UNIVERSIDAD NACIONAL DE COLOMBIA

Tamaño de muestra para funciones de tablas de contingencia vía algoritmos genéticos

Sample size for functions of a contingency table via genetic algorithms

Oveida Rosa Bustos Polo

Universidad Nacional de Colombia, Sede Medellín
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2020

Tamaño de muestras para funciones de tablas de contingencia vía algoritmos genéticos

Sample size for functions of a contingency table via genetic algorithms

Oveida Rosa Bustos Polo

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de Magister en Ciencias Estadística
Magister en Ciencias Estadística

Director(a):
Ph.D., Juan Carlos Correa Morales
Profesor Asociado
Escuela de Estadística
Universidad Nacional de Colombia, sede Medellín

Universidad Nacional de Colombia, Sede Medellín
Facultad Ciencias, Escuela de Estadística
Medellín, Colombia

2020

Este trabajo es dedicado a mi hija Daniela por ser mi fuente de motivación e inspiración para poder superarme cada día más y así poder luchar para que la vida nos depare un futuro mejor.
A mi familia, en especial a mi padre Leonardo quien es mi ejemplo de constancia y dedicación, a mi madre Oveida quien es mi fortaleza e inspiración diaria, a mis hermanos Jorge, Carmen y Leonardo, por sus palabras de aliento, comprensión, enseñanzas y en especial de amor.
A Oscar Daniel por el apoyo y los momentos compartidos al lado de nuestra hija.
A mi asesor de tesis Juan Carlos Correa, por su ayuda y paciencia, ya que sin ellas no habría podido conseguir este logro.
A mis amigos y muy especialemente a Oscar Giraldo, quienes sin esperar nada a cambio compartieron sus conocimientos, alegrías y tristezas. Y a todas aquellas personas que estuvieron a mi lado apoyando para que este sueño se hiciera realidad.

La preocupación por el hombre y su destino siempre debe ser el interés primordial de todo esfuerzo técnico. Nunca olvides esto entre tus diagramas y ecuaciones.

Albert Einstein

Resumen

La metodología GSK propuesta por Grizzle, Starmer y Koch (1969), enmarca su metodología en el análisis de regresión y el método de mínimos cuadrados ponderados, para la estimación de parámetros de un modelo donde las variables respuestas son funciones generadas de una tabla de contingencia, y en particular para modelar variables respuestas que tienen una distribución multinomial. Las principales ventajas de esta metodología son: su flexibilidad en la construcción de funciones respuestas y su facilidad de cálculos.

El método de estimación por mínimos cuadrados ponderados, debido a la naturaleza asintótica de sus pruebas estadísticas, necesita un tamaño muestral mínimo para obtener resultados confiables. Prácticamente cada procedimiento estadístico necesita su propio desarrollo para el cálculo del tamaño muestral requerido. En el análisis de tablas de contingencia, dependiendo del interés del investigador, se debe proceder a realizar cálculos específicos al problema de interés. Por ejemplo, los tamaños muestrales para una razón de odds son diferentes a los tamaños para un riesgo relativo. La metodología GSK permite desarrollar soluciones a muy diversos problemas en el caso de tablas de conteo. La determinación de tamaños muestrales en este caso es un problema complejo, ya que se involucran varias poblaciones multinomiales y funciones de los parámetros de estas poblaciones, que son el objetivo del investigador.

En este trabajo se propone una metodología no desarrollada aún en la literatura para tabla de contingencias, la cual consiste en la determinación del tamaño muestral mediante la utilización de algoritmos genéticos a fin de desarrollar la metodología GSK para la estimación de modelos para datos en tabla de contingencia. Se debe encontrar un vector de tamaños muestrales que cumplan con algún criterio de optimización establecido. Se implementa esta metodología en el software estadístico R.

keywords: Tablas de Contigencia, Tamaño muestrales, Método GSK, Funciones respuesta.

Abstract

The GSK methodology proposed by Grizzle, Starmer and Koch (1969), frames their methodology. a in the regression analysis and the weighted least squares method, for the estimation of parameters of a model where the response variables are generated functions of a contingency table, and in particular to model responses variables that they have a multinomial distribution. The main advantages of this methodology are: its exibility in the construction of response functions and their ease of calculations.

The method of estimation by weighted least squares, due to the asymptotic nature of your statistical tests, you need a minimum sample size to get results with ables. Practically every statistical procedure needs its own development to calculating the required sample size. In the analysis of contingency tables, depending of the researcher's interest, calculations specific to the problem should be carried out of interest is. For example, the sample sizes for an

odds ratio are different from sizes for relative risk. The GSK methodology allows developing solutions to very various problems in the case of counting tables. Determination of sample sizes in this case it is a complex problem, since several multinomial populations are involved and functions of the parameters of these populations, which are the researcher's objective. In this work we propose a methodology not developed to one in the literature for table of contingencies, which consists of determining the sample size by using of genetic algorithms to develop the GSK methodology for estimating of models for data in contingency table. A vector of sizes must be found Show them that they meet some established optimization criteria. This is implemented methodology in the statistical software R.

keywords: Contingency Tables, Sample sizes, GSK method, Response functions.

Índice general

1 Introducción

Grizzle, Starmer y Koch (1969) presentaron una metodología de ahora en adelante denotada por GSK, para análisis de tablas de contingencia que es muy similar a los modelos de regresión lineal para las respuestas del tipo continuo, esta metodología permite respuestas correlacionadas y no requiere varianza constante, y ha demostrado ser una poderosa herramienta para modelar y analizar tablas de contingencia, debido a su naturaleza flexible y de fácil implementación para gran cantidad de modelos usando un lenguaje vectorizado (Correa, 2020)[6].

La determinación del tamaño mínimo de muestra requerido por estrato de una tabla de contingencia, para estimar parámetros de una distribución multinomial, tradicionalmente se realiza controlando el error absoluto entre el estimador y el parámetro desconocido. Es importante tener en cuenta que los tamaños de muestra por estrato deben ser grandes para que los resultados de tipo asintótico sigan siendo válidos. Por tanto este problema ha sido analizado por varios autores, entre ellos Tortora (1978)[24],Thompson (1987)[23] y Cochran (1977) [5]. Infortunadamente existe poca literatura para cálculo del tamaño muestral para funciones de tablas de contigencias usando el método GSK. Rochon (1989)[21] usó la metodología de mínimos cuadrados ponderados GSK sobre variables respuestas y utiliza los procedimientos descritos por la metodología para generar el tamaño mínimo de muestra requerido usando pruebas de hipótesis, para demostrar un efecto específico en los niveles de precisión del modelo y compara con métodos propuestos por otros autores, para el cálculo de:

- Comparación de probabilidades
- Modelos Log-Lineales
- Diseño Estratificados
- Medidas experimentales Repetidas

Y éste encuentra ventajas en el uso de la metodología GSK, ya que las pequeñas diferencias de los resultados obtenidos usando el enfoque GSK contrastados con los obtenidos usando el enfoque de otros autores no son sustanciales e indica que son los mismos, y puesto que en la literatura existe contradicciones con respecto al uso de algún método (Por ejemplo, el criterio menos riguroso el uso del estadístico χ^2 lo convencional ha dictado que el valor esperado

debe superar a 5 en cada celda de una tabla de contingencia, sin embargo, Conhra (1977)[5] sugirió que debe exceder en 1 en cada celda de una tabla de contingencia), el enfoque GSK gana puntos ya que sus raíces están basadas en una metodología estadística bien establecida.

En este trabajo se propone un algoritmo para el cálculo del tamaño mínimo muestral por estrato de una tabla de contingencia para un parámetro del modelo GSK mediante la utilización de un algoritmo genético, el cual es un procedimiento heurístico propuesto para optimizar una función (Haupt y Haupt, 1997)[13] e imita una evolución natural y se ha utilizado para resolver problemas estadísticos como: selección de variables, series de tiempo, etc. Se desarrolla la metodología GSK para la estimación del modelo de interés de una tabla de contingencia y se incluye la implementación del software.

Este trabajo se desarrolla de la siguiente forma:
En el capítulo 2, se presenta una revisión del enfoque GSK, mientras que en el capítulo 3 se realiza una revisón de las técnicas para el cálculo de tamaños muestrales de una distribución multinomial y se proporciona el algoritmo a desarollar en este trabajo. El capítulo 4 está dedicado a la aplicación de la metodología GSK de una situación específica, el capítulo 5 proporciona los resultados de la aplicación del algoritmo, esto es, el cálculo de los tamaños de muestra para los estratos de la Tabla de contigencia utilizando la metodología de algoritmos genéticos y utilizados para la estimación del modelo vía GSK, finalmente conclusiones, recomendaciones a implementar y código del algoritmo en el lenguaje de programación R.

2 El modelo GSK

Grizzle, Starmer y Koch(1969) [12] definen un modelo lineal para análisis de datos categóricos, conocido como *GSK*, éste se basa en conceptos como los modelos lineales, regresión múltiple y mínimos cuadrados ponderados para analizar tablas de contingencia.

La metodología GSK está definida en tres pasos:

1. **Definición de variables respuestas**: En esta metodología, la variable dependiente no se refiere a individuos o probabilidad sino funciones de probabilidades de la tabla de contigencia.

2. **Definición del modelo**: La definición del modelo puede dividirse en dos categorías.
   - Si solo tiene una variable respuesta.
   - Si hay variables independientes que definen estratos o subpoblaciones, se podría considerar construir una matriz de diseño respecto a las variables independientes.

3. **Estimación y Validación**: En esta parte, es necesario considerar los tamaños de muestras disponibles para asegurar resultados asintóticos.

Utilizando la metodología GSK, para modelar variables categóricas, se pueden utilizar funciones de la variable dependiente del modelo, llamada función respuesta. Para el caso de variables categóricas multinomiales, se basa en probabilidades de la variable respuesta o en la tasa de ocurrencia.
Para la formación de las funciones de la variable respuesta, se inicia con la tabla de frecuencia de un conjunto de datos, donde se tienen R categorías con S factores o poblaciones.

2.1. La tabla teórica o poblacional

La distribución de la población es un producto de S distribuciones multinomiales independientes que son definidas sobre R categorías, llamadas respuestas y que pueden ser combinaciones de varias variables categóricas.

Subpoblación	1	2	...	r	...	R	Total
1	π_{11}	π_{12}	...	π_{1r}	...	π_{1R}	1.0
2	π_{21}	π_{22}	...	π_{2r}	...	π_{2R}	1.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	π_{s1}	π_{s2}	...	π_{sr}	...	π_{sR}	1.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S	π_{S1}	π_{S2}	...	π_{Sr}	...	π_{SR}	1.0

Tabla 2-1: Tabla Teórica o Poblacional

La **Tabla2-1** se denota como la matriz $\mathbf{\Pi}$, donde π_{sr} es la probabilidad de que una observación tomada al azar tenga respuesta r dado que pertenece a la subpoblación s.
Con la condición que:

$$\sum_{j=1}^{R} \pi_{ij} = 1 \text{ para } i = 1, 2, \dots S \tag{2-1}$$

2.2. La tabla muestral

Subpoblación	1	2	...	r	...	R	Total
1	n_{11}	n_{12}	...	n_{1r}	...	n_{1R}	n_1
2	n_{21}	n_{22}	...	n_{2r}	...	n_{2R}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	n_{s1}	n_{s2}	...	n_{sr}	...	n_{sR}	n_s
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S	n_{S1}	n_{S2}	...	n_{Sr}	...	n_{SR}	n_S

Tabla 2-2: Tabla muestral

Donde n_i es el tamaño de muestra de la $i-$ésima subpoblación y n_{ij} = representa el número de individuos observados en la celda (i, j).
La **Tabla 2-2** se denota por $\boldsymbol{N}$, tal que

$$\sum_{j=1}^{R} n_{ij} = n_i \text{ para } i = 1, 2, \dots S \tag{2-2}$$

2.3. Representación del vector π

Se asume que $X_i, i = 1, 2, \ldots, S$ es una muestra aleatoria de una distribución multinomial $MN(n_i, (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iR})^T)$ con $\sum_{j=1}^{R} \pi_{ij} = 1$ (Nótese que en la ecuación anterior se usó la notación "MN", frecuente en la literatura estadística, como abreviatura de la frase "Distribución Multinomial"), se denota por $\boldsymbol{\pi}$ el siguiente vector:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \vdots \\ \pi_{1R} \\ \pi_{21} \\ \pi_{22} \\ \vdots \\ \pi_{2R} \\ \vdots \\ \pi_{S1} \\ \pi_{S2} \\ \vdots \\ \pi_{SR} \end{pmatrix}$$

Sea $\boldsymbol{\pi^T} = (\boldsymbol{\pi_1^T}, \boldsymbol{\pi_2^T}, \ldots \boldsymbol{\pi_S^T})$, donde $\boldsymbol{\pi_i^T} = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iR})$, con $i = 1, 2, \ldots, S$.

La función verosimilitud para $\boldsymbol{\pi}$ es:

$$L(\pi_{i1}, \pi_{i2} \ldots \pi_{iR}) = P(X_{i1} = n_{i1}, ..., X_{iR} = n_{iR}; \pi_{i1}, \pi_{i2} \ldots \pi_{iR}) = \frac{n_i!}{n_{i1}! n_{i2}! \ldots n_{iR}!} \pi_{i1}^{n_{i1}} \pi_{i2}^{n_{i2}} \ldots \pi_{iR}^{n_{iR}} \tag{2-3}$$

tal que

$$\sum_{j=1}^{R} n_{ij} = n_i, \text{ con } i = 1, 2, \ldots, S \tag{2-4}$$

Aplicando *log* a la función de verosimilitud se obtiene:

$$l = log\Big(L(\pi_{i1}, \pi_{i2} \dots \pi_{iR})\Big) \tag{2-5}$$

$$= log\Big(\frac{n!}{n_{i1}!n_{i2}! \dots n_{iR}!}\pi_{i1}^{n_{i1}}\pi_{i2}^{n_{i2}} \dots \pi_{iR}^{n_{iR}}\Big) \tag{2-6}$$

$$= log\Big(\frac{n_i!}{n_{i1}!n_{i2}! \dots n_{iR}!}\Big) + log\Big(\pi_{i1}^{n_{i1}}\pi_{i2}^{n_{i2}} \dots \pi_{iR}^{n_{iR}}\Big) \tag{2-7}$$

$$= log\Big(\frac{n_i!}{n_{i1}!n_{i2}! \dots n_{iR}!}\Big) + \sum_{j=1}^{R} log\Big(\pi_{ij}^{n_{ij}}\Big) \tag{2-8}$$

$$= log\Big(\frac{n_i!}{n_{i1}!n_{i2}! \dots n_{iR}!}\Big) + \sum_{j=1}^{R} n_{ij} log\Big(\pi_{ij}\Big) \tag{2-9}$$

con $i = 1, 2, \dots, S$.
Los estimadores de máxima verosimilitud se encuentran de la siguiente forma (Correa 2020)([6]).

Se denota por l^* a:

$$l^* = log(L(\pi_{i1}, \pi_{i2} \dots \pi_{iR})) \tag{2-10}$$

y la condición

$$\sum_{j=1}^{R} \pi_{ij} = 1 \text{ para } i = 1, 2, \dots S. \tag{2-11}$$

Utilizando el método de multiplicadores de Lagrange, se tiene que:

$$\nabla l^*(\pi_{i1}, \pi_{i2} \dots \pi_{iR}) = \lambda \nabla\Big(\sum_{j=1}^{R} \pi_{ij} - 1\Big) \tag{2-12}$$

$$\Big(\frac{\partial}{\partial \pi_{i1}} l^*(\pi_{i1}, \pi_{i2} \dots \pi_{iR}), \frac{\partial}{\partial \pi_{i2}} l^*(\pi_{i1}, \pi_{i2} \dots \pi_{iR}), ..., \frac{\partial}{\partial \pi_{iR}} l^*(\pi_{i1}, \pi_{i2} \dots \pi_{iR})\Big) = \lambda\Big(1, 1, ..., 1\Big) \tag{2-13}$$

donde ∇ denota el gradiente de la función y se obtiene el siguiente sistema de ecuaciones homogéneo:

$$\begin{cases} \frac{n_{i1}}{\pi_{i1}} - \lambda = 0 \\ \frac{n_{i2}}{\pi_{i2}} - \lambda = 0 \\ \vdots \quad \vdots \quad \vdots \\ \frac{n_{iR}}{\pi_{iR}} - \lambda = 0 \\ \sum_{j=1}^{R} \pi_{ij} - 1 = 0 \end{cases} \tag{2-14}$$

donde:

$$\widehat{\pi_{ij}} = \frac{n_{ij}}{n_i} \text{ para } i = 1, 2, \ldots, S \text{ y } j = 1, 2, \ldots, R$$

se denota por $\widehat{\pi_{ij}} = \dfrac{n_{ij}}{n_i}$ la proporción de la i-ésima subpoblación que pertenece a la j-ésima categoría sobre n_i el total de observaciones de la i-ésima subpoblación y $E[\widehat{\pi_{ij}}] = \pi_{ij}$. Esto es:

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} \widehat{\pi_1} \\ \widehat{\pi_2} \\ \vdots \\ \widehat{\pi_s} \end{pmatrix}$$

2.4. Matriz de varianzas y covarianzas de $\boldsymbol{\pi}$

El análisis de mínimos cuadrados ponderados, requiere la estimación de la matriz de varianzas y covarianzas para $\boldsymbol{\pi}$.
Sea Σ_π la matriz de varianzas y covarianzas para $\boldsymbol{\pi}$ de tamaño $SR \times SR$. Un estimador $\widehat{\Sigma_\pi}$ de Σ_π es:

$$\widehat{\Sigma_\pi} = \begin{bmatrix} \widehat{\Sigma_{\pi_1}} & 0 & \cdots & 0 \\ 0 & \widehat{\Sigma_{\pi_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\Sigma_{\pi_S}} \end{bmatrix}$$

una matriz de bloque diagonal con $\widehat{\Sigma_{\pi_i}}$ en la diagonal principal, donde:

$$\widehat{\Sigma_{\pi_i}} = \begin{bmatrix} \widehat{\pi_{i1}}(1-\widehat{\pi_{i1}}) & -\widehat{\pi_{i1}}\widehat{\pi_{i2}} & \cdots & -\widehat{\pi_{i1}}\widehat{\pi_{iS}} \\ -\widehat{\pi_{i2}}\widehat{\pi_{i1}} & \widehat{\pi_{i2}}(1-\widehat{\pi_{i2}}) & \cdots & -\widehat{\pi_{i2}}\widehat{\pi_{iS}} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\pi_{iS}}\widehat{\pi_{i1}} & -\widehat{\pi_{iS}}\widehat{\pi_{i2}} & \cdots & \widehat{\pi_{iS}}(1-\widehat{\pi_{iS}}) \end{bmatrix}$$

y $\Sigma_{\widehat{\pi_i}}$ es el estimador de Σ_{π_i} matriz de varianzas y covarianzas de la i-ésima subpoblación para $i = 1, 2, \ldots, S$.

2.5. Intervalos de confianza para la proporción de una distribución multinomial

Sea $\mathbf{n} = (n_1, n_2, ..., n_s)^T$ un vector s-dimensional de una distribución multinomial $MN(\boldsymbol{\pi}^T, N)$, donde $N = \sum n_i$ y $\boldsymbol{\pi}^T = (\pi_1, \pi_2, ..., \pi_s)^T$ un vector s-dimensional tal que $\sum \pi_i = 1$ con $i = 1, 2, ..., s$, entonces:

$$\hat{\boldsymbol{\pi}} = \frac{\mathbf{n}}{N} = \left(\frac{n_1}{N}, \frac{n_2}{N}, ..., \frac{n_s}{N}\right) = \left(\hat{\pi_1}, \hat{\pi_2}, ..., \hat{\pi_s}\right) \tag{2-15}$$

En la literatura se tienen varias propuestas de intervalos de confianza simultáneos para los parámetros de una distribución multinomial, los cuales se verán en la siguiente sección.

2.5.1. Intervalos de confianza Quesenberry & Hurts

Quesenberry y Hurts (1964)[20] propusieron el intervalo para π_i dado por:

$$\frac{\chi^2_{k-1,1-\alpha} + 2n_i \pm \sqrt{\chi^2_{k-1,1-\alpha}(\chi^2_{k-1,1-\alpha} + 4\frac{n_i}{N}(N - n_i))}}{2(N + \chi^2_{k-1,1-\alpha})} \tag{2-16}$$

del $(1-\alpha)100\,\%$ de confianza.
Otra propuesta es el intervalo basado en el teorema del límite central (Correa 2020)[6]. Si el tamaño de la muestra es grande, éste está dado por:

$$\left(\widehat{\pi_i} - z_{\alpha/2}\sqrt{\frac{\widehat{\pi_i}(1-\widehat{\pi_i})}{n_i}}, \widehat{\pi_i} + z_{\alpha/2}\sqrt{\frac{\widehat{\pi_i}(1-\widehat{\pi_i})}{n_i}}\right) \tag{2-17}$$

Dada la importancia de la construcción de intervalos de confianza para los parámetros de una distribución multinomial, Gonzáles y Correa (2010),[7] compararon entre los intervalos de confianza más reconocidos, para determinar cuál procedimiento es el mejor, llegando a la conclusión que los dos mejores son: Intervalos de confianza basados en la distribución F e Intervalos de confianza basados en la razón de verosimilitud, los cuales se describen en la siguiente sección.

2.5.2. Intervalos de Confianza basado en la razón de verosimilitud

Kalbfleish(1985) [16] propuso el intervalo de confianza basado en la razón de verosimilitud, descrito a continuación:
Sea $L(\theta)$ la función de verosimilitud del parámetro θ, se define la función de verosimilitud relativa como:

$$R(\theta) = \frac{L(\theta)}{\text{máx}_\theta\, L(\theta)} = \frac{L(\theta)}{L(\widehat{\theta})} \tag{2-18}$$

donde $\widehat{\theta}$ es el valor del parámetro que maximiza $L(\theta)$ y se le llama estimador de máxima verosimilitud de θ, por tanto un intervalo de confianza basado en la razón de verosimilitud con un nivel α, se define como:

$$Intervalo\, Verosimilitud\, = \{\theta : R(\theta) \geq \alpha\} \tag{2-19}$$

donde $0 \leq \alpha \leq 1$. En el caso del vector de parámetros π de una distribución multinomial se obtiene al encontrar un par de raíces tales que:

$$R(\pi_1, \pi_2, \ldots, \pi_k) = \frac{L(\pi_1, \pi_2, \ldots, \pi_k)}{L(\hat{\pi_1}, \hat{\pi_2}, \ldots, \hat{\pi_k})} \geq K(k, \alpha) \tag{2-20}$$

La solución se encuentra con métodos numéricos (Correa 2020) [6] y (González y Correa, 2010)[7]

2.5.3. Método exacto basado en la distribución F

Construir un intervalo de confianza del $(1-\alpha)100\,\%$ para $\boldsymbol{\pi}$, Lewia & Trivedi (1996), muestran dos procedimientos para calcular los límites del intervalo, siendo L_I el límite inferior y L_S el límite superior tal que $P(Y \geq y|\pi = L_I) = \alpha/2$ y $P(Y \leq y|\pi = L_I) = \alpha/2$, donde el intervalo es:

$$\left(\left(1 + \frac{n - n_i + 1}{n_i F_{2n_i, 2(n-n_i+1), 1-\alpha/2k}}\right)^{-1}, \left(1 + \frac{n - n_i + 1}{(n_i+1) F_{2n_i, 2(n-n_i+1), 1-\alpha/2k}}\right)^{-1}\right)$$

2.6. Definición de la función respuesta

La definición de respuestas en la metodología GSK, no está definida por una variable como lo es en los modelos lineales tradicionales, en esta metodología, son funciones que establecen relaciones entre las probabilidades de la "tabla" de probabilidades, para luego modelar la respuesta.
Se define a

$$\boldsymbol{f}^T(\boldsymbol{\pi}) = \left[f_1(\pi), f_2(\pi), ..., f_u(\pi)\right]$$

como una función de los elementos de $\boldsymbol{\pi}$, estos elementos tienen derivadas continuas hasta el segundo orden con respecto a π_{ij} con $i = 1, 2, \ldots, S$ y $j = 1, 2, \ldots, R$, y $\boldsymbol{f}(\boldsymbol{\pi})$ es un vector con u funciones respuesta , $u \leq (S-1)R$.
Sea $\boldsymbol{f}(\hat{\boldsymbol{\pi}})$ el estimador de máxima verosimilitud de $\boldsymbol{f}$ con

$$\left(\boldsymbol{f}(\hat{\boldsymbol{\pi}})\right)^T = \left(f_1(\widehat{\boldsymbol{\pi}}), f_2(\widehat{\boldsymbol{\pi}}), ..., f_u(\widehat{\boldsymbol{\pi}})\right) \text{ (Grizzle et al., 1969)[12]}$$

donde $Var(\widehat{\boldsymbol{f}})$ la matriz de varianzas y covarianzas de $\widehat{\boldsymbol{f}}$ es calculada por el método Delta (Correa 2020)[6].

$$Var(\widehat{\boldsymbol{f}}) \sim \widehat{\Sigma_{\boldsymbol{f}}} = H\widehat{\Sigma_{\hat{\boldsymbol{\pi}}}}H^T \tag{2-21}$$

donde

$$H_{u\times RS} = \left[\frac{\partial \boldsymbol{f}_m(\pi)}{\partial \pi_{ij}}|\pi_{ij} = \hat{\pi}_{ij}\right] \tag{2-22}$$

la cual es una matriz de las derivadas parciales de primer orden de $\boldsymbol{f}$ para $m = 1, ..., u$ con m la m-ésima función $\boldsymbol{f}$ construida y todas las RS combinaciones (i, j). (Agresti, 1996a) [1] (Grizzle et al., 1969)[12] demostraron que se puede encontrar H bajo la aplicación de tres tipos de transformaciones (lineal, logarítmica y exponencial).
En general, funciones de respuesta lineales pueden obtenerse del conjunto básico de probabilidades observadas como:

$$\boldsymbol{f} = A \times \boldsymbol{\pi} \tag{2-23}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1SR} \\ a_{21} & a_{22} & \ldots & a_{2SR} \\ \vdots & \vdots & \vdots & \vdots \\ a_{u1} & a_{u2} & \ldots & a_{uSR} \end{bmatrix}$$

donde :
$\boldsymbol{f}$ es un vector de u componentes
A es una matriz de dimensión $u \times SR$.
$\boldsymbol{\pi}$ es un vector con SR componentes.
Al emplear la Tabla (2-1), se tiene que:

$$\widehat{\boldsymbol{f}} = A\widehat{\boldsymbol{\pi}} \tag{2-24}$$

2.7. Definición del modelo

Muchas veces el interés es modelar respuestas de la forma:

$$\boldsymbol{f(\pi)} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2-25}$$

donde $\boldsymbol{f(\pi)}$ es vector de tamaño $u \times 1$.
X : es una $u \times v$ matriz de diseño de rango v
$\boldsymbol{\beta}$: es un $v \times 1$ vector.
$\boldsymbol{\epsilon}$: es un $u \times 1$ vector aleatorio no observable

2.7.1. Estimación y Validación

La regresión ponderada produce el mejor estimador asintóticamente normal de $\boldsymbol{\beta}$ (McCullang, 1989)[17] dado por:

$$\widehat{\boldsymbol{\beta}} = (X^T\Sigma_{\widehat{\boldsymbol{f}}}^{-1}X)^{-1}(X^T\Sigma_{\widehat{\boldsymbol{f}}}^{-1}\widehat{\boldsymbol{f}}) \tag{2-26}$$

Para muestras grandes, aplicando el teorema central del límite para funciones de $\boldsymbol{f}$, se tiene un estimador consistente para la covarianzas de $\widehat{\boldsymbol{\beta}}$, el cual es:

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) \approx \widehat{\sum}_{\widehat{\boldsymbol{\beta}}} = (X^T\widehat{\sum}_{\widehat{\boldsymbol{f}}}X)^{-1} \tag{2-27}$$

Donde la prueba de bondad y ajuste usa el término residual

$$\widehat{\boldsymbol{f}^T\Sigma_{\widehat{\boldsymbol{f}}}^{-1}}\widehat{\boldsymbol{f}} - \widehat{\boldsymbol{\beta}^T}\big(X^T\widehat{\Sigma_{\widehat{\boldsymbol{f}}}^{-1}}X\big)\widehat{\boldsymbol{\beta}} \tag{2-28}$$

Bajo la hipótesis nula $H_0 : \boldsymbol{f}(\boldsymbol{\pi}) - X\boldsymbol{\beta} = 0$ el estadístico de prueba es asintóticamente χ^2 con $u \times v$ grados de libertad (Agresti) [2].

