UNIVERSIDAD DE SALAMANCA
Departamento de Estadística

EL MODELO DE CLASES LATENTES EN PRESENCIA DE TABLAS POCO OCUPADAS: APLICACIÓN DEL MÉTODO BOOTSTRAP

CARLOMAGNO ARAYA ALPIZAR

Directora:

Prof. Dra. Rosa A. Sepúlveda Correa.

- 2008 -

U N I V E R S I D A D D E S A L A M A N C A
D e p a r t a m e n t o d e E s t a d í s t i c a

EL MODELO DE CLASES LATENTES EN PRESENCIA DE TABLAS POCO OCUPADAS: APLICACIÓN DEL MÉTODO BOOTSTRAP

Trabajo de investigación correspondiente al 2do año del bienio 2006-2007 del Programa de Doctorado en Estadística Multivariante Aplicada; presentado por:

Carlomagno ARAYA ALPÍZAR

Salamanca, 2008

V^{o} B^{o}

Directora

ÍNDICE

INTRODUCCIÓN 1

CAPÍTULO 1: EL MODELO DE CLASES LATENTES

1.1 Introducción 6

1.2 Planteamiento del modelo 8

1.3 Asignación de los elementos 12

1.4 Modelo con variables manifiestas binarias 13

1.5 El Modelo de clases latentes como un modelo logarítmico lineal 14

1.6 Estimación de los parámetros del modelo 15

1.7 Identificabilidad del modelo 18

1.8 Bondad de ajuste del modelo 19

1.9 Selección de modelos 23

1.10 Modelos con más de una variable latente 27

CAPÍTULO 2: EL MÉTODO BOOTSTRAP

2.1 Introducción 30

2.2 Estimación bootstrap 33

2.3 Los métodos bootstrap 38

2.4 Ejemplo de estimación bootstrap para la magnitud del error de muestreo 40

2.5 Errores en los estimadores bootstrap 44

CAPÍTULO 3: ALTERNATIVAS PARA LA SOLUCIÓN DEL PROBLEMA DE TABLAS POCOS OCUPADAS

3.1 Introducción 51

3.2 Añadir una constante a cada celda 51

3.3 Imponer restricciones a los parámetros del modelo 52

3.4 Utilización de los métodos de remuestreo 54

CAPÍTULO 4: APLICACIONES Y RESULTADOS

4.1 Datos 58

4.2 Estimación del modelo: datos originales 62

4.3 Estadísticos de bondad de ajuste 67

4.4 Pruebas de normalidad de los estadísticos

4.4.1 Modelo con una variable latente 69

4.4.2 Modelo con dos variables latentes 71

4.5 Selección del modelo

4.5.1 Bootstrap paramétrico 74

4.5.2 Bootstrap no-paramétrico 74

4.5.3 Índice de disimilaridad 76

CONCLUSIONES 77

BIBLIOGRAFÍA 82

INTRODUCCIÓN

Un problema que puede surgir en la aplicación de los modelos de clases latentes son las tablas de contingencia “poco ocupadas” (conocidas como “sparse”), esto es, tablas en las cuales muchas celdas tienen frecuencias bajas (AGRESTI, 1990); una tabla “poco ocupada” es una tabla de contingencia en la cual aproximadamente el 20 por ciento de los patrones de respuesta presentan frecuencias bajo 5. El efecto negativo que más se ha estudiado de estas, es el incumplimiento de las propiedades asintóticas de los estadísticos de bondad de ajuste y su repercusión en las pruebas de hipótesis.

Otro problema se relaciona con la no existencia de estimadores máximos verosímiles y de los errores estándar asintóticos para ciertos parámetros log-lineales. Más específicamente, las estimaciones del parámetro toman a veces valores extremadamente grandes (o infinitos). En tales casos, el algoritmo de Newton-Raphson puede no converger.

En la presente investigación se utilizan los resultados del estudio “*Factores que inciden en el consumo de drogas, población juvenil. Región Central de Occidente*” realizado en Costa Rica, noviembre de 2006. La población en estudio está formada por 13.428 jóvenes pertenecientes a 135 centros de enseñanza y 17 Equipos Básicos de Atención Integral de Salud (EBAIS). La muestra de 7.553 jóvenes se seleccionó utilizando un

muestreo de conglomerados completos (sin submuestreo) proporcionales al tamaño.

En este estudio el posible número de patrones de respuesta es de

$$\mathbf{J = m^k = 2^{13} = 8.192 \text{ patrones}}$$

Aun cuando el tamaño de muestra es muy grande (**n=7.553**), la mayor parte de los patrones no fueron observados[1], de modo que las frecuencias de la tabla de contingencia son sobre todo ceros (98%), es decir, estamos en presencia de una tabla poco ocupada.

En tal situación, las frecuencias esperadas serán extremadamente pequeñas, la regla de que deben ser mayores o iguales a 5 es violada y por tanto, la aplicación de los tests clásicos de bondad de ajuste Chi-cuadrado no es conveniente.

Los objetivos del presente trabajo son:

(1) Estudiar los problemas que surgen en el modelo de clases latentes (MCL) cuando hay muchas celdas con frecuencias bajas.

[1] De los 8.192 patrones de respuestas posibles fueron observados 163, de los cuales el 66% (107) se presentaron sólo una vez.

(2) Exponer los fundamentos teóricos básicos del método de remuestreo bootstrap y su aplicación en un MCL.

(3) Examinar la distribución de los estadísticos de bondad de ajuste, mediante el método bootstrap, bajo condiciones de de tablas poco ocupadas.

La idea del método bootstrap, introducido por Efron en 1979, es tomar muchas muestras con reemplazamiento de la muestra original, para generar variabilidad del estimador y una estimación de la distribución empírica que a su vez es un estimador de la verdadera función de distribución. Esto permitiría hacer inferencias sin tener que hacer suposiciones acerca de la distribución poblacional. El bootstrap es una técnica de estadística no-paramétrica.

CAPÍTULO 1

EL MODELO DE CLASES LATENTES

1.1 INTRODUCCIÓN

El modelo de clases latentes fue introducido por HENRY & LAZARSFELD (1968). Por otro lado, los problemas de estimación e identificación han sido tratados por ANDERSON (1954) y MCHUGH (1956). GOODMAN (1974) conectó estos modelos con la teoría moderna de las tablas de contingencia y finalmente, se puede citar a distintos autores que han desarrollado estas técnicas, como AGRESTI (1984), ANDERSEN (1991), BARTHOLOMEW (1987), CLOGG (1993), entre otros.

El modelo de clases latentes es una técnica estadística que permite estudiar la existencia de una (o varias) variable(s) latente(s) a partir de un conjunto de variables explicativas observadas, y definir a partir de sus clases una clasificación o tipología de los individuos analizados. Las relaciones de dependencia entre las variables categóricas de una tabla de contingencia en muchos casos están provocadas por la existencia de una asociación entre cada una de ellas y otra variable no observable directamente, llamada ***variable latente***.

Por lo tanto, se distinguen en el modelo dos tipos de variables. Las variables que pueden ser directamente observadas, v***ariables manifiestas***[1], conforman un vector de $\boldsymbol{p}$ componentes $\boldsymbol{X}' = (X_1, \cdots, X_p)$. Las variables latentes son representadas por $\boldsymbol{Y}$, se expresan mediante del vector $\boldsymbol{Y}' = (Y_1, \cdots, Y_q)$, donde $\boldsymbol{q} < \boldsymbol{p}$. Las variables observadas y las latentes se consideran variables categóricas con dos o más categorías.

La relación entre las variables manifiestas debe verificar el principio de *Independencia local.* Este supone que dentro de cada categoría de la variable latente, las variables observadas son estadísticamente independientes, es decir, las variables de la tabla de contingencia son condicionadamente independientes dada una clase determinada de la variable latente. Toda la asociación observada entre las variables manifiestas, está medida o explicada por las variables latentes.

El incumplimiento del principio trae como consecuencia que los estadísticos de ajuste del modelo (χ^2 ó G^2) sean demasiado grandes, los valores de las estimaciones de los parámetros del modelo se distorsionan, los errores estándar para los estimadores serán grandes, y los estimadores de las diferencias de sus varianzas serán no consistentes. Además, la falta

[1] También son llamadas variables indicadoras, salida, resultantes, endógenas o ítems.

de ajuste en un modelo de clases latentes, es causada por la violación del principio de independencia local (VERMUT & MAGIDSON, 2002).

Otro de los supuestos en un modelo de clases latentes es el que las clases latentes son ***internamente homogéneas***, es decir, todos los miembros de una clase latente tienen la misma distribución de probabilidades con respecto a la variable latente. Ésta será distinta de la distribución de probabilidades para los individuos pertenecientes a otra clase, por lo que individuos de diferentes clases presentarán características diferentes. Este hecho sirve para diferenciar a los individuos pertenecientes a diferentes clases y poder caracterizar tanto la variable latente como las clases latentes.

1.2 PLANTEAMIENTO DEL MODELO

Supongamos que se tiene una matriz conteniendo información de *p* variables categóricas sobre una muestra de *n* individuos,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

donde cada fila de la matriz, contiene el conjunto de respuestas de un individuo dado para cada una de las variables, conocido como ***respuesta*** o ***patrón de respuesta***.

Para el planteamiento teórico del modelo es suficiente considerar sólo una variable latente, ya que modelos con más de una variable latente, **q>1**, pueden ser desarrollados considerando ***q*=1** bajo adecuadas restricciones a los parámetros del modelo (GOODMAN, 1974). Por esta razón, se presentará el modelo de clases latentes considerando una única variable latente ***Y*** con ***C*** categorías o clases latentes.

Supongamos un conjunto de $\boldsymbol{p}$ variables manifiestas $\boldsymbol{X_1, \cdots, X_p}$ que se consideran indicadoras de una variable latente $\boldsymbol{Y}$; y que estas variables conforman un modelo de clases latentes con $\boldsymbol{C}$ clases o categorías. Sea $\pi_{\boldsymbol{X}}(\boldsymbol{X})$ la densidad conjunta de las variables manifiestas $\boldsymbol{X'} = (X_1, \cdots, X_p)$, donde $\boldsymbol{x'} = \left(x_1, x_2, \cdots, x_p\right)$ denota un determinado patrón de respuesta en el cual cada una de las x_i toma valores dependiendo de las categorías de la correspondiente variable manifiesta. Estas variables conforman un tabla de contingencia múltiple con $\prod_{i=1}^{p} I_i$ patrones de respuesta, tal que cada X_i contiene I_i categorías.

Por el principio de independencia local, la densidad condicional $P(\boldsymbol{X} = \boldsymbol{x}/Y = c)$ está dada por,

$$\pi_{\boldsymbol{X}/Y(c)}(x) = \prod_{i=1}^{p} \pi_{X_i/Y(c)}(x_i) \tag{1.1}$$

donde: $\pi_{X_i/Y(c)}(x_i) = P(X_i = x_i/Y = c)$

$x_i = 1, \cdots, I_i$

$c = 1, \cdots, C$

La distribución conjunta de **X** e Y está dada por,

$$\pi_{\boldsymbol{X},Y}(x, c) = \pi_Y(c)\pi_{\boldsymbol{X}/Y(c)}(x) \tag{1.2}$$

donde $\pi_y(c) = P(Y = c)$, representa la proporción de elementos que se encuentran en la clase latente ***c***, también conocida como probabilidad ***a priori***. Utilizando las expresiones anteriores, el Modelo de Clases Latentes se expresa como,

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{c=1}^{C} \pi_Y(c) \prod_{i=1}^{p} \pi_{X_i/Y_{(c)}}(x_i) \tag{1.3}$$

donde

$\pi_Y(c)$= probabilidad de la clase latente **c**, para $c = 1, \cdots, C$

$\pi_{X_i/Y(c)}(x_i)$= probabilidad de respuesta condicional de cada una de las variables manifiestas dentro de la clase latente c, para $c = 1, \cdots, C$; $i = 1, \cdots, p$; $x_i = 1, \cdots, I_i$

Los parámetros del modelo están sujetos a las siguientes restricciones:

$$(1) \quad \sum_{c=1}^{C} \pi_Y(c) = 1$$
$$(2) \quad \sum_{x_i=1}^{I_i} \pi_{x_i/Y(c)}(x_i) = 1 \tag{1.4}$$

Las probabilidades condicionales (2) son comparables a las cargas o "loadings" del análisis factorial. La expresión en (1.3) implica que la población puede ser dividida en ***C*** clases latentes exhaustivas y exclusivas, por lo tanto, la probabilidad conjunta de las variables manifiestas se obtiene sumando sobre la dimensión latente. En este sentido, esta expresión implica la existencia de la variable latente.

La representación gráfica del modelo de clases latentes descrito anteriormente se encuentra en la **Figura 1.1**. Las variables manifiestas (X_i) no están conectadas directamente y la dirección de las flechas, indica que la variable latente (Y) explica toda la posible asociación entre las variables

manifiestas. Las variables ***indicadoras*** reflejan un aspecto del concepto subyacente (latente) que estamos midiendo; los cambios en la variable latente se ven ***reflejados*** en las variables indicadoras sin que esto signifique que exista un efecto causal.

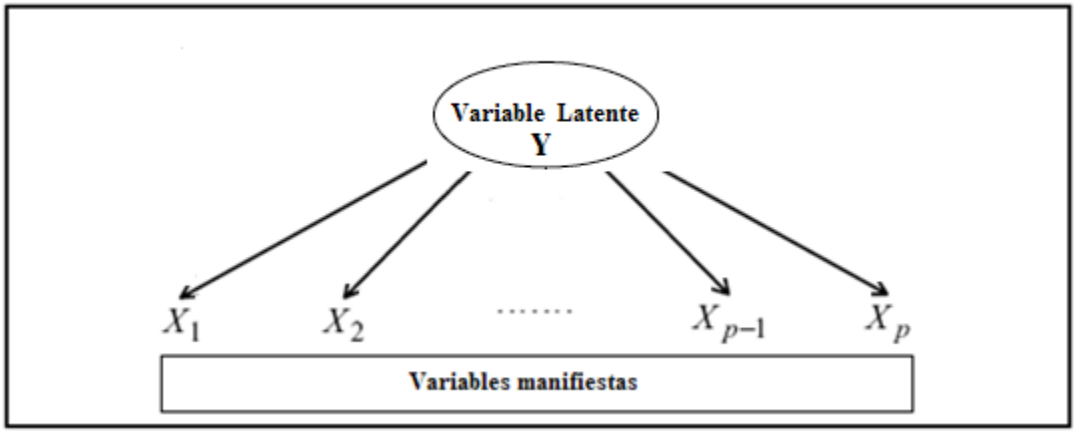

Figura 1.1. Representación gráfica de un Modelo de Clases Latentes con una variable latente y **p** variables manifiestas.

