Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Facultad de Ciencias Matemáticas
Escuela Académico Profesional de Estadística

Uso de la información auxiliar para el ajuste de muestras probabilísticas

MONOGRAFÍA

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Dora Esmeralda ORTEGA ASENCIOS

ASESOR

Julio César RAMOS RAMÍREZ

Lima, Perú

2008

Referencia bibliográfica

Ortega, D. (2008). *Uso de la información auxiliar para el ajuste de muestras probabilísticas*. Monografía para optar el título profesional de Licenciada en Estadística. Escuela Académico Profesional de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

Dedicatoria

A mis padres Genoveva y Walter, por su ejemplo de lucha, apoyo, sacrificio y amor incondicional que hicieron posible el alcanzar una nueva meta en mi vida.

AGRADECIMIENTOS

- A Dios por darme la posibilidad de estar culminando un objetivo de mi vida.

- A mis padres, Walter y Genoveva, por enseñarme el verdadero significado de la vida y por el principal legado: la educación y la cultura; a mis hermanos Javier, Yessica y Susan que junto con mis padres son la fuerza que me impulsa a seguir día a día.

- Un agradecimiento muy especial a mi tutor Lic. Julio C. Ramos Ramírez que me orientó en todo momento con sus conocimientos y experiencia para poder culminar esta monografía.

- A las profesoras de la escuela de estadística quienes me impartieron conocimiento que es el soporte de mi desarrollo profesional y muy en especial a la Mg. Estela Ponce Aruneri, quien siempre ha estado dispuesta a absolver mis dudas y sobre todo a aconsejarme, más que una profesora la considero una amiga.

- A todas las personas que de una u otra manera me han apoyado con sus consejos, amor y paciencia y que han sido un ejemplo para mi vida profesional y sobre todo para mi vida personal.

RESUMEN

USO DE LA INFORMACIÓN AUXILIAR PARA EL AJUSTE DE MUESTRAS PROBABILÍSTICAS

Dora E. Ortega Asencios

MAYO – 2008

Asesor : Lic. Julio C. Ramos Ramírez

Título Obtenido : Licenciado en Estadística

La fase de ajuste y expansión de muestras tiene por objetivo pasar de la muestra a la población, utilizando para ello toda la información disponible y haciendo compatibles los resultados entre diferentes encuestas u otras fuentes.

En la presente monografía se describe la metodología de los "Métodos con Máxima Información Auxiliar" utilizada para el ajuste estadístico de muestras y una introducción de la metodología basada en "procedimientos iterativos" utilizados cuando se dispone de información auxiliar univariante.

Palabras Clave: Ajuste de muestras, Elevadores, Información auxiliar máxima, Estimadores de Raking Ratio.

ABSTRACT

USE OF THE AUXILIARY INFORMATION FOR THE ADJUSTMENT OF SAMPLES PROBABILÍSTICAS

Dora E. Ortega Asencios

MAYO – 2008

Adviser : Lic. Julio C. Ramos Ramírez

Academic Degree: Licensed in statistics

The phase of adjustment and expansion of samples has for aim(lens) happen from the sample to the population, using for it all the available information and making the results compatible between(among) different surveys or other sources(fountains).

In the present monograph there is described the methodology of the " Methods by Maximum Auxiliary Information " used for the statistical adjustment of samples and an introduction of the methodology based on " iterative procedures " used when univariante arranges of auxiliary information.

Key Word: Adjustment of samples, Elevators, Auxiliary maximum information, Estimator of raking ratio.

INDICE

INTRODUCCIÓN

En esta monografía abordaremos un tema frecuente en la estadística, que se describe en los siguientes párrafos.

Tenemos una población o universo de tamaño N, sobre la que queremos estimar el total de una cierta característica Y. Al ser imposible, en general, encuestar a todos los elementos de una población, lo que se hace es tomar una muestra de tamaño n de dicha población, y se estudia el comportamiento de dicha característica en estos individuos. Pero evidentemente, el total de Y sobre los elementos observados no va a ser el mismo que su total poblacional y para conseguir obtener una estimación de dicho total, considerando las observaciones muestrales obtenidas para Y, se utilizarán unos elevadores (Factores de Expansión). Con esta elevación conseguiremos una primera estimación del total poblacional de la variable Y.

La elevación también es usada para compensar la pérdida de muestra debida a la no-respuesta. Esta última compensación se puede realizar tanto en censos como en encuestas muestrales, pero se ha de hacer siempre que los respondientes no tengan un perfil específico y diferenciado.

En general, además de la variable objetivo Y (variable de interés), dispondremos de otras características observadas en la población y en la muestra, denominadas variables auxiliares, suponiendo una cierta relación entre la variable objetivo y las auxiliares.

Gracias a esta suposición, modificaremos las elevaciones o pesos iniciales de los individuos muéstrales con el fin de ajustar a la distribución poblacional conocida, las variables auxiliares.

Entre los métodos de ajuste y elevación empleados en la muestra, unos utilizan información auxiliar máxima y otra información auxiliar marginal.

Los métodos descritos en esta monografía son utilizados en la actualidad en el Instituto Vasco de Estadística(EUSTAT), en donde han elaborado un cuaderno de trabajo que ha sido el punto de partida de esta monografía, en este documento se describe la metodología utilizada para el ajuste estadístico de muestras, diferenciando dos grupos de metodologías, según la cantidad de información auxiliar disponible:

- Métodos con Máxima Información Auxiliar: utilizados cuando se dispone de la distribución auxiliar multivariante y,
- Métodos iterativos: utilizados en el caso de que la información auxiliar disponible sea la univariante. Solo se considera este tipo de ajuste cuando las variables auxiliares son cualitativas. Se presentan dos métodos iterativos de ajuste: el Raking y una variante del mismo, utilizada en el procedimiento Redre del SPAD.

Debe señalarse que la idea de estimar la frecuencia de una casilla de una tabla de contingencia a partir de los totales marginales, se debe a Demming and Stephen(1940)en el paper denominado ***"On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known"***,

en donde muestra un modo de asegurar la consistencia entre la data poblacional y la data muestral del Censo de Población de 1940 en U.S. al cual llamaron procedimiento ajuste proporcional iterativo(Raking Ratio), desde entonces los avances y las modificaciones han sido numerosas.

La estimación de la razón por el procedimiento iterado ha sido estudiado por J.N.K. RAO y por Brackstone en ***"AN INVESTIGATION OF RAKING RATIO ESTIMATORS"***, en donde compara el error estándar del estimador Raking ratio del total poblacional, con el error estándar generado por el estimador máximo verosímil, para lo cual realiza un estudio empírico en donde se muestra la reducción del error estándar obtenido por el raking(iteraciones sucesivas) para una variedad de características.

En cuanto al procedimiento Redre que también se menciona en esta monografía no existen muchas investigaciones en esta parte.

En el capítulo I, se describen los conceptos básicos de estimación en el muestreo probabilistico. En el capitulo II, analizaremos el ajuste de muestras bajo el método de máxima información auxiliar. Luego en el capitulo III, se presenta una introducción sobre los métodos iterativos para el ajuste de muestras. Finalmente en el capitulo IV, se presenta una aplicación con el método de información auxiliar máxima, sobre una encuesta de características Toxicológicas realizada en la comunidad de AYASH de la provincia de Huari del departamento de Ancash.

CAPITULO I

CONCEPTOS BÁSICOS DE ESTIMACIÓN EN EL MUESTREO PROBABILÍSTICO

1.1 Estimador Y Estimación

Un **estimador** es un estadístico (esto es, una función de la muestra aleatoria) usado para estimar un parámetro desconocido de la población. Por ejemplo, si se desea conocer el precio medio de un artículo (el parámetro desconocido) se recogerán observaciones del precio de dicho artículo en diversos establecimientos (la muestra) y la media aritmética de las observaciones puede utilizarse como estimador del precio medio. Es decir, si las unidades de la población son $\{u_1, u_2, \ldots, u_N\}$ cuyos valores respectivos de la variable Y son $\{Y_1, Y_2, \ldots, Y_N\}$, entonces el parámetro media poblacional es $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$.

Si las unidades de la muestra son $\{s_1, s_2, \ldots, s_n\}$ cuyas variables aleatorias son $\{y_1, y_2, \ldots, y_n\}$, entonces el estimador natural mas simple del parámetro $\bar{Y}$ es la media muestral $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Para cada parámetro pueden existir varios estimadores diferentes. En general, escogeremos el estimador que posea mejores propiedades que los restantes, como insesgadez, consistencia, convergencia y eficiencia.

Se llama **estimación** al valor aproximado de un parámetro de una población obtenido a partir de los datos proporcionados por una muestra en base a un estimador.

1.2 Formas de Estimación en el Muestreo

1.2.1 Estimación sin Uso de Información Auxiliar

Es aquella forma de estimación en que se parte de unos estimadores básicos de los estadísticos objetivo de estudio, sin considerar la existencia de información auxiliar alguna. Uno de estos estimadores bastantes utilizados en el muestreo probabilistico es el estimador de Horvitz–Thompson, que se describe a continuación.

- **Estimador de Horvitz–Thompson**

Son aquellos estimadores de los parámetros objetivo de estudio que toman como pesos los inversos de las probabilidades de inclusión en la muestra para cada elemento muestral. Son, por lo tanto, pesos asignados a elementos en función a las probabilidades.

Se considera que tenemos una población denotada por U y una muestra de la misma denotada por S:

$U = \{u_1, u_2, \ldots, u_k, \ldots, u_N\}$ con $k = 1, \ldots, N$

$S = \{s_1, s_2, \ldots, s_k, \ldots, s_n\}$ con $k = 1, \ldots, n$

Definimos la variable objetivo Y como un vector Nx1; en la población:

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_k \\ \dots \\ Y_N \end{bmatrix}$$

El parámetro total poblacional a estimar es:

$$Y = \sum_{k=1}^{N} Y_k$$

A cada elemento se le asigna un peso inicial:

$$Z_K = 1 \big/ \pi_K \qquad \text{con k =1,...,n.}$$

Donde π_K es la probabilidad de inclusión del elemento muestral k en la muestra S, en el caso de un muestreo probabilístico.

El estimador de HORVITZ – THOMPSON para el total poblacional de la variable objetivo se define como:

$$\hat{Y}_\pi = \sum_{k=1}^{n} z_k y_k$$

Otra forma de expresar $\hat{Y}_\pi$ es:

$$\hat{Y}_\pi = \sum_{k=1}^{N} z_k y_k \tau_k$$

Donde:

τ_k: Número de veces que aparece la unidad U_k en la muestra S de tamaño n(variable aleatoria con distribución de probabilidad Bernoulli).

$$\tau_k \to B(\pi_k) \qquad \forall\, k = 1,2,...,N$$

Además:

$E(\tau_k) = \pi_k$
$V(\tau_k) = \pi_k(1-\pi_k)$
$COV(\tau_k \tau_l) = \pi_{kl} - \pi_k \pi_l$

π_{Kl} es la probabilidad de inclusión de la unidad U_k y U_l en la muestra S.

El estimador $\hat{Y}_\pi$ es insesgado del total poblacional $Y = \sum_{k=1}^{N} Y_K$:

$$E(\hat{Y}_\pi) = E(\sum_{k=1}^{n} z_K y_K) = \sum_{i=1}^{N} Z_K Y_K E(\tau_k) = \sum_{k=1}^{N} \frac{1}{\pi_k} Y_k \pi_k = \sum_{k=1}^{N} Y_k = Y$$

$$\Rightarrow E(\hat{Y}_\pi) = Y$$

La varianza de $\hat{Y}_\pi$ es:

$$V(\hat{Y}_\pi) = V(\sum_{k=1}^{N} \frac{Y_k}{\pi_k} t_k) = \sum_{k=1}^{N} \frac{Y^2{}_k}{\pi^2{}_k} V(\tau_k) + \sum_{k=1}^{N} \sum_{l \neq k}^{N} \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l} \operatorname{cov}(\tau_k \tau_l)$$

$$V(\hat{Y}_\pi) = \sum_{k=1}^{N} \frac{Y^2{}_k}{\pi^2{}_k} \pi_k (1-\pi_k) + \sum_{k=1}^{N} \sum_{l \neq k}^{N} \frac{Y_k\, Y_l}{\pi_k \pi_l} (\pi_{kl} - \pi_k \pi_l)$$

$$V(\hat{Y}_\pi) == \sum_{k=1}^{N} \frac{y^2{}_k}{\pi_k} (1-\pi_k) + \sum_{k=1}^{N} \sum_{l \neq k1}^{N} \frac{y_k y_l}{\pi_k \pi_l} (\pi_{kl} - \pi_k \pi_l)$$

El estimador de la varianza de $\hat{Y}_\pi$ es:

$$\hat{V}(\hat{Y}_\pi) == \sum_{k=1}^{n} \frac{y^2{}_k}{\pi^2{}_k} (1-\pi_k) + \sum_{k=1}^{n} \sum_{l \neq k}^{n} \frac{y_k y_l}{\pi_k \pi_l} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}}$$

es un estimador insesgado de la varianza de $\hat{Y}_\pi$:

$$E[(\hat{V}(\hat{Y}_\pi)] = \sum_{k=1}^{N} Y^2{}_k \frac{(1-\pi_k)}{\pi_k{}^2} E(\tau_k) + \sum_{k=1}^{N} \sum_{l \neq k1}^{N} \frac{Y_k Y_l}{\pi_k \pi_l} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} E(\tau_k \tau_l)$$

como $E(\tau_K) = \pi_K$ y $E(\tau_K \tau_l) = \pi_{Kl}$; entonces:

$$E(\hat{V}(\hat{Y}_\pi) = \sum_{k=1}^{N} Y^2{}_k \frac{(1-\pi_k)}{\pi_k} + \sum_{k=1}^{N} \sum_{l \neq k}^{N} \frac{(\pi_{kl} - \pi_k \pi)_l}{\pi_k \pi_l} Y_k Y_l$$

En el caso particular de que el muestreo sea sin reemplazo y con igual probabilidad de selección para todos los elementos, tenemos que estas probabilidades de inclusión:

$$\pi_k = \frac{n}{N} \qquad \forall\ k = 1,2,3,....,N \qquad \pi_{kl} = \frac{n}{N} * \frac{n-1}{N-1} \qquad \forall\ k,l = 1,...,N \qquad k \neq l$$

y las expresiones anteriores toman la forma:

$$\hat{Y}_\pi = \sum_{k=1}^{n} \frac{N}{n} y_k$$

$$Var(\hat{Y}_\pi) = N^2(1-f)\frac{S^2}{n} \qquad \text{con} \quad f = \frac{n}{N}$$

donde: $S^2 = \dfrac{\sum_{k=1}^{N}(Y_k - \bar{Y})}{N-1}$ y $\bar{Y} = \sum_{k=1}^{N}\dfrac{Y_k}{N}$

$$\hat{Var}(\hat{Y}_\pi) = N^2(1-f)\frac{s^2}{n}$$

donde: $s^2 = \dfrac{\sum_{k=1}^{n}(y_k - \bar{y})}{n-1}$ y $\bar{y} = \sum_{k=1}^{n}\dfrac{y_k}{n}$

- **Método corregido de Horvitz-Thompson**

Es utilizado con el fin de corregir la no-respuesta. El peso muestral utilizado es el inverso del producto de la probabilidad de inclusión y la tasa de respuesta, considerada en los estratos homogéneos. En este caso, la ponderación es por celdas, siendo necesaria igual probabilidad de inclusión en la muestra para elementos de una misma celda. Es decir, el peso muestral considerando la No-respuesta es:

$$\frac{1}{\pi_k} * \frac{1}{TR} \quad ; \qquad \text{TR= tasa de respuesta}$$

Los métodos de estimación que usan información auxiliar, se pueden agrupar dentro del modelo general de regresión múltiple.

1.2.2 Estimación Con Uso De Información Auxiliar

La clave en el ajuste de muestras está en aprovechar la información auxiliar, básica o complementaria, relativa a una variable o característica correlacionada con la que se estudia, para conseguir estimadores mas precisos que los calculados a partir de la muestra.

Como información auxiliar pueden utilizarse observaciones obtenidas con muestras grandes, pero no probabilísticas; o probabilísticas, pero de tamaño excesivamente pequeño, o bien relativas a la población en estudio, pero en fechas anteriores, como ocurre cuando se dispone de resultados de un censo de hace varios meses o años, o en ultimo caso de estimaciones relativas a una población diferente, pero correlacionada con la que se investiga.

Representamos por Y_i, la variable que constituye el objeto del estudio propiamente dicho y por X_i, una variable auxiliar que suponemos correlacionada con la primera. Admitimos que se conoce el total población de la variable auxiliar X_i, esto es: $\sum_{i=1}^{N} X_i$.

Por consiguiente (X_i, Y_i) representan un par de valores o medidas de las dos variables en la i-ésima unidad de la población o de la muestra.

Supongamos que se trata de estimar el total para la variable Y_i. Además del estimador directo, ampliado, expandido o inflado: $\hat{Y} = N \bullet \overline{Y}$ .Pueden utilizarse las siguientes estimaciones indirectas, casos particulares de la expresión general:

$$\hat{Y}_G = \hat{Y} + b_0(X - \hat{X})$$

en donde b_0 puede interpretarse como un coeficiente de corrección para mejorar el estimador $\hat{Y}$.

Para valores particulares de b_0 tenemos los siguientes casos:

a) $b_0 = 0$; $\hat{Y}_G = \hat{Y}$ (estimador directo, expandido o inflado).

b) $b_0 = \dfrac{\hat{Y}}{\hat{X}} = \hat{R}$

$\hat{Y}_G = \hat{Y} + \dfrac{\hat{Y}}{\hat{X}} \cdot (X - \hat{X}) = \hat{Y}_R = \hat{R} \cdot X = \hat{Y} \cdot \dfrac{X}{\hat{X}} = \hat{Y} \cdot \dfrac{N\overline{X}}{N\overline{x}} = N\overline{y} \cdot \dfrac{\overline{X}}{\overline{x}}$ (estimador de razón o por cociente).

c) $b_0 = 1$; $\hat{Y}_G = \hat{Y} + (X - \hat{X}) = \hat{Y}_D$ (estimador por diferencia).

d) $b_0 = b$; $\hat{Y}_{rg} = \hat{Y} + b(X - \hat{X})$ (estimador de regresión).En donde b representa el coeficiente de regresión de Y_i en X_i tal como se define en cualquier texto de estadistica.

- **El estimador de la Razón:**

El método de estimación de la razón trata de mejorar la acuracidad de un estimador simple expansión, utilizando información sobre una variable auxiliar X_i que supone correlacionada con la variable de estudio Y_i.

Si los puntos (X_i , Y_i) estuviesen situados sobre una línea recta que pasase por el origen de coordenadas se verificaría:

$$\frac{Y_i}{X_i} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{Y}{X} = R = \hat{R}$$

Vemos pues que, en este supuesto, cualquiera de las muestras posibles, en un esquema sin reposición y probabilidades iguales, proporcionaría la estimación:

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} \cdot X \quad [1]$$

que coincidiría con el valor población Y. Estaríamos en un caso ideal de estimador insesgado con varianza cero.

Por supuesto que las condiciones de este modelo no se verifican en la práctica. No obstante podemos suponer que un estimador del tipo [1] será tanto mas apropiado cuanto mas próximos estén los puntos (X_i, Y_i) a una recta que pase por el origen.

El estimador de la razón puede escribirse en la forma:

$$\hat{Y}_R = N \cdot \bar{y} \cdot \frac{\overline{X}}{\bar{x}} = \hat{Y} \cdot \frac{\overline{X}}{\bar{x}}$$

es decir, como un estimador de simple expansión ajustado, de una posible estimación por defecto o exceso, por el factor $\overline{X}/\bar{x}$ mayor o menor a 1.

Puesto que, en general la esperanza de un cociente de variables aleatorias no es igual al cociente de las esperanzas, tenemos:

$$E(\hat{R}) = E\left(\frac{\bar{y}}{\bar{x}}\right) \neq \frac{E(\bar{y})}{E(\bar{x})} = \frac{\overline{Y}}{\overline{X}} = R$$

y el estimador $\hat{R}$ suele ser sesgado.

Hartley y Ross(1954) obtuvieron un valor exacto del sesgo considerando la covarianza de $\hat{R}$ e $\bar{x}$.

$$\operatorname{cov}(\hat{R}, \bar{x}) = E(\hat{R}\bar{x}) - E(\hat{R}) \cdot E(\bar{x}) = \rho \sigma_{\hat{R}} \sigma_{\bar{x}} = \overline{X} \cdot (R - E(\hat{R}))$$

de donde se deduce:

$$\rho \sigma_{\hat{R}} \sigma_{\bar{x}} = R - E(\hat{R}) \qquad \frac{|B|}{\sigma_{\hat{R}}} = \rho \cdot \sigma_{\bar{x}} \leq \sigma_{\bar{x}}$$

El sesgo estimado a partir de los valores muestrales:

$$\hat{B} \approx (1-f)\hat{R}\left[\frac{1}{\bar{x}^2}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x}}{n(n-1)}-\frac{1}{\bar{y}\bar{x}}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n(n-1)}\right]$$

La varianza aproximada es:

$$V(\hat{R}) \approx \frac{1-f}{\bar{X}^2 n}\cdot(S_y^2+R^2S_x^2-2RS_{yx})$$

$$V(\hat{R}) \approx \frac{1-f}{\bar{X}^2 n(N-1)}\cdot\left[\sum_{i=1}^{N}Y_i^2+R^2\sum_{i=1}^{N}X_i^2-2R\sum_{i=1}^{N}Y_iX_i\right]$$

Supone tomar un modelo de regresión lineal en el que el parámetro independiente es nulo, es decir, se toma una recta de regresión que pasa por el origen. El elevador que se obtiene es por celdas y es la razón entre el total poblacional y el total muestral de una variable cuantitativa en dicha celda.

- **El estimador de Regresión:**

Supone una relación entre la variable objetivo Y y las variables auxiliares. El modelo de regresión generalmente supuesto entre estas variables es el modelo de regresión lineal. Este estimador de regresión es uno de los procedimientos de estimación más generales, en el que se pueden utilizar varias variables auxiliares.

