

Inferencia Estadística: Una Visión General

Gabriel Escarela Pérez

México D.F., octubre del 2003

Resumen

Este trabajo presenta una sinopsis de principios y aspectos de la inferencia estadística. Los temas centrales están principalmente enfocados a la función de verosimilitud y a la familia exponencial dentro de un marco frecuentista. A pesar de la aparente diversidad de las técnicas para las diferentes familias de distribuciones, se logra unificar la teoría para el estudio de datos provenientes de prácticamente cualquier naturaleza. Se enfatiza el uso de información concomitante y se describe el modelo lineal generalizado. Los métodos se ilustran con unos datos de riesgos contendientes, los cuales representan cierta complejidad para su análisis.

1 Introducción

El propósito fundamental de la inferencia estadística es el de describir aspectos específicos de los artificios aleatorios que representan la ocurrencia de cierto evento dado un conjunto de datos. En la mayoría de los casos el interés principal de la inferencia se concentra en el valor de uno o varios parámetros desconocidos referentes a ciertas características de una población.

Existen tres tipos principales de inferencia, estos son: estimación puntual, estimación por intervalos y pruebas de hipótesis. En estimación puntual, por cada parámetro desconocido de interés se calcula un solo valor a partir de los datos, y entonces es usado como estimación del parámetro.

Las estimaciones puntuales proveen un reporte conciso de los datos muestrales, pero no contienen información acerca de su precisión. Un intervalo de confianza cubre un rango de valores que tienen una probabilidad predeterminada de incluir al valor verdadero del parámetro que sigue desconocido. Así, los intervalos de confianza ayudan a cuantificar la incertidumbre asociada con los estimadores.

Las pruebas de hipótesis establecen hipótesis específicas acerca de los parámetros de interés y evalúan la plausibilidad de cualquier hipótesis al comprobar si los datos observados apoyan o rechazan la hipótesis; aunque las pruebas de hipótesis muchas veces pueden ser artificiales en el sentido de que ninguna de las hipótesis será completamente correcta, es una forma conveniente de proceder y representa una parte sustancial de la investigación científica.

Así como hay diferentes tipos de inferencia, también hay diferentes métodos. El esquema más ampliamente usado es el frecuentista, el cual parte de la suposición de que uno puede tomar aleatoriamente muestras repetidas de datos de la población, bajo las mismas condiciones en que se tomó la única muestra que se tiene a la mano. Las propiedades de los estimadores puntuales, los estimadores de intervalos, y las pruebas de hipótesis se derivan bajo esta suposición de muestreo repetido.

El principal método alternativo a la inferencia frecuentista es el esquema Bayesiano. Aquí no hay necesidad de suponer muestreo repetido; en vez de tomar los parámetros como fijos, en el esquema Bayesiano es necesario especificar una distribución de probabilidad para los parámetros de interés. La inferencia Bayesiana permite atacar el problema de inferencia en una forma más informativa; sin embargo, sufre de dificultades prácticas tales como seleccionar una distribución de probabilidad “adecuada” para los parámetros.

Un tercer esquema de la inferencia se basa en la teoría de las decisiones, en el cual una inferencia es vista como la toma de una decisión entre estimadores o hipótesis contendientes. Como en la inferencia Bayesiana, se asigna una distribución de probabilidad a los parámetros de interés, pero además se necesita una función de utilidad. Tal función mide qué tan aceptable es alguna decisión estadística (del estimador o la hipótesis) para cada conjunto de valores posibles de los parámetros.

El objetivo primordial de este trabajo no es el de discutir en detalle las diferencias metodológicas entre los diversos esquemas de inferencia, el cual se encuentra bien documentado en Barnett (1982). El propósito de este artículo es el de exponer la implementación de los tipos de inferencia dado un conjunto de datos con una estructura potencialmente compleja. Como máxima verosimilitud es uno de los métodos más recurridos y flexibles, se adopta el esquema frecuentista para la descripción de los principales aspectos de la inferencia estadística paramétrica.

La sección 2 a continuación revisa los principales principios de reducción de datos y algunas propiedades deseables de los estimadores, introduciendo y mencionando aspectos importantes sobre la familia exponencial de distribuciones. La sección 3 describe la técnica de máxima verosimilitud y métodos para encontrar pruebas de hipótesis y regiones de confianza, haciendo énfasis en aplicaciones para muestras de tamaño grande. La sección 4 extiende las ideas presentadas en el marco general para la situación en que se requiere incluir los efectos de una o varias variables explicativas, poniendo particular atención al

modelo lineal generalizado. El trabajo concluye en la sección 5 con un ejemplo de inferencia estadística al aplicar los métodos presentados a unos datos de riesgos contendientes.

2 Principios de la Reducción de Datos

Como se ha mencionado, se usa la información de una muestra aleatoria $\mathbf{X} = (X_1, X_2, \dots, X_n)$ para hacer inferencias acerca de un parámetro desconocido θ o de un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$. Si la muestra de tamaño n es grande, entonces la muestra observada $\mathbf{x} = (x_1, x_2, \dots, x_n)$ es una lista larga de números que puede ser difícil de interpretar. Tal vez se desee resumir la información contenida en una muestra al determinar unas cuantas características claves de los valores muestrales. Esto típicamente se hace al estudiar un *estadístico*, que es una función de la muestra $T(\mathbf{X})$. Por ejemplo, la media muestral, la varianza muestral, la observación más grande y la observación más chica son cuatro estadísticos que pueden ser usados para resumir algunas características claves de la muestra.

Cualquier estadístico define una forma de reducción de los datos, i.e. un resumen de los datos. La reducción de datos puede ser vista como una partición del espacio muestral S . Sea $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ para alguna } \mathbf{x} \in S\}$ la imagen de S bajo $T(\mathbf{x})$. Entonces $T(\mathbf{x})$ está particionando el espacio muestral en conjuntos A_t , $t \in \mathcal{T}$, definidos por $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. El estadístico resume los datos al reportar que $T(\mathbf{x}) = t$ o equivalentemente a que $\mathbf{x} \in A_t$. Por ejemplo, si $T(\mathbf{x}) = \sum_{i=1}^n x_i$, entonces $T(\mathbf{x})$ no reporta los valores de la muestra sino sólo el de la suma.

La meta principal de la *reducción de los datos* es encontrar métodos que no excluyan información importante acerca del parámetro desconocido θ y métodos que tienen la virtud de excluir exitosamente información que es irrelevante en el conocimiento de θ . Las ventajas y consecuencias de este tipo de reducción de datos son los tópicos de esta primera sección.

Este estudio se concentrará en tres principios de reducción de datos, que son: *i*) El principio de suficiencia, el cual promueve un método de reducción de datos que no excluye información acerca de θ y además logra cierto compendio de los datos; *ii*) El principio de verosimilitud, que describe una función del parámetro determinada por la muestra observada, la cual contiene toda la información sobre θ y que está disponible; *iii*) El principio de invariancia, que preserva algunas características importantes del modelo.

2.1 El Principio de Suficiencia

Un estadístico suficiente de un parámetro θ es un estadístico que - en cierto sentido - captura toda la información sobre θ contenida en la muestra. Cualquier información adicional en

la muestra, aparte del valor del estadístico suficiente, no contiene más información acerca de θ . Estas consideraciones conllevan a la técnica de reducción de datos conocida como el *principio de suficiencia*.

