CONVENIO

INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

DE LA UNIVERSIDAD NACIONAL DE ROSARIO

LA PRUEBA DE INDEPENDENCIA EN TABLAS BIDIMENSIONALES CUANDO LA MUESTRA ES DE DISEÑO COMPLEJO

DRA. ELSA SERVY

EST. LETICIA HACHUEL

LIC. DANIEL WOJDYLA

1993

PRÓLOGO

Los servicios de información estadística tienen la doble tarea de analizar los datos que recogen, proveyendo al público de interpretaciones repecto de diversos aspectos de interés general, así como de presentar datos en forma resumida sobre temas más específicos a fin de ofrecer material básico a las investigaciones de usuarios externos.

Este segundo cometido, plantea un dilema entre la brevedad y la minuciocidad, que equivale a decidir qué información es relevante y cuál no lo es.

En diversos estudios (muestra del Censo de 1980, por ejemplo) el Instituto Nacional de Estadística y Censos, INDEC, presenta los datos cuantitativos resumidos en tablas, cuyas casillas contienen estimaciones de las frecuencias, calculadas de acuerdo con el diseño de la muestra , acompañadas de información sobre los efectos de diseños marginales.

Dado un conjunto de estimaciones, emergentes de una muestra con diseño complejo (estratificado, multi-etápica, etc.), los efectos de diseño dan una idea de la medida en que se puede asimilar esas estimaciones a las que resultan de muestras aleatorias simples, y por extensión, aplicar metodologías creadas a partir de distribuciones multinomiales con mayor o menor grado de confianza a datos recogidos según diseños complejos.

El "test" chi-cuadrado de Pearson es una de las estadísticas de las que se ha hecho mayor uso y abuso. Muchos investigadores, sin embargo, han constatado su falta de robustez en el análisis de muestras de diseño complejo. Esto ha llevado a la consideración de estadísticas alternativas, de las que son especialmente importantes las de tipo Wald (Koch, Freeman y Freeman (1977)) y las de tipo de Rao y Scott (1979, 1980, 1981).

Una de las estadísticas que puede considerarse como perteneciente al segundo grupo y que reviste especial interés, por ser directamente aplicable a las tablas, del tipo antes mencionado que publica el INDEC, es la que corrige el chi-cuadrado de Pearson por una constante que es el mínimo de los promedios de los efectos de diseño marginales de las tablas. Sin embargo, sobre el comportamiento de esta estadística no existe suficiente información. De allí que en este estudio se haya dirigido, en parte, a realizar simulaciones que ayudasen a ilustrar sobre ese problema.

Este particular interés condujo a la búsqueda de un modelo para generar tablas bivariadas que dependiesen de dos parámetros: uno para cada uno de los efectos de diseño de sus distribuciones marginales, el cual, además, sirvió para estudiar las estadísticas de tipo Wald y de Rao y Scott.

Como consecuencia del estudio, se ha obtenido un "software" para el análisis de tablas de contingencia, que se pone a disposición del INDEC.

Creemos que las conclusiones de esta investigación, así como el mencionado "software" permitirán que, tanto el INDEC como los consumidores de sus datos, puedan realizar análisis con mayor facilidad y un mayor grado de conciencia sobre las bases que sustentan las metodologías utilizadas.

Este informe alega, principalmente, contra el uso indiscriminado de "recetas estadísticas" que aceptan o rechazan hipótesis como reglas absolutas sin tomar en cuenta los verdaderos niveles de significación y las potencias que les son propias.

El proyecto, "Análisis de muestras complejas", se desarrolla en el marco del convenio entre el INDEC y la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario.

Agradecemos especialmente la colaboración de la Dra. Norma Pizarro de Pereira, Jefa de la Dirección de Estudios Estadísticos del INDEC, que hizo aportes en la discusión del proyecto y acompañó su desarrollo, apoyándonos con su valiosa experiencia. También extendemos nuestro agradecimiento al Licenciado Gerardo Mitas, que nos brindó sugerencias, bibliografía y "software" para el desarrollo de la investigación.

INDICE

1. INTRODUCCIÓN

La inferencia estadística, a partir de datos de encuestas, requiere la verificación de tres supuestos básicos:

a) La estimación muestral del parámetro poblacional es aproximadamente insesgada.

b) Se puede computar a través de la muestra una estimación de la variancia del estimador, aproximadamente insesgada.

c) La distribución de la razón entre, el estimador muestral menos su valor esperado, y su desviación estandard, es aproximadamente normal.

Si el diseño muestral es una selección simple al azar de elementos, si el estimador es una media, una proporción ó un total y si la muestra es grande, a, b y c se cumplen. Sin embargo, no todas las muestras son grandes y peor aún, el supuesto de muestreo al azar irrestricto no se verifica en la mayoría de diseños, donde por razones económicas y/o prácticas se utilizan conglomerados, lo que introduce una correlación intra-clase entre sus elementos.

En la medida que se consideren diseños muestrales que difieren del muestreo simple al azar y/o las estimaciones muestrales sean diferentes a los parámetros clásicos descriptivos, la validez de cada uno de los supuestos a, b y c se tornan cuestionables. La información recogida de cada unidad respondente es, en la mayoría de los casos, de naturaleza multivariada y su análisis se complica si por ejemplo, se pretende realizar estimaciones en dominios, calcular regresiones lineales ó logísticas, analizar tablas de contingencia, etc.

En particular, una muestra compleja provoca dificultades para el cálculo de la variancia del estimador pues la fórmula depende de las probabilidades de selección de cada unidad, y aunque teoricamente ésta pueda explicitarse, su aplicación a datos concretos da lugar a programas de cómputo complicados.

Un pilar importante en la emergencia de ideas y teoría en torno a las encuestas complejas es el concepto de "efecto de diseño" (DEFF) debido a Kish (1965). El efecto de diseño se define como la razón entre la variancia real de una estadística bajo un diseño específico y la variancia que se hubiera alcanzado si la muestra hubiese sido simple al azar y del mismo tamaño. El concepto de "efecto de diseño" ha sido especialmente útil en el análisis de datos de encuestas que involucran conglomeración y estratificación.

En el área de investigaciones socio-económicas, es bien conocida la necesidad de realizar análisis de datos categóricos presentados en forma de tablas de contingencia. Los métodos más difundidos que se utilizan para tal fin, descansan en el supuesto de que los datos se obtienen por muestreo simple al azar de una o más poblaciones. Cuando se satisface este supuesto, los datos de las tablas de conteo se distribuyen de acuerdo con una multinomial o producto de multinomiales. Sin embargo, la mayoría de los estudios a gran escala que se llevan a cabo actualmente poseen diseños que suelen llamarse complejos, y que incorporan estratificación y posiblemente más de una etapa de selección.

En los últimos años se han realizado numerosos progresos en el desarrollo de métodos para el análisis de contingencia con datos provenientes de estudios complejos. Se han propuesto distintos "tests" alternativos que, de alguna manera, tienen en cuenta el diseño complejo de la muestra; entre ellas se encuentran las estadísticas de tipo Wald y las correcciones propuestas por Rao y Scott al Chi-Cuadrado tradicional.

Este trabajo compara el comportamiento de las estadísticas de Wald, Rao y Scott y

Fellegi mediante un estudio de simulación a partir de un modelo cuya importancia radica en que permite generar una tabla bi-variada con efectos de diseño diferentes para filas y columnas y un determinado grado de asociación entre las variables que la definen. Se presenta en la sección 2 una reseña de los principales métodos para el cómputo de la variancia en diseños complejos y en la sección 3 y 4 los diferentes "tests" que se utilizan en el análisis de tablas de contingencia y el resultado de sus comportamientos en estudios previos, respectivamente.

La sección 5 incluye una descripción de los distintos procedimientos llevados a cabo para el cálculo y comparación de las estadísticas. La sección 6 se dedica a la presentación del modelo creado para la generación de datos. La sección 7 describe los procedimientos seguidos en el estudio de Montecarlo mientras que las secciones 8 y 9 se destinan a resultados y conclusiones.

2. CÓMPUTO DE LA VARIANCIA PARA DISEÑOS MUESTRALES COMPLEJOS

Un importante aspecto de la inferencia es que cada estimador debe tener asociado un estimador de su variancia.

Un total es una función lineal simple de las observaciones y es posible derivar expresiones algebraicas explícitas para las variancias estimadas de tales funciones lineales, aún cuando se debe tener en cuenta que si el tipo de diseño es complejo, el estimador de la variancia para el total debe tomar en consideración todas las etapas de selección, requiriendo el cómputo de las probabilidades conjuntas de selección para todas las unidades muestreadas a través de las etapas.

En estudios en gran escala, frecuentemente se supone que los conglomerados de la primera etapa se han seleccionado con reemplazamiento, aún cuando el esquema de selección realmente usado haya sido sin reemplazamiento. Este conduce a una sobreestimación de la variancia pero el sesgo relativo es chico si la fracción de muestreo de la primera etapa es pequeña. Este supuesto, por otro lado, permite computar la variancia del estimador consideran-

do sólo la primera etapa de selección (Des Raj 1968).

Para funciones no lineales del vector de observaciones se dispone de diferentes métodos para aproximar la variancia. Los mismos se describen brevemente a continuación.

2.1. *Método de replicaciones independientes*

Requiere la extracción de varias muestras independientes de la misma población con el fin de obtener estimaciones independientes de la misma estadística θ.

Sean ellas, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$. La media estimada es,

$$\tilde{\theta} = \sum_{i=1}^{M} \frac{\hat{\theta}_i}{M}, \qquad (2.1.1)$$

y un estimador de su variancia:

$$v_r(\tilde{\theta}) = \sum_{i=1}^{M} \frac{(\hat{\theta}_i - \tilde{\theta})^2}{M(M-1)}, \qquad (2.1.2)$$

En la práctica $\tilde{\theta} \neq \hat{\theta}$ y se hace necesario suponer $v_r(\hat{\theta}) = v_r(\tilde{\theta})$.

Esta técnica impone la restricción de un tamaño de muestra grande, ya que cada muestra independiente es mucho menor que la muestra total factible. El número de replicaciones es habitualmente pequeño (2-8) y de aquí que la estimación de la variancia tiene pocos grados de libertad y tiende a ser inestable. (Existe una alternativa para muestras extraídas de la población sin reemplazo, es decir cuando las replicaciones son dependientes).

2.2. *Método de replicaciones de mitades balanceadas*

Fue sugerido por McCarthy (1966) y está diseñado para estudios con exactamente dos unidades muestrales primarias por estrato en la muestra. Para un diseño con L estratos, se elige un subconjunto ortogonal de mitades de muestras entre las 2^L mitades de muestras posibles, eligiendo al azar una unidad primaria en cada estrato. Se forma una estimación $\hat{\theta}_i$ para cada miembro del subconjunto ortogonal y se computa la variancia como,

$$v_B(\hat{\theta}) = \sum_{i=1}^{S} \frac{(\hat{\theta}_i - \hat{\theta})^2}{S}, \qquad (2.2.1)$$

donde $L+1 \leq S \leq L+4$. Este método también impone restricciones sobre el diseño muestral.

2.3. *Método "jackknife"*

Originalmente sugerido por Quenouille (1956) es otro método aplicable a muestras complejas. La correspondiente estadística se computa de la siguiente manera en un contexto estratificado.

Sea θ un parámetro no lineal de interés, $\hat{\theta}$ su estimador a partir de la muestra completa y $\hat{\theta}_{(hi)}$ el estimador de θ después de omitir la i-ésima unidad dentro del h-ésimo estrato. Se define:

$$\hat{\theta}_{(h)} = \sum_{i=1}^{n_h} \frac{f_{(hi)}}{n_h}, \qquad (2.3.1)$$

(donde n_h es el número de unidades dentro del h-ésimo estrato).

Entonces el estimador de la variancia jackknife de $\hat{\theta}$ es,

$$v_j(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_{h-1}}{n_h} \sum_{i=1}^{n_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2 , \qquad (2.3.2)$$

2.4. *Método de linearización*

Expresa el parámetro θ como una función g(Y) de $Y = (Y_1,.., Y_p)$ donde Y_j es un total de una variable dada en una población.

Un estimador consistente de g(Y) es $g(\hat{Y})$ donde $\hat{Y} = (\hat{Y}_1,..., \hat{Y}_p)$ es el estimador de Y obtenido de acuerdo al diseño.

Si se supone ahora que,

$$\hat{Y}_j = \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi}\, y_{hij} , \qquad (2.4.1)$$

se tiene que:

$$\begin{aligned} \text{Var}\left[g(\hat{Y})\right] &\doteq V\left[\sum_{j=1}^{p} (\hat{Y}_j - Y_j)\, g^{(j)}(Y)\right] \\ &= V\left[\sum_{j=1}^{p} \hat{Y}_j\, g^{(j)}(Y)\right] \\ &= \sum_{h=1}^{L} V\left[\sum_{i=1}^{n_h} w_{hi}\, z_{hi}\right] \end{aligned} \qquad (2.4.2)$$

donde:

$$z_{hi} = \sum_{j=1}^{p} y_{hij}\, g^{(j)}(\hat{Y}) \quad y \quad g^{(j)}(\hat{Y}) = \frac{\partial g(Y)}{\partial \hat{Y}_j}\Big|_{\hat{Y}_j = Y_j} . \tag{2.4.3}$$

Un estimador consistente de $V\ [g(\hat{Y})]$ es:

$$v[g(\hat{Y})] = \sum_{h=1}^{L} v(\hat{z}_h) , \tag{2.4.4}$$

donde,

$$\hat{z}_h = \sum_{i=1}^{n_h} w_{hi}\, \hat{z}_{hi} \quad y \quad \hat{z}_{hi} = \sum_{j=1}^{p} y_{hij}\, g^{(j)}(\hat{Y}) . \tag{2.4.5}$$

Esta formulación de la variancia, por medio de la aproximación de Taylor, fue dada por Woodruff (1971) y luego mejorada computacionalmente por Fuller (1975) y Binder (1983). La ventaja de este método sobre los otros es que no impone restricciones sobre el diseño muestral, no es difícil computacionalmente, y puede usarse para estimar componentes de variancia. Computacionalmente, sólo se necesita calcular la combinación lineal dada por $\hat{z}_{hi}$ y se pueden usar los algoritmos usuales para la variancia de totales para diseños con varias etapas de selección.

Debido a que, en el caso de estadísticos no lineales, todos los métodos involucran aproximaciones, se ha estudiado el sesgo y la precisión de la variancia comparando estos métodos en términos de su potencial para estimar bien las variancias de interés. Así, diversos estudios indicaron que todos los métodos conducen a buenos estimadores de la variancia para varios estadísticos: razones y medias post-estratificadas, coeficientes de regresión y correlación simple y parcial. -

Estas conclusiones resultaron válidas tanto para pequeñas como para grandes muestras.

La elección de un método depende, por lo tanto, de la flexibilidad del diseño muestral, disponibilidad de programas, etc. Además, estos métodos pueden adaptarse para el cómputo de matrices de covariancias.

3. ANÁLISIS DE TABLAS DE CONTINGENCIA

En el caso de análisis de tablas de contingencia multidimensionales, el "test" chi-cuadrado (o el "test" de la razón de verosimilitud) se usa frecuentemente para la evaluación y selección de modelos parcos ("parsimonious") que describen las probabilidades de las celdas poblacionales. Este método puede conducir a inferencias erróneas ya que la tasas de error de tipo I de los "tests" de hipótesis pueden llegar a ser mucho mayor que los niveles nominales.

A continuación se presenta la teoría asintótica que da lugar a la postulación de diversos "tests" que tienen en cuenta el diseño muestral.

3.1. *Efectos de Diseño y Teoría Asintótica*

Para extender el concepto de DEFF al conjunto de proporciones que aparecen en una tabla de contingencia, Rao y Scott (1979, 1981) definen la matriz,

$$D = P^{-1}V \quad , \tag{3.1.1}$$

donde V y P son las matrices de covariancias del conjunto de proporciones estimadas de la tabla, tomando en cuenta el diseño muestral y sin tenerlo en cuenta, respectivamente. Estos autores proponen como DEFF generalizados a los valores característicos de la matriz D, los cuales se estiman por los valores característicos de:

$$\hat{D} = \hat{P}^{-1}\,\hat{V}, \tag{3.1.2}$$

donde,

$$\hat{P} = diag(\hat{\pi}) - \hat{\pi}\hat{\pi}', \tag{3.1.3}$$

y $\hat{\pi}$ es el vector que contiene a las estimaciones de las proporciones de la tabla.

Se presenta a continuación una descripción breve de la teoría asintótica que avala los procedimientos que luego se utilizan.

Sea, ahora, una tabla de r filas y c columnas y sea $\pi = (\pi_{00},\ldots,\ \pi_{(r-1)(c-1)})'$ el vector de probabilidades, $\left(\sum\sum\pi_{ij} = 1\right)$. Sea el vector de probabilidades estimadas $\hat{\pi} = (\hat{\pi}_{00},\ldots,\hat{\pi}_{(r-1)(c-1)})$ tal que, si existe alguna versión apropiada del teorema central del límite, se tiene:

$$\sqrt{n}\ (\hat{\pi} - \pi) \rightarrow N(0,\ V)\ . \tag{3.1.4}$$

Nótese que debido a que $\hat{\pi}$ tiene dimensión rc y $\sum\sum \hat{\pi}_{ij} = 1$, la matriz de covariancias V, será singular.

Se definen, además, los vectores de probabilidades marginales:

$\pi_r = (\pi_{0+}, \ldots, \pi_{(r-1)+})'$ donde $\pi_{i+} = \sum_{j=0}^{(c-1)} \pi_{ij}$, i = 0,..,(r-1), y,

$\pi_c = (\pi_{+0}, \ldots, \pi_{+(c-1)})'$ donde $\pi_{+j} = \sum_{i=0}^{(r-1)} \pi_{ij}$, j = 0,.., (c-1).

Considérese una hipótesis general sobre los parámetros π_{ij}, formalizada como

$$H_o: h_l(\pi) = 0 \qquad (l = 1,\ldots, b) \tag{3.1.5}$$

donde las $\{h_l(\pi), l = 1,..,b\}$ son funciones que se anulan si la hipótesis se verifica.

Sea $H(\pi)$ la matriz de rango b de derivadas parciales:

$$H(\pi) = \left(\frac{\partial h_l(\pi)}{\partial \pi} \right)_{l=1\ldots b} . \tag{3.1.6}$$

Bajo los supuestos antes mencionados,

$$\sqrt{n}\ (h(\hat{\pi}) - h(\pi)) \tag{3.1.7}$$

es asintóticamente normal con media 0 y matriz de variancias y covariancias HVH' donde $h(\pi) = (h_1(\pi),\ldots, h_b(\pi))'$ y $H = H(\pi)$.

Si se dispone de un estimador consistente de V se puede realizar el "test" a través de la estadística del tipo de las de Wald:

$$X^2_W = n\, h(\hat{\pi})' \, (\hat{H}\hat{V}\hat{H}')^{-1} \, h(\hat{\pi}) \,, \qquad (3.1.8)$$

que se distribuye asintóticamente bajo Ho como χ^2_b . Es posible también utilizar una versión modificada de (3.1.8) con V (y H) estimada bajo Ho.

Cuando no se dispone de una estimación de V, es común ignorar la estructura muestral y usar la matriz de variancias multinomial P en lugar de V, esto es:

$$X^2 = n\, h'(\hat{\pi}) \, (\hat{H}\hat{P}\hat{H}')^{-1} \, h(\hat{\pi}) \qquad (3.1.9)$$

El comportamiento asintótico de X^2 bajo un esquema muestral general es relativamente simple de obtener ya que se puede expresar como:

$$X^2 = \sum_{l=1}^{(r-1)(c-1)} \delta_l \, z_l^2 \qquad (3.1.10)$$

donde los $\{z_l\}$ son asintoticamente independientes N(0,1) bajo H_o y los $\{\delta_l\}$, son los autovalores de:

$$D = (HPH')^{-1} \, (HVH') \,. \qquad (3.1.11)$$

Los δ_l se pueden interpretar como efectos de diseño (en este caso de las componentes de $H(\hat{\pi})$).

3.2. *Estadísticas de tipo Wald*

Un buen ejemplo del uso del enfoque de Wald aparece en el artículo de Grizzle, Starmer y Koch (1969), quienes utilizan mínimos cuadrados ponderados (WLS) en el análisis de datos categorizados. Presentan en su trabajo un método unificado que permite realizar "tests"

de bondad de ajuste, independencia, análisis de modelos anidados, análisis de "logits", etc. A estas siguieron muchas publicaciones usando el enfoque en nuevos problemas relacionados, como en el artículo antes mencionado, con la distribución multinomial.

En 1977, Koch, Freeman y Freeman presentan el mismo enfoque para el análisis de variables tanto cualitativas como cuantitativas que provienen de muestras complejas, es decir, muestras que incluyen selección de conglomerados, estratificación y selección con etapas múltiples, y que, además están realizadas en gran escala, como ocurre con las encuestas conducidas por los Sistemas Nacionales de Estadísticas de diferentes países. En particular, el interés recae sobre la estimación de características en sub-dominios a través de razones apropiadas.

Dado que las muestras en tales situaciones son generalmente muy amplias, se puede suponer que las estimaciones de diversas características pertenecientes a diferentes subpoblaciones tienen, aproximadamente, una distribución normal multivariada, con una matriz de variancias y covariancias que puede ser estimada, ya sea directamente o mediante métodos de re-muestreo.

El núcleo del artículo está expuesto en el ítem 2.8 sobre el "análisis de estimaciones para datos provenientes de encuestas con diseño complejo". Su resumen se da a continuación.

Sea $F = \phi(\hat{\pi})$ un vector de estadísticas de dimensión g, tales como estimadores de medias ($\bar{y}$) en diferentes dominios, o estimadores por razón u otras funciones de ellos. Sea V_F una estimación consistente de la matriz de covariancias de F, obtenida por métodos directos de estimación o re-muestreo.

La relación entre F y ciertos aspectos de interés acerca de la naturaleza de varias subpoblaciones o dominios puede caracterizarse por

$$F = X b, \qquad (3.2.1)$$

donde X es una matriz pre-fijada de diseño, con coeficientes conocidos, y rango completo u, y b un conjunto de u parámetros.

Las estadísticas que se proponen para estimar el modelo y su adecuación a los datos,

son:

$$\hat{b} = V_b X' V_F^{-1} F \quad \text{con} \quad V_b = (X' V_F^{-1} X)^{-1} , \tag{3.2.2}$$

$$Q = (F - X \hat{b})' V_F^{-1} (F - X \hat{b}) , \tag{3.2.3}$$

donde V_F se supone no singular.

Una vez que se determinó que el modelo es adecuado, se pueden formular hipótesis lineales del tipo $Cb = 0$ donde C es una matriz de orden (d x u), y juzgarlas usando la estadística,

$$Q_C = b' C' \left| C' V_b^{-1} C \right|^{-1} C b , \tag{3.2.4}$$

que se distribuye aproximadamente, como un χ^2 con d grados de libertad bajo la hipótesis nula.

Finalmente, los valores predichos se pueden calcular así,

$$\hat{F} = X \hat{b} = X (V_b)^{-1} X' V_F^{-1} F , \tag{3.2.5}$$

El trabajo de Koch y colaboradores presenta luego algunas aplicaciones a problemas surgidos en relación a datos del National Center for Health Statistics.

3.3. *Estadísticas de Rao y Scott*

Rao y Scott (1979,1980,1981) realizaron un estudio sistemático sobre el impacto del diseño del estudio sobre el chi-cuadrado estándar de bondad de ajuste. Encontraron que, con esquemas muestrales generales, se obtiene una buena corrección de la estadística (3.1.9), calculando,

$$\chi^2_{R_1} = \frac{X^2}{\hat{\delta}.} \quad , \quad \hat{\delta}. = \sum_{l=1}^{b} \frac{\hat{\delta}_l}{b} \tag{3.3.1}$$

que se distribuye como χ^2_b bajo Ho. $\{\hat{\delta}_l,\ l=1,..,b\}$ son los autovalores de la matriz $\hat{D}$ que estima a la matriz definida en (3.1.11).

Se estima $\delta.$ a partir de las estimaciones de las variancias y covariancias de las $\{\ h_l(\hat{\pi}), l=1,...,b\ \}$ que puede ser a veces computacionalmente compleja.

Los autores (1981) proponen, para sustituir adecuadamente el cálculo de $\hat{\delta}.$, el uso de :

$$\chi^2_{R_2} = \frac{X^2}{\hat{\lambda}.} \quad , \hat{\lambda}. = \sum_{t=1}^{(rc-1)} \frac{\hat{\lambda}_t}{(rc-1)}, \tag{3.3.2}$$

donde con $\hat{\lambda}_t$ se simboliza un autovalor de $\hat{P}^{-1}\hat{V}$, matriz que es más simple que la (3.1.11) y de uso más general pues sirve para el "test" de más de una hipótesis respecto de los $\{\pi_{ij}$; i = 0,.., (r-1); j = 0,.., (c-1)$\}$.