1.3 ASIGNACIÓN DE LOS ELEMENTOS

Un análisis posterior en el modelo de clases latentes está relacionado con la ubicación de los elementos que pertenecen a una clase determinada. Los elementos se clasifican dentro de la clase latente más probable. Aplicando la definición de probabilidad condicional se tiene,

$$\pi_{Y/\boldsymbol{X}(\boldsymbol{x})}(c) = \frac{\pi_{\boldsymbol{X},Y}(\boldsymbol{x}, c)}{\pi_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{\pi_Y(c)\pi_{\boldsymbol{X}/Y(c)}(\boldsymbol{x})}{\pi_{\boldsymbol{X}}(\boldsymbol{x})} \qquad (1.5)$$

En la práctica, para cada patrón de respuesta **x** se inspecciona este conjunto de probabilidades, y se asigna el individuo a la clase latente en la cual esta probabilidad es mayor.

1.4 MODELO CON VARIABLES MANIFIESTAS BINARIAS

El modelo de clases latentes se simplifica en caso que las variables manifiestas sean binarias. Solamente tenemos dos niveles de respuesta para las variables manifiestas (0,1), las probabilidades condicionales $\pi_{X_i/Y(c)}$ siguen una distribución Bernoulli y por el principio de independencia local:

$$\pi_{X_i/Y(c)}(x) = \prod_{i=1}^{p} \pi_{ic}^{x_i}(1-\pi_{ic})^{1-x_i} \tag{1.6}$$

donde: $c = 1, \cdots, C$

$x_i = 0,1$

π_{ic} = probabilidad de obtener una respuesta positiva en la variable $\boldsymbol{i}$,

Así, el modelo de clases latentes toma la forma,

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{c=1}^{C} \pi_Y(c) \prod_{i=1}^{p} \pi_{ic}^{x_i}(1-\pi_{ic})^{1-x_i} \tag{1.7}$$

y la probabilidad conjunta de que un individuo seleccionado aleatoriamente tenga patrón de respuesta **x** y esté en la clase latente $\boldsymbol{c}$ es,

$$\pi_{\boldsymbol{X},Y}(\boldsymbol{x}, c) = \pi_Y(c) \prod_{i=1}^{p} \pi_{ic}^{x_i} (1 - \pi_{ic})^{1-x_i} \qquad (1.8)$$

Los parámetros a estimar son las probabilidades de cada clase latente, $\pi_Y(c)$, y las probabilidades condicionales, π_{ic}. Se deben estimar $C - 1$ probabilidades para las clases latentes y $p \times C$ probabilidades condicionales.

1.5 EL MODELO DE CLASES LATENTES COMO UN MODELO LOGARÍTMICO LINEAL

Una alternativa de parametrización es la expresión en términos de un modelo log-lineal donde las variables manifiestas son localmente independientes dada la variable latente $\boldsymbol{Y}$. HABERMANN (1979) demostró que el modelo de clases latentes, es formalmente idéntico al modelo log-lineal jerárquico $[X_1 Y][X_2 Y] \cdots [X_p Y]$,

$$\log m_{\boldsymbol{X},Y}(x, c) = \mu + \lambda_{i_1}^{X_1} + \cdots + \lambda_{i_p}^{X_p} + \lambda_c^Y + \lambda_{i_1 c}^{X_1 Y} + \cdots + \lambda_{i_p c}^{X_p Y} \qquad (1.9)$$

donde: $i_i = 1, \cdots, I_i$ $i = 1, \cdots, p$ $c = 1, \cdots, C$

La ecuación anterior, además de la media general y los términos de una variable, contiene sólo los términos de interacción entre la variable latente ***Y*** y las variables manifiestas. Como las variables manifiestas son independientes entre si dada la clase latente, no aparecen los términos de interacción entre las variables observadas.

Utilizando el hecho de que un MCL se puede considerar como un modelo log-lineal, es posible obtener una parametrización logit para las probabilidades condicionales,

$$\pi_{X_i/Y(c)}(x_i) = \frac{exp\left(\lambda_{i_i}^{X_i} + \lambda_{i_i c}^{X_i Y}\right)}{\sum_{i_i=1}^{I_i} exp\left(\lambda_{i_i}^{X_i} + \lambda_{i_i c}^{X_i Y}\right)}; \quad i = 1, \cdots, p \qquad (1.10)$$

1.6 ESTIMACIÓN DE LOS PARAMÉTROS DEL MODELO

Para realizar las estimaciones de las probabilidades se utilizan procedimientos iterativos basados en estimaciones de máxima verosimilitud. Los algoritmos más conocidos son Newton-Raphson y el algoritmo ***EM*** [1] (HARTLEY, 1958).

[1] EM: expectación-maximización.

El algoritmo de **Newton-Raphson** es un procedimiento iterativo para la solución de sistemas de ecuaciones lineales, en el cual se utiliza la matriz de primeras derivadas parciales de la función a optimizar. En tanto, **EM** es un método que permite encontrar los estimadores máximo verosímiles de los parámetros de la distribución subyacente de un conjunto de datos, cuando los datos son incompletos o existen datos faltantes. Puede aplicarse en muchas situaciones en las que se desea estimar un conjunto de parámetros $\boldsymbol{\theta}$, dado únicamente una parte observada de los datos completos producidos por la distribución. Cada iteración del algoritmo consiste de dos pasos:

Paso 1. ***Paso de estimación (E)***. Se calculan los valores esperados dados los observados para estimar la distribución de probabilidad (o parámetros del modelo).

Paso 2. ***Paso de maximización (M)***. Se maximiza la función de verosimilitud de los datos a partir de los valores esperados.

En un MCL los "*datos faltantes*" son las clases donde pertenecen los individuos y los "estimadores" de estos valores son las probabilidades ***a posteriori***,

$$\pi_{Y/\boldsymbol{X}(\boldsymbol{x})}(c) = P(Y = c/X = x) \qquad (1.11)$$

La estimación de las probabilidades constituye el paso **E** del algoritmo y la inserción de estos valores para obtener los estimadores mejorados constituyen el paso **M**.

El algoritmo ***EM*** es sencillo tanto en la teoría como en el cálculo, y generalmente los valores iniciales elegidos aleatoriamente son suficientes para llegar a una solución. El número de iteraciones necesarias para la convergencia del método será menor si estos valores se encuentran "cercanos" a los alcanzados por los estimadores máximo verosímiles (EMV).

En relación con el criterio de convergencia para detener el proceso, se puede considerar detener el proceso cuando la distancia euclídea entre sucesivos estimadores del vector de parámetros $\pi_{Y/\boldsymbol{X}(\boldsymbol{x})}(c)$ se menor a un valor determinado. Otro criterio que puede considerarse es que el cambio en el logaritmo de verosimilitud, $\mathcal{L}$, entre sucesivas iteraciones sea menor a un valor dado (SEPÚLVEDA, 2004).

1.7 IDENTIFICABILIDAD DEL MODELO

Según GOODMAN (1974), desde una perspectiva matemática, una condición suficiente para la identificabilidad del modelo de clases latentes es que la matriz de covarianzas teóricas de las estimaciones máximo-verosímiles sea positiva, es decir, que no haya colinealidad entre los parámetros.

Una manera de contrastar si un modelo es identificable con el algoritmo ***EM*** consiste en estimar el modelo con diferentes valores iniciales. Si con esos valores iniciales distintos, el modelo proporciona el mismo valor del logaritmo de la función de verosimilitud pero distintas estimaciones de los parámetros, el modelo es no identificable.

Un modelo es identificable cuando existe una solución única, y no identificable cuando tiene infinitas soluciones. Hay dos tipos de no identificabilidad: *intrínseca* y *empírica.* La no identificabilidad *intrínseca* no depende de los datos, sino del modelo en sí, y es común en modelos con muchas variables latentes.

La no identificabilidad *empírica* es la que ocurre con ciertas estructuras de los datos, es decir, el mismo modelo con otros datos puede

ser identificable. En general, la no identificabilidad empírica está relacionada con muestras pequeñas y tablas poco ocupadas.

El caso más simple de un modelo no identificable intrínsecamente es cuando está mal especificado, por ejemplo, muchas clases latentes pueden causar el problema, pues el máximo número de parámetros estimables está limitado por el número de grados de libertad disponibles.

Una condición ***necesaria*** para que un modelo sea identificable es que el número de parámetros independientes (a ser estimados) no exceda el número de frecuencias observadas, o equivalentemente, que los grados de libertad sean no negativos (SEPÚLVEDA, 2004).

1.8 BONDAD DE AJUSTE DEL MODELO

La calidad del ajuste de un modelo log-lineal concreto, en particular el modelo de clases latentes, puede determinarse con la comparación de las frecuencias observadas para cada patrón de respuesta, $\boldsymbol{f_x}$, con las frecuencias estimadas, $\hat{\boldsymbol{f}}_{\boldsymbol{x}}$, mediante el contraste Chi-cuadrado de Pearson o la razón de verosimilitud $\mathbf{G^2}$**.** Las frecuencias esperadas están dadas por la expresión:

$$\hat{f}_x = n \times \left[\sum_{c=1}^{C} \hat{\pi}_Y (c) \prod_{i=1}^{p} \hat{\pi}_{X_i/Y(c)}(x_i) \right] \quad (1.12)$$

Luego, se calcula $\boldsymbol{\chi^2}$ y $\boldsymbol{G^2}$ como,

$$\chi^2 = \sum_x \frac{\left(f_x - \hat{f}_x\right)^2}{\hat{f}_x} \quad (1.13)$$

$$G^2 = 2 \sum_x f_x ln \left(\frac{f_x}{\hat{f}_x} \right) \quad (1.14)$$

Si el modelo es válido para la población, ambos estadísticos siguen asintóticamente una distribución Chi-cuadrado. Para cada modelo el número de grados de libertad (*gl.*) de la distribución se obtiene a partir de la diferencia del número de celdas menos el número de parámetros independientes, o de igual forma,

$$gl = \prod_{i=1}^{p} I_i - C \left[\sum_{i=1}^{p} (I_i - 1) + 1 \right] \quad (1.15)$$

Si el rango de la matriz de varianzas y covarianzas es menor a ***m,*** el número de parámetros independientes a estimar en el modelo, entonces los grados de libertad deben ser restringidos para que no se dé el caso de un modelo no identificable.

Si algunas frecuencias esperadas estimadas son ceros estructurales o no pueden calcularse algunos parámetros al existir ceros en algunos estadísticos suficientes, CLOGG & GOODMAN (1984) mostraron que el número de grados de libertad pasaría a ser: gl. = *número de celdas sin ceros - número de parámetros estimables.*

El estadístico $\boldsymbol{G^2}$ tiene una ventaja sobre el de Pearson porque puede descomponerse en distintas componentes referidas a diferentes efectos, submodelos o subgrupos. Esta propiedad es muy interesante cuando buscamos un modelo que se ajuste bien y, simultáneamente, sea reducido.

Otro estadístico de bondad de ajuste es el estadístico de Freeman-Tukey (BISHOP, FIENBERG & HOLLAND, 1975) definido como:

$$FT^2 = 4\sum_x \left(\sqrt{f_i} - \sqrt{\hat{f}_i}\right)^2 \qquad (1.16)$$

Cuando se tienen frecuencias esperadas menores a 5, otra alternativa a los test anteriores, es presentada por READ & CRESSIE (1988). Ellos proponen la utilización de una versión generalizada del estadístico Chi-cuadrado. La forma general del estadístico Read-Cressie es,

$$CR(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_x f_x \times \left[\left(\frac{f_x}{\hat{f}_x} \right)^{\lambda} - 1 \right] \qquad (1.17)$$

Dependiendo del valor de $\boldsymbol{\lambda}$ este estadístico toma diferentes formas (***Tabla 1.1***). Si $\boldsymbol{\lambda = 0}$, se obtiene el estadístico $\mathbf{G^2}$; si $\boldsymbol{\lambda = 1}$ se tiene el estadístico $\boldsymbol{\chi^2}$ y cuando $\lambda = -1/2$ resulta la estadística de Freeman Tukey. Read y Cressie recomiendan considerar $\boldsymbol{\lambda = 2/3}$, ya que cuando se trabaja con "grandes" tablas de datos se obtiene un estadístico más apropiado que $\boldsymbol{\chi^2}$ o $\mathbf{G^2}$ (VERMUNT, 1997a).

Estadístico	Cálculo	λ
$\boldsymbol{\chi^2}$ =	$\sum_x \frac{(f_x - \hat{f}_x)^2}{\hat{f}_x}$ =	$CR(1)$
CR =	$\frac{2}{\lambda(\lambda+1)} \sum_x f_x \times \left[\left(\frac{f_x}{\hat{f}_x} \right)^{\lambda} - 1 \right]$ =	$CR\left(\frac{2}{3}\right)$
$\boldsymbol{G^2}$ =	$2 \sum_x f_x ln\left(\frac{f_x}{\hat{f}_x} \right)$ =	$CR(0)$
$\boldsymbol{FT^2}$ =	$4 \sum_x \left(\sqrt{f_i} - \sqrt{\hat{f}_i} \right)^2$ =	$CR\left(-\frac{1}{2}\right)$

Tabla 1.1. Estadísticos de bondad de ajuste y su cálculo en términos del estadístico Read-Cressie.

1.9 SELECCIÓN DE MODELOS

El objetivo es encontrar el mejor modelo que explique las relaciones existentes entre las variables en la población que generan los datos observados. Por tanto, los errores posibles al seleccionar un modelo se producirán cuando éste contenga más parámetros de los necesarios o se excluyan algunos parámetros que forman parte del mejor modelo.

En el proceso lógico de la modelización estadística, se parte de unas hipótesis o supuestos *a priori* que se reflejan en una formulación determinada del modelo. Dichas hipótesis naturalmente deben basarse en las ideas que el investigador tenga sobre las relaciones existentes entre las variables en la población, es decir, es conveniente utilizar los conceptos teóricos relacionados con el problema que intentamos resolver.

Estadísticamente, pueden existir cientos de modelos para un solo conjunto de datos, que se ajusten con la misma calidad. Si no seguimos la orientación proporcionada por el problema teórico que queremos resolver, es difícil dilucidar qué modelo elegir. Si los supuestos de partida llevan a un único modelo log-lineal no saturado, el proceso es fácil, dado que se limitaría a la aplicación de los estadísticos mostrados en la ecuación (1.13) y (1.14). Sin embargo, como hemos expuesto anteriormente, la dificultad

comienza a la hora de descubrir cuál es el mejor modelo dentro de una gama.

Si los modelos están anidados jerárquicamente, pueden utilizarse contrastes de G^2 condicionados. Dos modelos están anidados jerárquicamente cuando el modelo restringido contiene sólo un subconjunto de los efectos presentes en el modelo libre. Este estadístico condicionado sigue una distribución Chi-cuadrado si el modelo libre es válido, la muestra es grande y la aproximación es buena. Incluso en aquellas situaciones como las muestras pequeñas, en que el contraste no condicionado tiene problemas (HABERMAN, 1978).

A partir de la teoría de la información, es posible desarrollar una forma de seleccionar el modelo más adecuado. El objetivo no es descubrir el modelo "verdadero", sino aquel que proporciona mayor información sobre la realidad. Por un lado, las frecuencias esperadas estimadas deben ser parecidas a las observadas y, por otro, el modelo debe ser tan reducido como sea posible. La idea principal es que, dados dos modelos con igual valor en la función verosimilitud, el mejor modelo es el que tiene el menor número de parámetros.