Un estadístico suficiente de θ se define formalmente como aquel estadístico $T(\mathbf{X})$ tal que la distribución condicional de la muestra \mathbf{X} dado el valor de $T(\mathbf{X})$ no depende de θ (e.g. Lehmann, 1986). Como consecuencia de esta definición, el *principio de suficiencia* ha sido propuesto; éste indica que las mismas conclusiones deben de obtenerse a partir de todos los conjuntos de observaciones con los mismos estadísticos suficientes.

Es común, sin embargo, encontrar varios estadísticos suficientes para un mismo problema. De esta forma, es deseable encontrar un estadístico que logre la mayor reducción de los datos y que además retenga toda la información sobre θ . A este estadístico se le conoce como *estadístico mínimo suficiente*. Formalmente, un estadístico suficiente $T(\mathbf{X})$ es mínimo suficiente si, para cualquier otro estadístico suficiente $T'(\mathbf{X})$, $T(\mathbf{X})$ es una función de $T'(\mathbf{X})$; en este caso, decir que $T(\mathbf{x})$ es una función de $T'(\mathbf{x})$ significa que si $T'(\mathbf{x}) = T'(\mathbf{y})$ entonces $T(\mathbf{x}) = T(\mathbf{y})$.

A diferencia de un estadístico suficiente, un *estadístico auxiliar* no contiene información acerca de θ . Un estadístico auxiliar $T_2(\mathbf{X})$ es una observación de una variable aleatoria cuya distribución es fija y conocida, y además no tiene relación con θ . Paradójicamente, cuando se usa junto con otros estadísticos, un estadístico auxiliar puede contener información importante para obtener inferencias sobre θ . Un estadístico auxiliar y un estadístico mínimo suficiente pueden estar relacionados; más aún, el estadístico auxiliar muchas veces puede dar información acerca de la precisión de un estimador de θ (e.g. Efron y Hinkley, 1978). De esta forma, si un estadístico auxiliar existe, la inferencia sobre el parámetro θ debe basarse en la distribución condicional del estadístico mínimo suficiente $T(\mathbf{X})$ dado $T_2(\mathbf{X})$ (principio de auxiliaridad).

Para muchas situaciones importantes, sin embargo, la intuición de que un estadístico mínimo suficiente es independiente de cualquier estadístico auxiliar es correcta. Sea $f(t|\theta)$ una familia de funciones de densidad para un estadístico $T(\mathbf{X})$. La familia de distribuciones de probabilidad se llama *completa* si cuando $E_\theta[g(T)] = 0$ para todo θ implica que $P_\theta\{g(T) = 0\} = 1$ para todo θ . En este caso, $T(\mathbf{X})$ es llamado un *estadístico completo*. El Teorema de Basu demuestra que si $T(\mathbf{X})$ es un estadístico mínimo y suficiente, entonces $T(\mathbf{X})$ es independiente de cualquier estadístico auxiliar. Este resultado es importante pues permite decidir la independencia de los estadísticos $T(\mathbf{X})$ y $T_2(\mathbf{X})$, y así evita la necesidad de encontrar la distribución conjunta de los dos estadísticos (Lehman, 1981).

Considérese la *familia exponencial de k -parámetros* definida por la siguiente función de

densidad

$$f(x|\boldsymbol{\theta}) = \exp \left\{ \sum_{j=1}^k A_j(\boldsymbol{\theta}) B_j(x) + C(x) + D(\boldsymbol{\theta}) \right\}, \quad (1)$$

donde $A_1(\boldsymbol{\theta}), A_2(\boldsymbol{\theta}), \dots, A_k(\boldsymbol{\theta})$ y $D(\boldsymbol{\theta})$ son funciones de $\boldsymbol{\theta}$ unicamente, y $B_1(x), B_2(x), \dots, B_k(x)$, y $C(x)$ son funciones de x unicamente. Miembros populares de la familia exponencial son las distribuciones $\text{BIN}(n, p)$, $\text{POI}(\mu)$, $\text{EXP}(\mu)$, $\text{N}(0, \sigma^2)$, $\text{N}(\mu, 1)$ y $\text{N}(\mu, \sigma^2)$ entre otras. Si X tiene una distribución que pertenece a la familia exponencial de k -parámetros entonces es posible demostrar que

$$\mathbf{T} = \left(\sum_{i=1}^n B_1(X_i), \sum_{i=1}^n B_2(X_i), \dots, \sum_{i=1}^n B_k(X_i) \right)$$

es un estadístico completo (mínimo) suficiente. Aquí, mínimo está en paréntesis pues suficiencia y completos implica suficiencia mínima suficiente automáticamente, aunque la posición converso no es verdadera (Zacks, 1971). El vector de estadísticos suficientes \mathbf{T} del vector de parámetros $\boldsymbol{\theta}$ existe si y sólo si la distribución de X pertenece a la familia exponencial de k parámetros (e.g. Garthwaite *et. al.*, 1995).

La suficiencia juega un rol importante en la búsqueda de estimadores del parámetro θ con propiedades muy deseables tales como insesgamiento y varianza mínima. Supóngase que T es un estadístico suficiente de θ y que $\hat{\theta}$ es cualquier estimador insesgado de θ , i.e. $E[\hat{\theta}] = \theta$. Sea $\hat{\theta}_T = E[\hat{\theta}|T]$, entonces (Teorema Rao-Blackwell)

- (a) $\hat{\theta}_T$ es una función de T únicamente;
- (b) $E[\hat{\theta}_T] = \theta$;
- (c) $\text{Var}[\hat{\theta}_T] \leq \text{Var}[\hat{\theta}]$.

De esta forma $\hat{\theta}_T$ es un *estimador insesgado de varianza mínima* (EIVM). Es posible demostrar que si existe un EIVM de θ , entonces este estimador debe ser una función del estadístico mínimo suficiente de θ el cual es un EIVM. Se puede demostrar, por ejemplo, que si se tiene a la mano un miembro de la familia exponencial de $k = 1$ parámetro, y si $A(\theta)$ es una función estrictamente monótona de θ , entonces T es un EIVM; recíprocamente, para la existencia de un EIVM de una función de θ la función de distribución de X necesita ser miembro de la familia exponencial.

2.2 Los Principios de Verosimilitud e Invariancia

Un concepto importante que aparece en la inferencia es la función de verosimilitud. Sea $f(\mathbf{x}|\theta)$ la función de densidad conjunta de la muestra aleatoria $\mathbf{X} = (X_1, X_2, \dots, X_n)$. En-

tonces, dado que $\mathbf{X} = \mathbf{x}$ se observa, la función de θ definida por

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

es llamada la *función de verosimilitud*.

El *principio de verosimilitud* especifica cómo la función de verosimilitud debe de usarse para la reducción de datos. Si \mathbf{x} e \mathbf{y} son dos puntos muestrales tales que $L(\theta|\mathbf{x})$ es proporcional a $L(\theta|\mathbf{y})$, esto es: existe una constante $C(\mathbf{x}, \mathbf{y})$ tal que

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y})L(\theta|\mathbf{y}) \quad \text{para todo } \theta,$$

entonces las conclusiones obtenidas con \mathbf{x} e \mathbf{y} deben ser idénticas.

