} \frac{d_j}{\sum_{j \in R_l} \exp(Z_j \hat{\underline{\beta}})}$$

en donde $d_j$ = número de fallas al tiempo $t_j$ y $\sum_{j \in R_l} \exp\{Z_j \hat{\underline{\beta}}\}$ es el efecto loglineal total de las covariables de los individuos en el j-ésimo conjunto en riesgo.

Observese que si la población es homogénea ($\underline{Z} = 0$) este estimador se reduce al estimador derivado del Kaplan-Meier.

## 5.5 Estimación y Pruebas de Hipótesis para los Modelos de Vida Acelerada

Como ya vimos, la forma básica de este modelo es

$$y = \alpha + \underline{\beta}\underline{Z} + \varepsilon, \text{ con } E(\varepsilon) = 0 \text{ y } V(\varepsilon) = \sigma^2$$

Reescribiremos el modelo como

$$y = \alpha + \underline{\beta}\underline{Z} + \sigma e \qquad \ldots(6)$$

de tal manera que ahora $V(e) = 1$, con $f(e)$ función de densidad para el error independiente de $\underline{Z}$

Primeramente, supongamos que $f(e)$ es una función paramétrica y que estamos interesados en probar la hipótesis $\underline{\beta} = \underline{\beta}_0$.

Sea $w = y - \underline{Z}\underline{\beta}_0$ los residuos bajo la hipótesis nula. Definimos $w_{(1)} < \ldots < w_{(k)}$ los residuos ordenados no censurados con covariables $Z_{(i)}$ y $w_{(i1)} \ldots w_{(im_i)}$ los residuos censurados (los residuos asociados a las observaciones censuradas), en el intervalo $[w_{(i)}, w_{(i+1)})$ para $i=0,1,\ldots,k$ con covariables $Z_{(ij)}$.

Si $S(\cdot)$ denota la función de supervivencia del error, entonces la función de log-verosimilitud puede escribirse como

$$\log L = \sum_{i=0}^{k} \left[ \log f\left( \{w_{(i)} - \alpha - (\beta - \beta_0) Z_{(i)} \} \sigma^{-1} \right) \right.$$

$$\left. + \sum_{j=1}^{m_i} \log S\left( \{w_{(ij)} - \alpha - (\beta - \beta_0) Z_{(ij)} \} \sigma^{-1} \right) \right] - k \log \sigma$$

que puede reescribirse de la siguiente manera:

$$\log L = \sum_{i=0}^{k} \left\{ \log f\left(\tau_{(i)} - \gamma Z_{(i)}\right) + \sum_{j=1}^{m_i} \log S\left(\tau_{(ij)} - \gamma Z_{(ij)}\right) \right\} - k \log\sigma$$

con $\tau_{(i)} = (w_{(i)} - \alpha)\sigma^{-1}$, $\tau_{(ij)} = (w_{(ij)} - \alpha)\sigma^{-1}$ y $\gamma = (\beta - \beta_0)\ \sigma^{-1}$

Una prueba de puntajes para probar $\beta = \beta_0$ o equivalente $\gamma = 0$, utiliza la estadística

$$\left[\frac{d \log L}{d\gamma}\right]_{\gamma=0} = \sum_{i=0}^{k} \left( Z_{(i)} c_{(i)} + \sum_{j=1}^{m_i} Z_{(ij)} C_{(ij)} \right) \qquad \ldots(7)$$

$$\text{con } c_{(i)} = \left( \frac{-d \log f(\tau_{(i)})}{d\tau_{(i)}} \right) \quad \text{y}$$

$$C_{ij} = \left( \frac{-d \log S(\tau_{(ij)})}{d\tau_{(ij)}} \right) \qquad \ldots(8)$$

este hecho se desprende de que:

$$\frac{d}{d\gamma} \log f(\tau_{(i)} - \gamma Z_{(i)})\big|_{\gamma=0} = \frac{Z_{(i)} f'(\tau_{(i)})}{f(\tau_{(i)})}$$

$$= Z_{(i)} \left( \frac{-d \log f(\tau_{(i)})}{d\tau_{(i)}} \right) = Z_{(i)} c_i$$

de manera similar se concluye que:

$$\frac{d}{d\gamma} \log S(\tau_{(ij)} - \gamma Z_{(ij)})\big|_{\gamma=0} = Z_{(ij)} C_{ij}$$

Los puntajes (8) dependen de los valores de $\tau$, $\alpha$ y $\sigma$ que pueden ser estimados antes de calcular la estadística (7) . Por lo general se usan los estimadores de máxima verosimilitud de $\alpha$ y $\sigma$ sujetos a $\beta=\beta_0$, en cuyo caso la estadística (7) es asintóticamente eficiente bajo el supuesto de f (Prentice 1978).