Los contrastes más conocidos basados en la teoría de la información son el *criterio de información de Akaike* (***AIC***) (AKAIKE, 1987) y el *criterio de información bayesiano* (***BIC***) (RAFTERY, 1986). El primero, penalizando al modelo según su grado de complejidad, determina hasta qué punto un modelo concreto se desvía de la realidad. Su expresión es:

$$AIC = -2\ln(\ell) + 2m \qquad (1.18)$$

donde $\boldsymbol{\ell}$ representa el valor de la función de verosimilitud y $\boldsymbol{m}$ el número de parámetros desconocidos.

SCHWARZ (1978) utiliza el criterio bayesiano para desarrollar una medida consistente asintóticamente (***BIC***), basada en el logaritmo de la función de verosimilitud ($\boldsymbol{\ell}$), el número de parámetros independientes a ser estimados ($\boldsymbol{m}$) y el tamaño muestral.

$$BIC = -2\ln(\ell) + m \times \ln(n) \qquad (1.19)$$

Cuanto menores sean los valores de los estadísticos mejor será el modelo, porque mayor información contendrá. Ambos criterios pueden calcularse a partir del estadístico $\boldsymbol{G^2}$ de la siguiente forma:

$$AIC^* = G^2 - 2v$$
$$BIC^* = G^2 - v \times \ln(n) \tag{1.20}$$

donde ***v*** son los correspondientes grados de libertad.

Por tanto, y refiriéndonos al ***BIC***, al ser el más adecuado en los modelos log-lineales, se podrá calcular a partir del estadístico $\mathbf{G^2}$ no condicionado. Un valor negativo indica que el modelo es preferible al modelo saturado y además, debe elegirse aquel modelo con el menor valor. Este criterio elimina los problemas del ajuste por exceso y por defecto.

Un modelo que no se ajuste bien a las frecuencias observadas tendrá un $\boldsymbol{G^2}$ elevado. Por otra parte, si un modelo se ajusta muy bien porque posee un gran número de parámetros, al tener una cantidad muy pequeña de grados de libertad, el valor del criterio será muy elevado.

Según ANDRADE, FAJARDO, PÉREZ y CORRALES (2002), cuando los tamaños de muestra son grandes, $\boldsymbol{\chi^2}$ y $\mathbf{G^2}$ tienden a rechazar modelos no saturados, debido a que todos los efectos del modelo saturado son significativos. En estas condiciones, es preferible utilizar el índice ***BIC***.

Investigaciones empíricas en modelos de clases latentes (LIN & DAYTON, 1997) sugieren que se debe utilizar el ***AIC*** cuando los tamaños de muestra sean pequeños o los modelos estimados tenga poco parámetros.

1.10 MODELOS CON MÁS DE UNA VARIABLE LATENTE

La especificación de los modelos de clases latentes con más de una variable latente se propuso por GOODMAN (1974) y HABERMAN (1979). Supongamos un modelo con dos variables latentes Y_1 e Y_2, cada una con dos niveles de respuesta, $r, s = 1,2$. Se considerará que las manifiestas, X_1 y X_2 son indicadoras de la variable latente Y_1, y las variables X_3 y X_4 son indicadoras de la variable latente Y_2. Además, supongamos que ambas variables latentes están relacionadas como aparece en la **Figura 1.2**. La expresión de dicho modelo será la siguiente,

$$\pi_X(X) = \sum_{r,s=1}^{2} \pi_{Y_1,Y_2}(r,s) \prod_{i=1}^{4} \pi_{X_i/Y_1(r),Y_2(s)}(x_i) \qquad (1.21)$$

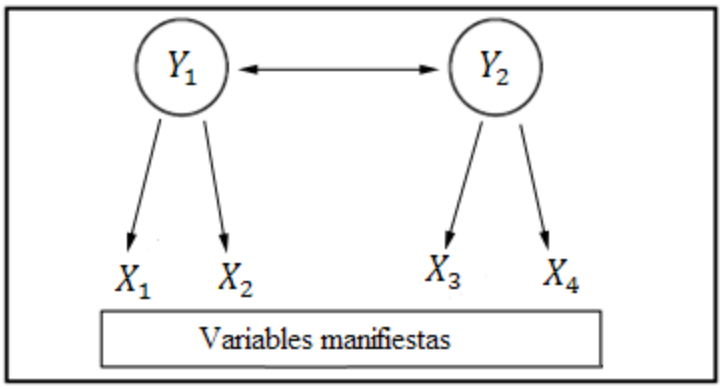

Figura 1.2. Modelo de Clases Latentes con dos variables latentes y cuatro variables manifiestas.

El modelo expresado en la **Figura 1.2**, se corresponde con el modelo log-lineal $[Y_1Y_2][X_1Y_1][X_2Y_1][X_3Y_2][X_4Y_2]$, es decir,

$$\log \mathrm{m}_{\mathbf{X},\mathrm{Y}_1,\mathrm{Y}_2}(\mathbf{X},\mathrm{r},\mathrm{s}) = \mu + \lambda_{\mathrm{i}_1}^{\mathrm{X}_1} + \cdots + \lambda_{\mathrm{i}_4}^{\mathrm{X}_4} + \lambda_{\mathrm{r}}^{\mathrm{Y}_1} + \lambda_{\mathrm{s}}^{\mathrm{Y}_2} + \lambda_{\mathrm{i}_1\mathrm{r}}^{\mathrm{X}_1\mathrm{Y}_1} + \lambda_{\mathrm{i}_2\mathrm{r}}^{\mathrm{X}_2\mathrm{Y}_1} + \lambda_{\mathrm{i}_3\mathrm{s}}^{\mathrm{X}_3\mathrm{Y}_2} + \lambda_{\mathrm{i}_4\mathrm{s}}^{\mathrm{X}_4\mathrm{Y}_2} + \lambda_{\mathrm{rs}}^{\mathrm{Y}_1\mathrm{Y}_2} \tag{1.22}$$

donde: $i_p = 1, \cdots, I_p$ $p = 1, \cdots, 4$ $r = 1,2$ $s = 1,2$

Es posible imponer distintos tipos de restricciones sobre los parámetros del modelo log-lineal o sobre las probabilidades condicionales Además, el modelo puede ser visto como un modelo con una única variable latente **Y**, formada por cuatro clases latentes, que son la combinación de los niveles de las variables latentes Y_1 y Y_2. Sin embargo, esta alternativa de parametrización, tiene la desventaja que no impone restricciones acerca de las relaciones entre las variables latentes.

CAPÍTULO 2

EL MÉTODO BOOTSTRAP

2.1 INTRODUCCIÓN

En 1979, BRADLEY EFRON (EFRON,1979) desarrolla y publica el análisis formal del *método bootstrap*, término que procede de la expresión inglesa "*to pull oneself up by one's bootstrap*" (que podría traducirse por: *levantarse mediante el propio esfuerzo*), tomada de una de las aventuras del Barón Munchausen, personaje del siglo XVIII creado por el escritor Rudolph Erich Raspe, en la cual el barón había caído al fondo de un lago profundo y, cuando creía que todo estaba perdido, tuvo la idea de ir subiendo tirando hacia arriba de los cordones (*bootstrap*) de sus propias botas.

El *método bootstrap* se basa en la analogía entre la muestra y la población de la cual la muestra es extraída. De acuerdo con EFRON y TIBSHIRANI (1986), dada una muestra con ***n*** observaciones el estimador no paramétrico de máxima verosimilitud de la distribución poblacional es la función de densidad de probabilidad que asigna una masa de probabilidad ***1/n*** a cada una de las observaciones. La idea central es que muchas veces puede ser mejor extraer conclusiones sobre las características de la población a partir de los datos obtenidos en la muestra, que haciendo supuestos poco realistas sobre la población.

La esencia del *método bootstrap* consiste en que en ausencia de otra información, los valores de una muestra aleatoria son la mejor representación de la distribución de la población y remuestrar la muestra nos proporciona la mejor información sobre lo que sucedería si remuestrearamos la población (EFRON & TIBSHIRANI, 1993; MANLY, 1997).

Los procedimientos basados en los *métodos bootstrap* implican obviar los supuestos sobre la distribución teórica que siguen los estadísticos. En su lugar, la distribución del estadístico se determina simulando un número elevado de muestras aleatorias construidas directamente a partir de los datos observados. Es decir, utilizamos la muestra original para generar a partir de ella nuevas muestras que sirvan de base para estimar inductivamente la forma de la distribución muestral de los estadísticos, en lugar de partir de una distribución teórica asumida ***a priori***.

Este enfoque tiene su antecedente inmediato en las técnicas de simulación Monte Carlo, las cuales consisten en extraer un número elevado de muestras aleatorias de una población conocida, para calcular a partir de ellas el valor del estadístico cuya distribución muestral pretende ser estimada.

Sin embargo, en la práctica no solemos conocer la población y lo que manejamos es una muestra extraída de ella. El investigador parte de un conjunto de datos observados, que constituyen una muestra extraída de la población que pretende estudiar. Cuando las técnicas Monte Carlo son aplicadas a la resolución de problemas estadísticos, partiendo de datos observados en una muestra, reciben más apropiadamente la denominación de «*técnicas de remuestreo*».

El *método bootstrap* es simple y directo para calcular los sesgos aproximados, desviaciones estándar, intervalos de confianza, etc., en casi cualquier problema de estimación no paramétrico. Debido a que el sustento teórico matemático-estadístico del *bootstrap* es bastante complejo, hasta finales de la década del ’80 del siglo pasado, la eficiencia del método era probada de manera empírica, es decir, en el terreno de la práctica.

Los procedimientos de remuestreo en general, han comenzado a centrar la atención de los estadísticos a partir de la década de los ochenta, cuando el desarrollo de la informática allanó los obstáculos prácticos unidos a la simulación de un número elevado de muestras. A finales de esta década, la utilización del método *bootstrap* para el contraste de hipótesis empezaba a ser

considerada una alternativa a los tests paramétricos y no paramétricos convencionales (NOREEN, 1989).

2.2 ESTIMACIÓN *BOOTSTRAP*

La idea básica, en síntesis, es tratar la(s) muestra(s) como si fuera la población, (debido a la analogía entre muestra y población) y a partir de ella extraer con reposición un gran número de muestras de tamaño ***n***. Así, aunque cada "remuestra" tendrá el mismo número de elementos que la muestra original, mediante el remuestreo con reposición cada una podría incluir algunos de los datos originales más de una vez.

Como resultado cada remuestra, será muy probablemente, algo diferente de la muestra original; con lo cual, un estadístico $\widehat{\boldsymbol{\theta}}^*$, calculado a partir de una de esas remuestras tomará un valor diferente del que produce otra remuestra y del $\widehat{\boldsymbol{\theta}}$ observado. La afirmación fundamental del *bootstrap* es que una distribución de frecuencias de esos $\widehat{\boldsymbol{\theta}}^*$ calculados a partir de las remuestras es una estimación de la distribución muestral de $\widehat{\boldsymbol{\theta}}$ (MOONEY & DUVAL, 1993).

Sea $\mathbf{X} = (x_1, x_2, \dots, x_n)$ una muestra aleatoria de tamaño $\boldsymbol{n}$, se designa con $F(\boldsymbol{x}) = Pr(\boldsymbol{X} \leq \boldsymbol{x})$ a la función de distribución común de las variables aleatorias x_i, lo cual en forma simbólica se escribe $(x_1, x_2, \dots, x_n) \sim F(\boldsymbol{x})$ o simplemente, $x_i \sim F(x)$. Cuando el valor del parámetro $\boldsymbol{\theta}$ de una población es desconocido y, en consecuencia, se desea utilizar un estimador $\widehat{\boldsymbol{\theta}} = \mathrm{f}(\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_n, \theta)$ del mismo, la distribución de $\widehat{\boldsymbol{\theta}}$ es aproximada generando una muestra de los resultados independientes $\boldsymbol{\theta}_\mathbf{j}^*$ para $\mathbf{j} = \mathbf{1}, \mathbf{2}, \dots, \mathbf{b}$ y construyendo la distribución empírica $\boldsymbol{F}_{\widehat{\boldsymbol{\theta}}}$.

La ***Figura 2.1***[1] muestra el proceso de estimación de estadísticos *bootstrap* a partir de una muestra. Consideremos que disponemos de una muestra (**X**) a partir de la cual calculamos un estadístico de interés que estime algún parámetro poblacional (f(**X**)).

Las propiedades estadísticas de este estimador pueden ser calculadas mediante remuestreo *bootstrap* de la siguiente manera:

[1] El esquema se basa en la ***Figura 2.2,*** página 32 de la Tesis de Doctoral de CERVIÑO (2004).

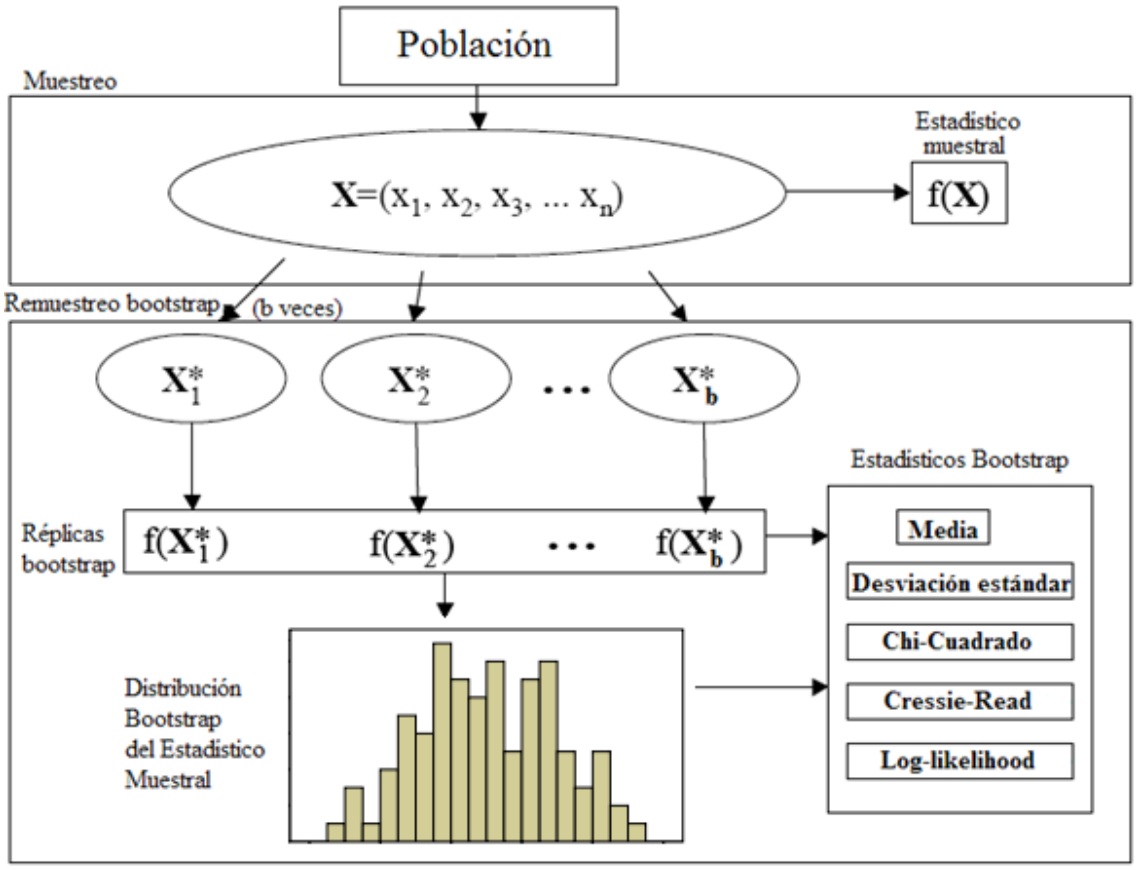

Figura 2.1. Esquema del proceso de estimación de estadísticos bootstrap.