En el caso especial $C(\mathbf{x}, \mathbf{y}) = 1$, el principio de verosimilitud indica que si dos puntos muestrales resultan en la misma función de verosimilitud entonces contienen la misma información sobre θ . De hecho, el principio de verosimilitud va más lejos pues indica que, aún cuando los dos puntos muestrales sólo tienen verosimilitudes proporcionales, estos contienen información equivalente sobre θ . Este es el razonamiento: la función de verosimilitud se usa para comparar la plausibilidad de varios valores del parámetro; si $L(\theta_2|\mathbf{x}) = 2L(\theta_1|\mathbf{x})$ entonces θ_2 es el doble de plausible de lo que es θ_1 . Si el principio de verosimilitud es verdadero, entonces $L(\theta_2|\mathbf{y}) = 2L(\theta_1|\mathbf{y})$. De esta forma, si se observa \mathbf{x} o \mathbf{y} se concluye de igual forma que θ_2 es el doble de plausible de lo que es θ_1 .

Nótese que se ha seleccionado cuidadosamente la palabra “plausible” en vez de “probable” ya que se debe considerar a θ como un valor fijo; aunque $f(\mathbf{x}|\theta)$ es una función de densidad, no existe garantía alguna de que $L(\theta|\mathbf{x})$ sea una función de densidad también.

El principio de verosimilitud no especifica que la inferencia estadística únicamente debe basarse en este principio. Es un principio relativo en el sentido de que compara modelos pero no provee ningún conocimiento acerca de alguno de los modelos posibles. Esto es lógico si ningún modelo puede ser verdadero. Sólo se desea el mejor modelo de entre aquellos disponibles para ayudar a entender cómo pudieron haber sido producidos los datos.

El principio de invariancia describe la reducción de datos algo diferente a como se ha visto con los principios de suficiencia y verosimilitud. Este principio combina dos consideraciones: 1) la inferencia que se hace no debe de depender de la escala de medición que se usa con los datos; 2) Si dos problemas de inferencia tienen la misma estructura formal en términos del modelo matemático usado, entonces el mismo procedimiento de inferencia debe ser usado en ambos problemas.

El *principio de invariancia* restringe a la inferencia al determinar que otras inferencias deben de hacerse en puntos muestrales relacionados. De hecho, los tres principios predeterminan relaciones similares entre inferencias hechas en diferentes puntos muestrales. El

principio de suficiencia indica que la inferencia debe de ser la misma en dos puntos cualesquiera que dan el mismo valor del estadístico suficiente. El principio de verosimilitud hace la misma indicación para dos puntos cualesquiera que dan funciones de verosimilitud proporcionales. De esta forma, las tres técnicas de reducción de datos restringen el conjunto de inferencias permisibles y de esta forma simplifican el análisis del problema.

3 Verosimilitud

Estimación de máxima verosimilitud es el mejor conocido, más ampliamente usado y más importante de todos los métodos de estimación. Todo consiste en determinar la función de verosimilitud $L(\boldsymbol{\theta}|\mathbf{x})$, y entonces se procede a encontrar el valor $\hat{\boldsymbol{\theta}}$ el cual maximiza $L(\boldsymbol{\theta}|\mathbf{x})$, con lo cual se obtiene el *estimador de máxima verosimilitud* (EMV). Encontrar el EMV se convierte, entonces, en un problema de optimización.

En muchas circunstancias es más conveniente trabajar con el logaritmo natural de $L(\boldsymbol{\theta}|\mathbf{x})$. En este estudio denotamos a esta función *log-verosimilitud* con $l_n(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}|\mathbf{x})$. En práctica, la maximización de $l_n(\boldsymbol{\theta})$ es más sencilla que la maximización directa de $L(\boldsymbol{\theta}|\mathbf{x})$; como $\ln(\cdot)$ es una función monótona, el mismo valor $\hat{\boldsymbol{\theta}}$ maximiza a $l_n(\boldsymbol{\theta})$ y a $L(\boldsymbol{\theta}|\mathbf{x})$.

Entre más angosta sea la gráfica de la función de verosimilitud, o de log-verosimilitud, mayor información se estará obteniendo acerca del parámetro. De esta forma, la anchura indica la precisión del estimador. Una función de verosimilitud con la gráfica de una función aproximadamente constante indica que los datos no proveen información alguna.

En el caso de la familia exponencial de distribuciones, el EMV es un estadístico suficiente del parámetro; en otros casos, es simplemente un estimador puntual (Lindsey, 1996). En general, la ventaja más sobresaliente del método de máxima verosimilitud es que un problema de n dimensiones se puede reducir a uno de p dimensiones, donde p representa la dimensión del vector $\boldsymbol{\theta}$. Si el EMV no es suficiente, entonces se pierde información acerca del parámetro. Esta pérdida depende mucho de la cercanía del estimador con el parámetro.

En general, el uso de un EMV es atractivo cuando se tienen muestras relativamente grandes pues sus propiedades asintóticas hacen que la pérdida de información sobre el parámetro se reduzca en forma significativa. Estas propiedades son (e.g. Garthwaite *et. al.*, 1995):

Consistencia $\Pr\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| < \epsilon\} \rightarrow 0$ cuando $n \rightarrow \infty$ para cualquier $\epsilon > 0$; i.e. la función de densidad de $\hat{\boldsymbol{\theta}}$ se centra asintóticamente alrededor de $\boldsymbol{\theta}$ para n grande.

Insesgamiento, Normalidad y Eficiencia Cuando $n \rightarrow \infty$ entonces $\hat{\boldsymbol{\theta}} \sim \text{NMV}(\boldsymbol{\theta}, [\mathbf{I}_n(\boldsymbol{\theta})]^{-1})$, donde NMV denota la distribución normal p -variada y $\mathbf{I}_n(\boldsymbol{\theta})$ es la matriz de *infor-*

mación de Fisher definida como

$$\mathbf{I}_n(\boldsymbol{\theta}) = \mathbb{E} \left[-\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right].$$

En forma equivalente, se tiene el siguiente resultado “estudentizado”:

$$[\mathbf{I}_n(\boldsymbol{\theta})]^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \text{NMV}(\mathbf{0}, I_p), \quad \text{cuando } n \rightarrow \infty, \quad (2)$$

donde I_p es la matriz identidad de $p \times p$. Este resultado muestra: 1) que $\hat{\boldsymbol{\theta}}$ es *asintóticamente insesgado*, i.e. $\mathbb{E}[\hat{\boldsymbol{\theta}}] \rightarrow \boldsymbol{\theta}$; 2) la normalidad asintótica de $\hat{\boldsymbol{\theta}}$; y 3) que $\hat{\boldsymbol{\theta}}$ es *asintóticamente eficiente*, i.e. $\hat{\boldsymbol{\theta}}$ tiende a alcanzar la varianza mínima posible para un estimador insesgado de $\boldsymbol{\theta}$.

En práctica, el resultado de la normalidad aproximada de los EMV para muestras grandes no es un resultado útil pues $\mathbf{I}_n(\boldsymbol{\theta})$ es desconocida al igual que $\boldsymbol{\theta}$. Sin embargo, la matriz de *información observada*, definida como:

$$\tilde{\mathbf{I}}_n(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (3)$$

es un estimador consistente de $\mathbf{I}_n(\boldsymbol{\theta})$; de esta forma, se tiene que

$$[\tilde{\mathbf{I}}_n(\hat{\boldsymbol{\theta}})]^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \text{NMV}(\mathbf{0}, I_p), \quad \text{cuando } n \rightarrow \infty, \quad (4)$$

el cual es un resultado que sí se puede usar pues el estimador y la matriz de información observada pueden calcularse a partir de la muestra. Esta relación indica que, aproximadamente en muestras grandes, el EMV $\hat{\boldsymbol{\theta}}$ se distribuye normal p -variada con media $\boldsymbol{\theta}$ y matriz de covarianzas $[\tilde{\mathbf{I}}_n(\hat{\boldsymbol{\theta}})]^{-1}$ (e.g. Vu *et. al.*, 1996). El uso principal de esta propiedad es el cálculo de regiones de confianza para entradas individuales o conjuntos de componentes vectoriales de $\boldsymbol{\theta}$.