Sin embargo, regularmente se tiene incertidumbre para la elección de f, además de que existe la posibilidad de que algunas observaciones aberrantes puedan tener un efecto dominante en (7), estas son razones importantes para buscar alternativas para dicha estadística.

Las pruebas no paramétricas basadas en rango son generalmente robustas bajo malas elecciones de f y son resistentes contra observaciones aberrantes. Una situación atractiva en las pruebas por rangos es que los puntajes correspondientes a (8) son independientes de $\alpha$ y $\sigma$ por lo que el problema de estimar estos parámetros no se presenta.

Como la construcción de la estadística de prueba para datos censurados es esencialmente la misma que para datos no censurados, es conveniente realizar, en primer lugar, el desarrollo de la prueba para este último caso.

### 5.5.1 Caso sin Censura

Suponga que no hay posibilidad de censura, es decir $m_i=0$ para toda i, además el tamaño total de muestra es n=k . El vector de rangos r=r(w) está dado por las etiquetas correspondientes a los residuos ordenados; esto es, r(w) = {(1),...,(n)} y la probabilidad del vector de rangos es

$$P(r) = \int_R \prod_{i=1}^{n} f(\tau_{(i)}-\gamma z_{(i)})d\tau_{(i)} \qquad (9)$$

con R la región $\tau_{(1)}<\ldots<\tau_{(n)}$. Puede desarrollarse una prueba basada en la estadística de puntajes para $\gamma=0$, es decir $\beta=\beta_0$, derivada de (9), de la manera siguiente:

$$v = \left[\frac{d \log P(r)}{d\gamma}\right]_{\gamma=0} = \frac{1}{P(r)} \frac{dP(r)}{d\gamma} \Bigg|_{\gamma=0}$$

$$= \frac{1}{\int_R \prod_{j=1}^{n} f(\tau_{(j)})} \int_R \left\{-Z_{(i)} \frac{df(\tau_{(i)})}{d\tau_{(i)}}\right\} \prod_{j=1}^{n} \left\{f(\tau_{(j)})d\tau_{(j)}\right\}$$

$$= \sum_{i=1}^{n} Z_{(i)} c_i$$

Como $\int_R \prod_{j=1}^{n} (\tau_{(j)}) = \frac{1}{n!}$ entonces

$$c_i = n! \int_R \left\{ \frac{df(\tau_{(i)})}{d\tau_{(i)}} \right\} \prod_{j=1}^{n} \left\{f(\tau_{(j)})d\tau_{(j)}\right\}$$

si hacemos el cambio de variable

$$u_i = 1-S(\tau_{(i)})$$

entonces $du_i = f(\tau_{(i)})d\tau_{(i)}$

y $\tau_{(i)} = S^{-1}(1-u_i)$, por lo que

$$c_i = n! \int_R \Phi(u_i) \prod_{i=1}^{n} du_i = E\left\{\Phi(u_i)\right\}$$

Note que $u_i$ es la i-ésima estadística de orden de una muestra uniforme (0,1) de tamaño n, y $\Phi(u)$ $(0<u<1)$ está dada por

$$\Phi(u) = \Phi(u,f) = -d \log f(\tau_{(i)})/d\tau_{(i)} = \frac{f'(S^{-1}(1-u_i))}{f(S^{-1}(1-u_i))}$$

El hecho que $E(u_i) = i(n+1)^{-1}$ lleva a un sistema de puntajes asintóticamente equivalente dados por

$$c_i = \Phi\{i(n+1)^{-1}\}$$

Algunos ejemplos de interés de estos sistemas son:

1) Densidad logística:

$$f(\tau) = \exp\{\tau\}(1+\exp\{\tau\})^{-2}$$

$$\frac{-f'(\tau)}{f(\tau)} = \frac{2\exp\{\tau\}}{(1+\exp\{\tau\})} -1 = 2F(\tau) -1$$

$$c_i = E(2F(\tau)-1) = 2i(n+1)^{-1}-1$$

con F la función de distribución.