Para la prueba de hipótesis $H_0 : C\boldsymbol{\beta} = 0$ vs $H_1 : C\boldsymbol{\beta} \neq 0$ se utiliza el estadístico de prueba :

$$W_\lambda = \widehat{\boldsymbol{\beta}}^T C^T\big[C\big(X^T\widehat{\Sigma_{\widehat{\boldsymbol{f}}}^{-1}}X^{-1}\big)C^T\big]^{-1}C\widehat{\boldsymbol{\beta}} \tag{2-29}$$

el cual tiene asintóticamente una distribución χ^2 con d grados de libertad si H_0 es cierta (Grizzle et al., 1969)[12] (Agresti, 1990) [1] (Grizzle et al., 1969).

3 Tamaño de muestra

La determinación del tamaño muestral, es de gran importancia en el análisis de datos categóricos, ya que este juega un papel de importancia, dado que se ve reflejado directamente en el costo de un estudio de investigación, un tamaño de muestra muy grande aumenta la posibilidad que la muestra sea más representativa, pero es un lujo que pocos pueden permitirse, mientras que un tamaño muestral demasiado pequeño privará a un diseño de investigación el poder detectar diferencias significativas en la población cuando realmente existía.
Como consecuencia, para el caso de una población multinomial, muchos autores han estudiado el problema del tamaño muestral y estimación de intervalos de confianza simultáneos para un modelo multinomial y también discuten el tamaño muestral mínimo cuando se realizan pruebas chi-cuadrado en tablas de contingencia y es importante notar que el tamaño de muestra por estrato debe ser grande para obtener resultados de tipo asíntotico.

3.1. Tamaños de muestra de una población multinomial

La determinación del tamaño de muestra necesario para la estimación simultánea de proporciones multinomiales es equivalente a la construcción de intervalos de confianzas simultáneos para variables con distribución multinomial, donde la diferencia radica en que para el cálculo del tamaño de muestra los límites del intervalo de confianza son fijados a priori por el investigador, con el objetivo de controlar la probabilidad que el verdadero valor del parámetro esté en el intervalo. Como consecuencia, muchos autores han estudiado el problema del tamaño muestral y estimación de intervalos de confianza simultáneos para un modelo multinomial, por ello en la literatura se encuentran diferentes técnicas para el calculo del tamaño mínimo muestral, algunas de ellas descritas a continuación:

Goodman (1965) [10] propuso una metodología basada en la aproximación de la distribución binomial a la distribución normal, utilizando la desigualdad de Bonferroni y el teorema del límite central, obteniendo así intervalos de confianza simultáneos, dados por:

$$L_i = \pi_i - \sqrt{\frac{B\pi_i(1-\pi_i)}{n}} \quad \text{y} \quad L_s = \pi_i + \sqrt{\frac{B\pi_i(1-\pi_i)}{n}} \quad \text{con} \quad i = 1, 2..., k \qquad \text{(3-1)}$$

donde L_i y L_s son límite inferior y límite superior respectivamente, siendo B el percentil superior $\left(\frac{\alpha}{k} \times 100\right)$ de una distribución chi-cuadrado con un grado de libertad ($B \sim \chi^2_1$).

Yarnold(1970)[25] propuso un criterio para el cálculo del tamaño mínimo n en una celda tal que se garantice $n\pi_i \geq 5q$ para todo $i = 1, 2, ..., k$, donde q es la proporción de la celda en la cual $n\pi_i < 5$.

Eaton (1978) [8] propuso un algoritmo donde se garantiza el criterio de Yarnold, el algoritmo es:
Sea un vector π_h de probabilidades, se realiza lo siguiente:

- Ordenar las probabilidades de π_h de mayor a menor.

$$\pi_k \geq ... \geq \pi_i \geq ... \geq \pi_1$$

- Encontrar m, el máximo i con $i = 1, 2, ..., k$ tal que $\pi_{(i)} \leq k\pi_1$
- $k\frac{\pi_i}{0} = +\infty$ y $c_i = \frac{k\pi_1}{i-1}$
- Se compara π_i con c_i hasta que se cumpla que $\pi_i \leq c_i$, y sea s el primer i que satisface la condición.
- $n = \begin{cases} \frac{5}{c_{s+1}} & \text{si } \pi_s \leq \pi_{(s+1)}, c_{(s+1)} \text{y } c_{(s)} \\ \frac{5}{\pi_{(s)}} & \text{en otros casos} \end{cases}$

Tortora (1978) [24] determina el tamaño muestral requerido de una población multinomial, considerando una muestra aleatoria simple de una población en la que cada variable se clasifica de una de las k categorías mutuamente excluyente y exhaustiva, este método se basa en crear intervalos de confianza simultáneos debido a Goodman (1965)[10] y requiere la precisión absoluta b_i para cada celda de una población multinomial, para un valor específico de α, se desea obtener un conjunto de intervalos S_i, $i = 1, ..., k$, tal que

$$Pr\left\{\bigcap_{i=1}^{k}(\pi_i \in S_i)\right\} \geq 1 - \alpha \tag{3-2}$$

donde π_i, $i = 1, .., k$ denota la proporción de la población en la i-ésima categoría en una muestra aleatoria simple de tamaño n de la población, donde L_i límite inferior y L_s límite superior del intervalo de confianza para π_i, los cuales son dados por:

$$L_i = \pi_i - \sqrt{B\pi_i(1-\pi_i)/n} \tag{3-3}$$

$$L_s = \pi_i + \sqrt{B\pi_i(1-\pi_i)/n} \tag{3-4}$$

la desviación estándar para la i-ésima celda de la población multinomial es

$$\sigma_i = [\pi_i(1-\pi_i)/n]^{1/2} \tag{3-5}$$

para el cálculo del tamaño de muestra, se requiere una precisión absoluta de b_i para cada celda y a partir de las ecuaciones (3-3) y (3-4), se tiene que:

$$\pi_i - b_i = \pi_i - \sqrt{B\pi_i(1-\pi_i)/n} \tag{3-6}$$

$$\pi_i + b_i = \pi_i + \sqrt{B\pi_i(1-\pi_i)/n} \tag{3-7}$$

despejando b_i de la ecuación (3-6), se obtiene:

$$b_i = \sqrt{B\pi_i(1-\pi_i)/n} \tag{3-8}$$

y de (3-8) se tiene que:

$$n_i = \frac{B\pi_i(1-\pi_i)}{b_i^2} \tag{3-9}$$

siendo B el $\frac{\alpha}{k}$-ésimo percentil superior de una χ^2 con un grado de libertad y $n = max\{n_i\}$ con $i = 1, 2, ..., k$.

Bromaghin (1993)[4] hace una revisión a la propuesta de Tortora [24], la cual está basada en la construcción k $(k > 2)$ intervalos de confianza simultáneos de Goodman [10] y propone determinar el tamaño de muestra mínimo n, dado por:

$$n_i = \frac{z^2_{\alpha_i/2}}{2d_i^2}\big(\pi_i(1-\pi_i) - 2d_i^2 + \sqrt{\pi_i^2\big(1-\pi_i\big)^2 - 2d_i^2\big[4\pi_i\big(1-\pi_i\big)\big]}\big) \tag{3-10}$$

donde d_i es el ancho deseado para la i-ésima probabilidad, α_i es el nivel de confianza de cada intervalo y n se elige de manera tal que $n = min\{n_i, i = 1, 2, \ldots, k\}$

Thompson (1987) [23] propone un método para el cálculo del tamaño de muestra, el objetivo de éste es seleccionar el n más pequeño para una muestra aleatoria de una población

multinomial, basado en la distribución marginal binomial. Para cada parámetro se elige un intervalo con un nivel α_i especificado con $i = 1, 2, \ldots, k$, tal que

$$\alpha_i = P\left(|z_i| > z_{1-\frac{\alpha_i}{2}} \sqrt{\frac{\pi_i^0(1-\pi_i^0)}{n_i}} \right) \tag{3-11}$$

donde $d_i = z_{1-\frac{\alpha_i}{2}} \sqrt{\frac{\pi_i^0(1-\pi_i^0)}{n}}$ es la longitud media del intervalo, k el número de categorías y π_i proporción en la i-ésima categoría, el propósito es encontrar un n tal que

$$\sum_{i=1}^{k} \alpha_i \leq \alpha \tag{3-12}$$

dadas las longitudes $2d_i$ especificadas. y

$$n = \max_i n_i \tag{3-13}$$

Cochran (1977) [5] presenta una propuesta para la estimación de un intervalo de confianza para la proporción de la i-ésima categoría de una distribución multinomial, con $i = 1, 2, ..., k$, dado por:

$$\hat{\pi}_i - error_i \leq \pi_i \leq \hat{\pi}_i + error_i \tag{3-14}$$

donde

$$error_i = z_{(1-\alpha)} \sqrt{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}} \tag{3-15}$$

y despejando n_i de la ecuación (3-15), se tiene que:

$$n_i = z^2_{(1-\alpha)} \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{error_i} \tag{3-16}$$

y

$$n = \Sigma_{i=1}^{k} n_i \tag{3-17}$$

con $error_i$ es es error absoluto requerido y $z_{(1-\alpha)}$ el percentil de $(1-\alpha)$ de una distribución normal estándar.

3.1.1. Tamaño de muestra utilizando la metodología GSK

Rochon (1989) [21] considera una tabla análoga a la (**Tabla 2-1**) con S subpoblaciones independientes que poseen una distribución multinomial, y utiliza los procedimientos de la metodología GSK descrita en el capítulo anterior para calcular el tamaño mínimo requerido de la subpoblación al realizar una prueba de los efectos específicos de los parámetros del modelo. Primero específica el vector $\boldsymbol{\pi}$ y sustituye $\widehat{\boldsymbol{\pi}}$ por $\boldsymbol{\pi}$, así construye una función respuesta $\boldsymbol{f}(\widehat{\boldsymbol{\pi}})$ (una función sobre valores de la tabla), y estima la matriz de varianzas y covarianzas de $\widehat{\Sigma_{\widehat{f}}}$ de $\widehat{\boldsymbol{f}(\boldsymbol{\pi})}$, plantea el modelo requerido $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}$, donde X matriz de diseño y donde la estimación del vector $\boldsymbol{\beta}$ dada por la metodología GSK, es:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\widehat{\Sigma_{\widehat{\boldsymbol{f}}}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\Sigma_{\widehat{\boldsymbol{f}}}}^{-1}\widehat{\boldsymbol{f}} \tag{3-18}$$

Por consiguiente, para evaluar los efectos específicos de $\widehat{\boldsymbol{\beta}}$, construye una matriz C de rango completo c y define un vector h, y contrasta:

$$H_0 : C\beta = h \text{ vs } H_1 : C\beta \neq h \tag{3-19}$$

donde el estadístico de prueba es:

$$Q_H = (C\beta - h)^T(C(X^T\widehat{\sum_f}^{-1}X)C^T)^{-1}(C\beta - h) \sim \chi^2_c \tag{3-20}$$

con grados de libertad c igual al rango de la matriz C y el parámetro de no centralidad λ

$$\lambda = n(C\widehat{\beta} - h)^T(C(X^T\widehat{\Sigma_{\widehat{\boldsymbol{f}}}}^{-1}X)C^T)^{-1}(C\widehat{\beta} - h) \tag{3-21}$$

$$n = \frac{(C\beta - h)^T(C(X^THRH^TX)C^T)^{-1}(C\beta - h)}{\lambda} \tag{3-22}$$

Siendo $\lambda = (z_{1-\alpha/2} + z_{1-\delta})^2$, con α error de tipo I y $1-\delta$ la potencia establecidos previamente, los cuales están asociados al estadístico de prueba Q_H.

3.2. Algoritmos genéticos

En 1970 Holland [14] introdujo los algoritmos genéticos (**AG**) y no son más que abstracciones de la teoría de la evolución y genética, los cuales son usados en la estadística como herramientas de optimización estocástica para modelos y son capaces de dar soluciones óptimas

para funciones de varias variables sin los requisitos matemáticos como continuidad estricta, diferenciabilidad, convexidad y otras condiciones (Haupt & Haupt 1998) [13] .
Los **AG** son una línea de la inteligencia artificial más promisoria, estos son llamados así porque se inspiran en la evolución biológica y propuestos para optimizar funciones complejas (Haupt and Haupt, 1998).[13]

3.2.1. Ventajas

La naturaleza de los algoritmos genéticos reune ventajas frente a otros tipos de algoritmos de búsqueda, lo cual los convierte en eficientes y eficaces.
Alguna de estas características son mencionadas por Moujahid et al (2008) [18] y Gil (2006) [19].

- No están sujetos a restricciones del problema a solucionar.
- Los conceptos de codificación son sencillos.
- Utilizan operadores probabilísticos para la selección del "mejor" individuo, en vez de operadores determinísticos.
- Es ideal para la búsqueda de solución de problema de grandes dimensiones.
- No requiere conocimientos a priori sobre el problema a resolver.
- Son intrínsecamente paralelos, esto es, manipulan simultáneamente varios parámetros obteniendo así varias soluciones buenas para el problema a resolver.
- Los algoritmos genéticos resultan con un mejor desempeño, en comparación con otros métodos, esto es, en los problemas que incluyen una función objetivo discontinua, ruidosa, cambiante en el tiempo o que posee varios óptimos locales presentan un mejor desempeño.

3.2.2. Desventajas

Aunque los **AG** muestran una eficiencia para la solución de problemas dada su robustez, estos también presentan ciertas falencias, Gil (2006) [19] enuncia una serie de desventajas y a su vez recursos para abordarlas.

- Definir una representación del problema. El lenguaje utilizado para especificar soluciones candidatas debe ser robusto, debe ser capaz de tolerar cambios aleatorios que no produzcan constantemente errores fatales o resultados sin sentido. Se puede solucionar mediante la definición de los individuos como listas de números donde cada número representa algún aspecto de la solución candidata.

- Pueden tardar mucho en converger, o no converger en absoluto, dependiendo en cierta medida de los parámetros que se utilicen como: tamaño de la población, número de generaciones, el ritmo de mutación y cruzamiento, el tipo y fuerza de la selección
- Un problema muy común que puede surgir se conoce como la convergencia prematura. Si un individuo que es más apto que la mayoría de sus competidores emerge muy pronto en el curso de la ejecución, se puede reproducir tan abundantemente que merme la diversidad de la población demasiado pronto, provocando que el algoritmo converja hacia el óptimo local que representa ese individuo, en lugar de rastrear el paisaje adaptativo lo bastante a fondo para encontrar el óptimo global. Esto es un problema especialmente común en las poblaciones pequeñas, donde incluso una variación aleatoria en el ritmo de reproducción puede provocar que un genotipo se haga dominante sobre los otros.

3.2.3. Algoritmo genético binario

Esquema de funcionamiento

Los algoritmos genéticos **AG** son algoritmos de búsquedas que imitan los mecanismos de selección natural y genética natural (Goldberg 1975) [9]. Las especies buscan la mejor adaptación al medio, donde el mejor individuo, es aquel que logra adaptarse, sobrevivir y reproducirse, transmitiendo así su genética a la siguiente generación.

Componentes algoritmo genético binario

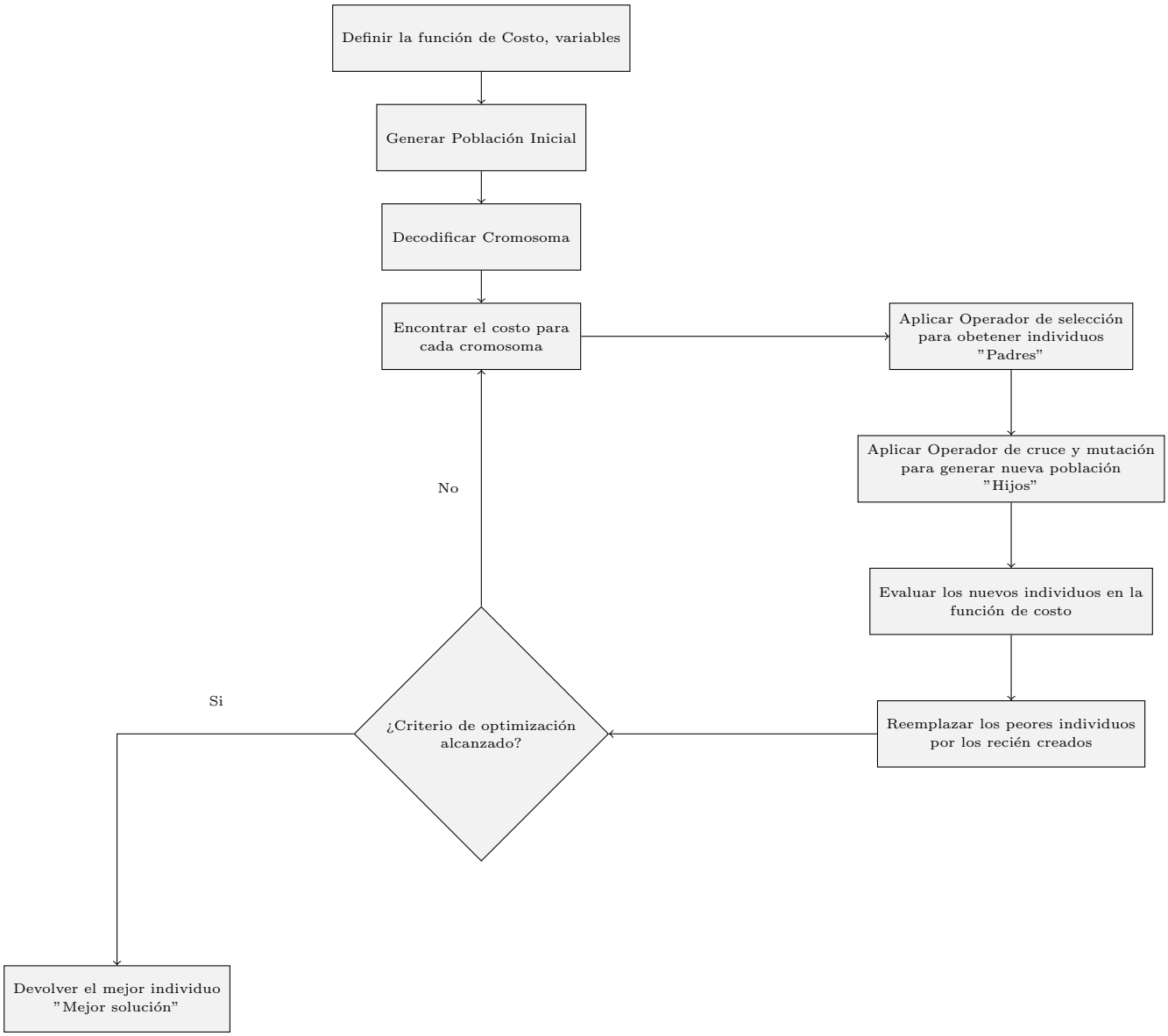


Figura (3-1) Componentes algoritmo genético binario

Las componentes de un algoritmo genético son:

1. **Cromosoma**: Se refiere a un candidato a solución del problema. Los genes codifican un elemento particular del candidato a solución (por ejemplo, en el caso de la optimización de una función multiparamétrica, un parámetro particular, se considera un gen).

2. **Función Fitness o Función Costo**: El algoritmo requiere una función de capacidad o potencial que asigna una puntuación (la capacidad) a cada cromosoma de la población actual. La capacidad o el potencial de un cromosoma depende de cómo resuelva ese cromosoma el problema a tratar.

3. **Población**: Conjunto de cromosomas conocidos.

4. **Selección natural**

   Los **AG** siguen el mismo modelo genético de recombinación y selección natural, estos representan variables como una cadena binaria codificada o con variables continuas para minimizar el costo.
   El **AG** inicia con una población compuesta por individuos y cada individuo está formado por una serie de variables las cuales se definen como un cromosoma o una formación de la variable para ser optimizada. El cromosoma es evaluado en la función de costo, donde la función de costo debe ser diseñada para cada problema de manera específica y ésta asigna a un cromosoma particular un valor real, el cual refleja el nivel de adaptación al problema a optimizar representado por el cromosoma. Si el cromosoma tiene N_{var} variables entonces es escrito como:

$$cromosoma = [p_1, p_2, ..., p_{N_{var}}]$$

   donde el

$$costo = f(cromosoma) = f[p_1, p_2, ..., p_{N_{var}}], \text{ tal que } f := \text{función costo}$$

   La fórmula matemática para codificar y decodificar las n-variables p_n son: (Haupt & Haupt, 1998) [13]
   Para codificar:

$$p_{norm} = \frac{p_n - p_{lo}}{p_{hi} - p_{lo}}$$

$$gene[m] = round\{p_{norm} - 2^{-m} - \sum_{p=1}^{m-1} gene[p]2^{-p}\}$$

   Para decodificar:

$$p_{quant} = \sum_{m=1}^{N_{gene}} gene[m]2^{-m} + 2^{-(M+1)} \; q_n = p_{quant}(p_{hi} - p_{lo}) + p_{lo}$$

donde:
p_{norm}: Variable normalizada, $0 \leq p_{norm} \leq 1$
p_{hi}: Valor máximo de la variable
p_{lo}: Valor mínimo de la variable
N_{gene}: Número de bits de un gen
$gene[m]$: Versión binaria de p_n
$round\{.\}$: Redondea al entero más cercano
p_{quant}: Versión cuantificada de p_{norm}
q_n: Versión cuantificada de p_n donde $q_n = gene \times Q^T$, tal que $Q^T = (2^{-1}, 2^{-1}, ..., 2^{N_{gene}})$ y $gene = (b_1, b_2, ..., b_{N_{gene}})$ con b_i =1 o 0 , $i = 1, 2, ...N_{gene}$ donde $N_{bits} = N_{gene} \times N_{var}$, con N_{bits}: número de bits que posee un cromosoma.

5. **La población**

   Los **AG** inician con una población de N cromosomas, la población tiene N_{pop} cromosomas, donde se forma una matriz de tamaño $N_{pop} \times N_{bits}$ de unos y ceros, donde los cromosomas son las columnas de dicha matriz.