Supongamos que los puntos (Y_i, X_i) con i=1,2,...,N donde X_i representa una variable auxiliar correlacionada con la Y_i estuviesen situados sobre una línea recta que no pasa por el origen.

$$Y_i = a + bX_i$$

Se verificaría:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = a + b\frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \bar{y} = a + b\bar{x}$$

$$\frac{1}{N}\sum_{i=1}^{n} Y_i = a + b \cdot \frac{1}{N}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \bar{Y} = a + b\bar{X}$$

Y por consiguiente:

$$\bar{y} - \bar{Y} = b(\bar{x} - \bar{X}) \quad \Rightarrow \quad \bar{Y} = \bar{y} + b(\bar{X} - \bar{x})$$

De la expresión anterior se deduce:

1. Si $\bar{x} = \bar{X}$, el estimador de regresión $\bar{y}_{rg}$ coincide con la media poblacional $\bar{Y}$ siendo por lo tanto $V(\bar{y}_{rg}) = 0$.
2. Si $\bar{x}$ es menor o mayor que $\bar{X}$ es de esperar, por la correlación postulada, que $\bar{y}$ es menor o mayor que $\bar{Y}$ siendo $b(\bar{x} - \bar{X})$ el sumando de ajuste para un posible estimación por defecto o por exceso.

El razonamiento anterior sugiere intentar una ganancia en acuracidad, cuando la relación entre Y_i e X_i, sea aproximadamente lineal, sin pasar la recta por el origen, utilizando el denominado estimador de regresión.

$$\bar{y}_{rg} = \bar{y} + b_0\left(\bar{X} - \bar{x}\right)$$

Que para el total seria $\hat{Y}_{rg} = N\bar{y}_{rg}$.

El estimador de regresión es, en general sesgado con un sesgo igual a la covarianza de las variables b_0 , $\bar{x}$.

En efecto:

$$E(\bar{y}_{rg}) = E(\bar{y}) + XE(b_0) - E(b_0\bar{x}) = E(\bar{y}) - (E(b_0\bar{x}) - E(b_0)\,.\,E(\bar{x}))$$
$$= \bar{Y} - \mathrm{cov}(b_0\ ,\ \bar{x})$$

Si b_0 es una constante, el estimador es insesgado y su varianza viene dada por la expresión:

$$V(\bar{y}_{rg}) = \frac{1-f}{n}\,.\left(S_y^2 + b_0^2\,S_x^2 - 2b_0\,S_{yx}\right)$$

Y el estimador insesgado de la varianza es:

$$\hat{V}(\bar{y}_{rg}) = \frac{1-f}{n}\,.\left(\hat{S}_y^2 + b_0^2\,\hat{S}_x^2 - 2b_0\,\hat{S}_{yx}\right)$$

1.3 Factores De Expansión O Elevadores

Definición: Los factores de expansión son valores numéricos calculados para cada elemento de la muestra. Cada valor representa, aproximadamente, las veces que un elemento de la muestra se repite en la población. También se les denomina elevadores o pesos muestrales.

Se denota con w_i para todo i =1,2,…,n. Los factores de expansión w_i, se calculan a partir de las probabilidades de inclusión de las unidades muestrales.

1.3.1 Factores de expansión en muestreo con Probabilidades Iguales

- ***En el muestreo con reposición***

Sea una población de tamaño N denotada por $U_1, U_2, \ldots, U_N$ cuyos valores respectivos son $Y_1, Y_2, \ldots, Y_N$.

Sean $p_i=1/N$ para todo i, las probabilidades de selección asignados por un procedimiento de selección como el Muestreo Aleatorio Simple, por ejemplo.

El parámetro Y (total poblacional) definido por $Y = \sum_{i=1}^{N} Y_i$, es estimado insesgadamente por:

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{np_i} = \sum_{i=1}^{n} w_i y_i$$

En donde el factor de expansión está dado por: $w_i = \frac{1}{np_i} = \frac{N}{n}$ que se mantiene constante para todas las unidades. Se cumple que:

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{N}{n} = N$$

- ***En el muestreo sin reposición***

Las probabilidades de inclusión de primer y segundo orden para cada unidad de la muestra están dadas por: π_i y π_{ij}, respectivamente. Donde $\pi_i = \frac{n}{N}$ y $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$.

El parámetro Y(total poblacional) definido por: $Y = \sum_{i=1}^{N} Y_i$, es estimado insesgadamente según el estimador de Horvitz-Thompson, que fue presentado anteriormente, propone el estimador:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

En donde el factor de expansión esta dado por: $w_i = \frac{1}{\pi_i} = \frac{N}{n}$ que se mantiene constante para todas las unidades de la muestra(i=1,2,...,n).

La varianza del estimador esta dado por:

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i \; y_j = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1}\right)$$

1.3.2 Factores de expansión en el Muestreo con probabilidades Desiguales

Sea una población de tamaño N denotada por $U_1, U_2, \ldots, U_N$ cuyos valores respectivos son $Y_1, Y_2, \ldots, Y_N$ y las medidas de tamaño son: $M_1, M_2, \ldots, M_N$

- ***En el muestreo con reposición***

Las probabilidades de selección para cada unidad de la población es dada por: $p_i = \frac{M_i}{\sum M_i}$ para todo i.

El parámetro Y (total poblacional) definido por $Y = \sum_{i=1}^{N} Y_i$, es estimado insesgadamente por:

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{np_i} = \sum_{i=1}^{n} w_i y_i$$

En donde el factor de expansión está dado por $w_i = \frac{1}{np_i}$ para todo i =1,2,...,N.

Es obvio que este factor no se mantiene constante para todas las unidades, pues depende de los tamaños de las unidades. Las unidades más grandes tienen mas probabilidad de selección que las pequeñas, ya que:

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{1}{np_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{p_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{M}{M_i} = \frac{M}{n} \sum_{i=1}^{n} \frac{1}{M_i} = \frac{N\overline{M}}{n} \sum_{i=1}^{n} \frac{1}{M_i}$$

- ***En el muestreo con reposición***

Las probabilidades de inclusión del primer y segundo orden para cada unidad de la muestra están dadas por: π_i y π_{ij}, respectivamente. En este esquema, las probabilidades π_i y π_{ij} son complicadas de calcular, sin embargo, existen algunos métodos que permiten el calculo aproximado de las mismas.

El parámetro Y (total poblacional) definido por $Y = \sum_{i=1}^{N} Y_i$, es estimado insesgadamente según el estimador de Horvitz-Thompson, que fue presentado anteriormente, propone el estimador:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

En donde el factor de expansión esta dado por: $w_i = \frac{1}{\pi_i}$ que se mantiene constante para todas las unidades de la muestra(i=1,2,...,n).

La varianza del estimador esta dada por:

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + \sum_{i=1}^{N} \sum_{i \neq j}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i \; y_j$$

1.4 POST-ESTRATIFICACIÓN

La post-estratificación es realizada después de haber obtenido los resultados del estudio realizado, es decir la postestratificación entra en juego luego de que nosotros ya contamos con los datos obtenidos en campo de nuestro estudio. Antes de realizar el estudio no se conocen los estratos con detalle, en la postestratificación utilizaremos variables obtenidas en campo.

La post-estratificación se utiliza con más frecuencia para corregir los efectos de la ausencia de respuestas, sin embargo también puede ser utilizada para realizar imputaciones y para el ajuste de los pesos o ponderaciones en base a los datos reales de población.

Cuando se manejan determinadas variables de estratificación puede ocurrir que no se conozca el estrato a que pertenece una unidad sino hasta después de recoger los datos.

Ejemplos típicos son las características personales como la edad, el sexo, la estatura y el nivel de educación. Los tamaños de los estratos N_h se pueden obtener de manera bastante exacta a partir de las estadísticas oficiales, pero las unidades se pueden clasificar en estratos solamente después de conocer los datos de la muestra. Por lo tanto puede suponerse que los W_h y los N_h son conocidos.

Este método se utiliza cuando se desconocen a priori las unidades que pertenecen a cada estrato. Obtenida la muestra, las unidades se asignan al estrato correspondiente. Si los pesos de estos son conocidos, se puede utilizar el estimador insesgado:

$$\hat{\bar{Y}} = \sum_{h=1}^{L} W_h \bar{y}_h$$

Cuya precisión es similar a la obtenida con la afijación proporcional, siempre que todos los n_h sean grandes, por ejemplo superiores a 30 unidades.

CAPÍTULO II

AJUSTE DE LA MUESTRA A PARTIR DE INFORMACIÓN AUXILIAR MÁXIMA

2.1 ¿Por qué se utiliza el Ajuste de la Muestra?

Se utiliza el ajuste de la muestra con la información auxiliar disponible y que este correlacionada con la variable objetivo o variable de estudio, con el propósito de mejorar la estimación de la(s) variable(s) objetivo(s) al incorporar en la estimación la información auxiliar que se tiene disponible.

2.2 Los Métodos de Ajuste de la Muestra

Existen distintos procedimientos para el ajuste, según el tipo de variables auxiliares, así como de la cantidad de información auxiliar disponible.

2.2.1 Según el tipo de Variable Auxiliar

Los métodos de ajuste pueden diferenciarse en dos bloques de variables:

- Métodos de Ajuste con Variable Auxiliar Cualitativa, cuando las variables auxiliares tienen un número de modalidades o categorías finito. Como ejemplo de este tipo de variables, tenemos el sexo, la relación con el mercado de trabajo (con modalidades como estar ocupado, parado, inactivo), estado civil, etc.

- Métodos de Ajuste con Variable Auxiliar Cuantitativa cuando las variables auxiliares son continuas que toman valores en un intervalo y que, por lo tanto, estos valores pueden ser infinitos. Como ejemplo de este tipo de variables está monto de facturación, gasto de consumo, edad, etc.

2.2.2 Según la Cantidad de Información Disponible

La siguiente clasificación vendrá determinada por la información auxiliar disponible. En todas las encuestas existe una estratificación de la población en celdas o estratos, utilizando para ello el cruce multivariante de variables cualitativas. Se pueden dar dos situaciones:

- Métodos de Ajuste con Información Auxiliar Máxima, cuando las variables auxiliares disponibles son cualitativas, la información que se define en los estratos es el número de efectivos de cada uno de estos. En el caso de poseer Máxima Información Auxiliar se dispone del número de efectivos poblacionales de cada una de las celdas de la estratificación.

Tabla N° 2.1: Distribución del número de elementos de la Población según Sexo y Nivel de Instrucción

Sexo	Nivel de Instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	4	5	4	13
Mujer	9	10	8	27
Total	13	15	12	40

Cuando las variables auxiliares son cuantitativas, la información es referente a la distribución de las variables auxiliares a lo largo de los estratos obtenidos. En el caso de Máxima Información Auxiliar, la información es el total, media,... poblacionales de las variables auxiliares sobre cada uno de los estratos.

Tabla N° 2.2: Distribución de ingresos mensual en la población según sexo y Nivel de instrucción

Sexo	Nivel de Instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	75,000	250,000	275,000	600,000
Mujer	120,000	220,000	250,000	590,000
Total	195,000	470,000	525,000	1190,000

- Métodos de Ajuste con Información Auxiliar Mínima, cuando las variables auxiliares disponibles son cualitativas, la información es referente a la distribución de las variables auxiliares a lo largo de

los estratos obtenidos. En el caso de poseer No-máxima Información Auxiliar lo que se dispone es sólo la distribución marginal univariante de las variables auxiliares.

Tabla N° 2.3: Distribución del número de elementos en la población según sexo y nivel de instrucción

Sexo	Nivel de Instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	?	?	?	13
Mujer	?	?	?	27
Total	13	15	12	40

Cuando las variables auxiliares disponibles son cuantitativas, la información que se define en los estratos es el número de efectivos de cada uno de estos y en el caso de No-máxima Información Auxiliar, la información que se tiene es el total poblacional, media,... de la variable auxiliar en cada una de las modalidades de las variables auxiliares de estratificación univariante.

Tabla N° 2.4: Distribución de ingresos mensuales en la población según sexo y nivel de instrucción

Sexo	Nivel de Instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	?	?	?	600,000
Mujer	?	?	?	590,000
Total	195,000	470,000	525,000	1190,000

2.2.3 Esquema de Clasificación según el Tipo de Variable Auxiliar y la Cantidad de Información Disponible

Un cuadro resumen de las clasificaciones de los métodos de ajuste de muestras, es el siguiente:

Tipo de Variable Auxiliar	**Información Disponible**	
	Máxima	***Mínima***
Cualitativa	Se dispone del número de Efectivos Poblacionales de cada una de las celdas de estratificación.	Se dispone solo de la distribución marginal univariante de las variables auxiliares.
Cuantitativa	Se dispone del Total, media,..,poblacionales de las variables auxiliares por cada uno de los estratos o celdas de estratificación bivariante.	Se dispone del total poblacional, media,... de la variable auxiliar en cada una de las modalidades de las variables auxiliares de estratificación univariante.

2.3 Métodos de Ajuste con Máxima Información Auxiliar

Los métodos que a continuación vamos a analizar se pueden utilizar tan sólo cuando se dispone de MÁXIMA información auxiliar, esto es, conocemos los totales de las variables auxiliares sobre cada una de las celdas de la estratificación bivariante.

Se plantean una serie de restricciones con el fin de que la distribución muestral ponderada sea igual a la distribución poblacional conjunta. A continuación vamos a introducir la notación para representar estas restricciones, diferenciando el caso, respecto del tipo de información auxiliar, en el que estemos.

2.3.1 Ajuste con Variables Auxiliares Cualitativas

La expresión matricial de las restricciones es:

$$\hat{X} \bullet w = \mathrm{X} \bullet \mathrm{I}$$

Donde el vector de pesos $w = (w_1,...,w_k,...,w_n)$ y el vector I es un vector Nx1 de unos:

$$\mathrm{I} = \begin{bmatrix} 1 \\ .. \\ 1 \\ .. \\ 1 \end{bmatrix}$$

Las matrices $\hat{X}$ y X de variables auxiliares son las matrices muestral y poblacional de dimensión (m+1)xn y (m+1)xN, respectivamente. Para la definición de las componentes de estas matrices, definimos una serie de variables:

Tenemos L variables auxiliares cualitativas, cada una con $L_1, L_2,..., L_l,..., L_L$ modalidades (categorías). Se realiza la estratificación de la población mediante el cruce multivariante de estas L variables, obteniendo $L_1 \bullet L_2 \bullet ... \bullet L_l \bullet ... \bullet L_L = m$ estratos.

Definimos las variables identificadoras o dicotómicas de los estratos:

$$X_{h1..hl..hL}(k) = \begin{cases} 1, si\ k \in (h_1,...,h_l,..,h_L) \\ 0, en\ \ caso\ contrario \end{cases}$$

$\Rightarrow L_1 \bullet ... \bullet L_l \bullet ... \bullet L_L = m$

siendo m el número de variables dicotómicas

Tenemos un vector poblacional fila

$X_{h1..hl...hL} = (X_{h1...hl...hL}(1),..., X_{h1...hl...hL}(k),..., X_{h1...hl...hL}(N))$ de dimensión (1xN)

y un vector muestral fila

$\hat{X}_{h1..hl...hL} = (X_{h1...hl...hL}(1),..., X_{h1...hl...hL}(k),..., X_{h1...hl...hL}(n))$ de dimensión (1xn)

Para asegurar que:

$$w_1 + ... + w_k + ... + w_n = N$$

Tomamos la variable X_O idénticamente 1 y se representa mediante:

$Xo = (Xo(1),..., Xo(k),..., Xo(N)) = (1,...,1,...,1)$ Vector fila poblacional de dimensión (1xN).

$\hat{X}o = (Xo(1),..., Xo(k),..., Xo(n)) = (1,...,1,...,1)$ Vector fila muestral de dimensión (1xn)

Las matrices X y $\hat{X}$ están formadas por estos vectores fila:

$$X = \begin{bmatrix} X_o \\ X_1 \\ .... \\ X_h \\ .... \\ X_m \end{bmatrix} \qquad \text{y} \qquad \hat{X} = \begin{bmatrix} \hat{X}_o \\ \hat{X}_1 \\ .... \\ \hat{X}_h \\ .... \\ \hat{X}_m \end{bmatrix}$$

En el ejemplo 1, se describe con valores las notaciones anteriores con 2 variables auxiliares cualitativas.

Ejemplo 1:

Supongamos que tenemos dos variables auxiliares (L=2) Sexo y nivel de instrucción:

- ***Sexo:*** con dos categorías (hombres =1 y mujeres =2)
- ***Nivel de instrucción:*** con tres categorías (primaria =1, secundaria =2 y universitario =3).

El tamaño de la población es de 20 individuos(N =20):

Los datos auxiliares son:

Ind(k)	Sexo	Estudio
1	1	1
2	1	1
3	1	1
4	1	1
5	2	1
6	2	1
7	2	1
8	1	2
9	1	2
10	1	2
11	1	2
12	2	2
13	2	2
14	2	2
15	2	2
16	2	2
17	1	3
18	2	3
19	2	3
20	2	3

Según la notación dada anteriormente podemos decir que:

$L_1 = 2$ y $L_2 = 3$

$\Rightarrow L_1 \bullet L_2 = 6 = m$ Celdas o estratos.

Luego el vector fila: $X_{h1,h2} = \left[X_{h1h2}(1),....,X_{h1h2}(20)\right]$ con $1 \le h_1 \le 2$, $1 \le h_2 \le 3$

Entonces la matriz X esta formado por los vectores:

$$X = \begin{bmatrix} X_0 \\ X_{11} \\ X_{12} \\ X_{13} \\ X_{21} \\ X_{22} \\ X_{23} \end{bmatrix}$$

que simplificando seria:

$$X = \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}_{7X1}$$

Definimos las variables identificadoras o dicotómicas de cada estrato (celda):

$$X_{h1h2}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (\text{h}_1, \text{h}_2) \\ 0, \text{en caso contrario} \end{cases}$$

Entonces, para $h_1 = 1$ y $h_2 = 1$ tenemos:

$$X_{11}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (1,1) \\ 0, \text{en caso contrario} \end{cases}$$

si k =1 $\Rightarrow$ $X_{11}(1) = 1 \Rightarrow$ individuo 1 es sexo =1 y estudio =1

k =2 $\Rightarrow$ $X_{11}(2) = 1 \Rightarrow$ individuo 2 es sexo =1 y estudio =1

k =3 $\Rightarrow$ $X_{11}(3) = 1 \Rightarrow$ individuo 3 es sexo =1 y estudio =1

k =4 $\Rightarrow$ $X_{11}(4) = 1 \Rightarrow$ individuo 4 es sexo =1 y estudio =1

K=5 $\Rightarrow$ $X_{11}(5) = 0 \Rightarrow$ individuo 5 es sexo =2 y estudio =1

.

.

.

k =20 $\Rightarrow$ $X_{11}(20) = 0 \Rightarrow$ individuo 20 es sexo =2 y estudio =3

El vector $X_{11}(k)$ es:

$$\Rightarrow X_{11}(k) = [1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ \ 0\ 0 \ldots 0]$$

y así sucesivamente con los demás vectores, obtendremos:

$$X_{12}(k) = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ \ldots 0]$$
$$X_{13}(k) = [0\ 0\ 0\ 0\ 0\ \ldots\ 1\ 0\ 0\ 0 \ldots 0]$$
$$X_{21}(k) = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0 \ldots 0]$$
$$X_{22}(k) = [0\ 0\ 0 \ldots 1\ 1\ 1\ 1\ 1\ 0\ 0 \ldots 0]$$
$$X_{23}(k) = [0\ 0\ 0\ 0\ 0\ 0 \ldots\ 0\ 0 \ldots 1\ 1\ 1]$$

Por lo tanto la matriz poblacional de datos auxiliares **X** es:

$$X = \begin{bmatrix} 1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,11111\,11\,1\,11 \\ 1\,1\,1\,1000000000000000 \\ 0000000\,1\,1\,1\,1000000000 \\ 00000000000000000\,1000 \\ 0000\,1\,1\,10000000000000 \\ 000000000001\,1\,1\,110000 \\ 0000000000000000001\,11 \end{bmatrix}$$

Supongamos que de la población anterior tomamos una muestra de tamaño 12 (n =12), sean los datos de la muestra:

Ind(k)	Sexo	Estudio
1	1	1
2	2	1
3	2	1
4	2	1
5	1	2
6	2	2
7	2	2
8	2	2
9	1	3
10	2	3
11	2	3
12	2	3

Siguiendo el mismo procedimiento de la matriz población X , obtenemos:

$$\hat{X}_{11}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (1,1) \\ 0, \text{en caso contrario} \end{cases}$$

si k =1 $\Rightarrow$ $\hat{X}_{11}(1) = 1 \Rightarrow$ individuo 1 es sexo =1 y estudio =1

k =2 $\Rightarrow$ $\hat{X}_{11}(2) = 0 \Rightarrow$ individuo 2 es sexo =2 y estudio =1

k =3 $\Rightarrow$ $\hat{X}_{11}(3) = 0 \Rightarrow$ individuo 3 es sexo =2 y estudio =1

k =4 $\Rightarrow$ $\hat{X}_{11}(4) = 0 \Rightarrow$ individuo 4 es sexo =2 y estudio =1

k =5 $\Rightarrow$ $\hat{X}_{11}(5) = 0 \Rightarrow$ individuo 5 es sexo =1 y estudio =2

.

.