Rao y Scott (1981) proporcionan una forma de cálculo de $\hat{\lambda}.$ que depende solo de información parcial de la matriz de variancias y covariancias de $\hat{\pi}$. Sin embargo, aún así $\hat{\lambda}.$ requiere el cálculo de los efectos de diseño $\{\hat{d}_{ij}\ i=0,..,(r-1);\ j=0,..,(c-1)\}$ de todos las proporciones estimadas en la tabla $\{\hat{\pi}_{ij}\ i=0,..,(r-1);\ j=0,..,(c-1)\}$, y esta información

puede no estar disponible. En el mejor de los casos, la única información que se dispone en la práctica son las variancias de $\{\hat{\pi}_{i+},\hat{\pi}_{+j} \quad i=0,..,(r-1);j=0,..,(c-1)\}$, es decir de las estimaciones de las probabilidades marginales de la tabla, lo que ha llevado a considerar la estadística que se describe más adelante, (3.4).

El "test" corregido (3.3.1) es asintoticamente válido en el caso de efectos de diseño de conglomerados constantes y su comportamiento es bueno cuando la variabilidad de los $\{\delta_l;\ l = 1,..,\ b\}$ es pequeña.

Pueden encontrarse aplicaciones del enfoque de Rao-Scott a una variedad de situaciones.

i) Bondad de ajuste en tablas unidimensionales

ii) Tests de homogeneidad de proporciones y de asociación en tablas bidimensionales

iii) Análisis de tablas tridimensionales y de mayor orden usando modelos log-lineales (Hidiroglou y Rao 1987).

3.4. *Estadísticas de Fellegi y Corrección por el "Deff" promedio marginal mínimo*

Fellegi (1978) sugirió el uso del promedio de los efectos de diseño de las estimaciones de las proporciones de las celdas de una tabla de contingencia, como divisor de la estadística chi-cuadrado.

$$\chi^2_{R_3} = \frac{X^2}{d_\bullet}, \quad d_\bullet = \sum_{ij} \frac{d_{ij}}{rc}. \tag{3.4.1}$$

Otra corrección consiste en calcular:

$$\chi^2_{R_4} = \frac{X^2}{d_m}, \tag{3.4.2}$$

siendo d_m el mínimo de $\left\{ \sum_{i=1}^{(r-1)} \frac{d_{i+}}{r}, \sum_{j=0}^{(c-1)} \frac{d_{+j}}{c} \right\}$, donde $\{d_{i+}, i = 0,.., (r-1)\}$ son los efectos de diseño marginales correspondientes a $\{\hat{\pi}_{i+}, i = 0,..,(r-1)\}$, mientras que los $\{d_{+j}, j = 0,.., (c-1)\}$ tienen igual definición para el otro márgen.

3.5. *Otras versiones de las estadísticas mencionadas*

A continuación se presentan otras estadísticas, ó versiones diferentes de las ya mencionadas, que suelen aparecer en publicaciones sobre el tema. Ellas son las expresadas en las fórmulas (3.5.1/11).

$$F_W = \frac{(n - f + 1)}{fn} \chi^2_w , \qquad (3.5.1)$$

donde f son los grados de libertad de la variable chi-cuadrado y n es el número de conglomerados en la muestra.

F_w es una modificación de χ_w^2 ,(Koch, Freeman y Freeman,1977) y su distribución es la de una variable F de Snedecor con f y (n-f+1) grados de libertad.

$$G^2 = 2\, n\, k \sum \hat{\pi} \ln\left(\frac{\hat{\pi}}{\pi}\right) , \qquad (3.5.2)$$

que se denomina cociente de máxima verosimilitud y se distribuye aproximadamente como un chi-cuadrado con f grados de libertad. La corrección de Rao y Scott aplicada a dicha estadística, da lugar a otra, para el mismo "test," que se indica con G^2_R (3.5.3).

Se obtienen "tests" alternativos calculando:

$$FX_R^2 = \frac{X_R^2}{f}, \tag{3.5.4}$$

$$FG_R^2 = \frac{G_R^2}{f}, \tag{3.5.5}$$

y usando como distribución aproximada la de la variable F con f y (n-1) f grados de libertad.

Una corrección de segundo orden que tiene en cuenta la variabilidad de los $\{\delta_l; l = 1,.., b\}$, se obtiene aplicando la aproximación de Satterthwaite a la suma ponderada de variables independientes χ^2 (Rao y Scott 1984). Dichos "tests" requieren el conocimiento completo de la matriz de variancias y covariancias estimada de $\hat{\pi}$.

Si se utiliza dicha corrección, se tienen:

$$\chi^2_{RS} = \frac{X_R^2}{(1 + \hat{a}^2)} \tag{3.5.6}$$

$$G^2_{RS} = \frac{G_R^2}{(1 + \hat{a}^2)} \tag{3.5.7}$$

donde $\hat{a}$, es el coeficiente de variación de los autovalores estimados de la matriz de "deffs" generalizados. Dichos "tests" nuevamente requieren el conocimiento de la matriz completa estimada de covariancias de $\hat{\pi}$

Además,

$$\chi^2_{Ro},\ G^2_{Ro},\ \chi^2_{RSo},\ G^2_{RSo} \tag{3.5.8}$$

son estadísticas similares a $\chi^2_R,\ G^2_R,\ \chi^2_{RS},\ G^2_{RS}$, pero utilizando como matriz de variancias y covariancias, dicha matriz bajo la hipótesis nula.

Se presentan además, la versión "jackknifed" del X^2 de Pearson,

$$X^2_J \tag{3.5.9}$$

versión "jackknifed" del cociente de verosimilitud, y

$$G^2_J \tag{3.5.10}$$

que tienen la misma definición que X^2 y G^2, pero utilizando como matriz de variancias y covariancias la estimación obtenida por el método "jackknife". En tanto que,

$$FX^2_J \quad , \quad FG^2_J \tag{3.5.11}$$

son idénticas a (3.5.9) y (3.5.10) pero utilizando la transformación indicada en (3.5.4) y (3.5.5) que remite a las tablas de la distribución F.

3.6. *Otras estadísticas*

Existen otras estadísticas que provienen de considerar el "test" chi-cuadrado bajo modelos para muestreo por conglomerados.

Entre ellos se encuentran la estadística de Cohen (1976) para conglomerados de tamaño 2 y su extensión para conglomerados de mayor tamaño realizada por Altham (1976).

Estos autores proponen un modelo para formalizar la idea de que existe independencia entre conglomerados y dependencia (positiva) entre los individuos de un mismo conglomerado. El modelo depende de una constante "a", que proporciona una medida de la asociación entre los individuos dentro del conglomerado.

La estadística que propone Cohen para el "test" de la hipótesis de independencia entre dos variables que conforman una tabla de contingencia de dimensión rxc es:

$$W^2 = \frac{X^2}{1 + \hat{a}} \qquad (3.6.1)$$

donde X^2, es la estadística convencional para llevar a cabo el "test" de independencia de dos variables categóricas y $\hat{a}$ es la estimación máximo-verosímil de la constante que aparece en el modelo de dependencia entre los individuos de un conglomerado. Bajo estas condiciones W^2 se distribuye asintóticamente como χ^2 con (r-1)(c-1) grados de libertad.

Altham extiende el procedimiento para el caso de conglomerados de tamaño fijo k. Sin embargo, el modelo propuesto por los autores resulta muy restricto ya que el mismo implica el mismo "deff" para todas las celdas individuales y combinaciones lineales de las mismas.

En la misma linea, Brier (1978) considera un modelo de dependencia para muestreo a dos etapas (con un número constante de elementos muestreados dentro de cada conglomerado) que involucra distribuciones Dirichlet-multinomial, $DM_t(m,\pi,v)$. La estadística que modifica el chi-cudrado convencional es:

$$\{(v + 1) / (v + m)\} X^2 \qquad (3.6.2)$$

que tiene una distribución asintótica χ^2 con b grados de libertad, bajo Ho.

Nuevamente, este modelo de dependencia se traduce en "deff" constantes para las celdas y para las combinaciones lineales de las mismas y desafortunadamente el supuesto de efecto

de diseño único no se corresponde con las situaciones más frecuentes que se presentan en la vida real.

Por ejemplo, Kish (1976) registró efectos diseño promedio para características socio-económicas en el rango de 4 a 8, miestras que en el área demográfica el rango es de 1 a 1.6. Así, una tabla bidimensional formada por una variable socio-económica y otra demográfica no tendría un efecto de diseño común.

Aunque escasos, hay estudios sobre el comportamiento de $\delta.$, en los cuales éste tiende a ser menor que los efectos de las celdas individuales. Los resultados de estudios empíricos confirman esta tendencia.

Otros aportes más recientes se deben a Scott, Rao y Thomas (1989) quienes han desarrollado un método unificado para modelos singulares y no singulares. En una publicación posterior, Rao, Kumar y Roberts (1989) exponen ese método en el caso de variables de tipo categórico, aunque tal vez su alcance pueda ser más amplio y extenderse a otros tipos de variables. Según este artículo,

$$F = \Phi(\hat{\pi}) \tag{3.6.3}$$

donde $\hat{\pi}$ es una estimación muestral de las probabilidades de las celdas y Φ una función derivable. El modelo puede escribirse así:

$$\hat{\Phi} = \hat{\Phi}(\hat{\pi}) = X\beta + \delta \tag{3.6.4}$$

donde δ es el término de error tal que $P(\lim \delta) = 0$ y $\hat{\Phi}$ tiene una matriz de covariancia asintótica que es singular, $V_\phi = H\, V_\pi\, H'$, estimada consistentemente por $V_{\hat{\pi}}$, con $H = (\partial\Phi / \partial\pi)$ evaluada en $\hat{\pi}$.

4. COMPORTAMIENTOS DE LOS "TESTS"

4.1. *Introducción*

En el caso del "test" basado en el X^2 convencional, es bien conocido que sobreestima el nivel de significación cuando el plan de muestreo utiliza conglomerados de unidades que estan positivamente correlacionadas entre sí. Sin embargo, Rao y Scott (1980) muestran que el X^2 estándar puede ser conservador en caso de un plan de muestreo estratificado que no usa conglomeración, si el número de estratos es igual a 2.

No existe un conocimiento exhaustivo del comportamiento de los "tests" alternativos. Se han estudiado especialmente sus efectivos niveles de significación pero sus potencias son poco conocidas o explicitadas, especialmente cuando el tamaño de la muestra es chico o moderado.

Se han realizado principalmente dos tipos de estudios para conocer el comportamiento de los "tests". Ellos se basan en :

1. Estudios empíricos que consisten en el análisis de encuestas reales de gran escala y utilizan los métodos aproximados para estimar los niveles de significacion de los "tests" asintóticos.

2.- Estudios de simulación para estudiar los niveles de significacion y la potencia de diversos "tests" cuando la muestra es de tipo complejo pero su tamaño es chico o moderado.

4.2. *Investigaciones empíricas*

Dos encuestas del Reino Unido han servido de base para los estudios empíricos llevados a cabo por Holt y col. (1980) y Rao y Scott (1981). Ellas son: la Encuesta General de Hogares (GHS) de 1971 y el Estudio de elecciones británicas (BES) de 1974.

Ambas utilizaron una muestra multietápica estratificada de más de 13.000 hogares y 2.500 individuos respectivamente.

4.2.1. Caso del "test" de independencia.

Para tener idea sobre el tamaño de δ. (3.3.1) , y su relación con los efectos de diseño por celda y marginales, Holt y col.(1980)examinaron un gran número de variables de los dos grandes estudios mencionados.

En todos los casos la variabilidad de $\{\delta_l,\ l=1,..,b\}$ fué pequeña, por lo que sólo publicaron el valor de δ. (3.3.1). Muestran valores de d. (3.4.1) calculados a partir de las variancias por celda $\{v(\hat{\pi}_{ij}),\ i=0,..,(r-1),\ j=0,..,(c-1)\}$,

TABLA 4.1

Efecto de diseño para variables de BES y niveles de significación para diferentes procedimientos con un nivel nominal de 5%. (las variables del estudio se denominan B2, B3, B5, B7, B10, B12, B14, B15 y B16)

Variables	g.l.	δ.	d.	Valor de χ^2_w	Tam X^2	Tam $\chi^2_{R1\cdot}$	Tam $\chi^2_{R3\cdot}$	Tam $\chi^2_{R4\cdot}$
B2xB3	3	1.30	1.88	41.40	11	5	1	1
B2xB5	2	1.29	1.76	127.28	10	5	2	3
B2xB7	3	1.41	1.81	5.56	13	5	2	4
B2xB10	2	1.22	1.72	32.59	9	5	2	4
B2xB12	2	1.17	1.73	0.31	8	5	1	4
B2xB14	5	1.18	1.39	33.76	9	5	2	6
B2xB15	1	1.16	2.08	1.77	7	5	1	4
B2xB16	1	0.96	1.98	0.02	5	5	0	4
B3xB5	6	1.03	1.31	10.81	6	5	2	1
B3xB7	9	1.13	1.37	6.86	10	6	2	1
B3xB10	6	1.03	1.30	7.84	6	5	2	2
B3xB12	6	1.07	1.33	11.53	7	5	2	3
B3xB14	15	1.01	1.15	49.83	6	6	2	3
B3xB15	3	1.16	1.47	5.79	9	5	2	4
B3xB16	3	1.09	1.44	5.23	7	5	2	6
B5xB7	6	0.97	1.19	20.59	5	5	2	1
B5xB10	4	1.07	1.24	19.74	7	5	3	2
B5xB12	4	1.07	1.21	7.02	7	5	3	3
B5xB14	10	1.09	1.14	29.20	9	5	5	5
B5xB15	2	0.84	1.21	2.24	3	5	2	1
B5xB16	2	1.12	1.30	146.75	7	5	3	6
B7xB10	6	1.03	1.19	17.89	6	5	3	2
B7xB12	6	0.94	1.16	18.35	4	5	2	1
B7xB14	15	1.07	1.12	653.96	8	6	4	4
B7xB15	3	1.00	1.26	0.58	5	5	3	3
B7xB16	3	1.02	1.25	2.52	6	5	3	5
B10xB12	4	1.02	1.14	6.89	5	5	3	2
B10xB14	10	1.05	1.11	40.10	7	5	5	5
B10xB15	2	0.97	1.13	4.94	5	5	3	2
B10xB16	2	1.02	1.15	18.68	5	5	4	5
B12xB14	10	1.01	1.07	39.38	6	5	5	3
B12xB15	2	1.03	1.14	9.03	6	5	4	3
B12xB16	2	0.96	1.09	7.93	5	5	4	4
B14xB15	5	0.99	1.07	7.30	5	5	3	3
B14xB16	5	0.94	1.04	16.07	4	5	3	3
B15xB16	1	1.02	1.14	2.39	5	5	4	5

y además valores de la estadística de Wald para dar alguna idea del grado de dependencia de las variables. La Tabla 4.1 muestra los valores de $\delta.$, $d.$ y X^2_w para cruces de variables del estudio BES. La tabla también contiene los tamaños estimados de los "tests" basados en X^2, $\chi^2_{R3} = X^2/ d.$ y $\chi^2_{R4} = X^2/ d_m$ donde d_m es el menor de los promedios de los dos efectos marginales.

La principal razón para considerar $X^2/d.$ y X^2/d_m es que $\delta.$ es raramente disponible. La conclusión que extraen es que ambos "tests" son conservadores y que el "test" basado en d. es ligeramente mejor que aquel basado en los efectos marginales. Si se corrige por el menor de los efectos marginales, el "test" es muy conservador y como esta modificación por lo general es severa, se pierde potencia. Esta pérdida de potencia puede no ser muy importante si el tamaño muestral es grande, pero es crucial en una investigación pequeña.

Rao y Scott (1981) usan los datos de los estudios mencionados para investigar el comportamiento de los "tests" que proponen. La Tabla 4.2 muestra algunos de los resultados parciales que obtuvieron. Presentan los niveles de significación efectivos para las estadísticas X^2 convencional (3.1.9), $\chi^2_{R_2} = \frac{X^2}{\hat{\lambda}.}$ (3.3.2) de Rao y Scott y $\chi^2_{R_1} = \frac{X^2}{\hat{\delta}.}$ (3.3.1) de los mismos autores. La hipótesis que consideran en los tres casos es la de independencia, formalizada como $h_1(\pi) = \pi_{ij} - \pi_{i+} . \pi_{+j} = 0$.

Los verdaderos niveles de significación fueron aproximados por el método de Solomon y Stephens (1977). Estos autores consideran las distribuciones de formas cuadráticas del tipo:

$$Q_k = \sum_{j=1}^{k} c_j (x_j + a_j)^2 \qquad (4.2.1.1)$$

donde los x_j son variables independientes e identicamente distribuidas como normales estandarizadas (media cero y variancia uno) y donde c_j y a_j son constantes no negativas.

TABLA 4.2

Tamaños asintóticos estimados de los "tests" basados en X^2, $X^2 / \hat{\delta}_.$, $X^2 / \hat{\lambda}_.$ para clasificaciones cruzadas de variables seleccionadas de la Encuesta General de Hogares del Reino Unido de 1971. Tamaño nominal : 0.05. (Las variables del estudio se denominan G1, G2, G3, G4, G5 y G6)

Clasificación Cruzada	r x c	$\hat{\delta}_.$	$\hat{\lambda}_.$	Tamaño X^2	Tamaño $\chi^2_{R_1} = X^2/\hat{\delta}_.$	Tamaño $\chi^2_{R_2} = X^2/\hat{\lambda}_.$
G1 x G2	2 x 2	1.99	3.16	.16	.05	.01
G1 x G3	2 x 3	1.97	2.36	.22	.05	.03
G1 x G4	2 x 3	1.24	1.98	.09	.05	.01
G1 x G5	2 x 6	.91	1.23	.04	.05	.02
G1 x G6	2 x 3	.97	1.75	.05	.05	.01
G2 x G3	2 x 3	1.94	2.49	.21	.05	.03
G2 x G4	2 x 3	1.41	1.86	.12	.05	.02
G2 x G5	2 x 6	1.02	1.18	.06	.05	.03
G2 x G6	2 x 3	1.13	1.61	.08	.05	.02
G3 x G4	3 x 3	1.26	1.72	.11	.05	.01
G3 x G5	3 x 6	.93	1.14	.03	.05	.02
G3 x G6	3 x 3	.96	1.51	.05	.05	.01
G4 x G5	3 x 6	.94	1.05	.05	.05	.03
G4 x G6	3 x 3	.93	1.21	.04	.05	.02
G5 x G6	6 x 3	.85	.94	.03	.05	.04

Este tipo de distribución surge en los problemas estadísticos concernientes con la teoría asintótica del X^2 de Pearson, cuando se permite que los datos de las celdas sean dependientes.

Proponen dos aproximaciones para Q_x:

1. Ajustar una curva de Pearson con los mismos cuatro momentos que Q_x.
2. Ajustar $Q_x = A\ w^g$ donde w tiene una distribución chi-cuadrado con v grados de libertad y A, g y v se determinan igualando los tres primeros momentos de Q_x a

los de Aw^g.

Rao y Scott utilizan este último método para obtener la verdadera distribución de X^2, considerando que éste equivale a la combinación lineal $\sum_{l=1}^{b} \delta_l z_l^2$ donde $\{z_l;\ l = 1,..,b\}$ son variables aleatorias estandarizadas independientes. Luego, la distribución de esa combinación lineal se aproxima a la de una variable $A\ w^g$ donde A y g son constantes y w es una variable χ^2 con v grados de libertad. La aproximación se obtiene igualando los momentos de ambas variables, lo que permite determinar los valores de A, g y v que corresponden a los $\{\delta_l;\ l = 1,.., b\}$, y por su intermedio a la tabla de datos que dió origen a los mismos.

Se puede observar en la Tabla 4.2, que el X^2 es demasiado liberal (especialmente si los efectos de diseño son importantes), el $\frac{X^2}{\hat{\lambda}.}$ demasiado conservador y el $\frac{X^2}{\hat{\delta}.}$ es el que mejor comportamiento tiene, pero a la vez es el que requiere mayor conocimiento de los datos originales.

4.2.2. *Caso de bondad de ajuste : Ho: $\pi = \pi_\sigma$*

Los resultados aparecen en la Tabla 4.3.

TABLA 4.3

Tamaños asintóticos estimados de los "tests" basados en X^2 y $\chi^2_{R_2}$ para algunos items de la Encuesta General de Hogares del Reino Unido de 1971; Tamaño nominal 0.05.

Variable	N° de Categorias	k	$\hat{\lambda}.$	Tam (X^2)	Tam (χ^2_{R2})
G1: Antiguedad Construcción	3	33.1	3.42	.41	.05
G2: Tipo de Dueño	3	33.4	2.54	.37	.06
G3: Tipo de Comodidades	4	27.7	2.17	.30	.06
G4: Número de Habitaciones	10	34.6	1.19	.14	.06
G5: Ingreso Bruto Semanal del	6	26.6	1.14	.10	.06
G6: Edad del Jefe del Hogar	3	34.6	1.26	.10	.05

Se analizan varias variables destacando el número medio de elementos (k) por conglomerado y la media de los autovalores $\hat{\lambda}_{\cdot}$, como un indicador de los efectos de diseño. Las estadísticas que se comparan son el X^2 estándard de bondad de ajuste y el χ^2_{R2} (que en el caso de bondad de ajuste coincide con χ^2_{R1}). El primero tiene un comportamiento mucho más distorsionado en este caso de bondad de ajuste que en el anterior, de prueba de independencia. χ^2_{R2} se muestra como una estadística aceptable.

4.3. *Simulaciones*

Thomas y Rao (1987) realizaron un estudio con métodos Monte Carlo para determinar el nivel de significación y la potencia de estadísticas simples de bondad de ajuste bajo muestreo en conglomerados en pequeñas muestras. Ellos compararon las siguientes estadísticas:

- "Test" X^2 estándar de Pearson.
- "Test" basados en cocientes de verosimilitud G^2 (3.5.2).
- "Tests" basados en los mínimos cuadrados ponderados con estadísticas de tipo χ^2_W (Koch y col., 1975).
- Idem con la corrección F_w, según (3.5.1).

  que se compara con variables F con (T-1) y (n-T-1) grados de libertad, donde T es el número de clases y n es el número de conglomerados en la muestra.
- "Tests" basados en la estadística chi-cuadrado usando "jackknife" y estrategias de repetición. Usa dos versiones χ^2_J (3.5.9) y G^2_J (3.5.10). Ambas toman como punto de referencia a

$$\sqrt{2(\chi^2_{T-1})^{1/2} - (T-1)^{1/2}}$$

- Dos "tests" alternativos basados en χ^2_R (G^2_R) (3.3.2- 3.5.3) y $\chi^2_{RS}(G^2_{RS})$ (3.5.6-3.5.7).

- "Tests" basados en χ^2_R (G^2_R) divididos por (T-1), (3.5.4- 3.5.5), lo que implica utilizar las tablas de $F_{(T-1),(n-1)(T-1)}$

- "Tests" basados en las estadísticas definidas en (3.5.8).

Los parámetros que controlan son:

i) α, el nivel de significación del "test".

ii) π, el vector de probabilidades del modelo.

iii) T, el número de categorías.

iv) n, el número de conglomerados en la muestra.

v) m, el número (constante) de elementos extraídos por conglomerado.

vi) la media de los efectos de diseño, $\lambda_\cdot = \sum_{l=1}^{(T-1)} \frac{\lambda_l}{T-1}$ (tratándose del caso de bondad de ajuste, $\{\delta_l = \lambda_l; l = 1,.., (T-1)\}$).

vii) El coeficiente de variación de los valores característicos $\{\lambda_l, l = 1...(T-1)\}$, que denominan a, (3.5.6/7).

El grado de conglomeración está representado el par $(\lambda_\cdot, a)$, como sigue:

a) muestreo multinomial $(\lambda_\cdot = 1, a = 0)$.

b) efecto de diseño constantes $(\lambda_\cdot > 1, a = 0)$.

c) efecto de diseño no constante $(\lambda_\cdot > 1, a > 0)$.

El plan de muestreo que se considera en el estudio es bi-etápico y consiste en extraer una muestra de n conglomerados y, a posteriori, m unidades dentro de ellos, clasificables en T categorías.

Los datos están generados de manera tal que en la segunda etapa de muestreo dentro de cada conglomerado, la distribución sea condicionalmente multinomial dependiendo de vectores (p) obtenidos en la primera etapa a partir de una distribución que es una mezcla de 2 distribuciones Dirichlet con parámetros $\nu_j = \nu\pi_j$, $\nu>0$, $j=1,2$, siendo $\pi_j = (\pi_{1j},...,\pi_{(T-1)j})'$.