3.1 Pruebas de Hipótesis

La formulación más usual de un problema de hipótesis es cuando se tiene que decidir entre dos hipótesis sobre uno o más parámetros $\boldsymbol{\theta}$. Las dos hipótesis son la *hipótesis nula*, la cual se denota con $H_0 : \boldsymbol{\theta} \in \omega$ y la *hipótesis alternativa* que se denota con $H_1 : \boldsymbol{\theta} \in \Omega - \omega$. Aquí, Ω es el conjunto de todos los valores posibles de $\boldsymbol{\theta}$ el cual es conocido como el *espacio de parámetros*, y ω es algún subconjunto de Ω .

Para poder escoger entre H_0 y H_1 se selecciona algún subconjunto C de valores posibles de \mathbf{X} y se rechaza H_0 si y solo si $\mathbf{X} \in C$. Usualmente C se define en términos de valores extremos de algún estadístico $T(\mathbf{X})$. T es el *estadístico de prueba* y C es la *región crítica* o la *región de rechazo*; C^c , i.e. el complemento de C , es la *región de aceptación*.

Una hipótesis es *simple* si ésta especifica un solo valor de $\boldsymbol{\theta}$, i.e. cuando ω u $\Omega - \omega$ contienen solo un punto; de otra forma la hipótesis es *compuesta*.

El rechazo de H_0 cuando es verdadera es llamado el *error de tipo I*; la aceptación de H_0 cuando es falsa es un *error de tipo II*. Las probabilidades del error tipo I y del error tipo II se denotan con α y β respectivamente. Así mismo, α es conocido como el *nivel de significancia* o como el *tamaño* de la prueba.

Idealmente se desea que la región crítica haga a α y β lo más pequeños posible, pero es claro que si uno decrece el otro se incrementa. Se tiene que $\alpha = \Pr\{\mathbf{X} \in C|H_0\}$ y $\beta = \Pr\{\mathbf{X} \in C^c|H_1\}$; para decrecer a α se necesita quitar puntos de C , y entonces estos puntos se suman a C^c . Esta transferencia causa que β se incremente. De esta forma, el esquema *clásico* se basa en fijar α en algún nivel estándar predeterminado (en general se escoge $\alpha = 0.05, 0.01, 0.10$ o 0.001) y entonces se procede a encontrar una prueba que minimiza β .

Una prueba que minimiza β para un α fijo es llamada la prueba *más potente* o la *mejor* prueba de tamaño α . Cuando ambas hipótesis en cuestión son simples. i.e. cuando se desea probar $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, el lema Neyman-Pearson indica que la mejor prueba de tamaño α de H_0 vs. H_1 tiene región crítica de la forma

$$\frac{L(\boldsymbol{\theta}_1|\mathbf{x})}{L(\boldsymbol{\theta}_0|\mathbf{x})} \geq A \quad \text{para alguna constante no negativa } A.$$

La prueba propuesta por este lema suele ser llamada la prueba del *cociente de verosimilitud*.

Cuando se considera el marco general de hipótesis simples y compuestas, es posible generalizar las ideas vistas para cuando H_0 y H_1 son simples. La idea fundamental es la de encontrar una región crítica que es producto de la maximización de la *función potencia*, la cual se define como $\eta(\boldsymbol{\theta}) = \Pr\{\mathbf{X} \in C|\boldsymbol{\theta}\}$, para toda $\boldsymbol{\theta} \in \Omega - \omega$ y con un nivel de significancia α fijo. Cuando se alcanza esta optimización se tiene entonces una prueba *uniformemente más potente* (UMP). Si H_0 es simple pero H_1 es compuesta y si se puede encontrar la misma mejor prueba a partir del lema Neyman-Pearson de H_0 vs. $H_1' : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, donde $\boldsymbol{\theta}_1$ toma algún valor de $\boldsymbol{\theta}$ en $\Omega - \omega$, entonces esta prueba es UMP.

En general, cuando se considera el marco más general para probar $H_0 : \boldsymbol{\theta} \in \omega$ vs. $H_1 : \boldsymbol{\theta} \in \Omega - \omega$ es posible usar una *prueba del cociente de máxima verosimilitud* (PCMV), también conocida como la *prueba del cociente de verosimilitud generalizado*, la cual tiene región crítica de la forma:

$$\lambda = \left\{ \frac{\max_{\boldsymbol{\theta} \in \omega} L(\boldsymbol{\theta}|\mathbf{x})}{\max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}|\mathbf{x})} \right\} \leq A, \quad \text{para alguna constante } A. \quad (5)$$

La forma de esta prueba es intuitivamente plausible; se tiene que $0 \leq \lambda \leq 1$, y λ debe de estar cerca de 1 si $\boldsymbol{\theta} \in \omega$, pero lejos de 1 si $\boldsymbol{\theta} \notin \omega$. Además, cuando H_0 y H_1 son simples,

la PCMV se reduce a la mejor prueba dada por el lema Neyman-Pearson. De hecho, si H_0 es simple, entonces λ da una prueba UMP (si esta existe).

El uso del PCMV es en general atractivo por sus propiedades asintóticas. Esto se debe a que muchas veces λ no tiene una distribución conocida ni tampoco algún estadístico de prueba equivalente. En tales casos, y cuando n es relativamente grande, es posible usar el estadístico del *cociente de log-verosimilitud*, definido como

$$D_n(\boldsymbol{\theta}) = 2[l_n(\hat{\boldsymbol{\theta}}) - l_n(\boldsymbol{\theta})], \quad (6)$$

donde $\hat{\boldsymbol{\theta}}$ es el EMV de $\boldsymbol{\theta}$. Usando la aproximación de normalidad de los EMV y aplicando este resultado, es posible demostrar que la distribución asintótica de $D_n(\boldsymbol{\theta})$ bajo H_0 es la ji-cuadrada con p grados de libertad, i.e. $D_n(\boldsymbol{\theta}) \sim \chi_p^2$, donde p es la diferencia en dimensión entre H_0 y $H_0 \cup H_1$. Hay algunas familias paramétricas para las cuales este resultado es exacto.

Si se tiene una muestra aleatoria de observaciones de una distribución que pertenece a la familia exponencial de k -parámetros, entonces se tiene que

$$D_n(\tilde{\boldsymbol{\theta}}) = 2 \left\{ \sum_{j=1}^k \sum_{i=1}^n B_j(x_i) [A_j(\hat{\boldsymbol{\theta}}) - A_j(\tilde{\boldsymbol{\theta}})] + n [D(\hat{\boldsymbol{\theta}}) - D(\tilde{\boldsymbol{\theta}})] \right\},$$

donde $\tilde{\boldsymbol{\theta}}$ y $\hat{\boldsymbol{\theta}}$ son los EMV de $\boldsymbol{\theta}$ bajo H_0 y $H_0 \cup H_1$ respectivamente.

El cociente de log-verosimilitud es no negativo, lo cual es importante para comparar el ajuste que diferentes modelos pueden dar a los datos. Este estadístico es una función de las observaciones y también del parámetro desconocido. Entre más grande sea $D_n(\boldsymbol{\theta})$ más lejos se encuentra el modelo en consideración del modelo más plausible dado el conjunto de datos.