1. A partir de la muestra original $\boldsymbol{X} = (x_1, x_2, \dots, x_n)$ se extrae una nueva muestra $\boldsymbol{x_i^*} = (x_1^*, x_2^*, \dots, x_n^*)$. Cada $\boldsymbol{x_i^*}$ para $1 \leq i \leq b$[1], de esta muestra se obtiene independientemente (con *reemplazamiento*). Es decir, tras la

[1] Teóricamente, la magnitud de **b** depende de las pruebas que se van a aplicar a los datos. Se ha afirmado que **b** debería estar entre 50 y 200, para estimar el error típico de $\widehat{\boldsymbol{\theta}}$ y debería ser de al menos de 1000 para estimar intervalos de confianza para $\widehat{\boldsymbol{\theta}}$ por el método del percentil (EFRON y TIBSHIRANI, 1986, 1993). Sin embargo, esto tiene reducida importancia en la actualidad, pues los ordenadores son tan rápidos que no tiene sentido tener un afán especial en trabajar con valores bajos de **b** y, por otra parte, nunca es pernicioso que **b** sea demasiado grande. Por lo general, con **b**=1000 se suelen conseguir buenos resultados y valores de **b** superiores a 5000 no suponen ninguna ventaja adicional.

extracción de un primer elemento, éste se repone en la muestra original de tal forma que podría ser elegido como segundo elemento de la muestra extraída. De este modo, cada observación individual tiene una probabilidad ***1/n*** de ser elegido cada vez, como si el muestreo se realizara sin reposición en un universo infinitamente grande construido a partir de la información que provee la muestra. La notación $\boldsymbol{x_b^*}$ indica que nos referimos a la ***b-ésima*** muestra *bootstrap*, la cual de forma genérica, podemos designar así:

$$\mathbf{x_j^*} = \left(x_{1j}^*, \cdots, x_{nj}^*\right) \tag{2.1}$$

2. Para la muestra obtenida se calcula el valor de un determinado estadístico $\boldsymbol{\theta_j^*}$ que se utiliza como estimador del parámetro poblacional $\boldsymbol{\theta}$, en cuyo estudio estamos interesados.

3. Repetimos los dos pasos anteriores, hasta obtener un elevado número de estimaciones $\boldsymbol{\theta_j^*}$. Aunque para obtener el número total de tales posibles muestras *bootstrap* ($\boldsymbol{n^n}$), el tiempo requerido de ordenador puede ser considerable, en la práctica no es necesario *extraer* tal número total de

muestras ya que, a veces se logra convergencia con aproximadamente 1000 muestras, o incluso con menos.

4. Se construye una distribución empírica del estadístico $\widehat{\boldsymbol{\theta}}$, que representa una buena aproximación a la verdadera función de probabilidad para ese estadístico. Es decir, se determina de este modo la distribución muestral de un estadístico sin haber hecho suposiciones sobre la distribución teórica a la que ésta se ajusta y sin manejar fórmulas analíticas para determinar los correspondientes parámetros.

De acuerdo con la idea central en que se basa el método *bootstrap*, el procedimiento supone utilizar la muestra considerando que en si misma contiene la información básica sobre la población. Por tanto, la adecuación de este método será mayor, cuando más información aporte la muestra sobre la población.

Una consecuencia directa es que a medida que aumenta el tamaño de la muestra mejor será la estimación que podemos hacer sobre la distribución muestral de un estadístico. No obstante, incluso con muestras pequeñas, entre 10 y 20 casos, el método *bootstrap* puede ofrecer resultados correctos

(BICKEL & KRIEGER, 1989), juzgándose inconvenientes para muestras de tamaño inferior a 5 (CHERNICK, 1999). Con un tamaño suficientemente grande, el incremento en el número de muestras procurará una mejora en la estimación de la distribución muestral.

2.3 LOS MÉTODOS *BOOSTRAP*

En términos generales, los métodos *bootstrap* son aquellos que se basan en el muestreo con reemplazamiento de una muestra para estudiar las propiedades estadísticas de los estimadores derivados de esa muestra.

En el método *bootstrap paramétrico* se supone un modelo paramétrico predeterminado a partir del cual se realiza la simulación, es decir, se crean nuevos datos; los datos de entrada en el modelo son sustituidos por su función de densidad. El modelo se repite un número suficientemente grande de veces y las propiedades estadísticas de las salidas del modelo se analizan a través de su distribución. Su efectividad depende de la suposición sobre qué distribuciones estadísticas son las que mejor se ajustan a los parámetros o variables que deseamos simular.

El método *bootstrap no-paramétrico* se lleva a cabo por medio de la distribución obtenida directamente de los datos. La idea consiste en generar observaciones a partir de la distribución de una muestra aleatoria independiente obtenida de la población de estudio.

La diferencia está en función de que el remuestreo se produzca sobre una distribución teórica o una distribución empírica. También, pueden estar condicionado al ajuste del modelo, es decir, se remuestrean los residuos del modelo en vez de los datos observados, entonces es llamado *bootstrap condicionado*.

En resumen, las diferentes versiones del *bootstrap* se distinguen por el estimador $\hat{F}$ que utilizan:

- Bootstrap paramétrico. Si se supone que F pertenece a un modelo paramétrico $\{F_\theta : \theta \in \Theta\}$, entonces $\hat{F} = F_{\hat{\theta}}$.

- Bootstrap no paramétrico. Si no se hace ninguna hipótesis sobre F, entonces $\hat{F} = F_n$, donde F_n es la función de distribución empírica.

2.4 EJEMPLO DE ESTIMACIÓN BOOTSTRAP PARA LA MAGNITUD DEL ERROR DE MUESTREO

El error de muestro de un estadístico es la diferencia entre el valor verdadero (parámetro poblacional) y el valor estimado[1]; éste se puede aproximar mediante el método *bootstrap* como la diferencia entre la media del estimador obtenida con un gran número de muestras y el estimador obtenido de la muestra original (MANLY, 1997).

Supongamos que a 8 personas que acaban de tener un ataque cardíaco se les ha medido su colesterol (o concentración de glucosa en sangre en miligramos por decilitro (mg/dl)) por medio de una muestra de sangre; los niveles inferiores a 100 son buenos y los niveles mayores que 160 son considerados de alto riesgo para la enfermedad cardíaca[2]. Se obtuvieron los siguientes resultados:

[1] Las cantidades que se usan para caracterizar una población son llamados parámetros y se representan por $\boldsymbol{\theta}$. Las cantidades que se calculan usando la muestra tomada y que se espera reflejen el comportamiento del parámetro son llamados estimadores y se representan por $\widehat{\boldsymbol{\theta}}$. Propiamente $\widehat{\boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{X_1}, \cdots, \boldsymbol{X_n})$ donde $(\boldsymbol{X_1}, \cdots, \boldsymbol{X_n})$ representa la muestra aleatoria y $\mathbf{T}$ es una función óptima que se usa para estimar $\boldsymbol{\theta}$.

[2] Cuando los niveles de colesterol comienzan a elevarse, se adhiere a las paredes de las arterias, y si este proceso continua por varios años, las paredes arteriales llegarán a engrosarse tanto que se dificultaría la circulación sanguínea, a tal punto, de obstruirlas totalmente, en este minuto se produce el infarto (muerte del tejido), por falta de oxígeno y nutrientes.

233 259 215 322 289 220 276 299

El promedio aritmético de la muestra resulta ser $\hat{\theta} = \bar{x} = 264.125$ (mg/dl). Pero el interés real es caracterizar el colesterol del total de personas que tiene ataques cardíacos y con ello, calcular el error de muestreo. **¿Cómo resolver el problema mediante el método *bootstrap?***

Lo primero que tenemos que hacer es construir un universo hipotético de niveles de colesterol en pacientes con problemas cardíacos y de éste seleccionar un gran número de *remuestras* con reemplazamiento, calcular el estadístico de interés para cada *remuestra* (en el ejemplo que nos ocupa la media del colesterol en la sangre), y finalmente calcular la magnitud del error de muestreo.

Los siguientes pasos nos conducen operativamente a solucionar el problema[1]:

[1] Comandos en MATLAB: » coles=[233 259 215 322 289 220 276 299]
» [mediab, muestind]=bootstrp(1000,'mean',coles)

1. Seleccionar aleatoriamente y **con reemplazamiento** 8 pacientes (de esta forma se genera un universo infinito de valores) (**Ver Tabla 2.1**).

2. Calcular la media del colesterol de esa *remuestra*, $\widehat{\boldsymbol{\theta}}^{*}$**.**

3. Repetir ***b=1000*** veces los pasos 1 y 2.

4. Calcular el promedio de las medias aritméticas obtenidas en el paso 2, $\bar{\theta}^{*}$ (**ver *Figura 3.2***).

5. Calcular el error de muestreo, $d = \left|\hat{\theta} - \bar{\theta}^{*}\right|$.

Muestras Bootstrapping									
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
299	276	299	259	322	276	215	322	276	276
259	322	299	259	299	233	259	276	220	289
289	289	322	259	322	220	259	276	215	215
322	276	299	289	322	322	220	289	215	220
299	299	233	215	276	276	215	322	215	289
276	220	215	259	289	289	289	299	289	322
322	259	276	233	259	220	259	276	220	220
233	322	233	220	220	322	220	220	215	289
Medias aritméticas									
287.37	282.87	272.0	249.12	288.62	269.75	242.00	285.00	233.12	265.00

Tabla 2.1. Muestras bootstrap y medias aritméticas (primeras **10** muestras de un total de **b=1000**).

Con lo cual podemos afirmar que la magnitud del error de muestreo al estimar el verdadero promedio aritmético del nivel de colesterol en la sangre de la población de pacientes con problemas cardíacos es,

$$d = |264.125 - 264.590| = 0.465 \text{ (mg/dl)}$$

Sin embargo, EFRON y TIBSHIRANI (1993) no recomiendan usar esta estimación del error para corregir el estimador debido a la alta variabilidad del error estimado mediante remuestreo, aunque reconocen que ofrece una visión razonable de su tendencia y en algunos casos puede ser útil para corregir los intervalos de confianza.

DAVISON y HINKLEY (1997) muestran que, en el caso de una media simple, el método *bootstrap* subestima su varianza en un factor de *(n-1)/n*, siendo *n* el tamaño de la muestra. Cuando el tamaño de la muestra es grande, este factor es inapreciable pero en muestras pequeñas puede provocar una subestimación de la varianza de la media.

Finalmente, la **Figura 2.2** presenta la distribución de las medias obtenidas con 1000 remuestras bootstrap, se puede visualizar que la distribución de probabilidad es aproximadamente normal.

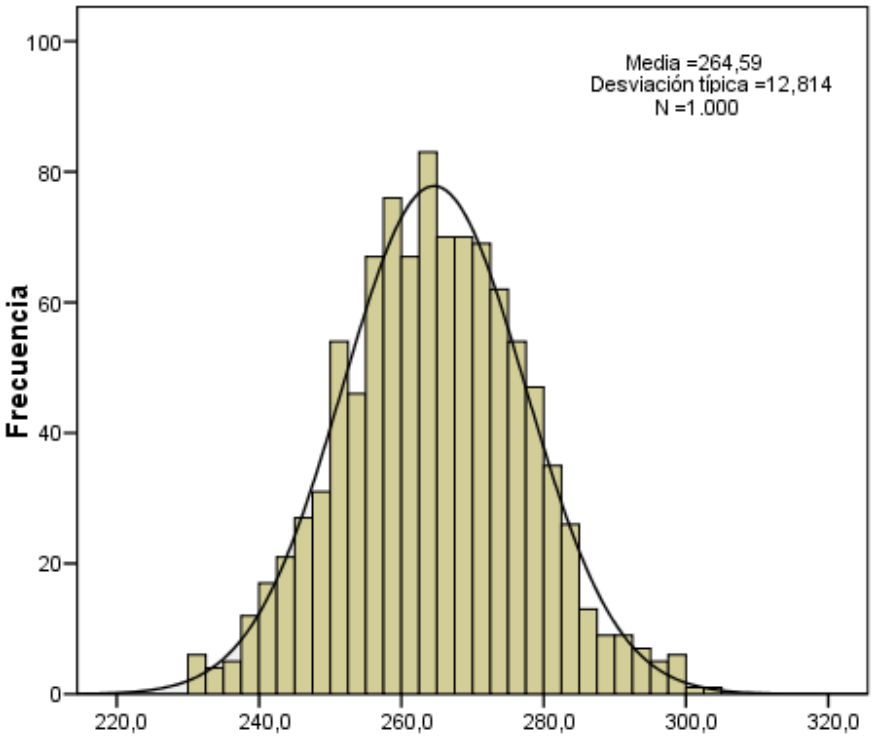

Figura 2.2. Distribución de los niveles de colesterol en la sangre de b=1000 muestras bootstrap.

2.5 ERRORES EN LOS ESTIMADORES *BOOTSTRAP*

Se ha descrito como se pueden usar los métodos *bootstrap* para evaluar le exactitud estadística de un estimador, sin embargo, los estadísticos

B*ootstrap* no son exactos ya que pueden tener una varianza sustancial. Esta varianza puede tener dos orígenes distintos (EFRON & TIBSHIRANI, 1993): por un lado está el error debido a que analizamos una muestra y no la población entera; es decir, el error estadístico o la variabilidad de muestreo, y por otra parte, el error debido a que no podemos realizar infinitas submuestras, es decir, el error de simulación o la variabilidad del remuestreo (***Figura 3.3***). El primer objetivo de un análisis mediante métodos *bootstrap* es reducir al máximo ambos tipos de errores.