.

k =12 $\Rightarrow$ $\hat{X}_{11}(12) = 0 \Rightarrow$ individuo 12 es sexo =2 y estudio =3

$$\Rightarrow \hat{X}_{11}(k) = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ \ 0\ 0 \ldots 0]$$

y así sucesivamente con todos los vectores, obtendremos :

$$\hat{X}_{12}(k) = [0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ \ldots 0]$$

$$\hat{X}_{13}(k) = [0\ 0\ 0\ 0\ \ \ldots\ 1\ 0\ 0\ 0 \ldots 0]$$

$$\hat{X}_{21}(k) = [0\ 1\ 1\ 1\ 0\ 0\ \ 0\ 0\ 0\ 0\ 0 \ldots 0]$$

$$\hat{X}_{22}(k) = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 0 \ldots 0]$$

$$\hat{X}_{23}(k) = [0\ 0\ 0\ 0\ 0 \ldots\ \ 0\ 0 \ldots 1\ 1\ 1]$$

Por tanto la matriz muestral de los datos auxiliares $\hat{X}$ es:

$$\hat{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Continuando con el desarrollo teórico, la expresión matricial:

$$\hat{X} \bullet w = \mathrm{X} \bullet \mathrm{I}$$

Se puede desarrollar como:

$$\begin{bmatrix} \hat{X}_o \\ \hat{X}_1 \\ .... \\ \hat{X}_h \\ .... \\ \hat{X}_m \end{bmatrix} \bullet \begin{bmatrix} w_1 \\ w_2 \\ .... \\ .... \\ .... \\ w_n \end{bmatrix} = \begin{bmatrix} X_o \\ X_1 \\ .... \\ X_h \\ .... \\ X_m \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ ... \\ ... \\ ... \\ 1 \end{bmatrix}$$

Obtenemos el siguiente sistema de ecuaciones:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet Xo(k) = \sum_{k=1}^{N} Xo(k) \Leftrightarrow w_1 + ... + w_k + ... + w_n = N \\ \sum_{k=1}^{n} w_k \bullet X_{h1...hl...hL}(k) = \sum_{k=1}^{N} X_{h1...hl...hL}(k), \ \forall 1 \le h_1 \le L_1, ..., 1 \le h_l \le L_l, ... 1 \le hL_L \le L_L \end{cases}$$

Al ser las variables identificadoras de los estratos o celdas, los datos auxiliares totalizados toman la forma:

$$\text{a nivel poblacional}: \ \sum_{k=1}^{N} X_{h1...hl...hL}(k) = N_{h1..hl..hL}$$

$$\text{a nivel muestral}: \ \sum_{k=1}^{n} X_{h1...hl...hL}(k) = n_{h1..hl..hL}$$

En el ejemplo 1: $\sum_{k=1}^{20} X_{h1h2}(k) = N_{h1h2}$, si $h_1 = 1, h_2 = 1 \ \Rightarrow \ \sum_{k=1}^{20} X_{11}(k) = 4 = N_{11}$.

Además $\sum_{h1=1}^{L1} \sum_{h2=1}^{L2} N_{h1h2} = N$, en el ejemplo $\sum_{h1=1}^{2} \sum_{h2=1}^{3} N_{h1h2} = 20$.

Resolvemos el sistema, teniendo en cuenta que es un ajuste de celdas y por lo tanto, todos los individuos de un mismo estrato tienen el mismo peso[1], por lo que el sistema pasa a tener como incógnitas $W_{h1...hl...hL}$: elevadores asignados a las celdas. El sistema es el siguiente:

$$\begin{cases} W_{h1...hl...hL} \bullet n_{h1...hl...hL} = N_{h1...hl...hL} \quad \forall 1 \le h_1 \le L_1, ..., 1 \le h_l \le L_l, ..., 1 \le h_L \le L_L \\ \sum_{h1..hl..hL} W_{h1..hl...hL} \bullet n_{h1...hl...hL} = N \end{cases}$$

(1) Los indicadores de las variables se han simplificado a un solo índice, sin más que establecer un orden en las variables

Se trata de un sistema de $L_1 \bullet ... \bullet L_l \bullet ... \bullet L_L = m$ incógnitas y $L_1 \bullet ... \bullet L_l \bullet ... \bullet L_{L+1}$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema mxm, en el que la matriz asociada es una matriz A_{mxm} diagonal. Los elementos diagonales de esta matriz son el número de efectivos muéstrales en cada celda de la estratificación.

Tras lo anterior, el sistema tiene una única solución $\Leftrightarrow n_{h1...hl...hL} \neq 0 \quad \forall\, h_1..h_l..h_L$

Bajo estas condiciones, la solución del sistema es:

$$\hat{W}_{h1..hl..hL} = \frac{N_{h1..hl..hL}}{n_{h1..hl..hL}}$$

$$\forall\ 1 \le h_1 \le L_1,...,1 \le h_l \le L_l,...,1 \le h_L \le L_L$$

Luego tenemos que:

$$Valor\ ajustado = \sum Ponderacion \bullet Valor\ muestral$$

La estimación resultante tras utilizar estos elevadores es la estimación de Estratificación, en la que el muestreo se supone aleatorio estratificado. Esta estimación resulta del ajuste a los datos poblacionales que teníamos y al ajuste en tales condiciones le llamamos ajuste de ESTRATIFICACIÓN. Un caso particular de este tipo de ajuste es cuando la estratificación de la población no se realiza antes de la muestra, sino tras la muestra. En este último ajuste la

estimación que se obtiene se denomina estimación de Post-estratificación y el método de POST-ESTRATIFICACIÓN. En ambos, los elevadores de las celdas son el cociente entre el número de efectivos poblacionales y muéstrales para cada una de ellas.

2.3.1.2 Procedimiento de estimación

Analicemos las propiedades del estimador que hemos considerado en el caso de máxima información auxiliar, es decir, para el estimador de post-estratificación. El estudio lo realizaremos tomando como estadístico la Media Poblacional de la variable objetivo.

El estimador que se obtiene en el muestreo aleatorio estratificado para la media poblacional es:

$$\hat{Y}_{st} = \sum_{h=1}^{m} W_h \bullet \bar{y}_h \quad , \quad \text{donde } W_h = \frac{N_h}{N}^2 \quad , \quad \bar{y}_h = \frac{\sum_{i=1}^{n_h} y_i}{n_h}$$

Siendo $\bar{y}_h$ la media muestral sobre el estrato h. Esta estimación se obtiene de forma análoga tanto en el caso de una estratificación Pre-muestral como en el de estratificación post-muestral (post-estratificación).

Consideramos en un primer momento que tenemos una estratificación Pre-muestral y los valores n_h son, por lo tanto, fijos.

(2) Tomamos una numeración de los estratos resultantes del cruce, en la que intervendrá ya un sólo subíndice.

Para considerar este estimador necesitamos conocer $W_h = \frac{N_h}{N}$, o, lo que es lo mismo, N_h (N es constante) $\forall h = 1,.....,m$.

Conocidos estos datos, estimamos la media poblacional de Y sobre cada celda: $\bar{y}_h = \frac{y_h}{n_h}$, donde y_h se considera ahora el total muestral sobre cada celda de la variable objetivo Y.

Sustituimos esta expresión en el estimador y obtenemos la estimación:

$$\hat{Y}_{st} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{n_h}$$

Ésta es la estimación de estratificación de la media poblacional de Y. A partir de ésta se obtiene fácilmente la estimación del total poblacional:

$$\hat{Y}_{st} = \hat{\bar{Y}}_{st} \bullet N$$

Sustituyendo de nuevo en la expresión anterior, obtenemos:

$$\frac{\hat{Y}_{st}}{N} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{n_h} \quad \Leftrightarrow \quad \hat{Y}_{st} = \sum_{h=1}^{m} \frac{N_h}{n_h} \bullet y_h$$

Vamos a estudiar ahora el error que se produce al tomar la estimación de estratificación para la media poblacional:

$$\text{var}(\hat{Y}_{st}) = \text{var}(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet \text{var}(\bar{y}_h) + 2 \bullet \sum_{h=1}^{m} \sum_{j>h}^{m} \text{cov}(\bar{y}_h, \bar{y}_j)$$

y el error estándar es: $\sqrt{\text{var}(\hat{\bar{Y}}_{st})}$.

Al tratarse de un muestreo aleatorio estratificado, tenemos que el término de las covarianzas desaparece, debido a la independencia del muestreo en cada uno de los estratos y la expresión de la varianza toma la forma:

$$\text{var}(\hat{\bar{Y}}_{st}) = \text{var}(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet \text{var}(\bar{y}_h)$$

El muestreo es aleatorio simple en cada uno de los estratos y la varianza en cada uno de ellos es:

$$\text{var}(\bar{y}_h) = \frac{S_h^2}{n_h} \bullet \frac{N_h - n_h}{N_h}$$

De esta forma, la expresión final de la varianza es

$$\text{var}(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{m} \frac{W_h^2 \bullet S_h^2}{n_h} - \frac{\sum_{h=1}^{m} W_h \bullet S_h^2}{N}$$

Siendo S_h la varianza verdadera de la variable Y sobre el estrato h.

Estimamos la varianza en cada estrato por:

$$\hat{\text{var}}(\bar{y}_h) = \frac{s_h^2}{n_h} \bullet \frac{N_h - n_h}{N_h}$$

Con s_h la varianza muestral de Y sobre el estrato h. Y a continuación obtenemos una estimación de la varianza del estimador de estratificación para la media poblacional:

$$\hat{\text{var}}(\hat{Y}_{st}) = \sum_{h=1}^{m} \frac{W_h^2 \bullet s_h^2}{n_h} - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet s_h^2$$

En la expresión de la estimación del error se ve, claramente, que el error de la estimación depende de la varianza de la variable objetivo Y en cada estrato, por lo que si estos estratos son homogéneos respecto de la variable objetivo, la varianza total será pequeña, siempre que todos los n_h sean lo suficientemente grandes. De esta forma se podrá controlar relativamente el error.

El control de la varianza de la variable objetivo sobre las celdas se puede conseguir tomando las variables de estratificación lo más altamente correlacionadas con la variable objetivo. Los estratos resultantes mediante el cruce multivariante de las variables serán de esta forma homogénea respecto de la variable objetivo.

De forma análoga consideramos el **estimador de post-estratificación.** En este caso los valores $\hat{n}_h$ son aleatorios, dependen de la muestra, y los denotamos como $\hat{n}_h$. Una vez tomada la muestra, estos valores $\hat{n}_h$, que exceden todos a cero, son fijos.

El estimador que obtenemos de la media poblacional toma forma análoga al estimador de estratificación:

$$\bar{Y}_w = \sum_{h=1}^{m} W_h \bullet \bar{Y}_h$$

En cada muestra, tomamos la media muestral en cada estrato $\bar{y}_h = \frac{y_h}{\hat{n}_h}$ y obtenemos el estimador de post-estratificación:

$$\hat{\bar{Y}}_W = \sum_{h=1}^{m} W_h \bullet \bar{y}_h = \sum_{h=1}^{m} W_h \bullet \frac{y_h}{\hat{n}_h} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{\hat{n}_h} = \frac{1}{N} \bullet \sum_{h=1}^{m} \frac{N_h}{\hat{n}_h} \bullet y_h$$

Una vez realizada la muestra, como ya hemos dicho, estos $\hat{n}_h$ son fijos, por lo que la estimación de la varianza para este estimador es igual que para el estimador de estratificación y la expresión que tenemos es:

$$\mathrm{var}(\hat{\bar{Y}}_W) = \mathrm{var}(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet \mathrm{var}(\bar{y}_h) = \sum_{h=1}^{m} \frac{W_h^2 \bullet S_h^2}{\hat{n}_h} - \frac{1}{N} \sum_{h=1}^{m} W_h \bullet S_h^2$$

Ahora bien, esta varianza no es fija, desde el momento que los $\hat{n}_h$ no son fijos y debemos hallar por lo tanto un valor esperado de dicha varianza. Este valor esperado se halla sin más que tomar la esperanza:

$$E\left[\text{var}(\hat{Y}_W)\right] = E\left[\sum_{h=1}^{m} \frac{W_h^2 \bullet S_h^2}{\hat{n}_h} - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet S_h^2\right]$$

$$= \sum_{h=1}^{m} W_h^2 \bullet S_h^2 \bullet E\left[\frac{1}{\hat{n}_h}\right] - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet S_h^2$$

Ahora bien $E\left[\frac{1}{\hat{n}_h}\right] = \frac{1}{n \bullet W_h} + \frac{1 - W_h}{n^2 \bullet W_h^2}$, esto fue demostrado por Stephan(1945) con lo que la expresión anterior se reduce a:

$$E\left[\text{var}\left(\hat{\bar{Y}}_w\right)\right] = \frac{1-f}{n} \sum_{h=1}^{m} W_h \bullet S_h^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m} (1 - W_h) \bullet S_h^2$$

El primer término corresponde a la varianza del estimador de estratificación $\text{var}(\bar{Y}_{st})$, siendo el segundo el incremento producido en la varianza al tomar el estimador de postestratificación. Desarrollando este segundo término obtenemos la expresión:

$$\frac{1}{n^2} \bullet \sum_{h=1}^{m} (1 - W_h) \bullet S_h^2 = \frac{1}{n \bullet \bar{n}_h} \bullet \bar{S}_h^2 - \frac{1}{n^2} \bullet \sum_{h=1}^{m} W_h \bullet S_h^2$$

Donde $\bar{n}_h = \frac{n}{m}$ es el número promedio de unidades muéstrales por estrato y $\bar{S}_h^2$ es el promedio de las S_h^2. Se obtiene finalmente, que el incremento de la varianza obtenido al tomar el estimador de post-estratificación se mantiene pequeño siempre que $\bar{n}_h$ se mantenga razonablemente grande.

La estimación del promedio de la varianza que se obtiene es:

$$\hat{E}\left[\text{var}\left(\hat{\bar{Y}}_W\right)\right] = \frac{1-f}{n}\sum_{h=1}^{m} W_h \bullet s_h^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m}\left(1-W_h\right) \bullet s_h^2$$

En este caso de post-estratificación, el control de la varianza de la variable objetivo sobre las celdas se puede conseguir tomando las variables auxiliares cualitativas altamente correlacionadas con la objetivo. Los estratos resultantes mediante el cruce multivariante de las variables auxiliares serán de esta forma homogéneos respecto de la variable objetivo.

Concluimos que el error cometido, tanto con la estimación de estratificación como con la postestratificación, es directamente proporcional a la varianza de la variable sobre las celdas e inversamente proporcional al tamaño muestral de los estratos.

Ahora si lo que deseamos es estimar la varianza del total de la variable objetivo, solo debemos multiplicar por el cuadrado del tamaño de la población para obtener la varianza del total a partir de la varianza de la media, es decir:

$$N^2 \hat{E}\left[\text{var}\left(\hat{Y}_W\right)\right] = N^2 * \frac{1-f}{n} \sum_{h=1}^{m} W_h \bullet s_h^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m} \left(1 - W_h\right) \bullet s_h^2$$

Veamos el ejemplo 2, en donde se realiza una estimación basada en los datos auxiliares del ejemplo 1.

Ejemplo 2:

Consideraremos que el objetivo de la investigación es estimar el ingreso promedio mensual y el porcentaje de trabajadores dependientes, para ello haremos uso de dos variables auxiliares: el sexo y el nivel de estudios(ejemplo 1), entonces tenemos las variables objetivos:

- Y_1: Ingreso mensual por trabajador.
- Y_2: Tipo de trabajo: donde (0: dependiente y 1: independiente).

Las variables auxiliares son:

- Sexo: tiene dos categorías L_1 =2
- Nivel de estudios: tiene tres categorías L_2 =3

Se tiene una población de 20 individuos y se toma de esta población una muestra de tamaño 12 (n =12).

Obtenemos por lo tanto $L_{1x}L_2$ =6 celdas, para lograr el objetivo de la investigación se definirá 6 variables dicotómicas.

$$X_{h1h2}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (\text{h}_1, \text{h}_2) \\ 0, \text{en caso contrario} \end{cases}$$

Los datos muéstrales son:

Ind(k)	Sexo	N. Estudios	Ingreso Mensual	Tipo de Trabajo
1	1	1	6000	0
2	2	1	100000	1
3	2	1	110000	1
4	2	1	90000	1
5	1	2	90000	1
6	2	2	100000	0
7	2	2	137500	1
8	2	2	137500	1
9	1	3	150000	0
10	2	3	200000	0
11	2	3	200000	0
12	2	3	200000	0

Al agrupar los datos anteriores en celdas obtenemos:

Tabla N° 2.5: Variables auxiliares muestrales $\left(n_{hh'}\right)$

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	1	1	1	3
Mujer	3	3	3	9
Total	4	4	4	12

Tabla N° 2.6: Variable Objetivo T_{y_1}: Ingreso total mensual de los trabajadores

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	60,000	90,000	150,000	300,000
Mujer	300,000	375,000	600,000	1,275,000
Total	360,000	4,650,000	750,000	1,575,000

Tabla N° 2.7: Número total de trabajadores independientes donde(0: dependiente y 1: independiente)

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	0	1	0	1
Mujer	3	2	0	5
Total	0	3	0	6

Además, a partir del ejemplo 1, tenemos las variables auxiliares poblacionales organizadas en celdas:

Tabla N° 2.8: Variables auxiliares en la población ($N_{hh'}$)

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	4	4	1	9
Mujer	3	5	3	11
Total	7	9	4	20

La matriz que obtendremos para las variables auxiliares es una matriz de 7xN.

$$X = \begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}_{7X1}$$

Luego, obtenemos las restricciones:

$$\begin{bmatrix} N \\ \sum_{k=1}^{N} X_{11}(k) \\ \cdots\cdots \\ \sum_{k=1}^{N} X_{hh'}(k) \\ \\ \cdots\cdots \\ \\ \sum_{k=1}^{N} X_{23}(k) \end{bmatrix} = \begin{bmatrix} 20 \\ 4 \\ 4 \\ \\ 1 \\ 3 \\ 3 \\ \\ 5 \\ 3 \\ \end{bmatrix}$$

Para obtener los elevadores respectivos aplicaremos la siguiente formula (9) a partir de las tablas N° 2.8 y 2.5.

$$\hat{W}_{hh'} = \frac{N_{hh'}}{n_{hh'}}$$

Tabla N° 2.9: Elevadores

Sexo	Nivel de Estudios		
	Primaria	Secundaria	Universitario
Hombre	4=4/1	4=4/1	1=1/1
Mujer	1=3/3	1.667=5/3	1=3/3

Ahora, estimaremos el ingreso total mensual, multiplicando la tabla N° 2.6 * tabla N° 2.9, el cual resultara:

Tabla N° 2.10: Estimación de la variable ingreso mensual

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	240,000	360,000	150,000	750,000
Mujer	300,000	625,125	600,000	1,525,125
Total	540,000	985,125	750,000	2,275,125

Luego, haremos lo mismo para el total de personas independientes tabla N° 2.7* tabla N° 2.9:

Tabla N° 2.11: Estimación de la variable total de personas independientes

Sexo	Nivel de Estudios			Total
	Primaria	Secundaria	Universitario	
Hombre	0	4	0	4
Mujer	3	3.34	0	6
Total	3	7	0	10.34

Con lo que tenemos que las estimaciones del total tanto para la variable objetivo cuantitativa INGRESO TOTAL como para la variable objetivo cualitativa TRABAJADORES INDEPENDIENTE, son:

$\hat{Y}_1 =$ 2,275,125, es decir el ingreso total estimado de la población es 2,275,125. El número total de trabajadores independientes($\hat{Y}_2$) es aproximadamente de 10 individuos en la población. Así mismo podemos obtener fácilmente que el ingreso promedio estimado en la población es de 1,113,756 y la proporción de trabajadores independientes en la población es de 50%(0.5). Siguiendo el ejemplo, realizaremos la estimación convencional sin uso de información auxiliar, es decir sin utilizar ningún ajuste, entonces obtendremos:

$$T_{y_1} = \sum_{i=1}^{N} Y_i$$

el parámetro total poblacional a estimar. El estimador del total poblacional es:

$$\hat{T}_{y_1} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

$$\Rightarrow \hat{T}_{y_1} = \frac{20}{12}(1521000) \quad \hat{T}_{y_1} = 2,535,000 \text{(ingreso total estimado)}$$

Con este resultado podemos observar que el ajuste que realizamos utilizando información auxiliar difiere de la estimación tradicional, se espera que el ajuste mejore la estimación del parámetro.

2.3.2 Variables Auxiliares Cuantitativas

En este caso, supondremos que se tiene una única variable auxiliar cuantitativa X. La estratificación de la población se realiza mediante variables antes del muestreo: Pre-estratificación.

Las restricciones toman la misma forma matricial, con unas nuevas matrices, que se definirán a continuación:

$$(\hat{X}^*) \bullet w = (X^*) \bullet \mathrm{I}$$

donde, como antes, el vector de pesos es $w' = (w_1,...,w_k,...,w_n)$ y el vector I es un vector Nx1 de 1′:

$$\mathrm{I} = \begin{bmatrix} 1 \\ ... \\ 1 \\ ... \\ 1 \end{bmatrix}$$

Para definir las matrices X^* y $\hat{X}^*$ definimos las variables auxiliares:

$$X^*_{h1...hl...hL}(k) = X_{h1...hl...hL}(k) \bullet X(k) \;\; \forall \;\; 1 \le h_1 \le L_1,...,1 \le h_l \le L_l,...1 \le h_L \le L_L$$

con $X_{h1...hl...hL}(k)$ la variable dicotómica identificadora de cada uno de los estratos resultantes y X(k) es el valor de la variable cuantitativa para cada uno de los elementos poblacionales.

Definimos también: $X^{*}{}_{o}(k) = X_{o}(k) \bullet X(k) = X(k)$

Obtenemos los vectores:

$X^{*}{}_{h1\ldots hl\ldots hL} = (X^{*}{}_{h1..hl..hL}(1),\ldots, X^{*}{}_{h1..hl..hL}(k),\ldots, X^{*}{}_{h1..hl..hL}(N)$ Vector fila de dimensión (1xN)

$\hat{X}^{*}{}_{h1\ldots hl\ldots hL} = (\hat{X}^{*}{}_{h1..hl..hL}(1),\ldots, \hat{X}^{*}{}_{h1..hl..hL}(k),\ldots, \hat{X}^{*}{}_{h1..hl..hL}(n)$ Vector muestral de dimensión (1xn)

Estos vectores son las componentes filas de las matrices X^{*} y $\hat{X}^{*}$. La forma que toman es:

$$X^{*} = \begin{bmatrix} X_{o}^{*} \\ X_{1}^{*} \\ \ldots\ldots \\ X^{*}{}_{k} \\ \ldots \\ X^{*}{}_{m} \end{bmatrix} \qquad \hat{X}^{*} = \begin{bmatrix} \hat{X}_{o}^{*} \\ \hat{X}_{1}^{*} \\ \ldots\ldots \\ \hat{X}^{*}{}_{k} \\ \ldots \\ \hat{X}^{*}{}_{m} \end{bmatrix}$$

En el ejemplo 3, se describe la notación anterior utilizando una variable auxiliar cuantitativa.