6. **Selección natural**

   La selección del mejor cromosoma de la población, se da utilizando la función costo, esto es, se descarta el cromosoma con el costo más alto. La selección natural ocurre cada generación o cada iteración del algoritmo. Del cromosoma N_{pop} en una generación, solo el N_{keep} sobreviven para el apareamiento y el $N_{pop} - N_{keep}$ se descartan, para esto, primero se ordena los costos de cada cromosoma del más bajo al más alto, de ahí se selecciona los mejores o los que tiene un costo más bajo que algún umbral dado.

$$N_{keep} = X_{rate}.N_{pop} \text{ donde } X_{rate} = s, \text{ con } 0 \leq s \leq 1 \tag{3-23}$$

7. **Selección** La selección de los cromosomas puede ser hecha usando varios métodos
   - *a*) **Emparejamiento de arriba a abajo.** Combina filas impares con filas pares de la matriz $N_{pop} \times N_{bits}$.
   - *b*) **Emparejamiento Aleatorio.** Este emparejamiento utiliza un generador de números aleatorios uniforme para seleccionar un cromosoma.
   - *c*) **Emparejamiento aleatorio ponderado.** Las probabilidades asignadas a los cromosomas en el grupo de apareamiento son inversamente proporcionales a su costo. El cromosoma con el costo más bajo tiene la mayor probabilidad de apareamiento, mientras que el cromosoma con el costo más alto tiene la menor probabilidad de apareamiento. Un número aleatorio determina qué cromosoma se selecciona. Este tipo de ponderación a menudo se denomina ponderación de ruleta. Hay dos técnicas: ponderación de rango y ponderación de costos:

1) Ponderación de rango: Este enfoque es independiente del problema y encuentra la probabilidad del rango n, del cromosoma:

$$P_n = \frac{N_{keep} - n + 1}{\sum_{n=1}^{N_{keep}} n} \tag{3-24}$$

2) Ponderación de costos: La probabilidad de selección se calcula a partir del costo del cromosoma en lugar de su rango en la población. Un costo normalizado se calcula para cada cromosoma restando el más bajo costo de los cromosomas descartados ($c_{N_{keep+1}}$) del costo de todos los cromosomas en el grupo de apareamiento, esto es,

$$C_n = c_n - c_{N_{keep+1}} \tag{3-25}$$

tal que

$$P_n = \left| \frac{C_n}{\sum_{m}^{N_{keep+1}} C_m} \right| \tag{3-26}$$

d) **Selección de torneo**: Consiste en elegir aleatoriamente un pequeño subconjunto de cromosomas (dos o tres) del grupo de apareamiento, y el cromosoma con el costo más bajo en este subconjunto se convierte en uno de los padres.

8. **Apareamiento**
El apareamiento es la creación de uno o más descendientes de los padres seleccionados. La forma más común de apareamiento involucra a dos padres que producen dos crías, para esto, se selecciona un punto crossover al azar de los cromosomas de los padres.
Primero, el padre 1 pasa su código binario a la izquierda de ese punto de cruce a la descendencia 1. Análogamente el padre 2 pasa su código binario a la izquierda del punto de cruce a la descendencia 2 y a continuación, el código binario a la derecha del punto de cruce del padre 1 va a la descendencia 2 y el padre 2 pasa su código binario a la derecha del punto de cruce a la descendencia 1. Por tanto las descendencias contienen códigos binarios de ambos padres. Aquí se ha producido un total de $N_{pop} - N_{keep}$ descendientes.

9. **Mutación**
La mutación altera cierto porcentaje de los bits en la lista de cromosomas, esto puede introducir rasgos que no están en la población original y evita que el **AG** converja rápidamente sin muestrear toda la superficie de costo. Los puntos de mutación se seleccionan al azar de la matriz de $N_{pop} \times N_{bits}$ y la cantidad de mutaciones es elegida por:

$$\#mutaciones = \mu \times (N_{pop} - 1) \times N_{bits} \tag{3-27}$$

donde μ = tasa de mutación.

10. **Nueva Generación**
    Después que se producen las mutaciones, los costos asociados a cada descendencia y cromosomas mutados se calculan y se eliminan los cromosomas con mayores costos, obteniendo así una nueva poblacioón con $N_{pop} \times N_{bits}$.

3.3. Determinación de tamaño de muestra vía Simulación

En este trabajo se propone el uso de simulación para calcular el tamaño de muestra por cada estrato de un parámetro del modelo GSK.

Algoritmo vía simulación

1. Establecer las probabilidades de cada subpoblación $\pi_{i1}, ..., \pi_{iR}$ con $i = 1, 2, \ldots, S$.
   a) Fijar los $n_1, n_2, n_3, ..., n_S$, siendo S = Número de filas de la tabla.
      1) Generar una muestra simulada para los i- estratos (i-subpoblación) con $i = 1, 2, ..., S$ de una distribución multinomial $MN(n_i, \pi_{i1}, ..., \pi_{iR})$.
      2) Se estima el parámetro de interés θ del modelo propuesto $\hat{\theta}$.
      3) Se repite los dos pasos anteriores, el número de repeticiones es igual al número de simulaciones establecidas.
      4) Se construye una región de confianza del $(1 - \alpha)100\,\%$ para θ.
   b) Se repite el proceso con varios valores para los $n_i's$. Estos diferentes valores son controlados por un algoritmo genético, esto es:
      1) Se genera una población inicial de cromosomas de tamaño n_i correspondiente al i-estrato de la Tabla con $i = 1, 2..., S$.
      2) Para el i-estrato de la Tabla, se evalúa la función de costo para cada cromosoma de la población inicial de tamaño n_i con $i = 1, 2..., S$.
      3) Se selecciona dos individuos (cromosomas) de cada población existente, que proporcionan el mejor costo.
      4) Se cruzan los dos individuos seleccionados en cada población, y se generan descendientes.
      5) Se aplica un proceso de mutación sobre el individuo nuevo de cada población.
      6) Se añade el nuevo individuo en cada población respectivamente.
      7) Se reemplaza cada población antigua (población inicial) por cada una de las nueva poblaciones.

2. Se escogen los $n_i's$ que satisfagan el criterio establecido,

$$Criterio = ((\boldsymbol{L}_i)_{(S\times 1)} - (\boldsymbol{L}\boldsymbol{*}_i)_{(S\times 1)})^2 \tag{3-28}$$

donde $(\boldsymbol{L}_i)_{(S\times 1)}$ y $(\boldsymbol{L}\boldsymbol{*}_i)_{(S\times 1)}$ vectores tal que sus componentes son las longitudes medias de los intervalos de confianza para las componentes de θ_{obs} y $\widehat{\theta}$ respectivamente, con θ_{obs}: valor del parámetro observado.

3.3.1. Esquema del Algoritmo

Para estimar los n_i, tamaños de muestra de cada subpoblación de la tabla con $i = 1, 2, ..., S$, se siguen las siguientes estrategias:

- Se implementa la metodología de los algoritmos genéticos para estimar los n_i de cada subpoblación con $i = 1, 2, ...S$.
- Se calcula la función de interés (dada por el investigador).
- Se aplica la metodología **GSK** para la estimación de la función de interés.

El código del algoritmo, creado utilizando el software estadístico R se encuentra en el (**Anexo**) y el diagrama de flujo, el cual esboza el funcionamiento y facilita comprensión de éste puede ser observado en las Figuras 3-2 y 3-3.

A continuación se ilustra el funcionamiento del algoritmo.
Para dar inicio al algoritmo, se requiere establecer un mínimo de argumentos, estos son:

1.
   - $N_{var} = R$: Número de variables (número de categorías).
   - $Nbits$: Número de bits del gen.
   - $\boldsymbol{N} = \left(n_1^*, n_2^*, \ldots n_S^*\right)$, n_i: Tamaño establecido de la población inicial de cromosomas de la i-ésima subpoblación de la tabla.
   - **mínimos**=$(a_1, a_2, \ldots, a_{N_{var}})$: Vector con valores mínimos para cada parámetro.
   - **máximos**=$(b_1, b_2, \ldots, b_{N_{var}})$: Vector con valores máximos para cada parámetro.
   - $n - generaciones$: Número de generaciones
   - $Costo()$: Función de costo
   - $a.optimizar()$: Criterio de optimización
   - $Nsim$: Número de simulaciones
   - $\boldsymbol{errores} = (d_1, d_2, ..., d_{N_{var}})$: Vector de longitudes establecidas
   - $\boldsymbol{P}_{S\times R}$: Matriz con probabilidades establecidas para las S subpoblaciones
   - $\boldsymbol{Pobs}_{R\times S}$: Matriz con probabilidades observadas paras las S subpoblaciones.

- $\boldsymbol{A}$: Matriz para el cálculo de la función respuesta.
- $\boldsymbol{K}$: Matriz para el cálculo de la función respuesta.
- $\boldsymbol{X}$: Matriz de diseño del modelo.

2. El código inicia leyendo un vector numérico $\boldsymbol{N} = (n_1^*, n_2^*, n_3^*, \ldots, n_S^*)^T$ cuyas componentes son los tamaños de las S subpoblaciones, el tamaño de este vector será S número de filas de la tabla.
   Para la i-ésima subpoblación de la Tabla, n_i^* es el tamaño de la población inicial de cromosomas, donde cada cromosoma representa una posible solución al problema.

$$cromosoma_k = \big(gen_1, gen_2, \ldots, gen_{Nvar}\big) \text{ con } k = 1, 2, ...n_i^* \tag{3-29}$$

   y cada gen_j es una representación binaria con $Nbits$, $j = 1, 2, ..., N_{var}$. Puesto que cada cromosoma de la población inicial está escrito en código binario, se aplica la función $parametros()$ la cual convierte cada cromosoma de la población (decodifica valores binarios) en un vector con componentes discretas.

$$parametros(cromosoma_k) = \big((n_{i1})_k, (n_{i2})_k, ..., (n_{iN_{var}})_k\big), k = 1, 2, ..., n_i^*. \tag{3-30}$$

   donde $a_j \leq n_{ij} \leq b_j$, con a_j y b_j la j-ésima componente de los vectores **mínimos** y **máximos** respectivamente, con $j = 1, 2, ..., N_{var}$.
   De ahí, para la i-ésima subpoblación de la tabla se tiene una población inicial con n_i^* vectores, donde cada componente n_{ij} representa el posible número de elementos de la celda, en la i-ésima subpoblación y j-ésima categoría, con $j = 1, 2, ..., Nvar$, $i = 1, 2..., S$.

3. Otro de los atributos requerido por el algoritmo es la función de costo, esta función es la fortaleza (fitness o costo) de cada individuo (cromosoma). El costo está relacionado con el valor de la función de costo para cada individuo, en este trabajo se requiere el costo mínimo, esto es, para la i-ésima subpoblación de la tabla con $i = 1, 2, ..., S$, utilizando método de Cochran descrito en la **Sección 3-1**, se calcula un intervalo de confianza del $(1 - \alpha)\,\%$ para cada parámetro $\pi_{i1}, \pi_{i2}, ...\pi_{iR}$, dado $\boldsymbol{errores} = \boldsymbol{d_i} = (d_{i1}, d_{i2}, ..., d_{iR})$ vector de longitudes medias de los R intervalos de confianza para la i-ésima subpoblación.
   Por tanto el costo de un cromosoma es:

$$costo(cromosoma_k) = costo\big((n_{i1}, n_{i2}, ..., n_{iN_{var}})_k\big) = (\delta, \delta, \delta, ..., \delta, volumen, n_i) \tag{3-31}$$

con $k = 1, 2..., n_i^*$

$$\delta = \begin{cases} 0, & si \quad l_{ij} \leq d_{ij} \\ 1, & si \quad l_{ij} > d_{ij} \end{cases} \tag{3-32}$$

$$volumen = \prod_{j=1}^{R} l_{ij} \quad y \quad n_i = \sum_{j=1}^{R} n_{ij} \quad con \quad i = 1, 2, ..., S$$

con l_{ij} = longitud media del intervalo de confianza para π_{ij}

$$l_{ij} = \sqrt{\frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}}} \tag{3-33}$$

con $j = 1, 2, ..., R \ y \ i = 1, 2, 3, \ldots, S$.

En este punto, en cada una de las $n - generaciones$, para la i-ésima subpoblación, con $i = 1, 2, \ldots S$ se genera una población de cromosomas de tamaño n_i^*, es decir, un conjunto de n_i^* posibles soluciones. Cada solución es evaluada en la función de costo, con el fin de determinar la solución que posea el costo mínimo, esto es: $costo_{minimo} = (0, 0, 0, ..., 0, volumen_{min}, (n_i)_{min})$, con $volumen_{min}$: volumen mínimo y $(n_i)_{min}$: tamaño de muestra mínimo para la i-ésima subpoblación.

Esta función costo fue creada con el fin de determinar los mejores n_{ij} de la i-ésima subpoblación con $j = 1, 2, ..., R$ y así poder aplicar el operador de selección, cruce y mutación para generar una nueva población para la siguiente generación, reemplazando los peores individuos (cromosomas) por los nuevos creados, así hasta satisfacer con el número de generaciones definidas, obteniendo la última población considerada la mejor.
La función ***nis.subpoblacion()*** arroja un vector con los n_i, $i = 1, 2, ..., S$ tamaño muestras para cada subpoblación.

4. Al obtener vía algoritmo genético las estimaciones de los n_i , $i = 1, 2, \ldots, S$ tamaños de muestra de cada estrato o subpoblación, se genera mediante simulación muestras de una distribución multinomial $MN(n_i, \pi_{i1}, \pi_{i2}, ..., \pi_{iR})$, donde los $\pi_{i1}, \pi_{i2}, ..., \pi_{iR}$ son los valores establecidos para la subpoblación i.

5. Se estima la función de interés $\boldsymbol{f}$ de la tabla de probabilidades estimadas, la cual será obtenida mediante la función $\boldsymbol{f-est}()$ y se calcula la función de interés $\boldsymbol{f_{obs}}$ para la probabilidades observadas $\boldsymbol{Pobs}_{S\times R}$, mediante la función $\boldsymbol{f-obs}()$.

6. Utilizando la metodología **GSK**, la cual relaciona las funciones de interés con un conjunto de covariables para cada subpoblación

$$\boldsymbol{f(\pi)} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3-34}$$

donde $\boldsymbol{X}$ es la matriz de diseño del modelo, $\boldsymbol{\beta}$ vector de parámetros desconocidos y $\boldsymbol{\epsilon}$ vector de errores, se estima $\widehat{\boldsymbol{f*}}$ mediante la función $\boldsymbol{f-gorro}()$.

7. Se calcula un intervalo de confianza para cada i-ésima componente de $\widehat{\boldsymbol{f*}}$ y $\boldsymbol{f_{obs}}$ de $(1-\alpha)100\,\%$ (descrito en el siguiente capítulo), dado el $n_i's$ obtenido vía algoritmo genéticos, con $i = 1, 2, \ldots, S$.

8. El criterio de optimización utilizado en este trabajo, se resume así: La función $a.optimizar()$, se encarga de medir que tan distante se encuentran las longitudes medias de los intervalos de confianza para cada parámetro estimado y el verdadero valor, a través de la diferencia al cuadrado, esto es, sean $(\boldsymbol{L}_i)_{(S\times 1)}$ y $(\boldsymbol{L*}_i)_{(S\times 1)}$ vectores donde sus componentes son las longitudes medias de los intervalos de confianza para las componentes de $\boldsymbol{f_{obs}}$ y $\widehat{\boldsymbol{f*}}$ respectivamente, por tanto:

$$\boldsymbol{a.optimizar}\boldsymbol{((L_i)_{(S\times 1)}, L*_i)_{(S\times 1)})} = ((\boldsymbol{L}_i)_{(S\times 1)} - (\boldsymbol{L*}_i)_{(S\times 1)})^2 \tag{3-35}$$

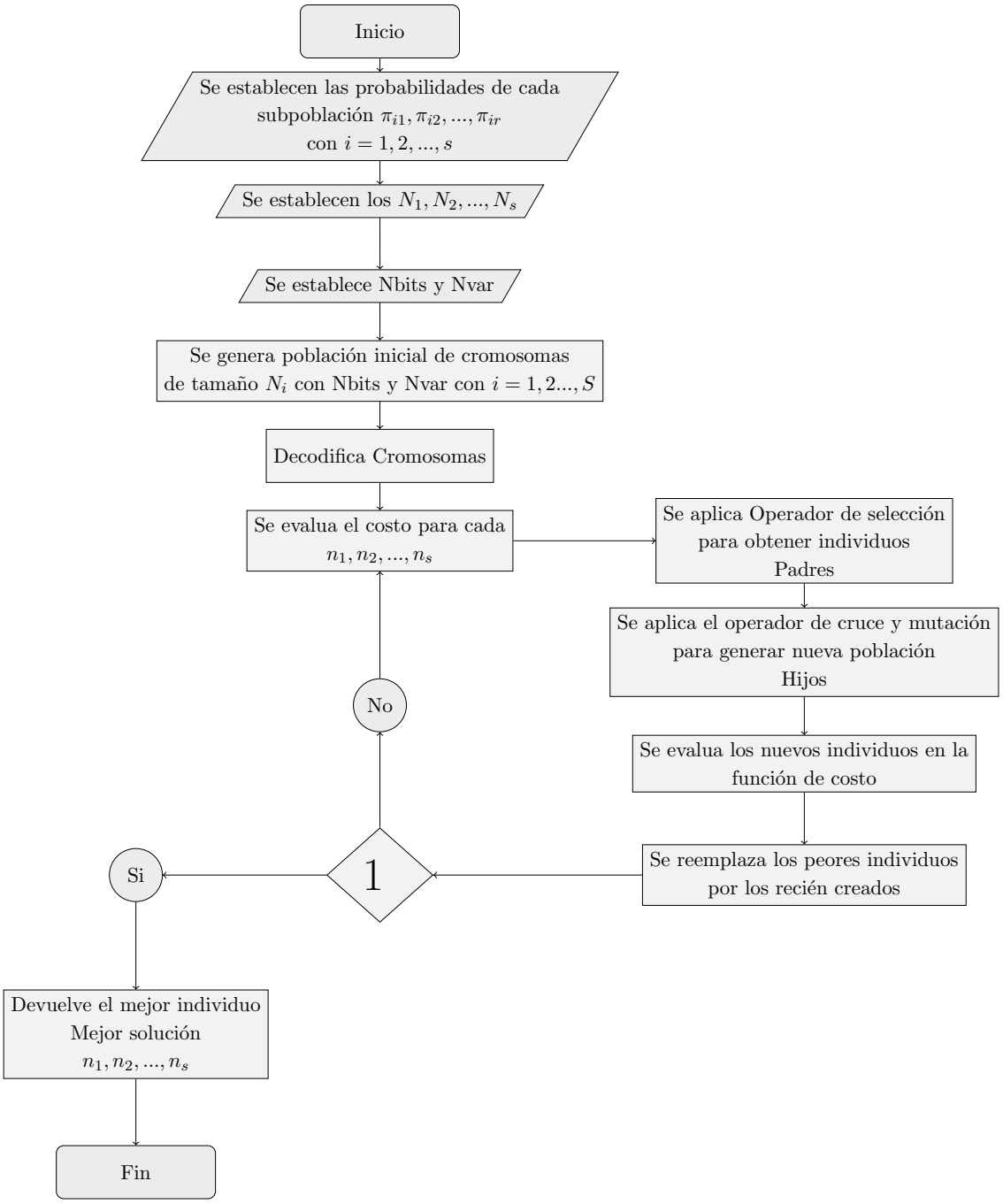
donde la función $\boldsymbol{a.optimizar()}$ debe arrojar un resultado ideal, un vector nulo o con componentes cercanas a cero, cada elemento de $\boldsymbol{a.optimizar}$ corresponde al error de estimación, se toma como valores aceptables los errores menores o iguales a 0.05. El resultado es calculado mediante la función $\boldsymbol{a.optimizar}()$.

9. Si $\boldsymbol{a.optimizar}()$ satisface el criterio establecido por:

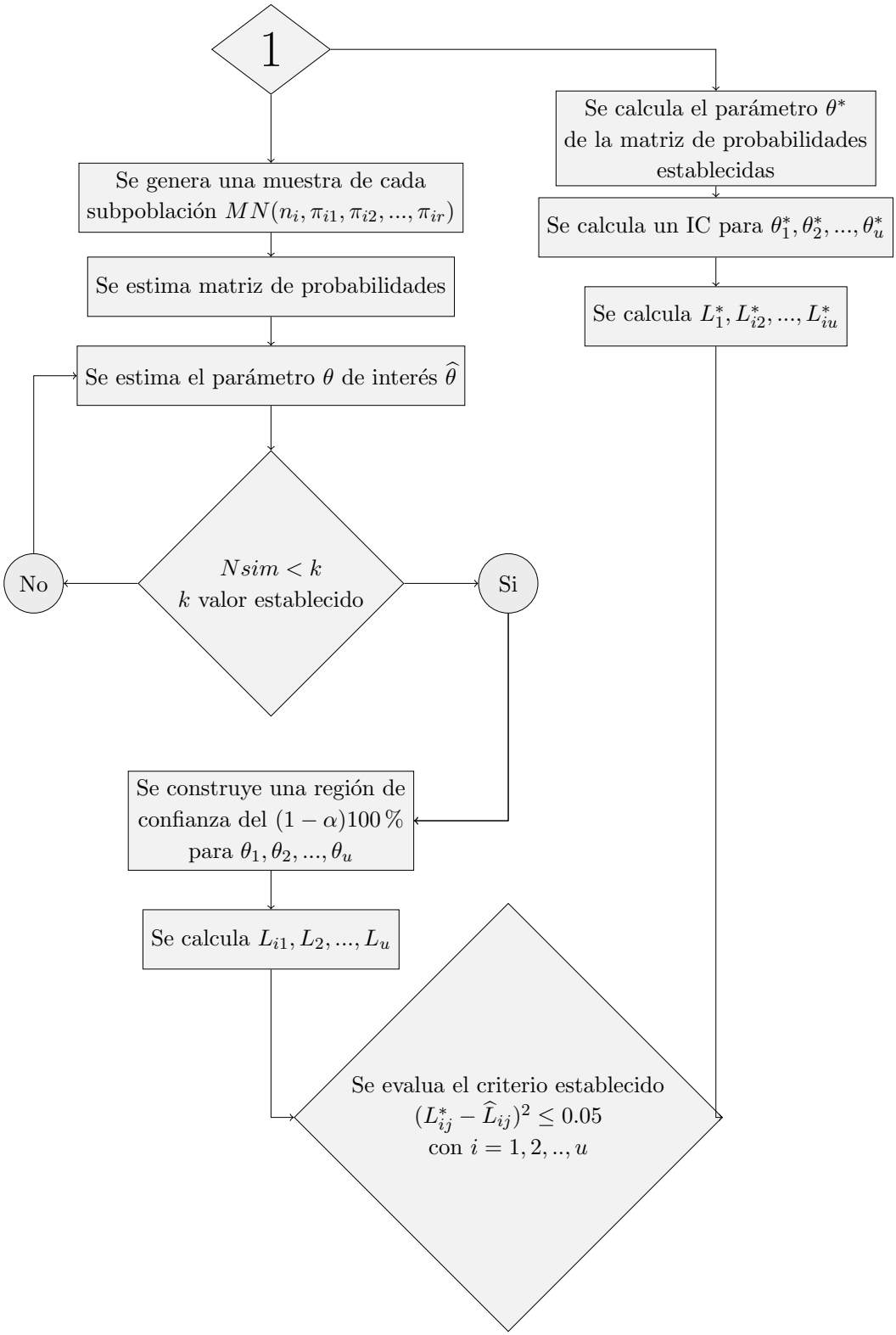
$$\boldsymbol{a.optimizar}\boldsymbol{((L_i)_{(S\times 1)}, L*_i)_{(S\times 1)})} = ((\boldsymbol{L}_i)_{(S\times 1)} - (\boldsymbol{L*}_i)_{(S\times 1)})^2 \leq 0.05 \tag{3-36}$$

entonces, el algoritmo concluye y arroja el vector de los $(n_i)_{optimos}$, el cuál es el vector de valores n_i obtenidos mediante el algoritmo genético, si no se satisface el criterio el algoritmo realiza nuevamente el procedimiento.

3.4. Diagrama de flujo



Figura(3-2) Diagrama de flujo del código para la estimación de los tamaños de muestras de los estratos de la tabla de contingencia. PARTE 1



Figura(3-3) Diagrama de flujo del código para la estimación de los tamaños de muestras de los estratos de la tabla de contingencia. PARTE 2

4 Aplicación de la Metodología GSK

En este capítulo se describe la estimación del modelo utilizando la metodología **GSK**, dado los n_i tamaño de muestra para la i-ésima subpoblación de la tabla con $i = 1, 2, ..., S$ (Grizzle et al, 1969)[12](los n_i son obtenidos por medio del algoritmo genético **sección 3-3**)

Los datos para ilustrar la metodología fueron tomados de la encuesta de ENSIN (Encuesta Nacional de la Situación Nutricional en Colombia, 2005)[15]. Este estudio tiene como objetivo calcular el IMC (Indice de Masa Corporal) para cada uno de los 33 departamentos de Colombia, la población de estudio son hombres y mujeres no embarazadas de edades entre 18 y 64 años.
De este estudio se han seleccionado las variables **Delgadez**, **Normal** y **Sobrepeso**, ya que son de importancia para describir la situación nutricional de los colombianos.
En la siguiente tabla se muestra la clasificación nutricional por índice de masa corporal (IMC) de hombres y mujeres no embarazadas de 18 a 64 años, por departamentos de Colombia en el 2005.