Bajo este modelo, se tiene:

$$\pi = \beta_1\pi_1 + \beta_2\pi_2$$

y, conduce a:

$$\lambda_. = \sum_{l=1}^{T-1} \frac{\lambda_l}{T-1} = \frac{m + \nu}{1 + \nu} + \frac{(m - 1)\ \nu\ \delta}{(T-1)\ (1 + \nu)}$$

$$a = \frac{\left[\sum_{l=1}^{T-1}(\lambda_l-\lambda_.)^2\right]^{1/2}}{\lambda_.(T-1)^{1/2}} = \frac{(T-2)^{1/2}\ (m-1)\ \nu\ \delta}{[(T-1)(m + \nu) + (m - 1)\ \nu\ \delta]}$$

donde $\{\lambda_l, l=1,..,(T-1)\}$ son los efectos diseño,

$$0 \le \delta \le 1;\ \delta = \beta_1\beta_2(\pi_1 - \pi_2)\ \Delta^{-1}\ (\pi_1 - \pi_2)$$

β_1, β_2, ($\beta_1 + \beta_2 = 1$) son los coeficientes en la mezcla de las distribuciones. Los autores adoptaron este modelo por ser simple y porque permite generar valores relativamente altos de a para valores fijos de $\lambda_.$, conduciendo a efectos de diseño que no son constantes.

Los resultados del estudio se muestran en las Tablas 4.4 a 4.9.

Las tablas 4.4, 4.5, 4.6 y 4.7 se refieren al nivel de significación. Las tablas 4.8 y 4.9 hacen referencia a la potencia. Dado el gran número de parámetros a controlar, no fue posible examinar un conjunto de combinaciones que conformasen un factorial completo. Luego, las simulaciones relacionadas con el error de tipo I se llevaron en general, a cabo para $\lambda_.=2$, un valor de $\alpha = 5\%$ y bajo el caso equiprobable $\boldsymbol{\pi}_0 = (1/T,...,1/T)'$. De igual manera los resultados de la potencia que se publican tienen como parámetros fijos a T=5, $\lambda_.=2$. Además el vector $\boldsymbol{\pi}$ de las alternativas se restringió a la clase $\pi(T,q,\varepsilon)$, definido por el vector de elementos $\pi_j(T,q,\varepsilon) = 1/T + \varepsilon$, $j=1...T$ y $\pi_j(T,q,\varepsilon) = 1/T + q\varepsilon/(T-q)$, $j=q+1,...,T$ con ε positivo o negativo.

La elección de los valores $\varepsilon = -0.1$ y 0.1, que aparecen en la tabla 4.8, dan potencias de Pitman de $F\chi^2_{Co}$, χ^2_J, F_W iguales a 90% para $\lambda_. = 2$, a = 0 y n = 50. La potencia de Pitman se obtiene a partir de un χ^2_{T-1} no central con parámetro

$$(n/\bar{\lambda}) \sum_{j=1}^{T} (\pi_j - T^{-1})^2/T$$

TABLA 4.4

Niveles de significación (NivSig) reales (%) para los "tests" sin ajustar X^2 y G^2. n = 50, α = 5%, $\pi = (1/T,....,1/T)'$, $\pi_j = 1/T$ j = 1,..,T.

N° de Categ.	$\lambda_.$	a	m	NivSig(X^2)	NivSig(G^2)
3	1.5	.0	10	13.4	13.7
3	1.5	.29	10	12.7	12.7
3	2.0	.0	10	23.3	23.0
3	2.0	.29	10	21.0	20.8
5	2.0	.0	10	31.7	32.1
5	2.0	.5	10	28.3	28.7
10	2.0	.0	20	50.3	50.0
10	2.0	.82	20	44.3	44.3
10	1.0	.0	20	6.6	6.2

TABLA 4.5

Comparación de los niveles de significación (NivSig) reales (%) de los "tests" χ^2_W y F_W, α = 5%, $\lambda_.$ = 2.0, $\pi = (1/T,..., 1/T)'$, $\pi_k = 1/T$; j = 1,.., T.

N° de Categ.	m	a	n	NivSig(χ^2_w)	NivSig(F_w)
3	10	.29	50	6.8	5.2
3	10	.29	30	7.5	4.9
3	10	.29	10	15.7	7.4
5	10	.5	50	9.1	5.9
5	10	.5	30	12.6	7.7
5	10	.5	10	37.4	10.8
10	20	.82	50	18.1	9.0
10	20	.82	30	31.5	10.3
10	20	.82	20	49.0	10.6
10	20	.82	10	94.4	5.3
10	20	.0	50	19.5	7.0
10	20	.0	30	29.1	6.5
10	20	.0	20	47.2	8.1
10	20	.0	10	95.5	4.1

TABLA 4.6

Comparación de los niveles de significación (%) (NivSig) de los "tests" $\chi^2_{R_20}$, $F\chi^2_{R_20}$, $\chi^2_{R_3}$

α = 5%, $\lambda_.$ = 2.0, $\pi = (1/T,...,1/T)'$, $\pi_j = 1/T$, j = 1,..T.

N° de Categ.	m	a	n	NivSig(χ^2_{R20})	NivSig($F\chi^2_{R2o}$)	NivSig(χ^2_{R3})
3	10	.0	50	5.5	5.1	5.4
3	10	.0	30	4.7	4.0	4.4
3	10	.0	10	6.7	4.7	6.3
3	10	.29	50	6.0	5.6	5.8
3	10	.29	30	6.3	5.4	6.2
3	10	.29	10	9.2	6.2	8.5
10	20	.0	50	5.5	5.3	5.9
10	20	.0	30	4.7	4.1	5.1
10	20	.0	10	5.0	3.5	5.4
10	20	.82	50	9.9	9.2	9.9
10	20	.82	30	10.7	10.6	11.4
10	20	.82	10	13.1	10.9	15.0

TABLA 4.7

Niveles de significación (%) (NivSig) para los "tests" de Rao-Scott y Fay. α = 5%, $\lambda_.$ = 2.0, $\pi = (1/T,...,1/T)$, $\pi_j = 1/T$, j = 1,.., T.

T	m	a	n	NivSig (χ^2_{RSo})	NivSig (G^2_{RSo})	NivSig (χ^2_J)	NivSig (G^2_J)	NivSig ($F\chi^2_{R2o}$)	NivSig (FG^2_{Ro})
3	10	.29	50	5.5	5.7	5.6	5.5	5.6	5.7
3	10	.29	30	5.9	5.8	5.5	5.4	5.4	5.8
3	10	.29	10	7.4	8.2	9.2	9.2	6.2	6.8
5	10	.5	50	5.1	6.2	4.8	5.0	6.2	6.5
5	10	.5	30	5.1	4.9	5.8	6.0	6.2	6.7
5	10	.5	10	7.8	7.6	10.4	10.9	9.0	9.3
10	20	.82	50	5.4	6.1	4.9	4.5	9.2	9.4
10	20	.82	30	7.4	7.4	7.5	7.5	10.6	10.2
0	20	.82	10	5.5	6.4	11.1	13.1	10.9	12.8
10	20	.0	50	3.6	3.8	5.5	5.9	5.3	4.7
10	20	.0	30	1.6	2.5	4.5	5.0	4.1	4.3
10	20	.0	10	1.5	1.8	4.6	5.7	3.5	4.0

TABLA 4.8

Potencia (%) de $F\chi^2_{R2o}, \chi^2_{RSo}, \chi^2_J, F_W$ como funciones de n, a y ε , para T = 5, $\lambda_{\cdot} = 2$, m = 10.

a	ε	n	$F\chi^2_{R2o}$	χ^2_{RSo}	χ^2_J	F_w
.0	-.1	50	94.1	94.0	95.2	94.6
.0	-.1	30	71.5	69.9	77.0	76.8
.0	-.1	20	47.7	44.9	57.8	59.0
.0	-.1	10	22.4	19.3	32.6	31.7
.0	+.1	50	85.3	84.8	84.6	76.4
.0	+.1	30	61.8	60.8	61.1	54.7
.0	+.1	20	39.7	38.4	40.4	33.6
.0	+.1	10	21.8	20.4	23.8	17.6
.5	-.1	50	89.0	88.1	84.9	75.4
.5	-.1	30	69.1	68.8	65.0	53.9
.5	-.1	20	55.1	54.3	52.0	40.3
.5	-.1	10	34.4	32.7	39.8	34.8
.5	+.1	50	73.8	67.9	63.8	45.3
.5	+.1	30	50.4	45.1	41.1	28.8
.5	+.1	20	36.0	29.3	27.2	19.1
.5	+.1	10	16.9	13.6	14.7	11.7

TABLA 4.9

Una comparación de las tendencias en las potencias de los "tests"

$F\chi^2_{R2o}, \chi^2_{RSo}, \chi^2_J, F_W$, con $\lambda_{\cdot} = 2$, m = 10 para T = 3, m = 20 para T = 10 y n = 30.

T	a	ε	$F\chi^2_{R2o}$	χ^2_{RSo}	χ^2_J	F_w
3	.0	-.11	72.7	74.0	74.5	71.9
3	.0	+.11	65.7	67.5	66.5	60.1
10	.0	-.6	73.0	65.3	79.3	81.3
10	.0	+.6	60.9	54.4	58.6	35.5
3	.29	-.11	71.6	72.7	69.8	63.8
3	.29	+.11	60.3	59.4	57.4	48.8
10	.82	-.06	70.2	66.9	60.9	47.7
10	.82	+.06	36.7	23.5	20.4	12.7

La estadística modificada $F\chi^2_{R2o}$ se comporta mejor que la χ^2_{R2o} ajustada por λ. en el control del error de tipo I. Desde que $F\chi^2_{R2o}$ requiere la misma cantidad de información que χ^2_{R2o}, es más recomendable la primera. Entre las estadísticas χ^2_{RSo}, χ^2_J, F_W, que requieren información igualmente detallada, las dos primeras se comportan mejor que F_W. El comportamiento de esta última, bajo Ho, es sensible a la asimetría de la hipótesis nula y su comportamiento con respecto a la potencia es sensible a la alternativa π. Además, la potencia de F_W es marcadamente menor que la de χ^2_{RSo} ó χ^2_J en caso de efectos de diseños no constantes.

El estudio de Monte Carlo muestra que F_W es mejor que χ^2_W. Los comportamientos de χ^2_J y χ^2_{RSo} son similares aunque el último tiene una leve ventaja en el caso de efectos de diseño no constantes.

5. Definiciones y Procedimientos

5.1. *Estadísticas que se estudian en este trabajo*

Se ha prestado especial atención a las estadísticas de tipo Wald y a las de Rao y Scott. Las razones son las siguientes:

- Las estadísticas de tipo Wald tienen propiedades asintóticas óptimas y son aplicables a innumerables clases de problemas. Por otro lado, en la Escuela de Estadística (1) se habían realizado estudios de simulación para el caso de la independencia de dos variables binarias, usando una estadística de Wald, y se habían obtenido, para valores moderados del tamaño de la muestra, resultados promisorios, tanto desde el punto de vista del nivel de significación como de la potencia.

---

(1) Servy y col. (1989) "Análisis de Tablas de Contingencia a partir de Datos provenientes de Muestras de Diseño Complejo". Convenio INDEC-Facultad de Ciencias Económicas, Universidad Nacional de Rosario.

- Las estadísticas de Rao y Scott también demostraron buen comportamiento en las simulaciones de Thomas y Rao (1987), para pequeñas muestras, mientras que Rao y Scott encuentran el mismo resultado para muestras grandes, usando la aproximación de Solomon y Stephens (1977).

La estadística que corrige el X^2 (3.1.9) por el mínimo de los promedios de los efectos de diseño marginales (que se puede considerar del tipo de las de Rao y Scott) se incluye en el estudio por la facilidad de su aplicación a muchas de las tablas publicadas por el INDEC, gracias al cálculo que realiza el Instituto de los efectos de diseños para los márgenes de las mismas. Esta estadística no ha sido estudiada en profundidad, de modo que un mejor conocimiento de la misma ha sido uno de los objetivos de este trabajo.

5.2. *Problema específico considerado*

Se busca investigar los "tests" de independencia de tablas de dos variables. En su forma general, la hipótesis de independencia puede escribirse como:

$$\boldsymbol{h}(\boldsymbol{\pi}) = \mathbf{0} \tag{5.2.1}$$

donde $\boldsymbol{h}(\boldsymbol{\pi})$ es un vector de funciones específicas aplicadas al vector $\boldsymbol{\pi}$ que contiene a las proporciones poblacionales que corresponden a las casillas de la tabla. En particular, en una tabla de contingencia 2x2, conformada por las proporciones $\{\pi_{00},\pi_{01},\pi_{10},\pi_{11}\}$, la hipótesis de independencia se puede escribir así:

$$\ln\pi_{00}-\ln\pi_{01}-\ln\pi_{10}+\ln\pi_{11} = 0 \tag{5.2.3}$$

5.3. *Plan de muestreo*

Se considera que la población finita está formada por conglomerados, distribuidos en estratos de tamaño variable, y que el plan de muestreo consiste en tomar muestras de conglomerados sin reposición en cada estrato, cuyos tamaños son proporcionales al de los estratos. Todos los elementos de los conglomerados extraídos se incluyen en la muestra.

Este plan es equivalente, cuando la fracción de muestreo es pequeña, al muestreo sin reposición. En datos de grandes encuestas, la diferencia entre los casos con reposición y sin reposición tiene, en general, poco impacto sobre las estimaciones y sus variancias. Sin embargo, desde el punto de vista académico, vale la pena realizar la distinción, para conocer las discrepancias que ambos métodos pueden presentar, si las fracciones de muestreo no son despreciables en su magnitud.

A pesar que el planteo se ha realizado en forma bastante general, en las simulaciones las muestras fueron extraídas con reposición, dejando para una etapa posterior, la consideración del otro caso.

5.4. *Modelo para la teoría asintótica*

Se necesita un marco de referencia para la teoría asintótica en la que se basan los "tests" tanto de Wald como de Rao y Scott, ya que ellos se han diseñado para muestras grandes.

Se debe dar precisión a frases como: "la población tiende a infinito"; ¿es el número de estratos el que tiende a infinito, ó las poblaciones dentro de los mismos?; ó ambos a la vez?.

Se supondrá, como en Fuller (1975,1984) que la sucesión de poblaciones finitas está generada por una población infinita que se usa como base para las generalizaciones que están más allá de la población finita particular.

Para formalizar este concepto, se supone que: $\{\zeta_t, t=1,...,\infty\}$, es una sucesión de poblaciones finitas estratificadas, con $L_t > L_{t-1}$ estratos; la población finita del estrato h de ζ_t es una muestra aleatoria de N_{th} ($> N_{(t-1)h}$) conglomerados seleccionados de una superpoblación infinita.

La respuesta multivariada del v-ésimo elemento del u-ésimo conglomerado del estrato h, se simboliza por un vector -tipo columna-

$$\{Y_{thuv},\ h=1,..,L_t,\ u=1,..,N_{th},\ v=1,..,M_{thu}\}. \qquad (5.4.1)$$

Para el caso de las tablas de contingencia, las coordenadas de Y_{huv} (de aquí en más se

suprimirá el subindice t para simplificar la escritura) están identificadas por un doble índice, (i,j) y su definición es la que sigue:

$$Y_{huv}^{(ij)} = \begin{cases} 1 & \text{si el u-ésimo individuo del v-ésimo conglomerado del h-ésimo estrato pertenece a la categoría i-ésima de la primer variable y a la j-ésima de la segunda variable.} \\ 0 & \text{en otro caso.} \end{cases}$$

i=0,...,(r-1) , j=0,...,(c-1). Es decir, la dimensión del vector Y_{huv} es rxc , para todo h, u y v.

Para la t-ésima población finita se define al parámetro vectorial π de dimensión r x c, cuyas coordenadas son :

$$\pi_{ij} = \frac{\sum_{h=1}^{L} W_h N_h^{-1} \sum_{u=1}^{N_h} \left[\sum_{v=1}^{M_{hu}} Y_{huv}^{(ij)} \right]}{\sum_{h=1}^{L} W_h N_h^{-1} \sum_{u=1}^{N_h} M_{huv}} = \frac{\bar{Y}_{ij}}{\bar{X}} \qquad i=0,..,(r-1)\ ,\ j=0,..,(c-1) \qquad (5.4.2)$$

con $W_h = \frac{N_h}{N_.}$, $N_. = \sum N_h$.

$\bar{Y}_{ij}$ es la media de elementos en la población asociados a la combinación (i,j) de las dos variables, mientras que $\bar{X}$ es la media del número de elementos por conglomerados de la población.

5.5. *Fórmulas utilizadas para el cálculo de las estadísticas*

Se supone que para una población finita fija, se extrae una muestra estratificada. Del h-ésimo estrato se extrae una muestra de n_h conglomerados, cuyo tamaño es proporcional al tamaño del estrato, para todo h=1,..,L. Bajo este esquema, los parámetros $\{\pi_{ij}\ ,i=0,...,(r-1),\ j=0,...,(c-1)\}$ se estiman por la razón combinada:

$$\hat{\pi}_{ij} = \frac{\sum_{h=1}^{L} W_h\, n_h^{-1} \sum_{l=1}^{n_h} \left[\sum_{j=1}^{M_{hlv}} Y_{hlv}^{(ij)} \right]}{\sum_{h=1}^{L} W_h\, n_h^{-1} \sum_{l=1}^{n_h} M_{hlv}} = \frac{\bar{y}_{ij}}{\bar{x}} \qquad i=0,..,(r-1)\ ,\ j=0,.,(c-1) \qquad (5.5.1.0)$$

Nuevamente, el denominador es la media del tamaño de los conglomerados, pero esta vez, estimada por la muestra.

5.5.1. Estadística de tipo Wald

Se utilizó como vector de funciones que definen la independencia a:

$$\boldsymbol{h}(\boldsymbol{\pi}) = \boldsymbol{K} \ln \boldsymbol{\pi} = \boldsymbol{K} \ln \bar{\boldsymbol{Y}} \qquad (5.5.1.1)$$

donde $\bar{\boldsymbol{Y}}$ es el vector de coordenadas $\{\bar{Y}_{ij}\ ;\ i=0,..,(r-1)\ ,\ j=0,..,(c-1)\}$. En las tablas 2x2, $\boldsymbol{h}(\boldsymbol{\pi})$ es un escalar ya que $\boldsymbol{K} = (1\ ,-1\ ,-1\ ,\ 1)$ De modo que:

$$\begin{aligned} h\ (\pi) &= \ln\pi_{00} - \ln\pi_{01} - \ln\pi_{10} + \ln\pi_{11} \\ &= \ln\bar{y}_{00} - \ln\bar{y}_{01} - \ln\bar{y}_{10} + \ln\bar{y}_{11} \end{aligned} \qquad (5.5.1.2)$$

En las tablas 3x3, $\boldsymbol{h}(\boldsymbol{\pi})$ es un vector de dimensión 4, pues en este caso:

$$K = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix} \qquad (5.5.1.3)$$

y

$$\pi' = [\pi_{00}\ \pi_{01}\ \pi_{02}\ \pi_{10}\ \pi_{11}\ \pi_{12}\ \pi_{20}\ \pi_{21}\ \pi_{22}] \qquad (5.5.1.4)$$

Si la tabla tiene dimensión r x c, $h(\pi)$ es un vector de dimensión (r-1)(c-1) x 1 y K es una matriz de contrastes apropiados de orden (r-1)(c-1).

La matriz de variancias y covariancias de $h(\hat{\pi})$ donde $\hat{\pi}$ está dado por (5.5.1.0), resulta, después de utilizar el método de linearización por Taylor, aproximadamente igual a:

$$S = K\, D^{-1} V(\hat{\pi}) D^{-1} K' \qquad (5.5.1.5)$$

siendo $D = diag(\hat{\pi})$ y $V(\hat{\pi})$ la matriz de covariancias de las estimaciones que se definen más adelante.

Debido a que K es una matriz de contrastes y los denominadores de las coordenadas de $\hat{\pi}$ son todos iguales entre sí, resulta

$$h(\hat{\pi}) = h(\bar{y}) \qquad (5.5.1.6)$$

donde $\bar{y}$ es la estimación de $\bar{Y}$ cuyos elementos se definen en (5.4.3). De allí, resulta que la matriz de covariancias aproximada de (5.5.1.6) se puede escribir como en (5.5.1.5) pero sustituyendo los $\{ \hat{\pi}_{ij} ;\ i=0,..,(r-1)\ ,\ j=0,..,(c-1)\}$ por los $\{ \bar{y}_{ij} ;\ i=0,..,(r-1)\ ,\ j=0,..,(c-1)\}$.

La matriz de variancias y covariancias de $\bar{y}$ se computa de la siguiente manera:

$$V(\bar{y}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) n_h^{-1} \sum_h \qquad (5.5.1.7)$$

donde,

$$\Sigma_h = (n_h-1)^{-1} \sum_{u=1}^{n_h} (y_{hu.}-\bar{y}_{h..})(y_{hu.}-\bar{y}_{h..})' \qquad (5.5.1.8)$$

siendo $\bar{y}_{hu.}$ el vector de medias de las variables respuesta en el u-ésimo conglomerado de la muestra del estrato h-ésimo y $\bar{y}_{h..}$ tiene igual definición para el estrato h.

Bajo el modelo antes descripto y ciertas condiciones de regularidad respecto de:

i) los límites de funciones de $\{ N_h, W_h, n_h ; h=1,..,L \}$,

ii) los momentos de los totales de los conglomerados,

y si $\sum_h$ es positiva definida, es posible demostrar (Fransisco (1987)), que el vector $\bar{y}$ debidamente estandarizado, se distribuye asintóticamente según la distribución normal de parámetros ($\mathbf{0}, I$).

La estadística del "test" $h(\pi) = 0$ es la siguiente:

$$h(\bar{y})[K\ D^{-1}\ v(\bar{y})D^{-1}K']\ h'(\bar{y}) = h(\hat{\pi})[K\ D^{-1}\ v(\hat{\pi})D^{-1}K']\ h'(\hat{\pi}) \qquad (5.5.1.9)$$

y su distribución asintótica es la de un χ^2 con tantos grados de libertad como relaciones linealmente independientes conforman $h(\hat{\pi})$.

5.5.2. Estadísticas de Rao y Scott

Para las estadísticas de Rao y Scott se tiene en cuenta el esquema de muestreo descripto en la sección (5.3).

La hipótesis de independencia se escribe según (5.5.1.1) y la estadística X^2 (antes de corregir) responde a la fórmula (3.1.9), donde, para el caso de interés, $\hat{\pi}$ se obtiene de (5.5.1.0).

La estimación de la variancia de $\hat{\pi}$, bajo la distribución multinomial, se calcula de la manera siguiente:

$$\hat{P} = diag(\hat{\pi}) - \hat{\pi}' \hat{\pi} \qquad (5.5.2.1)$$

y las derivadas $H(\pi) = \left[\frac{\partial h_l(\pi)}{\partial \pi}\right]_{l=1,..,b}$ conforman, en este caso, la inversa de la matriz $D = diag(\hat{\pi})$.

Para estimar el valor de δ_*, (3.3.1) o el de λ. (3.3.2), se utiliza la matriz $\hat{V}(\hat{\pi}) = ((v(\hat{\pi}_{ij})))_{\substack{i=0,..,(r-1)\\ j=0,..,(c-1)}}$ donde

$$\hat{V}(\hat{\pi}) = \sum_{h=1}^{L} n_h(1-\frac{n_h}{N_h})\frac{S_h}{\bar{x}^2}, \qquad (5.5.2.2)$$

siendo:

$$S(h) = \frac{1}{n_h-1}\sum_{u=1}^{n_h}(z_{hu}-\bar{z}_h)(z_{hu}-\bar{z}_h)', \qquad (5.5.2.3)$$

donde los elementos generales de los vectores z_{hu} (de dimensión r x c) son:

$$z_{hu}^{(ij)} = \phi_{hu}(Y_{hu_{ij}} - \hat{\pi}_{ij}M_{hu}) \quad , \quad \bar{z}_h = \sum_{u=1}^{n_h}\frac{z_{hu}}{n_h} \qquad (5.5.2.4)$$

$$i=0,..,(r-1) \quad j=0,...,(c-1)$$

respectivamente, y ϕ_{hu} la probabilidad de inclusión en la muestra que corresponde al conglomerado u-ésimo del h-ésimo estrato. En este caso, $\phi_{hu} = 1/n$ para todo h y u, porque la muestra es autoponderada.