Ejemplo 3:

Supongamos que tenemos una variable auxiliar:

X(k): Ingreso total mensual.

La estratificación de la población la realizaremos en función a dos variables auxiliares:

- ***Sexo***: tiene dos categorías (1: hombre, 2: mujer) L_1 =2
- ***Nivel de estudios:*** tiene 3 categorías, L_2=3 (1:primaria, 2:secundaria y 3:superior).

Consideraremos el tamaño de la población N =20 y el tamaño de la muestra n =12, obtenemos $L_1 * L_2$ =6 variables con $1 \le h_1 \le 2,\ 1 \le h_2 \le 3$. Los datos de la población son:

ind(k)	Sexo	N. Estudios	Ingreso
1	1	1	110,000
2	1	1	110,000
3	1	1	110,000
4	1	1	110,000
5	2	1	140,000
6	2	1	140,000
7	2	1	100,000
8	1	2	110,000
9	1	2	130,000
10	1	2	110,000
11	1	2	110,000
12	2	2	90,000
13	2	2	90,000
14	2	2	70,000
15	2	2	80,000
16	2	2	90,000
17	1	3	420,000
18	2	3	140,000
19	2	3	140,000
20	2	3	120,000

Según la notación anterior, se define:

$$X^{*}_{h1h2}(k) = X_{h1h2}(k) \bullet X(k) \quad k = 1,2,3,....,20$$

donde:

$$X_{h1h2}(k) = \begin{cases} 1 \text{ si k } \in \text{a la celda } (h_1, h_2) \\ 0 \text{ caso contrario} \end{cases}$$

Sea $h_1 = 1, h_2 = 1$

$X_{11}(1) = 1 \quad X_{11}(2) = 1 \quad X_{11}(3) = 1 \quad X_{11}(4) = 1 \quad X_{11}(5) = 0 \ldots\ldots\ldots X_{11}(20) = 0$

y

$X_{11}^*(1) = X_{11}(1) \bullet X(1) = 1 * 110000 = 110000$

$X_{11}^*(2) = X_{11}(2) \bullet X(2) = 1 * 110000 = 110000$

$X_{11}^*(3) = X_{11}(3) \bullet X(3) = 1 * 110000 = 110000$

$X_{11}^*(4) = X_{11}(4) \bullet X(4) = 1 * 110000 = 110000$

$X_{11}^*(5) = X_{11}(5) \bullet X(5) = 0 * 110000 = 0$

$\hat{X}_{11}^*(20) = \hat{X}_{11}(20) \bullet X(20) = 0 * 120000 = 0$

$X_{11}^*(k) = (110000\ 110000\ 110000\ 110000 \quad 0\ 0\ 0 \ldots\ldots\ldots\ldots\ldots 0)$

y así sucesivamente, obtendremos:

$X_{21}^*(k) = (0 \quad 0 \quad 0 \quad 0 \ldots\ldots\ldots\ldots 140000\ 140000\ 100000\ 0 \quad 0\ 0 \ldots\ldots\ldots\ldots\ldots 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)$

$X_{12}^*(k) = (0 \quad 0 \quad 0 \quad 0 \ldots\ldots\ldots\ldots 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 140000\ 160000\ 140000 \ldots\ldots\ldots\ldots\ldots 0 \quad 0 \quad 0 \quad)$

$X_{22}^*(k) = (0 \quad 0 \quad 0 \quad 0 \ldots\ldots\ldots\ldots 0 \quad 0 \quad 0 \quad 0 \quad 9000 \quad 9000\ 7000\ 7000\ 8000\ 9000 \ldots\ldots\ldots\ldots\ldots \quad 0)$

$X_{23}^*(k) = (0 \quad 0 \quad 0 \quad 0 \ldots\ldots\ldots\ldots 0 \quad 0 \quad 0\ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 140000\ 140000\ 120000)$

Supongamos que se toma una muestra (n =12), sean los datos de la muestra:

ind(k)	Sexo	Nivel de Estudios	Ingreso	Ahorro	Tipo de Trabajo
1	1	1	75,000	20,000	0
2	2	1	400,000	90,000	1
3	2	1	400,000	80,000	1
4	2	1	400,000	80,000	1
5	1	2	250,000	175,000	1
6	2	2	80,000	600,000	1
7	2	2	80,000	700,000	1
8	2	2	60,000	450,000	0
9	1	3	275,000	200,000	0
10	2	3	80,000	20,000	0
11	2	3	80,000	30,000	0
12	2	3	90,000	20,000	0

Según la notación anterior, se define:

$$\hat{X}^{*}_{h1h2}(k) = \hat{X}_{h1h2}(k) \bullet \hat{X}(k) \quad k = 1,2,3,....,12$$

Donde:

$$\hat{X}_{h1h2}(k) = \begin{cases} 1 \text{ si k } \in \text{ a la celda } (\mathrm{h}_1, \mathrm{h}_2) \\ 0 \text{ caso contrario} \end{cases}$$

Sea $h_1 = 1, h_2 = 1$

$$\hat{X}_{11}(1) = 1 \quad \hat{X}_{11}(2) = 1 \quad \hat{X}_{11}(3) = 1 \quad \hat{X}_{11}(4) = 1 \quad \hat{X}_{11}(5) = 0.....\hat{X}_{11}(20) = 0$$

y

$$\hat{X}^{*}_{11}(1) = \hat{X}_{11}(1) \bullet X(1) \qquad \hat{X}^{*}_{11}(2) = \hat{X}_{11}(2) \bullet X(2) \qquad \hat{X}^{*}_{11}(3) = \hat{X}_{11}(3) \bullet X(3)$$
$$= 1 * 750000 = 750000 \qquad = 0 * 400000 = 0 \qquad = 0 * 400000 = 0$$

$$\hat{X}^{*}_{11}(4) = \hat{X}_{11}(4) \bullet X(4) \quad X^{*}_{11}(5) = X_{11}(5) \bullet X(5) \quad \cdots\cdots\cdots \quad \hat{X}^{*}_{11}(12) = \hat{X}_{11}(12) \bullet X(12)$$
$$= 0 * 400000 = 0 \qquad = 0 * 110000 = 0 \qquad = 0 * 90000 = 0$$

$$X^{*}_{11}(k) = \left(750000 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0.....0\right)$$

y así sucesivamente, obtendremos:

$$X^{*}_{21}(k) = (0 \quad 40000 \quad 40000 \quad 250000.....0\,0\,0\,0.....0\,0 \quad 0\,0\,0.....0)$$

$$X^{*}_{12}(k) = (0 \quad 0 \quad 0 \quad 0.....0 \quad 80000 \quad 80000 \quad 6000\,0 \quad 0 \quad 0 \quad 0.....0)$$

$$X^{*}_{22}(k) = (0 \quad 0 \quad 0 \quad 0.....0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 2750000 \quad 0\,0\,0\,0\,0.....0)$$

$$X^{*}_{23}(k) = (0 \quad 0 \quad 0 \quad 0.....0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 80000\,80000\,90000)$$

Desarrollamos la expresión matricial de las restricciones y obtenemos el sistema:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet X^{*}{}_{h1\ldots hl\ldots hL}(k) = \sum_{k=1}^{N} X^{*}{}_{h1..hl..hL}(k) \;\forall 1 \le h_1 \le L_1,...,1 \le h_l \le L_l,...1 \le hL_L \le L_L \\ \sum_{k=1}^{n} w_k \bullet X^{*}_{o}(k) = \sum_{k=1}^{N} X^{*}_{o}(k) \Leftrightarrow \sum_{k=1}^{n} w_k \bullet \hat{X}^{*}_{0}(k) = \sum_{k=1}^{N} X(k) \end{cases}$$

Resolvemos el sistema teniendo en cuenta que es un ajuste por celdas y por lo tanto todos los individuos tienen el mismo peso por lo que el sistema pasa a tener como incógnitas $\hat{W}_{h1..hl..hL}$ elevadores asignados a las celdas. El sistema es el siguiente:

$$\begin{cases} \sum_{k=1}^{n} W_k \bullet \hat{X}^{*}_{0}(k) = \sum_{k=1}^{N} X(k) \\ \sum_{k=1}^{n} W_{h1.....hL} \bullet \hat{X}^{*}{}_{h1\ldots hl\ldots hL}(k) = \sum_{k=1}^{N} X^{*}{}_{h1..hl..hL}(k) \;\forall 1 \le h_1 \le L_1,...,1 \le h_l \le L_l,...1 \le hL_L \le L_L \\ W_{h1....hL} \bullet \sum_{k=1}^{n} \hat{X}^{*}_{h1....hL}(k) = \sum_{k=1}^{N} X^{*}_{h1....hL}(k) \end{cases}$$

Es un sistema de ecuaciones $L_1 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L = m$ incógnitas y $L_1 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_{L+1}$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema m*m, en el que la matriz asociada es una matriz $A \in (mxm)$ diagonal:

$$\sum_{k=1}^{n} X^*_{h1\ldots hl..hL}(k)$$

que serán no nulos si: La suma de la variable auxiliar X sobre cada estrato es distinta de cero, lo que se puede conseguir pidiendo que la variable sea >0(positivas). Los efectivos muestrales en cada estrato son no nulos. La primera restricción no influye, ya que normalmente, las variables cuantitativas que se analizan son variables positivas, por ejemplo las económicas: monto de facturación, gastos de consumo, etc.

Verificadas estas condiciones de no-singularidad para la matriz del sistema, tenemos que la única solución que se obtiene es:

$$\hat{W}^*_{h1..hl..hL} = \frac{\sum_{k=1}^{N} \hat{X}^*_{h1..hl..hL}}{\sum_{k=1}^{n} \hat{X}^*_{h1..hl..hL}}$$

Estos $\hat{W}^*_{h1..hl..hL}$ vuelven a ser los elevadores.

El método resultante es el método de la RAZÓN , en el que tomamos como factores de elevación de las celdas los cocientes del total poblacional de una variable cuantitativa en dicha celda entre su total muestral. Hemos visto entonces que en ambos casos, tanto con información auxiliar cualitativa como cuantitativa, el método de ponderación nos da un único elevador para cada estrato, siempre que en cada estrato de la estratificación exista algún elemento muestral.

2.3.2.1 Procedimientos de Estimación

Analicemos las propiedades del estimador que hemos considerado en el caso de máxima información auxiliar, es decir, para el estimador de la razón. El estudio lo realizaremos tomando como parámetro de interés la Media Poblacional de la variable objetivo.

Hay dos tipos de estimadores para la media poblacional basada en la razón, en muestreo aleatorio estratificado:

1) Estimador de la media basado en la razón separada

2) Estimador de la media basado en la razón combinada

El estimador que vamos a tomar es el estimador de razón separada. Este estimador corresponde a un modelo de regresión en el que la pendiente de la recta de regresión no es necesariamente la misma para cada una de las celdas de la estratificación [(3)] .

[(3)]El estimador de razón combinada, sin embargo, supone la misma pendiente de regresión para todas las celdas. Nosotros consideraremos el Estimador de la Razón Separada.

La media poblacional es:

$$\bar{Y} = \sum_{h=1}^{m} W_h^* \bullet \bar{Y}_h = \sum_{n=1}^{m} \frac{N_h}{N} \bar{Y}_h$$

siendo $\bar{Y}_h$ la media poblacional sobre el estrato h:

$$\bar{Y}_h = \sum_{k=1}^{N_h} \frac{Y_{hk}}{N_h}$$

El estimador de la media basado en la razón separada, es:

$$\hat{\bar{Y}}_{RS} = \sum_{h=1}^{m} W_h \hat{\bar{Y}}_{R_h}$$

Donde:

$$\hat{\bar{Y}}_{RS} = \sum \hat{R}_h \bar{X}_h^* , \qquad \bar{X}_h^* = \frac{\sum_{K=1}^{N_h} X_{hk}^*}{N_h} , \qquad \hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h^*} \quad , \qquad \bar{x}_h^* = \frac{\sum_{k=1}^{n_h} x_{hk}^*}{n_h} ,$$

$$W_h = \frac{N_h}{N}$$

Con $\bar{y}_h = \sum_{k=1}^{n_h} \frac{y_{hk}}{n_h}$

Respecto a la varianza de este estimador, tenemos que en el caso de muestreo aleatorio estratificado con tamaños muéstrales suficientemente grandes en todos los estratos, la varianza de este estimador es:

$$\operatorname{var}\left(\hat{\bar{Y}}_{Rs}\right) = \sum_{h=1}^{m} W_h \frac{(1-f_h)}{n_h} \bullet \left(Sy_h^2 + R_h^2 Sx_h^2 - 2R_h \rho_h Sy_h Sx_h\right) = \sum_{h=1}^{m} W_h \; Var(\hat{\bar{Y}}_{Rh})$$

Dado que se utiliza el estimador de la razón en cada celda de la estratificación, se toma la varianza de dicho estimador sobre cada una de estas celdas, bajo un muestreo aleatorio simple, de la forma:

$$\mathrm{var}(\hat{\bar{Y}}_{Rh}) = \frac{1-f_h}{n_h} \bullet \left[\frac{\sum_{k=1}^{N_k} (Y_{hk} - R_h x_{hk})^2}{N_h - 1} \right]$$

con R_h la razón entre medias muestrales de Y y X en el estrato h, es decir, la pendiente correspondiente a la recta de regresión entre la variable Y y X para el estrato h :

$$R_h = \frac{\sum_{k=1}^{N_h} Y_{hk}}{\sum_{k=1}^{N_h} X_{hk}} = \frac{\bar{Y}_h}{\bar{X}_h}$$

Al igual que en el caso de variables auxiliares cualitativas, concluimos que:

La varianza del estimador de la Razón Separado es inversamente proporcional al tamaño muestral de las celdas de la estratificación y directamente proporcional a la varianza del estimador de la razón sobre cada una de estas celdas.

El estimador de la varianza de $\hat{\bar{Y}}_{RS}$ es:

$$\hat{Var}(\hat{\bar{Y}}_{RS}) = \sum_{h=1}^{m} W_h^2 \left(\frac{1-f_h}{n_h}\right) \frac{\sum_{i=1}^{n_h} (y_{hi} - \hat{R}_h x_h)^2}{n_h - 1}$$

Veamos el ejemplo 4, en donde se realiza una estimación basada en el ejemplo 3.

EJEMPLO 4:

Consideraremos dos variables objetivos:

Y_1: Ahorro mensual

Y_2:Número total de trabajadores independientes.

La variable auxiliar es Ingresos Mensuales: X(k). La estratificación de la población la realizaremos en función a dos variables:

- Sexo: tiene dos categorías (1: hombre, 2: mujer)
- Nivel de instrucción: tiene 3 categorías(1: Primaria, 2: Secundaria y 3: Universitario).

Tabla N° 2.12: Distribución de ingresos en la muestra según sexo y nivel de instrucción

Sexo	**Nivel de instrucción**			**Total**
	Primaria	**Secundaria**	**Universitario**	
Hombre	75,000	25,000	275,000	600,000
Mujer	120,000	220,000	250,000	590,000
Total	195,000	470,000	525,000	1,190,000

Tabla N° 2.13: Distribución de ingresos en la población según sexo y nivel de instrucción

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	440,000	460,000	420,000	1,320,000
Mujer	380,000	420,000	400,000	1,200,000
Total	820,000	880,000	820,000	2,520,000

Tabla N° 2.14: Distribución de ahorro mensual en la muestra según sexo y nivel de instrucción

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	20,000	175,000	200,000	395,000
Mujer	25,000	175,000	70,000	270,000
Total	45,000	350,000	270,000	665,000

Tabla N° 2.15: Distribución del número de trabajadores independientes en la muestra según sexo y nivel de instrucción

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0	1	0	1
Mujer	3	2	0	5
Total	3	3	0	6

Los elevadores que se obtienen dividiendo el total poblacional (tabla N° 2.13) entre el total muestral por celdas para la variable auxiliar (tabla N° 2.12) son:

Tabla N° 2.16: Tabla de Elevadores

$W_{hh'}$	Nivel de instrucción		
Sexo	Primaria	Secundaria	Universitario
Hombre	5.86	1.84	1.52
Mujer	3.16	1.90	1.60

Por ejemplo, en la celda 1: $\frac{440000}{75000} = 5.867$.

Tabla N° 2.17: Distribución de la variable ahorros mensuales ajustados

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	117,340	322,000	305,400	744,740
Mujer	79,175	334,075	112,000	525,250
Total	196,515	656,075	417,400	1,269,990

Luego,

Tabla N° 2.18: Distribución de la variable número de trabajadores independientes ajustada

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0	2	0	2
Mujer	9	4	0	13
Total	9	6	0	15

Las estimaciones del total, para la variable cuantitativa Y_1 como para la variable cualitativa Y_2 son:

$\hat{Y}_1$:1,269,990 es el ahorro mensual total estimado en la población.

$\hat{Y}_2$:15 es el número estimado de trabajadores independientes en la población.

El número de trabajadores independientes en la población es estimado en 15 individuos, además estimamos que el ahorro promedio mensual de los trabajadores en la población es de 63,499 soles al mes.

CAPITULO III

INTRODUCCIÓN A LOS PROCEDIMIENTOS ITERATIVOS CON INFORMACIÓN AUXILIAR CUALITATIVA

Los procedimientos iterativos que vamos a presentar en este capítulo son el RAKING USUAL y una variación de éste utilizada en el procedimiento REDRE. Ambos se utilizan cuando tenemos variables auxiliares cualitativas y tenemos máxima información en las distribuciones univariantes de estas variables auxiliares. La estratificación se realiza con el cruce multivariante de estas variables auxiliares.

3.1 Introducción a los estimadores Raking (Raking ratio estimator)

El estimador de la razón por el procedimiento iterado ha sido estudiado por J.N.K. Rao(1974) y por Brackstone y Rao(1979). Para su aplicación se parte de dos caracteres básicos para los que se dispone de datos en un censo. Por ejemplo, grupos de edad y tipos de vivienda o, sexo y estado civil. Es un método que se resuelve de forma iterativa.

Se designa por N_{ij} la frecuencia absoluta o total censal de observaciones correspondientes a la casilla(i,j). Los totales marginales son, para la fila i-esima:

$$N_{i\cdot} = \sum_{i} N_{ij}$$

y para la columna j-esima:

$$N_{\cdot j} = \sum_{j} N_{ij}$$

Verificandose:

$$\sum_{i}\sum_{j} N_{ij} = N$$

Se trata ahora de estimar las frecuencias absolutas de otros caracteres. Por ejemplo, nivel de educación, migración o desempleo, para los cuales solo se dispone de una muestra de tamaño n. El número de observaciones muestrales en los caracteres básicos que corresponden a la casilla (i,j), antes mencionada, se designa por n_{ij}. Sea x_{ij} la frecuencia absoluta muestral. La frecuencia absoluta estimada del nuevo carácter se representa por $\hat{X}_{ij}$ y para la iteración t-esima, el estimador será:

$$X_{ij}^{(t)} = W_{ij}^{(t)} \cdot x_{ij}$$

En donde W_{ij} es una ponderación que se obtiene como a continuación se indica. La sucesión de iteraciones puede iniciarse por filas o por columnas. Si se empieza por filas se tiene, en función de la iteración anterior:

$$W_{ij}^{(t)} = \frac{W_{ij}^{(t-1)} \cdot N_{i\cdot}}{\sum_{j} n_{ij} W_{ij}^{(t-1)}} \quad , \qquad \text{si t es impar;}$$

$$W_{ij}^{(t)} = \frac{W_{ij}^{(t-1)} \cdot N_{\cdot j}}{\sum_{i} n_{ij} W_{ij}^{(t-1)}} , \qquad \text{si t es par.}$$

La constante inicial depende del diseño. En el muestreo aleatorio simple es $W^{(1)} = \frac{N}{n}$. Por consiguiente, para la primera y segunda iteración se verifica:

$$W_{ij}^{(1)} = \frac{N_{i\cdot}}{n_{i\cdot}} \quad , \quad W_{ij}^{(2)} = \frac{\frac{N_{i\cdot}}{n_{i\cdot}} \cdot N_{\cdot j}}{\sum_i n_{ij} \left(\frac{N_{i\cdot}}{n_{i\cdot}} \right)}$$

Por ultimo, la iteración t-esima antes mencionada (por ejemplo del total de desempleados o de inmigrantes), proporciona el total estimado:

$$\hat{X}^{(t)} = \sum_i \sum_j \hat{X}_{ij}^{(t)} = \sum_i \sum_j \hat{W}_{ij}^{(t)} X_{ij}$$

Debe señalarse que la idea de estimar la frecuencia de una casilla de una tabla de contingencia a partir de los totales marginales se debe a W.E. Deming y F.F. Stephan, que llamaron al procedimiento ajuste proporcional iterativo(iterative proporcional procedure).

En muchas aplicaciones prácticas se ha observado que los resultados de las iteraciones tienden a estabilizarse rápidamente, por ejemplo, a partir de la segunda iteración. Se ha observado asimismo que las diferencias al empezar por filas o por columnas, son pequeñas.

Expresiones de varianza y sesgos para el estimador por iteración, con diferentes diseños muestrales, pueden verse en el mencionada trabajo de Brackstone Y Rao.

Nos encontramos con dos tipos de métodos Raking:

1. La información auxiliar es cualitativa y los datos de que disponemos son los efectivos marginales. Los elevadores de celdas resultantes son los cocientes entre los efectivos marginales poblacionales y los marginales aproximados obtenidos tras las iteraciones. Este es el Raking usual. Como modelo de regresión es un caso particular del estimador usual de la Razón, en el que se toman las variables auxiliares de forma univariante.

2. La información auxiliar es cuantitativa. Los datos en este caso, son totales marginales de las variables cuantitativas en la estratificación. Como modelo de regresión es un caso particular del estimador de la Razón.

· MIXTOS: Existen variaciones de los dos métodos anteriores, métodos que son una combinación de los anteriores. Entre estos, mencionaremos un método denominado estimador modificado del Raking Ratio.