Supóngase que se desea comparar el ajuste de un modelo simple M_0 con parámetro $\boldsymbol{\theta}_0$ de dimensión r con un modelo más general M_1 con parámetro $\boldsymbol{\theta}_1$ de dimensión p . Aquí, el modelo M_0 está *anidado* en M_1 ; i.e. $\boldsymbol{\theta}_0 \subset \boldsymbol{\theta}_1$. Es posible demostrar que, aproximadamente, $D_n(\boldsymbol{\theta}_1) - D_n(\boldsymbol{\theta}_0) \sim \chi_{p-r}^2$ y que, consecuentemente, $d_n = 2[l_n(\hat{\boldsymbol{\theta}}_1) - l_n(\hat{\boldsymbol{\theta}}_0)] \sim \chi_{p-r}^2$. De esta forma, el *estadístico desviación* d_n (*deviance* en inglés) forma las bases para la siguiente prueba de hipótesis para modelos anidados:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\theta_1, \theta_2, \dots, \theta_r)^T \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1 = (\theta_1, \theta_2, \dots, \theta_p)^T \quad (r < p < n).$$

Si M_0 y M_1 describen bien los datos y d_n es consistente con la aproximación de la ji-cuadrada, en general se escoge el modelo M_0 correspondiente a H_0 pues es más simple. Por otra parte, si el valor de d_n se encuentra en la región crítica (i.e. si $d_n > \chi_{p-r, 1-\alpha}^2$, donde α es el nivel de significancia), entonces se rechaza H_0 en favor de M_1 a partir de que

M_1 provee una descripción de los datos significativamente mejor (aún si el mismo modelo no ajusta los datos particularmente bien).

Cuando se está interesado en un sólo parámetro θ_j , es posible usar el *estadístico de Wald* que es equivalente al uso del conocido estadístico de prueba de la normal estándar:

$$Z = \frac{\hat{\theta}_j}{\sqrt{\text{Var}[\hat{\theta}_j]}},$$

donde $\hat{\theta}_j$ es la j -ésima entrada de $\boldsymbol{\theta}^T$ y $\text{Var}[\hat{\theta}_j]$ es el (j, j) -ésimo componente de la matriz de covarianzas; en este caso, la cantidad $\sqrt{\text{Var}[\hat{\theta}_j]}$ es conocida como el *error estándar* (ee). Bajo la hipótesis nula de que $\theta_j = 0$, Z se distribuye asintóticamente $N(0, 1)$.

3.2 Estimación por intervalos

Es muy común tratar de dividir valores del parámetro $\boldsymbol{\theta}$ en una “región” plausible y una región “menos plausible”. En general se desea construir la región plausible con una probabilidad determinada de incluir el verdadero valor de $\boldsymbol{\theta}$. Más formalmente, se desea encontrar $\mathbf{S}_{\mathbf{X}}$, un subconjunto de Ω el cual depende de \mathbf{X} , tal que

$$\Pr\{\mathbf{X} : \mathbf{S}_{\mathbf{X}} \supset \boldsymbol{\theta}\} = 1 - \alpha.$$

En este caso se dice que $\mathbf{S}_{\mathbf{X}}$ es un *conjunto de confianza* de $\boldsymbol{\theta}$ con coeficiente de confianza $1 - \alpha$; cuando $\boldsymbol{\theta}$ es un escalar, el conjunto de confianza toma la forma de un intervalo de la recta real, y por ello se le llama *intervalo de confianza*.

3.2.1 Cantidades Pivotaes

Una *cantidad pivotal* es una función de un estadístico suficiente y un parámetro el cual tiene una distribución independiente del parámetro. En general, la cantidad pivotal sólo existe para variables aleatorias continuas y debe ser una función monótona de $\boldsymbol{\theta}$. Las cantidades pivotaes solo existen para funciones de distribución de probabilidad que pertenecen a la familia de transformación, y en particular para la familia de escala y localidad¹ (Lindsey, 1996).

¹La familia de distribuciones de escala y localidad se define por

$$f(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}; 0, 1\right),$$

donde $f(\cdot)$ es la función de densidad correspondiente. En esta familia, los dos parámetros corresponden a cambios en localidad y escala de las mediciones.

Las cantidades pivotaes pueden usarse para construir conjuntos de confianza. La idea básica es que la distribución de la cantidad pivotal g se usa para escribir la probabilidad del siguiente evento

$$\Pr\{g_1 \leq g \leq g_2\} = 1 - \alpha,$$

donde $1 - \alpha$ es el *coeficiente de confianza*. Para el escalar θ , y dada la monoticidad de g , las desigualdades pueden ser despejadas llegando a:

$$\Pr\{\theta_1(\mathbf{X}) \leq \theta \leq \theta_2(\mathbf{X})\} = 1 - \alpha.$$

De esta forma, el intervalo $[\theta_1(\mathbf{X}), \theta_2(\mathbf{X})]$ representa un intervalo de confianza.

Cuando se tiene una distribución que no pertenece a la familia de distribuciones de escala y localidad, se pueden usar algunas cantidades pivotaes aproximadas. Es posible verificar que las cantidades en las ecuaciones (2) y (4) son cantidades pivotaes aproximadas. Otra cantidad pivotal se puede obtener a partir del estadístico del cociente de log-verosimilitud definido en la ecuación (6). En este caso, se tiene que la región toma la forma aproximada $D_n(\boldsymbol{\theta}) \leq \chi_{p,1-\alpha}^2$ o en forma equivalente:

$$l_n(\boldsymbol{\theta}) \geq l_n(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{p,1-\alpha}^2.$$

Así, un intervalo de confianza para $\boldsymbol{\theta}$, con un coeficiente de confianza aproximado de $1 - \alpha$, consiste de todos los valores para los cuales $l_n(\boldsymbol{\theta}) \geq l_n(\hat{\boldsymbol{\theta}}) - \frac{1}{2} \chi_{p,1-\alpha}^2$. En otras palabras, todos los valores de $\boldsymbol{\theta}$ los cuales tienen una log-verosimilitud dentro de la banda de $\frac{1}{2} \chi_{p,1-\alpha}^2$ del máximo de la log-verosimilitud están incluidos en el intervalo de confianza.

4 El Uso de Información Concomitante

4.1 Modelado de variables explicativas

En general, el proceso de inferencia abarca el estudio de de una variable respuesta Y y varias variables explicativas $\mathbf{z}^T = (1, z_1, z_2, \dots, z_q)$. El conocimiento del contexto en que los datos se obtuvieron puede usarse para diseñar el modelo. De esta forma, el modelo tiene entonces dos componentes: 1) La distribución de probabilidad de Y , e.g. $Y \sim N(\mu, \sigma^2)$, y 2) Una ecuación que liga el valor esperado de Y , o el valor de algún parámetro de importancia, con una función de las variables explicativas, típicamente esta función es una combinación lineal; e.g. $E[Y] = \alpha + \beta z$ o $\ln E[Y] = \beta_0 + \beta_1 \text{sen}(\alpha z)$.