La fórmula (5.5.2.2) surge de aplicar el método de linearización de Taylor para el caso de un estimador por razón (5.5.1.0). No se aplicó aquí la simplificación que consiste en reemplazar

$V(h(\hat{\pi}))$ por $V(h(\bar{y}))$ porque en alguna de las estadísticas de tipo Rao las correcciones requieren el cómputo de la matriz de covariancias de $\hat{\pi}$ y no de las funciones de $\hat{\pi}$.

No obstante que las fórmulas presentadas y los programas efectuados consideran que los conglomerados pueden ser de tamaño desiguales, las simulaciones se realizaron para la situación particular en que $M_{hu} = k$, para la cual $V(\hat{\pi})$ y $V(h(\hat{\pi}))$ sólo difieren en una constante. Además, el muestreo que se utilizó fue con reposición ya que éste resulta un caso particular del que se efectúa sin reposición cuando n_h / N_h es despreciable.

Para el cálculo de $\chi^2_{R_2}$, se necesita previamente conocer $\hat{\lambda}_.$, el cual se obtiene de la siguiente manera:

$$\hat{\lambda}_. = \sum_{i=0}^{(r-1)} \sum_{j=0}^{(c-1)} (1 - \hat{\pi}_{ij}) \frac{\hat{d}_{ij}}{(rc-1)} \qquad (5.5.2.5)$$

donde, $\hat{d}_{ij}$, es el "deff" correspondiente a la celda (i,j).

5.5.3. Estadística de Fellegi

El cálculo de $\chi^2_{R_3}$, requiere la obtención de $\hat{d}_.$ a partir de:

$$\hat{d}_. = \frac{1}{rc} \sum_{ij=0}^{(r-1)(c-1)} \frac{n v(\hat{\pi}_{ij})}{\hat{\pi}_{ij} (1 - \hat{\pi}_{ij})} \qquad (5.5.3.1)$$

donde, $v(\hat{\pi}_{ij})$ es un elemento de la matriz estimada de variancias y covariancias de $\hat{\pi}$, y $\hat{\pi}_{ij}(1-\hat{\pi}_{ij})/n$ es un elemento de la diagonal principal de la matriz $\hat{P}$ (5.5.2.1).

5.5.4. Corrección por el "deff" mínimo

La estadística $\chi^2_{R_4}$ requiere el cálculo previo de los "deffs" promedio para cada una de las variables de la tabla, para luego elegir el mínimo de los dos como denominador de X^2.

El "deff" promedio para las filas se obtiene de:

$$\sum_{i=0}^{(r-1)} \frac{(1-\hat{\pi}_{i+})\hat{d}_{i+}}{r-1} \quad , \quad \hat{d}_{i+} = \frac{nv(\hat{\pi}_{i+})}{\hat{\pi}_{i+}(1-\hat{\pi}_{i+})} \tag{5.5.4.1}$$

y, en forma análoga para la variable columna:

$$\sum_{j=0}^{(c-1)} \frac{(1-\hat{\pi}_{+j})\hat{d}_{+j}}{c-1} \quad , \quad \hat{d}_{+j} = \frac{nv(\hat{\pi}_{+j})}{\hat{\pi}_{+j}(1-\hat{\pi}_{+j})} \tag{5.5.4.2}$$

6. Modelo para la Simulación de los Datos

6.1. *Características Generales*

El modelo permite simular datos tales que:

i) Responden a tipos de poblaciones y esquemas de muestreo descriptos en la Sección (5.3)

ii) Provienen de poblaciones bivariadas (cuyas variables se designan x e y), que representan ya sea la hipótesis nula o las alternativas.

iii) los conglomerados en que esas poblaciones se agrupan posean matrices de correlaciones intra-conglamerados distintas.

iv) cuando las variables x e y se consideran individualmente, las correlaciones intra-conglomerados pueden diferir entre sí, produciendo efectos de diseño marginales de diferente magnitud.

Para describir el modelo se supone que todos los conglomerados tienen tamaño k aunque

el modelo podría extenderse fácilmente al caso de tamaños variables. Dado que cada individuo está identificado por dos variables discretas (x e y) cuyos recorridos son {0, 1,.., (r-1)} y {0, 1,.., (c-1)}, cada conglomerado posee una respuesta de dimensión 2k, la cual puede tomar un número finito de estados $((rc)^k)$.

Los respuestas de los individuos se identifican, a veces por dos subíndices, (i, j), aludiendo a las categorías de ambas variables, x e y, pero cuando es conveniente se los indica con un sólo subíndice que toma valores en {1,2,.., rc}.

La población que corresponde al h-ésimo estrato es una muestra de N_h conglomerados, donde cada conglomerado se escoge independientemente del otro y la probabilidad de que un conglomerado escogido posea un estado determinado es $P_{k,\upsilon(1),\ldots,\upsilon(k)}$, donde $\upsilon(v) = 1,..,$ rc es un indicador que identifica al vector de valores de la respuesta del v-ésimo individuo del conglomerado, v = 1,.., k.

Los estados de los conglomerados tienen pues, en el estrato h-ésimo una distribución multinomial de parámetros $\{N_h, P_{k,\upsilon(1),\ldots,\upsilon(k)}, \upsilon(v) = 1,.., rc, v = 1,.., k\}$.

Las probabilidades $P_{k,\upsilon(1),\ldots,\upsilon(k)}$, a su vez, responden a las de una cadena de Markoff, con un número finito k de pasos , que está determinada por probabilidades iniciales $\{a_{\upsilon(1)}, \upsilon(1) = 1,.., rc\}$ y una matriz de transición que depende de dos parametros, η_1 y η_2.

Las probabilidades $\{a_{\upsilon(1)}, \upsilon(1) = 1,.., rc\}$ se eligen de forma tal que los individuos (cuando se ignoran los conglomerados) puedan constituir una tabla de frecuencias bivariada de probabilidades $\{\pi_{ij}, i = 0,..., (r-1), j = 0,..., (c-1)\}$ prefijadas.

Si se desea considerar las respuestas de los conglomerados en forma ordenada, descartando toda información sobre cuál es el orden (primero, segundo, etc.) del individuo que posee el par (x, y), los estados posibles de las respuestas de los conglomerados se reducen a $(rc)^k / k!$ y las probabilidades de las respuestas ordenadas se obtienen sumando las probabilidades de las k! respuestas distinguibles que dan lugar a la misma respuesta ordenada.

Como las estadísticas que nos interesan estudiar son invariantes con respecto a la permutación de los individuos del conglomerado, a los efectos de nuestro trabajo resulta indiferente utilizar los conglomerados distinguibles u ordenados.

El modelo propuesto puede ampliarse para casos más complejos:

i) Se puede permitir que los conglomerados sean de diferente tamaño en los diferentes estratos o aún dentro de ellos. En este caso, la distribución de las respuestas de los conglomerados sería una mezcla de multinomiales.

ii) Pueden generarse, dentro de un mismo estrato, conglomerados de diferentes DEFF's para una misma variable.

6.2. *Modelo para Generación de los Datos*

Se desea generar conglomerados que estén idéntica e independientemente distribuidos, y que los individuos que contienen no sean independientes entre sí. En el caso de dos variables cualitativas, con r y c categorías respectivamente, los elementos de un conglomerado genérico se pueden ubicar en una matriz 2 x k, donde cada fila está formada por los valores de cada variable, x e y, que corresponden a los k individuos del conglomerado, los cuales se identifican con las columnas de la matriz.

Dado que los posibles valores de x son {0,1,..,(r-1)} y los de y son {0,1,..,(c-1)}, el primer individuo o columna de la matriz puede tomar valores en el conjunto: 00, 01,..., 0(c-1),...,(r-1)(c-1), y es escogido de acuerdo con las probabilidades $\{a_{00}, a_{01},..., a_{0(c-1)},..., a_{(r-1)(c-1)}\} = \{a_{v(1)}, v(1) = 1,..,rc\}$ Los individuos siguientes se simulan según un proceso de Markoff, del cual las constantes $\{a_{v(1)}, v(1) = 1,..,rc\}$ son las probabilidades iniciales. Cada conglomerado es una cadena de Markoff de longitud k. Para formar otro conglomerado, se inicia otra cadena similar independiente de la primera.

En cuanto a la matriz de transición, $p_{\alpha\beta}$, designa a la probabilidad condicional de que $v(v) = \alpha$ dado que $v(v-1) = \beta$ para v = 2,.., k y está definida así:

$$P_{\alpha\beta} = \begin{cases} \eta_1\,\eta_2 & \text{si el par de valores que corresponde a } \alpha \text{ es igual al par que corresponde a } \beta. \\ \eta_1\,[\xi_2 / (c-1)] & \text{si los pares que corresponden a } \alpha \text{ y } \beta \text{ tienen los primeros elementos iguales y los segundos diferentes.} \\ [\xi_1 / (r-1)]\,\eta_2 & \text{si los pares que corresponden a } \alpha \text{ y } \beta \text{ tienen sus segundos elementos iguales y los primeros diferentes.} \\ [\xi_1 / (r-1)][\xi_2 / (c-1)] & \text{si los dos pares que designan } \alpha \text{ y } \beta \text{ tienen respectivamente sus primeros y segundos elementos diferentes entre si.} \end{cases} \qquad (6.2.1)$$

siendo $\eta_1 + \xi_1 = 1$, $\eta_2 + \xi_2 = 1$.

Se vera a continuación las probabilidades asociadas con cada conglomerado, y alguna notación que se utilizará más adelante.

La probabilidad de un conglomerado genérico de tamaño k es,

$$P_{k,\nu_{(1)},\nu_{(2)},\ldots,\nu_{(k)}} = a_{\nu(1)}\, P_{\nu(1)\nu(2)}\, P_{\nu(2)\nu(3)}\cdots P_{\nu(k-1)\nu(k)}$$
$$\nu_{(\nu)} = 1,2,\ldots, rc \qquad \nu = 1,\ldots, k \qquad (6.2.2)$$

Otras probabilidades de interés son las probabilidades de que el primer individuo del conglomerado porte el par de valores identificado con α y el v-ésimo porte el identificado con β

$$P_{k,\,\nu_{(1)}=\alpha,\nu_{(\nu)}=\beta} = \sum_{\substack{\nu_{(1)}\cdots\nu_{(k)} \\ \alpha\nu_{(1)}\nu_{(\nu)}}} P_{k,\,\nu_{(1)}=\alpha,\nu_{(2)},\ldots,\nu_{(\nu)}=\beta,\nu_{(\nu+1)},\ldots,\nu_{(k)}} \qquad (6.2.3)$$

$\alpha, \beta = 1, 2,.., rc.$

En términos de cadenas de Markoff (6.2.3) es la probabilidad de pasar del estado inicial α al estado β en exactamente (v-1) pasos.

$$P_{k,\, v(1)=\alpha,\, v(v)=\beta} = a_\alpha\, p_{\alpha\beta}^{(v-1)} \qquad \begin{matrix} \alpha, \beta = 1,..., rc \\ v = 1,.., k \end{matrix} \tag{6.2.4}$$

Con notación matricial:

$$((P_{k,\, v(1)=\alpha, v(v)=\beta}))_{\alpha,\beta\, =\, 1,.., rc} = D\,(a_\alpha)\, P^{(v-1)}, \tag{6.2.5}$$

donde D (a_α) es la matriz diagonal cuyos elementos no nulos son las probabilidades iniciales y el segundo factor es la potencia (v-1)-ésima de la matriz de transición.

Ademas,

$$\begin{aligned} P_{k,\, v(v)\, =\, \alpha} &= \sum_{\substack{v(1),..,\, v(k) \\ \neq\, v(v)}}^{rc} P_{k,\, v(1),..,\, v(v)=\alpha,..,\, v(k)} = \\ &= \sum_{\beta\, =\, 1}^{rc} P_{k,\, v(1)=\beta, v(v)=\alpha} = \sum_{\beta=1}^{rc} a_\beta\, p_{\beta\alpha}^{(v-1)} \end{aligned} \tag{6.2.9}$$

Las probabilidades $P_{k,\, v(1),..,\, v(k)}$ son fácilmente estimables, porque la muestra de conglomerados es de tipo simple al azar.

Las probabilidades que corresponden a los conglomerados no son, en forma directa, el objeto de este estudio. Las probabilidades de interés son las probabilidades $\{\pi_\alpha,\ \alpha = 1,.., rc\}$ de que al extraer un individuo al azar de la población, el mismo posea las características 00, 01, 0(r-1),..., (c-1)0,.., (c-1)(r-1) que corresponden a $\alpha = 1,.., rc$.

Pero es posible obtener las segundas a través de las primeras. Así, la probabilidad de obtener un individuo con características correspondientes a $\alpha = 1,2,.., rc$, cuando la población está formada por conglomerados de tamaño k es:

$$\pi_\alpha = \sum_{\nu(1),\ldots,\nu(k)} P_{k,\,\nu(1),\ldots,\nu(k)}\ Pr\ (\alpha \mid \nu(1),\ldots,\ \nu(k)). \qquad (6.2.10)$$

$Pr\ (\alpha \mid \nu_{(1)},\ldots,\ \nu_{(k)})$ es la probabilidad de escoger un individuo en la categoria identificada por α cuando se extrajo el conglomerado identificado por $(\upsilon(1),\ldots,\ \upsilon(k))$. Siendo la extracción de tipo aleatorio simple ella coincide con el cociente entre, $m(\alpha/\upsilon(1),..,\ \upsilon(k))$, el número de individuos de tipo α en el conglomerado, y k, el tamaño del mismo. O sea,

$$k\ \pi_\alpha = \sum_{\nu(1),\ldots,\nu(k)} P_{k,\,\nu(1),\ldots,\nu(k)}\ m(\alpha/,\ \nu(1),..,\ \nu(k)) \quad \alpha = 1,\ldots,rc, \qquad (6.2.11)$$

que alternativamente, se puede escribir:

$$k\ \pi_\alpha = \sum_{\nu=1}^{k} P_{k,\,\nu(\nu)=\alpha} = \sum_{\nu=1}^{k}\sum_{\beta=1}^{rc} a_\beta\ p_{\beta\alpha}^{(\nu-1)}., \qquad (6.2.12)$$

La equivalencia de la primera igualdad de (6.2.12) y la expresada en (6.2.11) queda justificada por el siguiente argumento:

Sea α el número que identifica a valores fijos de las variables (x,y). $P_{k,\,\alpha,\,\alpha,\ldots,\alpha}$ está incluído según la definición (6.2.9) en $P_{k,\,\nu(1)=\alpha}$, $P_{k,\,\nu(2)=\alpha},\ldots,$ $P_{k,\,\nu(k)=\alpha}$. Luego en la primera suma de (6.2.12) $P_{k,\,\alpha,\,\alpha,\ldots,\,\alpha}$ está presente k veces. Por otro lado si consideramos un conglomerado con los "s" primeros individuos portadores de α-ésimo par, su probabilidad está incluída en $P_{k,\nu(1)=\alpha}$, $P_{k,\nu(2)=\alpha},\ldots,$ $P_{k,\nu(s)=\alpha}$ y sólo en ellas por lo que $P_{k,\alpha,\ldots,\alpha,\ \nu(s+1),\ldots,\nu(k)}$ aparece en la primera suma de (6.2.12) "s" veces.

La segunda igualdad en (6.2.12) es consecuencia de (6.2.9), y puede escribirse, poniendo en evidencia los valores de las $\{a_\alpha;\ \alpha = 1,..,\ rc\}$

$$k\ \pi_\alpha = a_\alpha + a_1 \sum_{\nu=1}^{(k-1)} p_{1\alpha}^{(\nu)} + a_2 \sum_{\nu=1}^{(k-1)} p_{2\alpha}^{(\nu)} + \ldots + a_{rc} \sum_{\nu=1}^{(k-1)} p_{rc\alpha}^{(\nu)}, \qquad (6.2.13)$$

$$\alpha = 1,2,..,\ rc$$

Estas ecuaciones pueden ser construídas iterativamente pues,

$$k\,\pi_\alpha = a_\alpha + \sum_{\beta=1}^{rc} a_\beta \sum_{\gamma=1}^{(k-2)} p_{\alpha\beta}^{(\gamma)} + \sum_{\beta=1}^{rc} a_\beta\, p_{\beta\alpha}^{(k-1)} \qquad \alpha = 1,..,\ rc, \qquad (6.2.14)$$

Es decir,

$$k\,\boldsymbol{\pi} = \{\{(k-1)\ \boldsymbol{\pi}\}\} + P^{(k-1)}\, a, \qquad (6.2.15)$$

donde a es un vector columna que tiene por coordenadas a $\{a_\alpha, \alpha = 1,..,rc\}$ y $\{\{(k\text{-}1)\pi\}\}$ debe interpretarse como el conjunto de rc formas lineales en las $\{a_\alpha, \alpha = 1,..,rc\}$ que constituían los segundos miembros de las ecuaciones utilizadas en la determinación de las probabilidades iniciales para conglomerados de extensión (k-1).

El sistema (6.2.15) es crucial en las simulaciones. Para generar datos que representen una población con parámetros $\{\pi_\alpha: \alpha = 1,..., rc\}$ prefijados y estén relacionados según un proceso de Markoff dado, cuya matriz de transición esta definida en (6.2.1), se comienza por resolver las ecuaciones (6.2.15) para las $\{a_\alpha; \alpha = 1,.., rc\}$.

Luego, utilizando las soluciones como probabilidades iniciales y la matriz de transición (6.2.1) se generan cadenas independientes de un número fijo de pasos, que constituyen la población de conglomerados deseados.

No todas las soluciones del sistema son satisfactorias como probabilidades iniciales. La estructura del sistema garantiza que $\sum_{\alpha=1}^{rc} a_\alpha = 1$, pero no garantiza la positividad de las soluciones. El método puede no funcionar para algunas combinaciones de k, r, c, π, ξ y η.

6.3. *Efectos de Diseño*

A partir de las ecuaciones (6.2.14) es evidente que, para un tamaño de conglomerado fijo, k, dado los valores de $\{\pi_{ij}; i = 0,.., (r\text{-}1); j = 0,..., (c\text{-}1)\}$ las probabilidades $P_{k,w(1),..,w(k)}$ dependen exclusivamente de η_1 y η_2. Por lo tanto las distribuciones de los estimadores $\{\hat{\pi}_{ij};\ i = 0,..,(r-1),\ j = 0,..,(c-1)\}$ dependen sólo de esos dos parámetros y en consecuencia, lo

mismo ocurre con sus momentos y los efectos de diseño asociados.

Ahora bien, si se ignora una de las variables, sea ella la y, la población pasa a ser univariada, así como los conglomerados que la constituyen. Los nuevos conglomerados se identifican con k subíndices, x(1),..., x(k) donde x(v) es el valor que la variable x toma en el v-ésimo individuo, v = 1,.., k. Las probabilidades correspondientes se designan

$$\{P_{k,\, x(1),\ldots,\, x(k)};\ x(v) = 0,..,\ (r\text{-}1),\ v = 1,..,\ k\}$$

y éstas, dadas las probabilidades poblacionales marginales

$$\{\pi_{i+} = \sum_{j=0}^{(c-1)} \pi_{ij}\ ,\ i = 0,..,(r-1)\},$$

dependen exclusivamente de η_1.

En efecto, los conglomerados se pueden pensar como generados a partir de la matriz de transición según la cual, la probabilidad condicional de que un individuo tenga la característica " i' ", dado que la anterior es de tipo " i ", es igual a η_1 si i = i', o es igual a $\xi_1 = (1 - \eta_1)$ si $i \neq i'$.

Luego, para un conjunto dado de probabilidades marginales $\{\pi_{i+};\ i = 0,..,\ (r\text{-}1)\}$, los efectos de diseño son funciones de un único parámetro η_1.

De alli que, en las simulaciones se puedan manipular los efectos de diseño, sean ellos asociados a las dos variables o a cada una de variables individualmente, a través de cambios en los parámetros η_1 y η_2.

7. DISEÑO DEL ESTUDIO DE MONTECARLO

7.1. *Descripción*

Se generaron datos para estudiar tablas de dos dimensiones (2x2, 2x3 y 3x3) considerando distintas poblaciones y efectos de diseño variables, que permitieron evaluar los niveles de significación reales y la potencia de los "tests".

Los parámetros que se controlaron en las simulaciones fueron:

- π, el vector de proporciones en la tabla que corresponde o no con la hipótesis nula.
- η_1 y η_2, que determinan los efectos de diseño para las variables asociadas con filas y columnas respectivamente.

En cada uno de los casos, el proceso de simulación consistió en:

- Determinar a partir de los parámetros π, η_1 y η_2 los valores de k, tamaño de conglomerado, para los cuales las probabilidades iniciales -soluciones de las ecuaciones (6.2.13)- resultan positivas (ver a manera de ejemplo la Tabla 8.1) .

- En base al resultado anterior, determinar las combinaciones de los parámetros (π, η_1, η_2 y k) a utilizar en las simulaciones. El procedimiento no establece restricciones para los tamaños de muestra, los cuales se eligieron atendiendo a los utilizados en estudios previos.

- Generar las muestras de acuerdo a los parámetros establecidos. Los conglomerados fueron generados según el procedimieto explicado en 6.2 independientemente unos de otros.

- Calcular las estadísticas señaladas en 5.1 según las fórmulas presentadas en 5.5.

- Repetir los dos últimos pasos mil veces y calcular, en el caso de cumplirse la hipótesis nula, indicadores de la variabilidad de las estadísticas de los "tests". (Tablas 8.3-8.5).

- Evaluar los distintos efectos de diseño utilizados en las correcciones para diferentes valores de η_1 y η_2, cuando se verifica la hipótesis nula en tablas 2x2 con conglomerados de tamaño 5, (Tabla 8.2).

- Calcular en los casos correspondientes a las hipótesis de independencia, los niveles de significación reales para un nivel nominal del 5%. (Tablas 8.6-8.8-8.10).

- Recalcular para los "tests" que se mostraron más conservadores en el punto anterior, los niveles de significación reales para un nivel nominal del 7.5%.

Para los estudios de niveles reales de significación, los valores de π fueron los siguientes:

- Tablas 2x2: {0.16 0.24, 0.24 0.36}
- Tablas 2x3: {0.160 0.096 0.144, 0.240 0.144 0.216}
- Tablas 3x3: {0.109714 0.118857 0.114286,
  0.118857 0.128762 0.123810,
  0.091428 0.099048 0.095238}

- Calcular, en los casos correspondientes a las hipótesis alternativas, los porcentajes de rechazo asociados a un error de tipo I del 5%, para las estadísticas de Wald y $\chi^2_{R_1}$ mientras que para las restantes $(\chi^2_{R_2}, \chi^2_{R_3}, \chi^2_{R_4})$ considerar un nivel nominal del 7.5%.

- Calcular los porcentajes de rechazo para un tamaño de conglomerado igual a 5 y para los siguientes valores de los parámetros:

  - Tablas 2x2: {0.35 0.15, 0.15 0.35}
  - Tablas 2x3: {0.200 0.250 0.100, 0.100 0.100 0.250}
  - Tablas 3x3: {0.125 0.100 0.075, 0.100 0.125 0.075, 0.075 0.100 0.225}

  (Tablas 8.7-8.9-8.11).

- Realizar gráficos de resumen de algunas de las tablas anteriores. (Graficos A.I.1-A.I.12 del Anexo I.

7.2. *Programas de Cómputo*

Para la generación de las diversas poblaciones y el cálculo posterior de las distintas estadísticas, se elaboraron programas empleando el paquete estadístico SAS y en especial su módulo IML (Intercative Matrix Language).

Los programas realizados incluyen:

- Cálculo de las Probabilidades Iniciales (INI_?.PRG): Se utiliza este programa para resolver el sistema de ecuaciones planteado en (6.2.13). Los parámetros que requiere el programa son los valores del vector π, η_1, η_2 y el tamaño de los conglomerados. Los valores de las probabilidades iniciales obtenidas como resultado las coloca en el archivo INI_?.LOG. Dichos valores deben ser siempre positivos, lo cual permite determinar el valor máximo de k, para una combinación dada de η_1 y η_2 .

- Generación de la Muestra (CRE_?.PRG): Este programa genera las muestras utilizadas en el proceso de-simulación. Los parámetros requeridos son las probabilidades inicales,

calculadas por INI_?.PRG (que dependen de los valores de π), η_1, η_2 , el tamaño de los conglomerados (k) y el tamaño de la muestra (n). Produce como resultado un archivo SAS temporal, denominado MUE_A.SSD que sirve de fuente al programa de cálculo de los "tests".

- Cálculo de las Estadísticas (TES_?.PRG): Aplica los "tests" y calcula los efectos de diseño para la muestra obtenida por CRE_?.PRG. El único archivo requerido es MUE_A.SSD y los resultados son almacenados en un archivo SAS permanente denominado RES_?.SSD

- Realización de las Simulaciones (SIM_?.PRG): Es una macro que reúne los dos programas anteriores, y permite especificar el número de repeticiones a realizar. Produce como resultado un archivo SAS permanente con los resultados de los "tests" y efectos de diseño de cada una de las repeticiones denominado TES_?.SSD. Además se obtiene un archivo en formato ASCII con los mismos resultados. Por último, produce estadísticas resumenes a traves de todas las repeticiones para los "tests" y efectos de diseño calculados, el cual se almacena en el archivo UNI_?.LST.