3.2 Ajuste con el método iterativo

Se plantean unas restricciones con el fin de que la distribución univariante ponderada de la muestra iguale a la población. La representación matricial de estas restricciones sigue siendo la misma:

$$\hat{X} \bullet w = X \bullet \mathrm{I}$$

La representación de las matrices auxiliares X y $\hat{X}$ la damos a continuación:

Tenemos L variables auxiliares cualitativas, cada una con $L_1, L_2, ..., L_L$ modalidades. Se realiza la estratificación de la población mediante el cruce multivariante de estas L variables, obteniendo $L_1 \bullet .... \bullet L_l \bullet ... L_L$ estratos. El número total de modalidades es $L_1 + L_2 + ... + L_l + ... + L_L = m$

En este caso, tomamos las variables identificadoras o dicotómicas de las modalidades:

$$X_{h_l}(k) = \begin{cases} 1, \text{si k tiene la modalidad} \\ 0, \text{en caso contrario} \end{cases} \qquad 1 \le h_l \le m$$

con k =1,..,n en la muestra y k =1,..,N en la población.

$\Rightarrow L_1 + ... + L_l + ... + L_L = m$ variables dicotómicas.

Tomamos los vectores fila poblacionales:

$X_{h_l} = (X_{h_l}(1), ..., X_{h_l}(k), ..., X_{h_l}(N))$ de dimensión (1xN), con $1 \le h_l \le m$

y los muéstrales:

$\hat{X}_{h_l} = (\hat{X}_{h_l}(1),...,\hat{X}_{h_l}(k),...,\hat{X}_{h_l}(n))$ de dimensión (1xn), con $1 \le h_l \le m$

Para asegurar que:

$$w_1 + ... + w_k + ... + w_n = N$$

Tomamos la variable X_O idénticamente 1 y se representa mediante:

$Xo = (Xo(1),...,Xo(k),...,Xo(N)) = (1,...,1,...,1)$ vector fila poblacional de dimensión (1xN)

$\hat{X}o = (Xo(1),...,Xo(k),...,Xo(n)) = (1,...,1,...,1)$ vector fila muestral de dimensión (1xn)

Las matrices X y $\hat{X}$ están formadas por estos vectores fila:

$$X = \begin{bmatrix} X_o \\ X_1 \\ .... \\ X_h \\ .... \\ X_m \end{bmatrix} \quad \text{y} \quad \hat{X} = \begin{bmatrix} \hat{X}_o \\ \hat{X}_1 \\ .... \\ \hat{X}_h \\ .... \\ \hat{X}_m \end{bmatrix}$$

Es decir, son dos matrices de dimensión (m+1)xN y (m+1)xn

Desarrollando la expresión matricial, obtenemos:

$$\hat{X} \bullet w = \mathrm{X} \bullet \mathrm{I}$$

$$\begin{bmatrix} \hat{X}_o \\ \hat{X}_1 \\ .... \\ \hat{X}_h \\ .... \\ \hat{X}_m \end{bmatrix} \bullet \begin{bmatrix} w_1 \\ w_2 \\ .... \\ .... \\ .... \\ w_n \end{bmatrix} = \begin{bmatrix} X_o \\ X_1 \\ .... \\ X_h \\ .... \\ X_m \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ ... \\ ... \\ ... \\ 1 \end{bmatrix}$$

Luego, el sistema de ecuaciones que se obtiene es:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet Xo(k) = \sum_{k=1}^{N} Xo(k) \quad \Leftrightarrow w_1 + ... + w_k + ... + w_n = N \\ \sum_{k=1}^{n} w_k \bullet X_h(k) = \sum_{k=1}^{N} X_h(k), \quad \forall \ 1 \le h \le m \end{cases}$$

Vamos a simplificar la notación, tomando únicamente dos variables auxiliares con un número de modalidades a y b. Numeramos las modalidades, de tal forma que si los datos que tenemos son los marginales $N_{h.}$ y $N_{.h'}$ tenemos que:

$$\begin{cases} \mathrm{N_h} = \mathrm{N_{h.}} \quad \text{con} \quad 1 \le \mathrm{h} \le \mathrm{a} \ \text{y} \\ \mathrm{N_{h'}} = \mathrm{N._{h'}} \ \text{con} \ \mathrm{a} + 1 \le \mathrm{h'} \le \mathrm{a} + \mathrm{b} \end{cases} \qquad \text{siendo a+b =m}$$

y de forma análoga, definimos la notación para los datos muéstrales de la distribución univariante:

$$\begin{cases} \mathrm{n_h} = n_{\mathrm{h.}} \ \text{con} \quad 1 \le \mathrm{h} \le \mathrm{a} \ \text{y} \\ \mathrm{n_{h'}} = \mathrm{n._{h'}} \ \text{con} \ \mathrm{a} + 1 \le \mathrm{h'} \le \mathrm{a} + \mathrm{b} \end{cases}$$

Tomando la representación con variables auxiliares de estos datos, tenemos:

$$\begin{cases} \sum_{k=1}^{N} \mathrm{X_h}(k) = N_h \\ \sum_{k=1}^{n} \mathrm{X_h}(k) = n_h \end{cases} \text{ con } 1 \le h \le m$$

Resolvemos el sistema

El ajuste es por celdas, teniendo todos los individuos de una misma celda el mismo peso. Las incógnitas pasan a ser entonces, los elevadores de las celdas $W_{hh'}$ con $1 \le h \le a$, $a+1 \le h' \le a+b$. Y el sistema tiene la forma:

$$\begin{cases} \sum_{h=1}^{a} \sum_{h=a+1'}^{a+b} W_{hh'} = N \\ \sum_{h'=a+1}^{a+b} W_{hh'} \bullet n_{hh'} = \sum_{k=1}^{N} X_h(k) = N_h = N_{h.,} \quad \forall\, 1 \le h \le a \\ \sum_{h=1}^{a} W_{hh'} \bullet n_{hh'} = \sum_{k=1}^{N} X_{h'}(k) = N_{h'} = N_{.h'}, \forall\, a+1 \le h' \le a+b \end{cases}$$

En general se obtiene un sistema de $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L$ incógnitas y $L_1 + \ldots + L_l + \ldots + L_L + 1 = m + 1$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema de $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L$ incógnitas y $L_1 + \ldots + L_l + \ldots + L_L = m$ ecuaciones. En el caso de sólo dos variables auxiliares, tenemos un sistema de a*b incógnitas y a+b ecuaciones.

En general hay más incógnitas que ecuaciones, por lo que el sistema tendrá infinitos elevadores posibles. Para evitar esto se toma un criterio de minimización para una función, con lo que se obtendrá la unicidad de la solución. La función que se tomará será la función distancia del vector de pesos iniciales al final. Según qué función distancia se tome, se obtendrán distintas soluciones.

3.3 Descripción del procedimiento

Vamos a describir a continuación los procedimientos iterativos que se siguen, tanto con el método RAKING USUAL como con su variación en el REDRE.

Antes de describir la fórmula iterativa de cada uno de los dos procesos vamos a presentar la situación, respecto de la información que se dispone, que es común para los dos procesos:

La tabla bivariante con la información poblacional es:

A \ B	1	2		h'		b	Total
1	W_{11}	W_{12}		$W_{1h'}$		W_{1b}	$W_{1.} = W_1$
2	W_{21}	W_{22}		$W_{2h'}$		W_{2b}	$W_{2.} = W_2$
h	W_{h1}	W_{h2}		$W_{hh'}$		W_{hb}	$W_{h.} = W_h$
a	W_{a1}	W_{a2}		$W_{ah'}$		W_{ab}	$W_{a.} = W_a$
Total	$W_{.1} = W_{a+1}$	$W_{.2} = W_{a+2}$		$W_{.h'} = W_{a+h'}$		$W_{.b} = W_{a+b}$	1

De esta tabla sólo conocemos los datos de las distribuciones univariantes marginales y que corresponden a las celdas sombreadas.

Respecto a los datos muéstrales, la tabla que se obtiene toma la siguiente forma:

A \ B	1	2		h'		b	Total
1	q_{11}	q_{12}		$q_{1h'}$		q_{1b}	$q_{1.} = q_1$
2	q_{21}	q_{22}		$q_{2h'}$		q_{2b}	$q_{2.} = q_2$
h	q_{h1}	q_{h2}		$q_{hh'}$		q_{hb}	$q_{h.} = q_h$
a	q_{a1}	q_{a2}		$q_{ah'}$		q_{ab}	$q_{a.} = q_a$
Total	$q_{.1} = q_{a+1}$	$q_{.2} = q_{a+2}$		$q_{.h'} = q_{a+h'}$		$q_{.b} = q_{a+b}$	1

En este caso se conoce la distribución conjunta de únicamente dos variables auxiliares cualitativas y bajo esta suposición vamos a presentar ambos procesos iterativos. Para ello, vamos a ilustrar ambos procesos iterativos aplicándolos a unas tablas sencillas imaginarias, a continuación la simulación del procedimiento.

El tamaño de la *población es 20* y la *muestra tiene 10* individuos.

Los datos poblacionales que conocemos son los siguientes $(W_{hh'})$

c	1	2	3	TOTAL
1	W_{11}	W_{12}	W_{13}	$W_{1.} = 0.75$
2	W_{21}	W_{22}	W_{23}	$W_{2.} = 0.25$
TOTAL	$W_{.1} = 0.2$	$W_{.2} = 0.4$	$W_{.3} = 0.4$	1

La tabla de datos muéstrales es $(q_{hh'})$

A B	1	2	3	TOTAL
1	$q_{11} = 0.1$	$q_{12} = 0.2$	$q_{13} = 0.2$	$q_{1.} = 0.5$
2	$q_{21} = 0.2$	$q_{22} = 0.1$	$q_{23} = 0.2$	$q_{2.} = 0.5$
TOTAL	$q_{.1} = 0.3$	$q_{.2} = 0.3$	$q_{.3} = 0.4$	1

3.3.1 RAKING USUAL

El algoritmo iterativo que define el método RAKING es:

$$w(k,m) = w(k,m-1) * \frac{W_{h0}}{q_{h0,m-1}} * \frac{W_{h_0'}}{q'_{h_0',m-1}} \quad \text{con}$$

$$\begin{cases} q'_{h',0} = \sum_{h=1}^{a} q'_{hh',0} = \sum_{h=1}^{a} q_{hh',0} \bullet \frac{W_{h.}}{q_{h,0}} \\ q'_{h',m} = \sum_{h=1}^{a} q'_{hh',m} = \sum_{h=1}^{a} q_{hh',m} \bullet \frac{W_{h}}{q_{h,m}}, \quad con \quad m \geq 1 \end{cases} \qquad para\ a+1 \leq h' \leq a+b$$

y el elemento k en la celda $\left(h_0, h_0'\right)$.

Seguidamente vamos a aplicar la fórmula iterativa al ejemplo simulado. Para ello, presentamos tablas en las que presentaremos los pesos obtenidos, así como la distribución muestral ponderada obtenida al elevarla por estos pesos.

Comenzamos las iteraciones, recordando que cada iteración se compone de 2 pasos, al haber únicamente dos variables de ajuste:

1. Se ajusta la distribución respecto de la distribución poblacional marginal de las modalidades de la primera variable.

2. En la segunda fase, con esta nueva distribución muestral ponderada se ajustara la distribución a las modalidades de la segunda característica, finalizando así la iteración.

Se repetirá el procedimiento hasta que la tabla muestral quede ajustada a la tabla poblacional.

En primer lugar tomamos la tabla número de porcentajes muestrales ($q_{hh'}$):

A \ B	1	2	3	TOTAL
1	$q_{11}=0.1$	$q_{12}=0.2$	$q_{13}=0.2$	$q_{1.}=0.5$
2	$q_{21}=0.2$	$q_{22}=0.1$	$q_{23}=0.2$	$q_{2.}=0.5$
TOTAL	$q_{.1}=0.3$	$q_{.2}=0.3$	$q_{.3}=0.4$	1

El primer peso que se asigna a los individuos es el de Horvitz– Thompson, es decir, $\frac{N}{n}$. PESOS(0)=20/10(El peso inicial se toma en base al diseño muestral, en este caso, es un muestreo aleatorio simple).

2	2	2
2	2	2

muestral ponderada por los *pesos(0)* $(q_{hh'}, 0)$:

A \ B	1	2	3	Total	$W_{h.}$	Elevadores
1	0.2	0.4	0.4	1	0.75	0.75
2	0.4	0.2	0.4	1	0.25	0.25
Total	0.6	0.6	0.8	2		

El primer paso de la iteración es el ajuste de la muestra respecto a la primera característica auxiliar o de estratificación. Los factores de elevación(elevadores) que se obtienen debido al ajuste son:

0.75	0.75	0.75
0.25	0.25	0.25

Y obtenemos la siguiente tabla muestral en porcentajes ponderada $(q'_{hh'}, 0)$, multiplicando los elevadores obtenidos por la tabla de datos muestral:

A \ B	1	2	3	Total
1	0.15	0.3	0.3	0.75
2	0.1	0.05	0.1	0.25
Total	0.25	0.35	0.4	1
$W_{.h}$	0.2	0.4	0.4	
Elevadores	0.8	1.143	1	

El *segundo paso* de esta *primera iteración* consiste en ajustar la distribución muestral a la segunda variable de estratificación. Los factores de elevación de muestra en este segundo paso de la iteración son:

0.8	1.143	1
0.8	1.143	1

La *tabla muestral en porcentajes* obtenida al *elevar* la muestra por estos nuevos elevadores es $(q_{hh',1})$:

A \ B	1	2	3	Total	$W_{h.}$	Elevadores
1	0.12	0.343	0.3	0.763	0.75	0.983
2	0.08	0.057	0.1	0.237	0.25	1.055
Total	0.2	0.4	0.4	1		

La tabla muestral se ha *desajustado ahora respecto de la primera variable* de estratificación, por lo que procedemos a la *segunda iteración* del procedimiento:

El *primer paso de la segunda iteración*. Los *factores de elevación* que se obtienen debido al ajuste son:

0.983	0.983	0.983
1.055	1.055	1.055

Y obtenemos la siguiente *tabla muestral en porcentajes ponderada* $(q^{'}_{hh'},1)$:

A \ B	1	2	3	Total
1	0.118	0.337	0.295	0.75
2	0.084	0.06	0.106	0.25
Total	0.202	0.397	0.401	1
$W_{.h}$	0.2	0.4	0.4	
Elevadores	0.99	1.008	0.998	

El *segundo paso de la segunda iteración*. Los factores de son:

0.99	1.008	0.998
0.99	1.008	0.998

La *tabla muestral en porcentajes* que se obtiene al *elevar* la muestra por estos nuevos elevadores es $(q_{hh',2})$:

A \ B	1	2	3	Total
1	0.117	0.34	0.294	0.751
2	0.083	0.06	0.106	0.249
Total	0.2	0.4	0.4	1

La tabla obtenida es la de *porcentajes muéstrales ajustados a la población*. Para obtener la tabla de número de *efectivos poblacionales* estimados en cada celda de la estratificación, basta con multiplicar la tabla por el tamaño de la población (N =20). La tabla estimada de efectivos que obtenemos es:

A \ B	1	2	3	Total
1	2	7	6	15
2	2	1	2	5
Total	4	8	8	20

La fórmula iterativa surge del siguiente proceso:

- Tomamos la distribución univariante de la primera variable auxiliar. Se ponderan todas las celdas de esa fila h por el factor $\frac{W_{h.}}{qh.,0} = \frac{W_h}{qh,0}$, con el fin de ajustar la distribución marginal auxiliar ponderada de la muestra a la población. La distribución muestral obtenida tras este primer paso se denota por $q'_{hh'},0$.

- Se ajusta la distribución muestral obtenida con el fin de ajustar a la distribución marginal de la segunda variable auxiliar. Para ello se toma el factor de elevación $\frac{W_{.h'}}{q'_{.h'},0} = \frac{W_{h'}}{q'_{h'},0}$.

Con los pasos anteriores se finaliza una iteración, tras lo cual, la distribución muestral queda elevada por $\frac{W_h}{q_{h.,0}} \bullet \frac{W_{.h'}}{q'_{.h',0}} = \frac{W_h}{q_{h,0}} \bullet \frac{W_{h'}}{q'_{h',0}}$.

La distribución resultante ajusta exactamente respecto de la distribución univariante para la segunda variable auxiliar, pero debido al segundo paso se ha desajustado ligeramente respecto de la distribución univariante de la primera variable auxiliar.

Para una iteración "m" cualquiera obtenemos en el ajuste a la primera variable los factores $\frac{W_{.h.}}{q_{h.,m-1}} = \frac{W_h}{q_{h,m-1}}$ y en el segundo paso $\frac{W_{.h'}}{q'_{.h',m-1}} = \frac{W_{h'}}{q'_{h',m-1}}$, con lo que los pesos finales son:

$$w(k,m) = w(k,m-1) \bullet \frac{W_h}{q_{h',m-1}} \bullet \frac{W_{.h'}}{q'_{h',m-1}} \quad \text{con k en el estrato } (h,h')$$

El proceso se sigue iterando hasta un *número fijo de iteraciones*, o hasta que se da la *convergencia* a la distribución poblacional univariante. Este proceso iterativo resulta de tomar el sistema de restricciones y dividirlo en *dos sistemas*:

1. Un primer sistema en el que se plantean las restricciones correspondientes a la *primera variable auxiliar* y en el que se consideran que los pesos resultantes son de la forma:

$$w'(k,m-1) = w(k,m-1) * c_h$$

Resulta entonces un sistema de *a* incógnitas y *a* ecuaciones, con matriz asociada diagonal, de elementos diagonales el número de efectivos muéstrales obtenidos para cada modalidad de la primera variable auxiliar. Este sistema tiene *única solución* siempre que dichos elementos diagonales sean no nulos. El primer paso de cada iteración resulta entonces de la resolución de este sistema.

2. El segundo paso surge análogamente de tomar los pesos de la forma:

$$w(k,m) = w'(k,m-1) * c_{h'}$$

y se plantea entonces un sistema de *b* ecuaciones y *b* restricciones, en el que las incógnitas son $c_{h'}$.

Este sistema tiene como matriz asociada una matriz diagonal con elementos diagonales el número de efectivos muéstrales marginales obtenidos para las modalidades de esta segunda variable auxiliar. Bajo la hipótesis de no nulidad de estos elementos, el sistema tiene una *única solución*, con lo que se obtiene la distribución ponderada tras el segundo paso de la iteración y por lo tanto, al tener únicamente dos variables auxiliares, los pesos finales para la iteración.

En el caso de tener más variables auxiliares, tendríamos más sistemas planteados en cada iteración, tantos como variables. La solución que se obtiene mediante el procedimiento iterativo RAKING, coincide además con la solución que se obtiene con el método de *Ajuste de Mínimo Cuadrados*: se plantea un sistema con unas restricciones, que corresponden a

la consistencia con respecto de la información auxiliar poblacional de la distribución muestral ponderada, y una función objetivo en el que se mide la exactitud de dicha distribución muestral ponderada tomando como distancia una función de la forma $D(x,y)=\frac{(x-y)^2}{x}$.

Así, se toma la función distancia como la suma de las distancias de las distribuciones muéstrales univariantes con respecto a las poblacionales correspondientes a cada modalidad, obteniendo así la función distancia a minimizar.

Para resolver este problema de minimización de una función objetivo con variables sujetas a una serie de restricciones se emplea el método de los *multiplicadores de Lagrange* y la resolución del mismo mediante dichos multiplicadores nos lleva a la misma solución del Raking, lo que pasa que, mediante este último el proceso es más rápido.

Respecto al Raking, podemos además obtener el ajuste para una celda que sea para nosotros de especial interés de una forma relativamente rápida, sin más que mediante una compresión de las celdas restantes en celdas colindantes a la celda en cuestión.

Como vemos el Raking es un procedimiento iterativo que ajusta la distribución marginal de las modalidades por variables auxiliares. Resaltar que cuando alguna de las celdas en la estratificación no tiene ningún elemento Muestral $(n_{hh'}=0)$, entonces en el Raking se le asigna un número de efectivos poblacionales aproximado $\tilde{N}_{hh'}=0$ y respectivamente $\tilde{W}_{hh'}=0$.

El Raking converge bajo ciertas condiciones de regularidad. Además se obtiene que en estos casos de convergencia, los estimadores de los $N_{hh'}$, los $\tilde{N}_{hh'}$ son asintóticamente insesgados, de distribución normal y de mínima varianza, es decir, son estimadores BAN (best asymptocally normal estimators).

3.3.2 Redre

Es una variación del método RAKING USUAL. La fórmula iterativa que lo define es:

$$w(k,m) = w(k,m-1) * \frac{\left[\sum\sum X_{hh'}(k) * \left[\frac{\frac{\sum_{k=1}^{N} X_h(k)}{N}}{\frac{\sum_{k=1}^{n} X_h(k) * w(k,m-1)}{n}} + \frac{\frac{\sum_{k=1}^{N} X_{h'}(k)}{N}}{\frac{\sum_{k=1}^{n} X_h'(k) * w(k,m-1)}{n}} \right] \right]}{2}$$

Con: $w(k,m)$) peso asignado al elemento k en la iteración m-ésima: $w(k,0) = 1$.