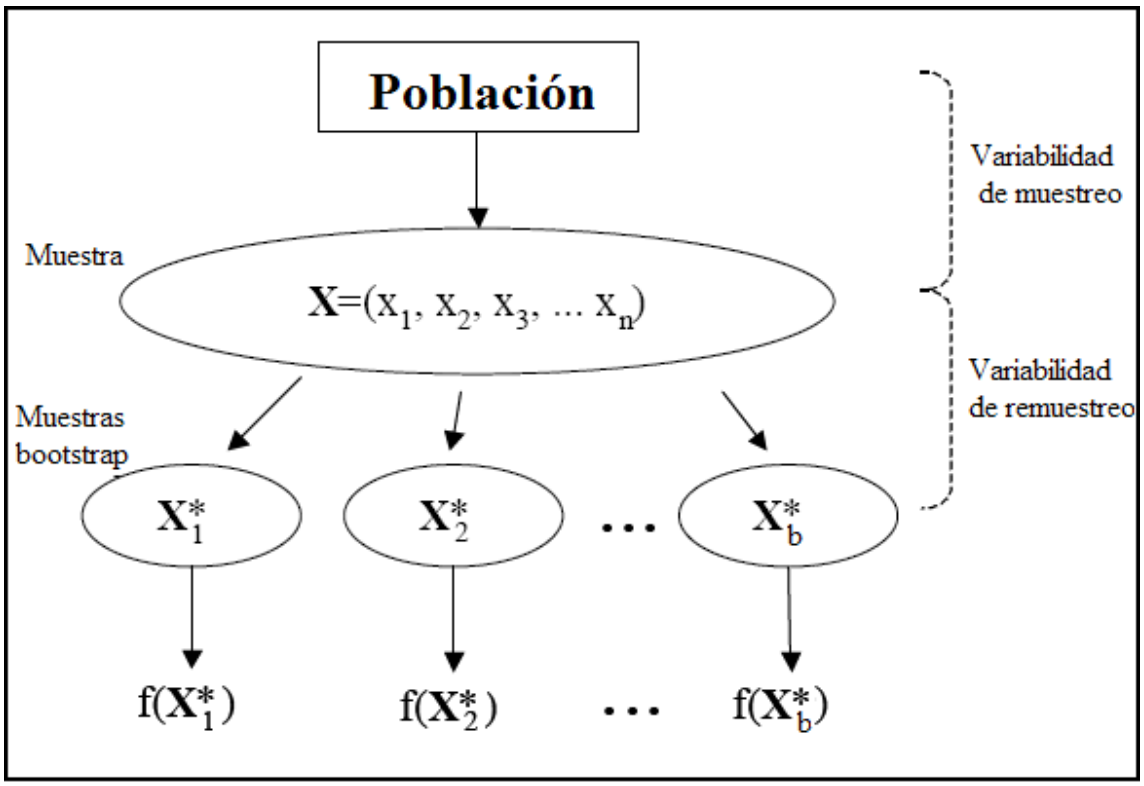

Figura 2.3. Esquema de un proceso bootstrap mostrando la variabilidad de muestreo y de remuestreo.

El error de simulación se puede reducir aumentando el número de réplicas. La teoría nos dice que los mejores estimadores se obtienen con infinitas réplicas, algo que en la práctica es imposible e innecesario. La cuestión de interés es el número de réplicas *bootstrap n*ecesarias para obtener un estadístico de suficiente precisión para nuestro propósito. Sabemos que conforme aumenta el número de réplicas se reduce el error de simulación. EFRON y TIBSHIRANI (1986) encuentran que, dependiendo de la complejidad del estadístico de interés, entre 25 y 200 réplicas pueden ser suficientes para estimar su error estándar.

En cualquier caso es necesario comprobar que estamos trabajando con un número suficiente, y hay dos maneras de hacerlo: la primera sería repetir el proceso *bootstrap* varias veces para comprobar si convergen en el mismo resultado; una segunda manera sería siguiendo los estadísticos de interés (media, Chi-cuadradas, Cressie-Read, etc.) conforme aumenta el número de réplicas hasta que se alcanza un nivel de estabilidad que consideremos suficiente (MANLY, 1997).

¿Cuándo falla el "bootstrap?

Los métodos *bootstrap* pueden proporcionar en situaciones complicadas errores estándar e intervalos de confianza que de una manera analítica serían intratables, sin embargo no siempre funcionan correctamente. Ya hemos visto que la varianza y la covarianza *bootstrap* pueden estar sesgadas en un factor de (n-1)/n; este sesgo es inapreciable con tamaños de muestra mayores de 20 aunque con muestras pequeñas debe ser tenido en cuenta. Aparte de esto, el *bootstrap* puede fallar debido a sus propiedades asintóticas, inexactitud inherente de la muestra y presencia de casos atípicos (DIXON, 2001).

Las propiedades asintóticas se refieren a la facultad del método de converger hacia un determinado valor según aumenta el número de réplicas; se ha observado que en determinadas situaciones esta convergencia se produce más lentamente de lo deseable y el método falla si se detiene en un número insuficiente de réplicas.

En determinadas situaciones el *bootstrap* falla debido a una característica intrínseca al problema. DIXON (2001) proporciona como ejemplo el cálculo de la riqueza de especies; el conteo del número de especies de una muestra es habitualmente una infravaloración del número de especies

de la población debido a que algunas especies raras pueden no estar incluidas en la muestra; cuando se realiza un *bootstrap* en estas muestras nunca se obtendrán submuestras con un número de especies mayor que el de la muestra original y por lo tanto los intervalos de confianza fallarán. Problemas similares se pueden encontrar en la determinación de máximos o mínimos.

El *bootstrap* puede fallar debido a la presencia de casos atípicos: puesto que el *bootstrap* asume que la distribución de la muestra representa la distribución de la población, si la muestra es inusual también lo serán los estimadores *bootstrap* derivados de ella. Un claro ejemplo de este problema es con poblaciones con distribuciones muy asimétricas donde es fácil que los valores extremos estén submuestreados.

En resumen, las técnicas *bootstrap* nos ofrecen la posibilidad de evaluar la incertidumbre asociada a un estimador mediante el cálculo automático de su error estándar, sus intervalos de confianza, su sesgo y su distribución de frecuencias. Tienen la ventaja, sobre los métodos paramétricos, de que no depende de ningún supuesto acerca de la distribución estadística asociada a los datos; cuando no existen dudas acerca de la distribución subyacente, los métodos paramétricos son la mejor opción.

En situaciones de incertidumbre sobre la distribución de un parámetro, los métodos *bootstrap* proporcionan estimadores más robustos (EFRON y TIBSHIRANI, 1993). También pueden ser útiles en situaciones en que se conoce el modelo del error pero el parámetro a estimar implica procesos complejos y el cálculo analítico de su error no es sencillo. Por otra parte, las principales desventajas de los métodos bootstrap son: la necesidad de desarrollar programas de ordenador adecuados a las circunstancias particulares de cada caso y el tiempo que se emplea en los cálculos, que depende de la complejidad del problema y del número de réplicas.

Puesto que el único supuesto de los métodos *bootstrap* es que la distribución de la muestra conserva las propiedades estadísticas de la distribución de la población, el *bootstrap* fallará cuando la distribución muestral no sea representativa de la distribución poblacional; esta última característica no hace que el método sea inferior a otros ya que no hay ninguno suficientemente robusto para este problema.

CAPÍTULO 3

ALTERNATIVAS PARA LA SOLUCIÓN DEL PROBLEMA DE TABLAS POCO OCUPADAS

3.1 INTRODUCCIÓN

En la aplicación de los modelos de clases latentes se pueden presentar problemas con tablas poco ocupadas, ya que cuando la frecuencia observada de los patrones de respuesta es cero, los valores esperados tienden a ser muy pequeños. Por esta razón, no se cumplen las propiedades asintóticas de los estadísticos de bondad de ajuste, los estimadores son inestables y existe menor precisión de los parámetros. A continuación se presentan alternativas para la solución de los problemas citados.

3.2 AÑADIR UNA CONSTANTE A CADA CELDA

Una solución de uso frecuente, es agregar una constante pequeña a cada frecuencia de los patrones de respuestas antes del análisis. Uno de las consecuencias de agregar una constante es que las estimaciones de los parámetros log-lineales tienden a cero. Asimismo, el tamaño de muestra se incrementa al sumarle una constante a las frecuencias. GOODMAN (1974) recomendó usar de este procedimiento sólo para los modelos saturados.

AGRESTI (1990) propone realizar un análisis de la sensibilidad para constantes de varios tamaños. Puede incluso ser adecuado agregar una

constante extremadamente pequeña, tal como 1×10^{-8}, a las "celdas" vacías.

3.3 IMPONER RESTRICCIONES A LOS PARÁMETROS DEL MODELO

Otra alternativa es utilizar modelos con restricciones en los parámetros, que tengan sentido en términos de este tipo de problemas, con el inconveniente de la posible existencia de más de un extremo local, en la función de verosimilitud y el proceso de estimación quedará afectado al liberar tantos grados de libertad como parámetros se restrinjan.

En la práctica, el hecho de imponer restricciones a los parámetros del modelo es equivalente a realizar una ***prueba de hipótesis*** respecto a los valores de los parámetros involucrados en la restricción. Para un conjunto de variables manifiestas binarias, las restricciones implican que sólo son admisibles ciertos patrones de respuesta, los cuales comprenden determinadas secuencias de 0's y 1's.

A manera de ejemplo, supongamos que tenemos treces variables manifiestas binarias, $(X_1, X_2, \ldots, X_{13})$ y un modelo con dos variables latentes Y_1 e Y_2, con r=2 y s=2 niveles de respuesta, respectivamente.

Además, que (X_1, X_2, X_3, X_4) son indicadoras de Y_1, y las variables $(X_5, X_6, \dots, X_{13})$ son indicadoras de la variable latente Y_2.

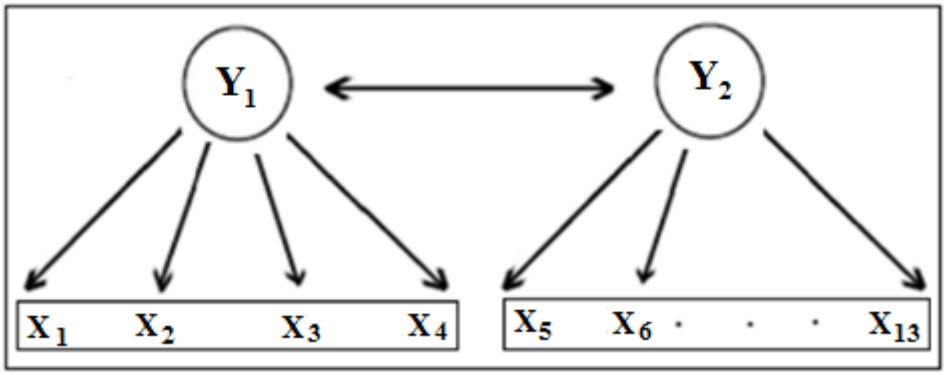

Figura 3.1. Modelo con dos variables latentes relacionadas.

En base a la **Figura 3.1**, se imponen restricciones a las probabilidades condicionales para considerar el modelo de dos variables latentes como un modelo de una variable latente con cuatro categorías (clases latentes). Como las variables (X_1, X_2, X_3, X_4) sólo son afectadas por Y_1, se consideran las siguientes restricciones,

$$\pi_{X_1/Y(1)}(x_1) = \pi_{X_1/Y(2)}(x_1) \qquad \pi_{X_1/Y(3)}(x_1) = \pi_{X_1/Y(4)}(x_1)$$
$$\pi_{X_2/Y(1)}(x_2) = \pi_{X_2/Y(2)}(x_2) \qquad \pi_{X_2/Y(3)}(x_2) = \pi_{X_2/Y(4)}(x_2)$$
$$\pi_{X_3/Y(1)}(x_3) = \pi_{X_3/Y(2)}(x_3) \qquad \pi_{X_3/Y(3)}(x_3) = \pi_{X_3/Y(4)}(x_3)$$
$$\pi_{X_4/Y(1)}(x_4) = \pi_{X_4/Y(2)}(x_4) \qquad \pi_{X_4/Y(3)}(x_4) = \pi_{X_4/Y(4)}(x_4)$$

Las variables $(X_5, X_6, \dots, X_{13})$ sólo son afectadas por Y_2, por lo tanto se consideran las siguientes restricciones,

$$\pi_{X_5/Y(1)}(x_5) = \pi_{X_5/Y(3)}(x_5) \qquad \pi_{X_5/Y(2)}(x_5) = \pi_{X_5/Y(4)}(x_5)$$
$$\pi_{X_6/Y(1)}(x_6) = \pi_{X_6/Y(3)}(x_6) \qquad \pi_{X_6/Y(2)}(x_6) = \pi_{X_6/Y(4)}(x_6)$$
$$\pi_{X_7/Y(1)}(x_7) = \pi_{X_7/Y(3)}(x_7) \qquad \pi_{X_7/Y(2)}(x_7) = \pi_{X_7/Y(4)}(x_7)$$

$$\pi_{X_8/Y(1)}(x_8) = \pi_{X_8/Y(3)}(x_8) \qquad \pi_{X_8/Y(2)}(x_8) = \pi_{X_8/Y(4)}(x_8)$$

$$\pi_{X_9/Y(1)}(x_9) = \pi_{X_9/Y(3)}(x_9) \qquad \pi_{X_9/Y(2)}(x_9) = \pi_{X_9/Y(4)}(x_9)$$

$$\pi_{X_{10}/Y(1)}(x_{10}) = \pi_{X_{10}/Y(3)}(x_{10}) \qquad \pi_{X_{10}/Y(2)}(x_{10}) = \pi_{X_{10}/Y(4)}(x_{10})$$

$$\pi_{X_{11}/Y(1)}(x_{11}) = \pi_{X_{11}/Y(3)}(x_{11}) \qquad \pi_{X_{11}/Y(2)}(x_{11}) = \pi_{X_{11}/Y(4)}(x_{11})$$

$$\pi_{X_{12}/Y(1)}(x_{12}) = \pi_{X_{12}/Y(3)}(x_{12}) \qquad \pi_{X_{12}/Y(2)}(x_{12}) = \pi_{X_{12}/Y(4)}(x_{12})$$

$$\pi_{X_{13}/Y(1)}(x_{13}) = \pi_{X_{13}/Y(3)}(x_{13}) \qquad \pi_{X_{13}/Y(2)}(x_{13}) = \pi_{X_{13}/Y(4)}(x_{13})$$

Por lo tanto, el modelo considerado en la **Figura 3.1**, puede ser estimado considerando sólo una variable latente *Y*, bajo adecuadas restricciones a los parámetros del modelo.

3.4 UTILIZACIÓN DE LOS MÉTODOS DE REMUESTREO

Mediante el remuestreo o "*bootstrap*" se puede resolver el problema de seleccionar el modelo de clases latentes en tablas poco ocupadas (EFRON, 1979). Las muestras *bootstrap* son utilizadas para re-estimar los parámetros del modelo y los estadísticos de bondad de ajuste (***EBA***). Seguidamente, los ***EBA*** de los datos verdaderos son evaluados comparándolos con la distribución empírica de los ***EBA*** obtenidos a través de las muestras generadas por medio del método *bootstrap*.