- Aplicación de los "Tests" a un Conjunto de Datos (APL_?.PRG): Permite el cálculo de las estadísticas de los distintos "tests" con cualquier conjunto de datos. Requiere que se confeccione previamente un archivo en formato ASCII (DAT_?.DAT) incluyendo número de conglomerado al que pertenece el individuo, y el valor del mismo en las dos variables de interés. Produce como resultado un archivo ASCII con los valores de todos los "tests" y efectos de diseño (RES_?.LST).

Existen versiones de estos programas para aplicar a:

Tablas 2x2: INI_A.PRG, CRE_A.PRG, TES_A.PRG, SIM_A.PRG,APL_A.PRG

Tablas 2x3: INI_B.PRG, CRE_B.PRG, TES_B.PRG, SIM_B.PRG, APL_B.PRG

Tablas 3x3: INI_C.PRG, CRE_C.PRG, TES_C.PRG, SIM_C.PRG, APL_C.PRG

En el Anexo II se presenta un listado completo de todos los programas mencionados.

8. RESULTADOS

8.1. *Determinación del Tamaño de Conglomerado Máximo*

La Tabla 8.1 presenta los tamaños máximos de conglomerado parla el caso de Tablas 2x2 bajo la hipótesis de independencia para distintos valores de η_1 y η_2. Para este caso el máximo tamaño de conglomerado resultó k = 12 cuando los parámtros η_1 y η_2 toman ambos el valor 9/10. Los valores máximos de k bajo la hipótesis alternativa resultan siempre más bajos.

TABLA 8.1.

Tamaños de Conglomerados Máximos para Tablas 2 x 2. {π = 0.16 0.24, 0.24 0.36}.

		η_2			
		9/10	5/6	3/4	2/3
η_1	9/10	12	10	8	5
	5/6	10	8	5	5
	3/4	8	5	5	
	2/3	5	5		

8.2. *Relación entre los parámetros del modelo y los efectos de diseño*

La relación entre los parámetros η_1 y η_2 muestra que los efectos de diseño, ya sea por fila o columna aumentan cuando el parámetro de tipo η crece.

Como fuera expresado por Holt (1980) el valor promedio de los autovalores δ. es menor que los efectos de diseño promedio tanto de las filas como de las columnas; λ. y d. en cambio, tienen valores intermedios entre los "deff" marginales.

Cuando los "deff" marginales son iguales ($\eta_1 = \eta_2 = 5/6$), las correcciones de χ^2_{R2} y χ^2_{R3} son iguales, comentario que fue hecho por Holt y col. (1980).

Cuando aumentan los valores de η_1 y η_2, aumentan también δ. ,λ. y d.. Esto es válido para n=10 ó n=50.

TABLA 8.2.

Relación entre los Parámetros η_1 y η_2 y los Diferentes "DEFFS" utilizados en las Correciones (Bajo Independencia , k=5, Tablas 2x2)

n	η_1	η_2	"Deff" Fila	"Deff" Columna	δ.	λ.	d.
10	2/3	5/6	1.31	2.43	1.04	1.57	1.59
	3/4	5/6	1.75	2.42	1.22	1.77	1.80
	5/6	5/6	2.40	2.41	1.43	2.05	2.08
	9/10	5/6	3.09	2.42	1.62	2.34	2.39
	9/10	3/4	3.05	1.75	1.33	2.02	2.05
	9/10	2/3	3.07	1.31	1.08	1.80	1.82
50	2/3	5/6	1.50	2.75	1.28	1.84	1.85
	3/4	5/6	2.04	2.74	1.56	2.11	2.12
	5/6	5/6	2.74	2.73	1.86	2.44	2.45
	9/10	5/6	3.47	2.77	2.19	2.80	2.81
	9/10	3/4	3.46	2.04	1.71	2.39	2.41
	9/10	2/3	3.47	1.52	1.38	2.12	2.13

8.3. *Indicadores de las poblaciones generadas*

χ_W y χ_{R1} son muy inestables cuando la muestra es chica. χ_{R2} es mejor, en este aspecto, cuando la tabla es 2x2, pero en las de mayor dimensión su inestabilidad es muy marcada.

Los comportamientos mejoran, casi siempre, cuando el número de elementos por conglomerado (k), pasa de 5 a 10 y n=10, y sin excepción cuando el número de conglomerados aumenta.

El X^2 tradicional es relativamente estable cuando las muestras son pequeñas y desmejora para valores grandes de n. A mayores dimensiones de las tablas, como las otras estadísticas, se torna más inestable. χ_{R2}, χ_{R3} y χ_{R4} son bastante estables. χ_{R2} y χ_{R3} aumentan en algo su variabilidad con el aumento en la dimensión de la tabla pero están poco afectadas por el tamaño del conglomerado ó el tamaño de la muestra.

TABLA 8.3

Variancia de las Estadisticas en 1000 Repeticiones en Tablas 2 x 2. (Bajo Independencia)

k	n	X^2	χ^2_W	$\chi^2_{R_1}$	$\chi^2_{R_2}$	$\chi^2_{R_3}$	$\chi^2_{R_4}$
5	10	5.62	42.32	131.03	1.24	1.12	1.07
	20	8.53	4.35	6.24	1.34	1.27	1.25
	30	9.71	2.64	3.02	1.45	1.40	1.33
	50	9.60	2.09	2.26	1.27	1.25	1.25
10	10	11.12	4.40	6.75	1.43	1.32	1.67
	20	10.17	2.24	2.75	0.96	0.93	1.14
	30	13.22	2.56	2.93	1.14	1.11	1.23
	50	16.06	2.74	2.97	1.23	1.21	1.29

TABLA 8.4

Variancia de las Estadisticas en 1000 Repeticiones en Tablas 2 x 3. (Bajo Independencia)

k	n	X^2	χ^2_W	$\chi^2_{R_1}$	$\chi^2_{R_2}$	$\chi^2_{R_3}$	$\chi^2_{R_4}$
5	10	10.27	169.98	40.79	2.26	2.04	1.08
	20	16.07	48.10	5.59	2.47	2.36	1.51
	30	24.75	8.36	6.13	3.36	3.25	2.13
	50	22.06	4.23	4.21	2.67	2.62	1.83
10	10	20.57	95.57	10.91	2.55	2.36	1.43
	20	29.63	20.58	5.87	2.37	2.28	1.27
	30	30.66	5.36	4.63	2.25	2.19	1.31
	50	38.24	5.06	4.88	2.52	2.48	1.45

TABLA 8.5

Variancia de las Estadisticas en 1000 Repeticiones en Tablas 3 x 3. (Bajo Independencia)

k	n	X^2	χ^2_W	$\chi^2_{R_1}$	$\chi^2_{R_2}$	$\chi^2_{R_3}$	$\chi^2_{R_4}$
5	10	12.60	821.96	33.32	3.87	3.42	0.48
	20	27.69	117.69	8.63	4.29	4.10	0.75
	30	35.80	65.20	7.90	4.63	4.51	0.90
	50	51.40	11.30	8.72	5.83	5.74	1.19
10	10	36.81	579.29	17.13	4.28	3.95	0.66
	20	63.73	47.10	10.47	4.65	4.49	0.85
	30	85.93	21.67	11.22	5.49	5.36	1.07
	50	97.37	13.17	10.18	5.55	5.48	1.17

8.4. *Niveles de significación y Porcentajes de Rechazo*

En ningún caso se observó una influencia apreciable del tamaño del conglomerado en los estudios tanto de niveles de significación y potencia de los distintos "tests". Por tanto sólo se analiza las diferencias en los comportamientos provocados por variaciones en el tamaño de la muestra n.

• Tablas 2x2:

Los resultados se presentan en las Tablas 8.6 y 8.7 y en los Gráficos A.I.1, A.I.2, A.I.7 y A.I.8 del Anexo 1.

Los niveles reales de significación de las estadísticas χ^2_w y $\chi^2_{R_1}$ están en la franja del 10-15% cuando el tamaño de la muestra es 10. A partir de n = 30 los niveles de significación de ambas se acercan al 5%, pero practicamente se igualan los niveles reales y nominales para n=50.

Los "tests" $\chi^2_{R_2}, \chi^2_{R_3}$ y $\chi^2_{R_4}$ resultan conservadores ya que sus niveles de significación son menores que los nominales; aún para n = 50 estos niveles están próximos a 2 ó 3%.

Por tal razón, se elevó el nivel de significación nominal de estos tres "tests" al 7,5%, hecho que provocó que los niveles verdaderos de significación se aproximasen al 5%.

Se estudió entonces la potencia de los "tests" para niveles semejantes de significación reales.

Se encontró que, para todos los "tests" la potencia aumenta al aumentar n, hasta alcanzar un nivel del 85% para la alternativa planteada.

(No se incluye en los gráficos el porcentaje de rechazos del X^2 debido a su mal comportamiento bajo Ho).

TABLA 8.6

Niveles de significación reales (%) de los "tests" para un nivel nominal de α =5% cuando se verifica Ho. [π = {0.16 0.24, 0.24 0.36}, η_1 = 9/10, η_2 = 5/6, Tablas 2 x 2].

k	n	X^2	χ_W^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	17.6	11.2	14.1	3.4	2.8	2.2
	20	17.9	6.1	6.9	3.5	3.3	3.1
	30	21.3	5.8	7.0	3.7	3.3	3.3
	50	19.4	4.6	4.9	2.8	2.8	2.9
8	10	22.9	10.6	14.4	3.2	2.7	4.1
	20	24.2	7.0	8.4	2.9	2.9	3.2
	30	22.6	6.3	6.8	2.8	2.6	3.1
	50	21.7	4.4	5.1	2.0	1.9	2.1
10	10	23.7	9.9	13.1	3.4	2.9	3.7
	20	22.7	5.6	6.7	2.0	1.8	2.2
	30	24.0	6.1	6.8	2.3	2.0	2.6
	50	25.0	7.0	7.5	2.5	2.5	3.2

TABLA 8.7

Porcentaje de Rechazo de los "tests" para un nivel nominal de α =5% cuando no se verifica Ho. [π = {0.35 0.15, 0.15 0.35}, η_1 = 9/10, η_2 = 5/6, Tabla 2 x 2]

k	n	X^2	χ_W^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	51.8	34.1	40.1	23.1	21.3	18.1
	20	81.5	60.0	65.2	49.0	48.4	47.4
	30	89.8	74.5	77.1	64.7	63.9	64.5
	50	98.5	94.7	95.1	90.0	89.8	89.5

• Tablas 2x3:

Los resultados, presentados en las Tablas 8.8 y 8.9 y los Gráficos A.I.3, A.I.4, A.I.9 y A.I.10 del Anexo 1, son semejantes a los encontrados para las tablas 2 x 2. La estadística de Wald, en este caso, tiene un comportamiento peor que la de Rao, en cuanto al nivel de significación real, para valores de n menores que 30. Los "tests" $\chi^2_{R_2}$,$\chi^2_{R_3}$ y $\chi^2_{R_4}$ mantienen su comportamiento conservador: para n=10 tienen un valor muy bajo y aún para n=50, $\chi^2_{R_4}$ no alcanza el nivel de los otros dos.

Para el estudio de los porcentajes de rechazo nuevamente se consideraron niveles de significación del 7,5% para las tres estadísticas más conservadoras, buscando llevar al 5%, el error de tipo I real. Sin embargo,este cambio no produjo el resultado esperado.

Los porcentajes de rechazo son bajos para n=10; $\chi^2_{R_2}$ y $\chi^2_{R_3}$ alcanzan el 80% a partir de n=30, si bien el porcentaje de rechazo de $\chi^2_{R_3}$ se presenta como el más bajo. $\chi^2_{R_1}$ y $\chi^2_{R_2}$ tienen aproximadamente los mismos porcentajes de rechazo, mientras χ^2_w aparece como con un porcentaje mayor. Para n=50 todos los porcentajes de rechazo se aproximan al 85%.

TABLA 8.8

Niveles de significación reales (%) de los "tests" para un nivel nominal de α =5% cuando se verifica Ho. [π = {0.160 0.096 0.144, 0.240 0.144 0.216}, η_1 = 9/10, η_2 = 5/6, Tabla 2 x 3]

k	n	X^2	χ_w^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	21.6	27.6	13.2	1.4	1.2	0.3
	20	28.5	11.7	7.3	2.3	2.3	1.4
	30	32.2	9.2	8.3	4.4	4.3	2.9
	50	33.0	5.1	4.9	2.6	2.6	1.8
8	10	30.5	20.1	13.4	2.4	2.2	0.6
	20	37.4	10.4	9.7	2.3	2.0	1.0
	30	36.8	7.7	7.3	2.4	2.3	1.4
	50	36.1	6.4	6.9	2.3	2.3	1.3
10	10	34.1	17.9	14.8	2.1	1.9	1.3
	20	37.5	8.2	8.3	2.2	2.2	0.9
	30	40.0	6.6	7.2	1.5	1.4	1.0
	50	36.0	6.5	5.9	2.5	2.5	1.6

TABLA 8.9

Porcentaje de Rechazo de los "tests" para un nivel nominal de α =5% cuando no se verifica Ho. [π = {0.20 0.25 0.10, 0.10 0.10 0.25}, η_1 = 9/10, η_2 = 5/6, Tabla 2 x 3]

k	n	X^2	χ_w^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	58.5	53.1	45.4	14.7	11.4	3.7
	20	89.6	66.3	67.2	47.4	45.9	35.5
	30	97.2	84.4	84.3	72.9	72.1	66.4
	50	99.6	97.6	97.5	94.9	94.7	92.6

• Tablas 3x3:

Los resultados aparecen en las Tablas 8.10 y 8.11 y los Gráficos A.I.5, A.I.6, A.I.11 y A.I.12 del Anexo 1.

En este caso, el χ^2_w para n= 10 alcanza un error de tipo I del 65%; un 10% para n=30 y para n = 50 es de aproximadamente del 7,5%.

La estadística de Rao es superior en el sentido de que para n=10 el nivel es del 18%, se estabiliza para n = 30, para terminar en un nivel próximo al 5% en n=50.

Aún para un nivel de significación nominal del 7,5%, no se logra niveles satisfactorios de los verdaderos niveles en las tres estadísticas más conservadoras.

Las potencias no son comparables en niveles bajos de n debido a la gran diferencia en los niveles reales de significación.

Para la alternativa planteada, los porcentajes de rechazo de χ^2_w y $\chi^2_{R_1}$, alcanzan un nivel del 70%; le siguen $\chi^2_{R_2}$ y $\chi^2_{R_3}$, y $\chi^2_{R_4}$ es el que tiene menor porcentaje (63%).

TABLA 8.10

Niveles de significación reales (%) de los "tests" para un nivel nominal de α =5% cuando se verifica Ho. [π = {0.109714 0.118857 0.114286, 0.118857 0.128762 0.123810, 0.091428 0.099048 0.095238}, η_1 = 9/10, η_2 = 5/6, Tabla 3 x 3]

k	n	X^2	χ_W^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	20.9	65.9	17.1	0.8	0.5	0.1
	20	41.7	19.4	6.8	0.9	0.9	0.4
	30	45.6	10.3	4.8	1.8	1.7	1.0
	50	47.3	7.4	6.1	3.0	2.8	1.2
8	10	40.8	54.1	16.6	1.3	1.1	0.2
	20	55.3	17.3	6.7	0.8	0.7	0.0
	30	58.8	12.9	7.0	1.5	1.4	0.6
	50	60.6	8.8	6.0	2.7	2.7	1.6
10	10	51.3	49.5	15.2	1.6	1.3	0.1
	20	57.7	18.6	8.6	1.3	1.3	0.8
	30	61.1	12.4	8.1	2.5	2.4	1.5
	50	58.3	9.6	7.6	2.2	2.5	1.3

TABLA 8.11

Porcentaje de Rechazo de los "tests" para un nivel nominal de α =5% cuando no se verifica Ho. [π = {0.125 0.100 0.075, 0.100 0.125 0.075, 0.075 0.100 0.225}, η_1 = 9/10, η_2 = 5/6, Tabla 3 x 3]

k	n	X^2	χ_W^2	χ_{R1}^2	χ_{R2}^2	χ_{R3}^2	χ_{R4}^2
5	10	36.0	77.2	29.4	3.1	2.0	0.0
	20	74.9	49.1	33.9	16.5	15.4	5.9
	30	88.9	53.3	48.4	31.4	30.3	21.3
	50	96.4	74.5	74.0	62.7	62.2	56.6

9. CONCLUSIONES

En la literatura sobre el tema no existen antecedentes de modelos donde es posible manipular los efectos de diseños marginales a través de parámetros del mismo. Esa es la característica que distingue al modelo descripto en la sección 6.

Por esa razón, se lo puede utilizar para simular situaciones que sirven para relacionar los comportamientos de los "tests" descriptos en la sección 5, tanto con los efectos de diseño de la tabla (en forma conjunta) como con los efectos de las distribuciones marginales de la misma.

9.1 *Efectos de diseño de la tabla y comportamiento de los "tests"*

En este aspecto, las simulaciones realizadas corroboran, en general, las conclusiones obtenidas por otros autores (Rao y Scott (1980), Thomas y Rao (1987), Holt y col. (1980))

en el caso de otros modelos y las mismas u otras hipótesis. Por cierto, existen algunos matices que las separan. En este trabajo se obtuvieron los resultados que siguen:

La estadística X^2 (chi-cuadrado tradicional) conduce a errores de tipo I mucho mayores que el indicado por su valor nominal. La magnitud real del error tipo I, aumenta cuando la muestra crece.

χ^2_w lleva en general a errores de tipo I también altos pero el problema desaparece cuando la muestra es grande (mayor de 30 ó 50 según el número de celdas de la tabla).

Para tablas 2x2, χ^2_w tiene mejor comportamiento que sus competidores aún para tamaños de muestras moderados ó chicos.

Además, en lo que respecta a inestabilidad, cuando las tablas son de tipo 2x2, resulta una estadística aceptable. Sin embargo, se muestra bastante inestable para tablas de mayor tamaño (su variancia alcanza valores cercanos a 822 en el caso de tablas de tipo 3x3) en las 1000 muestras simuladas.

En general, el comportamiento de χ^2_w está afectado negativamente por el aumento de las celdas de la tabla; parece que este aumento no llega a compensarse con aumentos moderados del tamaño de la muestra. Este es un problema que Koch, Freeman y Freeman (1977) no destacaron suficientemente en el momento de su creación. Los ejemplos que presentan utilizan muestras grandes para tablas de gran dimensión, sin que se cuestionen la validez de los procedimientos.

$\chi^2_{R_1}$ tiene, bajo la hipótesis nula, un comportamiento errático en las tablas 2x2 y conduce a errores de tipo I altos. Sólo para muestras de 50 ó más se acerca al nivel nominal. Este resultado tampoco ha sido suficientemente aclarado en la literatura sobre el tema. Cuando el número de celdas aumenta, $\chi^2_{R_1}$ se vuelve más adecuado que el χ^2_w, para el "test" de independencia, especialmente para tamaños de muestra menores que 30.

$\chi^2_{R_2}$, $\chi^2_{R_3}$ y $\chi^2_{R_4}$ son "tests" conservadores. Sus niveles de significación son menores que

el nivel nominal del 5%. Cuanto mayor es el número de celdas de la tabla, tanto menor el verdadero error de tipo I. En las tablas 2x2 la pérdida se sitúa en un margen del 2%, pero para las tablas mayores esa pérdida es mayor y ni siquiera el aumento de la muestra soluciona el problema.

Se aconseja fijar en 7,5% el valor nominal del error de tipo I para $\chi^2_{R_2}$. Si la tabla es 2x2, se aconseja el mismo valor para $\chi^2_{R_3}$ o $\chi^2_{R_4}$. Para tablas mayores aún usando un nivel nominal de 7,5% sus errores de tipo I caen por debajo del 5%. Deberían realizarse nuevas simulaciones para determinar cuáles deberían ser los errores nominales para alcanzar un error de tipo real próximo al 5%.

Las potencias de los diversos "tests" sólo son comparables si los errores de tipo I que los afectan tiene un valor real común. Es por eso que, si la tabla es 2 x 2, n=50 y el valor nominal para $\chi^2_{R_2}$, $\chi^2_{R_3}$ y $\chi^2_{R_4}$ se toma igual al 7,5%, todos los "tests" estudiados pueden cotejarse y del examen surge que son, en cuanto a potencia, equivalentes.

Para las tablas 2x3 y 3x3, las potencias de $\chi^2_{R_1}$ y χ^2_W son semejantes entre sí (si la muestra es de 50 conglomerados ó más). No se puede compararlos con los otros tres "tests", ya que estos son conservadores. Con respecto a la potencia de estos últimos, $\chi^2_{R_2}$ y $\chi^2_{R_3}$ aparecen como igualmente potentes. El más conservador de los tres es $\chi^2_{R_4}$ para el cual el porcentaje de rechazo es más bajo que el de los otros dos.

9.2 *Efectos de diseño marginales y comportamiento de los "tests"*

Con respecto a las consecuencias de los cambios en los márgenes de las tablas sobre las estadísticas estudiadas, se corrobora que a valores crecientes de η_1, se obtienen efectos de diseño mayores en la variable identificada por las filas y que orto tanto ocurre con η_2 y las columnas.

En todos los casos estudiados $\hat{\delta}_.$, promedio de los autovalores obtenidos de la estimación de (3.1.11), resultó menor que el menor de los promedios de los "deffs" de filas y columnas. Este hecho ya habia sido conjeturado por Rao y Scott(1980).

Un hecho relacionado con el anterior es que el X^2 tradicional- al menos para tablas 2x2- es menos herético cuando al menos uno de los márgenes de la tabla tiene, en promedio, efectos de diseño chicos. Esta observación se extiende a todos los tamaños de muestra simulados.

9.3 *Recomendaciones*

Resumiendo, si la tabla es del tipo 2x2, la estadística de Wald tiene mejor comportamiento que sus rivales. En este caso, si uno de los márgenes de la tabla tiene, en promedio, efectos de diseño bajos, el X^2 tradicional no es demasiado liberal.

En las tablas 2x3 y 3x3, la estadística $\chi^2_{R_1}$ es la que debe preferirse (si se tiene acceso a la matriz de covariancias de las estimaciones). Si esta última condición no se cumple, se pueden usar $\chi^2_{R_2}$, $\chi^2_{R_3}$ y $\chi^2_{R_4}$ pero con valores nominales, para los errores de tipo I superiores o iguales al 7,5%.

Se debería descartar el uso del X^2 sin corregir, en el análisis de tablas de contingencia, salvo que la situación sea la antes mencionada.

Las muestras deben tener, al menos, 30 conglomerados.

Seria conveniente que el INDEC generalizara, en lo posible, la práctica de publicar los efectos de diseño marginales junto a las tablas obtenidas a través de muestras complejas, ya que $\chi^2_{R_4}$ es una estadística que demostró ser aceptable cuando los niveles de significación se modifican adecuadamente y resulta fácil de calcular. Esta práctica mejoraría los análisis de tablas de contingencia de los usuarios con un costo computacional semejante al del X^2 tradicional.

9.4 *Investigación futura*

El modelo presentado en la sección 6, puede ser más explotado, realizando nuevas simulaciones que ilustren mejor sobre la influencia individual de los efectos de diseño de cada variable sobre las estadísticas de los diversos "tests".