$\frac{\sum_{k=1}^{N} X_h(k)}{N}$ designa el porcentaje poblacional de elementos con la modalidad h, siendo esta modalidad una de las de la primera variable auxiliar. Tomando la nomenclatura del número:

$$\begin{cases} \frac{N_h}{N} & con\ 1 \le h \le a \\ \frac{N_{h'}}{N} & con\ a+1 \le h' \le b \end{cases}$$

Para simplificar aún más la notación denotamos estos porcentajes poblacionales como son los porcentajes muéstrales ponderados mediante los pesos obtenidos tras la iteración "m". Son porcentajes marginales correspondientes a las modalidades de la primera y la segunda variable auxiliar, respectivamente. Para simplificar las expresiones tomamos la notación:

$$W_h = \frac{\left(\sum_{k=1}^{N} X_h(k)\right)}{N} = \frac{N_h}{N} \qquad \text{y} \qquad W_{h'} = \frac{\left(\sum_{k=1}^{N} X_{h'}(k)\right)}{N} = \frac{N_{h'}}{N}$$

$\frac{\left[\sum_{k=1}^{n} X_{h'}(k) * w(k,m-1)\right]}{n}$ y $\frac{\left[\sum_{k=1}^{n} X_h(k) * w(k,m-1)\right]}{n}$ son los porcentajes muéstrales ponderados mediante los pesos obtenidos tras la iteración "m". Son porcentajes marginales correspondientes a las modalidades de la primera y la segunda variable auxiliar, respectivamente. Para simplificar las expresiones tomamos la notación:

$$q_{h,m-1} = \frac{\left[\sum_{k=1}^{n} X_h(k) * w(k,m-1)\right]}{n} \quad \text{y} \quad q_{h',m-1} = \frac{\left[\sum_{k=1}^{n} X_{h'}(k) * w(k,m-1)\right]}{n}$$

En el caso de m =1 estos porcentajes muéstrales ponderados son:

$$\begin{cases} \left[\frac{n_h}{n}\right] & con\ 1 \le h \le a \quad y \\ \left[\frac{n_{h'}}{n}\right] & con\ a+1 \le h' \le a+b \end{cases}$$

Estas expresiones se deducen directamente de tomar $w(k,0)=1$.

- El valor 2, representa el número de variables auxiliares.

- $\sum_{h=1}^{a}\sum_{h'=1}^{b} X_{hh'}(k)$ es un filtro para hallar los factores que intervendrán en el peso obtenido en cada iteración, al ser las variables $X_{hh'}(k)$ $con\, 1 \le h \le a$ y $1 \le h' \le b$ dicotómicas asociadas a cada celda de la estratificación. Es decir, según la notación que hemos tomado en el problema, $X_{hh'}(k) = X_h(k) * X_{h'}(k)$.

Tras todo lo mencionado la fórmula iterativa para REDRE es:

$$w(k,m) = w(k,m-1) * \left[\sum_{h=1}^{a}\sum_{h'=1}^{b} X_{hh'}(k) * \left[\frac{\frac{W_h}{q_{h,m-1}} + \frac{W_{h'}}{q_{h',m-1}}}{2} \right] \right]$$

Generalizamos la fórmula iterativa anterior, al caso general de más variables auxiliares y obtenemos:

$$w(k,m) = w(k,m-1) * \left[\frac{suma\left[\frac{W_h}{q_{h,m-1}} \right]}{n^{\circ}\ \text{var}iables\ auxiliares} \right]$$

Siendo la suma en todas las modalidades en las que está el elemento k de la muestra, para todas las variables auxiliares.

Estos pesos se multiplican luego por $\frac{N}{n}$, con el fin de que $\sum_{k=1}^{n} w(k,m) = N$. Así, esto es equivalente a tomar como pesos iniciales $w(k,0) = \frac{N}{n}$, es decir, corresponde a tomar como pesos iniciales los inversos de la probabilidad de inclusión en la muestra para cada individuo, es decir, corresponde a tomar el estimador inicial de HORVITZ-THOMPSON en los muestreos probabilistas con igual probabilidad de inclusión $z_k = \frac{n}{N}$ para todo individuo muestral.

Es un ajuste iterativo por celdas, en el que a todos los individuos de una misma celda se les asigna el mismo peso.

Vemos cómo actúa el REDRE aplicado a nuestro ejemplo de simulación:

El tamaño de la población es 20 y la muestra tiene 10 individuos.

Los datos poblacionales que conocemos son los siguientes $(W_{hh'})$

A \ B	1	2	3	TOTAL
1	W_{11}	W_{12}	W_{13}	$W_{1.} = 0.75$
2	W_{21}	W_{22}	W_{23}	$W_{2.} = 0.25$
TOTAL	$W_{.1} = 0.2$	$W_{.2} = 0.4$	$W_{.3} = 0.4$	1

La tabla de datos muéstrales es $(q_{hh'})$

A \ B	1	2	3	TOTAL
1	$q_{11} = 0.1$	$q_{12} = 0.2$	$q_{13} = 0.2$	$q_{1.} = 0.5$
2	$q_{21} = 0.2$	$q_{22} = 0.1$	$q_{23} = 0.2$	$q_{2.} = 0.5$
TOTAL	$q_{.1} = 0.3$	$q_{.2} = 0.3$	$q_{.3} = 0.4$	1

Primera iteración

Elevadores, que se obtienen $w(k,0) = \dfrac{(\dfrac{W_{h.}}{q_{h.}} + \dfrac{W_{.h'}}{q_{.h'}})}{2}$, el ajuste se hace simultamente para la primera y segunda variable auxiliar.

=(0.75/0.5+0.2/0.3)/2=**1.083**	=(0.75/0.5+0.4/0.3)/2=**1.417**	=(0.75/0.5+0.4/0.4)/2=**1.25**
=(0.25/0.5+0.2/0.3)/2=**0.583**	=(0.25/0.5+0.4/0.3)/2=**0.917**	=(0.25/0.5+0.4/0.4)/2=**0.75**

Tabla muestral en porcentajes ajustada a la población, se obtiene de multiplicar la tabla de elevadores por la tabla de datos muestrales $(q_{hh'},1)$:

A \ B	1	2	3	Total
1	0.108	0.283	0.25	0.641
2	0.117	0.092	0.15	0.359
Total	0.225	0.375	0.41	1

Segunda iteración: *Para lo cual utilizamos:*

La tabla de datos poblaciones $(W_{hh'})$

A \ B	1	2	3	TOTAL
1	W_{11}	W_{12}	W_{13}	$W_{1.} = 0.75$
2	W_{21}	W_{22}	W_{23}	$W_{2.} = 0.25$
TOTAL	$W_{.1} = 0.2$	$W_{.2} = 0.4$	$W_{.3} = 0.4$	1

La tabla muestral de porcentajes ya ajustada $(q_{hh'},1)$

A \ B	1	2	3	Total
1	0.108	0.283	0.25	0.641
2	0.117	0.092	0.15	0.359
Total	0.225	0.375	0.41	1

Elevadores $(w(k,1))$

=(0.75/0.641+0.2/0.225)/2=**1.029**	=(0.75/0.641+0.4/0.375)/2=**1.118**	=(0.75/0.641+0.4/0.4)/2=**1.085**
=(0.25/0.359+0.2/0.225)/2=**0.793**	=(0.25/0.359+0,4/0.375)/2=**0.882**	=(0.25/0.359+0.4/0.4)/2=**0.848**

Tabla muestral en porcentajes ajustada a la población $(q_{hh'},2)$:

A \ B	1	2	3	Total
1	0.111	0.316	0.271	0.698
2	0.093	0.081	0.127	0.301
Total	0.204	0.397	0.398	0.999

Y así sucesivamente,

Tercera iteración

Elevadores

=(0.75/0.698+0.2/0.204)/2=**1.027**	=(0.75/0.698+0.4/0.397)/2=**1.041**	=(0.75/0.698+0.4/0.398)/2=**1.04**
=(0.25/0.301+0.2/0.204)/2=**0.905**	=(0.25/0.301+0.4/0.397)/2=**0.919**	=(0.25/0.301+0.4/0.398)/2=**0.918**

Tabla muestral en porcentajes ajustada a la población

A \ B	1	2	3	Total
1	0.114	0.329	0.282	0.725
2	0.084	0.074	0.117	0.275
Total	0.198	0.403	0.399	1

Cuarta iteración

Elevadores

=(0.75/0.725+0.2/0.198)/2=**1.022**	=(0.75/0.725+0.4/0.403)/2=**1.014**	=(0.75/0.725+0.4/0.399)/2=**1.018**
=(0.25/0.275+0.2/0.198)/2=**0.96**	=(0.25/0.275+0.4/0.403)/2=**0.951**	=(0.25/0.275+0.4/0.399)/2=**0.956**

Tabla muestral en porcentajes ajustada a la población

A \ B	1	2	3	Total
1	0.117	0.334	0.287	0.738
2	0.081	0.07	0.112	0.263
Total	0.198	0.404	0.399	1.001

Quinta iteración

Elevadores

=(0.75/0.738+0.2/0.198)/2=**1.013**	=(0.75/0.738+0.4/0.404)/2=**1.003**	=(0.75/0.738+0.4/0.399)/2=**1.009**
=(0.25/0.263+0.2/0.198)/2=**0.98**	=(0.25/0.263+0.4/0.404)/2=**0.97**	=(0.25/0.263+0.4/0.399)/2=**0.977**

Tabla muestral en porcentajes ajustada a la población

A \ B	1	2	3	Total
1	0.119	0.335	0.29	0.744
2	0.079	0.068	0.109	0.256
Total	0.198	0.403	0.399	1

Sexta iteración

Elevadores

=(0.75/0.744+0.2/0.198)/2 =**1.009**	=(0.75/0.744+0.4/0.403)/2 =**1**	=(0.75/0.744+0.4/0.399)/2 =**1.005**
=(0.25/0.256+0.2/0.198)/2 =**0.993**	=(0.25/0.256+04/0.403)/2 =**0.985**	=(0.25/0.256+0.4/0.399)/2 =**0.99**

Tabla muestral en porcentajes ajustada a la población

A \ B	1	2	3	Total
1	0.12	0.335	0.291	0.746
2	0.078	0.067	0.108	0.253
Total	0.198	0.402	0.399	0.999

Séptima iteración

Elevadores

=(0.75/0.746+0.2/0.198)/2 =**1.008**	=(0.75/0.746+0.4/0.402)/2 =**1**	=(0.75/0.746+0.4/0.399)/2 =**1.004**
=(0.25/0.253+0.2/0.198)/2 =**0.999**	=(0.25/0.253+0.4/0.402)/2 =**0.992**	=(0.25/0.253+0.4/0.398)/2 =**0.995**

Tabla muestral en porcentajes ajustada a la población

A \ B	1	2	3	Total
1	0,121	0.335	0.292	0.748
2	0.078	0.066	0.107	0.251
Total	0.199	0.401	0.399	0.999

La distribución ha quedado ya ajustada con una *tolerancia de 0.002,* tras siete iteraciones.

La tabla con el número de efectivos estimados tras el ajuste a la distribución univariante de las variables de estratificación es:

A \ B	1	2	3	Total
1	2	7	6	15
2	2	1	2	5
Total	4	8	8	20

La tabla estimada para nuestro ejemplo simulado obtenida con ambos métodos es la misma. Este método de ajuste, el REDRE lo que plantea es evitar tantas modificaciones de los pesos en cada iteración, por lo que toma el factor de modificación de los pesos como una media de los factores resultantes de los ajustes a las modalidades a las que pertenece cada elemento. Así, lo que hace es plantear un único sistema a solucionar, en el que obtenemos a+b incógnitas y ecuaciones.

Los pesos resultantes toman entonces la forma

$$w(k,m) = w(k,m-1) * \sum_{hh'} \left(X_h(k) * c_h\right) * \left(X_{h'}(k) * c_{h'}\right)$$

con $1 \le h \le a \quad y \quad a+1 \le h' \le a+b+1$

Los factores $c_h \quad y \quad c_{h'}$ se hallan de resolver el sistema de restricciones planteado y el peso final para cada elemento resulta de elevar el peso inicial por la media de los factores $c_h \quad y \quad c_{h'}$.

En el ejemplo 5, veremos los procedimientos antes descritos para el método de Raking Usual.

EJEMPLO 5:

Consideraremos dos variables objetivos, correspondientes a los dos tipos de variables objetivos que pueden dar: cualitativa y cuantitativa.

Y_1: Número total de elementos de la población independiente.

Y_2: Es el ingreso total mensual de la población.

Las variables auxiliares son dos:

- ***Sexo***: tiene dos categorías L_1 =2
- ***Nivel de instrucción:*** tiene tres categorías L_2 =3

Tamaño de la población N =20 y muestral n =12

Consideraremos un muestreo probabilistico, con igual probabilidad de inclusión para todos los elementos de la muestra 12/20.

Variable auxiliares:

Tabla N°3.19: Datos poblacionales $\left(N_{hh'}\right)$

Sexo $\left(\sum_{k=1}^{N} X_{hh'}(k)\right)$	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	?	?	?	9
Mujer	?	?	?	11
Total	7	9	4	20

Tabla N°3.20: Datos muestrales $\left(n_{hh'}\right)$

Sexo $\left(\sum_{k=1}^{n} X_{hh'}(k)\right)$	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	1	1	1	3
Mujer	3	3	3	9
Total	4	4	4	12

Variable objetivo Y_1.

Tabla N°3.21: Número de trabajadores independientes muestrales en cada estrato de la población $\left(Y_{1,hh'}\right)$

Sexo $\left(\sum_{k=1}^{n} Y_1(k)\right)$	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0	1	0	1
Mujer	3	2	0	5
Total	3	3	0	6

Variable objetivo Y_2

Tabla N° 3.22: Ingreso total mensual en cada estrato de la muestra $\left(Y_{2,hh'}\right)$

Sexo $\left(\sum_{k=1}^{n} Y_2(k)\right)$	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	75,000	250,000	275,000	600,000
Mujer	12,000	220,000	250,000	590,000
Total	195,000	470,000	525,000	1,190,000

Presentaremos estos datos para las variables auxiliares en forma de porcentajes:

Tabla N°3.23: Distribución Poblacional porcentual marginal($W_{hh'}$)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	?	?	?	0.45
Mujer	?	?	?	0.55
Total	0.35	0.45	0.2	1

Tabla N° 3.24: Distribución muestral porcentual ($q_{hh'}$)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.0833	0.0833	0.0833	0.25
Mujer	0.25	0.25	0.25	0.75
Total	0.333	0.333	0.333	1

Resolveremos entonces por Raking Usual y seguidamente lo haremos por el Redre.

Comenzamos las iteraciones, recordando que cada iteración se compone de 2 pasos, al haber únicamente dos variables de ajuste:

1 Se ajusta la distribución respecto de la distribución poblacional marginal de las modalidades de la primera variable.

2 En la segunda fase, con esta nueva distribución muestral ponderada se ajustara la distribución a las modalidades de la segunda característica, finalizando así la iteración.

En primer lugar tomamos la tabla número de porcentajes muestrales $q_{hh'}$ (Tabla N°3.24)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.0833	0.0833	0.0833	0.25
Mujer	0.25	0.25	0.25	0.75
Total	0.333	0.333	0.333	1

El primer peso que se le asigna a los individuos es el Horvitz – Thompson, es decir N/n=1.6667, la tabla que se obtiene al ponderar por estos pesos se denomina tabla expandida.

Tabla Nº 3.25: Tabla expandida por los Elevadores o Expansores (Horvitz-Thompson)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.1388	0.1388	0.1388	0.4164
Mujer	0.417	0.417	0.417	1.2501
Total	0.5555	0.5555	0.5555	1.6665

El primer paso de la primera iteración: ajuste de la distribución a la distribución univariante de la variable Nivel de instrucción.

PASO 1: Primera Iteración:

Sexo	Nivel de instrucción			Total $(q_{hh'}, 0)$
	Primaria	Secundaria	Universitario	
Hombre	0.1388	0.1388	0.1388	0.4164
Mujer	0.4167	0.4167	0.4167	1.2501
Total	0.5555	0.5555	0.5555	1.6665
$W_{hh'}$	0.35	0.45	0.2	
Elevadores	0.6301	0.8101	0.36	

Ponderamos la tabla muestral con los elevadores hallados y se procede al segundo paso: Al ajuste a la distribución univariante de la variable Sexo.

PASO 2: Primera Iteración:

Sexo	Nivel de instrucción			Total	$W_{hh'}$	Elevadores
	Primaria	Secundaria	Universitario			
Hombre	0.087	0.112	0.05	0.2499	0.45	1.8010
Mujer	0.262	0.337	0.15	0.7501	0.55	0.7332
Total	0.35	0.45	0.2	1		

Ponderamos la tabla muestral con los elevadores hallados y obtenemos la tabla final:

Tabla N° 3.26: Distribución de la población en porcentaje

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.1575	0.2025	0.0900	0.45
Mujer	0.1925	0.2475	0.11	0.55
Total	0.35	0.45	0.2	1

Tras la única iteración la distribución univariante de las variables Nivel de instrucción y Sexo ha quedado ajustada.

La tabla de contingencia muestral, ya ajustada que se obtiene es:

Tabla N° 3.27: Distribución muestral ajustada

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	1	1	1	3
Mujer	3	4	2	9
Total	4	5	3	12

Así mismo, la tabla de contingencia poblacional ajustada que obtenemos es:

Tabla N° 3.28: Distribución poblacional ajustada

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	3	4	2	9
Mujer	4	5	2	11
Total	7	9	4	20

Se obtiene la siguiente tabla de pesos para cada individuos muestral:

Tabla N° 3.29: Distribución de pesos muestrales para cada individuo

Sexo	Nivel de instrucción		
	Primaria	Secundaria	Universitario
Hombre	3	4	2
Mujer	1.3333333	1.6666667	0.6666667

Las tablas con los totales estimados sobre las celdas para cada variable objetivo son las siguientes:

Tabla N° 3.30: Distribución de los trabajadores independiente estimada

Sexo	Nivel de instrucción			Total $(\tilde{Y}_{1,hh'})$
	Primaria	Secundaria	Universitario	
Hombre	0	4	0	4
Mujer	4	3	0	7
Total	4	7	0	11

Tabla N°3.31 : Distribución de los ingresos totales estimada

Sexo	Nivel de instrucción			Total $(\tilde{Y}_{2,hh'})$
	Primaria	Secundaria	Universitario	
Hombre	225,000	1,000,000	550,000	1,775,000
Mujer	160,000	366,667	166,667	693,334
Total	385,000	1,366,667	716,667	2,468,334

Por lo tanto, obtenemos que:

- El total de trabajadores independientes ($\hat{Y}_1 = 11$) son 11 trabajadores y el número de trabajadores promedio en la población es de 1 trabajador.

- El ingreso total mensual de los trabajadores ($\hat{Y}_2 = 2{,}468{,}333$) es de 2,468,333 soles y el ingreso promedio mensual de los trabajadores es de 123,417.

A continuación en el ejemplo 6, se mostrara el procedimiento del método REDRE.

EJEMPLO 6:

Hemos utilizado el procedimiento interactivo Redre que es una modificación del procedimiento Raking usual para los mismos datos del ejemplo 5, en este caso se realizaran 10 interacciones para obtener un número de tolerancia de 0.001.

Tabla N° 3.32: Distribución de la población marginal en porcentaje ($W_{hh'}$)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	?	?	?	0.45
Mujer	?	?	?	0.55
Total	0.35	0.45	0.2	1

Tabla N° 3.33: Distribución de la muestra en porcentaje ($q_{hh'}$)

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.0833	0.0833	0.0833	0.25
Mujer	0.25	0.25	0.25	0.75
Total	0.333	0.333	0.333	1

PASO 1: Primera iteración (elevadores)

1.425	1.575	1.2
0.89166667	1.04166667	0.66666667

Tabla N° 3.34: Porcentajes de la muestra ajustada a la población

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.1187	0.1312	0.1000	0.35
Mujer	0.2229	0.2604	0.1667	0.65
Total	0.342	0.392	0.267	1.00

PASO 2 : Segunda iteración(elevadores)

1.15538073	1.21765949	1.01817065
0.93534326	0.99762202	0.79813318

Tabla N°3.35: Porcentajes de la muestra ajustada a la población

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.1371	0.1598	0.1018	0.40
Mujer	0.2085	0.2598	0.1330	0.60
Total	0.346	0.420	0.235	1.00

Y así sucesivamente, seguimos realizando el ajuste a la tabla muestral y en la décima iteración, obtenemos la tabla de los elevadores seguida por la tabla muestral, las tablas son las siguientes:

PASO 10 :Décima iteración (elevadores)

1.00063497	1.00095947	0.99964128
0.99957821	0.99990271	0.99858451

Tabla N°3.36: Porcentajes de la muestra ajustada a la poblacion

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0.1573	0.1943	0.0981	0.45
Mujer	0.1927	0.2556	0.1021	0.55
Total	0.350	0.450	0.200	1.00

La tabla muestral ajustada como resultado de la última iteración es:

Tabla N° 3.37: Tabla muestral ajustada

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	2	3	1	6
Mujer	2	3	1	7
Total	4	6	2	12

Redondeado al entero siguiente.

La tabla poblacional ajustada = tabla N° 3.37 *(20/12).

Tabla N° 3.38: Tabla poblacional ajustada

Sexo	Nivel de instrucción		
	Primaria	Secundaria	Universitario
Hombre	4	5	2
Mujer	1	1.333	0.667

La tabla de pesos para cada elemento de la muestra según a que estrato pertenece es:

Tabla N° 3.39: Tabla de pesos por cada elemento según estrato

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	4	5	2	11
Mujer	3	4	2	9
Total	7	9	4	20

Con esto hemos obtenido y la distribución muestral ponderada ajustada a la distribución poblacional marginal, que no es exacta, dado que hemos empleado una tolerancia de 0.001 y con un número máximo de iteraciones de 10.

Tabla N° 3.40: Tabla ajustada de trabajadores independientes

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	0	5	0	5
Mujer	3	3	0	6
Total	3	8	0	11

Tabla N°3.41: Tabla ajustada del total de ingresos

Sexo	Nivel de instrucción			Total
	Primaria	Secundaria	Universitario	
Hombre	235,972	1,137,999	539,808	1,913,779
Mujer	154,134	374,814	170,083	699,031
Total	390,105	1,512,814	709,892	2,612,810

Las estimaciones resultantes para las dos variables objetivos que tenemos son:

- El total de trabajadores independientes ($\hat{Y}_1 = 11$) es de 11 y el promedio de trabajadores independientes es de 1 trabajador.

- El total de ingresos mensuales de los trabajadores ($\hat{Y}_2 = 2{,}612{,}810$) es de 2,612,810 soles y el ingreso promedio de los trabajadores es de 130,641 soles.

CAPITULO IV

APLICACIÓN DEL AJUSTE DE MUESTRAS CON INFORMACIÓN AUXILIAR

Para aplicar el método de ajuste con Máxima Información Auxiliar, utilizaremos los datos de una encuesta Socio-Demográfica y de Salud realizada en los centros poblados de la Cuenca del río Ayash de la provincia de Huari del departamento de Ancash, en el año 2004.