El método *bootstrap* se puede considerar como la manera de encontrar la distribución desconocida F_T de un ***EBA*** *T*, cuando *T* es función de los datos observados $(x_1, \cdots, x_n)$ y de un vector $\boldsymbol{\theta}$ de

parámetros, así $t = T(x_1, \cdots, x_n, \boldsymbol{\theta})$. La distribución de T es aproximada generando una muestra de los resultados independientes t_j^* para $j = 1, \cdots, b$ y construyendo la distribución empírica $\hat{F}_{t^*}$. Esto puede conseguirse simulando b muestras $\boldsymbol{x_j^*} = \left(x_{1j}^*, \cdots, x_{nj}^*\right)$ y estimando el vector de parámetros $\widehat{\boldsymbol{\theta}}$.

COLLINS et al. (1993) investigaron por medio de una simulación de datos con problemas de "sparse" (tablas poco ocupadas) la distribución de G^2, X^2 y *CR*; y encontraron que la media de la distribución X^2 es aproximadamente la Chi-cuadrada, pero la desviación estándar de X^2 es considerablemente más grande que la G^2 y *CR*, y aún mayor que la desviación estándar[1] de la distribución teórica χ^2. Por esta razón, recomiendan no utilizar el estadístico Chi-cuadrado en modelos de clases latentes en presencia de tablas poco ocupadas.

LANGEHEINE et al. (1996) estudiaron con el *bootstrap paramétrico* modelos de clases latentes, concluyen que los "sparseness" usualmente no causan problemas en la estimación de los parámetros, pero la evaluación del ajuste del modelo se puede dificultar por el hecho de que la verdadera distribución del ***EBA*** es una aproximación inadecuada de la

[1] La desviación estándar de una variable con distribución Chi-cuadrado es la raíz cuadrada de dos por los grados de libertad ($\sqrt{2k}$).

distribución teórica χ^2. También afirman que un modelo es incorrecto, si la proporción α de los G^2 obtenidos con *bootstrap* es más grandes que el valor original[2].

Según DAVIER (1997), a un nivel de significación (α), si al menos $(n\alpha)$ muestras *bootstrap paramétricas,* tienen un estadístico de bondad de ajuste mayor al observado con los datos originales, el modelo es rechazado con un nivel de $(1-\alpha)\%$ de confianza. Concluye que los estadísticos de bondad de ajuste *bootstrap* Chi-cuadrado y Creesie-Read, proporcionan resultados similares aun cuando los datos sean muy escasos. En contraste, los estadísticos FT (Freeman-Tukey) y G^2 pueden conducir a decisiones incorrectas en diversas condiciones del "sparseness".

DAVIER (1997), plantea que cuando a los datos originales se le ajusta un modelo apropiado, el valor $t = T(x_1, \cdots, x_n, \hat{\theta})$ no presenta diferencias significativas en relación a las muestras del *bootstrap* $(x_j^*)_{j=1,\cdots,b}$. En caso contrario, con un modelo incorrecto, el valor $\boldsymbol{t}$ (que es una medida de bondad de ajuste de los datos originales) deben presentar diferencias significativas con los valores t^* de las muestras *bootstrap.*

[2] Por cualquier duda en la traducción, la regla en inglés es: "**Reject the model if the proportion α of bootstrap G^2 that are larger than the original G^2 is very small**" (LANGEHEINE et al., 1996, página 495).

CAPÍTULO 4

APLICACIONES Y RESULTADOS

4.1 DATOS

En la presente investigación son utilizados los resultados del estudio "*Factores que inciden en el consumo de drogas, población juvenil. Región Central de Occidente*" de Costa Rica. La encuesta fue realizada por el Instituto Costarricense sobre Drogas, la Asociación Ramonense Pro Bienestar de la Comunidad, la Caja Costarricense de Seguro Social, el Ministerio de Educación, el Instituto de Alcoholismo y Fármaco dependencia y la Universidad de Costa Rica durante noviembre del 2006.

La población en estudio está formada por 13.428 jóvenes perteneciente a 135 centros de enseñanza y 17 Equipos Básicos de Atención Integral de Salud (EBAIS). La muestra de 7.553 jóvenes se seleccionó utilizando un muestreo por conglomerados (sin submuestreo) proporcionales al tamaño.

Esta encuesta se diseño principalmente para proporcionar información sobre el consumo de drogas, actividades de tiempo libre, participación de los jóvenes en actividades de la comunidad, condiciones de vida y niveles de información de los jóvenes.

El módulo de interés son las variables respecto a "***Cuando fue que, por primera vez utilizó o consumió drogas***"; se tienen así 13 drogas (variables o ítems) con respuestas binarias (0=no, 1=Sí). El listado de variables del cuestionario utilizado en la presente investigación es el siguiente.

1- ¿Fumó cigarrillos?

2- ¿Bebió cerveza?

3- ¿Bebió vino?

4- ¿Consumió bebidas fuertes (whisky, vodka, ginebra, guaro, etc.)?

5- ¿Fumó marihuana?

6- ¿Consumió cocaína?

7- ¿Utilizó pastillas (estimulantes, tranquilizantes) para drogarte?

8- ¿Uso inhalantes?

9- ¿Utilizó alucinógenos?

10- ¿Uso heroína?

11- ¿Uso éxtasis?

12- ¿Consumió crack?

13- ¿Utilizó otras drogas?

En la **Tabla 4.1** se presentan las distribuciones de frecuencias marginales y bivariadas del consumo de drogas. Tanto en filas como en columnas las drogas se identifican con números de 1 a 13 según el orden expuesto en la anterior página. En la diagonal de la matriz se presenta la distribución de frecuencias de las drogas; por ejemplo, 2.441 jóvenes tienen el hábito de fumar cigarrillos y 166 han consumido cocaína.

DROGAS	DROGAS													
		1	2	3	4	5	6	7	8	9	10	11	12	13
	1	**2.441**												
	2	**2.300**	4.573											
	3	3.239	4.202	5.588										
	4	831	2.840	2.886	3.016									
	5	498	576	580	554	601								
	6	145	155	154	154	**147**	**166**							
	7	125	129	131	127	109	94	139						
	8	121	133	137	128	110	95	92	145					
	9	98	103	102	104	98	84	87	86	110				
	10	79	80	82	83	79	79	75	76	78	89			
	11	82	91	91	92	85	80	82	81	81	81	99		
	12	96	100	101	102	97	92	87	88	81	83	85	109	
	13	135	145	145	144	109	89	91	92	82	80	88	92	157

TABLA 4.1. Matriz de frecuencias marginales y bivariadas de factores que inciden en el consumo de drogas.

Las frecuencias bivariadas nos revelan por ejemplo, que 2.300 personas tienen el hábito de fumar y beber cerveza. También que 147 jóvenes consumen cocaína y marihuana. Se observa una asimetría de la

matriz, determinada por la droga cinco (fumó marihuana), que resulta como una frontera entre las drogas débiles (o legales) y las fuertes (las prohibidas).

Los resultados serán el fundamento para plantear posteriormente, un modelo con dos variables latentes. Existen drogas legales e ilegales. Las ilegales son las que son penalizadas por la ley, tales como: la marihuana, la cocaína, el éxtasis, etc. Las legales son las que se pueden comprar en diferentes negocios, entran en la categoría: el cigarrillo, la cerveza y el alcohol. En otras palabras, las drogas legales son aquellas cuya utilización no está prohibida por la ley, y las ilegales son las restantes, cuyo consumo constituye un delito.

Por otra parte, considerando que se tiene 13 variables manifiestas, el número posible de patrones de respuesta es 8.192. El tamaño de muestra es 7.553, sin embargo, el 98% de los patrones no fueron observados. En la **Tabla 4.2**, se presentan los patrones de respuestas observados, que equivalen aproximadamente a un 2% del total de patrones. Por ejemplo el patrón (0000000000000) nos indica que 1.513 personas no consumen ninguna de la treces drogas y por lo contrario, 56 jóvenes consumen todas las drogas, que es el patrón (1111111111111).

Patrones de respuesta	Frecuencia	Porcentaje	Porcentaje acumulado
0000000000000	1513	20,0	20,0
1111000000000	1214	16,1	36,1
1000000000000	1168	15,5	51,6
1100000000000	1015	13,4	65,0
1110000000000	925	12,2	77,3
1110000000000	400	5,3	82,5
1111100000000	281	3,7	86,3
1000000000000	189	2,5	88,8
1100000000000	111	1,5	90,2
1100000000000	69	,9	91,2
1111000000000	63	,8	92,0
1101000000000	59	,8	92,8
1010000000000	56	,7	93,5
1111111111111	56	,7	94,3
1000000000000	43	,6	94,8
1010000000000	32	,4	95,2
1111110000000	29	,4	95,6
1011000000000	23	,3	95,9
1110100000000	18	,2	96,2
1000000000000	17	,2	96,4
1111000000001	16	,2	96,6
1111100000001	14	,2	96,8
1101000000000	13	,2	97,0
1111100010000	9	,1	97,1
1001000000000	8	,1	97,2
1111000100000	8	,1	97,3
1111001000000	8	,1	97,4
1111101000000	6	,1	97,5
1111111111111	6	,1	97,6

Patrones de respuesta	Frecuencia	Porcentaje	Porcentaje acumulado
1110000000000	5	,1	97,6
1100001000000	5	,1	97,7
1111100100000	5	,1	97,8
1000000000100	4	,1	97,8
1110001000000	4	,1	97,9
1100100000000	4	,1	97,9
1111010000000	4	,1	98,0
1010000000000	3	,0	98,0
1000100000000	3	,0	98,1
1101100000000	3	,0	98,1
1110001000000	3	,0	98,1
1111110000010	3	,0	98,2
1111111000000	3	,0	98,2
1111111100011	3	,0	98,3
1111111110000	3	,0	98,3
1111111111110	3	,0	98,3
1000010000000	2	,0	98,4
1111111111100	2	,0	98,4
1100100000000	2	,0	98,4
1110100000000	2	,0	98,4
1111100000000	2	,0	98,5
1100000000001	2	,0	98,5
1110000100000	2	,0	98,5
1111000001111	2	,0	98,5
1111001000001	2	,0	98,6
1111101000101	2	,0	98,6
1000000000000	1	,0	98,6
Otros*	**107**	**1,4**	100,0
Total	7553	100,0	

*** Patrones de respuesta con una frecuencia de 1.**

Tabla 4.2. Distribución de los patrones de respuestas observados.

4.2 ESTIMACIÓN DEL MODELO: DATOS ORIGINALES

Se realizaron múltiples ajustes de modelos de clases latentes[1] y finalmente se escogieron los dos que resultaron más apropiados para describir la relación entre las drogas (variables manifiestas) y las variables

[1] Las muestras bootstrap fueron seleccionadas mediante un algoritmo implementado en MATLAB (MATHWORKS, 2007).Para realizar los cálculos de los modelos se utilizaron los programas LEM (VERMUT, 1997b) para el bootstrap no-paramétrico y WINMIRA, versión 2001 (DAVIER, 2001) en el bootstrap paramétrico.

latentes, utilizando como criterio de decisión los estadísticos de bondad de ajuste: Chi-cuadrado, razón de verosimilitud y Cressie-Read. Los modelos seleccionados fueron:

- **Modelo 1 (M1)**. Una variable latente con ocho clases latentes.

- **Modelo 2 (M2)**. Dos variables latentes relacionadas, cada una de ellas con dos clases latentes.

A continuación, se estudiarán los estadísticos de bondad de ajuste de **M1** y **M2** con los datos originales (sin remuestreo).

El valor del χ^2 de **M1** resulta significativamente superior a **M2**; sin embargo, esta relación se invierte en G^2, al disminuir el valor de 2667,89 a 575,55 de **M2** a **M1**, respectivamente. Respecto al estadístico ***CR***, aunque existe una pequeña discrepancia, está no resulta tan marcada entre ambos modelos. La diferencia entre los estadísticos es una clara consecuencia de trabajar con tablas poco ocupadas (**Tabla 4.3**).

Estadísticos	Modelos	Valor
Chi-cuadrado (χ^2)	M1	66.131,62
	M2	28.917,28
Razón de verosimilitud (G^2)	M1	575,55
	M2	2.667,89
Cressie-Read (CR)	M1	5.602,28
	M2	7.632,13

Tabla 4.3. Estadísticos de bondad de ajuste (n=7.553).

El modelo de una variable latente (**M1**) proporciona la siguiente definición de las clases (**Tabla 4.4**). La primera clase latente se puede interpretar como la correspondiente al grupo de individuos "alcohólicos", que tienen el hábito de consumir cerveza, vino y bebidas fuertes (guaro, whisky, etc.), conteniendo aproximadamente un 21% de la población.

La segunda y sexta clase latente está compuesta por 5% de individuos; y representan los que consumen bebidas alcohólicas, fuman cigarrillos y marihuana, pero en la segunda clase el consumo de marihuana se encuentra más asentado. En tanto, los sujetos que componen la tercera clase (el 28% del total), se puede interpretar como la correspondiente al grupo de individuos "normales", que no consumen ningún tipo de droga. La cuarta clase (21%), son personas que sólo beben vino.

Variables manifiestas	C1	C2	C3	C4	C5	C6	C7	C8
	0,2075	**0,0397**	**0,2842**	**0,2148**	**0,0198**	**0,011**	**0,2129**	**0,0099**
X1	0,1652	0,9711	0,0000	0,053	1,000	0,8604	0,9428	0,8937
X2	0,8890	0,9972	0,1075	0,4982	0,5653	0,9637	1,0000	0,9069
X3	0,9696	0,9950	0,2017	0,9994	0,1824	0,8987	0,9593	0,9336
X4	0,8162	1,0000	0,0090	0,0000	0,1345	0,9393	0,7760	0,9467
X5	0,0563	0,9983	0,0006	0,0013	0,0491	0,6848	0,0457	0,9577
X6	0,0025	0,1317	0,0000	0,0014	0,0000	0,5018	0,0035	0,9711
X7	0,0003	0,0275	0,0006	0,0010	0,0000	0,5065	0,0077	0,9701
X8	0,0053	0,0310	0,0010	0,0039	0,0000	0,4376	0,0059	0,9736
X9	0,0010	0,0338	0,0000	0,0003	0,0000	0,2633	0,0004	1,0000
X10	0,0000	0,0000	0,0000	0,0000	0,0000	0,1786	0,0000	0,9878
X11	0,0000	0,0000	0,0000	0,0006	0,0000	0,2756	0,0000	1,0000
X12	0,0000	0,0082	0,0000	0,0000	0,0066	0,3660	0,0000	1,0000
X13	0,0000	0,0464	0,0007	0,0027	0,0121	0,5423	0,0119	0,9469

Tabla 4.4. Parámetros del modelo clásico de ocho clases latentes

Los que tienen el hábito de fumar y tomar cerveza son la quinta clase y constituyen el 2% de los individuos. La séptima clase consumen las drogas legales: cigarros, cerveza, vino y bebidas fuertes. Los casos más graves son sujetos politoxicómanos (1%) que definen la octava clase latente (coordenada *C8 de la* ***Figura 4.1***). La **politoxicomanía** o **policonsumo** se produce cuando una persona se administra una variada gama de drogas.