Cuando la distribución de Y pertenece a la familia exponencial de distribuciones de probabilidad y la ecuación del segundo componente tiene la siguiente forma general:

$$g(E[Y]) = \mathbf{z}^T \boldsymbol{\beta}, \quad \boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_q), \quad (7)$$

entonces se tiene un *modelo lineal generalizado*; aquí, $\mathbf{z}^T \boldsymbol{\beta}$ se conoce como el *componente lineal*, $\boldsymbol{\beta}$ es el vector de parámetros o los coeficientes de regresión, y la función g es llamada la *función liga*. Por ejemplo, si $g(a) = a$, que es la *liga identidad*, y si $Y \sim N(\mu, \sigma^2)$, entonces se tiene el bien conocido modelo de regresión lineal. En general es posible usar la misma metodología para incluir información covariada a respuestas cuya función de distribución no sea necesariamente un miembro de la familia exponencial. De hecho, el segundo componente no necesariamente debe incluir una sola liga en un sólo parámetro pues es posible modelar los diferentes componentes de la distribución de Y , estos modelos son conocidos como *modelos no lineales*.

4.2 Pruebas de Hipótesis

Cuando se tiene información covariada, ciertas pruebas de hipótesis sobre los parámetros son útiles para medir la utilidad del modelo. Es posible establecer una prueba para determinar si existe una relación entre la variable respuesta Y y un conjunto de variables explicativas. Esta prueba de hipótesis es la *prueba de significancia*:

$$H_0 : \boldsymbol{\beta}_{-1} = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_{-1} \neq \mathbf{0},$$

donde $\boldsymbol{\beta}_{-1} = (\beta_1, \dots, \beta_q)$. El rechazo de H_0 indica que al menos uno de los coeficientes de las variables z_1, \dots, z_q contribuye significativamente al modelo. El procedimiento de prueba, en general, hace uso del estadístico desviación; en muy pocos casos es posible encontrar un estadístico de prueba exacto.

Frecuentemente, un investigador puede estar interesado en probar hipótesis acerca de coeficientes individuales. Tales pruebas pueden ser útiles para determinar el valor de cada una de las variables regresoras. Por ejemplo, el modelo puede ser más efectivo con la inclusión de variables adicionales, o tal vez con la eliminación de una o más variables en el modelo.

La hipótesis para probar la significancia de cualquier coeficiente de regresión individual β_j es

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Si H_0 no se rechaza, entonces esto indica que z_j se puede quitar del modelo. En general, el estadístico de prueba que se usa es el estadístico de Wald; de esta forma, la hipótesis nula H_0 se rechaza al nivel de significancia α si $|Z| > z_{1-\alpha/2}$. Nótese que esta es en realidad una prueba marginal pues el coeficiente $\hat{\beta}_j$ depende de todas las variables explicativas z_j que se encuentran en el modelo.

4.3 Modelos Lineales Generalizados

Sean Y_1, Y_2, \dots, Y_n variables aleatorias independientes tales que su función de densidad puede expresarse como:

$$f(y_i; \theta_i, \phi) = \exp \{ [y_i \theta_i - b(\theta_i)] / a_i(\phi) + c(y_i, \phi) \}, \quad (8)$$

con funciones especificadas $a_i(\cdot)$, $b(\cdot)$, y $c(\cdot)$. Si ϕ es conocido entonces se tiene un *modelo lineal de la familia exponencial* con *parámetros canónicos* θ_i , $i = 1, 2, \dots, n$. Nótese que la densidad en la ecuación (8) es una reparametrización de la densidad en la Ecuación (1) que corresponde a la familia exponencial de distribuciones.

Es común encontrarse con $a_i(\phi) = a_i \phi$, donde a_i es conocido y ϕ es llamado el *parámetro de dispersión* o de *escala*. Por ejemplo, cuando cada Y_i es la media de n_i variables aleatorias independientes distribuidas normalmente y con varianza constante σ^2 , entonces $a_i(\phi) = \sigma^2 / n_i$; i.e. $\phi = \sigma^2$ y $a_i = 1/n_i$.

El *modelo lineal generalizado* (MLG) se define cuando las Y_i se distribuyen con función de densidad como en la ecuación (8) y cuando la función liga tiene la forma presentada en la ecuación (7) (e.g McCullagh y Nelder, 1989). El MLG tal vez no es muy “generalizado” en el contexto más amplio. El MLG supone que ϕ se mantiene constante y no permite que $\mathbf{z}^T \boldsymbol{\beta}$ se reemplace por una función de $\boldsymbol{\beta}$. A pesar de estas limitaciones, el MLG forma una clase de modelos muy útil y versátil.

Un ejemplo de un MLG es cuando se considera a Y_1, Y_2, \dots, Y_n variables aleatorias independientes binomiales tales que $Y_i \sim \text{BIN}(n_i, \pi_i)$; i.e.

$$\Pr\{Y_i = y_i\} = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \quad i = 1, 2, \dots, n, \quad \text{y} \quad E[Y_i] = \mu_i = n_i \pi_i.$$

Tales variables aleatorias pueden ser observadas en un conjunto de realizaciones para evaluar la efectividad de un insecticida con n dosis d_1, d_2, \dots, d_n . En la i -ésima realización n_i insectos se exponen a una dosis d_i y entonces se procede a registrar el número de *insectos* que “responden”. La probabilidad de observar una respuesta es π_i .

La función liga más usada con tales datos es la *logit* de μ_i , i.e.

$$g(\mu_i) = \ln \left(\frac{\mu_i}{n_i - \mu_i} \right) = \ln \left(\frac{\pi_i}{1 - \pi_i} \right),$$

la cual corresponde a la función de distribución de una distribución logística. Una razón de la popularidad de esta liga es que con la distribución binomial es que es la única que admite estadísticos suficientes para los parámetros en el modelo lineal.

Otras funciones liga que podrían ser consideradas son la *probit* y la *log-log-complementaria* (l-l-c). La liga probit se puede expresar en términos de π_i como $g(\pi_i) = \Phi^{-1}(\pi_i)$, donde Φ

es la función de distribución normal estándar; la liga l-l-c es $g(\pi_i) = \ln[-\ln(1 - \pi_i)]$, la cual se deriva de la distribución de los valores extremos.

Como la binomial pertenece a la familia exponencial entonces se tiene un MLG. Usando la liga logit, se puede considerar ajustar las diferentes cantidades de dosis con una línea recta:

$$g(\mu_i) = g(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 d_i.$$

Un modelo algo más complicado podría definir una regresión polinomial. Si se define $\mathbf{z}_i^T = (1, d_i, d_i^2, \dots, d_i^q)$, entonces este modelo se puede escribir como

$$g(\pi_i) = \sum_{k=0}^q \beta_k z_{ik} = \mathbf{z}_i^T \boldsymbol{\beta},$$

donde $z_{i0} = 1$ es la variable que corresponde a la ordenada β_0 . Así al aplicar la transformación inversa se obtiene que:

$$\pi_i = \frac{\exp\{\mathbf{z}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{z}_i^T \boldsymbol{\beta}\}}.$$

Es muy conveniente estimar los parámetros dentro del modelo usando máxima verosimilitud. La función log-verosimilitud en el marco general del MLG se puede expresar como:

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{[y_i \theta_i - b(\theta_i)]}{a_i(\phi)} + \sum_{i=1}^n c(y_i, \phi).$$

Para obtener los estimadores de máxima verosimilitud de β_j , $j = 0, 1, \dots, q$, se debe maximizar esta función. Para esto, se necesita resolver el sistema de ecuaciones formado por

$$\frac{\partial l_n(\boldsymbol{\theta})}{\partial \beta_j} = 0, \quad j = 0, 1, \dots, q.$$

Este proceso se puede culminar usando algún método numérico tal como el Newton-Raphson. Es importante mencionar que, a menos de que sea conocido, es necesario estimar ϕ para obtener los estimadores de $\boldsymbol{\beta}$.