2) Densidad del valor extremo:

$$f(\tau) = \exp(\tau-\exp\{\tau\})$$

$$\frac{-f'(\tau)}{f(\tau)} = \exp\{\tau\} -1$$

$$c_i = E(\exp\{\tau_{(i)}\} -1)$$

como $T = \exp\{\tau\}$ tiene distribución exponencial con media 1, entonces $E(\exp\{\tau_{(i)}\})$ es el valor esperado de la i-ésima estadística de orden en una muestra de tamaño n de la distribución exponencial estándar, dada por

$$c_i = \sum_{k=1}^{i} \frac{1}{n-k+1} - 1$$

3) Densidad Normal estándar

$c_i = \Phi^{-1}\{i(n+1)^{-1}\}$ con $\Phi$ la función de distribución de una normal estándar.

Pese a que todos estos puntajes tienen expresiones matemáticas diferentes, para cada n fija, sus valores numéricos son muy parecidos si n es grande, con lo que se observa el hecho de que esta estadística es robusta bajo malas elecciones de f. El cambio de una prueba paramétrica por una que no lo es tiene sus desventajas, porque la potencia de la prueba disminuye, pero a cambio tenemos robustez y resistencia con la prueba no paramétrica.

Finalmente observamos que:

$$E(v) = E\left(\sum_{i=1}^{n} c_i Z'_{(i)}\right) = \sum_{i=1}^{n} E(c_i)\overline{Z}'_{(i)} = 0$$

ya que $E(c_i) = 0$, porque $c_i$'s son puntajes.

La matriz de varianzas-covarianzas es

$$V = E(vv') = \sum\sum_{i \neq j} c_i c_j E(Z_{(i)} Z'_{(j)})$$

ahora como

$$\left(\sum c_i\right)^2 = \sum c_i^2 + \sum_{i \neq j} c_i c_j = 0,$$

entonces

$$\sum_{i \neq j} c_i c_j = -\sum c_i^2 \quad y$$

también

$E(Z_{(i)}Z'_{(j)}) = [n(n-1)]^{-1}\left(\sum Z_i Z'_i - n^2 \bar{Z}\,\bar{Z}'\right)$, por lo que

tenemos

$$V = \frac{1}{n-1}\sum c_i^2 Z\,Z'$$

donde Z (nxp) es la matriz de covariables corregidas por su media. Si Z es de rango completo, en general bajo $\beta = \beta_0$ la estadística v es asintóticamente normal de donde v'Vv será asintótica $\chi^2$ con p grados de libertad.

### 5.5.2 Caso con Censura

En este caso no es directa la definición de rangos dado que para las observaciones censuradas sólo se sabe que exceden a un cierto valor. Prentice (1978) propone una extensión para este caso definiendo un vector de rangos generalizado r=r(w)= [(1),...,(k); (i1),...,(im$_i$)], i=0,...,k con probabilidad

$$p(r) = \int_R \prod_{i=1}^{k}\left\{f(\tau_{(i)}-\gamma Z_{(i)})\prod_{j=1}^{m_i} S(\tau_{(i)} - \gamma Z_{(ij)})\right\} \qquad \ldots(10)$$

en donde (1),...,(k) etiquetan las fallas y (i1)...(im$_i$) etiquetan las censuras en el intervalo $[W_i, W_{i+1})$ i=0,...,k.

Por lo que, de manera similar al caso sin censura, es posible generar una estadística de puntajes para probar $\beta=\beta_0$ a partir de (10) dada por

$$v = \left[\frac{d \log p(r)}{d\gamma}\right]_{\gamma=0} = \sum_{i=1}^{k} (z_{(i)}c_i + ss_{(i)}c_i)$$

donde $ss_{(i)} = z_{(i1)}+\ldots+z_{(im_i)}$, $c_i$ el puntaje correspondiente a $w_{(i)}$ y $C_i$ puntaje correspondiente para cada $w_{(i1)},\ldots,w_{(im_i)}$, con las expresiones para los puntajes dadas por:

$$c_i = \int_R \left\{-\frac{d f(\tau_{(i)})}{d\tau_{(i)}}\right\} \prod_{j=1}^{k} \left\{n_j S^{m_j}(\tau_{(j)}) f(\tau_{(j)}) d\tau_{(j)}\right\}$$

$$C_i = \int_R \left\{-\frac{d S(\tau_{(i)})}{d\tau_{(i)}}\right\} \prod_{j=1}^{k} \left\{n_j S^{m_j}(\tau_{(j)}) f(\tau_{(j)}) d\tau_{(j)}\right\}$$

con R la región $\tau_{(1)}<\ldots<\tau_{(k)}$ .