Índice de Masa Corporal (IMC)				
Departamentos	**Delgadez**	**Normal**	**Sobrepeso**	Número de Personas
La Guajira	0.048	0.500	0.452	2253
Cesar	0.054	0.541	,0.406	2331
Magdalena	0.057	0.522	0.422	2180
Atlántico	0.07	0.500	0.429	3499
San Andrés	0.042	0.364	0.593	764
Bolívar	0.077	0.546	0.377	2335
Sucre	0.062	0.515	0.423	2394
Córdoba	0.068	0.559	0.373	2425
Norte de Santander	0.041	0.506	0.453	1947
Santander	0.028	0.502	0.47	1871
Boyacá	0.014	0.513	0.473	1489
Cundinarmarca	0.038	0.439	0.523	1746
Meta	0.029	0.463	0.507	1674
Bogotá	0.019	0.496	0.486	3629
Antioquia	0.035	0.536	0.428	3874
Caldas	0.032	0.519	0.45	1853
Risaralda	0.029	0.509	0.461	1871
Quindío	0.039	0.505	0.457	2184
Tolima	0.017	0.435	0.543	1791
Huila	0.035	0.502	0.463	2239
Caquetá	0.034	0.497	0.469	1509
Valle	0.042	0.456	0.502	4268
Cauca	0.033	0.576	0.392	2214
Nariño	0.017	0.476	0.507	2581
Chocó	0.031	0.497	0.472	1652
Arauca	0.026	0.440	0.536	1130
Casanare	0.023	0.467	0.509	994
Guainía	0.014	0.460	0.526	1158
Vichada	0.019	0.416	0.565	939
Amazonas	0.016	0.419	0.566	1224
Putumayo	0.015	0.438	0.546	959
Guaviare	0.016	0.439	0.544	1191
Vaupés	0.019	0.51	0.47	1010

Tabla 4-1: Clasificación nutricional por índice de masa corporal de hombres y mujeres no embarazadas de 18 a 64 años en el 2005

La **Tabla 4-1** se puede plantear como en la **Tabla 4-2**. En esta se observa cómo se clasifican las unidades de la población en I subpoblaciones o estratos. Las celdas contienen las probabilidades π_{ij}, donde:

1. i se refiere al estrato o subpoblación que pertenece el individuo, con $i = 1, 2, \ldots 33$.
2. j se refiere a la categoría, con $j = 1, 2, 3$.

Por tanto, π_{ij} es la probabilidad de que un sujeto de la i-ésima subpoblación esté en la j-ésima categoría. Esto se puede reescribir como una multinomial con $j = 1, 2, 3$ categorías.

Subpoblación	**1**	**2**	**3**
1	π_{11}^{1}	π_{12}^{1}	π_{13}^{1}
2	π_{11}^{2}	π_{12}^{2}	π_{13}^{2}
3	π_{11}^{3}	π_{12}^{3}	π_{13}^{3}
4	π_{11}^{4}	π_{12}^{4}	π_{13}^{4}
5	π_{11}^{5}	π_{12}^{5}	π_{13}^{5}
6	π_{11}^{6}	π_{12}^{6}	π_{13}^{6}
7	π_{11}^{7}	π_{12}^{7}	π_{13}^{7}
8	π_{11}^{8}	π_{12}^{8}	π_{13}^{8}
9	π_{11}^{9}	π_{12}^{9}	π_{13}^{9}
10	π_{11}^{10}	π_{12}^{10}	π_{13}^{10}
11	π_{11}^{11}	π_{12}^{11}	π_{13}^{11}
12	π_{11}^{12}	π_{12}^{12}	π_{13}^{12}
13	π_{11}^{13}	π_{12}^{13}	π_{13}^{13}
14	π_{11}^{14}	π_{12}^{14}	π_{13}^{14}
15	π_{11}^{15}	π_{12}^{15}	π_{13}^{15}
16	π_{11}^{16}	π_{12}^{16}	π_{13}^{16}
17	π_{11}^{17}	π_{12}^{17}	π_{13}^{17}
18	π_{11}^{18}	π_{12}^{18}	π_{13}^{18}
19	π_{11}^{19}	π_{12}^{19}	π_{13}^{19}
20	π_{11}^{20}	π_{12}^{20}	π_{13}^{20}
21	π_{11}^{21}	π_{12}^{21}	π_{13}^{21}
22	π_{11}^{22}	π_{12}^{22}	π_{13}^{22}
23	π_{11}^{23}	π_{12}^{23}	π_{13}^{23}
24	π_{11}^{24}	π_{12}^{24}	π_{13}^{24}
25	π_{11}^{25}	π_{12}^{25}	π_{13}^{25}
26	π_{11}^{26}	π_{12}^{26}	π_{13}^{26}
27	π_{11}^{27}	π_{12}^{27}	π_{13}^{27}
28	π_{11}^{28}	π_{12}^{28}	π_{13}^{28}
29	π_{11}^{29}	π_{12}^{29}	π_{13}^{29}
30	π_{11}^{30}	π_{12}^{30}	π_{13}^{30}
31	π_{11}^{31}	π_{12}^{31}	π_{13}^{31}
32	π_{11}^{32}	π_{12}^{32}	π_{13}^{32}
33	π_{11}^{33}	π_{12}^{33}	π_{13}^{33}

Tabla 4-2: Distribución de probabilidades para I subpoblaciones generadas

Se tiene que, la distribución de la población es conceptualmente como el producto de I subconjuntos multinomiales, donde cada subpoblación tiene igual número de categorías, así se tienen 3 categorías e I subpoblaciones, con:

$$\pi_{11}^{i} + \pi_{12}^{i} + \pi_{13}^{i} = 1 \tag{4-1}$$

la muestra tomada de la población se representa en la **Tabla 4-4**, la cual tiene la misma estructura que la tabla poblacional.
Para el modelo poblacional, se asume que cada subpoblación es independiente de las otras subpoblaciones, de ahí que la tabla muestral asume también esta condición, es decir, cada subpoblación se toma de una muestra aleatoria e independiente.

Subpoblación	**1**	**2**	**3**	Total
1	n_{11}^{1}	n_{12}^{1}	n_{13}^{1}	n_1
2	n_{11}^{2}	n_{12}^{2}	n_{13}^{2}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{11}^{i}	n_{12}^{i}	n_{13}^{2}	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S	n_{11}^{S}	n_{12}^{S}	n_{13}^{S}	n_S

Tabla 4.4 Tabla muestral

por tanto, a la **Tabla 4-4** se le ajusta un modelo multinomial para la i-ésima subpoblación con función de masa de probabilidad, dada por:

$$P\left(n_{11}^{i}, n_{12}^{i}, n_{13}^{i} | \pi_{11}^{i}, \pi_{12}^{i}, \pi_{13}^{i}\right) = \frac{n_i}{(n_{11}^{i}!)(n_{12}^{i}!)(n_{13}^{i}!)} {\pi_{11}^{(i)}}^{n_{11}^{(i)}} {\pi_{12}^{(i)}}^{n_{12}^{(i)}} {\pi_{13}^{(i)}}^{n_{13}^{(i)}} \tag{4-2}$$

Las n_i observaciones fijas en la i-ésima subpoblación satisfacen que:

$$n_{11}^{i} + n_{12}^{i} + n_{13}^{i} = n_i \tag{4-3}$$

Como las subpoblaciones son independientes la función de probabilidad conjunta para todo el conjunto de datos es el producto de funciones multinomiales dada por:

$$P\left(n_{11}^{i}, n_{12}^{i}, n_{13}^{i} | \pi_{11}^{i}, \pi_{12}^{i}, \pi_{13}^{i}\right) = \prod_{i=1}^{I} \frac{n_i}{(n_{11}^{i}!)(n_{12}^{i}!)(n_{13}^{i}!)} {\pi_{11}^{(i)}}^{n_{11}^{(i)}} {\pi_{12}^{(i)}}^{n_{12}^{(i)}} {\pi_{13}^{(i)}}^{n_{13}^{(i)}} \tag{4-4}$$

con $i = 1, 2, \ldots S$

4.1. Definición de la variable respuesta

La **Tabla 4-2** puede ser representada como un vector donde las tres primeras componentes corresponden a la primera subpoblación y así hasta las tres últimas que corresponden a la última subpoblación.
Este vector bajo ciertas operaciones genera la función respuesta de interés que permite modelar la variable categórica bajo el enfoque **GSK**, así se tiene que:

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{12}^{(1)} \\ \pi_{13}^{(1)} \\ \pi_{11}^{(2)} \\ \pi_{12}^{(2)} \\ \pi_{13}^{(2)} \\ \pi_{11}^{(3)} \\ \pi_{12}^{(3)} \\ \pi_{13}^{(3)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{12}^{(i)} \\ \pi_{13}^{(i)} \\ \vdots \\ \pi_{11}^{(33)} \\ \pi_{12}^{(33)} \\ \pi_{13}^{(33)} \end{bmatrix}$$

Se multiplica el vector $\boldsymbol{\pi}$ por una matriz $\boldsymbol{A}$ de dimensión $S \times 3S$ que es un matriz de bloque y cuya estructura depende del problema que se esté abordando. El resultado es una vector de $3S \times 1$,
Existen muchas funciones respuestas (Correa 2020) [6]. Generalmente las funciones respuesta lineal y logarítmicas son las más importantes.
Las funciones respuesta lineal se obtiene del conjunto de probabilidades observadas como:

$$\boldsymbol{f} = A\boldsymbol{\pi} \tag{4-5}$$

donde $\boldsymbol{f}$ es un vector de u componentes, A es una matriz de tamaño $u \times RS$ y $\boldsymbol{\pi}$ es el vector con rs componentes.
Cuando se trabaja con la Tabla muestral, entonces f es estimada por

$$\widehat{\boldsymbol{f}} = A\widehat{\boldsymbol{\pi}} \tag{4-6}$$

En este trabajo, se considera la información de la **Tabla 4-1** donde cada subpoblación multinomial contiene 3 categorías: **Delgadez, Normal y Sobrepeso** y la tabla tiene 99 componentes, para cada i subpoblación con $i = 1, 2, 3, \ldots, 33$ la distribución será:

Suppoblación	Delgadez	Normal	Sobrepeso
i	π_{11}^i	π_{12}^i	π_{13}^i

Tabla 4-3: Distribución i-ésima supoboblación

Para ilustrar la construcción de la matriz A, se considera la i-ésima subpoblación de la **Tabla 4-3**, donde se está interesado en dos funciones,

$$f_1^i = log\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right) \tag{4-7}$$

$$f_2^i = log\left(\frac{\pi_{12}^i}{\pi_{13}^i}\right) \tag{4-8}$$

Donde las funciones f_1^i y f_2^i se pueden considerar como funciones aditivas de la media de los efectos de las S subpoblaciones.

Se propone una relación logarítmica donde

$$\boldsymbol{f(\pi)} = \boldsymbol{K} ln(\boldsymbol{A\pi}) \tag{4-9}$$

con $\boldsymbol{A} = \boldsymbol{I}_{3S\times 3S}$, es decir, $\boldsymbol{A}$ es una matriz identidad de dimensiones $3S \times 3S$ y $\boldsymbol{K}$ definida por

$$\boldsymbol{K} = \begin{bmatrix}
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & 0 & -1 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & -1 & 0
\end{bmatrix}_{66\times 99}$$

y se obtiene una familia de funciones de la forma:

$$\boldsymbol{K\,Ln(A\pi)} =$$

$$\begin{bmatrix}
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 & -1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & 0 & -1 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & \vdots & & & & \dots & & \vdots & & \dots & & & & \vdots & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & -1 & 0
\end{bmatrix}
\begin{bmatrix}
log\left(\pi_{11}^{(1)}\right) \\
log\left(\pi_{12}^{(1)}\right) \\
log\left(\pi_{13}^{(1)}\right) \\
log\left(\pi_{11}^{(2)}\right) \\
log\left(\pi_{12}^{(2)}\right) \\
log\left(\pi_{13}^{(2)}\right) \\
log\left(\pi_{11}^{(3)}\right) \\
log\left(\pi_{12}^{(3)}\right) \\
log\left(\pi_{13}^{(3)}\right) \\
\vdots \\
log\left(\pi_{11}^{(i)}\right) \\
log\left(\pi_{12}^{(i)}\right) \\
log\left(\pi_{13}^{(i)}\right) \\
\vdots \\
log\left(\pi_{11}^{(33)}\right) \\
log\left(\pi_{12}^{(33)}\right) \\
log\left(\pi_{13}^{(33)}\right)
\end{bmatrix}$$

$$= \begin{bmatrix} log\left(\pi_{11}^{1}\right) - log\left(\pi_{13}^{1}\right) \\ log\left(\pi_{11}^{2}\right) - log\left(\pi_{13}^{2}\right) \\ \vdots \\ log\left(\pi_{11}^{i}\right) - log\left(\pi_{13}^{i}\right) \\ \vdots \\ log\left(\pi_{11}^{32}\right) - log\left(\pi_{13}^{32}\right) \\ log\left(\pi_{11}^{33}\right) - log\left(\pi_{13}^{33}\right) \\ log\left(\pi_{12}^{1}\right) - log\left(\pi_{13}^{1}\right) \\ log\left(\pi_{12}^{2}\right) - log\left(\pi_{13}^{2}\right) \\ \vdots \\ log\left(\pi_{12}^{i}\right) - log\left(\pi_{13}^{i}\right) \\ \vdots \\ log\left(\pi_{12}^{32}\right) - log\left(\pi_{13}^{32}\right) \\ log\left(\pi_{12}^{33}\right) - log\left(\pi_{13}^{33}\right) \end{bmatrix}$$

Aplicando propiedades de la función logaritmo, se tiene que:

$$\boldsymbol{f} = \begin{bmatrix} f_1^1 \\ f_1^2 \\ f_1^3 \\ \vdots \\ f_1^i \\ \vdots \\ f_1^{32} \\ f_1^{33} \\ f_2^1 \\ f_2^2 \\ f_2^3 \\ \vdots \\ f_2^i \\ \vdots \\ f_2^{32} \\ f_2^{33} \end{bmatrix} = \begin{bmatrix} \ln\left(\frac{\pi_{11}^{1}}{\pi_{13}^{1}}\right) \\ \ln\left(\frac{\pi_{11}^{2}}{\pi_{13}^{2}}\right) \\ \ln\left(\frac{\pi_{11}^{3}}{\pi_{13}^{3}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{i}}{\pi_{13}^{i}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{33}}{\pi_{13}^{33}}\right) \\ \ln\left(\frac{\pi_{12}^{1}}{\pi_{13}^{1}}\right) \\ \ln\left(\frac{\pi_{12}^{2}}{\pi_{13}^{2}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{12}^{i}}{\pi_{13}^{i}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{12}^{32}}{\pi_{13}^{32}}\right) \\ \ln\left(\frac{\pi_{12}^{33}}{\pi_{13}^{33}}\right) \end{bmatrix}$$

Um estimador de máxima verosimilitud para $\boldsymbol{\pi}$ es $\widehat{\boldsymbol{\pi}}$, donde:

$$\widehat{\pi^i_{11}} = \frac{n^i_{11}}{n_i}; \qquad \widehat{\pi^i_{12}} = \frac{n^i_{12}}{n_i} \qquad \text{y} \qquad \widehat{\pi^i_{13}} = \frac{n^i_{13}}{n_i} \tag{4-10}$$

Con n^i_{11}, n^i_{12} y n^i_{13} son las frecuencias observadas de la i-ésima subpoblación y $n_i = n^i_{11} + n^i_{12} + n^i_{13}$, el tamaño total de la i-ésima subpoblación. (Grizzle et al, 1969) [12].

Por tanto

$$\widehat{\boldsymbol{\pi}} = \begin{bmatrix} \widehat{\pi^{(1)}_{11}} \\ \widehat{\pi^{(1)}_{12}} \\ \widehat{\pi^{(1)}_{13}} \\ \vdots \\ \widehat{\pi^{(i)}_{11}} \\ \widehat{\pi^{(i)}_{12}} \\ \widehat{\pi^{(i)}_{13}} \\ \vdots \\ \widehat{\pi^{(33)}_{12}} \\ \widehat{\pi^{(33)}_{13}} \end{bmatrix} = \begin{bmatrix} \dfrac{n^{(1)}_{11}}{n_1} \\ \dfrac{n^{(1)}_{12}}{n_1} \\ \dfrac{n^{(1)}_{13}}{n_1} \\ \vdots \\ \dfrac{n^{(i)}_{11}}{n_i} \\ \dfrac{n^{(i)}_{12}}{n_i} \\ \dfrac{n^{(i)}_{13}}{n_i} \\ \vdots \\ \dfrac{n^{(33)}_{12}}{n_{33}} \\ \dfrac{n^{(33)}_{13}}{n_{33}} \end{bmatrix} \tag{4-11}$$

Los valores esperados para los estimadores de la i-ésima categoría son:

$$E\big(\hat{\pi}^i_{11}\big) = \pi^i_{11}, \quad E\big(\hat{\pi}^i_{12}\big) = \pi^i_{12} \quad \text{y} \quad E\big(\hat{\pi}^i_{13}\big) = \pi^i_{13} \tag{4-12}$$

Se puede demostrar por el teorema central del límite multivariado (Rao, 1973 p.128) que los $\widehat{\pi}$ sigue asintóticamente una distribución normal $AN(\pi, \Sigma_\pi)$.

La metodología GSK requiere estimar la varianzas y la covarianzas de $\widehat{\boldsymbol{\pi}}$. Para la i-ésima subpoblación la cual es multinomial, sea

$$\widehat{\boldsymbol{\pi}}_{(\boldsymbol{i})} = \begin{bmatrix} \hat{\pi}^{(i)}_{11} \\ \hat{\pi}^{(i)}_{12} \\ \hat{\pi}^{(i)}_{13} \end{bmatrix} \tag{4-13}$$

Para la i-ésima subpoblación de una muestra aleatoria de tamaño n_i, las varianzas de los estimadores son:

$$Var\big(\hat{\pi}_{11}^{(i)}\big) = \frac{1}{n_i}\big(\pi_{11}^i(1-\pi_{11}^i)\big), \quad Var\big(\hat{\pi}_{12}^{(i)}\big) = \frac{1}{n_i}\big(\pi_{12}^i(1-\pi_{12}^i)\big) \quad \text{y} \quad Var\big(\hat{\pi}_{13}^{(i)}\big) = \frac{1}{n_i}\big(\pi_{13}^i(1-\pi_{13}^i)\big) \tag{4-14}$$

Las covarianzas están dadas por:

$$Cov(\hat{\pi}_{11}^{(i)}, \hat{\pi}_{12}^{(i)}) = -\pi_{11}^i\pi_{12}^i, \quad Cov(\hat{\pi}_{11}^{(i)}, \hat{\pi}_{13}^{(i)}) = -\pi_{11}^i\pi_{13}^i \quad \text{y} \quad Cov(\hat{\pi}_{12}^{(i)}, \hat{\pi}_{13}^{(i)}) = -\pi_{12}^i\pi_{13}^i \tag{4-15}$$

En el caso de la distribución multinomial, la matriz de covarianzas no es de rango completo y frecuentemente requiere ser trabajada con una inversa generalizada (Tanabe and Sagae, 1992) [22].
La matriz de varianzas y covarianzas para $\widehat{\pi}_i$ es:

$$\Sigma_{\widehat{\pi}_i} = \begin{bmatrix} \pi_{11}^i(1-\pi_{11}^i) & -\pi_{11}^i\pi_{12}^i & -\pi_{11}^i\pi_{13}^i \\ -\pi_{11}^i\pi_{12}^i & \pi_{12}^i(1-\pi_{12}^i) & -\pi_{12}^i\pi_{13}^i \\ -\pi_{13}^i\pi_{11}^i & -\pi_{13}^i\pi_{12}^i & \pi_{13}^i(1-\pi_{13}^i) \end{bmatrix} \tag{4-16}$$

Por tanto la matriz de varianzas y covarianzas para $\boldsymbol{\pi}$ está dada por Σ_π, una matriz de bloque, donde en la diagonal principal tiene a $\Sigma_{\hat{\pi}_i}$, con $i = 1, 2, \ldots, 33$

$$\Sigma_{\widehat{\pi}} = \begin{bmatrix} \Sigma_{\widehat{\pi}_1} & 0 & 0 & \ldots & 0 & 0 \\ 0 & \Sigma_{\widehat{\pi}_2} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \Sigma_{\widehat{\pi}_3} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \Sigma_{\widehat{\pi}_{32}} & 0 \\ 0 & 0 & 0 & \ldots & 0 & \Sigma_{\widehat{\pi}_{33}} \end{bmatrix} \tag{4-17}$$

Como asintóticamente $\widehat{\boldsymbol{\pi}}$ se distribuye multinomial con media $\boldsymbol{\pi}$ y matriz de varianzas y covarianzas dada por $\Sigma_{\widehat{\pi}}$, entonces $\widehat{\boldsymbol{f}} = K\boldsymbol{Ln}(\boldsymbol{A}\widehat{\boldsymbol{\pi}})$ se distribuye asintóticamente multinormal con matriz de varianzas y covarianzas dada por $\Sigma_{\widehat{f}} = \boldsymbol{K}\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\pi}}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{D}^{-1}\boldsymbol{K}^{\boldsymbol{T}}$ (Grizzle et al., 1969)[12], donde $\boldsymbol{A} = I_{2S\times 2S}$ y $\boldsymbol{D}$ es una matriz diagonal dada por:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{a_1^T\pi} & 0 & 0 & \ldots & 0 & 0 \\ 0 & \boldsymbol{a_2^T\pi} & 0 & \ldots & 0 & 0 \\ 0 & 0 & \boldsymbol{a_3^T\pi} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \boldsymbol{a_{2(S-1)}^T\pi} & 0 \\ 0 & 0 & 0 & \ldots & 0 & \boldsymbol{a_{2S}^T\pi} \end{bmatrix} \tag{4-18}$$

donde $\boldsymbol{a_i^T}$ es la i-ésima fila de la matriz $\boldsymbol{A}$.
Por tanto

$$\boldsymbol{D} = \begin{bmatrix} \pi_{11}^1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \pi_{12}^1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \pi_{13}^1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \pi_{12}^{33} & 0 \\ 0 & 0 & 0 & \dots & 0 & \pi_{13^{33}} \end{bmatrix} \tag{4-19}$$

donde la matriz de varianzas y covarianzas estimada $\Sigma_{\widehat{\pi_i}}$ para la i-ésima subpoblación, con $i = 1, 2, ...33$ es:

$$\widehat{\Sigma_{\widehat{\pi_i}}} = \begin{bmatrix} \widehat{\pi_{11}^i}(1 - \widehat{\pi_{11}^i}) & -\widehat{\pi_{11}^i}\widehat{\pi_{12}^i} & -\widehat{\pi_{11}^i}\widehat{\pi_{13}^i} \\ -\widehat{\pi_{11}^i}\widehat{\pi_{12}^i} & \widehat{\pi_{12}^i}(1 - \widehat{\pi_{12}^i}) & -\widehat{\pi_{12}^i}\widehat{\pi_{13}^i} \\ -\widehat{\pi_{13}^i}\widehat{\pi_{11}^i} & -\widehat{\pi_{13}^i}\widehat{\pi_{12}^i} & \widehat{\pi_{13}^i}(1 - \widehat{\pi_{13}^i}) \end{bmatrix} \tag{4-20}$$

donde $\widehat{\Sigma_{\widehat{\pi_i}}}$, con $i = 1, 2, ..., 33$ es una matriz positiva y simétrica. (Grizzle 1969) [12].

y donde la matriz estimada de $\Sigma_{\widehat{f}}$ está dada por:

$$\widehat{\Sigma}_{\widehat{f}} = \boldsymbol{K}\widehat{\boldsymbol{D}}^{-\mathbf{1}}\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\pi}}}\boldsymbol{A}^{\boldsymbol{T}}\widehat{\boldsymbol{D}}^{-\mathbf{1}}\boldsymbol{K}^{\boldsymbol{T}} \tag{4-21}$$

con

$$\widehat{\Sigma_{\widehat{\pi}}} = \begin{bmatrix} \widehat{\Sigma}_{\widehat{\pi}_1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \widehat{\Sigma}_{\widehat{\pi}_2} & 0 & \dots & 0 & 0 \\ 0 & 0 & \widehat{\Sigma}_{\widehat{\pi}_3} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \widehat{\Sigma}_{\widehat{\pi}_{32}} & 0 \\ 0 & 0 & 0 & \dots & 0 & \widehat{\Sigma}_{\widehat{\pi}_{33}} \end{bmatrix} \tag{4-22}$$

y la matriz estimada para $\boldsymbol{D}$ dada por:

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} \widehat{\pi_{11}^1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \widehat{\pi_{12}^1} & 0 & \dots & 0 & 0 \\ 0 & 0 & \widehat{\pi_{13}^1} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \widehat{\pi_{12}^{33}} & 0 \\ 0 & 0 & 0 & \dots & 0 & \widehat{\pi_{13^{33}}} \end{bmatrix} \tag{4-23}$$

4.1.1. Intervalo de confianza para respuestas de la forma $\ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right)$

En este trabajo, se propone funciones respuestas de la forma $ln(\psi) = \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right)$, con $i = 1, 2..., S$ y $j, k = 1, 2..., R$, donde ψ es una razón de valores de la tabla. (Agresti 2010) [3] propone un intervalo de confianza con un nivel de $(1-\alpha)100\,\%$ para $\ln(\psi)$ dado por:

$$\left(\ln(\psi) - z_{1-\alpha/2}\sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{ik}}}, \quad \ln(\psi) + z_{1-\alpha/2}\sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{ik}}}\right) \tag{4-24}$$

4.2. Modelo lineal bajo la metodología GSK

El modelo lineal paramétrico que relaciona las variables de interés con un conjunto de covariables para cada subpoblación es:

$$\widehat{\boldsymbol{f}} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4-25}$$

donde $\widehat{\boldsymbol{f}}$ es la función respuesta muestral.

$$\widehat{\boldsymbol{f}} = \begin{bmatrix} \widehat{f}_1 \\ \widehat{f}_2 \\ \widehat{f}_3 \\ \widehat{f}_4 \\ \widehat{f}_5 \\ \widehat{f}_6 \\ \vdots \\ \widehat{f}_i \\ \vdots \\ \widehat{f}_{31} \\ \widehat{f}_{32} \\ \widehat{f}_{33} \end{bmatrix} \tag{4-26}$$

$\boldsymbol{X}$ es una matriz de diseño de de dimensiones 66×66, y $\boldsymbol{\beta}$ es un vector de tamaño 66×1 de parámetros desconocido y $\boldsymbol{\epsilon}$ un vector que se distribuye asintóticamente normal, con media $\mathbf{0}$ y $Var(\boldsymbol{\epsilon}) = \Sigma_{\widehat{\boldsymbol{\epsilon}}}$, esto es, $(\boldsymbol{\epsilon} \sim AN(0; \Sigma_{\widehat{\boldsymbol{\epsilon}}}))$, donde el mejor estimador asintóticamente normal para $\boldsymbol{\beta}$ es:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \widehat{\Sigma_{\widehat{\boldsymbol{f}}}}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \widehat{\Sigma_{\widehat{\boldsymbol{f}}}}^{-1} \widehat{\boldsymbol{f}}) \tag{4-27}$$

con $\widehat{\boldsymbol{\beta}}$ un estimador insesgado de $\boldsymbol{\beta}$, el cual minimiza la ecuación (4-27):

$$S(\boldsymbol{\beta}) = \left(\widehat{\boldsymbol{f}} - X\boldsymbol{\beta}\right)^T \Sigma_{\widehat{f}}^{-1} \left(\widehat{\boldsymbol{f}} - X\boldsymbol{\beta}\right) \tag{4-28}$$

La matriz de varianzas y covarianzas de $\widehat{\boldsymbol{\beta}}$ es $\Sigma_{\widehat{\boldsymbol{\beta}}}$ y su estimación es dada por:

$$\widehat{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \left(X^T \widehat{\Sigma}_{\widehat{\boldsymbol{f}}} X\right)^{-1} \tag{4-29}$$

Definición de la matriz de diseño del modelo

A continuación se presenta la matriz de diseño $\boldsymbol{X}$, ésta puede contener información del estrato o subpoblación expresada en términos de la variable categórica asociada al estrato o subpoblación.

$$\boldsymbol{X} = \left[\begin{array}{cc}
\begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \\
\begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}
\end{array}\right]$$

Matriz 4-1 : Elementos de la matriz de diseño $\boldsymbol{X}$

En el ejemplo a tratar sobre la encuesta realizada por ICBF (Instituto Colombiano de Bienestar Familiar) (Encuesta Nacional de la Situación Nuticional, 2005)[15], se genera la matriz

de diseño ilustrada por **Matriz 4-1**.
Por tanto el vector de parámetros de $\boldsymbol{\beta}$ estaría dado por:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{30} \\ \beta_{31} \\ \beta_{32} \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{30} \\ \alpha_{31} \\ \alpha_{32} \end{bmatrix} \tag{4-30}$$

Dado el objetivo propuesto en este trabajo, se asume que las funciones respuestas en los distintos niveles correspondientes a la **Tabla 4-3**, son independientes entre sí y que provienen de una distribución multinomial donde para la i-ésima respuesta Y_i con $i = 1, 2, \ldots, 66$

$$Y_i = \boldsymbol{X_i}\boldsymbol{\beta} \sim MULTINOMIAL(n_i, \boldsymbol{\pi_i}) \tag{4-31}$$

donde $\boldsymbol{X_i}$ es la i-ésima fila de la matriz de diseño X.