Como no existía información previa sobre los pobladores de la Cuenca del río Ayash, antes de la encuesta se procedió a realizar un censo para contar con información preliminar y posteriormente sacar una muestra probabilística para aplicar la encuesta de Salud epidemiológica.

4.1 Objetivos:

El Censo Local tuvo como objetivo conocer la realidad Socio-Demográfica y de Salud de los poblados de la Cuenca del río Ayash mediante la recopilación de información sobre variables personales, clínicas, ambientales y demográficas; para así establecer un perfil epidemiológico de la población.

La encuesta tuvo como objetivo la realización de pruebas clínicas de rigor para la estimación de prevalencias de enfermedades asociadas a la contaminación ambiental.

4.2 Diseño Muestral de la Encuesta

4.2.1 Población Objetivo

Está constituido por los pobladores residentes en los 10 centros poblados de estudio de la provincia de Huari del departamento de Ancash, cuya distribución poblacional se presenta en el cuadro 4.1.

Cuadro 4.1: Distribución poblacional de los Centros Poblados de estudio

Centro Poblado	Población
Atash	150
Ayash Huaripampa	535
Ayash Pichiu	333
Cambio 90	84
Centro Pichiu	152
Huamanín	109
Huancayoc	503
Huishllac	79
San Cristobal de Tambo	205
Vistoso	422
Total	2572

Fuente: Resultados del Censo de población y Salud – Marzo 2004

4.2.2 Población Muestreada

Está constituido por los pobladores residentes de los 10 centros poblados de estudio con 5 o más años cumplidos a la fecha de la Encuesta. Se ha considerado para la Encuesta aquellos pobladores que residen regularmente en alguno de los centros poblados de estudio. La distribución de familias y poblacional en los 10 centros poblados, después de las exclusiones mencionadas, se presenta en el cuadro 4.2.

Cuadro 4.2: Distribución familiar y poblacional de los Centros Poblados

Centro Poblado	Familias	Mujeres	Hombres	Población total
Atash	26	66	54	120
Ayash Huaripampa	118	219	237	456
Ayash Pichiu	74	133	148	281
Cambio 90	15	31	40	71
Centro Pichiu	36	59	59	118
Huamanín	21	45	42	87
Huancayoc	108	213	211	424
Huishllac	16	38	29	67
San Cristobal de Tambo	44	87	87	174
Vistoso	80	176	190	366
Total	538	1,067	1,097	2,164

Fuente: Resultados del Censo de población y Salud – Marzo 2004

4.2.3 Unidades

- **Unidad de Muestreo.** En la Encuesta Toxicológica, se define como unidad de muestreo al poblador residente de un centro poblado de estudio, con 5 o más años cumplidos.

- **Unidad de Análisis.** Para el análisis se definió dos unidades, a nivel individuo y a nivel familiar. Por tanto, la Encuesta determinó un cuestionario para recoger información a nivel individuo y otro a nivel familiar.

4.2.4 Marco Muestral

El Marco Muestral (MM) que se utilizó en la Encuesta Toxicológica es un marco de lista de los pobladores de los 10 centros poblados de estudio. Adicionalmente, se utilizó un material cartográfico consistente en mapas de los centros poblados de estudio. La información que contiene el MM y los mapas cartográficos, provienen del Censo de Población y Salud realizado en Marzo del 2004. La calidad de información contenida en el MM, permitió identificar, ubicar y estratificar adecuadamente las unidades muestrales.

El MM de lista permitió la selección de la muestra de pobladores residentes que formaron parte de la Encuesta, y con la ayuda de los mapas cartográficos se optimizó la ubicación de las viviendas de los pobladores seleccionados para la encuesta y facilitó la organización del trabajo de campo.

La preparación del Marco Muestral de lista de individuos fue realizado inmediatamente después del Censo de Población y Salud. Toda la información contenida en el Marco Muestral obtenida en el Censo, se vuelca en los mapas cartográficos que permitió identificar y ubicar las viviendas de los pobladores seleccionados para la Encuesta. El ámbito geográfico de la Encuesta Toxicológica abarca los 10 centros poblados de estudio ubicados en las cuencas del Ayash y del Huancayoc, de la provincia de Huari del departamento de Ancash.

4.2.5 Temas Investigados

El cuestionario consta de 5 partes o secciones:

- Sección I: Ubicación geográfica y datos generales de la vivienda y familia.
- Sección II: Información sobre morbilidad y uso de servicios de salud.
- Sección III: Información de salud reproductiva.
- Sección IV: Información de exposición a riesgos ambientales.
- Sección V: Información de mortalidad.

Las variables de la encuesta son mostradas en el Anexo A.5.

4.3 Ajuste con variables Auxiliares Cualitativas para la Estimación.

El parámetro objetivo que se desea estimar es el total de enfermos(Y)para lo cual utilizaremos dos variables auxiliares cualitativas, la primera de ellas es el Grupo de Edad (con 4 categorías) y el segundo es Estado de Exposición (con 2 categorías). Se realizó la estimación de la variable objetivo con el método de ajuste con información auxiliar para lo cual se utilizo un programa creado para tal fin en Matlab 7.0, el programa se puede ver en el Anexo A.1, y posteriormente con el método convencional para lo cual se utilizó el programa STATA 8.0, luego se realizó una comparación entre los dos ajustes obtenidos por ambos métodos, para lo cual se determinó el coeficiente de variación de la estimación del Total de enfermos.

4.3.1 Estimación con Ajuste de Información Auxiliar

El parámetro objetivo es:

Y : Número total de Enfermos.

Las Variables Auxiliares son dos:

A: Grupo de Edad, L_1 =4 categorías

1: De 5 a 9 años
2: De 10 a 17 años.
3: Mujer Adulta
4: Hombre Adulto.

B: Estado de Exposición, L_2 =2 categorías

1: Expuesto.
2: No Expuesto.

Tamaño de la población N =2,098 y el tamaño muestral n =308(de los 309 registro se ha eliminado un registro en la muestra, debido a que no formaba parte de la población muestreada).

Obtenemos $L_1 x L_2$ =8 celdas (por el cruce multivariante de las variables): 8 variables, para lograr el objetivo de la investigación se definirá 8 variables dicotómicas:

$$X_{h1h2}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (\text{h}_1, \text{h}_2) \\ 0, \text{en caso contrario} \end{cases}$$

Variables Auxiliares:

Tabla N°4.42: Distribución Poblacional según estado de Exposición y Grupo de Edad ($N_{hh'}$)

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	158	195	220	207	780
Si	257	308	374	379	1,318
Total	415	503	594	586	2,098

Variables Auxiliares:

Tabla N°4.43: Distribución Muestral según estado de Exposición y Grupo de Edad$\left(n_{hh'}\right)$

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	14	26	28	26	94
Si	40	56	58	60	214
Total	54	82	86	86	308

Tabla N°4.44: Distribución de la población de Enfermos en la muestra según estado de Exposición y Grupo de Edad(Y_1)

Estado de Exposición	Grupo de Edad				TOTAL
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	11	11	21	14	57
Si	31	29	52	38	150
TOTAL	42	40	73	52	207

Tabla N°4.45: Elevadores según estado de Exposición y Grupo de edad ($W_{h_1h_2} = \frac{N_{h_1h_h}}{n_{h_1h_2}}$)

Estado de Exposición	Grupo de Edad			
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos
NO	11.2857	7.5	7.8571	7.9615
SI	6.425	5.5	6.4483	6.3167

Tabla N°4.46: Distribución del total de enfermos ajustada con información auxiliar: Tabla N° 4.44 * Elevadores

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	124	83	165	111	483
Si	199	160	335	240	934
Total	323	242	500	351	1,417

El total de la población Enferma en la Cuenca del rió Ayash en la Provincia de Huari Departamento de Ancash es de 1,417 pobladores.

Cálculo de la Varianza

Primero vamos a calcular la varianza para la proporción de enfermos en la cuenca del rió Ayash, para luego encontrar la varianza del total de pobladores enfermos.

Varianza para la Proporción:

$$\hat{p}_h = \frac{enfermos\ \ en\ \ el\ \ estrato\ \ h}{total\ de\ pobladores\ \ en\ \ el\ \ estrato\ \ h} = \frac{n_{h_0}}{n_h}$$

$$\hat{p}_h = \frac{y_h}{n_h}$$

Como la varianza muestral es:

$$s^2 = \sum \frac{(y_i - \bar{y})^2}{n-1} = \frac{\sum y_i - n\bar{y}^2}{n-1}$$

Pero, $y_i = \begin{cases} 0\ individuo\ \ no\ enfermo \\ 1\ individuo\ \ enfermo \end{cases}$

Entonces.

$$s^2 = \frac{n\bar{y} - n\bar{y}^2}{n-1} = \frac{n\bar{y}(1-\bar{y})}{n-1} = \frac{np(1-p)}{n-1}$$

En cada estrato de la muestra, se tiene:

$$s_h^2 = \frac{np_h(1-p_h)}{n_h - 1}$$

Entonces la varianza estimada para la proporción de enfermos es:

$$\hat{E}\left[\mathrm{var}\left(\hat{Y}_W\right)\right] = \hat{V}ar(\hat{P}) = \frac{1-f}{n}\sum_{h=1}^{m} W_h \bullet s_h^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m}(1-W_h) \bullet s_h^2$$

Donde:

$f = \frac{n}{N}$, $w_h = \frac{N_h}{n_h}$ y $s_h^2 = \frac{n\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$ m= # de estratos.

Luego para estimar la varianza del total de enfermos, se utiliza:

$$\hat{V}ar(\hat{Y}) = N^2 * \hat{V}ar(\hat{P})$$

Al calcular la varianza por medio del programa[1] creado obtenemos: $\hat{V}ar(\hat{Y})$=2528.46729, y el error estándar es 50.2838671, bajo el ajuste con información auxiliar.

4.3.2 Estimación Convencional:

La estimación convencional del total de enfermos utilizando el programa STATA es:

		Error	Intervalo de		
Total	**Estimación**	**Estándar**	**Confianza al 95%**		**Deff**
Enfermos	1423.25	50.85	1323.178	1523.32	0.9689

[1]El programa esta realizado en Matlab y se encuentra en el anexo A.1.

El número total de pobladores Enfermos en la Cuenca del rió Ayash en la Provincia de Huari Departamento de Ancash bajo la estimación convencional es de 1423 pobladores enfermos y el error estándar es de 50.85.

4.3.3 Comparación de Estimaciones:

Para poder realizar las comparaciones entre el método de estimación con ajuste de información auxiliar y el método de estimación convencional necesitamos hacerlo en valores relativos, para lo cual calcularemos el coeficiente de variación(error relativo de muestreo) para cada método de estimación.

El coeficiente de variación de la estimación se calcula mediante:

$$\hat{C}v = \frac{\sqrt{\text{var}(\hat{Y})}}{\hat{Y}}$$

$\hat{C}v$ (Estimación convencional)= 0.0357 = 3.57%

$\hat{C}v$ (Estimación con ajuste) = 0.0355 = 3.55%

Del resultado anterior, se puede observar que la estimación con ajuste muestral genera un error relativo de muestreo, ligeramente menor que el método de estimación convencional.

Entonces podemos afirmar que el método de estimación con ajuste de información auxiliar proporciona una mejor estimación del total de enfermos en la cuenca del rió Ayash, esto se debe a que las variables auxiliares elegidas para realizar la post-estratificación están muy correlacionadas con la variable objetivo (población de enfermos) como se puede verificar en el Anexo A.3.

4.4 Ajuste con Variables Auxiliares Cuantitativas para la Estimación

El parámetro objetivo que se desea estimar es la concentración promedio de arsénico en la sangre ($\bar{Y}$) de los pobladores de la cuenca del rió Ayash, para lo cual se utilizó dos variables auxiliares para la post-estratificación de los datos, la primera de ellas es el Grupo de Edad (con 4 categorías) y el segundo es Estado de Exposición(con 2 categorías), y se utiliz como variable auxiliar para la estimación, la Edad. Se realizó la estimación del parámetro objetivo con el método de ajuste con información auxiliar para lo cual se utilizó un programa creado para tal fin en Matlab 7.0, el programa se puede ver en el Anexo A.2, y posteriormente con el método convencional para lo cual se utilizó el programa STATA 8.0, luego se realizó una comparación en cuanto a los dos ajustes obtenidos por ambos métodos para lo cual se calculó el coeficiente de variación y así determinar cual de los métodos de estimación produce una mejor estimación del promedio de arsénico en la sangre de los pobladores de la cuenca del rio Ayash, provincia de Huari departamento de Ancash.

4.4.1 Estimación con Ajuste de Información Auxiliar Cuantitativa.

El parámetro objetivo es:

$\bar{Y}$: Concentración promedio de Arsénico en la sangre.

La Variable Auxiliar cuantitativa es: Edad del poblador. Por tanto, en este caso se utilizó el estimador basado en la razón.

La post-estratificación de los datos se realizó en función de dos variables:

- A: **Grupo de Edad, L_1 =4 categorías**
  - 1: De 5 a 9 años
  - 2: De 10 a 17 años.
  - 3: Mujer Adulta
  - 4: Hombre Adulto.
- B: **Estado de Exposición, L_2 =2 categorías**
  - 1: Expuesto
  - 2: No Expuesto

Tamaño de la población N =2098 y muestral n =296(De los 308 registros se han eliminado 12 registros que no tenían datos para la variable nivel de arsénico en la sangre).

Obtenemos L_1 x L_2 =8 celdas o 8 variables, para lograr el objetivo de la investigación se definirá 8 variables dicotómicas:

$$X_{h1h2}(k) = \begin{cases} 1, \text{si k} \in \text{a la celda o estrato } (h_1, h_2) \\ 0, \text{en caso contrario} \end{cases}$$

Tabla N° 4.47: Distribución de la Edad total según estado de Exposición y Grupo de edad en la población

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	1,120	2,593	8,692	7,864	20,269
Si	1,779	4,033	13,852	13,702	33,366
Total	2,899	6,626	22,544	21,566	53,635

Tabla N° 4.48 : Distribución de la Edad total según estado de Exposición y Grupo de edad en la muestra

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	93	315	1,182	1,170	2,760
Si	258	675	2,352	2,082	5,367
Total	351	990	3,534	3,252	8,127

Tabla N° 4.49: Distribución del nivel total de arsénico según estado de Exposición y Grupo de edad en la muestra

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	30.588	48.25	62.686	53.41	194.934
Si	96.427	119.749	160.934	126.876	503.986
Total	127.015	167.999	223.62	180.286	698.92

Tabla N° 4.50: Elevadores según estado de exposición y Grupo de edad

Estado de Exposición	Grupo de Edad			
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos
No	12.0430	8.2317	7.3536	6.7214
Si	6.8953	5.9748	5.8895	6.5812

Tabla N°4.51: Niveles totales de Arsénico ajustada según estado de Exposición y Grupo de edad

Estado de Exposición	Grupo de Edad				Total
	De 5-9 años	De 10-17 años	Mujeres adultas	Hombres adultos	
No	368.372	397.182	460.970	358.988	1585.512
Si	664.898	715.478	947.814	834.993	3163.182
Total	1033.269	1112.660	1408.784	1193.981	4748.694

La concentración total de Arsénico en la sangre de los pobladores de la cuenca del rió Ayash del Departamento de Ancash es de 4748.694.Por lo que la concentración promedio de Arsénico en la sangre de los pobladores de la cuenca del rió Ayash es 2.2634 ug/dl.

Cálculo de la Varianza

Respecto a la varianza de este estimador, tenemos que en el caso de muestreo aleatorio estratificado con tamaños muéstrales suficientemente grandes en todos los estratos, la varianza de este estimador es:

$$\text{var}\left(\hat{\bar{Y}}_{Rs}\right) = \sum_{h=1}^{m} W_h \frac{(1-f_h)}{n_h} \bullet \left(Sy_h^2 + R_h^2 Sx_h^2 - 2R_h \rho_h Sy_h Sx_h\right) = \sum_{h=1}^{m} W_h \; Var(\hat{\bar{Y}}_{Rh})$$

Dado que se utiliza el estimador de la razón en cada celda de la estratificación, se toma la varianza de dicho estimador sobre cada una de estas celdas, bajo un muestreo aleatorio simple, de la forma:

$$\text{var}(\hat{\bar{Y}}_{Rh}) = \frac{1-f_h}{n_h} \bullet \left[\frac{\sum_{k=1}^{N_k} \left(Y_{hk} - R_h x_{hk}\right)^2}{N_h - 1} \right]$$

con R_h la razón entre medias muestrales de Y y X en el estrato h, es decir, la pendiente correspondiente a la recta de regresión entre la variable Y y X para el estrato h:

$$R_h = \frac{\sum_{k=1}^{N_h} Y_{hk}}{\sum_{k=1}^{N_h} X_{hk}} = \frac{\bar{Y}_h}{\bar{X}_h}$$

El estimador de la varianza de $\hat{\bar{Y}}_{RS}$ es:

$$\hat{V}ar(\hat{\bar{Y}}_{RS}) = \sum_{h=1}^{m} W_h^2 \left(\frac{1-f_h}{n_h}\right) \frac{\sum_{i=1}^{n_h} (y_{hi} - \hat{R}_h x_h)^2}{n_h - 1}$$

La estimación de la varianza de la concentración promedio de arsénico en la sangre es $\hat{V}ar(\hat{Y}_{RS})$=0.0081, el cual obtenemos por medio del programa(2) creado en matlab, para las estimaciones del parámetro objetivo, como de su varianza.

4.4.2 Estimación Convencional:

La estimación convencional para hallar el nivel promedio de arsénico en los pobladores de la cuenca del rió Ayash en el departamento de Ancash utilizando el programa STATA es:

		Error	Intervalo de		
Media	**Estimación**	**Estándar**	**Confianza al 95%**		**Deff**
Arsénico	2.3452	0.0747	2.1981	2.4923	1.0196

La concentración promedio de arsénico en la sangre de los pobladores de la Cuenca del rió Ayash en la Provincia de Huari Departamento de Ancash bajo la estimación convencional es de 2.34 ug/dl y el error estándar es de 0.0747.

(2)El programa se encuentra en el anexo A.2.

4.4.3 Comparación de Estimaciones:

Para poder realizar las comparaciones entre el método de estimación con ajuste de información auxiliar y el método de estimación convencional necesitamos hacerlo en valores relativos, para lo cual calcularemos el coeficiente de variación para cada método de estimación.

El coeficiente de variación se calcula mediante:

$$\hat{C}v = \frac{\sqrt{\text{var}(\hat{Y})}}{\hat{Y}}$$

$\hat{C}v$ (estimación convencional) = 0.319= 3.19%

$\hat{C}v$ (estimación con ajuste) = 0.396=3.96%

Podemos observar que el coeficiente de variación para la estimación convencional es menor que para la estimación con ajuste de información auxiliar, con lo que podemos afirmar que para este caso el método de estimación con ajuste de información auxiliar no proporciona una mejor estimación de la variable concentración promedio de arsénico en la sangre de los pobladores en la cuenca del rió Ayash, esto se debe a que la variable auxiliar elegida no esta muy correlacionada con la variable objetivo(nivel de arsénico)como se puede verificar en el anexo A.4.

De la misma forma que hemos estimado la concentración promedio de arsénico en la sangre para los pobladores de la cuenca del rió Ayash estimaremos las concentraciones promedio de cobre, plomo, zinc y mercurio en la sangre de los pobladores de la cuenca del rió Ayash y realizaremos igualmente la comparación con el método de ajuste con información auxiliar y con el método convencional para conocer cual método obtiene una mejor estimación de las variables objetivos. A continuación se muestra una tabla resumen de las estimaciones de los 4 metales por ambos métodos de estimación.

Tabla N°4.52: Tabla resumen de las estimaciones de las variables objetivos según el método de estimación

Niveles de metales en la sangre a estimar	**Método de estimación con información auxiliar**			**Método de estimación convencional**		
	$\hat{\bar{Y}}$ **(ug/dl)**	$EE(\hat{\bar{Y}})$	$\hat{C}v$	$\hat{\bar{Y}}$ **(ug/dl)**	$EE(\hat{\bar{Y}})$	$\hat{C}v$
Cobre	106.20	2.13	2.01%	110.18	1.09	0.99%
Plomo	3.98	0.11	2.73%	4.12	0.08	1.89%
Zinc	81.19	2.56	3.15%	83.70	2.10	2.51%
Mercurio	0.105	0.0079	7.57%	0.108	0.0073	6.67%

En el cuadro resumen podemos apreciar que la estimación bajo el método convencional para las diferentes concentraciones de metales en la sangre dan un mejor ajuste que si utilizamos la estimación con uso de información auxiliar, esto debido a que no existe correlación entre los parámetros objetivos y la variable auxiliar Edad del poblador, como se podrá observar en el Anexo A.4.

CONCLUSIONES Y RECOMENDACIONES

- En esta monografía hemos desarrollado el ajuste estadístico de muestras, considerando dos grandes bloques de estimadores:

  – Métodos utilizados con Máxima Información Auxiliar, en los que se obtienen elevadores celda a celda, estos métodos los hemos analizado en dos situaciones: estratificación y post-estratificación, cuando la información auxiliar es cualitativa los elevadores resultan del cociente entre el número de efectivos poblacionales por el número de efectivos muéstrales de cada celda. En el caso de información auxiliar cuantitativa, se ha presentado el Método de la Razón(Ratio), en el que los elevadores resultan del cociente entre los totales poblacionales y muéstrales de la variable auxiliar en las celdas obtenidas.

  – Procedimientos iterativos utilizados cuando se dispone únicamente de distribuciones auxiliares Univariantes, solo cuando las variables auxiliares son cualitativas, únicamente. Como métodos de ajuste se han presentado el Raking Usual y una variación de este.

- El total de la población Enferma estimada en la Cuenca rió Ayash en la Provincia de Huari Departamento de Ancash con ajuste con información auxiliar es de 1,417 pobladores, mientras con el ajuste convencional es de 1423 pobladores enfermos.

- Al realizar la comparación de ambas estimaciones observamos que la estimación con ajuste de información auxiliar proporciona una mejor estimación del total de enfermos que la estimación convencional, esto se debe a que las variables auxiliares elegidas para realizar la post-estratificación están correlacionadas con la variable objetivo (población de enfermos).