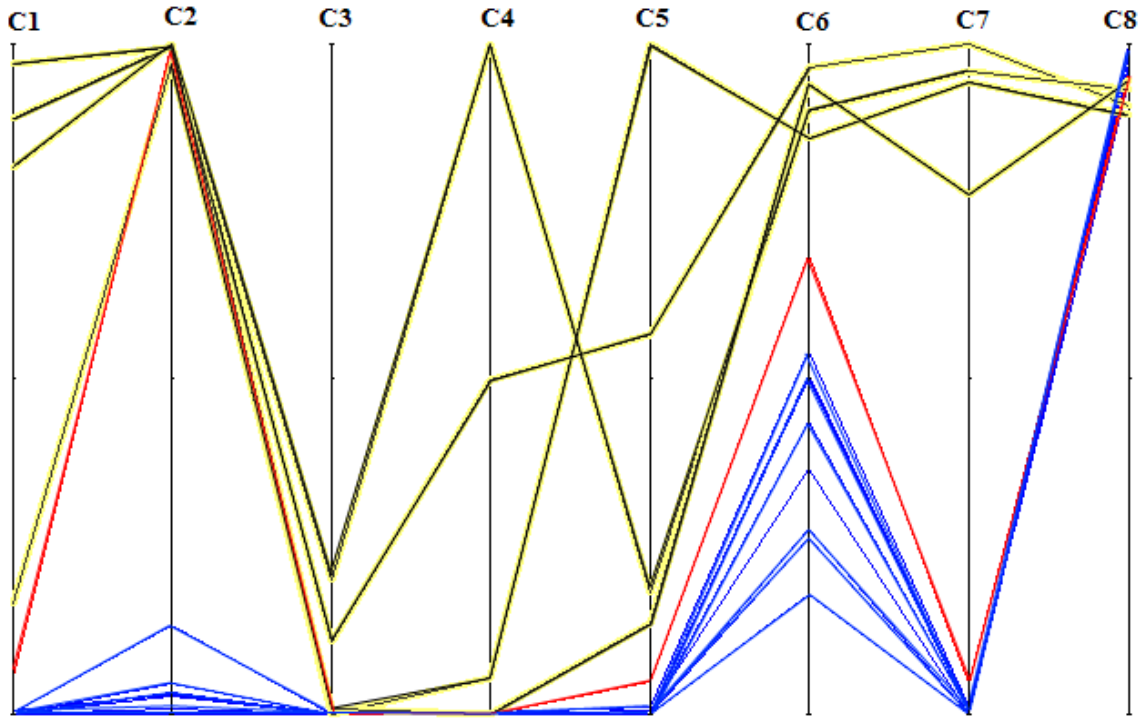

Figura 4.1. Representación en coordenadas paralelas de los parámetros del modelo de ocho clases latentes.

Examinando las probabilidades condicionales del modelo de dos variables latentes (**M2**) de la **Tabla 4.5**, podemos concluir que para la primera variable latente, las clases quedan definidas como: la primera por aquellos individuos que consumen las drogas legales: cigarros, cerveza, vino y guaro (52%) y la segunda por los que no consumen ningún tipo de droga (46%).

En tanto, para la segunda variable latente las clases quedan formadas de la siguiente manera: la primera por los individuos politoxicómanos (2%) y la segunda por aquellos que solamente consumen drogas ilegales (0.05%).

Variables manifiestas	Y1C1	Y1C2	Y2C1	Y2C2
	0.5182	**0.4613**	**0.0200**	**0.0005**
X1	0.5704	0.0350	0.5704	0.0350
X2	0.9690	0.1818	0.9690	0.1818
X3	0.9574	0.4863	0.9574	0.4863
X4	0.7190	0.0267	0.7190	0.0267
X5	0.0630	0.0630	0.8763	0.8763
X6	0.0063	0.0063	0.7736	0.7736
X7	0.0038	0.0038	0.7179	0.7179
X8	0.0046	0.0046	0.7191	0.7191
X9	0.0018	0.0018	0.6271	0.6271
X10	0.0000	0.0000	0.5770	0.5770
X11	0.0002	0.0002	0.6336	0.6336
X12	0.0002	0.0002	0.6956	0.6956
X13	0.0060	0.0060	0.7297	0.7297

Tabla 4.5. Parámetros del modelo de dos variables latentes.

4.3 ESTADÍSTICOS DE BONDAD DE AJUSTE

Las muestras *bootstrap* son utilizadas para re-estimar los parámetros del modelo y los estadísticos de bondad de ajuste (***EBA*)**. La **Tabla 4.6** muestra la media aritmética y desviación estándar de los estadísticos χ^2, G^2 y *CR*, para el modelo simulado para una y dos variables latentes, utilizando el *bootstrap* paramétrico y no paramétrico[1] con 100 repeticiones.

.

[1] **Bootstrap paramétrico**. Si se supone que F pertenece a un modelo paramétrico $\{F_\theta : \theta \in \Theta\}$, entonces $\hat{F} = F_{\hat{\theta}}$. **Bootstrap no paramétrico**. Si no se hace ninguna hipótesis sobre *F*, entonces $\hat{F} = F_n$, donde F_n es la función de distribución empírica

Estadísticos	Modelos	**Paramétrico**		**No Paramétrico**	
		Media	Desviación estándar	Media	Desviación estándar
Chi-cuadrado	M1	34.042,07	178.329,17	8.430,03	147,48
	M2	4.550,91	8.060,91	10.527,96	275,49
Razón de verosimilitud	M1	346,11	250,46	6.040,63	60,18
	M2	458,13	34,12	7.212,70	86,15
Cressie-Read	M1	2.768,98	11.110,85	6.887,60	91,67
	M2	1.006,03	464,88	8.459,81	157,84

Tabla 4.6. Estadísticos de bondad de ajuste con bootstrap (n=100).

Es evidente que con el *bootstrap paramétrico* la media y desviación estándar del estadístico χ^2 es mucho menor usando dos variables latentes. En contraste, el *bootstrap no paramétrico* aporta resultados inversos, ya que ambos estimadores -media y desviación estándar - son inferiores al utilizar una variable latente.

La media de la razón de verosimilitud es más pequeña al utilizar una clase para ambas remuestras *bootstrap*, no obstante la desviación estándar del *bootstrap paramétrico* es muy grande en relación al modelo de dos variables latentes.

La media del estadístico Cressie-Read con *bootstrap paramétrico* favorece a **M2** y tiene una desviación estándar pequeña, pero se invierte el criterio de mejor modelo al usar el *bootstrap no paramétrico,* la media y la desviación son menores en **M1**.

En resumen, los resultados del *bootstrap paramétrico* benefician al modelo de dos variables latentes (**M2)** y por otra parte, el *bootstrap no paramétrico* es concluyente en determinar que el mejor modelo es aquel con una variable latente y ocho clases (**M1)**.

4.4 PRUEBAS DE NORMALIDAD DE LOS ESTADÍSTICOS

Para corroborar que los estadísticos de bondad de ajuste tienen una distribución normal, se utiliza la prueba de Kolmogorov-Smirnov y Shapiro-Wilk. Es de interés el grado de acuerdo entre la distribución del conjunto de valores de la remuestras *bootstrap* y la distribución teórica normal.

4.4.1 MODELO CON UNA VARIABLE LATENTE

Los estadísticos de bondad de ajuste del *bootstrap no-paramétrico* para **M1**, presentan una distribución normal, como se observa en la ***Figura 4.2*** y los resultados de las pruebas de Kolmogorov-Smirnov y Shapiro-Wilk de la ***Tabla 4.7*** así lo ratifican.

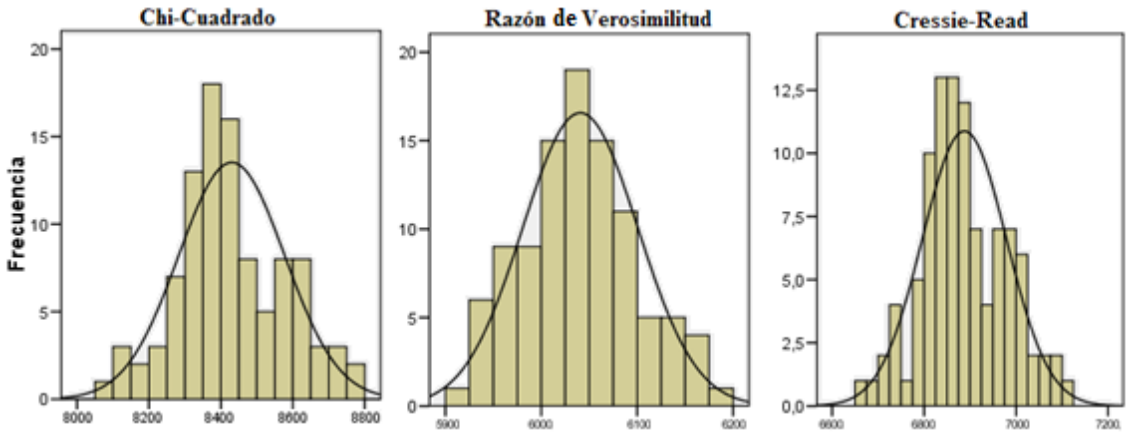

Figura 4.2. Distribución de los estadísticos de bondad de ajuste: bootstrap no-paramétrico una variable latente y ocho clases latentes.

Estadísticos	Kolmogorov-Smirnov			Shapiro-Wilk		
	Valor	Grados de Libertad	Significación	Valor	Grados de Libertad	Significación
Chi-cuadrado	**0,082**	**100**	**0,097**	**0,984**	**100**	**0,287**
Razón de verosimilitud	**0,040**	**100**	**0,200**	**0,989**	**100**	**0,592**
Cressie-Read	**0,079**	**100**	**0,127**	**0,987**	**100**	**0,448**

Tabla 4.7. Pruebas de normalidad de los estadísticos de bondad de ajuste bootstrap no-paramétrico una variable latente y ocho clases latentes.

Se observa que los tests K-S y Shapiro-Wilk son significativos (***Tabla 4.8***), por lo que podemos concluir que los estadísticos de la simulación *bootstrap paramétrico,* para una variable latente no siguen una distribución normal. Esto se verifica en la ***Figura 4.3***.

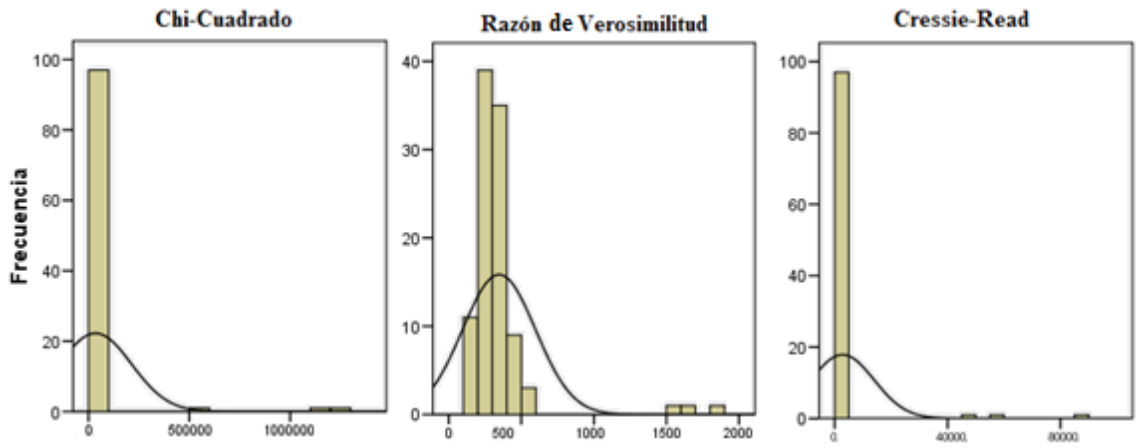

Figura 4.3. Distribución de los estadísticos de bondad de ajuste: bootstrap paramétrico.
Una variable latente y ocho clases latentes.

Estadísticos	Kolmogorov-Smirnov			Shapiro-Wilk		
	Valor	**Grados de Libertad**	**Significación**	**Valor**	**Grados de Libertad**	**Significación**
Chi-cuadrado	0,484	100	0,000	0,175	100	0,000
Razón de verosimilitud	0,279	100	0,000	0,457	100	0,000
Cressie-Read	0,473	100	0,000	0,190	100	0,000

Tabla 4.8. Pruebas de normalidad para el bootstrap paramétrico
Una variable latente y ocho clases latentes.

4.4.2 MODELO CON DOS VARIABLES LATENTES

Los resultados de las pruebas Kolmogorov-Smirnov y Shapiro-Wilk de **M2** resultan no significativos (***Tabla 4.9***) respecto a la distribución de los ***EBA*** de las re-muestras *bootstrap*, por tanto, podemos concluir que tienen una distribución normal. En la ***Figura 4.4*** podemos verificar que la distribución de los estadísticos: Chi-Cuadrado, Razón de verosimilitud y Cressie-Read se ajustan muy bien a la distribución normal.

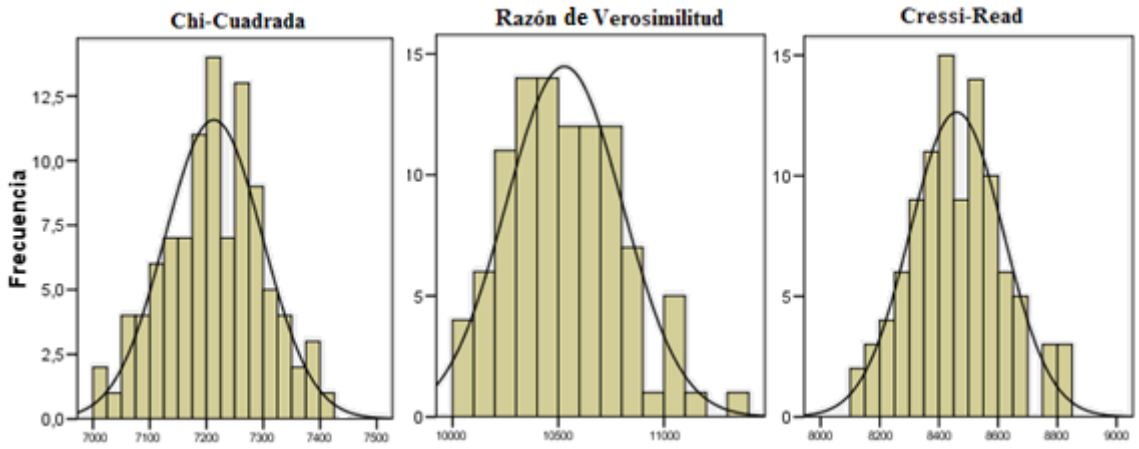

Figura 4.4. Distribución de los estadísticos de bondad de ajuste: bootstrap no-paramétrico. Dos variables latentes y dos clases latentes.