Entre los muchos modelos que se pueden desarrollar para este caso (Agresti 1977)[3], el modelo de regresión logística es de particular interés aquí.

$$Y_i = ln\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right) = \boldsymbol{X_i^T}\boldsymbol{\beta} \quad \text{con} \quad i = 1, 2, ..., 33 \tag{4-32}$$

$$Y_j = ln\left(\frac{\pi_{12}^j}{\pi_{13}^j}\right) = \boldsymbol{X_{j+33}^T}\boldsymbol{\beta} \quad \text{con} \quad j = 1, 2, ..., 33 \tag{4-33}$$

se tiene que:

$$\begin{bmatrix} \ln\left(\frac{\pi_{11}^1}{\pi_{13}^1}\right) \\ \ln\left(\frac{\pi_{11}^2}{\pi_{13}^2}\right) \\ \ln\left(\frac{\pi_{11}^3}{\pi_{13}^3}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{33}}{\pi_{13}^{33}}\right) \\ \ln\left(\frac{\pi_{12}^1}{\pi_{13}^1}\right) \\ \ln\left(\frac{\pi_{12}^2}{\pi_{13}^2}\right) \\ \vdots \\ \ln\left(\frac{\pi_{12}^i}{\pi_{13}^i}\right) \\ \vdots \\ \ln\left(\frac{\pi_{12}^{32}}{\pi_{13}^{32}}\right) \\ \ln\left(\frac{\pi_{12}^{33}}{\pi_{13}^{33}}\right) \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{30} \\ \beta_{31} \\ \beta_{32} \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{30} \\ \alpha_{31} \\ \alpha_{32} \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 \\ \beta_0 + \beta_2 \\ \beta_0 + \beta_3 \\ \vdots \\ \beta_0 + \beta_{32} \\ \beta_0 \\ \alpha_0 + \alpha_1 \\ \alpha_0 + \alpha_2 \\ \alpha_0 + \alpha_3 \\ \vdots \\ \alpha_0 + \alpha_{32} \\ \alpha_0 \end{bmatrix}$$

β_0 : Media de los $\ln\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right)$ de índice de masa corporal en los 33 departamentos

β_i : Diferencia de la media del IMC del i-ésimo departamento con respecto a β_0, con $i = 1, 2, \dots, 32$.

α_0 : Media de los $\ln\left(\frac{\pi_{12}^i}{\pi_{13}^i}\right)$ de índice de masa corporal en los 33 departamentos
α_i : Diferencia de la media del IMC del i-ésimo departamento con respecto a α_0, con $i = 1, 2, \ldots, 32$.

y el vector de error $\boldsymbol{\epsilon}$ sería:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{64} \\ \epsilon_{65} \\ \epsilon_{66} \end{bmatrix}$$

La estimación de $\widehat{\boldsymbol{f}}$ denotada por $\widehat{\boldsymbol{f}}*$ está dado por :

$$\widehat{\boldsymbol{f}}* = X\widehat{\boldsymbol{\beta}} \tag{4-34}$$

donde la matriz de varianza y covarianzas para $\widehat{\boldsymbol{f}}*$ denotada por $\widehat{\Sigma}_{\widehat{\boldsymbol{f}}*}$ está dada por:

$$\widehat{\Sigma}_{\widehat{\boldsymbol{f}}*} = X\left(X^T\widehat{\Sigma}_{\widehat{\boldsymbol{f}}}X\right)^{-1}X^T \tag{4-35}$$

estos resultados son asintóticos (Agresti, 1996a)[2], (Grizzle et al., 1996)[12].

4.3. Inferencia sobre el modelo

La prueba global para el modelo tiene en cuenta todos los coeficientes de $\widehat{\beta}$, excepto β_0 y α_0. Se plantea la hipótesis:

$$\begin{cases} H_0 : C\boldsymbol{\beta} = \mathbf{0} \qquad \text{equivalente} \qquad \beta_1 = ... = \beta_{32} = \alpha_1 = ... = \alpha_{32} \\ \\ \text{vs} \\ \\ H_1 : C\boldsymbol{\beta} \neq 0 \qquad \text{equivalente} \quad \text{al menos } \beta_i, \alpha_i \neq 0 \end{cases} \tag{4-36}$$

con $i = 1, 2, ..., 32$ y estadístico de prueba dado por:

$$\left(C\widehat{\boldsymbol{\beta}}\right)^T \left(C\widehat{\Sigma_{\widehat{\beta}}}C^T\right) C\widehat{\boldsymbol{\beta}} \sim \chi^2_v \tag{4-37}$$

con v= rango de la matriz C (Grizzle 1979) [11].
Para el modelo planteado en el ejemplo de este trabajo, la matriz C sería:

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

Para el caso que solo se quiera evaluar f_1, es decir, para :

$$\begin{cases} H_0 : C\boldsymbol{\beta} = \mathbf{0} \qquad \text{equivalente} \qquad \beta_1 = ... = \beta_{32} = 0 \\ \\ \text{vs} \\ \\ H_1 : C\boldsymbol{\beta} \neq 0 \qquad \text{equivalente} \quad \text{al menos } \beta_i \neq 0 \end{cases} \tag{4-38}$$

con $i = 1, 2, ..., 32$, la matriz C sería:

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 \end{bmatrix}_{32\times 66}$$

y para el caso que se quiera evaluar solo f_2, es decir:

$$\begin{cases} H_0 : C\boldsymbol{\beta} = \mathbf{0} & \text{equivalente} \quad \alpha_1 = ... = \alpha_{32} = 0 \\ \\ \text{vs} \\ \\ H_1 : C\boldsymbol{\beta} \neq 0 & \text{equivalente} \quad \text{al menos } \alpha_i \neq 0 \end{cases} \tag{4-39}$$

con $i = 1, 2..., 32$, la matriz C sería:

$$\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}_{32\times 66}$$

Para evaluar la significancia de cada parámetro estimado para el modelo, se debe generar un vector fila C con 1 en la i-ésima componente, tal que el estimador de interés está dado por $C\widehat{\boldsymbol{\beta}}$. Para ilustrar esto, si se desea evaluar el efecto de β_1, multiplicamos el vector $\widehat{\boldsymbol{\beta}}$ por C, donde

$$C = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 & 0 & ... & 0 & 0 \end{bmatrix}_{1\times 66}$$

o, si se desea evaluar el efecto de α_1, multiplicamos el vector $\widehat{\boldsymbol{\beta}}$ por C, donde

$$C = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 1 & 0 & ... & 0 & 0 \end{bmatrix}_{1\times 66}$$

con 1 en la posición 34, ya que el parámetro α_1 se encuentra en la posición 34 del vector $\widehat{\boldsymbol{\beta}}$. El estadístico de prueba es

$$\left(C\widehat{\boldsymbol{\beta}}\right)^T \left(C\widehat{\Sigma}_{\widehat{\beta}}C^T\right) C\widehat{\boldsymbol{\beta}} \sim \chi^2_1 \tag{4-40}$$

4.4. Residuales del modelo GSK

4.4.1. Residuales en el modelo GSK

El análisis de los residuos de ajuste del modelo nos permite determinar su adecuación. Para el modelo **GSK**, se puede definir residuos para:

- Función respuesta.
- Tabla de conteo sin procesar (original).

El análisis de residuales en el estudio de datos categóricos, se ha usado para establecer si existen valores atípicos en las celdas de la tabla de contingencia, ya que estos valores atípicos nos indican la calidad del ajuste del modelo a la tabla de contingencia.

Sean los vectores:
$\widehat{\boldsymbol{f}}$: Estimación del valor observado de **f**.
$\widehat{\boldsymbol{f}^*}$: Ajuste de la estimación del valor observado de **f**.
Y se define **e** como el vector resultante de la diferencia entre $\widehat{\boldsymbol{f}}$ y $\widehat{\boldsymbol{f}^*}$, donde e_t la t-ésima componente de **e** es:

$$e_t = f_t - f_t^* \tag{4-41}$$

$$e_t = \frac{f_t - f_t^*}{s.e(f_t^*)} \tag{4-42}$$

Donde f_t^* y f_t corresponden al t-ésimo elemento de los vectores $\widehat{\boldsymbol{f}^*}$ y $\widehat{\boldsymbol{f}}$ respectivamente.

4.4.2. Residuales para la Tabla

Teniendo en cuenta que el modelo genera un $\widehat{\boldsymbol{f}}$ que no es la probabilidad de las tablas sino una función de dichas probabilidades sobre las muestras, se debe reconvertir estas funciones **f** en las probabilidades de la tabla.
Como se estableció en la sección 4.2

$$f_i\left(\widehat{\pi_{11}^i}, \widehat{\pi_{12}^i}, \widehat{\pi_{13}^i}\right) = \ln\left(\frac{\widehat{\pi_{11}^i}^*}{\widehat{\pi_{13}^i}^*}\right) = \boldsymbol{X_i^T\beta} \tag{4-43}$$

$$f_i\left(\widehat{\pi_{11}^i}, \widehat{\pi_{12}^i}, \widehat{\pi_{13}^i}\right) = \ln\left(\frac{\widehat{\pi_{12}^i}^*}{\widehat{\pi_{13}^i}^*}\right) = \boldsymbol{X_{i+S}^T\beta} \tag{4-44}$$

con $i = 1, 2, \ldots, S$, aplicando la función *exp* en las ecuaciones (4-42) y (4-43), se tiene que:

$$\begin{pmatrix} \widehat{\pi_{11}^i}^* \\ \widehat{\pi_{13}^i}^* \end{pmatrix} = \exp(\boldsymbol{X_i^T \beta}) \tag{4-45}$$

$$\begin{pmatrix} \widehat{\pi_{12}^i}^* \\ \widehat{\pi_{13}^i}^* \end{pmatrix} = \exp(\boldsymbol{X_{i+S}^T \beta}) \tag{4-46}$$

Se sabe que $\sum_{j=1}^{3} \widehat{\pi_{1j}^i}^* = 1$, por tanto $\widehat{\pi_{13}^i}^* = 1 - \widehat{\pi_{11}^i}^* - \widehat{\pi_{12}^i}^*$ y realizando el cociente de las ecuaciones (4-44) y (4-45) se tiene que:

$$\frac{\widehat{\pi_{11}^i}^*}{\widehat{\pi_{12}^i}^*} = \frac{exp(\boldsymbol{X_i^T \beta})}{exp(\boldsymbol{X_{i+S}^T \beta})} \tag{4-47}$$

Despejando $\widehat{\pi_{11}^i}^*$ de la ecuación (4-46), se tiene que:

$$\widehat{\pi_{11}^i}^* = \left(1 - \widehat{\pi_{11}^i}^* - \widehat{\pi_{12}^i}^*\right) exp(\boldsymbol{X_i^T \beta}) \tag{4-48}$$

Ahora reemplazando la ecuación (4-47) en las ecuaciones (4-44) y (4-45) se obtiene que:

$$\widehat{\pi_{11}^i}^* = \frac{\exp(\boldsymbol{X_i^T \beta})}{\left(1 + exp(\boldsymbol{X_i^T \beta}) + exp(\boldsymbol{X_{i+S}^T \beta}\right)} \tag{4-49}$$

$$\widehat{\pi_{12}^i}^* = \frac{\exp(\boldsymbol{X_{i+S}^T \beta})}{\left(1 + exp(\boldsymbol{X_i^T \beta}) + exp(\boldsymbol{X_{i+S}^T \beta}\right)} \tag{4-50}$$

y reemplazando la ecuación (4-48) en (4-45) se tiene:

$$\widehat{\pi_{13}^i}^* = \frac{1}{\left(1 + exp(\boldsymbol{X_i^T \beta}) + exp(\boldsymbol{X_{i+S}^T \beta})\right)} \tag{4-51}$$

Por tanto los i-ésimos residuales para la i-ésima subpoblación son:

$$r_{\pi_{11}}^{(i)} = \widehat{\pi_{11}^i} - \widehat{\pi_{11}^i}^* \tag{4-52}$$

$$r_{\pi_{12}}^{(i)} = \widehat{\pi_{12}^i} - \widehat{\pi_{12}^i}^* \tag{4-53}$$

$$r_{\pi_{13}}^{(i)} = \widehat{\pi_{13}^i} - \widehat{\pi_{13}^i}^* \tag{4-54}$$

Como $\boldsymbol{\pi_i}$ tienen distribución multinomial, un residual pseudo estandarizado sería:

$$r_{\pi_{11}}^{(i)} = \frac{\widehat{\pi_{11}^i} - \widehat{\pi_{11}^i}^*}{\sqrt{\frac{1 - \widehat{\pi_{11}^i}^*}{n_{i1}}}} \tag{4-55}$$

$$r_{\pi_{12}}^{(i)} = \frac{\widehat{\pi_{12}^i} - \widehat{\pi_{12}^i}^*}{\sqrt{\frac{1 - \widehat{\pi_{12}^i}^*}{n_{i2}}}} \tag{4-56}$$

$$r_{\pi_{13}}^{(i)} = \frac{\widehat{\pi_{13}^i} - \widehat{\pi_{13}^i}^*}{\sqrt{\frac{1 - \widehat{\pi_{13}^i}^*}{n_{i3}}}} \tag{4-57}$$

5 Aplicaciones

En este capítulo se implementará el algoritmo que se trabajó en la **sección 3.3** y se aplica el modelo mencionado en el **capítulo 4** al ejemplo sobre la clasificación nutricional por índice de masa corporal de los colombianos hombres y mujeres no embarazadas de edades entre 18 y 64 años.

Como se mostró en la **Tabla 4-1**, se genera 33 subpoblaciones o estratos, en este trabajo se utiliza

$P=$

Subpoblación	**1**	**2**	**3**
1	0.1	0.5	0.4
2	0.1	0.5	0.4
3	0.1	0.5	0.4
⋮	⋮	⋮	⋮
31	0.1	0.5	0.4
32	0.1	0.5	0.4
33	0.1	0.5	0.4

Tabla 5-1: Matriz de distribución de Probabilidades establecidas para las 33 subpoblaciones

Se toma igual probabilidades por subpoblación para obtener un muestreo balanceado y a partir de la **Tabla 5-1** se procederá al cálculo de los tamaños de muestras óptimos para cada subpoblación vía algoritmos genéticos y la estimación del modelo utilizando la metodología GSK.

Para comenzar a ejecutar el algoritmo, se utilizan los siguientes atributos requeridos:
Dos vectores para el rango de búsqueda **mínimo**$= (60, 300, 240)$ y **máximo**$= (500, 2500, 2000)$ los cuales se establecieron con dichas características basados en los tamaños muestrales observados en la encuesta de ENSIN(Encuesta Nacional de la Situación Nutricional en Colombia, 2005)[15]. $Nsim = 1000$ número de simulaciones, $Nbits = 8$ número de bits para cada cromosoma, $Npar = 3$ número de variables, función de costo y función objetivo definidas previamente en la **sección 3.3**. En este punto para cada una de las $n - generaciones = 30$ generaciones,
con el objetivo de tener resultados balanceados, se estableció para la i-ésima subpoblación o estrato un vector $errores_i = (0.10, 0.14, 0.05)$ de longitudes medias de los parámetros y

se toma como valores para los tamaños de las poblaciones iniciales de las subpoblación, los valores de **Número de personas** de la **Tabla (4-1)**, esto es:

$$N = \begin{bmatrix} 2253 \\ 2331 \\ 2180 \\ 2499 \\ 764 \\ 2335 \\ 2394 \\ 2425 \\ 1947 \\ 1871 \\ 1489 \\ 1746 \\ 1674 \\ 3629 \\ 1853 \\ 1871 \\ 2184 \\ 1791 \\ 2239 \\ 1509 \\ 4268 \\ 2214 \\ 2581 \\ 1652 \\ 1130 \\ 994 \\ 1158 \\ 939 \\ 1224 \\ 954 \\ 1194 \\ 1010 \end{bmatrix}$$

Implementado el algoritmo, los valores obtenidos son:

Las frecuencias de la tabla y la matriz de varianzas y covarianzas por estrato calculadas con base en la **Tabla 5-1** de contingencia, estás son:

p estimado del Estrato 1		
0.04774971	0.50050174	0.45174855
Matriz de Varianzas y Covarianzas Estrato 1		
5.268792e-05	-2.769271e-05	-2.499521e-05
-2.769271e-05	2.896868e-04	-2.619941e-04
-2.499521e-05	-2.619941e-04	2.869893e-04

Tabla 5-2: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 1

p estimado del Estrato 2		
0.05372835	0.54066775	0.40560390
Matriz de Varianzas y Covarianza Estrato 2		
5.502340e-05	-3.143852e-05	-2.358488e-05
-3.143852e-05	2.687729e-04	-2.373344e-04
-2.358488e-05	-2.373344e-04	2.609192e-04

Tabla 5-3: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 2

p estimado del Estrato 3		
0.05717929	0.52123719	0.42158352
Matriz de Varianzas y Covarianzas Estrato 3		
6.003320e-05	-3.318928e-05	-2.684393e-05
-3.318928e-05	2.778942e-04	-2.447049e-04
-2.684393e-05	-2.447049e-04	2.715488e-04

Tabla 5-4: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 3

p estimado del Estrato 4		
0.07023692	0.50017587	0.42958721
Matriz de Varianzas y Covarianzas Estrato 4		
9.491816e-05	-5.106223e-05	-4.385593e-05
-5.106223e-05	3.633720e-04	-3.123098e-04
-4.385593e-05	-3.123098e-04	3.561658e-04

Tabla 5-5: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 4

p estimado del Estrato 5		
0.04193478	0.36447464	0.59359058
Matriz de Varianzas y Covarianzas Estrato 5		
1.455661e-04	-5.537741e-05	-9.018874e-05
-5.537741e-05	8.392496e-04	-7.838721e-04
-9.018874e-05	-7.838721e-04	8.740609e-04

Tabla 5-6: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 5

p estimado del Estrato 6		
0.07719825	0.54561516	0.37718659
Matriz de varianzas y Covarianzas Estrato 6		
2.076929e-04	-0.0001228004	-8.489255e-05
-1.228004e-04	0.0007227967	-5.999963e-04
-8.489255e-05	-0.0005999963	6.848888e-04

Tabla 5-7: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 6

p estimado del Estrato 7		
0.0621509	0.5147928	0.4230563
Matriz de varianzas y Covarianzas Estrato 7		
6.563983e-05	-3.603022e-05	-2.960961e-05
-3.603022e-05	2.812851e-04	-2.452549e-04
-2.960961e-05	-2.452549e-04	2.748645e-04

Tabla 5-8: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 7

p estimado del Estrato 8		
0.06790541	0.55957883	0.37251577
Matriz de Varianzas y Covarianzas Estrato 8		
1.425546e-04	-8.558204e-05	-0.0000569726
-8.558204e-05	5.550684e-04	-0.0004694863
-5.697260e-05	-4.694863e-04	0.0005264589

Tabla 5-9: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 8

p estimado del Estrato 9		
0.04059848	0.50620833	0.45319318
Matriz de Varianzas y Covarianzas Estrato 9		
1.475388e-04	-0.0000778458	-6.969302e-05
-7.784580e-05	0.0009468237	-8.689779e-04
-6.969302e-05	-0.0008689779	9.386709e-04

Tabla 5-10: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 9

p estimado del Estrato 10		
0.02779459	0.50320541	0.46900000
Matriz de Varianzas y Covarianzas Estrato 10		
4.868839e-05	-0.0000252007	-2.348768e-05
-2.520070e-05	0.0004504319	-4.252312e-04
-2.348768e-05	-0.0004252312	4.487189e-04

Tabla 5-11: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 10

p estimado del Estrato 11		
0.01368053	0.51404595	0.47227352
Matriz de Varianzas y Covarianzas Estrato 11		
1.476299e-05	-7.694112e-06	-7.068873e-06
-7.694112e-06	2.733071e-04	-2.656130e-04
-7.068873e-06	-2.656130e-04	2.726819e-04

Tabla 5-12: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 11

p estimado del Estrato 12		
0.0379138	0.4392422	0.5228440
Matriz de Varianzas y Covarianzas Estrato 12		
8.662157e-06	-3.954723e-06	-4.707433e-06
-3.954723e-06	5.849169e-05	-5.453696e-05
-4.707433e-06	-5.453696e-05	5.924440e-05

Tabla 5-13: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 12

p estimado del Estrato 13		
0.02899052	0.46290587	0.50810360
Matriz de Varianzas y Covarianzas Estrato 13		
8.891369e-06	-4.238750e-06	-4.652618e-06
-4.238750e-06	7.852938e-05	-7.429063e-05
-4.652618e-06	-7.429063e-05	7.894325e-05

Tabla 5-14: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 13

p estimado del Estrato 14		
0.01890061	0.49572908	0.48537031
Matriz de Varianzas y Covarianzas Estrato 14		
2.972648e-06	-1.502017e-06	-1.470631e-06
-1.502017e-06	4.007402e-05	-3.857201e-05
-1.470631e-06	-3.857201e-05	4.004264e-05

Tabla 5-15: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 14

p estimado del Estrato 15		
0.0351971	0.5367435	0.4280594
Matriz de Varianzar y Covarianzas Estrato 15		
7.702032e-06	-4.284829e-06	-3.417203e-06
-4.284829e-06	5.639599e-05	-5.211116e-05
-3.417203e-06	-5.211116e-05	5.552836e-05

Tabla 5-16: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 15

p estimado del Estrato 16		
0.03188759	0.51906496	0.44904745
Matriz de Varianzas y Covarianzas Estrato 16		
1.200730e-05	-6.437858e-06	-5.569445e-06
-6.437858e-06	9.709705e-05	-9.065920e-05
-5.569445e-06	-9.065920e-05	9.622864e-05

Tabla 5-17: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 16

p estimado del Estrato 17		
0.02905835	0.50975553	0.46118612
Matriz de Varianzas y Covarianzas Estrato 17		
1.076047e-05	-5.649373e-06	-5.111102e-06
-5.649373e-06	9.531077e-05	-8.966139e-05
-5.111102e-06	-8.966139e-05	9.477250e-05

Tabla 5-18: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 17

p estimado del Estrato 18		
0.03900471	0.50421735	0.45677794
Matriz de Varianzas y Covarianzas Estrato 18		
1.102451e-05	-5.784368e-06	-5.240144e-06
-5.784368e-06	7.352418e-05	-6.773981e-05
-5.240144e-06	-6.773981e-05	7.297996e-05

Tabla 5-19: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 18

p estimado del Estrato 19		
0.01704138 0	.43755112	0.54540750
Matriz de Varianzas y Covarianzas Estrato 19		
8.155294e-06	-3.630222e-06	-4.525072e-06
-3.630222e-06	1.198151e-04	-1.161848e-04
-4.525072e-06	-1.161848e-04	1.207099e-04

Tabla 5-20: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 19

p estimado del Estrato 20		
0.03517164	0.50195753	0.46287083
Matriz de Varianzas y Covarianzas Estrato 20		
7.505994e-06	-3.905037e-06	-3.600957e-06
-3.905037e-06	5.529665e-05	-5.139162e-05
-3.600957e-06	-5.139162e-05	5.499257e-05

Tabla 5-21: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 20

p estimado del Estrato 21		
0.03409311	0.49693478	0.46897211
Matriz de Varianzas y Covarianzas Estrato 21		
7.589484e-06	-3.904598e-06	-3.684885e-06
-3.904598e-06	5.761480e-05	-5.371020e-05
-3.684885e-06	-5.371020e-05	5.739508e-05