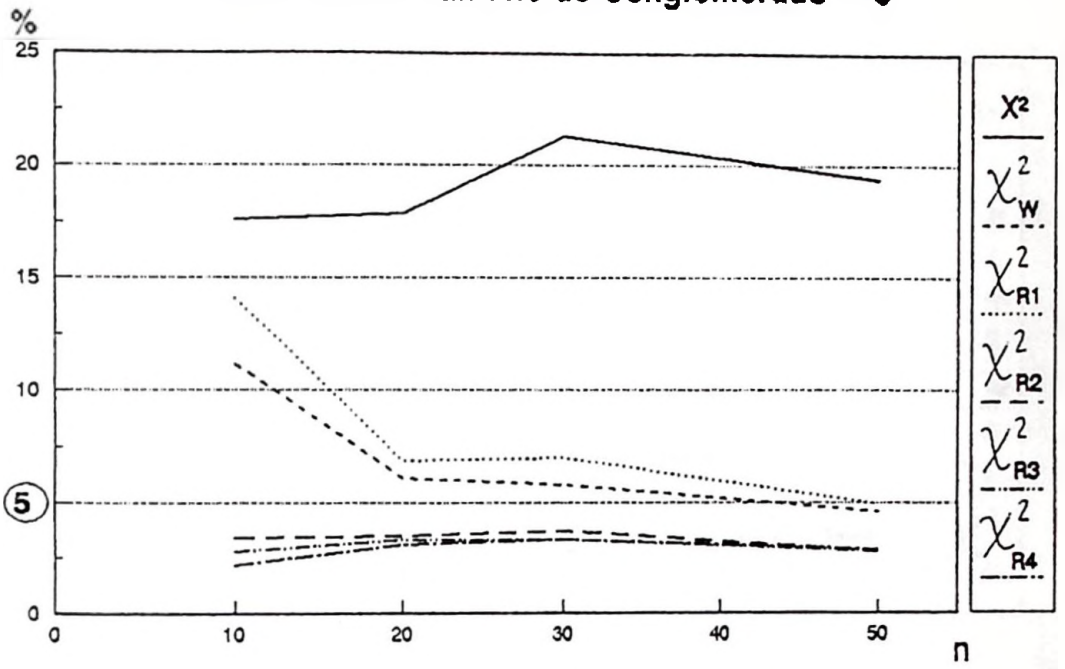
Por otro lado, el modelo es flexible y puede ser utilizado con estratificación, tamaños desiguales de los conglomerados, conglomerados con diferentes efectos de diseño dentro del mismo estrato, etc.

Aportaría una experiencia valiosa extender las simulaciones a estos casos más complejos, así como a tablas de mayor número de márgenes.

ANEXO I

Grafico A.I.1

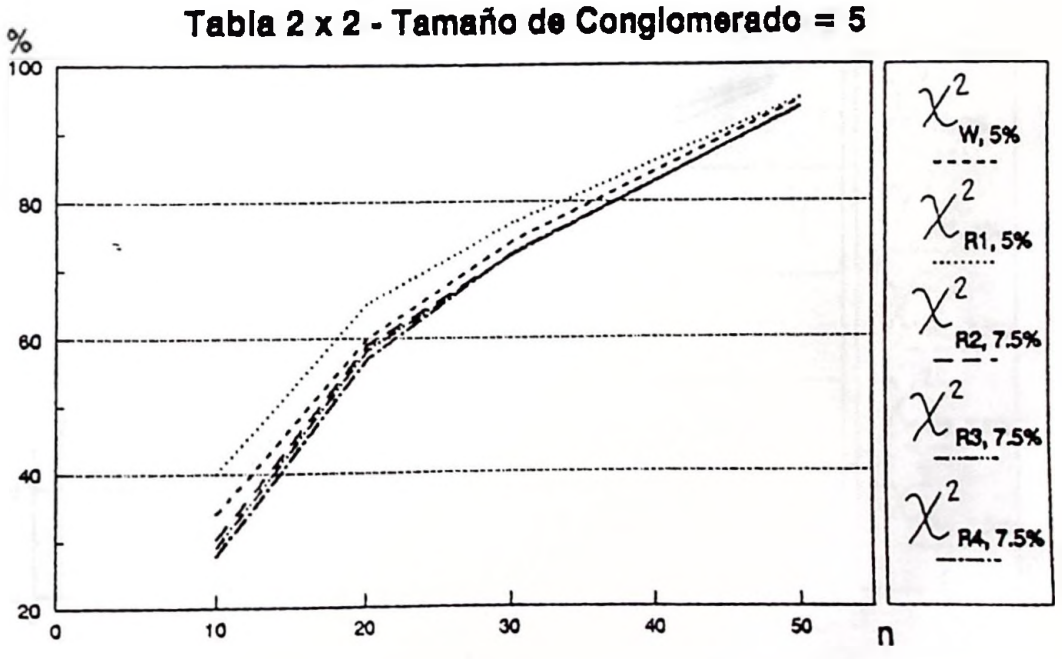
Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia
Tabla 2 x 2 - Tamaño de Conglomerado = 5



π = {0.16 0.24, 0.24 0.36}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.2

Porcentajes de Rechazo de los Tests para un Nivel Nominal del 5% y del 7.5% bajo la Alternativa
Tabla 2 x 2 - Tamaño de Conglomerado = 5

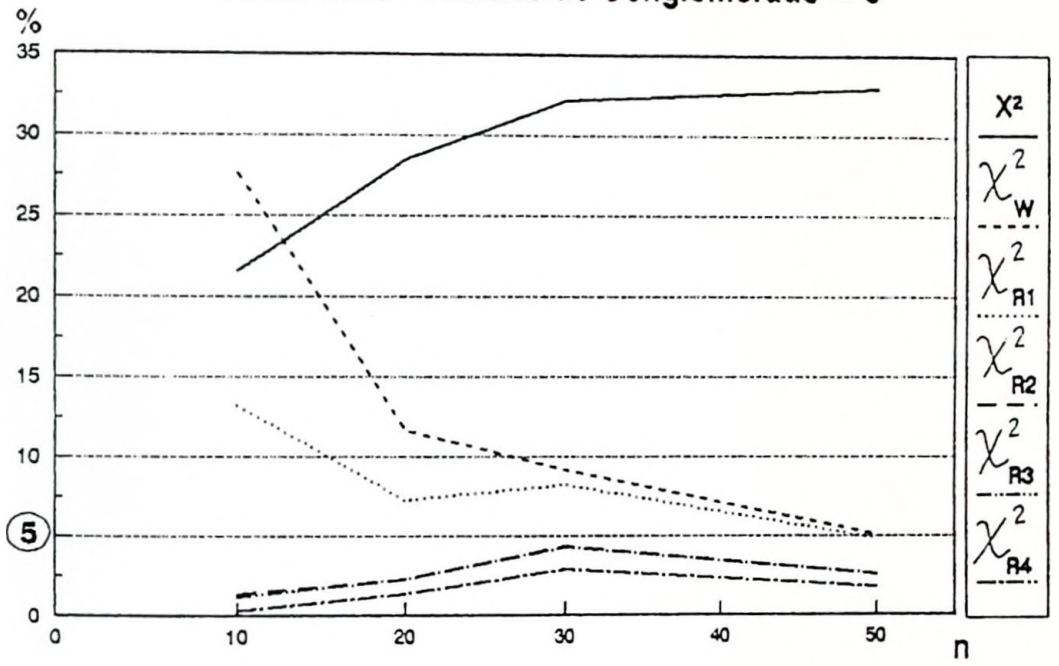


π = {0.35 0.15, 0.15 0.35}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.3

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 2 x 3 - Tamaño de Conglomerado = 5

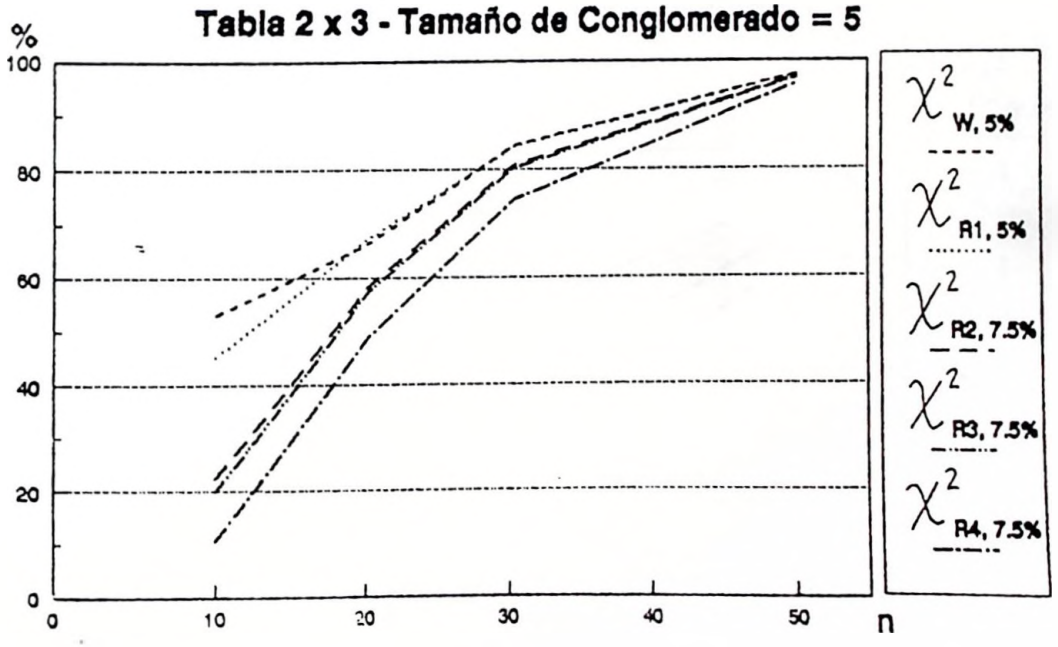


π = {0.160 0.096 0.144, 0.240 0.144 0.216}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.4

Porcentajes de Rechazo de los Tests para un Nivel Nominal del 5% y del 7.5% bajo la Alternativa

Tabla 2 x 3 - Tamaño de Conglomerado = 5

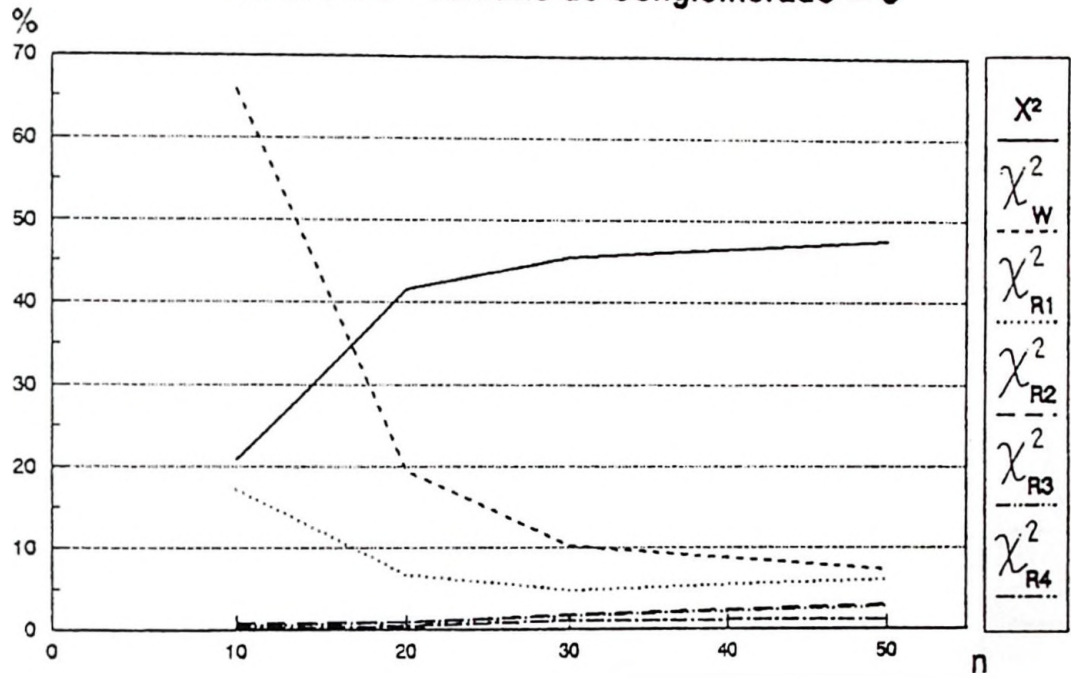


π = {0.200 0.250 0.100, 0.100 0.100 0.250}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.5

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 3 x 3 - Tamaño de Conglomerado = 5

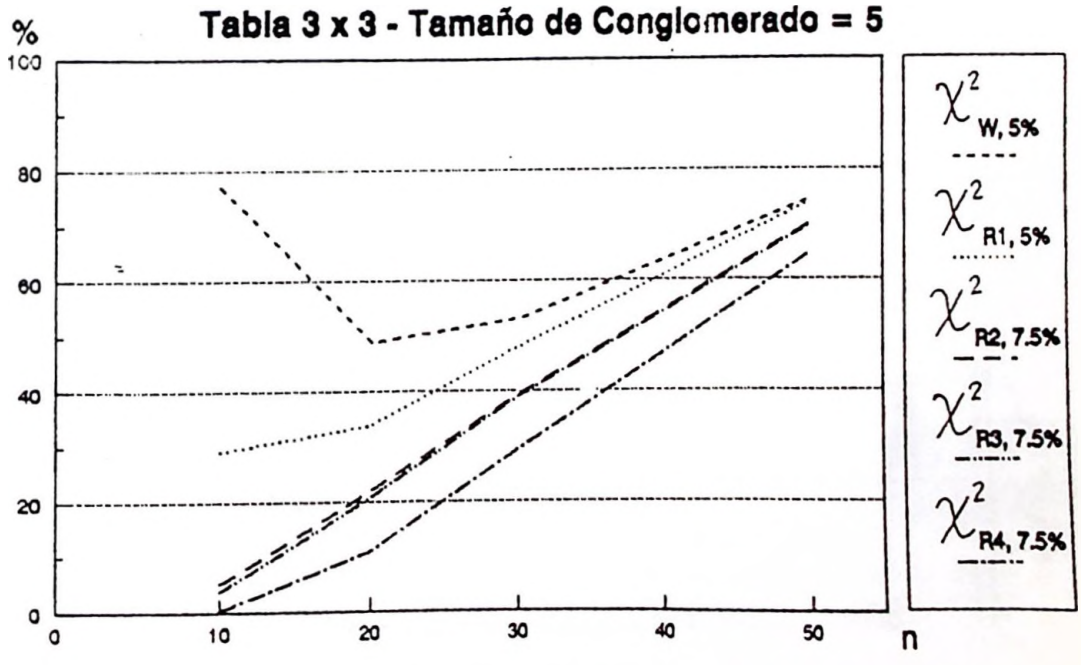


π = {0.109714 0.118857 0.114286, 0.118857 0.128762 0.123810, 0.091428 0.099048 0.095238}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.6

Porcentajes de Rechazo de los Tests para un Nivel Nominal del 5% y del 7.5% bajo la Alternativa

Tabla 3 x 3 - Tamaño de Conglomerado = 5

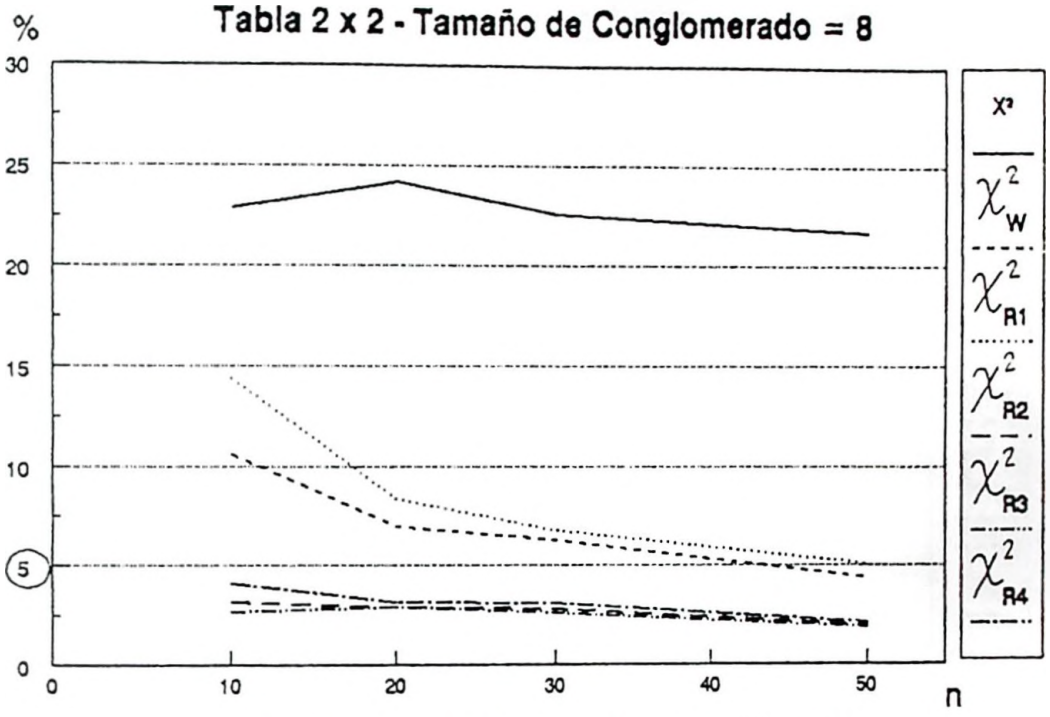


π = {0.125 0.100 0.075, 0.100 0.125 0.075, 0.075 0.100 0.225}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.7

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 2 x 2 - Tamaño de Conglomerado = 8

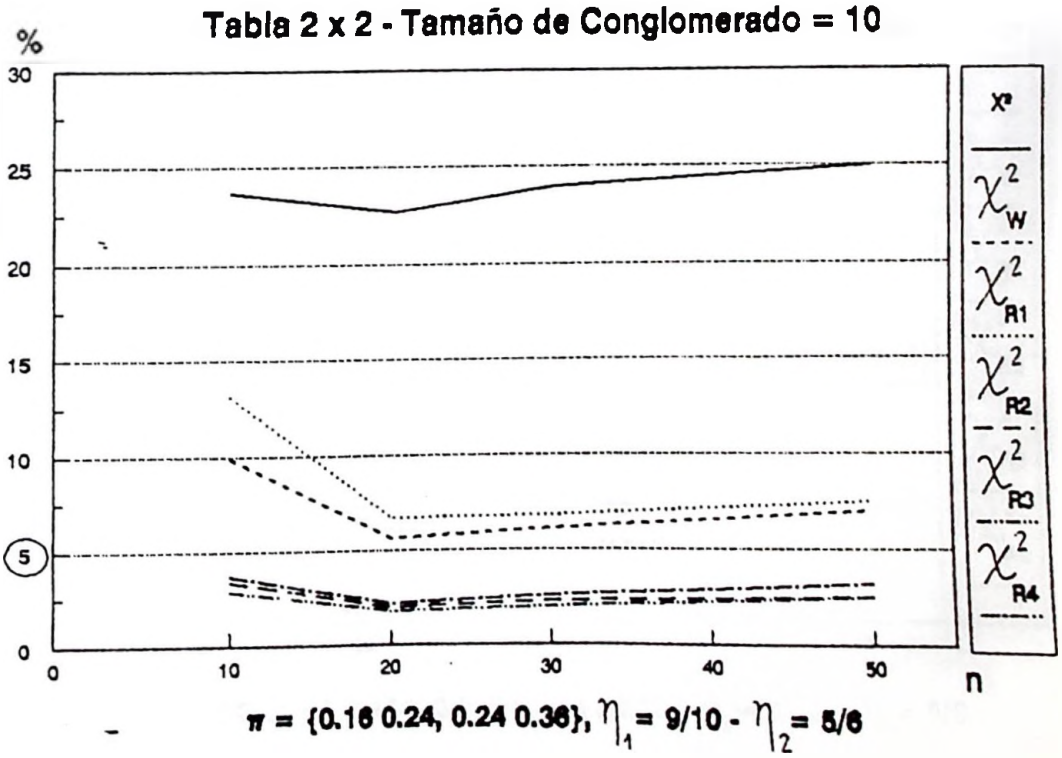


π = {0.16 0.24, 0.24 0.36}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.8

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 2 x 2 - Tamaño de Conglomerado = 10

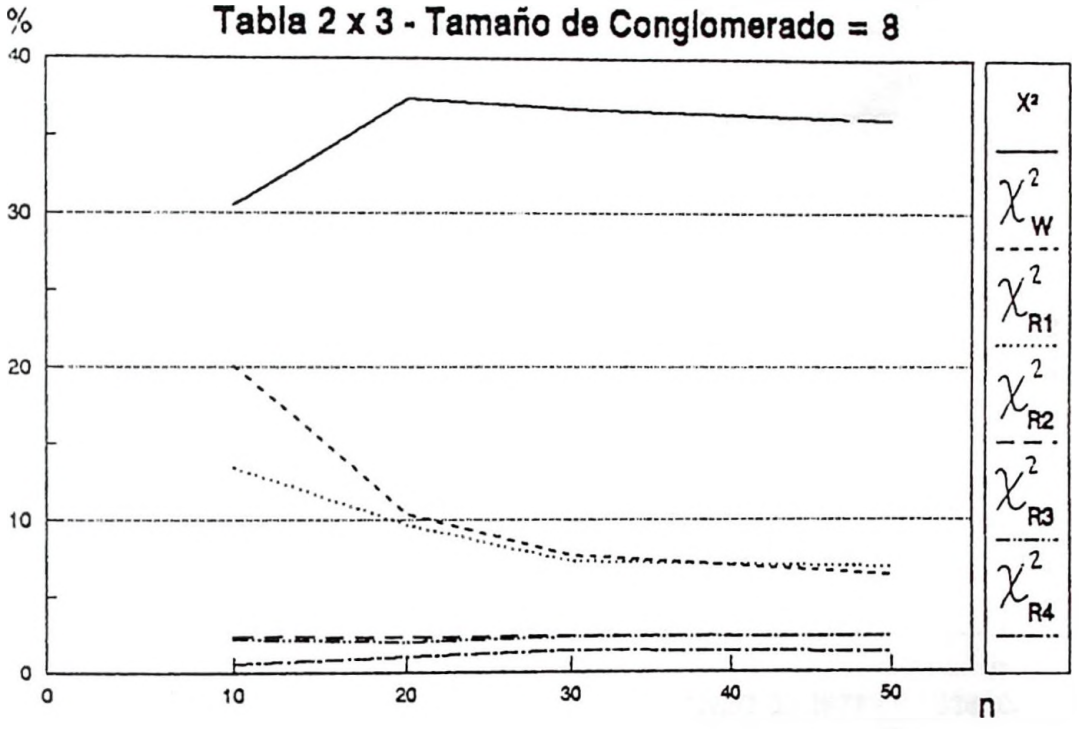


π = {0.16 0.24, 0.24 0.36}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.9

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 2 x 3 - Tamaño de Conglomerado = 8

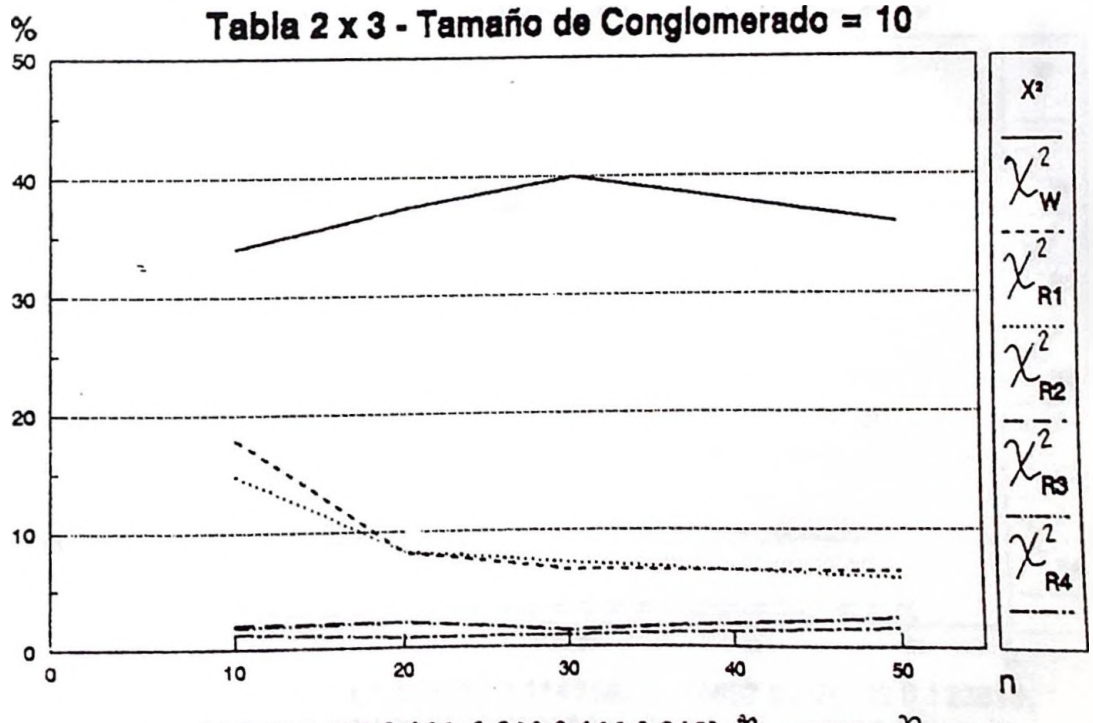


π = {0.160 0.096 0.144, 0.240 0.144 0.216}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.10

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 2 x 3 - Tamaño de Conglomerado = 10