Para deducir lo anterior obsérvese primeramente que

$$P(r)\Big|_{\gamma=0} = \frac{1}{n_1\ldots n_k} = \frac{1}{\prod_{j=1}^{k} n_j}$$

donde $n_i=(m_i+1) +\ldots+(m_k+1)$ es el número de individuos con valores de w que se sabe exceden a $w_{(i)}$. Además,

$$\prod_{j=1}^{m_i} S(\tau_{(i)}-\gamma z_{(ij)}) = S^{m_i}(\tau_{(i)}-\gamma z_{(ij)}),$$

que al evaluarla en $\gamma=0$, se obtiene $S^{m_i}(\tau_{(i)})$ .

Por lo que

$$\left.\frac{d \log p(r)}{d\gamma}\right|_{\gamma=0} = \frac{1}{p(r)} \left.\frac{d\, p(r)}{d\gamma}\right|_{\gamma=0}$$

$$= \frac{1}{\prod_{j=1}^{k} n_j} \left[ \int_R \left\{ z_{(1)} \frac{-d\, f(\tau_{(1)})}{d\tau_{(1)}} \right\} \prod_{j=1}^{k} \left\{ f(\tau_{(j)}) d\tau_{(j)} \prod_{j=1}^{m_1} S(\tau_{(1)}) \right\} \right.$$

$$\left. + \int_R \left\{ ss_{(1)} \frac{-d \log S(\tau_{(1)})}{d\tau_{(1)}} \right\} \prod_{j=1}^{k} \left\{ f(\tau_{(j)}) d\tau_{(j)} \prod_{j=1}^{m_1} S(\tau_{(1)}) \right\} \right]$$

$$= z_{(1)} \int_R \left\{ \frac{-d \log f(\tau_{(1)})}{d\tau_{(1)}} \right\} \prod_{j=1}^{k} \left\{ n_j S^{m_1}(\tau_{(j)}) f(\tau_{(j)}) d(\tau_{(1)}) \right\}$$

$$+ ss_{(1)} \int_R \left\{ \frac{-d \log S(\tau_{(1)})}{d\tau_{(1)}} \right\} \prod_{j=1}^{k} \left\{ n_j S^{m_1}(\tau_{(j)}) f(\tau_{(j)}) d(\tau_{(1)}) \right\}$$

$$= \sum_{1=1}^{k} \left[ z_{(1)} c_1 + ss_{(1)} c_1 \right] .$$

Si realizamos nuevamente el cambio de variable

$u_j = 1-S(\tau_{(j)})$ para $j=1,\ldots,k$ tenemos

$$\Phi(u) = \frac{f'\{S^{-1}(1-u)\}}{f\{S^{-1}(1-u)\}} \quad y$$

$$\phi(u) = (1-u)^{-1} f\{S^{-1}(1-u)\}$$

entonces, reescribimos los puntajes como

$$c_i = \int \Phi(u_i) \prod_{j=1}^{k} \{n_j(1-u_j)^{m_j} du\}$$

$$C_i = \int \phi(u_i) \prod_{j=1}^{k} \{n_j(1-u_j)^{m_j}\} du \ .$$

Ahora calcularemos, para el caso con censura, algunos de los sistemas de puntajes vistos anteriormente.

Primeramente definimos

$$J\{g(u_i)\} = \int g(u_i) \prod_{j=1}^{k} \{n_j(1-u_j)^{m_j} du_j\}$$

para una función arbitraria g. En particular, para $g(u)=(1-u)^l$ con l entero, tenemos que

$$J\{(1-u_i)^l\} = \prod_{j=1}^{l} \left(\frac{n_j}{n_j+1}\right) \qquad l=1,2...$$

Sea $S(w) = S\{(w-\alpha)\sigma^{-1}\}$ la función de supervivencia condicional en w, dado $\beta=\beta_0$. Entonces, bajo esta hipótesis

$$\tilde{S}(w) = \prod^{*} \frac{n_j}{n_j+1}$$

con $\prod^{*}$ el producto sobre $j|w_{(j)} \leq w$, es un estimador consistente de S(w), bajo condiciones muy generales de censura. De hecho $\tilde{S}(w)$ es aproximadamente igual al estimador Kaplan-Meier

$$\hat{S}(w) = \prod^{*} \frac{n_j-1}{n_j}$$

Utilizando lo anterior calculemos estos sistemas de puntajes

1) Densidad logística:

$\Phi(u)=2u-1$

como $S(\tau)= (1+\exp\{\tau\})^{-1}$ y $\dfrac{S'(\tau)}{S(\tau)} = 1-\exp\{\tau\}$, recordando que $S(\tau)=1-u$, tenemos

$$\phi(u)=u$$

por lo que los puntajes respectivos para $\Phi(u)$ y $\phi(u)$ son

$$c_i = 1-2\tilde{S}(w_{(i)}) \ .$$

$$c_i = 1-\tilde{S}(w_{(i)}) \ .$$

2) Densidad del valor extremo:

Como $f(\tau) = \exp(\tau-\exp\{\tau\})$ y $S(\tau) = \exp(-\exp\{\tau\})$, tenemos

$$\frac{-f'(\tau)}{f(\tau)} = \exp(\tau)-1 \quad y \quad \frac{-S'(\tau)}{S(\tau)} = \exp\{\tau\}$$

de donde, recordando que $S(\tau) = 1-u$, obtenemos

$$\Phi(u) = -\log(1-u)-1$$
$$\phi(u) = -\log(1-u) \ .$$

Si realizamos la integral tenemos que

$$J\{\log(1-u_{(i)})\} = \log\tilde{S}(w_{(i)})$$

por lo que los puntajes son

$$c_i = -\log\tilde{S}(w_{(i)})-1 \ .$$

$$c_i = -\log\tilde{S}(w_{(i)}) \ .$$

Por último, observemos que como en el caso sin censura

$$E(v) = E\left(\Sigma\, Z_{(i)}c_i + ss_{(i)}c_i\right) = 0$$

y que la matriz de información observada de Fisher

$$V_0 = \left.\frac{-d^2 \log P(\{r\})}{d\gamma^2}\right|_{\gamma=0}$$

provee un estimador de la varianza, generalmente apropiado, dado por

$$V_0 = \sum_{i=1}^{k}\left[Z_{(i)}Z'_{(i)}J\{\psi_1(u_i)\} + \sum_{i=1}^{m_i} Z_{(ij)}Z'_{(ij)}J\{\psi_2(u_i)\}\right]$$

$$-\{J(bb') - vv'\}$$

con $$\psi_1(u) = \left[\frac{d^2 \log f(\tau)}{d\tau^2}\right]_{(\tau=S^{-1}(1-u))}$$

$$\psi_2(u) = \left[\frac{d^2 \log S(\tau)}{d\tau^2}\right]_{(\tau=S^{-1}(1-u))}$$

y $$b = \sum_{i=1}^{k}\{Z_{(i)}\Phi(u_i) + s_{(i)}\phi(u_i)\}$$

que se deduce derivando dos veces en (10). Bajo $\beta = \beta_0$ y normalidad asintótica para v, tenemos que $v'V_0^{-1}v$ es asintóticamente $\chi^2$ con grados de libertad la dimensión de Z, suponiendo $V_0$ no singular. Esta última distribución también nos permite construir regiones de confianza para $\underline{\beta}$.

# CONSIDERACIONES FINALES

En este trabajo se han presentado solamente los procesos básicos del Análisis de Supervivencia. Se ilustró con situaciones sencillas algunos de los usos que hasta ahora tiene esta rama estadística. No obstante, esperamos que su consulta pueda resultar de utilidad para reconocer si esta herramienta es la apropiada para el análisis de la información que sobre algún fenómeno particular se tenga.

Por supuesto que aún queda mucho por dar a conocer en esta área, como muestra para motivar a quien se interese en el tema, diremos que en los modelos con covariables se analizó el caso en que éstas son vectores constantes y resta el caso en el que las covariables dependen del tiempo. No se mencionó nada acerca de datos multivariados, ni tampoco se desarrollaron los modelos mixtos.

Estos son sólo algunos temas que esperamos motiven un trabajo más extenso en el Análisis de Supervivencia.

# BIBLIOGRAFIA


-COX,D.R.(1975). Partial Likelihood. *Biometrika*, 62, 269-276.

-COX,D.R. and OAKES,D.(1984). *Analysis of Survival Data*. Chapman and Hall. Great Britain.

-KALBFLEISCH,J.D. and PRENTICE,R.L.(1980).*The Statistical Analysis of Failure Time Data*. John Wiley and Sons. New York.

-LAWLESS,J.F.(1982).*Statistical Models and Methods for Lifetime Data*. John Wiley and Sons. New York.

-MILLER,R.G.(1981).*Survival Analysis*. Wiley and Sons. New York.

-MOOD,A., GRAYBILL,F. and BOES,D.(1974).*Introduction to the Theory of Statistics*. Mc.Graw-Hill. Singapure.

-PRENTICE,R.L.(1978).Linear rank tests with right censored data. *Biometrika*, 65, 167-179.

-TREJO VALDIVIA,G.M.B.(1985).La Función de Influencia en el Análisis de Datos de Supervivencia. Tesis de maestría en Estadística e Investigación de Operaciones. UACPyP del CCH-UNAM.