5 Ilustración: Riesgos Contendientes

Para ilustrar el proceso de inferencia estadística, considérense los datos del cáncer de próstata publicados por Andrews y Herzberg (1985) y analizados recientemente por Escarela y Carrière (2003). La meta del análisis es la de comparar los niveles de diethylstilbestrol (DES), una medicina para tratar el cáncer de próstata, con respecto al tiempo de supervivencia de los pacientes. Como resultado de efectos secundarios del DES, los cuales son potencialmente mortales (e.g. muerte por causas cardiovascular o algún otro

cáncer), la evaluación del verdadero beneficio no solo debe tomar en cuenta el tiempo de vida hasta sufrir una muerte de cáncer de próstata, sino que también debe considerar el tiempo de vida hasta sufrir una muerte de otra naturaleza. Además, el estudio debe incluir la búsqueda de interacciones tratamiento \times covariable importantes las cuales pueden llevar a la definición de subconjuntos de pacientes en los cuales las diferencias del tratamiento son significativamente más marcadas o incluso revertidas.

En este estudio, los tipos de muerte se categorizan como cáncer de próstata y “otro”. Dentro de los $n = 483$ pacientes con datos completos, hubo 125 (26%) muertes de cáncer de próstata, 219 de “otras” causas (45%), y 139 observaciones censuradas (26%), i.e. 139 individuos seguían vivos al final del estudio. En este análisis se consideran las siguientes covariables: **RX**, el tratamiento (0=Placebo y 0.2 mg. estrógeno, el grupo de “dosis baja”, y 1=1.0 y 5.0 mg. estrógeno, el grupo de “dosis alta”); edad del paciente a la fecha del diagnóstico; **wt**, peso estandarizado (peso en Kg. $-$ estatura en cm. $+ 200$); **PF**, actividad diaria (0=actividad normal y 1= en cama al menos 50% del tiempo); **HX**, historia de condición cardiovascular (0=no, 1=si); **hg**, haemoglobina en $\mu\text{g}/100$ ml; **sz**, tamaño de la lesión estimada en cm^2 ; y **sg**, índice combinado de la etapa del tumor y del grado histológico.

Este tipo de datos son del tipo de *riesgos contendientes* con 2 tipos de muerte. Sea \mathbf{z}_i el vector de variables explicativas y T_{ji} el tiempo de vida a partir del diagnóstico hasta el j -ésimo tipo de muerte, donde $j = 1, 2$ e $i = 1, \dots, n$. En este caso, hacemos uso de los datos $\{X_i, c_{ij}\}$, donde $X_i = \min(T_i, C_i)$, y $T_i = \min\{T_{ij}, j = 1, 2\}$ es el *verdadero tiempo de supervivencia*, y C_i es el *tiempo de censura*; aquí, la *matriz indicadora del status* se define como $c_{ij} = I(T_i = T_{ij})$ y $\mathbf{c}_i = \sum_{j=1}^m c_{ij}$ es el vector de status de censura. Es claro que $T_{i1} \neq T_{i2}$.

El objetivo principal es el de modelar y estimar la función de supervivencia conjunta del vector aleatorio de tiempos de supervivencia de causa específica, también conocida como la *función de decremento múltiple* la cual se define como:

$$S(t_1, t_2) = \Pr\{T_1 > t_1, T_2 > t_2\}. \quad (9)$$

Para determinar la función de verosimilitud es importante definir algunas cantidades importantes. La *función de densidad cruda* se define como

$$f^{(j)}(t) = -\frac{\partial}{\partial t_j} S(t_1, t_2) \Big|_{t_1=t, t_2=t} \quad (j = 1, 2). \quad (10)$$

Esta es la función de densidad correspondiente al evento de experimentar una muerte de causa j dado que el individuo se encuentra vivo en el tiempo t . Otra cantidad importante es la función de supervivencia total definida como $S_T(t) = \Pr(T > t) = \Pr\left\{\bigcap_{j=1}^2 (T_j > t)\right\} = S(t, t)$, que es la probabilidad de que el individuo sobreviva a todos los tipos de muerte en el tiempo t .

Suponiendo que la variable de censura de un individuo i durante el experimento C_i y los tiempos de supervivencia de causa j , T_{ij} , son independientes, entonces la función de verosimilitud se puede escribir como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \left(\prod_{j=1}^2 [f^{(j)}(X_i)]^{c_{ij}} \right) [S_T(X_i)]^{1-c_i}. \quad (11)$$

Debido a que existe un problema de identificación (Tsiatis, 1975), no es posible seleccionar cualquier función de distribución bivariada para modelar la función de decremento múltiple. Sin embargo, Carrière (1995) ha demostrado que las *cóputas* forman familias de distribuciones bivariadas que resuelven adecuadamente el problema de indentificación.

Los modelos de cópula son clases de distribuciones de supervivencia bivariadas los cuales están especificados en términos de la funciones de supervivencia marginales y una función de cópula, la cual es una función de distribución bivariada continua con marginales uniformes (e.g. Nelsen, 1999). Una característica atractiva de la clase de cópulas es que la eliminación de las marginales a través de la cópula ayuda a modelar y entender la estructura de dependencia efectivamente, ya que la dependencia no tiene una relación con el comportamiento marginal de características individuales.

En el contexto de los riesgos contendientes, es posible especificar la función de decremento múltiple en términos de dos distribuciones de supervivencia marginales y una cópula que permite la inclusión de la relación de dependencia entre las variables correspondientes a cada riesgo de muerte. De esta forma, la función de distribución de supervivencia conjunta se expresa como:

$$S(t_1, t_2) = C(S_1(t_1), S_2(t_2)).$$

El análisis requiere que se escoja una clase de cópulas. En este estudio se adopta la familia de cópulas de Frank definida por

$$C_\theta(u, v) = -\frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)} \right), \quad \theta \neq 0. \quad (12)$$

El uso de la cópula de Frank es atractivo pues tiene la virtud de capturar el rango completo de dependencia. El grado de dependencia puede ser medido usando la τ de Kendall, que bajo el modelo de Frank es $\tau_\theta = 1 - \frac{4}{\theta} \left(1 - \frac{1}{\theta} \int_0^\theta \frac{t}{e^t - 1} dt \right)$.

Las funciones de densidad cruda definidas en la Ecuación (10) correspondientes al modelo de Frank se pueden escribir como:

$$f^{(j)}(t) = f_j(t) \frac{\exp\{-\theta S_j(t)\} [\exp\{-\theta S_{3-j}(t)\} - 1]}{(e^{-\theta} - 1) \exp\{-\theta S_T(t)\}} \quad (j = 1, 2), \quad (13)$$

donde $\theta \neq 0$,

$$S_T(t) = -\frac{1}{\theta} \log \left(1 + \frac{(\exp \{-\theta S_1(t)\} - 1)(\exp \{-\theta S_2(t)\} - 1)}{(e^{-\theta} - 1)} \right),$$

y f_j es la j -ésima función de densidad marginal. Cuando $\theta = 0$, $S(t_1, t_2) = S_1(t_1)S_2(t_2)$ y $f^{(j)}(t) = f_j(t)$, se forma el *modelo independiente* denotado por Π .

Para completar la construcción de la función de decremento múltiple es necesario seleccionar un modelo para las distribuciones marginales. En este estudio se considera la familia de distribuciones de supervivencia Weibull cuya función de supervivencia está especificada por $S_j(t_j) = \exp \{-(\lambda_j t_j)^{\alpha_j}\}$, donde $\lambda_j, \alpha_j > 0$. De esta forma, cuando se asigna al modelo de Frank las marginales Weibull, se obtiene una distribución Weibull bivariada con continuidad absoluta.