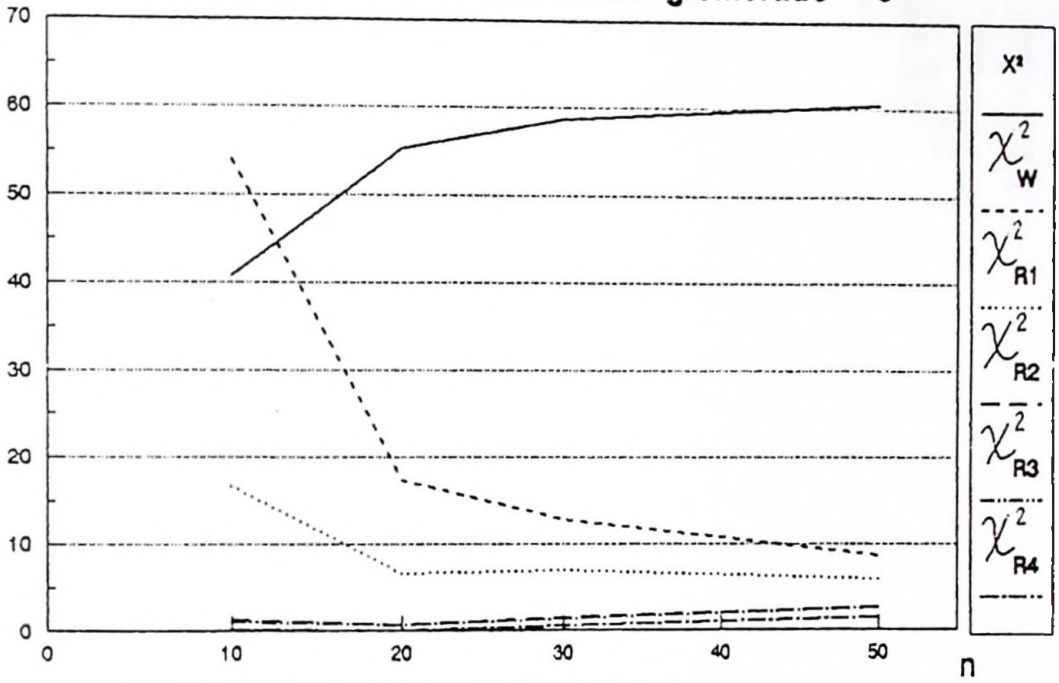


π = {0.160 0.096 0.144, 0.240 0.144 0.216}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.11

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 3 x 3 - Tamaño de Conglomerado = 8

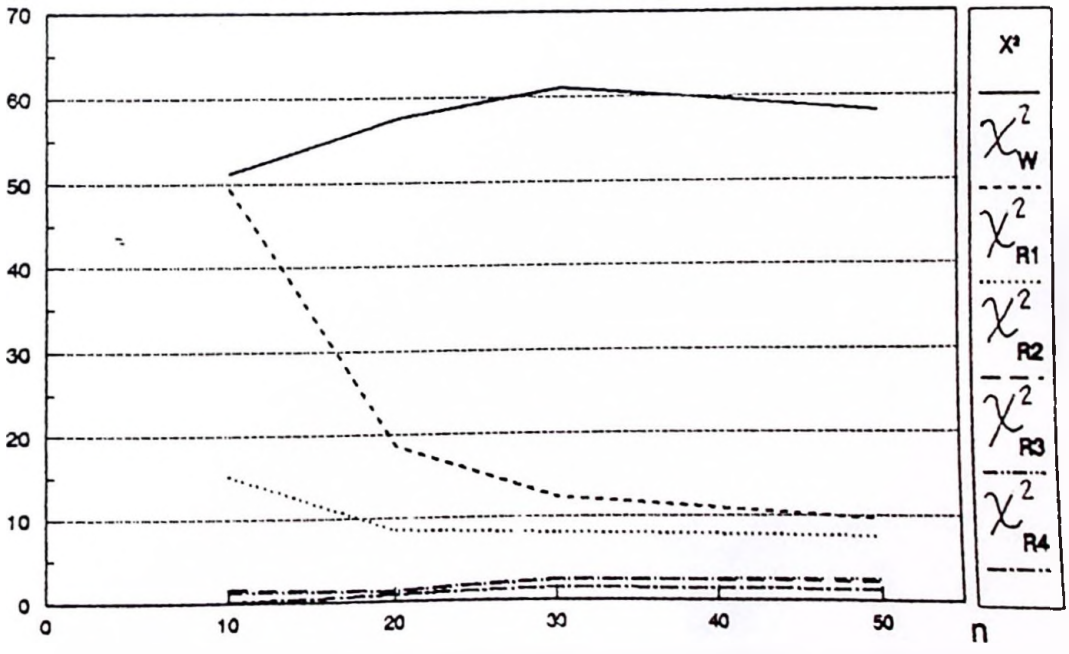


π = {0.109714 0.118857 0.114286, 0.118857 0.128762 0.123810, 0.091428 0.099048 0.095238}, η_1 = 9/10 - η_2 = 5/6

Grafico A.I.12

Niveles de Significación Reales de los Tests para un Nivel Nominal del 5% bajo Independencia

Tabla 3 x 3 - Tamaño de Conglomerado = 10



π = {0.109714 0.118857 0.114286, 0.118857 0.128762 0.123810, 0.091428 0.099048 0.095238}, η_1 = 9/10 - η_2 = 5/6

ANEXO II

II.1. *Tablas 2 x 2*

II.1.1. Programa: INI_A.PRG

```
* Programa de Calculo de las Probabilidades Iniciales. Tabla 2 x 2;
* Nombre del Programa: INI_A.PRG;
* Parametros de entrada: Valores de PI;
*                        Valores de ETA1 y ETA2;
*                        Valor de k (Tamaño del Conglomerado);
* Resultados: Valores de los a'es;
* Ultima Update: 16/12/93;
proc iml;
* Definicion de Parametros;
        k = 5;
        eta1 = {0.9 0.1, 0.1 0.9};
        eta2 = {0.8333 0.1667, 0.1667 0.8333};
        pi = {0.16, 0.24, 0.24, 0.36};
* Plantemiento del Sistema de Ecuaciones;
        m1 = I(4);
        m2 = eta1 @ eta2;
        m = m1 + m2;
        do j = 3 to k;
                ms = m2 ** (j-1);
                m = m + ms;
        end;
* Resolucion del Sistema de Ecuaciones
        im = inv(m);
        aes = inv(m) * (k # pi);
* Impresion de Resultados;
        print 'Valores de los Aes', aes;
quit;
```

II.1.2. Programa: CRE_A.PRG

```
* Programa de Generación de Muestras de Conglomerados para Tablas 2 x 2;
* Nombre del Programa: CRE_A.PRG;
* Parametros a especificar: Probabilidades Iniciales (A'es);
*                           Eta1 y Eta2 (Probabilidades de Cambio);
*                           Phi1 y Phi2 (1 - Eta's);
*                           Tamaño de Conglomerados (k);
*                           Numero de Conglomerados en la Muestra (n);
* Resultados: Muestra en 'MUE_A.SSD';
* Ultima Update: 15/12/93;
        data MUE_A;
                eta1 = 0.9; phi1 = 0.1;
                eta2 = 0.9; phi2 = 0.1;
                k = 5;
                n = 10;
                do i = 1 to n;
                        a1 = .0526856; a2 = .2550593;
                        a3 = .1697738; a4 = .5224812;
                        do j = 1 to k;
                                x = rantbl(0, a1, a2, a3, a4);
                                y1 = 0; y2 = 0; y3 = 0; y4 = 0;
                                select (x);
                                        when (1) do;
                                                a1 = eta1 * eta2;
                                                a2 = eta1 * phi2;
                                                a3 = phi1 * eta2;
                                                a4 = phi1 * phi2;
                                                y1 = 1;
                                                end;
                                        when (2) do;
                                                a1 = eta1 * phi2;
                                                a2 = eta1 * eta2;
                                                a3 = phi1 * phi2;
                                                a4 = phi1 * eta2;
                                                y2 = 1;
                                                end;
                                        when (3) do;
                                                a1 = phi1 * eta2;
                                                a2 = phi1 * phi2;
                                                a3 = eta1 * eta2;
                                                a4 = eta1 * phi2;
                                                y3 = 1;
                                                end;
                                        when (4) do;
                                                a1 = phi1 * phi2;
```

```
                                        a2 = phi1 * eta2;
                                        a3 = eta1 * phi2;
                                        a4 = eta1 * eta2;
                                        y4 = 1;
                                        end;
                            end;
                            keep i y1 y2 y3 y4;
                            output;
                    end;
            end;
run;
```

II.1.3. Programa: TES_A.PRG

```
* Programa de Calculo de los Tests Chi-Cuadrados en Tablas 2 x 2;
* Nombre del Programa: TES_A.PRG;
* Parametros a Especificar: Cantidad de Conglomerados en la Muestra (n);
* Archivo de Entrada: MUE_A;
* Archivos de Salida: Tests y Deffs de la Muestra en 'RES_A';
* Ultima Update: 15/12/93;

* Lectura de la Muestra y Calculo de PI Estimados;
      proc iml;
            n = 20;
            r = 2;
            c = 2;
            rc = 4;
            gl = 1;
            use MUE_A;
            do g = 1 to n;
                  read all var{y1 y2 y3 y4} into tabla where (i = g);
                  t = t || tabla[+,]`;
            end;
            close MUE_A;
            free tabla;
            pin = t[,:];
            kes = t[+,];
            pid = kes[:];
            nk = kes[+];
            do f = 1 to rc;
                  if (pin[f] = 0) then pin[f] = 0.01;
            end;
            pi = pin / pid;

* Variancia Multinomial;
            p = ((diag(pi)) - (pi*pi`))/ nk;

* Calculo de Sigma: Matriz de Variancias de los Numeradores;
            desvio = t - repeat(pin,1,n);
            sigma = (desvio * desvio`)/(n # (n - 1));
            free desvio;

* Calculo de la Variancia de los Pi Estimados;
            u = (1/n) # (t - (pi # repeat(t[+,],rc,1)));
            v = u[,:];
            wac = (u - repeat(v,1,n));
            var =  ((n - 1) / n) # ((wac * wac`) / (pid # pid));
            free u wac;
```

```
* Definicion de la Hipotesis;
                q = {1 -1 -1 1};

* Calculo del Chi-Cuadrado Tipo Wald;
                fw = q * log(pin);
                hw = q * diag(1/pin);
                sw = hw * sigma * hw`;
                chiwal = fw * inv(sw) * fw`;
                pawal = 1 - probchi(chiwal,gl);
                free fw hw sw;

* Calculo del Chi Cuadrado Comun;
                fc = q * log(pi);
                hc = q * diag(1/pi);
                sc = hc * p * hc`;
                chicom = fc * inv(sc) * fc`;
                pacom = 1 - probchi(chicom,gl);

* Calculo del Deff del Test;
                defft = inv(sc) * (hc * var * hc`);
                free fc hc sc;

* Calculo de la Correccion de Rao;
                chirao = chicom / defft;
                parao = 1 - probchi(chirao,gl);

* Calculo de los Deff Marginales;
                matpi = shape(pi,r,c);
                picol = matpi[,+];
                pirow = matpi[+,];
                jtot = j(rc,1);
                jcol = j(r,1);
                jrow = j(1,c);

                vpi0mas = var[1,1] + var[2,2] + (2 # var[1,2]);
                vpi1mas = var[3,3] + var[4,4] + (2 # var[3,4]);
                vpimas0 = var[1,1] + var[3,3] + (2 # var[1,3]);
                vpimas1 = var[2,2] + var[4,4] + (2 # var[2,4]);

                vpicol = vpi0mas // vpi1mas;
                vpirow = vpimas0 || vpimas1;

                free vpi0mas vpi1mas vpimas0 vpimas1;

                varianc = vecdiag(var);
                des = varianc / ((pi # (jtot - pi)) / nk);
```

```
                dcol = vpicol / ((picol # (jcol - picol)) / nk);
                drow = vpirow / ((pirow # (jrow - pirow)) / nk);

* Calculo del Deff por Celda y por Variable;
                defffell = des[:];
                defflamb = ((jtot - pi)` * des) / (rc - 1);
                deffcol = (jcol - picol)` * dcol / (r - 1);
                deffrow = (jrow - pirow) * drow` / (c - 1);
                deffmin = min(deffcol, deffrow);

* Calculo de la Correccion por el Deff Minimo (R4);
                chicol = chicom / deffmin;  pacol = 1 - probchi(chicol,gl);

* Calculo de la Correccion de Fellegi (R3);
                chico2 = chicom / defffell; paco2 = 1 - probchi(chico2,gl);

* Calculo de la Correccion por Lambda . (R2);
                chico3 = chicom / defflamb; paco3 = 1 - probchi(chico3,gl);

                resu = chiwal || pawal || chicom || pacom || chirao ||
                parao || chicol || pacol || chico2 || paco2 || chico3 ||
                paco3 || defft || deffcol || deffrow || defffell || defflamb;

                edit RES_A;
                append from resu;
                close RES_A;
        quit;
```

II.1.4. Programa: SIM_A.PRG

```
* Programa de Simulación en Tablas 2 x 2;
* Nombre del Programa: SIM_A.PRG;
* Parámetros a Especificar: Número de Repeticiones;
* Archivos de Entrada: CRE_A.PRG;
*                      TES_A.PRG;
* Archivos de Salida: TES_A.SSD (Archivo SAS con Tests y Deffs);
*                     RES_A.LST (Listado de Tests y Deffs);
*                     UNI_A.LST (Estadisticas Descriptivas de Tests y Deffs);
* Ultima Update: 15/12/93;

* Creación del Archivo de Resultados;
        proc iml;
                result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
                c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                     CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                     DEFFFELL DEFFLAMB};
                create RES_A from result [colname = c];
        quit;

        %macro ITE_A;
                %let ITERAC=1;
                %do ITERAC=1 %to 10;
                        %include 'CRE_A.PRG';
                        %include 'TES_A.PRG';
                %end;
        %mend ITE_A;

        %ITE_A;

* Tests y Deffs;
        data b2m.TES_A;
                set RES_A;
        run;
        proc printto new print="RES_A.LST";
        run;
        proc print data=RES_A;
                title  'RESULTADO DE nnnn SIMULACIONES. 15-10-93.';
                title2 'TESTS Y DEFFS.';
        run;

* Decisiones y Univariate;
        proc printto new print="UNI_A.LST";
        run;
        data tablas;
```

```
        set RES_A;
        if pawal > 0.05 then dwal = 'NR'; else dwal = 'R';
        if pacom > 0.05 then dcom = 'NR'; else dcom = 'R';
        if parao > 0.05 then drao = 'NR'; else drao = 'R';
        if pacol > 0.05 then dcol = 'NR'; else dcol = 'R';
        if paco2 > 0.05 then dco2 = 'NR'; else dco2 = 'R';
        if paco3 > 0.05 then dco3 = 'NR'; else dco3 = 'R';
run;
proc univariate data=RES_A;
        var chiwal chicom chirao chicol chico2 chico3
            defft deffcol deffrow defffell defflamb;
run;
proc freq data=tablas;
        tables dwal dcom drao dcol dco2 dco3;
run;
```

II.1.5. Programa: APL_A.PRG

```
* Programa de Aplicacion de los Tests en Tablas 2 x 2;
* Nombre del Programa: APL_A.PRG;
* Archivo de Entrada: 'DAT_A.DAT';
*                     'TES_A.PRG';
* Nota: El archivo de Entrada debe contener 3 variables:
*       . Identificación del Conglomerado;
*       . Variables de Interes (x1 y x2);
* Archivos de Salida: RES_A.LST (Listado de Tests y Deffs);
* Ultima Update: 21/12/93;


* Creación del Archivo de Resultados;
       proc iml;
              result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
              c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                   CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                   DEFFFELL DEFFLAMB};
              create RES_A from result [colname = c];
       quit;

* Lectura y Recodificación de Datos;
       data MUE_A;
              infile 'c:\b2m\dat_a.dat';
              input i x1 x2;
              y1 = 0; y2 = 0; y3 = 0; y4 = 0;
              if (x1 = 1) and (x2 = 1) then y1 = 1;
              if (x1 = 1) and (x2 = 2) then y2 = 2;
              if (x1 = 2) and (x2 = 1) then y3 = 3;
              if (x1 = 2) and (x2 = 2) then y4 = 4;
              keep i y1 y2 y3 y4;
              output;
       run;

       %include 'TES_A.PRG';

* Tests y Deffs;
       proc printto new print="RES_A.LST";
       run;
       proc print data=RES_A;
              title  'TESTS y DEFFS.';
       run;
```

II.2. *Tablas 2 x 3*

II.2.1. Programa: INI_B.PRG

```
* Programa de Calculo de las Probabilidades Iniciales. Tabla 2 x 3;
* Nombre del Programa: INI_B.PRG;
* Parametros de entrada: Valores de PI;
*                        Valores de ETA1 y ETA2;
*                        Valor de k (Tamaño del Conglomerado);
* Resultados: Valores de los a'es;
* Ultima Update: 21/12/93;

proc iml;
* Definicion de Parametros;
        k = 5;
        eta1 = {0.9 0.1, 0.1 0.9};
        eta2 = {0.8333 0.08335 0.08335, 0.08335 0.8333 0.08335,
                0.08335 0.08335 0.8333};
        pi = {0.12, 0.12, 0.16, 0.18, 0.18, 0.24};
* Plantemiento del Sistema de Ecuaciones;
        m1 = I(6);
        m2 = eta1 @ eta2;
        m = m1 + m2;
        do j = 3 to k;
                ms = m2 ** (j-1);
                m = m + ms;
        end;
* Resolucion del Sistema de Ecuaciones
        im = inv(m);
        aes = inv(m) * (k # pi);
* Impresion de Resultados;
        print 'Valores de los Aes', aes;
quit;
```

II.2.2. Programa: CRE_B.PRG

```
* Programa de Generación de Muestras de Conglomerados para Tablas 2 x 3;
* Nombre del Programa: CRE_B.PRG;
* Parametros a especificar: Probabilidades Iniciales (A'es);
*                           Eta1 y Eta2 (Probabilidades de Cambio);
*                           Phi1 y Phi2 (1 - Eta's);
*                           Tamaño de Conglomerados (k);
*                           Numero de Conglomerados en la Muestra (n);
* Resultados: Muestra en 'MUE_B.SSD';
* Ultima Update: 21/12/93;

        data MUE_B;
                eta1 = 0.9; phi1 = 0.1;
                eta2 = 0.8333; phi2 = 0.08335;
                k = 5;
                n = 20;
                do i = 1 to n;
                        a1 = .1572646; a2 = .0608386; a3 = .1331581;
                        a4 = .2853409; a5 = .1195137; a6 = .2438841;
                        do j = 1 to k;
                                x = rantbl(0, a1, a2, a3, a4, a5, a6);
                                y1 = 0; y2 = 0; y3 = 0;
                                y4 = 0; y5 = 0; y6 = 0;
                                select (x);
                                        when (1) do;
                                                a1 = eta1 * eta2;
                                                a2 = eta1 * phi2;
                                                a3 = eta1 * phi2;
                                                a4 = phi1 * eta2;
                                                a5 = phi1 * phi2;
                                                a6 = phi1 * phi2;
                                                y1 = 1;
                                                end;
                                        when (2) do;
                                                a1 = eta1 * phi2;
                                                a2 = eta1 * eta2;
                                                a3 = eta1 * phi2;
                                                a4 = phi1 * phi2;
                                                a5 = phi1 * eta2;
                                                a6 = phi1 * phi2;
                                                y2 = 1;
                                                end;
                                        when (3) do;
                                                a1 = eta1 * phi2;
                                                a2 = eta1 * phi2;
```

```
                                        a3 = eta1 * eta2;
                                        a4 = phi1 * phi2;
                                        a5 = phi1 * phi2;
                                        a6 = phi1 * eta2;
                                        y3 = 1;
                                        end;
                                when (4) do;
                                        a1 = phi1 * eta2;
                                        a2 = phi1 * phi2;
                                        a3 = phi1 * phi2;
                                        a4 = eta1 * eta2;
                                        a5 = eta1 * phi2;
                                        a6 = eta1 * phi2;
                                        y4 = 1;
                                        end;
                                when (5) do;
                                        a1 = phi1 * phi2;
                                        a2 = phi1 * eta2;
                                        a3 = phi1 * phi2;
                                        a4 = eta1 * phi2;
                                        a5 = eta1 * eta2;
                                        a6 = eta1 * phi2;
                                        y5 = 1;
                                        end;
                                when (6) do;
                                        a1 = phi1 * phi2;
                                        a2 = phi1 * phi2;
                                        a3 = phi1 * eta2;
                                        a4 = eta1 * phi2;
                                        a5 = eta1 * phi2;
                                        a6 = eta1 * eta2;
                                        y6 = 1;
                                        end;
                        end;
                        keep i y1 y2 y3 y4 y5 y6;
                        output;
                end;
        end;
run;
```

II.2.3. Programa: TES_B.PRG

```
* Programa de Calculo de los Tests Chi-Cuadrados en Tablas 2 x 3;
* Nombre del Programa: TES_B.PRG;
* Parametros a Especificar: Cantidad de Conglomerados en la Muestra (n);
* Archivo de Entrada: MUE_B;
* Archivos de Salida: Tests y Deffs de la Muestra en 'RES_B';
* Ultima Update: 21/12/93;

* Lectura de la Muestra y Calculo de PI Estimados;
      proc iml;
            n = 20;
            r = 2;
            c = 3;
            rc = 6;
            gl = 2;
            use MUE_B;
            do g = 1 to n;
                  read all var{y1 y2 y3 y4 y5 y6} into tabla
                     where (i = g);
                  t = t || tabla[+,]`;
            end;
            close MUE_B;
            free tabla;
            pin = t[,:];
            kes = t[+,];
            pid = kes[:];
            nk = kes[+];
            do f = 1 to rc;
                  if (pin[f] = 0) then pin[f] = 0.01;
            end;
            pi = pin / pid;

* Variancia Multinomial;
            p = ((diag(pi)) - (pi*pi`)) / nk;

* Calculo de Sigma: Matriz de Variancias de los Numeradores;
            desvio = t - repeat(pin,1,n);
            sigma = (desvio * desvio`)/(n # (n - 1));
            free desvio;

* Calculo de la Variancia de los Pi Estimados;
            u = (1/n) # (t - (pi # repeat(t[+,],rc,1)));
            v = u[,:];
            wac = (u - repeat(v,1,n));
          _ var =  ((n - 1) / n) # ((wac * wac`) / (k # k));
```

```
                free u wac;

* Definicion de la Hipotesis;
                q = {1 0 -1 -1  0 1,
                     0 1 -1  0 -1 1};

* Calculo del Chi-Cuadrado Tipo Wald;
                fw = q * log(pin);
                hw = q * diag(1/pin);
                sw = hw * sigma * hw`;
                chiwal = fw`* inv(sw) * fw;
                pawal = 1 - probchi(chiwal,gl);
                free fw hw sw;

* Calculo del Chi Cuadrado Comun;
                fc = q * log(pi);
                hc = q * diag(1/pi);
                sc = hc * p * hc`;
                chicom = fc` * inv(sc) * fc;
                pacom = 1 - probchi(chicom,gl);


* Calculo del Deff del Test;
                scinv = inv(sc);
                hcvar = (hc * var * hc`);
                aux1 = root(scinv);
                aux2 = aux1 * hcvar * aux1`;
                call eigen(auval,auvec,aux2);
                defft = auval[:];
                free fc hc sc scinv hcvar aux1 aux2 auval auvec;


* Calculo de la Correccion de Rao (R1);
                chirao = chicom / defft;
                parao = 1 - probchi(chirao,2);

* Calculo de los Deff Marginales;
                matpi = shape(pi,r,c);
                picol = matpi[,+];
                pirow = matpi[+,];
                jtot = j(rc,1);
                jcol = j(r,1);
                jrow = j(1,c);

                vpi0mas = var[1,1] + var[2,2] + var[3,3]
                        + (2 # var[1,2]) + (2 # var[1,3]) + (2 # var[2,3]);
```

```
                vpi1mas = var[4,4] + var[5,5] + var[6,6]
                        + (2 # var[4,5]) + (2 # var[4,6]) + (2 # var[5,6]);

                vpimas0 = var[1,1] + var[4,4] + (2 # var[1,4]);
                vpimas1 = var[2,2] + var[5,5] + (2 # var[2,5]);
                vpimas2 = var[3,3] + var[6,6] + (2 # var[3,6]);

                vpicol = vpi0mas // vpi1mas;
                vpirow = vpimas0 || vpimas1 || vpimas2;

                free vpi0mas vpi1mas vpimas0 vpimas1 vpimas2;

                varianc = vecdiag(var);
                des = varianc / ((pi # (jtot - pi)) / nk);

                dcol = vpicol / ((picol # (jcol - picol)) / nk);
                drow = vpirow / ((pirow # (jrow - pirow)) / nk);

* Calculo del Deff por Celda y por Variable;
                defffell = des[:];
                defflamb = ((jtot - pi)` * des) / (rc - 1);
                deffcol = ((jcol - picol)` * dcol) / (c - 1);
                deffrow = ((jrow - pirow) * drow`) / (c - 1);
                deffmin = min(deffcol, deffrow);

* Calculo de la Correccion por el Deff Minimo (R4);
                chicol = chicom / deffmin;  pacol = 1 - probchi(chicol,gl);

* Calculo de la Correccion de Fellegi (R3);
                chico2 = chicom / defffell; paco2 = 1 - probchi(chico2,gl);

* Calculo de la Correccion por Lambda . (R2);
                chico3 = chicom / defflamb; paco3 = 1 - probchi(chico3,gl);

                resu = chiwal || pawal || chicom || pacom || chirao ||
                parao || chicol || pacol || chico2 || paco2 || chico3 ||
                paco3 || defft || deffcol || deffrow || defffell || defflamb;

                edit RES_B;
                append from resu;
                close RES_B;
        quit;
```