Estadísticos	Kolmogorov-Smirnov			Shapiro-Wilk		
	Valor	**Grados de Libertad**	**Significación**	**Valor**	**Grados de Libertad**	**Significación**
Chi-cuadrado	0,054	100	0,200	0,983	100	0,217
Razón de verosimilitud	0,047	100	0,200	0,993	100	0,876
Cressie-Read	0,050	100	0,200	0,989	100	0,587

Tabla 4.9. Pruebas de normalidad para el bootstrap no-paramétrico
Dos variables latentes y ocho clases latentes.

Los estadísticos Chi-Cuadrado y Cressie-Read de **M2**, utilizando *bootstrap paramétrico* no provienen de una distribución normal. En tanto, no se rechaza la hipótesis de que la razón de verosimilitud tiene una distribución teórica normal (***Figura 4.5, Tabla 4.10***).

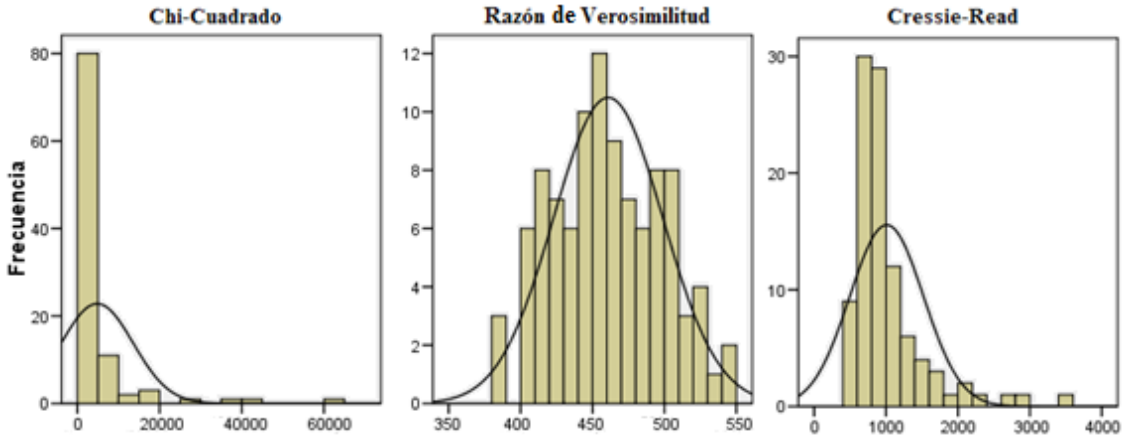

Figura 4.5. Distribución de los estadísticos de bondad de ajuste: bootstrap paramétrico. Dos variables latentes y dos clases latentes.

Estadísticos	Kolmogorov-Smirnov			Shapiro-Wilk		
	Valor	**Grados de Libertad**	**Significación**	**Valor**	**Grados de Libertad**	**Significación**
Chi-cuadrado	0,316	100	0,000	0,443	100	0,000
Razón de verosimilitud	0,059	100	0,200	0,985	100	0,305
Cressie-Read	0,207	100	0,000	0,742	100	0,000

Tabla 4.10. Pruebas de normalidad para el bootstrap paramétrico
Dos variables latentes y ocho clases latentes.

4.5 SELECCIÓN DEL MODELO

En esta sección, se trata de determinar el modelo más apropiado. La decisión está fundamentada en la distribución *bootstrap* del estadístico y con base a los criterios propuestos por DAVIER (1997) y LANGEHEINE et al. (1996).

4.5.1 BOOTSTRAP PARAMÉTRICO

La distribución del estadístico χ^2 del *bootstrap paramétrico* de **M2**, nos permite aceptarlo como correcto a un nivel de confianza del 97% o menos. Este nivel se incrementa respecto G^2 y *CR*, tal que no existen valores mayores al observado con los datos originales[1].

Por otra parte, con la distribución empírica de los estadísticos de **M1**, se concluye que resulta apropiado con un nivel de confianza inferior al 98%. Esto debido a que existen tres valores que son mayores al calculado con la muestra original.

4.5.2 BOOTSTRAP NO-PARAMÉTRICO

Las investigaciones sobre los efectos de las tablas poco ocupadas en los modelos de variables latentes, se ha dirigido exclusivamente a la simulación con *bootstrap paramétrico*. De esta manera, el presente estudio es pionero en abordar el problema con remuestreo *bootstrap no-paramétrico*.

[1] Un modelo es incorrecto, si la proporción α de los G^2 obtenidos con *bootstrap* es más grandes que el valor original (LANGEHEINE et al., 1996, página 495).

En la sección anterior, se demostró con las pruebas de normalidad que los tres ***EBA*** interés χ^2, G^2 y *CR*, tienen una distribución normal. La ***Tabla 4.11*** muestra los parámetros de la distribución de los estadísticos, el estimado con los datos originales y el valor Z[1].

Para ambos modelos **M1** y **M2**, el estadístico χ^2 de la muestra original es mayor que los valores obtenidos a partir de las muestras *bootstrap*. El valor Z calculado para una variable latente es 391,25 y 66,75 para el modelo de dos variables. Las probabilidades asociadas según la distribución normal estándar son muy pequeñas, por lo cual, los podemos considerar como estimadores pocos probables y concluir, que tanto **M1** y **M2** son apropiados para describir la relación entre los tipos de drogas y las variables latentes.

Estadísticos	Variables latentes	Media θ	Desviación estándar σ_θ	Estimador $\hat{\theta}$	Valor Z $z = \frac{\hat{\theta} - \theta}{\sigma_\theta}$
Chi-cuadrado	M1	8430,03	147,48	66131,62	391,25
	M2	10527,96	275,49	28917,28	66,75
Razón de verosimilitud	M1	6040,63	60,18	575,55	-90,81
	M2	7212,7	86,15	2667,89	-52,75
Cressie-Read	M1	6887,6	91,67	5602,28	-14,02
	M2	8459,81	157,84	7632,13	-5,24

Tabla 4.11. Estandarización de los estadísticos bootstrap no paramétricos.

[1] El *valor Z* es una medida de posición relativa. Describe la posición $\hat{\theta}$ relativa en unidades de la desviación estándar (σ_θ).

Respecto a los estimadores de bondad de ajuste ($\hat{\theta}$) Cressie-Read y razón de verosimilitud de los modelos bajo estudio, estos son más pequeños que las respectivas medias (θ) de las distribuciones empíricas y que todos los valores encontrados con el *bootstrap no paramétrico*. Por esta razón, los valores *Z* de la ***Tabla 4.11*** son negativos, de tal manera que las probabilidades que se presenten estos valores son mínimas. Se puede considerar los modelos **M1** y **M2** como correctos para análisis el problema del consumo de drogas.

4.5.3 ÍNDICE DE DISIMILARIDAD

Al considerar el *índice de disimilaridad* (en inglés, dissimilarity index)[1], el modelo de dos variables latentes y dos clases, daría lugar a una clasificación errónea de aproximadamente un 15% de los casos. Pero el modelo de una variable latente y ocho clases, corregiría mucho el índice y por esto podría seleccionarse como el mejor modelo, tal que reduce los sujetos mal clasificados al 2%.

[1] El índice de disimilaridad, I_D, es definido en términos de las frecuencias observadas y esperadas. Como regla general, valores de I_D menores a 0,05 son considerados pequeños.

CONCLUSIONES

Dado el carácter tanto teórico como práctico de este trabajo, hemos creído oportuno distinguir en este último apartado las aportaciones teóricas de los métodos *bootstrap* para la estimación de modelos de variables latentes en presencia de tablas pocos ocupadas y de las principales conclusiones sobre el consumo de drogas en los jóvenes.

Conclusiones teóricas:

- Existen diferencias en la distribución de probabilidad de los estadísticos de bondad de ajuste entre los *bootstrap* paramétricos y no-paramétricos en presencia de tablas poco ocupadas.

- Los métodos *bootstrap* proporcionan una estimación del sesgo de los estadísticos, como la diferencia entre la media *bootstrap* y el valor estimado por el modelo con los datos originales.

- El principal inconveniente del método *bootstrap* paramétrico proviene del problema de desarrollar un software para generar el remuestreo y las pocas opciones de programas informáticos disponibles en el mercado.

- Es necesario considerar con precaución los resultados obtenidos mediante el programa WINMIRA, ya que hay una suposición de normalidad de los estadísticos de bondad de ajuste siendo en realidad incorrecto.

- El análisis comparativo de la distribución de los estadísticos mediante métodos *bootstrap* paramétrico y no-paramétrico, permiten seleccionar con un mejor criterio, el modelo más apropiado de variables latentes.

- El método *bootstrap* no-paramétrico descansa en menos número de suposiciones al tomar los datos directamente de la muestra original.

Conclusiones sobre el consumo de drogas en los jóvenes:

El examen de los modelos, **M1** y **M2**, nos permite llegar a las siguientes conclusiones:

- Hay un grupo importante de jóvenes que no son consumidores de ningún tipo de droga.

- Un pequeño número de personas son adictos de todas las drogas.

- Existen un conjunto jóvenes que solamente consumen bebidas alcohólicas.

- La mayoría de personas utilizan drogas legales.

En cuanto a las aportaciones de este trabajo, mencionamos:

- Hemos podido demostrar la distribución de probabilidad de los estadísticos de bondad de ajuste para el *bootstrap* paramétrico y no paramétrico.

- Se ha establecido un criterio de decisión para la selección de modelos con el *bootstrap* no-paramétrico.

Tareas pendientes en modelos de clases latentes:

- Demostrar teóricamente la regla de decisión para la selección de los modelos, tanto para los *bootstrap* paramétrico como para los no paramétricos.

- Estudiar el comportamiento de los parámetros de los modelos de variables latentes en presencia de tablas poco ocupadas.

BIBLIOGRAFÍA

AGRESTI, A. (1984). *Analysis of Ordinal Categorical Data*. John Wiley and Sons, Nueva York.

_____ (1990). *Categorical Data Analysis*. New York, Wiley.

AKAIKE, H. (1987). Factor Analysis and AIC. *Psychometrika*, **52(3)**: 317-332.

ANDERSEN, E.B. (1991). *The Analysis of Categorical Data*. Springer-Verlag, Berlin.

ANDERSON, T.W. (1954). On Estimation of Parameters in Latent Structure Analysis. *Psychometrika*, **19**(1): 1-10.

ANDRADE, L.; FAJARDO, M.A.; PÉREZ, J.; CORRALES, N.M. (2002). *Los Modelos Markovianos de Variables Latentes*. ISBN 84-338-2878-9, 117-163. Universidad de la Rioja.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis*. 2ª edition, Oxford University Press, London.

BICKEL, P.J.; KRIEGER, A.M. (1989). Confidence Bands for a Distribution Function using the Bootstrap. *Journal of the American Statistical Association*, **84**(405): 95-100.

BISHOP, Y.M.; FIENBERG, S.E.; HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

CERVIÑO, L.S. (2004). *Estudio de la Incertidumbre Asociada a los Métodos de Evaluación de las Poblaciones de Peces*. Departamento de Ecología y Biología Animal, Universidad de Vigo. Tesis Doctoral.

CHERNICK, M.R. (1999). *Bootstrap Methods: A practitioner's Guide*. Wiley & Sons. Nueva York,

CLOGG, C.C. (1993). *Latent Class Models: Recent Developments and Prospects for the Future*. Handbook of statistical modeling in the social sciences. Plenum, New York.

CLOGG, C. C.; GOODMAN, L. A. (1984). Latent Structure Analysis of a set of Multidimensional Contingency Tables. *Journal of the American Statistical Association*, **79**:762-771.

COLLINS, L.M.; FIDLER, P.L.; WUGALTER, E. S.; LONG, J.D. (1993). Goodness-of-Fit Testing for Latent Class Models. *Multivariate Behavioral Research*, **28**(3): 375-389.

DAVIER, M. (1997). Bootstrapping Goodness-of-Fit Statics for Sparse Categorical Data – Results of a Monte Carlo Study. *Methods of Psychological Research Online*, **2**:29-48.

_____ (2001). WINMIRA 3.2 pro. *A program Systen for Analyses with the Rasch Model, with the Latent Class Analysis and with the Mixed Rasch Model*. Kiel: Institute for Science Education (IPN), Kiel, Germany.

DAVISON, A. C.; D.V. HINKLEY. (1997). *Bootstrap Methods and their Application*. Cambridge. Cambridge University Press.

DIXON, P.M. (2001). “The Jackknife and the Bootstrap”, en: S.M. Scheiner, y J. Gurevitch (eds). *The Design and Analysis of Ecological Experiments*. 2ª ed. Oxford, Oxford University press, 284-288.

EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. **7**: 1-26.

EFRON, B.; TIBSHIRANI, R.J. (1986). Bootstrap Methods for Standards Error, Confidence Intervals, and other Measures of Statistical Accuracy. *Statistical Science*. **1**:54-77.

_____ (1993). *An introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, No. 57. Chapman and Hall, London.

GOODMAN, L.A. (1974). Exploratory Latent Structure Analysis using both Identificable and Unidentificable Models. *Biometrical*, **61**(2): 215-231.

HABERMAN, S. J. (1978). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

_____ (1979). *Qualitative Data Analysis*: Vol. 2, New developments. Nueva York: Academic Press.

HARTLEY, H. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics*, **14**:174–194.

HENRY, N. W.; LAZARSFELD, P. F. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

LANGEHEINE, R.; PANNEKOEK, J.; VAN DE POL, F. (1996). Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis. *Sociological Methods & Research*, 24(4): 492-516.

LIN, T, H.; DAYTON, C. M. (1997). Model Selection Information Criteria for Non-nested Latent Class Models. *Journal of Educational and Behavioral Statistics*, **22**(3): 249-264.

MANLY, F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.

MATHWORKS (2007). *MATLAB: The Language of Technical Computing*. Massachusetts, USA.

MCHUGH, R.B. (1956). Efficient Estimation and Local Identification in Latent Class Analysis, *Psychometrika*, **21**:331-347.

MOONEY, C.Z.A.; DUVAL, R.D. (1993). *Bootstrapping: a Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage.

NOREEN, E. (1989). *Computer Intensive Methods for Testing Hypotheses*. New York: John Wiley & Sons, Ltd.

RAFTERY, A.E. (1986). Choosing Models for Cross-Classifications. *American Sociological Review*, **51**:145-146.

READ, T. R. C; CRESSIE, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springeer-Verlag, New York.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2): 461-464.

SEPÚLVEDA, R. A. (2004). *Contribuciones al Análisis de Clases Latentes en Presencia de Dependencia Local*. Tesis Doctoral, Universidad de Salamanca.

VERMUNT, J. K. (1997a). *Loglinear Models for Event Histories*. Thousand Oaks. Sage Publications.

_____ (1997b). *LEM: A General Program for the Analysis of Categorical Data*. Department of Methodology and Statistics, Tilburg University.

VERMUT, J. K.; MAGIDSON, J. (2002). *Latent Class Cluster Analysis*. In J.A. Hagenaars and A. L. McCutcheon. Advanced Latent Class analysis. Cambridge, UK: Cambridge University Press.