Para interpretar información concomitante en cada modelo marginal, se procede a usar la forma *log-lineal* del parámetro de escala λ_j :

$$\log \lambda_1 = \mathbf{b}_1^T \mathbf{u} \quad \text{y} \quad \log \lambda_2 = \mathbf{b}_2^T \mathbf{v}, \quad (14)$$

que involucra los vectores \mathbf{u} y \mathbf{v} de variables auxiliares y los vectores \mathbf{b}_1 y \mathbf{b}_2 de coeficientes de regresión. Nótese que los modelos en la ecuación (14) permiten la inclusión de diferentes conjuntos de variables para cada riesgo de muerte.

Una importante característica de este modelo de cópula con marginales específicas es que el proceso para encontrar un modelo parsimonioso sigue las ideas basadas en el cociente de verosimilitud vistas en la tercera sección de este documento. Este modelo no es lineal, así que se necesitan técnicas numéricas para encontrar el máximo de la función log-verosimilitud. Se usó la función `nllminb` del paquete estadístico S-PLUS para minimizar $-2 \times \log L_n$. A diferencia de muchos otros programas que aproximan la matriz de covarianzas, en este estudio se usa la matriz de información observada definida en la Ecuación (3); para esto, se usó el paquete matemático MAPLE para obtener la matriz Hessiano.

Para resolver la elección de variables auxiliares apropiadas para ser incluidas en la parte sistemática de cada distribución marginal de causa específica, se usó el procedimiento de *eliminación en reversa* para encontrar los mejores modelos. De esta forma se prueban diferencias del estadístico desviación de modelos anidados contra un valor crítico de $\chi_{0.95, gl=1}^2 = 3.84$.

Se procedió a ajustar los modelos de Frank e Independiente para comparar los estadísticos desviación y los coeficientes cuando se usan diferentes estructuras de dependencia y modelos marginales. Las cantidades $-2 \times \log L_n$, necesarias para calcular d_n , se muestran en la Tabla 1.

La historia de enfermedad cardiovascular no parece ser importante para el cáncer de próstata. En forma similar, ni la haemoglobina, ni el tamaño del tumor, ni el grado his-

Tabla 1: Cantidades minimizadas de $-2 \times \log L_n$ y τ de Kendall estimado para los datos de cáncer de próstata usando el modelo de cópula de Frank y el modelo independiente con marginales Weibull.

	Fórmula		Modelo	Np	$-2 \times \log L_n$	$\tau_{\hat{\theta}}$ y 95% IC
	Cáncer de Próstata	Otro				
1	Rx*(edad+Wt+PF+	Rx*(edad+Wt+PF+	Frank	35	3589.71	0.55 (0.34, 0.69)
	+Hx+Hg+Sz+Sg)	+Hx+Hg+Sz+Sg)	II	34	3593.18	0
2	Rx*(edad+Wt)+PF+	Rx*edad+Wt+PF+	Frank	24	3595.05	0.57 (0.32, 0.70)
	+Hx+Hg+Sz+Sg	+Hx+Hg+Sz+Sg	II	23	3600.37	0
3	Rx*(edad+Wt)+PF+	Rx*edad+Wt+PF+	Frank	21	3595.91	0.48 (0.31, 0.59)
	+Hx+Hg+Sz+Sg	Hx	II	20	3602.58	0
4	Rx*(edad+Wt)+Hg+	Rx*edad+Wt+Hx	Frank	18	3602.16	0.41 (0.29, 0.51)
	+Sz+Sg		II	17	3605.78	0
5	edad+Rx*Wt+Hg+	Rx+edad+Wt+Hx	Frank	16	3610.31	0.34 (0.20, 0.45)
	+Sz+Sg		II	15	3612.29	0
6	Rx+edad+Wt+Hg+	Rx+edad+Wt+Hx	Frank	15	3616.28	0.26 (0.11, 0.37)
	+Sz+Sg		II	14	3617.37	0
7	Rx	Rx	Frank	7	3819.46	-0.49 (-0.76, 0.42)
			II	6	3820.55	0
8	nulo	nulo	Frank	5	3828.84	-0.47 (-0.70, 0.15)
			II	4	3829.18	0

Np = número de parámetros

tológico parecieron ser de importancia para “otras” causas de muerte. Cuando se comparan los modelos 4 y 5 se encuentra que la interacción $\mathbf{RX} \times \mathbf{edad}$ muestra efectos significativos bajo ambas especificaciones. También se encontró que la interacción $\mathbf{RX} \times \mathbf{wt}$ es significativa sólo para el riesgo del cáncer de próstata. Todas las interacciones restantes mostraron una carencia de ajuste para ambos riesgos. La actividad diaria no mostró efectos significativos en ningún tipo de riesgo. De esta forma, usando cualquier estructura de dependencia, el modelo 4 es el más parsimonioso y es el que se escoge como el modelo final.

La Tabla 1 también muestra la τ de Kendall estimada e intervalos de confianza de 95%, basado en la aproximación $(\hat{\theta} - \theta)/\hat{e}_{\hat{\theta}} \sim N(0, 1)$, para cada modelo. Como el intervalo de confianza del modelo 4 no contiene a cero, entonces se rechaza $H_0 : \theta = 0$, y por tanto debe de incluirse este parámetro en el modelo. De esta forma, el modelo de Frank ofrece inferencias más precisas y confiables que las del independiente.

La inferencia presentada en esta ilustración ha sido completamente paramétrica, i.e. se conoce la forma de la función de decremento múltiple con la excepción de los valores

de los parámetros. En varias aplicaciones esta suposición puede ser poco realista. En la actualidad, el autor de este documento se encuentra construyendo un procedimiento de inferencia no-paramétrica para comparar y extender los resultados de los métodos aquí presentados.

Bibliografía

- Andrews, D.F.; Herzberg, A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.
- Barnett, V. (1982) *Comparative Statistical Inference* (2nd edn). London: Wiley.
- Carrière J.F. (1995) Removing cancer when it is correlated with other causes of death. *Biometrical Journal*, **37**: 339-50.
- Efron, B.F.; Hinkley, D.V. (1978) Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information. *Biometrika*, **65**, 457-487.
- Escarela, G.; Carrière, J.F. (2003) Fitting Competing Risks with an Assumed Copula. *Statistical Methods in Medical Research*, **12**:2, (aceptado para publicarse en el 2003).
- Garthwaite, P.H.; Jolliffe, I.A.; Jones, B. (1995) *Statistical Inference*. Hertfordshire: Prentice Hall.
- Lehmann, E.L. (1981) An Interpretation of Completeness and Basu's Theorem. *Journal of the American Statistical Association*, **76**, 335-340.
- Lehmann, E.L. (1986) *Testing Statistical Hypotheses* (2nd edn). New York: Wiley.
- Lindsey, J.K. (1996) *Parametric Statistical Inference*. New York: Oxford.
- McCullagh, P.; Nelder, J.A. (1989) *Generalized Linear Models* (2nd edn.) London: Chapman & Hall.
- Nelsen, R.B. (1999) *An Introduction to Copulas*. New York: Springer-Verlag.
- Tsiatis, A. (1975) A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, **72**: 20-2.
- Vu, H.T.V.; Maller, R.A.; Klass, M.J. (1996) On the Studentisation of random vectors. *Journal of Multivariate Analysis*, **57**, 142-155.
- Zacks, S. (1971) *The Theory of Statistical Inference*. New York: Wiley.