II.2.4. Programa: SIM_B.PRG

```
* Programa de Simulación en Tablas 2 x 3;
* Nombre del Programa: SIM_B.PRG;
* Parámetros a Especificar: Número de Repeticiones;
* Archivos de Entrada: CRE_B.PRG;
*                      TES_B.PRG;
* Archivos de Salida: TES_B.SSD (Archivo SAS con Tests y Deffs);
*                     RES_B.LST (Listado de Tests y Deffs);
*                     UNI_B.LST (Estadisticas Descriptivas de Tests y Deffs);
* Ultima Update: 21/12/93;

* Creación del Archivo de Resultados;
        proc iml;
                result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
                c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                     CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                     DEFFFELL DEFFLAMB};
                create RES_B from result {colname = c};
        quit;

        %macro ITE_B;
                %let ITERAC=1;
                %do ITERAC=1 %to 10;
                        %include 'CRE_B.PRG';
                        %include 'TES_B.PRG';
                %end;
        %mend ITE_B;

        %ITE_B;

* Tests y Deffs;
        data b2m.TES_B;
                set RES_B;
        run;
        proc printto new print="RES_B.LST";
        run;
        proc print data=RES_B;
                title  'RESULTADO DE nnnn SIMULACIONES. 21-12-93.';
                title2 'TESTS Y DEFFS.';
        run;

* Decisiones y Univariate;
        proc printto new print="UNI_B.LST";
        run;
        data tablas;
```

```
        set RES_B;
        if pawal > 0.05 then dwal = 'NR'; else dwal = 'R';
        if pacom > 0.05 then dcom = 'NR'; else dcom = 'R';
        if parao > 0.05 then drao = 'NR'; else drao = 'R';
        if paco1 > 0.05 then dco1 = 'NR'; else dco1 = 'R';
        if paco2 > 0.05 then dco2 = 'NR'; else dco2 = 'R';
        if paco3 > 0.05 then dco3 = 'NR'; else dco3 = 'R';
run;
proc univariate data=RES_B;
        var chiwal chicom chirao chico1 chico2 chico3
            defft deffcol deffrow defffell defflamb;
run;
proc freq data=tablas;
        tables dwal dcom drao dco1 dco2 dco3;
run;
```

II.2.5. Programa: APL_B.PRG

```
* Programa de Aplicacion de los Tests en Tablas 2 x 3;
* Nombre del Programa: APL_B.PRG;
* Archivo de Entrada: 'DAT_B.DAT';
*                     'TES_B.PRG';
* Nota: El archivo de Entrada debe contener 3 variables:
*       . Identificación del Conglomerado;
*       . Variables de Interes (x1 y x2);
* Archivos de Salida: RES_B.LST (Listado de Tests y Deffs);
* Ultima Update: 21/12/93;


* Creación del Archivo de Resultados;
       proc iml;
              result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
              c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                   CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                   DEFFFELL DEFFLAMB};
              create RES_B from result {colname = c};
       quit;

* Lectura y Recodificación de Datos;
       data MUE_B;
              infile 'c:\b2m\dat_b.dat';
              input i x1 x2;
              y1 = 0; y2 = 0; y3 = 0; y4 = 0; y5 = 0; y6 = 0;
              if (x1 = 1) and (x2 = 1) then y1 = 1;
              if (x1 = 1) and (x2 = 2) then y2 = 1;
              if (x1 = 1) and (x2 = 3) then y3 = 1;
              if (x1 = 2) and (x2 = 1) then y4 = 1;
              if (x1 = 2) and (x2 = 2) then y5 = 1;
              if (x1 = 2) and (x2 = 3) then y6 = 1;
              keep i y1 y2 y3 y4 y5 y6;
              output;
       run;

       %include 'TES_B.PRG';

* Tests y Deffs;
       proc printto new print="RES_B.LST";
       run;
       proc print data=RES_B;
              title  'TESTS y DEFFS.';
       run;
```

II.3. *Tablas 3 x 3*

II.3.1. Programa: INI_C.PRG

```
* Programa de Calculo de las Probabilidades Iniciales. Tabla 3 x 3;
* Nombre del Programa: INI_C.PRG;
* Parametros de entrada: Valores de PI;
*                        Valores de ETA1 y ETA2;
*                        Valor de k (Tamaño del Conglomerado);
* Resultados: Valores de los a'es;
* Ultima Update: 21/12/93;

proc iml;
* Definicion de Parametros;
        k = 5;
        eta1 = {0.9 0.05 0.05, 0.05 0.9 0.05, 0.05 0.05 0.9};
        eta2 = {0.9 0.05 0.05, 0.05 0.9 0.05, 0.05 0.05 0.9};
        pi = {0.109714, 0.118857, 0.114286,
              0.118857, 0.128762, 0.123810,
              0.091429, 0.099048, 0.095238};
* Plantemiento del Sistema de Ecuaciones;
        m1 = I(9);
        m2 = p1 @ p2;
        m = m1 + m2;
        do j = 3 to k;
                ms = m2 ** (j-1);
                m = m + ms;
        end;
* Resolucion del Sistema de Ecuaciones
        im = inv(m);
        aes = inv(m) * (k # pi);
* Impresion de Resultados;
        print 'Valores de los Aes', aes;
quit;
```

II.3.2. Programa: CRE_C.PRG

```
* Programa de Generación de Muestras de Conglomerados para Tablas 3 x 3;
* Nombre del Programa: CRE_C.PRG;
* Parametros a especificar: Probabilidades Iniciales (A'es);
*                           Eta1 y Eta2 (Probabilidades de Cambio);
*                           Phi1 y Phi2 (1 - Eta's);
*                           Tamaño de Conglomerados (k);
*                           Numero de Conglomerados en la Muestra (n);
* Resultados: Muestra en 'MUE_C.SSD';
* Ultima Update: 21/12/93;

        data MUE_C;
                eta1 = 0.9; phi1 = 0.05;
                eta2 = 0.8333; phi2 = 0.08335;
                k = 5;
                n = 50;
                do i = 1 to n;
                        a1 = .1078486; a2 = .1229329; a3 = .1153917;
                        a4 = .1199171; a5 = .1365451; a6 = .1282320;
                        a7 = .0837116; a8 = .0957103; a9 = .0897107;
                        do j = 1 to k;
                          x = rantbl(0, a1, a2, a3, a4, a5, a6, a7, a8, a9);
                                y1 = 0; y2 = 0; y3 = 0;
                                y4 = 0; y5 = 0; y6 = 0;
                                y7 = 0; y8 = 0; y9 = 0;
                                select (x);
                                        when (1) do;
                                                a1 = eta1 * eta2;
                                                a2 = eta1 * phi2;
                                                a3 = eta1 * phi2;
                                                a4 = phi1 * eta2;
                                                a5 = phi1 * phi2;
                                                a6 = phi1 * phi2;
                                                a7 = phi1 * eta2;
                                                a8 = phi1 * phi2;
                                                a9 = phi1 * phi2;
                                                y1 = 1;
                                                end;
                                        when (2) do;
                                                a1 = eta1 * phi2;
                                                a2 = eta1 * eta2;
                                                a3 = eta1 * phi2;
                                                a4 = phi1 * phi2;
                                                a5 = phi1 * eta2;
                                                a6 = phi1 * phi2;
```

```
            a7 = phi1 * phi2;
            a8 = phi1 * eta2;
            a9 = phi1 * phi2;
            y2 = 1;
            end;
    when (3) do;
            a1 = eta1 * phi2;
            a2 = eta1 * phi2;
            a3 = eta1 * eta2;
            a4 = phi1 * phi2;
            a5 = phi1 * phi2;
            a6 = phi1 * eta2;
            a7 = phi1 * phi2;
            a8 = phi1 * phi2;
            a9 = phi1 * eta2;
            y3 = 1;
            end;
    when (4) do;
            a1 = phi1 * eta2;
            a2 = phi1 * phi2;
            a3 = phi1 * phi2;
            a4 = eta1 * eta2;
            a5 = eta1 * phi2;
            a6 = eta1 * phi2;
            a7 = phi1 * eta2;
            a8 = phi1 * phi2;
            a9 = phi1 * phi2;
            y4 = 1;
            end;
    when (5) do;
            a1 = phi1 * phi2;
            a2 = phi1 * eta2;
            a3 = phi1 * phi2;
            a4 = eta1 * phi2;
            a5 = eta1 * eta2;
            a6 = eta1 * phi2;
            a7 = phi1 * phi2;
            a8 = phi1 * eta2;
            a9 = phi1 * phi2;
            y5 = 1;
            end;
    when (6) do;
            a1 = phi1 * phi2;
            a2 = phi1 * phi2;
            a3 = phi1 * eta2;
            a4 = eta1 * phi2;
```

```
                a5 = eta1 * phi2;
                a6 = eta1 * eta2;
                a7 = phi1 * phi2;
                a8 = phi1 * phi2;
                a9 = phi1 * eta2;
                y6 = 1;
                end;
        when (7) do;
                a1 = phi1 * eta2;
                a2 = phi1 * phi2;
                a3 = phi1 * phi2;
                a4 = phi1 * eta2;
                a5 = phi1 * phi2;
                a6 = phi1 * phi2;
                a7 = eta1 * eta2;
                a8 = eta1 * phi2;
                a9 = eta1 * phi2;
                y7 = 1;
                end;
        when (8) do;
                a1 = phi1 * phi2;
                a2 = phi1 * eta2;
                a3 = phi1 * phi2;
                a4 = phi1 * phi2;
                a5 = phi1 * eta2;
                a6 = phi1 * phi2;
                a7 = eta1 * phi2;
                a8 = eta1 * eta2;
                a9 = eta1 * phi2;
                y8 = 1;
                end;
        when (9) do;
                a1 = phi1 * phi2;
                a2 = phi1 * phi2;
                a3 = phi1 * eta2;
                a4 = phi1 * phi2;
                a5 = phi1 * phi2;
                a6 = phi1 * eta2;
                a7 = eta1 * phi2;
                a8 = eta1 * phi2;
                a9 = eta1 * eta2;
                y9 = 1;
                end;
end;
keep i y1 y2 y3 y4 y5 y6 y7 y8 y9;
output;
```

```
                end;
            end;
    run;
```

II.3.3. Programa: TES_C.PRG

```
* Programa de Calculo de los Tests Chi-Cuadrados en Tablas 3 x 3;
* Nombre del Programa: TES_C.PRG;
* Parametros a Especificar: Cantidad de Conglomerados en la Muestra (n);
* Archivo de Entrada: MUE_C;
* Archivos de Salida: Tests y Deffs de la Muestra en 'RES_C';
* Ultima Update: 21/12/93;

* Lectura de la Muestra y Calculo de PI Estimados;
        proc iml;
                n = 50;
                r = 3;
                c = 3;
                rc = 9;
                gl = 4;
                use MUE_C;
                do g = 1 to n;
                        read all var{y1 y2 y3 y4 y5 y6 y7 y8 y9}
                                into tabla where (i = g);
                        t = t || tabla[+,]`;
                end;
                close MUE_C;
                free tabla;
                pin = t[,:];
                kes = t[+,];
                pid = kes[:];
                nk = kes[+];
                do f = 1 to rc;
                        if (pin[f] = 0) then pin[f] = 0.01;
                end;
                pi = pin / pid;

* Variancia Multinomial;
                p = ((diag(pi)) - (pi*pi`)) / nk;

* Calculo de Sigma: Matriz de Variancias de los Numeradores;
                desvio = t - repeat(pin,1,n);
                sigma = (desvio * desvio`)/(n # (n - 1));
                free desvio;

* Calculo de la Variancia de los Pi Estimados;
                u = (1/n) # (t - (pi # repeat(t[+,],rc,1)));
                v = u[,:];
                wac = (u - repeat(v,1,n));
                var =  ((n - 1) / n) # ((wac * wac`) / (k # k));
```

```
                free u wac;

* Definicion de la Hipotesis;
                q = {1  0 -1  0  0  0 -1  0  1,
                     0  1 -1  0  0  0  0 -1  1,
                     0  0  0  1  0 -1 -1  0  1,
                     0  0  0  0  1 -1  0 -1  1};

* Calculo del Chi-Cuadrado Tipo Wald;
                fw = q * log(pin);
                hw = q * diag(1/pin);
                sw = hw * sigma * hw`;
                chiwal = fw`* inv(sw) * fw;
                pawal = 1 - probchi(chiwal,gl);
                free fw hw sw;

* Calculo del Chi Cuadrado Comun;
                fc = q * log(pi);
                hc = q * diag(1/pi);
                sc = hc * p * hc`;
                chicom = fc` * inv(sc) * fc;
                pacom = 1 - probchi(chicom,gl);


* Calculo del Deff del Test;
                scinv = inv(sc);
                hcvar = (hc * var * hc`);
                aux1 = root(scinv);
                aux2 = aux1 * hcvar * aux1`;
                call eigen(auval,auvec,aux2);
                defft = auval[:];
                free fc hc sc scinv hcvar aux1 aux2 auval auvec;

* Calculo de la Correccion de Rao (R1);
                chirao = chicom / defft;
                parao = 1 - probchi(chirao,gl);

* Calculo de los Deff Marginales;
                matpi = shape(pi,r,c);
                picol = matpi[,+];
                pirow = matpi[+,];
                jtot = j(rc,1);
                jcol = j(c,1);
                jrow = j(1,r);
                vpi0mas = var[1,1] + var[2,2] + var[3,3]
                        + (2 # var[1,2]) + (2 # var[1,3]) + (2 # var[2,3]);
```

```
                vpi1mas = var[4,4] + var[5,5] + var[6,6]
                        + (2 # var[4,5]) + (2 # var[4,6]) + (2 # var[5,6]);
                vpi2mas = var[7,7] + var[8,8] + var[9,9]
                        + (2 # var[7,8]) + (2 # var[7,9]) + (2 # var[8,9]);

                vpimas0 = var[1,1] + var[4,4] + var[7,7]
                        + (2 # var[1,4]) + (2 # var[1,7]) + (2 # var[4,7]);
                vpimas1 = var[2,2] + var[5,5] + var[8,8]
                        + (2 # var[2,5]) + (2 # var[2,8]) + (2 # var[5,8]);
                vpimas2 = var[3,3] + var[6,6] + var[9,9]
                        + (2 # var[3,6]) + (2 # var[3,9]) + (2 # var[6,9]);

                vpicol = vpi0mas // vpi1mas // vpi2mas;
                vpirow = vpimas0 || vpimas1 || vpimas2;
                free vpi0mas vpi1mas vpi2mas vpimas0 vpimas1 vpimas2;

                varianc = vecdiag(var);
                des = varianc / ((pi # (jtot - pi)) / nk);

                dcol = vpicol / ((picol # (jcol - picol)) / nk);
                drow = vpirow / ((pirow # (jrow - pirow)) / nk);

* Calculo del Deff por Celda y por Variable;
                defffell = des[:];
                defflamb = ((jtot - pi)` * des) / (rc - 1);
                deffcol = ((jcol - picol)` * dcol) / (r - 1);
                deffrow = ((jrow - pirow) * drow`) / (c - 1);
                deffmin = min(deffcol, deffrow);

* Calculo de la Correccion por el Deff Minimo (R4);
                chicol = chicom / deffmin;  pacol = 1 - probchi(chicol,gl);

* Calculo de la Correccion de Fellegi (R3);
                chico2 = chicom / defffell; paco2 = 1 - probchi(chico2,gl);

* Calculo de la Correccion por Lambda . (R2);
               chico3 = chicom / defflamb; paco3 = 1 - probchi(chico3,gl);

                resu = chiwal || pawal || chicom || pacom || chirao ||
                parao || chicol || pacol || chico2 || paco2 || chico3 ||
                paco3 || defft || deffcol || deffrow || defffell || defflamb;

                edit RES_C;
                append from resu;
                close RES_C;
        quit;
```

II.3.4. Programa: SIM_C.PRG

```
* Programa de Simulación en Tablas 3 x 3;
* Nombre del Programa: SIM_C.PRG;
* Parámetros a Especificar: Número de Repeticiones;
* Archivos de Entrada: CRE_C.PRG;
*                      TES_C.PRG;
* Archivos de Salida: TES_C.SSD (Archivo SAS con Tests y Deffs);
*                     RES_C.LST (Listado de Tests y Deffs);
*                     UNI_C.LST (Estadisticas Descriptivas de Tests y
Deffs);
* Ultima Update: 21/12/93;

* Creación del Archivo de Resultados;
        proc iml;
                result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
                c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                     CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                     DEFFFELL DEFFLAMB};
                create RES_C from result [colname = c];
        quit;

        %macro ITE_C;
                %let ITERAC=1;
                %do ITERAC=1 %to 10;
                        %include 'CRE_C.PRG';
                        %include 'TES_C.PRG';
                %end;
        %mend ITE_C;

        %ITE_C;

* Tests y Deffs;
        data b2m.TES_C;
                set RES_C;
        run;
        proc printto new print="RES_C.LST";
        run;
        proc print data=RES_C;
                title  'RESULTADO DE nnnn SIMULACIONES. 21-12-93.';
                title2 'TESTS Y DEFFS.';
        run;


* Decisiones y Univariate;
        proc printto new print="UNI_C.LST";
        run;
```

```
data tablas;
        set RES_C;
        if pawal > 0.05 then dwal = 'NR'; else dwal = 'R';
        if pacom > 0.05 then dcom = 'NR'; else dcom = 'R';
        if parao > 0.05 then drao = 'NR'; else drao = 'R';
        if pacol > 0.05 then dcol = 'NR'; else dcol = 'R';
        if paco2 > 0.05 then dco2 = 'NR'; else dco2 = 'R';
        if paco3 > 0.05 then dco3 = 'NR'; else dco3 = 'R';
run;
proc univariate data=RES_C;
        var chiwal chicom chirao chicol chico2 chico3
            defft deffcol deffrow defffell defflamb;
run;
proc freq data=tablas;
        tables dwal dcom drao dcol dco2 dco3;
run;
```

II.3.5. Programa: APL_C.PRG

```
* Programa de Aplicacion de los Tests en Tablas 3 x 3;
* Nombre del Programa: APL_C.PRG;
* Archivo de Entrada: 'DAT_C.DAT';
*                     'TES_C.PRG';
* Nota: El archivo de Entrada debe contener 3 variables:
*       . Identificación del Conglomerado;
*       . Variables de Interes (x1 y x2);
* Archivos de Salida: RES_C.LST (Listado de Tests y Deffs);
* Ultima Update: 21/12/93;

* Creación del Archivo de Resultados;
       proc iml;
              result = {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0};
              c = {CHIWAL PAWAL CHICOM PACOM CHIRAO PARAO CHICO1 PACO1
                   CHICO2 PACO2 CHICO3 PACO3 DEFFT DEFFCOL DEFFROW
                   DEFFFELL DEFFLAMB};
              create RES_C from result [colname = c];
       quit;
* Lectura y Recodificación de Datos;
       data MUE_C;
              infile 'c:\b2m\dat_c.dat';
              input i x1 x2;
              y1 = 0; y2 = 0; y3 = 0; y4 = 0; y5 = 0;
              y6 = 0; y7 = 0; y8 = 0; y9 = 0;
              if (x1 = 1) and (x2 = 1) then y1 = 1;
              if (x1 = 1) and (x2 = 2) then y2 = 1;
              if (x1 = 1) and (x2 = 3) then y3 = 1;
              if (x1 = 2) and (x2 = 1) then y4 = 1;
              if (x1 = 2) and (x2 = 2) then y5 = 1;
              if (x1 = 2) and (x2 = 3) then y6 = 1;
              if (x1 = 3) and (x2 = 1) then y7 = 1;
              if (x1 = 3) and (x2 = 2) then y8 = 1;
              if (x1 = 3) and (x2 = 3) then y9 = 1;
              keep i y1 y2 y3 y4 y5 y6 y7 y8 y9;
              output;
       run;
       %include 'TES_C.PRG';

* Tests y Deffs;
       proc printto new print="RES_C.LST";
       run;
       proc print data=RES_C;
              title  'TESTS y DEFFS.';
       run;
```

BIBLIOGRAFIA

Altham, P.A.E. (1976), "Discrete Variable Analysis for individuals Grouped into Families", Biometrika, 63, 263-269.

Binder, D.A. (1983): "On the variance of the asymptotically normal estimators from complex surveys". International Statistical Review, 51, 279-292.

Brier, S.S. (1978), "Discrete Data Models With Random Effecte", Technical Report, University of Minnesota, School of Statistics.

Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Cluster Sampling From Contingency Tables", Journal of the American Statistical Association, 71, 665-670.

Fay, R.E. (1985): "A jackknifed chi-squared test for complex samples", Journal of the American Statistical Association, 80, 148-157.

Fellegi, I.P. (1980), "Aproximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", Journal of the American Statistical Association, 75, 261-268.

Fuller, W.A. (1975), "Regression Analysis for Sample Surveys", Sankhya, C 37, 117-132.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data Linear Models", Biometrics, 25, 489-504.

Hidiroglou, M.A. and Rao, J.H.K. (1987), "Chi-Squared tests with categorical data from complex surveys": Part I, Part II. Journal of Official Statistics. Vol 3 Nro 2, 117-132, 133-140. Statistics Sweden.

Hidiroglou, M.A. and Paton, D.J. (1987); "Some experiences in computing estimates and their vairances using data from complex survey designs". In "Applied Probability, Statistics and Sampling Theory", Boston: D. Reidel Publishing Company.

Holt, D., Scott, A.J., and Ewings, P.O., (1980), "Chi-squared Tests With Survey Data", Journal of the Royal Statistical Society, Ser. A, 143, 302-320.

Judkins, D.R. (1990); "Fay's method for variance estimation". Journal of Official Statistics. Vol 6. 223-240. Statistics Sweden.

Kish, L. (1965); "Survey Sampling". New York: Wiley.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1971); "Strategies in the multivariate analysis of data from complex surveys". International Statistical Review, 43, 59-78.

Rao, J.N.K. and Scott, A.J: (1979), "Chi-squared Tests for Analysis of Categorical Data from Complex Surveys", Proceedings of the American Statistical Association, Section on Survey Research Methods, 58-66.

Rao, J.H.K. and Scott A.J. (1981); "The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables". Journal of the American Statistical Statistical Association, Vol 76. 221-230.

Rao, J.H.K. and Scott A.J. (1987); "On simple adjustments to chi-square tests with sample survey data". Annals of Statistics, 15, 385-397.

Rao, J.N.K., Kumar, S. and Roberts, G. (1989); "Analysis of sample survey data involving categorical response variable: Methods and Software". Survey Methodology. A journal of Statistics Canada, Vol 15, 161-185.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987); "Logistic Regresion analysis of sample survey data". Biometrika, 74, 1-12.

SAS Institute Inc., SAS/IML™ User's Guide, Release 6.03 Edition Cary, NC: SAS Institute Inc., 1988, 357 pp.

SAS Institute Inc., SAS® Language Guide for Personal Computer, Release 6.03 Edition Cary, NC: SAS Institute Inc., 1988, 558 pp.

SAS Institute Inc., SAS Guide to Macro Processing, Version 6 Edition, NC: SAS Institute Inc.1988.

Servy, E., Alonso, M., Arnesi, N., Boggio, G., Sanchez, S. (1989); "Análisis de Tablas de Contingencia a partir de Datos provenientes de Muestras de Diseño Complejo". Convenio INDEC- Facultad de Ciencias Económicas, UNR; Escuela de Estadística.

Solomon, H. and Stephens, M.A. (1977), "Distribution of a Sum of Weighted Chi-Square variables", Journal of American Statistical Association, 72, 881-885.

Thomas, D.R. and Rao, J.N.K. (1987); "Small sample comparisons of level and power for simple goodness of fit statistics under cluster sampling". Journal of the American Statistical Association, 82, 630-636.