Tabla 5-22: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 21

p estimado del Estrato 22		
0.04193302	0.45615509	0.50191190
Matriz de Varianzas y Covarianzas Estrato 22		
1.116583e-05	-5.316276e-06	-5.849550e-06
-5.316276e-06	6.894876e-05	-6.363248e-05
-5.849550e-06	-6.363248e-05	6.948203e-05

Tabla 5-23: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 22

p estimado del Estrato 23		
0.03318669	0.57549659	0.39131672
Matriz de Varianzas y Covarianzas Estrato 23		
9.933539e-06	-5.912949e-06	-4.020590e-06
-5.912949e-06	7.563476e-05	-6.972181e-05
-4.020590e-06	-6.972181e-05	7.374240e-05

Tabla 5-24: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 23

p estimado del Estrato 24		
0.01706982	0.47582284	0.50710734
Matriz de Varianzas y Covarianzas Estrato 24		
5.474206e-06	-2.649987e-06	-2.824219e-06
-2.649987e-06	8.137536e-05	-7.872537e-05
-2.824219e-06	-7.872537e-05	8.154959e-05

Tabla 5-25: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 24

p estimado del Estrato 25		
0.0311213	0.4968227	0.4720561
Matriz de Varianzas y Covarianzas Estrato 25		
8.667078e-06	-4.444313e-06	-4.222764e-06
-4.444313e-06	7.185683e-05	-6.741251e-05
-4.222764e-06	-6.741251e-05	7.163528e-05

Tabla 5-26: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 25

p estimado del Estrato 26		
0.02604171	0.43884991	0.53510838
Matriz de Varianzas y Covarianzas Estrato 26		
9.446385e-06	-4.256389e-06	-5.189996e-06
-4.256389e-06	9.171719e-05	-8.746081e-05
-5.189996e-06	-8.746081e-05	9.265080e-05

Tabla 5-27: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 26

p estimado del Estrato 27		
0.02302437	0.46785289	0.50912274
Matriz de Varianzas y Covarianzas Estrato 27		
1.015083e-05	-4.861019e-06	-5.289815e-06
-4.861019e-06	1.123495e-04	-1.074885e-04
-5.289815e-06	-1.074885e-04	1.127783e-04

Tabla 5-28: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 27

p estimado del Estrato 28		
0.01398051	0.45964339	0.52637610
Matriz de Varianzas y Covarianzas Estrato 28		
4.477122e-06	-2.087058e-06	-2.390064e-06
-2.087058e-06	8.066624e-05	-7.857918e-05
-2.390064e-06	-7.857918e-05	8.096924e-05

Tabla 5-29: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 28

p estimado del Estrato 29		
0.01891068	0.41610099	0.56498833
Matriz de Varianzas y Covarianzas Estrato 29		
4.707706e-06	-1.996639e-06	-2.711067e-06
-1.996639e-06	6.164957e-05	-5.965293e-05
-2.711067e-06	-5.965293e-05	6.236400e-05

Tabla 5-30: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 29

p estimado del Estrato 30		
0.0161079	0.4189470	0.5649451
Matriz de Varianzas y Covarianzas Estrato 30		
5.796794e-06	-2.468309e-06	-3.328486e-06
-2.468309e-06	8.903819e-05	-8.656988e-05
-3.328486e-06	-8.656988e-05	8.989836e-05

Tabla 5-31: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 30

p estimado del Estrato 31		
0.01505368	0.43865901	0.54628731
Matriz de Varianzas y Covarianzas Estrato 31		
4.600392e-06	-2.048846e-06	-2.551546e-06
-2.048846e-06	7.640003e-05	-7.435118e-05
-2.551546e-06	-7.435118e-05	7.690273e-05

Tabla 5-32: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 31

p estimado del Estrato 32		
0.01601912	0.43957937	0.54440151
Matriz de Varianzas y Covarianzas Estrato 32		
9.132391e-06	-4.079765e-06	-5.052626e-06
-4.079765e-06	1.427285e-04	-1.386487e-04
-5.052626e-06	-1.386487e-04	1.437013e-04

Tabla 5-33: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 32

p estimado del Estrato 33		
0.01897415	0.51079631	0.47022954
Matriz de Varianzas y Covarianzas Estrato 33		
5.727426e-06	-2.982132e-06	-2.745295e-06
-2.982132e-06	7.688721e-05	-7.390508e-05
-2.745295e-06	-7.390508e-05	7.665038e-05

Tabla 5-34: Frecuencias y matriz de varianzas y covarianzas para la Subpoblación 33

	f estimado	Intervalo de Confianza del 95 %	
i	f	Limite Inferior	Limite Superior
1	-0.161718207	-0.1861406798	-0.137295734
2	-0.095265367	-0.1221203297	-0.068410404
3	0.413197274	0.3822456679	0.444148880
4	-0.308274736	-0.3337662413	-0.282783230
5	-0.195176774	-0.2225732635	-0.167780284
6	-0.203259626	-0.2316345030	-0.174884749
7	-0.688292057	-0.7257308755	-0.650853239
8	0.906268137	0.8981466573	0.914389617
9	0.297584790	0.2678848679	0.327284712
10	-0.131666671	-0.1515220726	-0.111811270
11	-0.232037951	-0.2684127552	-0.195663147
12	-0.043737999	-0.0678681658	-0.019607832
13	-0.087237153	-0.1148972489	-0.059577056
14	0.018389491	-0.0102043854	0.046983368
15	0.134099700	0.1109595423	0.157239858
16	-0.146334506	-0.2165180052	-0.076151008
17	0.028616541	-0.0008452886	0.058078372
18	-0.129228151	-0.1603800965	-0.098076205
19	-0.043272843	-0.0651260057	-0.021419680
20	0.059992082	0.0282193629	0.091764800
21	0.094631645	0.0706354559	0.118627834
22	-0.154147600	-0.1783628609	-0.129932340
23	0.070538775	0.0499500087	0.091127542
24	0.130692090	0.1118398309	0.149544349
25	-0.069347539	-0.1235931767	-0.015101901
26	-0.106644203	-0.1347542105	-0.078534196
27	0.170022035	0.1435942365	0.196449833
28	-0.016534898	-0.0405127568	0.007442961
29	0.152616587	0.1296082187	0.175624956
30	-0.009735276	-0.0323148949	0.012844342
31	-0.075087211	-0.0990371100	-0.051137312
32	-0.072542982	-0.1007565692	-0.044329395
33	0.151675482	0.1285069253	0.174844039
34	-0.114170848	-0.1380883121	-0.090253383
35	0.510328442	0.4923721228	0.528284761
36	-0.092677465	-0.1349134977	-0.050441432
37	0.696635408	0.6843602111	0.708910604
38	0.419210523	0.4001125081	0.438308537
39	0.094859246	0.0702492029	0.119469289
40	-0.713379313	-0.7510952678	-0.675663359
41	0.244421757	0.2213632460	0.267480269
42	0.173544976	0.1413292465	0.205760705
43	-0.300504474	-0.3217895560	-0.279219392
44	-0.065117236	-0.0989383043	-0.031296167
45	-0.078450714	-0.1029788610	-0.053922568
46	0.316037014	0.2940984674	0.337975560
47	-0.081847586	-0.1118659127	-0.051829259
48	-0.021715053	-0.0468511163	0.003421011
49	-0.065857717	-0.1335328236	0.001817390
50	0.018284304	-0.0113337991	0.047902406
51	-0.139436409	-0.1707288452	-0.108143973
52	-0.232559755	-0.2563128027	-0.208806707
53	0.122857164	0.0921652631	0.153549064
54	0.100504664	0.0765864315	0.124422896
55	-0.003057701	-0.0256323716	0.019516970
56	0.108797113	0.0886365350	0.128957691
57	-0.322746026	-0.3460009447	-0.299491108
58	0.125740520	0.0766920392	0.174789000
59	-0.006465443	-0.0332729479	0.020342061
60	0.043660233	0.0152919010	0.072028566
61	-0.026999009	-0.0510999645	-0.002898053
62	0.165438244	0.1426046072	0.188271880
63	0.248010998	0.2285251824	0.267496813
64	-0.123840074	-0.1483269904	-0.099353157
65	-0.071659145	-0.0998611051	-0.043457185
66	0.004316269	-0.0207840177	0.029416555

Tabla 5-35: Valores estimados de $\widehat{f}$ e intervalos de confianza

La **Tabla 5-35** contiene la estimación de $\widehat{\boldsymbol{f}} = \boldsymbol{K}\ln(\boldsymbol{A}\widehat{\boldsymbol{\pi}})$ e intervalos de confianza del 95 % para los $\widehat{\pi}_i$, $i = 1, 2, ..., 66$, donde las componentes del vector $\widehat{\pi}$ se evidencian en la primera fila de las **Tabla 5-2** hasta la **Tabla 5-34**.

Matriz D estimada

$$\widehat{D} = \begin{bmatrix} 20.931462 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1.9997993 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 2.213621 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 62.425402 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1.957728 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 2.126621 \end{bmatrix}_{99\times 99}$$

Matriz de varianzas y covarianzas de $\widehat{\boldsymbol{f}}$:

$$\widehat{\Sigma_{\widehat{f}}} = \begin{bmatrix} 0.017781413 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0.010188726 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0.014748341 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0.00578119 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0.0197583 \end{bmatrix}_{66\times 66}$$

Matriz de varianzas y covarianzas de $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\Sigma_{\widehat{\beta}}} = \begin{bmatrix} 0.021832358 & -0.0218323587 & \dots & -0.0007797271 & -0.0008797271 \\ -0.0218323587 & 0.0284330454 & \dots & 0.0007797271 & 0.0007797271 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -0.0007797271 & 0.0007797271 & \dots & 0.0032691862 & 0.0015096541 \\ -0.0008797271 & 0.0007797271 & \dots & 0.0015096541 & 0.0030593089 \end{bmatrix}_{66\times 66}$$

Estimación β			**I.C. del 95 %**			
$\widehat{\boldsymbol{\beta}}$	Estimación	Error	Límite Inferior	Límite Superior	Valor chi	p-valor
β_0	-3.164448178	0.23735450	-3.629654452	-2.699241905	1.777465e+02	1.504769e-40
β_1	0.978440285	0.25480747	0.479026824	1.477853745	1.474499e+01	1.230744e-04
β_2	1.262774263	0.25666139	0.759727173	1.765821353	2.420641e+01	8.654387e-07
β_3	1.142315122	0.25306200	0.646322717	1.638307526	2.037596e+01	6.362421e-06
β_4	1.364525398	0.30748024	0.761875200	1.967175596	1.969377e+01	9.089730e-06
β_5	0.570203427	0.25830361	0.063937650	1.076469203	4.873026e+00	2.727963e-02
β_6	1.562634191	0.24780417	1.076946935	2.048321447	3.976467e+01	2.864803e-10
β_7	1.271579642	0.24695385	0.787558983	1.755600301	2.651280e+01	2.617983e-07
β_8	1.486338731	0.26577259	0.965434023	2.007243439	3.127629e+01	2.237964e-08
β_9	0.798756046	0.25306238	0.302762890	1.294749202	9.962610e+00	1.597514e-03
β_{10}	0.499560442	0.25870317	-0.007488449	1.006609334	3.728830e+00	5.348084e-02
β_{11}	-0.345852705	0.29145352	-0.917091116	0.225385706	1.408133e+00	2.353665e-01
β_{12}	0.652142554	0.48633048	-0.301047678	1.605332787	1.798134e+00	1.799382e-01
β_{13}	0.026348663	0.31316412	-0.587441728	0.640139055	7.079021e-03	9.329476e-01
β_{14}	-0.076406553	0.26782771	-0.601339226	0.448526119	8.138609e-02	7.754277e-01
β_{15}	0.704620288	0.25283931	0.209064349	1.200176227	7.766424e+00	5.322632e-03
β_{16}	0.651483576	0.28758249	0.087832256	1.215134895	5.131946e+00	2.348938e-02
β_{17}	0.285781293	0.26572592	-0.235031932	0.806594518	1.156644e+00	2.821626e-01
β_{18}	0.677616600	0.26241073	0.163301020	1.191932180	6.668143e+00	9.815141e-03
β_{19}	-0.567849896	0.32792384	-1.210568819	0.074869027	2.998618e+00	8.333559e-02
β_{20}	0.593019317	0.29437905	0.016046979	1.169991654	4.058111e+00	4.395966e-02
β_{21}	0.543967234	0.26362844	0.027264985	1.060669484	4.257562e+00	3.907597e-02
β_{22}	0.832932409	0.25933387	0.324647355	1.341217462	1.031576e+01	1.318993e-03
β_{23}	0.583963216	0.25662611	0.080985292	1.086941140	5.178087e+00	2.287350e-02
β_{24}	-0.331600141	0.29590351	-0.911560355	0.248360074	1.255825e+00	2.624428e-01
β_{25}	0.460223199	0.25639297	-0.042297786	0.962744183	3.221994e+00	7.265505e-02
β_{26}	0.319085862	0.28951791	-0.248358811	0.886530534	1.214687e+00	2.704059e-01
β_{27}	0.198754414	0.31505005	-0.418732341	0.816241168	3.979921e-01	5.281280e-01
β_{28}	-0.231178158	0.30759659	-0.834056393	0.371700076	5.648467e-01	4.523140e-01
β_{29}	-0.087305844	0.29119558	-0.658038689	0.483427002	8.989118e-02	7.643155e-01
β_{30}	-0.436833217	0.29912456	-1.023106590	0.149440157	2.132687e+00	1.441878e-01
β_{31}	-0.369408122	0.32281041	-1.002104898	0.263288653	1.309537e+00	2.524785e-01
β_{32}	-0.233147955	0.27676878	-0.775604787	0.309308876	7.096252e-01	3.995683e-01
α_0	0.117275550	0.06567771	-0.011450400	0.246001501	1.777465e+02	1.504769e-40
α_1	-0.024136924	0.07727713	-0.175597314	0.127323466	1.474499e+01	1.230744e-04
α_2	0.173654031	0.08049270	0.015891247	0.331416816	2.420641e+01	8.654387e-07
α_3	0.070210495	0.07720495	-0.081108427	0.221529418	2.037596e+01	6.362421e-06
α_4	0.103154295	0.11866525	-0.129425326	0.335733916	1.969377e+01	9.089730e-06
α_5	0.258267878	0.07581566	0.109671917	0.406863839	3.976467e+01	2.864803e-10
α_6	0.111088677	0.07352947	-0.033026442	0.255203796	2.651280e+01	2.617983e-07
α_7	0.350334891	0.08929011	0.175329490	0.525340293	3.127629e+01	2.237964e-08
α_8	0.052571832	0.07438144	-0.093213116	0.198356780	9.962610e+00	1.597514e-03
α_9	-0.030838655	0.07506566	-0.177964641	0.116287331	3.728830e+00	5.348084e-02
α_{10}	-0.022305064	0.07680354	-0.172837236	0.128227109	1.408133e+00	2.353665e-01
α_{11}	-0.187234139	0.17977716	-0.539590907	0.165122630	1.798134e+00	1.799382e-01
α_{12}	-0.183608105	0.08889171	-0.357832647	-0.009383563	7.079021e-03	9.329476e-01
α_{13}	-0.104888815	0.07393136	-0.249791615	0.040013984	8.138609e-02	7.754277e-01
α_{14}	0.097527369	0.07343811	-0.046408679	0.241463417	7.766424e+00	5.322632e-03
α_{15}	0.007073497	0.08963842	-0.168614587	0.182761581	5.131946e+00	2.348938e-02
α_{16}	-0.023105405	0.07592414	-0.171913992	0.125703182	1.156644e+00	2.821626e-01
α_{17}	-0.028542899	0.07845377	-0.182309472	0.125223674	6.668143e+00	9.815141e-03
α_{18}	-0.328373451	0.08359813	-0.492222766	-0.164524136	2.998618e+00	8.333559e-02
α_{19}	-0.012130343	0.09167551	-0.191811038	0.167550352	4.058111e+00	4.395966e-02
α_{20}	-0.083478171	0.07790766	-0.236174387	0.069218046	4.257562e+00	3.907597e-02
α_{21}	-0.200384013	0.07956527	-0.356329081	-0.044438944	1.031576e+01	1.318993e-03
α_{22}	0.239667454	0.07384339	0.094937068	0.384397840	5.178087e+00	2.287350e-02
α_{23}	-0.149958198	0.07862235	-0.304055176	0.004138781	1.255825e+00	2.624428e-01
α_{24}	-0.040412784	0.07381860	-0.185094573	0.104269004	3.221994e+00	7.265505e-02
α_{25}	-0.243395125	0.08680625	-0.413532244	-0.073258007	1.214687e+00	2.704059e-01
α_{26}	-0.128909738	0.09243523	-0.310079450	0.052259975	3.979921e-01	5.281280e-01
α_{27}	-0.196025901	0.08305546	-0.358811619	-0.033240182	5.648467e-01	4.523140e-01
α_{28}	-0.376727834	0.08214973	-0.537738341	-0.215717327	8.989118e-02	7.643155e-01
α_{29}	-0.419142258	0.07990194	-0.575747179	-0.262537336	2.132687e+00	1.441878e-01
α_{30}	-0.345864997	0.08589424	-0.514214605	-0.177515388	1.309537e+00	2.524785e-01
α_{31}	-0.314599206	0.07595101	-0.463460448	-0.165737963	7.096252e-01	3.995683e-01
α_{32}	-0.263962204	0.05547960	-0.37270021	-0.15522419	4.61472641	3.169855e-02

Tabla 5-36: Estimación de parámetros del Modelo, $\widehat{\boldsymbol{\beta}}$ e Intervalos de confianza

La **Tabla 5-36** contiene las estimaciones de los parámetros del modelo vía GSK, con la prueba χ^2 para $H_0 : \beta_i = 0$, $i = 1, 2, ..., 66$ e intervalo de confianza del 95 % construido para β_i.
El signo de los parámetros β_i representan una influencia positiva o negativa en la variable dependiente, en este caso, es la influencia sobre $f_1 = \ln\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right)$ y $f_2 = \ln\left(\frac{\pi_{12}^i}{\pi_{13}^i}\right)$.
Así, para f_1 la **Tabla 5-36** muestra que los parámetros β_{10}, β_{11}, β_{12}, β_{13}, β_{17}, β_{19}, β_{24}, β_{26}, β_{27}, β_{28}, β_{29}, β_{30}, β_{31}, β_{32} no son significativos, no siendo el caso para el resto.
Y para f_2, se muestra que los parámetros $\alpha_9, \alpha_{10}, \alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{16}, \alpha_{18}, \alpha_{23}, \alpha_{24}, \alpha_{25}, \alpha_{26}, \alpha_{27}, \alpha_{28}, \alpha_{29}, \alpha_{30}, \alpha_{31}$ no son significativos, no obstante los demás sí.

	f^* estimado	Intervalo de Confianza del 95%	
i	$\widehat{f^*}$	Limite Inferior	Limite Superior
1	-2.23379624	-2.52837626	-1.939216223
2	-2.01980542	-2.21794593	-1.821664903
3	-1.99462910	-2.16914155	-1.820116653
4	-1.81177816	-1.94776423	-1.675792094
5	-2.65183496	-2.85070572	-2.452964209
6	-1.60143723	-1.72828612	-1.474588336
7	-1.91976470	-2.11697230	-1.722557108
8	-1.70412740	-1.85133892	-1.556915888
9	-2.38788611	-2.62135016	-2.154422057
10	-2.80796137	-3.09095718	-2.524965565
11	-3.51938904	-3.94407386	-3.094704221
12	-2.61265001	-2.79361169	-2.431688324
13	-2.84726201	-3.10908079	-2.585443224
14	-3.21960013	-3.65036145	-2.788838813
15	-2.53966666	-2.78659935	-2.292733974
16	-2.65505172	-2.86709854	-2.443004895
17	-2.76832497	-3.01739484	-2.519255107
18	-2.44774596	-2.66069335	-2.234798573
19	-3.44246311	-3.77325190	-3.111674317
20	-2.58175798	-2.78247894	-2.381037009
21	-2.60972619	-2.82043160	-2.399020776
22	-2.49184625	-2.71596191	-2.267730586
23	-2.47230817	-2.71505961	-2.229556726
24	-3.40671305	-3.70342690	-3.109999197
25	-2.71885301	-2.96097452	-2.476731499
26	-3.02702665	-3.39472245	-2.659330846
27	-3.08967892	-3.34547867	-2.833879179
28	-3.63129765	-3.94410086	-3.318494450
29	-3.38764706	-3.67922786	-3.096066261
30	-3.56417598	-3.98206984	-3.146282108
31	-3.57112676	-3.99843632	-3.143817196
32	-3.52189424	-3.83404248	-3.209745992
33	-3.21512175	-3.47505121	-2.955192283
34	0.10755037	-0.01868898	0.233789715
35	0.28328483	0.19348965	0.373080003
36	0.21478727	0.13360870	0.295965837
37	0.15188456	0.08240389	0.221365227
38	-0.48856279	-0.57133987	-0.405785710
39	0.37348877	0.30597594	0.441001610
40	0.20530450	0.11030865	0.300300354
41	0.40054096	0.32590435	0.475177558
42	0.10524953	0.01192214	0.198576923
43	0.06544552	-0.02849316	0.159384193
44	0.08313794	-0.01667322	0.182949105
45	-0.17770270	-0.24777608	-0.107629315
46	-0.08433356	-0.17291580	0.004248685
47	0.02056170	-0.09825920	0.139382587
48	0.22735266	0.13784950	0.316855817
49	0.14365829	0.06942136	0.217895227
50	0.10427168	0.02081582	0.187727543
51	0.09800684	0.01504600	0.180967688
52	-0.22432728	-0.31171022	-0.136944335
53	0.08458689	0.01085375	0.158320025
54	0.05647279	-0.02044714	0.133392718
55	-0.09985258	-0.18974768	-0.009957477
56	0.38954910	0.30184344	0.477254762
57	-0.06913575	-0.14563353	0.007362028
58	0.05411145	-0.02992987	0.138152764
59	-0.19373432	-0.31135102	-0.076117627
60	-0.08462239	-0.16175117	-0.007493604
61	-0.13028862	-0.20377226	-0.056804986
62	-0.31108191	-0.39214655	-0.230017280
63	-0.30686190	-0.41338134	-0.200342454
64	-0.21342613	-0.31915194	-0.107700323
65	-0.21015248	-0.28918349	-0.131121473
66	0.08266715	0.01189252	0.153441772

Tabla 5-37: Valores estimados de $\widehat{f^*}$ e intervalos de confianzas

La **Tabla 5-37** contiene la estimación de $\widehat{f^*} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ e intervalos de confianza del 95 % para

f_i^*, $i = 1, 2, ..., 66$, donde $\boldsymbol{X}$ es la matriz de diseño definida en el capítulo 4.

	f observado	Intervalo de Confianza del 95 %	
i	Valor observado	Límite Inferior	Límite Superior
1	-2.24248117	-2.403550943	-2.081411395
2	-2.01736911	-2.278363035	-1.756375191
3	-2.00195405	-2.180077217	-1.823830875
4	-1.81296168	-2.006135996	-1.619787358
5	-2.64752478	-3.173622395	-2.121427166
6	-1.58843977	-1.774309587	-1.402569945
7	-1.92023779	-2.233246626	-1.607228962
8	-1.70207071	-1.837808053	-1.566333375
9	-2.40232006	-2.508543725	-2.296096393
10	-2.82052818	-2.940198373	-2.700857996
11	-3.52003806	-3.684254065	-3.355822053
12	-2.62199530	-2.740754625	-2.503235984
13	-2.86121517	-3.029873412	-2.692556935
14	-3.24176964	-3.389685862	-3.093853428
15	-2.50377513	-2.661069992	-2.346480276
16	-2.64351168	-2.752771028	-2.534252332
17	-2.76610221	-2.906519530	-2.625684896
18	-2.46112174	-2.645601200	-2.276642290
19	-3.46389598	-3.573645039	-3.354146913
20	-2.58237899	-2.811937353	-2.352820632
21	-2.62424224	-2.820948456	-2.427536031
22	-2.48093050	-2.592853994	-2.369007009
23	-2.47475428	-2.712383923	-2.237124633
24	-3.39529766	-3.798233664	-2.992361655
25	-2.72299178	-2.834107108	-2.611876454
26	-3.02603762	-3.153183491	-2.898891755
27	-3.09695380	-3.254427126	-2.939480475
28	-3.62624388	-3.767800667	-3.484687099
29	-3.39238675	-3.540713662	-3.244059842
30	-3.56600536	-3.704667851	-3.427342861
31	-3.59456877	-3.987540664	-3.201596885
32	-3.52636052	-3.704731412	-3.347989637
33	-3.20829372	-3.461173850	-2.955413581
34	0.10092592	-0.058871326	0.260723163
35	0.28706612	0.026974978	0.547157261
36	0.21266227	0.046195913	0.379128635
37	0.15315118	0.022793408	0.283508951
38	-0.48804053	-1.019155585	0.043074522
39	0.37037379	0.182139643	0.558607934
40	0.19679472	-0.025100587	0.418690030
41	0.40457105	0.241485103	0.567657004
42	0.11064454	-0.004770509	0.226059596
43	0.06586742	-0.057768071	0.189502921
44	0.08118046	-0.048484788	0.210845701
45	-0.17508205	-0.290876168	-0.059287934
46	-0.09078395	-0.306826932	0.125259033
47	0.02036730	-0.127431127	0.168165733
48	0.22501097	0.115085573	0.334936358
49	0.14265630	0.031223148	0.254089453
50	0.09904997	-0.040029901	0.238129848
51	0.09987504	-0.078815387	0.278565464
52	-0.22176329	-0.335672108	-0.107854470
53	0.08087307	-0.141727636	0.303473767
54	0.05798726	-0.079282253	0.195256768
55	-0.09610731	-0.211660775	0.019446155
56	0.38484582	0.155333515	0.614358127
57	-0.06309315	-0.829194390	0.703008091
58	0.05161104	-0.082298649	0.185520730
59	-0.19735943	-0.326719643	-0.067999225
60	-0.08611876	-0.212630585	0.040393067
61	-0.13407472	-0.275498345	0.007348898
62	-0.30614047	-0.463867676	-0.148413266
63	-0.30072316	-0.436033214	-0.165413102
64	-0.22040007	-0.381514357	-0.059285774
65	-0.21444983	-0.450909510	0.022009842
66	0.08363690	-0.161706614	0.328980405

Tabla 5-38: Valores observados e intervalos de confianza

La **Tabla 5-38** contiene los valores observados al evaluar las funciones $f_1 = \ln\left(\frac{\pi_{11}^i}{\pi_{13}^i}\right)$ y $f_2 = \ln\left(\frac{\pi_{12}^i}{\pi_{13}^i}\right)$, $i = 1, 2, ..., 33$ sobre los datos de la **Tabla 4-1**.Datos tomados de la encuesta de ENSIN 2005 [15] sobre el índice de masa corporal para cada uno de los 33 departamentos de Colombia e intervalos de confianza para f_1 y f_2, descrito en la ecuación (4-24).

Residuales crudos	
i	$f_{obs} - f^*$
1	0.007549198
2	-0.122842129
3	0.001435132
4	0.078453443
5	-0.001706638
6	-0.110795626
7	0.194510562
8	0.118315476
9	0.088636216
10	-0.051757718
11	0.302290548
12	0.109602185
13	-0.037728805
14	-0.082972816
15	-0.010053286
16	0.004928613
17	-0.170745978
18	0.074190460
19	-0.043921740
20	0.112248188
21	0.013794157
22	-0.138989276
23	0.075194798
24	0.363062352
25	0.011203700
26	0.031585439
27	0.122830702
28	-0.009225291
29	-0.028477312
30	-0.102624374
31	0.294697244
32	-0.375813340
33	0.114843120
34	-0.002697526
35	0.010833404
36	0.003236359
37	0.055953974
38	-0.004710287
39	-0.036413197
40	0.038987731
41	0.015185451
42	-0.032938693
43	-0.003584408
44	0.018694992
45	-0.007322499
46	0.026341408
47	-0.001527063
48	0.011099366
49	-0.017327090
50	0.044234969
51	-0.049758590
52	0.039959744
53	0.007086896
54	0.048836202
55	-0.002488320
56	-0.001118881
57	0.018040407
58	-0.022558374
59	0.027955671
60	0.038499573
61	0.057514174
62	0.003028142
63	-0.008201731
64	0.071553692
65	-0.056140452
66	0.009246078

Tabla 5-39: Residuales Crudos

La **Tabla 5-39** contiene la diferencia entre la función estimada y la función observada de la **Tabla 4-1**. Nótese que en la tabla se evidencia que los valores de la función estimada por el modelo en promedio es esencialmente los mismos que la función observada, sin considerar la significancia de los parámetros del modelo.

Diferencia al cuadrado entre $L_{f_{obs}}$ y L_{f^*}	
i	$\left(L_{f_{obs}} - L_{f^*}\right)^2$
1	8.092159e-03
2	4.056521e-03
3	3.056194e-04
4	1.265431e-02
5	5.027273e-03
6	2.903942e-02
7	2.054379e-05
8	1.695781e-04
9	2.263982e-03
10	1.015326e-02
11	1.976005e-02
12	1.832490e-02
13	5.078909e-03
14	2.786685e-02
15	4.656120e-03
16	5.718528e-03
17	8.141145e-04
18	8.107573e-03
19	3.592264e-03
20	3.054755e-04
21	1.294494e-02
22	2.479232e-02
23	2.452676e-02
24	2.437888e-02
25	3.591639e-03
26	3.457323e-03
27	1.605082e-02
28	9.130469e-03
29	2.009645e-02
30	9.301816e-04
31	4.459845e-05
32	2.076858e-02
33	2.889755e-02
34	6.687099e-04
35	3.156741e-05
36	2.607661e-04
37	1.259118e-02
38	7.006920e-05
39	7.343096e-03
40	4.193867e-03
41	1.706372e-04
42	4.021046e-02
43	1.135888e-05
44	6.027305e-05
45	4.359459e-02
46	4.407427e-06
47	2.895095e-06
48	1.804500e-05
49	9.410325e-06
50	3.067802e-06
51	2.516275e-05
52	4.877401e-03
53	7.944346e-06
54	2.592569e-04
55	9.971621e-03
56	1.063617e-04
57	2.450042e-05
58	7.256094e-03
59	3.097684e-05
60	5.545547e-06
61	2.906728e-04
62	1.099320e-04
63	4.914288e-08
64	6.016729e-06
65	9.080423e-06
66	7.238778e-05

Tabla 5-40: Diferencia al cuadrado entre las longitudes de los intervalos de confianza de f_{obs} y f^*

En la **Tabla 5-40** se evidencia el comportamiento de las longitudes de los intervalos de confianza para la función observada de la **Tabla 4-1** y la función estimada por el modelo dado los tamaños de muestra obtenidos mediante el algoritmo genético **Tabla 5-45**, y se

aprecia un comportamiento muy similar. Esta tabla se puede analizar también como los valores del criterio de optimización descrito en la sección 3.4 y se observa que los tamaños de muestra satisfacen el criterio establecido, es decir, $criterio = \left(L_{f_{obs}} - L_{f^*}\right)^2 \leq 0.05$, donde $L_{f_{obs}}$: longitud media de los intervalos de confianza para las funciones observadas de la **Tabla 4-1** y L_{f^*}: longitud media de los intervalos de confianza para las funciones estimada por el modelo.

f	Valor-chi	Grados de libertad	p-valor
f_1	780.6314	32	1.818712e-143
f_2	907.0679	32	6.024709e-170

Tabla 5-41: Prueba Global para f_1 y f_2

La **Tabla 5-41** contiene la prueba de los efectos específicos de los departamentos sobre $\widehat{f_1}$ y $\widehat{f_2}$ y se obtiene que $X^2 = 780.6314$ y $X^2 = 907.0679$ con 32 grados de libertad respectivamente, los cuales no superan el nivel de significancia de 0.05.

Dado que las funciones $\widehat{f_1}$ y $\widehat{f_2}$ son funciones de las probabilidades de la tabla y utilizando la equivalencia entre $\ln\left(\frac{\pi^i_{1j}}{\pi^i_{13}}\right)$ y la función lineal $x_i^T\boldsymbol{\beta}$, $i = 1, 2, ..., 33$, $j = 1, 2$, se empleó las ecuaciones (4-27), (4-28) y (4-29) para las estimaciones de$\widehat{\pi^1_{11}}^*$, $\widehat{\pi^1_{12}}^*$ y $\widehat{\pi^1_{13}}^*$ de la 1-subpoblación de tablas, las cuales son:

Estimación probabilidades 1- Subpoblación			
i	$\widehat{\pi^i_{11}}^*$	$\widehat{\pi^i_{12}}^*$	$\widehat{\pi^1_{13}}^*$
1	0.04345586	0.5012023	0.4553418

Tabla 5-42: Estimación probabilidades 1-subpoblación

Probabilidades Observadas 1- Subpoblación			
i	π^i_{11}	π^i_{12}	π^1_{13}
1	0.048	0.500	0.452

Tabla 5-43: Probabilidades Observadas 1-subpoblación

Residuales Pseudo Estandarizados Subpoblación 1			
i	π_{11}^{i}	π_{12}^{i}	π_{13}^{1}
1	0.0953324	-0.07467344	-0.1077277

Tabla 5-44: Residuales Pseudo Estandarizados Subpoblación 1

La **Tabla 5-44** corresponde a los residuales pseudo estandarizados de la subpoblación 1, encontrados utilizando las ecuaciones 4-55, 4-56 y 4-57 y los valores de la **Tabla 5-42** y **5-43**.

Subpoblación	n	$n_{algoritmo}$
Subpoblación 1	n_1	1991
Subpoblación 2	n_2	1345
Subpoblación 3	n_3	1531
Subpoblación 4	n_4	1931
Subpoblación 5	n_5	1256
Subpoblación 6	n_6	1336
Subpoblación 7	n_7	2951
Subpoblación 8	n_8	2348
Subpoblación 9	n_9	2580
Subpoblación 10	n_{10}	3189
Subpoblación 11	n_{11}	2479
Subpoblación 12	n_{12}	1313
Subpoblación 13	n_{13}	3483
Subpoblación 14	n_{14}	2854
Subpoblación 15	n_{15}	818
Subpoblación 16	n_{16}	3779
Subpoblación 17	n_{17}	3312
Subpoblación 18	n_{18}	1724
Subpoblación 19	n_{19}	897
Subpoblación 20	n_{20}	2652
Subpoblación 21	n_{21}	1399
Subpoblación 22	n_{22}	3695
Subpoblación 23	n_{23}	2883
Subpoblación 24	n_{24}	2583
Subpoblación 25	n_{25}	2696
Subpoblación 26	n_{26}	2107
Subpoblación 27	n_{27}	1632
Subpoblación 28	n_{28}	1979
Subpoblación 29	n_{29}	3292
Subpoblación 30	n_{30}	3319
Subpoblación 31	n_{31}	3322
Subpoblación 32	n_{32}	1227
Subpoblación 33	n_{33}	3208

Tabla 5-45: Tamaño de muestra subpoblación

Finalmente la **Tabla 5-45** contiene los tamaños de muestras óptimos estimados vía simulación utilizando algoritmos genéticos (descrito en la sección 3.4) para las funciones $f_1 = \ln\left(\frac{\pi_{11}^{i}}{\pi_{13}^{i}}\right)$ y $f_2 = \ln\left(\frac{\pi_{12}^{i}}{\pi_{13}^{i}}\right)$, π_{ij} probabilidades de una tabla de contingencia con $i = 1, 2, ..., 33$ y $j = 1, 2, 3$.

6 Conclusiones y recomendaciones

6.1. Conclusiones

- Para la comparación de las longitudes de los intervalos de confianza, se utilizó la diferencia al cuadrado entre la longitud media de los intervalos de confianza de la función observada y la longitud media de los intervalos de confianza de la función estimada, sin tener en cuenta la significancia de los parámetros del modelo. Se evidencia en la (tabla 5-40) que poseen esencialmente las mismas longitudes dado los tamaños de muestrales de cada subpoblación de la tabla de contingencia obtenidos mediante el algoritmo genético y además tienen un comportamiento relativo similar.
- Sin considerar la importancia de la significancia de los parámetros del modelo, la estimación de la función respuesta (tabla 5-37), redondeando al tercer decimal, en casi todos los casos coincide con la función observada (tabla 5-38).
- Dada la robustez de los algoritmos genéticos es poco frecuente encontrar estimaciones sesgadas debido a la influencia de mínimos locales.
- Los tamaños de muestras resultantes usando el algoritmo genético podría usarse como valores iniciales o valores de referencia para el uso de otras metodologías de estimación.
- La metodología de algoritmos genéticos posee muchas aplicación, se observa que una de ellas es la optimización de problemas.

6.2. Recomendaciones

- El código desarrollado en el software estadístico R puede mejorarse en su parte algorítmica y se puede crear una librería en R.
- Realizar diagnósticos del modelo.
- Realizar estudios de potencia, y de efectos de ceros muestrales en las tablas.

7 Anexos : Programa en R para determinar el tamaño de muestra de funciones de una tabla de contingencia vía algoritmos genético

7.1. Funciones Auxiliares

Dentro de la función principal hay una serie de funciones auxiliares que permiten generar los elementos necesarios para calcular los resultados de interés, estas funciones se describen a continuación:

7.1.1. La función crea.bloque

```
#############################################################################
Permite crear una matriz en bloque a partir de dos matrices
#############################################################################
crea.bloque<-function(mat1,mat2){
nf1<-nrow(mat1);nc1<-ncol(mat1)
nf2<-nrow(mat2);nc2<-ncol(mat2)
mat1<-cbind(mat1,matrix(0,nrow=nf1,ncol=nc2))
mat2<-cbind(matrix(0,nrow=nf2,ncol=nc1),mat2)
matdef<-rbind(mat1,mat2)
return(matdef) }
```

7.1.2. La función ceros.a.5

```
#############################################################################
corrige los ceros en 0,5. Esta correccion es comunmente
utilizada en analisis de tablas de contingencia (Agresti, 1996b) (Upton, 1978)
(Hanley, 1983)}
#############################################################################
ceros.a.5<-function(Tabla){
```

```
Nro.col<-ncol(Tabla)
if(any(Tabla==0))
Tabla<-matrix(ifelse
(as.vector(Tabla)==0,0.5,Tabla),ncol=Nro.col)
return(Tabla)
}
```

7.1.3. Función identidad

```
#############################################################################
La funcion identidad permite construir una matriz diagonal con k elementos
iguales a 1 en su diagonal principal (tambien puede utilizarse la funcion
diag existente en R).
#############################################################################
identidad<-function(k)diag(rep(1,k))
```

7.1.4. La función bloque

```
#############################################################################
La funcion bloque permite construir una matriz donde su primera fila esta
compuesta de ceros y debajo se adiciona una matriz diagonal con k elementos
iguales a 1 en su diagonal principal.
#############################################################################
bloque<-function(k)rbind(0,identidad(k))
```

7.1.5. La función repita

```
#############################################################################
La funcion repita toma dos matrices apartir de las cuales se genera una nueva
matriz de mayor dimension. La primera matriz aparece en las primeras columnas
y la segunda se adiciona ocupando las ultimas columnas de la nueva matriz
generada. Cada fila aparece repetida k veces una debajo de la otra.
#############################################################################
repita<-function(X,X2,k){
temp<-NULL
for(i in 1:nrow(X))temp<-rbind(temp,cbind
(matrix(rep(X[i,],k),ncol=ncol(X),byrow=T),X2))
return(temp)
}
```

7.1.6. Función para calcular la tabla de probabilidades estimada

```
##############################################################################
La funcion pi.est permite calcular las probabilidades de una muestra multinomial
M (ni, pi1, pi2, pi3, ...pik)
# ni.s = vector de tamanos muestrales de estratos de una tabla de contingencia
#pis. obs = tabla de probabilidades observadas
# Nsim = Numero de simulaciones
##############################################################################
pi.est <- function(N, P, Nsim){
aux <- NULL
res<- NULL
n.i <- rowSums(N)
piest<- NULL
muestra<- NULL
for(i in 1:nrow(N)){
muestra <- rmultinom(Nsim, n.i[i], P[i,])
res <- apply(muestra, 1, mean)
res <- res/n.i[i]
print("========================")
print(paste("pi estimados por estrato", i))
print(res)
print("========================")
aux <- matrix(res, ncol=1,byrow = T)
temp1<- aux%*%t(aux)
temp2<-diag(as.vector(aux))
varcov.p<-(temp2-temp1)/n.i[i]
print("========================")
print(paste("Matriz de varianzas y covarianzas estimada del estrato",i))
print(varcov.p)
print("========================")
#varcov.grande<-varcov.p
if(i==1){
varcov.grande<-varcov.p
piest<-res
}
else{
varcov.grande<-crea.bloque(varcov.grande,varcov.p)
piest<-rbind(piest,res)
}
```

```
}#fin for i
return(piest)
}
```

7.1.7. Función crea intervalos de confianza para la función de interés

```
I.Cfuncioninteres <- function(f,N){
N1 <- rowSums(N)
k <- nrow(N)
L.I <- NULL
L.S <- NULL
temp2<- NULL
temp1 <- NULL
L.I <- f- (1.64)*sqrt((f*(1-f))/N1)
L.S <- f+ (1.64)*sqrt((f*(1-f))/N1)
I.C <- rbind(L.I,L.S)
for(i in 1:k)
{
# temp1[i]<- ifelse(f[i]>I.C[1,i] & f[i]<I.C[2,i],1,0)
temp2[i]<-I.C[2,i]-I.C[1,i]
temp2[i] <-abs( temp2[i])/2
}#Fin for
#level<-ifelse(sum(temp1)==k,1,0) # Para saber si todas las pis caen en los IC
# print("====================================")
#print("Nivel")
#print(level)
return(temp2)
}
```

7.2. Algoritmo genético

7.2.1. Función que genera una población al azar de tamaño N

```
##############################################################################
# N= Tamano de la poblacion de cromosomas
# Nbits = Numero de bits de cada gen
# Nvar = Numero de variables del cromosoma.
##############################################################################
generar.poblacion<-function(N,Nbits,Npar){
# esta funcion genera una poblacion al azar de cromosomas de tamano N
```

```
temp<-c(0,1)
temp<-matrix(sample(temp,N*Nbits*Npar,replace=T),nrow=N)
temp
}
```

7.2.2. Función que decodifica binarios a enteros

```
#############################################################################
La funcion parametros decodifica cromosomas binarios a enteros
# Cromosoma= Matrix de poblacion de cromosomas binarios
#Nbits = Numero de bits de un gen
#Nvar= Numero de variables del cromosoma.
minimos = Vector establecido por el investigador
maximos = vector establecido por el investigador
#############################################################################
parametros<-function(cromosoma,Nbits,minimos,maximos){
# Esta funcion entrega un vector de parametros
Npar<-length(cromosoma)/Nbits
dos.a.la.i<-2^seq(0,(Nbits-1))
maximo<-sum(dos.a.la.i)
minimo<-0
parametro<-rep(NA,Npar)
genes<-matrix(cromosoma,nrow=Nbits)
parametros<-dos.a.la.i%*%genes/maximo
parametros<-(maximos-minimos)*parametros
}
```

7.2.3. Función costo

```
#############################################################################
La funcion COSTO, calcula el volumen de los IC y la suma de los nis
#############################################################################
costo <- function(P,Nbits, minimos, maximos, pis, errores){
nis <- NULL
Lis <- NULL
res <- NULL
nis <- parametros(P,Nbits, minimos, maximos)
Lis<-IC.F(nis, pis, errores)
Ni<- matrix(nis, ncol=3, byrow = T)
Ni<- round(apply(Ni,1, sum),0)
```

```
Volumen<- matrix(Lis, ncol=3, byrow=T )
Volumen<-apply(Volumen, 1,prod)
res <- cbind(Volumen,Ni)
return(res)
}
```

7.2.4. Función que ordena la población según la evaluación del costo

```
#############################################################################
La funcion oredena.poblacionXcosto ordena la poblacion de cromosomas de
acuerdo al costo.
#############################################################################
ordena.PoblacionXcosto<-function(Poblacion,costo){
temp<-cbind(costo,Poblacion)
m<-order(temp[,1])
temp<-temp[m,]
temp
}
```

7.2.5. Función que genera una nueva pobalción: genera hijos y mutantes

```
#############################################################################
Funcion que genera una nueva poblacion: genera hijos y mutantes
#############################################################################
genera.nueva.poblacion<-function(Poblacion,costo){
temp<-ordena.PoblacionXcosto(Poblacion,costo)
k<-ncol(temp)
Pobla<-temp[,2:k]
nueva.poblacion<-genera.hijos(Pobla)
nueva.poblacion<-mutaciones(nueva.poblacion)
}
```

Función que genera los mutantes

```
#############################################################################
Funcion que genera mutaciones
#############################################################################
mutaciones<-function(Poblacion,porcentaje=0.05){
n<-nrow(Poblacion)
```

```
n1<-ceiling(n*porcentaje)
k1<-sample(2:n,n1,replace=F)
k2<-sample(1:ncol(Poblacion),n1,replace=T) # 0.05 no debe ponerse muy alto
Poblacion[k1,k2]<-((Poblacion[k1,k2]+1) %% 2) # modulo 2
Poblacion
}
```

Función genera hijos

```
##############################################################################
Funcion que genera hijos de una poblacion
##############################################################################
genera.hijos<-function(Poblacion.ordenada){
n<-nrow(Poblacion.ordenada)
nmedio<-n%/%2
npar<-(nmedio%/%2)*2
padres<-Poblacion.ordenada[seq(1,npar,by=2),]
madres<-Poblacion.ordenada[seq(2,npar,by=2),]
hijos<-apply(rbind(padres,madres),1,generar.hijos.de.pareja)
hijos<-matrix(hijos,ncol=ncol(Poblacion.ordenada),byrow=T)
nueva.poblacion<-rbind(padres,madres,hijos)
nueva.poblacion
}
```

7.2.6. Función generar hijos de parejas

```
##############################################################################
La funcion generar.hijos.de.pareja Genera un par de hijos por la pareja
##############################################################################
generar.hijos.de.pareja<-function(padremadre){
n1<-length(padremadre)
n<-n1/2
padre<-padremadre[1:n]
madre<-padremadre[(n+1):n1]
k<-sample(1:(n-1),1)
hijo1<-c(padre[1:k],madre[(k+1):n])
hijo2<-c(madre[1:k],padre[(k+1):n])
result<-c(hijo1,hijo2)
result
}
```

7.2.7. Función algoritmo genético

```
algoritmo.genetico <-function(N, Nbits, Npar, minimos, maximos, pis, errores,
n_generaciones) {
Poblacion<- generar.poblacion(N,Nbits,Npar)
costo_pob <- costo(Poblacion,Nbits,minimos, maximos, pis, errores)
for (j in 1:n_generaciones){
print("========================")
print(paste("Generacion",j))
print("========================")
# N.Poblacion <- ordena.PoblacionXcosto(Poblacion,costo_pob)
k<-ncol(Poblacion)
Poblacion <- genera.nueva.poblacion(Poblacion, costo_pob)
costo_pob <- costo(Poblacion,Nbits,minimos, maximos, pis, errores)
print("========================")
print (cbind(costo_pob,Poblacion))
print("========================")
}# Fin for n-generaciones
res <- ordena.PoblacionXcosto(Poblacion, costo_pob)
nis<- parametros(res[1,3:ncol(res)],Nbits,minimos,maximos)
nis <- round(nis,0)
#nis<- rowSums(nis)
return(nis)
}
```

7.2.8. Función nis subpoblación

```
##############################################################################
#Funcion que aplica el algoritmo genetico a cada subpoblacion y
# entrega un vector con los tamanos de cada subpoblacion
##############################################################################
nis.subpoblacion <- function(N,Nbits,Npar,minimos,maximos,P,errores,
n_generaciones){
aux <- NULL
f <- nrow(P)
for(i in 1:f){
nis<- algoritmo.genetico(N[i,],Nbits,Npar,minimos,maximos,P[i,],errores,
n_generaciones)
aux <-rbind(nis,aux)
}# fin for i
return(aux)
```

```
}
```

7.3. Función principal

La función para calcular los estimadores de los parámetros de interés matriz de covarianza. Se requiere la librería MASS. Esta función tiene como componentes:

```
#####################################################################
# Funcion Principal
#############################################################
funcion.principal <- function(N,P,Pobs,A,minimos,maximos, errores, Nbits,
Npar,Nsim,n_generaciones){
y<- NULL
contar <- 0
repeat {
nisestratos <- nis.estratos(N,Nbits,Npar,minimos,maximos,P,errores,
n_generaciones)
#nisestratos1 <- matrix(t(nisestratos),ncol=3,byrow=T)
#nis obtenidos del algortimo genetico
nisestratos1 <- nisestratos
n.i<- rowSums(nisestratos1)
n.i <- matrix(n.i, ncol=1)
k <- nrow(n.i)
print(k)
aux <- NULL
res<- NULL
piest<- NULL
muestra<- NULL
for(i in 1:nrow(nisestratos1)){
muestra <- rmultinom(Nsim, n.i[i], Pobs[i,])
res <- apply(muestra, 1, mean)
res <- res/n.i[i]
print("========================")
print(paste("pi estimados por estrato", i))
print(res)
print("========================")
aux <- matrix(res, ncol=1,byrow = T)
temp1<- aux%*%t(aux)
temp2<-diag(as.vector(aux))
varcov.p<-(temp2-temp1)/n.i[i]
```

```
print("========================")
print(paste("Matriz de varianzas y covarianzas estimada del estrato",i))
print(varcov.p)
print("========================")
#varcov.grande<-varcov.grande
if(i==1){
varcov.grande<-varcov.p
piest<-res
}
else{
varcov.grande<-crea.bloque(varcov.grande,varcov.p)
piest<-rbind(piest,res)
}
}
piestimado <- matrix(piest, ncol=3, byrow=F)
#print(piestimado)
piest1 <- NULL
for(i in 1: nrow(piestimado)){
a <- piestimado[i,]
a <- t(a)
piest1 <- cbind(piest1,a)
}
piest1
piest1 <- t(piest1)
# print("pi estimado")
# print(piest1)
A <- identidad(nrow(piest1))
K <- matrix(c(1,0,-1,rep(0,96), rep(0,3),1,0,-1,rep(0,93),rep(0,6),1,0,-1,
rep(0,90),rep(0,9),1,0,-1,rep(0,87), rep(0,12),1,0,-1,rep(0,84),rep(0,15),
1,0,-1,rep(0,81),rep(0,18),1,0,-1,rep(0,78),rep(0,21),1,0,-1,rep(0,75),
rep(0,24),1,0,-1,rep(0,72),rep(0,27),1,0,-1,rep(0,69),rep(0,30),1,0,-1,
rep(0,66),rep(0,33),1,0,-1,rep(0,63),rep(0,36),1,0,-1,rep(0,60),rep(0,39),
1,0,-1,rep(0,57),rep(0,42),1,0,-1,rep(0,54),rep(0,45),1,0,-1,rep(0,51),rep(0,48)
,1,0,-1,rep(0,48),rep(0,51),1,0,-1,rep(0,45),rep(0,54),1,0,-1,rep(0,42),rep(0,57)
,1,0,-1,rep(0,39),rep(0,60),1,0,-1,rep(0,36),rep(0,63),1,0,-1,rep(0,33),rep(0,66)
,1,0,-1,rep(0,30),rep(0,69),1,0,-1,rep(0,27),rep(0,72),1,0,-1,rep(0,24),rep(0,75)
,1,0,-1,rep(0,21),rep(0,78),1,0,-1,rep(0,18),rep(0,81),1,0,-1,rep(0,15),rep(0,84)
,1,0,-1,rep(0,12),rep(0,87),1,0,-1,rep(0,9),rep(0,90),1,0,-1,rep(0,6),rep(0,93),
1,0,-1,rep(0,3),rep(0,96),1,0,-1,0,1,-1,rep(0,96), rep(0,3),0,1,-1,rep(0,93),
rep(0,6),0,1,-1,rep(0,90),rep(0,9),0,1,-1,rep(0,87),rep(0,12),0,1,-1,rep(0,84),
```

```
rep(0,15),0,1,-1,rep(0,81),rep(0,18),0,1,-1,rep(0,78),rep(0,21),0,1,-1,rep(0,75),
rep(0,24),0,1,-1,rep(0,72),rep(0,27),0,1,-1,rep(0,69),rep(0,30),0,1,-1,rep(0,66),
rep(0,33),0,1,-1,rep(0,63),rep(0,36),0,1,-1,rep(0,60),rep(0,39),0,1,-1,rep(0,57),
rep(0,42),0,1,-1,rep(0,54),rep(0,45),0,1,-1,rep(0,51),rep(0,48),0,1,-1,rep(0,48),
rep(0,51),0,1,-1,rep(0,45),rep(0,54),0,1,-1,rep(0,42),rep(0,57),0,1,-1,rep(0,39),
rep(0,60),0,1,-1,rep(0,36),rep(0,63),0,1,-1,rep(0,33),rep(0,66),0,1,-1,rep(0,30),
rep(0,69),0,1,-1,rep(0,27),rep(0,72),0,1,-1,rep(0,24),rep(0,75),0,1,-1,rep(0,21),
rep(0,78),0,1,-1,rep(0,18),rep(0,81),0,1,-1,rep(0,15),rep(0,84),0,1,-1,rep(0,12),
rep(0,87),0,1,-1,rep(0,9),rep(0,90),0,1,-1,rep(0,6),rep(0,93),0,1,-1,rep(0,3),
rep(0,96),0,1,-1),byrow = T,ncol=99)
#Funcion de interes f_estimado
pis <- A%*%piest1
log.pis <-log(pis)
f_est <- K%*%log.pis
#print("funcion de interes")
#print(f_est)

# Calcular (A*sigma pi-gorro*)
var.cov.pis<-A%*%varcov.grande%*%t(A)
#Calcular de matriz D-lineal
Di<-diag(1/as.vector(pis))
# print("Di")
#print(Di)

# Calculo de la matriz de varianzas y covarianzas Sigma f gorro
var.cov.f<-K%*%Di%*%var.cov.pis%*%t(K%*%Di)
print("var.cov.f")
print(var.cov.f)
alpha <- 0.05
#Intervalo de confianza f_est
#varcov.fg<-X%*%varcov.b%*%t(X)
error.f<-sqrt(diag(var.cov.f))
margen1 <- qnorm(alpha/2,mean = 0,sd=1,lower.tail = F)
lim_inferior1 <- f_est -margen1*error.f
lim_superior1 <- f_est + margen1*error.f
I.C.f.est <- cbind(lim_inferior1,lim_superior1)
#print("Intervalo de confianza f_est")
#print(I.C.f.gor)
l.f_est<- (lim_superior1-lim_inferior1)/2
# print("Longitud media de f_gorro")
```

```
#print(l.f.gor)
f.estimado <- cbind(f_est,I.C.f.est)
#print("f_estimado")
# print(f.estimado)

# # # #f_observado
piobs<- NULL
for(i in 1: nrow(Pobs)){
b <- Pobs[i,]
b <- t(b)
piobs <- cbind(piobs,b)
}
piobs
piobs <- t(piobs)
# # print("pi observado")
# # print(piobs)

pisobs <- A%*%piobs
log.pisobs <-log(pisobs)
f_obs <- K%*%log.pisobs
# print("funcion de interes")
#print(f_obs)

# n <- cbind(n.i,n.i)
# n<- matrix(n, ncol=1)
# L.IObs <- f_obs- (1.64)*sqrt((f_obs*(1-f_obs))/n)
# L.SObs <- f_obs+ (1.64)*sqrt((f_obs*(1-f_obs))/n)
# I.CObs <- rbind(L.IObs,L.SObs)
# print("f_obs, I.CObs")
# print(cbind(f_obs,I.CObs))
# l.f_obs <- I.Cfuncioninteres(f_obs,n.i)

#Matriz de diseno
X <- matrix(c(1, 1, rep(0,31), rep(0,33), 1, rep(0,1),1, rep(0,30),rep(0,33),
1, rep(0,2),1, rep(0,29),rep(0,33),1, rep(0,3),1, rep(0,28),rep(0,33),
1, rep(0,4),1, rep(0,27),rep(0,33),1, rep(0,5),1, rep(0,26),rep(0,33),
1, rep(0,6),1, rep(0,25),rep(0,33),1, rep(0,7),1, rep(0,24),rep(0,33),
1, rep(0,8),1, rep(0,23),rep(0,33),1, rep(0,9),1, rep(0,22),rep(0,33),
1, rep(0,10),1, rep(0,21),rep(0,33),1, rep(0,11),1, rep(0,20),rep(0,33),
1, rep(0,12),1, rep(0,19),rep(0,33),1, rep(0,13),1, rep(0,18),rep(0,33),
```

```
1, rep(0,14),1, rep(0,17),rep(0,33),1, rep(0,15),1, rep(0,16),rep(0,33),
1, rep(0,16),1, rep(0,15),rep(0,33),1, rep(0,17),1, rep(0,14),rep(0,33),
1, rep(0,18),1, rep(0,13),rep(0,33),1, rep(0,19),1, rep(0,12),rep(0,33),
1, rep(0,20),1, rep(0,11),rep(0,33),1, rep(0,21),1, rep(0,10),rep(0,33),
1, rep(0,22),1, rep(0,9),rep(0,33),1, rep(0,23),1, rep(0,8),rep(0,33),
1, rep(0,24),1, rep(0,7),rep(0,33),1, rep(0,25),1, rep(0,6),rep(0,33),
1, rep(0,26),1, rep(0,5),rep(0,33),1, rep(0,27),1, rep(0,4),rep(0,33),
1, rep(0,28),1, rep(0,3),rep(0,33),1, rep(0,29),1, rep(0,2),rep(0,33),
1, rep(0,30),1, rep(0,1),rep(0,33),1, rep(0,31),1,rep(0,33),1,rep(0,32),
rep(0,33),rep(0,33),1, 1, rep(0,31), rep(0,33), 1, rep(0,1),1, rep(0,30),
1, rep(0,2),1, rep(0,29),rep(0,33),1, rep(0,3),1, rep(0,28),rep(0,33),
1, rep(0,4),1, rep(0,27),rep(0,33),1, rep(0,5),1, rep(0,26),rep(0,33),
1, rep(0,6),1, rep(0,25),rep(0,33),1, rep(0,7),1, rep(0,24),rep(0,33),
1, rep(0,8),1, rep(0,23),rep(0,33),1, rep(0,9),1, rep(0,22),rep(0,33),
1, rep(0,10),1, rep(0,21),rep(0,33),1, rep(0,11),1, rep(0,20),rep(0,33),
1, rep(0,12),1, rep(0,19),rep(0,33),1, rep(0,13),1, rep(0,18),rep(0,33),
1, rep(0,14),1, rep(0,17),rep(0,33),1, rep(0,15),1, rep(0,16),rep(0,33),
1, rep(0,16),1, rep(0,15),rep(0,33),1, rep(0,17),1, rep(0,14),rep(0,33),
1, rep(0,18),1, rep(0,13),rep(0,33),1, rep(0,19),1, rep(0,12),rep(0,33),
1, rep(0,20),1, rep(0,11),rep(0,33),1, rep(0,21),1, rep(0,10),rep(0,33),
1, rep(0,22),1, rep(0,9),rep(0,33),1, rep(0,23),1, rep(0,8),rep(0,33),
1, rep(0,24),1, rep(0,7),rep(0,33),1, rep(0,25),1, rep(0,6),rep(0,33),
1, rep(0,26),1, rep(0,5),rep(0,33),1, rep(0,27),1, rep(0,4),rep(0,33),
1, rep(0,28),1, rep(0,3),rep(0,33),1, rep(0,29),1, rep(0,2),rep(0,33),
1, rep(0,30),1, rep(0,1),rep(0,33),1, rep(0,31),1,rep(0,33),1,rep(0,32)),
byrow = T, ncol=66)

#print("Matriz de diseno")
# print(X)

#### # Calcular los betas utilizando inversa generalizada y
#### # matriz de varianzas y covarianzas del modelo
temp<-solve(t(X)%*%ginv(var.cov.f)%*%X)
beta<-temp%*%t(X)%*%ginv(var.cov.f)%*%f_est
# print("Betas estimados")
# print(beta)
Sigma.beta<-temp
print("Matriz de varianzas y covarianzas del modelo")
print(Sigma.beta)
```

```
#####################################################################
######Inferencia sobre el modelo
#####################################################################
######Prueba para parametros individuales de f1
C <- matrix(c(0,1,rep(0,64),rep(0,2),1,rep(0,63),rep(0,3),1,rep(0,62),
rep(0,4),1,rep(0,61),rep(0,5),1,rep(0,60),rep(0,6),1,rep(0,59),
rep(0,7),1,rep(0,58),rep(0,8),1,rep(0,57),rep(0,9),1,rep(0,56),
rep(0,10),1,rep(0,55),rep(0,11),1,rep(0,54),rep(0,12),1,rep(0,53),
rep(0,13),1,rep(0,52),rep(0,14),1,rep(0,51),rep(0,15),1,rep(0,50),
rep(0,16),1,rep(0,49),rep(0,17),1,rep(0,48),rep(0,18),1,rep(0,47),
rep(0,19),1,rep(0,46),rep(0,20),1,rep(0,45),rep(0,21),1,rep(0,44),
rep(0,22),1,rep(0,43),rep(0,23),1,rep(0,42),rep(0,24),1,rep(0,41),
rep(0,25),1,rep(0,40),rep(0,26),1,rep(0,39),rep(0,27),1,rep(0,38),
rep(0,28),1,rep(0,37),rep(0,29),1,rep(0,36),rep(0,30),1,rep(0,35),
rep(0,31),1,rep(0,34),rep(0,32),1,rep(0,33)),ncol=66,byrow=T)
for(i in 1:nrow(C)){
tempf1<-matrix(C[i,],nrow=1)

S.temp<-tempf1%*%Sigma.beta%*%t(tempf1)
mi.chitempf1<-(tempf1%*%beta)%*%solve(S.temp)%*%(tempf1%*%beta)
gltempf1<-nrow(tempf1)
valor.ptempf1<-pchisq(mi.chitempf1,gltempf1,lower.tail=F)
# print("Coeficiente.")
# print(i)
# print("Matriz de Pruebas individuales sobre las variables f1")
# print(tempf1)
# print("Chi.obs.....Grados de libertad....Valor-p")
# print(c(mi.chitempf1,gltempf1,valor.ptempf1))
}

C2 <- matrix(c(rep(0,33),0,1,rep(0,31),rep(0,33),rep(0,2),1,rep(0,30),
rep(0,33),rep(0,3),1,rep(0,29),rep(0,33),rep(0,4),1,rep(0,28),
rep(0,33),rep(0,5),1,rep(0,27),rep(0,33),rep(0,6),1,rep(0,26),
rep(0,33),rep(0,7),1,rep(0,25),rep(0,33),rep(0,8),1,rep(0,24),
rep(0,33),rep(0,9),1,rep(0,23),rep(0,33),rep(0,10),1,rep(0,22),
rep(0,33),rep(0,11),1,rep(0,21),rep(0,33),rep(0,12),1,rep(0,20),
rep(0,33),rep(0,13),1,rep(0,19),rep(0,33),rep(0,14),1,rep(0,18),
rep(0,33),rep(0,15),1,rep(0,17),rep(0,33),rep(0,16),1,rep(0,16),
rep(0,33),rep(0,17),1,rep(0,15),rep(0,33),rep(0,18),1,rep(0,14),
rep(0,33),rep(0,19),1,rep(0,13),rep(0,33),rep(0,20),1,rep(0,12),
```

```
rep(0,33),rep(0,21),1,rep(0,11),rep(0,33),rep(0,22),1,rep(0,10),
rep(0,33),rep(0,23),1,rep(0,9),rep(0,33),rep(0,24),1,rep(0,8),
rep(0,33),rep(0,25),1,rep(0,7),rep(0,33),rep(0,26),1,rep(0,6),
rep(0,33),rep(0,27),1,rep(0,5),rep(0,33),rep(0,28),1,rep(0,4),
rep(0,33),rep(0,29),1,rep(0,3),rep(0,33),rep(0,30),1,rep(0,2),
rep(0,33),rep(0,31),1,rep(0,1),rep(0,33),rep(0,32),1),ncol=66,byrow=T)

for(i in 1:nrow(C2)){
tempf2<-matrix(C2[i,],nrow=1)
S.temp<-tempf2%*%Sigma.beta%*%t(tempf2)
mi.chitempf2<-(tempf2%*%beta)%*%solve(S.temp)%*%(tempf2%*%beta)
gltempf2<-nrow(tempf2)
valor.ptempf2<-pchisq(mi.chitempf2,gltempf2,lower.tail=F)
# print("Coeficiente.")
# print(i)
# print("Matriz de Pruebas individuales sobre las variables f2")
# print(tempf2)
# print("Chi.obs.....Grados de libertad....Valor-p")
#print(c(mi.chitempf2,gltempf2,valor.ptempf2))
}

#Pruebas chi cuadrado para los parametros individuales para f
chi.cua<-beta^2/diag(Sigma.beta)
#print("Pruebas individuales para beta")
#print(chi.cua)
Z<-nrow(beta)/2
#print("Z")
#print(Z)
chi.cuaf1<-chi.cua[c(1:Z)]
#print("Pruebas individuales para parametros f1")
#print(chi.cuaf1)
valor.p<-pchisq(chi.cua,1,lower=F)
valor.pf1<-valor.p[c(1:Z)]
#print("Valor-p f1")
#print(valor.pf1)
# print("Beta----Error--- chicuadrado-valorp")
# print(cbind(beta[c(1:Z)],sqrt(diag(Sigma.beta)),chi.cuaf1,valor.pf1))

#####################################################################
# Prueba para f1 y f2
```

```
#Prueba para f1
C1 <- matrix(c(0,1,rep(0,64),rep(0,2),1,rep(0,63),rep(0,3),1,rep(0,62),
rep(0,4),1,rep(0,61),rep(0,5),1,rep(0,60),rep(0,6),1,rep(0,59),
rep(0,7),1,rep(0,58),rep(0,8),1,rep(0,57),rep(0,9),1,rep(0,56),
rep(0,10),1,rep(0,55),rep(0,11),1,rep(0,54),rep(0,12),1,rep(0,53),
rep(0,13),1,rep(0,52),rep(0,14),1,rep(0,51),rep(0,15),1,rep(0,50),
rep(0,16),1,rep(0,49),rep(0,17),1,rep(0,48),rep(0,18),1,rep(0,47),
rep(0,19),1,rep(0,46),rep(0,20),1,rep(0,45),rep(0,21),1,rep(0,44),
rep(0,22),1,rep(0,43),rep(0,23),1,rep(0,42),rep(0,24),1,rep(0,41),
rep(0,25),1,rep(0,40),rep(0,26),1,rep(0,39),rep(0,27),1,rep(0,38),
rep(0,28),1,rep(0,37),rep(0,29),1,rep(0,36),rep(0,30),1,rep(0,35),
rep(0,31),1,rep(0,34),rep(0,32),1,rep(0,33)),ncol=66,byrow=T)

mi.chi1<-t(C1%*%beta)%*%solve(C1%*%Sigma.beta%*%t(C1))%*%C1%*%beta
gl<-nrow(C1)
valor.p<-pchisq(mi.chi1,gl,lower.tail=F)
print("Prueba f1 Chi.obs.....Grados de libertad....Valor-p")
print(c(mi.chi1,gl,valor.p))

#Prueba para f2
C2 <- matrix(c(rep(0,33),0,1,rep(0,31),rep(0,33),rep(0,2),1,rep(0,30),
rep(0,33),rep(0,3),1,rep(0,29),rep(0,33),rep(0,4),1,rep(0,28),
rep(0,33),rep(0,5),1,rep(0,27),rep(0,33),rep(0,6),1,rep(0,26),
rep(0,33),rep(0,7),1,rep(0,25),rep(0,33),rep(0,8),1,rep(0,24),
rep(0,33),rep(0,9),1,rep(0,23),rep(0,33),rep(0,10),1,rep(0,22),
rep(0,33),rep(0,11),1,rep(0,21),rep(0,33),rep(0,12),1,rep(0,20),
rep(0,33),rep(0,13),1,rep(0,19),rep(0,33),rep(0,14),1,rep(0,18),
rep(0,33),rep(0,15),1,rep(0,17),rep(0,33),rep(0,16),1,rep(0,16),
rep(0,33),rep(0,17),1,rep(0,15),rep(0,33),rep(0,18),1,rep(0,14),
rep(0,33),rep(0,19),1,rep(0,13),rep(0,33),rep(0,20),1,rep(0,12),
rep(0,33),rep(0,21),1,rep(0,11),rep(0,33),rep(0,22),1,rep(0,10),
rep(0,33),rep(0,23),1,rep(0,9),rep(0,33),rep(0,24),1,rep(0,8),
rep(0,33),rep(0,25),1,rep(0,7),rep(0,33),rep(0,26),1,rep(0,6),
rep(0,33),rep(0,27),1,rep(0,5),rep(0,33),rep(0,28),1,rep(0,4),
rep(0,33),rep(0,29),1,rep(0,3),rep(0,33),rep(0,30),1,rep(0,2),
rep(0,33),rep(0,31),1,rep(0,1),rep(0,33),rep(0,32),1),ncol=66,byrow=T)

mi.chi2<-t(C2%*%beta)%*%solve(C2%*%Sigma.beta%*%t(C2))%*%C2%*%beta
gl<-nrow(C2)
valor.p2<-pchisq(mi.chi2,gl,lower.tail=F)
```

```
print(" Prueba f2 Chi.obs.....Grados de libertad....Valor-p")
print(c(mi.chi2,gl,valor.p2))
#########################################################################
# calcular f*
f.gor<-X%*%beta
#print("f.gorro")
#print(f.gor)

alpha <- 0.05
#Intervalos de confianza Beta
margene<-qnorm(alpha/2,mean = 0,sd=1,lower.tail = F)*
sqrt(diag(Sigma.beta))
extrizqui<-beta-margene
extrdere<-beta+margene
intervalosdeconfianza<-cbind(extrizqui,extrdere)
betas <- cbind(beta,intervalosdeconfianza)
#print("betas")
#print(betas)
#print("Beta----Intervalo---Error--- chicuadrado-----valorp")
# print(cbind(beta,sqrt(diag(Sigma.beta)),intervalosdeconfianza, chi.cuaf1,
valor.pf1))

nigrande <- matrix(c(n.i,n.i),byrow = T,ncol=1)

#Intervalo de Confianza f_gorro
varcov.fg<-X%*%Sigma.beta%*%t(X)
error.fg<-sqrt(diag(varcov.fg))
margen <- qnorm(alpha/2,mean = 0,sd=1,lower.tail = F)
lim_inferior <- f.gor -margen*error.fg
lim_superior <- f.gor + margen*error.fg
I.C.f.gor <- cbind(lim_inferior,lim_superior)
# print("Intervalo de confianza f_gorro")
# print(I.C.f.gor)
l.f.gor<- (lim_superior-lim_inferior)/2

# print("Longitud media de f_gorro")
#print(l.f.gor)
f_gorro <- cbind(f.gor,I.C.f.gor)
# print("f-gorro")
# print(f_gorro)
```

```
#Intervalo de confianza f_obs
z <- nrow(f_obs)
B <- NULL
A <- NULL
for(i in 34:z){
L.If2<-f_obs[i,] -(1.96)*sqrt((1/nisestratos1[i-33,2])+(1/nisestratos1[i-33,3]))
L.Sf2<-f_obs[i,] +(1.96)*sqrt((1/nisestratos1[i-33,2])+(1/nisestratos1[i-33,3]))
B <- rbind(B, L.If2)
A <- rbind(A,L.Sf2)
}
lonf2<- (A-B)/2
I.Cf2 <-cbind(B,A)
B1 <- NULL
A1 <- NULL
for(i in 1:33){
L.If1 <- f_obs[i,] -(1.96)*sqrt((1/nisestratos1[i,1])+(1/nisestratos1[i,3]))
L.Sf1 <- f_obs[i,] +(1.96)*sqrt((1/nisestratos1[i,1])+(1/nisestratos1[i,3]))
B1 <- rbind(B1, L.If1)
A1 <- rbind(A1,L.Sf1)
}
lonf1 <- (A1-B1)/2
I.Cf1 <- cbind(B1,A1)
# print("I.f1")
# print(I.Cf1)
I.Cf_obs <- rbind(I.Cf1,I.Cf2)
print("f_obs e Intervalo")
print(cbind(f_obs,I.Cf_obs))
l.f_obs <- rbind(lonf1,lonf2)

#Residuales de la funcion respuesta
residuales_resp <- f_obs -f.gor
print("Residuales crudos de la respuesta")
print(residuales_resp)


#Probabilidades estimadas
a<- exp(t(X[1,])%*%beta)
b <- (1+ exp(t(X[1,])%*%beta))
```

```
c <- (1+exp(t(X[1+33,])%*%beta))
d <- (exp(t(X[1,])+t(X[1+33,]))%*%beta)
e <- exp(t(X[1+33,])%*%beta)
pi_11 <- a/(1+a+e)
pi_12 <- e /(1+a+e)
pi_13 <- 1/(1+a+e)
p1 <- cbind(pi_11, pi_12, pi_13)
print("p1")
print(p1)

optimizar <- a.optimizar(l.f_obs,l.f.gor)
print("criterio de optimizacion")
print(optimizar)
if(optimizar <= 0.05){
y <- n.i
break
} # Fin si
} # Fin repeat
print("nis optimos")
return(n.i)
}
```

Bibliografía

[1] Alan Agresti. *Categorical Data Analysis*. John Wile and Sons, 1990.

[2] Alan Agresti. *An Introduction to Categorical Data Analysis*. John Wile and Sons, 2007.

[3] Alan Agresti. *Analysis of Ordinal Categoriacal Data*. John Wile and Sons, 2010.

[4] Jeffrey F. Bromaghin. Sample size determination for interval estimation of multinomial probabilities. *American Statiscian*, 47:203, 1993.

[5] W.G. Cochran. *Sampling Technics*. John Wiley y Sons, 1977.

[6] Juan Carlos Correa Morales. *Analysis of Contingency Tables Via GSK Using R:Lecture Notes.*, volume 31. Universidad Nacional de Colombia Sede Medellín, 2020.

[7] González Gómez Difariney, Juan Carlos Correa Morales, and Jorge Iván Vélez. Comparación de 13 intervalos de confianza para los paraámetros de la distribucióm multinomial. *Revista Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín*, 4(2):150–163.

[8] P. W. Eaton. Yarnold's criterion and minimum sample size. *American Statiscian*, 32:102–103, 1978.

[9] Judith D. Goldberg. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association*, 70:561–567, 1975.

[10] Leo A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Tecnhometrics*, 7(2):247–254, 1965.

[11] James E. Grizzle and G. Koch. Some applications of categorical data analysis to epidemiological studies. *Brogan and Partners*, 32:169–179, 1979.

[12] James E. Grizzle, C. Frank Starmer, and Gary G. Koch. Analysis of categorical data by linear models. *Biometrics*, 25:489, 1969.

[13] Randy L. Haupt and Sue Ellen Haupt. *Practical Genetic Algorithms*. John Wiley and Sons, 1998.

[14] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press. A Bradfford Book, 1975.

[15] Instituto Colombiano de Bienestar Familiar ICBF. *Encuesta Nacional de la Situación Nutricional en Colombia, 2005*. 2005.

[16] J.G. Kalbeish. *Probability and Statistical Inference*, volume 2. Springer-Verlag, 1985.

[17] P. McCullang and T.A Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[18] Abdelmalik Moujahid, Iñaki Inza, and Pedro Larraï¿½aga. *Algoritmos Genéticos*. Universidad del País Vasco Euskal Herriko Unibertsitatea, 2008.

[19] Gil Londoño Natyhelem. Algoritmos genéticos. Master's thesis, Universidad Nacional de Colombia sede Medellín, Medellín, 2006.

[20] C. P. Quesenberry and D. C. Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Tecnhometrics*, 6(2):191–195, 1964.

[21] James Rochon. The application of the gsk method to the determination of minimum sample sizes. *Biometrics*, 45(1):193–205, 1989.

[22] K. Tanabe and M. Sagae. An exact cholesky descomposition and gneralized inverse of the variance-covariance matriz of multinomial distribution, with applications. *Journal of the Royal Statistical Society (Methodological)*, 54:211–219, 1992.

[23] Steven K. Thompson. Sample size for estimating multinomial proportions. *The American Statistician*, 41(1):42–46, 1987.

[24] Robert D. Tortora. A note on sample size estimation for multinomial populations. *The American Statistician*, 32(3):100–102, 1978.

[25] J.K Yarnold. The minimum expectation in χ^2 goodness of fit tests and the accuracy of aproximations for the null distribution. *Journal of the American Statistical Association*, 65(330):864–886, 1970.