- La concentración promedio de Arsénico en la sangre de los pobladores de la cuenca del rió Ayash es 2.2634 ug/dl con la estimación con información auxiliar, mientras que bajo la estimación convencional es de 2.34 ug/dl.

- Al realizar la comparación de ambas estimaciones observamos que la estimación convencional proporciona una mejor estimación de la concentración promedio de arsénico en la sangre de los pobladores de la cuenca del rio Ayash departamento de Ancash que la producida por la estimación con uso de información auxiliar, esto debido a que la variable auxiliar elegida(Edad total) no esta correlacionada con la variable objetivo(nivel de arsénico).

- Un aspecto importante para el ajuste de muestras es el uso de métodos iterativos, tanto con variables auxiliares cualitativas como cuantitativas. Por tanto, es conveniente desarrollar programas informativos para estos métodos que faciliten su uso en el análisis de datos por muestreo.

- A la hora de utilizar estos métodos de ajuste a menudo se presenta el problema de celdas vacías o muy pequeñas: el tamaño muestral de las celdas se establece en un mínimo de 20 o 25. Para evitar esto, se considera el colapso de celdas para todas aquellas que no superen un tamaño mínimo.

- La variable auxiliar a elegir debe estar lo mas correlacionada posible con la variable objetivo a ser estimada para que la estimación con uso de información auxiliar sea mas precisa que sin ella.

REFERENCIAS BIBLIOGRAFIA

[1] E. Bueno, A. Zarraga y A. Iztueta. Ajuste de muestras con información auxiliar –Documento de trabajo del Instituto Vasco de Estadística (EUSTAT) N° 9801.

[2] Raj D.Teoría del Muestreo. Fondo de cultura económica. México, 1984.

[3] Cochran, W.Técnicas de Muestreo. Cecsa. México, 1980.

[4] Kish L.Muestreo de Encuestas. Trillas. México, 1982.

[5] Azorín Francisco.Métodos y aplicaciones del muestreo, España, 1994.

[6] Pérez César.Técnicas de muestreo estadístico: Teoría, práctica y aplicaciones. España 2000.

[7] J. N. K Rao. "Raking Ratio Estimators", January 1974;

[8] Brackstone y Rao . AN INVESTIGATION OF RAKING RATIO ESTIMATORS. Sankhya –1979,Volume 41, Series C, pp. 97—114.

[9] Bolfarine H. Amostragem. 11vo. SINAPE. Brasil, 1993.

[10] Dalenius T.Elements of Surveys Sampling. Sarec. Suecia, 1985.

[11] Copeland, K. R. , Pertzmeier, F.K. & Hoy, C.E. An Alternative Method of Controlling Current Population Survey Estimates to Population Counts.Survey Methodology. December 1987 Vol.13 N°2.

[12] Dupont, F. Alternative Adjustments Where There Are Several Levels of Auxiliary Information.Survey Methodology. December 1995, Vol.21 N°2 pp. 125-135.

[13] Deming, W.E. & Stephan, F.F. On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. Annals of Mathematical Statistics. 1940, N°11 pp. 427-444.

ANEXOS

A.1.- PROGRAMA DEL CÁLCULO DEL ESTIMADOR Y LA VARIANZA CON USO DE INFORMACIÓN AUXILIAR: VARIABLE AUXILIAR CUALITATIVA.

```
clear;
clc;
a=input('ingrese el número de filas ');
b=input('ingrese el número de columnas ');
N=input('ingrese el tamaño de la población ');
n=input('ingrese el tamaño de la muestra ');
f=n/N;
F=(1-f)/n;
input('ingresar los datos de la variable auxiliar en la
poblacion');
for i=1:a
    for j=1:b
       X(i,j)=input('ingrese el valor de la celda "i" y
columna"j" ');
    end
end
input('ingresar los datos de la variable auxiliar en la
muestra');
for i=1:a
  for j=1:b
      J(i,j)=input('ingrese el valor de la celda "i" y
columna"j" ');
  end
end
X
J
for i=1:a
   for j=1:b
        w(i,j)=X(i,j)/J(i,j);
   end
end
w
input('ingresar los datos de la variable objetivo');
for i=1:a
    for j=1:b
       G(i,j)=input('ingrese el valor de la celda "i" y
columna"j" ');
```

```
    end
end
R=0;
 for i=1:a
     for j=1:b
         R=G(i,j)*w(i,j)+ R;
     end
 end
 O=ones(a,b);
 for i=1:a
     for j=1:b
         M(i,j)=G(i,j)/J(i,j);
         P(i,j)=(O(i,j)-M(i,j))*M(i,j);
         s(i,j)=J(i,j)*P(i,j)/(J(i,j)-O(i,j));
     end
  end

 for i=1:a
     for j=1:b
         W(i,j)=X(i,j)/N;
     end
 end
  E=0;
  Q=0;
  for i=1:a
     for j=1:b
         T(i,j)=O(i,j)-W(i,j);
         E=T(i,j)*s(i,j)+E;
         Q=W(i,j)*s(i,j)+Q;
     end
  end

  exp1=Q*F;
  exp2=E/(n^2);
  Var=exp1+exp2;
  input('el estimador es ');
   R
  input('  la variancia estimada es');
  V=N^2*Var
  input('  la desviación estandar es');
  SE=V^(0.5)
```

A.2.- PROGRAMA DEL CÁLCULO DEL ESTIMADOR Y LA VARIANZA CON USO DE INFORMACIÓN AUXILIAR: VARIABLE AUXILIAR CUANTITATIVA.

```
clear;
clc;
a=input('ingrese el número de filas ');
b=input('ingrese el número de columnas ');
N=input('ingrese el tamaño de la población ');
input('ingresar los datos de la variable auxiliar en la
población');
for i=1:a
    for j=1:b
       X(i,j)=input('ingrese el valor de la celda "i" y
columna"j" ');
        Nh(i,j)=input('ingrese   los   tamaños   de   la
poblacion para la celda "i" columna "j" ');
     end
end

input('ingresar los datos de la variable auxiliar en la
muestra');
for i=1:a
  for j=1:b
      J(i,j)=input('ingrese el valor de la fila "i" y
columna"j" ');
      nh(i,j)=input('ingrese los tamaños de la muestra
para la fila "i" columna "j"');
      fh(i,j)=nh(i,j)/Nh(i,j);
      exp1(i,j)=(1-fh(i,j))/nh(i,j);
      Wh(i,j)=Nh(i,j)/N;
      Wh2(i,j)=Wh(i,j)^2;
      exp2(i,j)=exp1(i,j)*Wh2(i,j);
  end
end
for i=1:a
   for j=1:b
        w(i,j)=X(i,j)/J(i,j);
   end
end

input('ingresar los datos de la variable objetivo');

for i=1:a
    for j=1:b
```

```
            G(i,j)=input('ingrese el valor de la celda "i" y
columna"j" ');
            Rh(i,j)=G(i,j)/J(i,j);
            nh1(i,j)=nh(i,j)-1;
          end
     end
     R=0;
      for i=1:a
          for j=1:b
              R=G(i,j)*w(i,j)+ R;
          end
      end
      sum=0;
       input('totales de las desviaciones ') ;
      for i=1:a
          for j=1:b
              S(i,j)=input('ingrese  los  totales  de  las
desviaciones de la fila "i" columna "j" ');
              exp3(i,j)=S(i,j)/nh1(i,j);
              exp4(i,j)=exp3(i,j)*exp2(i,j);
              sum=exp4(i,j)+sum;
          end
      end
      input('el estimador es ');
        R/N
     input('varianza de la estimación del promedio es ');
     sum
     input('la desviación estandar es ');
     sum^(0.5)
```

A.3.– ANÁLISIS DE CORRELACIÓN ENTRE LAS VARIABLES AUXILIARES GRUPO DE EDAD Y ESTADO DE EXPOSICIÓN Y LA VARIABLE OBJETIVO NÚMERO TOTAL DE ENFERMOS.

Para ver si existe correlación entre la variable objetivo Total de enfermos y las variables auxiliares cualitativas estado de exposición y grupo de edad, debemos realizar una prueba de hipótesis para verificar la existencia de correlación o dependencia, la prueba que utilizaremos para observar si existe o no correlación será la prueba Chi-Cuadrado de Independencia utilizada para variables cualitativas.

El parámetro objetivo es:

Y : Número total de enfermos.

Las variables auxiliares son:

- A: Grupo de Edad (L_1=4)

  1: De 5 a 9 años.
  2: De 10 a 17 años.
  3: Mujer Adulta.
  4: Hombre Adulto.

- B: Estado de Exposición (L_2=2)
  1: Expuesto
  2: No Expuesto

Tamaño de la población N=2098 y el tamaño muestral n=308.

Para que exista un ajuste de la estimación de la variable objetivo debe elegirse variables auxiliares que estén correlacionadas con la variable objetivo.

a) Prueba de independencia entre el estado de enfermedad y el estado de exposición:

H_o: El estado de la enfermedad no esta relacionado con el estado de exposición.
H_1: El estado de la enfermedad esta relacionado con el estado de exposición.

Tabla de contingencia de Estado de Exposición por Estado de Enfermedad.

Estado de Exposición	Estado de Enfermedad		**Total**
	Enfermo	Sano	
Expuesto	150	64	214
No expuesto	57	37	94
Total	207	101	308

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	2.649(b)	1	.104		
Corrección por continuidad(a)	2.238	1	.135		
Razón de verosimilitud	2.608	1	.106		
Estadístico exacto de Fisher				.115	.068
N de casos válidos	308				

a Calculado sólo para una tabla de 2x2.
b 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 30.82.

Como podemos observar el p_valor = 0.104 y con un nivel de significación de $\alpha = 0.15$, podemos concluir que (p_valor < α) se rechaza Ho, es decir el Estado de la enfermedad esta relacionado con el estado de exposición del paciente.

b) Prueba de independencia entre el estado de enfermedad y grupos de edades:

H_o: El estado de la enfermedad no esta relacionado con los grupos de edades.
H_1: El estado de la enfermedad esta relacionado con los grupos de edades.

Tabla de contingencia de Estado de la Enfermedad y Grupos de Edades

Grupo de Edad	Estado de la enfermedad		**Total**
	Sano	Enfermo	
De 5 a 9 años	12	42	54
De 10 a 17 años	42	40	82
Mujeres adultas	13	73	86
Hombres adultos	34	52	86
Total	101	207	308

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	29.338(a)	3	.000
Razón de verosimilitudes	30.428	3	.000
Asociación lineal por lineal	.338	1	.561
N de casos válidos	308		

a 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 17.71.

Como podemos observar el p_valor = 0.00 y con un nivel de significación de $\alpha = 0.15$, podemos concluir que (p_valor < α) se rechaza Ho, es decir el Estado de la enfermedad esta relacionado con los grupos de edades.

Podemos concluir que tanto el estado de exposición como los grupos de edades están correlacionados con el estado de enfermedad, es decir mediante estas dos pruebas hemos demostrado estadísticamente que la variable objetivo esta correlacionada con las variables auxiliares: Estado de exposición y grupos de edades.

A.4.– ANÁLISIS DE CORRELACIÓN ENTRE LA VARIABLE AUXILIAR EDAD Y LA VARIABLE OBJETIVO CONCENTRACIÓN DE ARSÉNICO EN LA SANGRE.

Para ver si existe correlación entre la variable objetivo concentración promedio de arsénico en la sangre y la variable auxiliar Edad total, como ambas variables son cuantitativas podemos utilizar el coeficiente de correlación de Pearson que es una medida de asociación lineal, para lo cual primero debemos inspeccionar los datos para detectar valores atípicos (que pueden producir resultados equívocos) y evidencias de una relación lineal que lo podríamos realizar con el grafico de dispersión de ambas variables, ya que dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación de Pearson no será un estadístico adecuado para medir su asociación y deberíamos utilizar otro coeficiente de correlación adecuado para cuantificar la relación entre ambas variables.

El parámetro objetivo es:

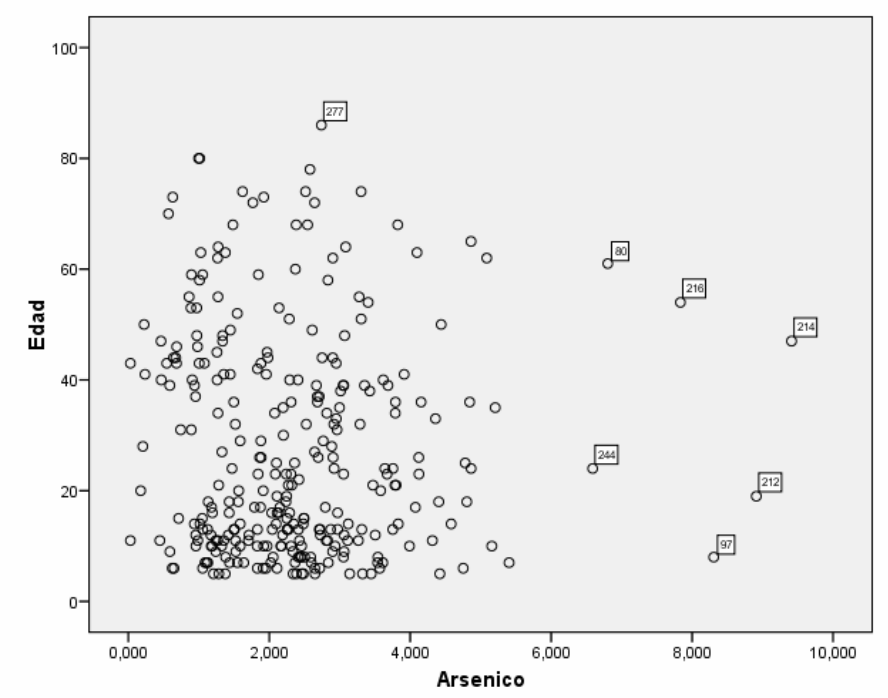
$\bar{Y}$: Concentración promedio de arsénico en la sangre.

La variable auxiliar es:

X: Edad del poblador

El gráfico de dispersión nos ayudara a determinar visualmente la relación que existe entre las dos variables cuantitativas.

Gráfico Nº A.4.1: Gráfico de dispersión entre la concentración de Arsénico en la sangre y la Edad.



Se puede observar claramente en el gráfico de dispersión entre la variable edad y concentración de arsénico en la sangre que no existen relación lineal ni otra relación que no sea lineal, también se muestra la presencia de datos atípicos, al calcular el coeficiente de correlación de Pearson entre la edad y la concentración de arsénico en la sangre como es lógico es cercano a cero(−0.029), con lo cual decimos que no existe relación entre ambas variables.

De la misma forma que hemos analizado la relación que podría existir entre la variable objetivo concentración de arsénico en la sangre con la variable auxiliar edad, analizaremos las relaciones de la variable edad con las demás variables objetivos: Concentración de cobre, zinc, plomo y mercurio en la sangre de los pobladores de la cuenca del rió Ayash.

A continuación se muestra una tabla resumen de la relación de los 4 metales con la variable auxiliar edad.

Tabla Nº A.4.1: Tabla resumen de la correlación entre las variables objetivos y la variable auxiliar edad

Variables Objetivos	**Variable auxiliar Edad**	
	Gráfico de dispersión	**Coeficiente de correlación**
Cobre	Nube de puntos sin ningún patrón especifico	0.002
Plomo	Nube de puntos sin ningún patrón especifico	−0.050
Zinc	Nube de puntos que muestra una tendencia constante	−0.050
Mercurio	Nube de puntos que muestra una tendencia constante	−0.031

Como conclusión de la tabla resumen podemos indicar que ninguna concentración de metales en la sangre esta relacionada con la variable edad(variable auxiliar).

A.5.– VARIABLES DE LA ENCUESTA TOXICOLÓGICA Y LOS NOMBRES DE CAMPOS DE LA BASE DE DATOS

CUADRO Nº A.5.2: CORRESPONDENCIA ENTRE VARIABLES DE LA ENCUESTA TOXICOLOGICA Y LOS NOMBRES DE CAMPOS DE LA BASE DE DATOS

	Nombre de la variable/pregunta	Nombre del campo en la BD
Datos generales		
1.	Nº de ficha	Codigo
2.	Código de Vivienda (para realizar el muestreo)	cod_viv
3.	Fecha de Encuesta	Fechenc
4.	Fecha de Nacimiento	Fechnac
5.	Edad	Edad
6.	Grupos de edad	grupeda
7.	Sexo	Sexo
8.	Estrato de exposición	dominio
9.	Estrato de edad y sexo	estrato
10.	Grado de Instrucción	Grinstru
11.	Domicilio	Domic
12.	Tiempo de Residencia	Tpresd
13.	Procedencia	Proced
Exposición a substancias químicas		
14.	Ocupación actual Nº 1	Ocu1
15.	Ocupación actual Nº 2	Ocu2
16.	Ocupación actual Nº 3	Ocu3
17.	Trabajo actual en CMA o Contratista	Wactual
18.	Tiempo de servicio en CMA o Contratista	Tpoactua
19.	Trabajo alguna vez en CMA o Contratista	Wprevio
20.	Puesto de trabajo anterior 1	Pue1
21.	Puesto de trabajo anterior 2	Pue2
22.	Tiempo de servicio en trabajo anterior en CMA	Tpoprevi
23.	Fumiga o prepara pesticidas	Fumiga
24.	Método de trabajo	Met1
25.	Uso de equipo de protección personal 1	Equip1
26.	Uso de equipo de protección personal 2	Equip2
27.	Uso de equipo de protección personal 3	Equip3
28.	Uso de equipo de protección personal 4	Equip4
29.	Manipula solventes	Solvente
30.	Usa agua de río	Aguario
31.	Que río	QueRio
32.	Formas de uso 1	Uso1
33.	Formas de uso 2	Uso2
34.	Formas de uso 3	Uso3
35.	Que agua usa para beber 1	Bebid1
36.	Que agua usa para beber 2	Bebid2

Nombre de la variable/pregunta	Nombre del campo en la BD
37. Que agua usa para beber 3	Bebid3
Antecedentes Patológicos	
38. Tiene antecedentes patológicos	Antenf
39. Enfermedad Crónica (antecedente)	Cronic1
40. Antecedentes de convulsión	Convul
41. Antecedente familiar de convulsión	Famconvu
42. Antecedente de Traumatismo Encefalocraneano	AntTEC
43. Antecedente de dermatosis	Dermato1
44. Tiempo de ocurrencia de la dermatosis	Tpoderma
45. Tipo de lesión	Les1
46. Antecedente de Enfermedad Infecciosa	Antinfe
47. Enfermedad Infecciosa declarada 1	Infec1
48. Enfermedad Infecciosa declarada 2	Infec2
49. Toma medicamentos en forma regular	Medici1
50. Medicamento declarado 1	Med1
51. Medicamento declarado 2	Med2
52. Consume alcohol	Alcohol
53. Frecuencia de consumo de alcohol	FrecOH
54. Consume Coca	Coca
55. Frecuencia de consumo de coca	Frecoca
56. Consume tabaco	Tabaco
57. Frecuencia de consumo de tabaco	Fretabac
Datos Perinatales (referencia personal)	
58. Antecedente de gestación normal	Antgesta
59. Qué problemas tuvo su gestación	Ges1
60. Tuvo Control Pre Natal	CPN
61. Que tipo de parto fue	TipoPart
62. Quien atendió el parto	Atendpar
63. Lugar de Nacimiento	Lugnac
Inmunizaciones y Desarrollo Psicomotor	
64. Estado de inmunización	Inmuniz
65. Desarrollo Psicomotor	Despsico
Datos Obstétricos – Reproductivos (solo mujeres)	
66. Edad de la Menarquia	Menarq
67. FUR – Gestación	Furgesta
68. Edad Gestacional	Edagesta
69. Número de Gestaciones	Numgesta
70. Número de partos	Numparto
71. Número de abortos	Numabort
72. Número de partos prematuros	Numprema
73. Número de hijos fallecidos	Numfalle
74. Toma anticonceptivos hormonales	Hormona
75. Antecedentes de malformaciones congénitas en hijos	Antmalfo
Enfermedad Actual	
76. Tiene sintomatología o enfermedad actualmente	Sintomas
77. Síntomas presentes 1	Sint1
78. Síntomas presentes 2	Sint2

Nombre de la variable/pregunta	Nombre del campo en la BD
79. Síntomas presentes 3	Sint3
80. Síntomas presentes 4	Sint4
81. Síntomas presentes 5	Sint5
82. Síntomas presentes 6	Sint6
83. Síntomas presentes 7	Sint7
84. Síntomas presentes 8	Sint8
85. Síntomas presentes 9	Sint9
86. Síntomas presentes 10	Sint10
87. Síntomas presentes 11	Sint11
88. Síntomas presentes 12	Sint12
89. Diagnóstico Clínico 1	Dxclin1
90. Diagnóstico Clínico 2	Dxclin2
91. Diagnóstico Clínico 3	Dxclin3
92. Diagnóstico Clínico 4	Dxclin4
93. Enfermedad Común	Enfcomun
94. Enfermedad Ambiental	Enfambie
Evaluación Dermatológica	
95. Tuvo evaluación de dermatológica	Evaderma
96. Motivo de consulta dermatológica 1	Cder1
97. Motivo de consulta dermatológica 2	Cder2
98. Factores provocadores 1	Factpro1
99. Factores provocadores 2	Factpro2
100. Diagnóstico presuntivo dermatológico 1	Dxderm1
101. Diagnóstico presuntivo dermatológico 2	Dxderm2
102. Diagnóstico presuntivo dermatológico 3	Dxderma3
103. Diagnóstico presuntivo dermatológico 4	Dxderma4
104. Se le realizó biopsia de piel	Bppiel
105. Diagnóstico histopatológico dermatológico	Dermpat1
106. Transferencia al MINSA	TransfMinsa
107. Ficha llenada por:	Doctor
Diagnóstico Nutricional	
108. Diagnóstico Nutricional 1	DxNut1
109. Diagnóstico Nutricional 2	DxNut2
110. Diagnóstico Nutricional 3	DxNut3
Resultados del Estudio parasitológico	
111. Presencia de parásito 1	Parasit1
112. Presencia de parásito 2	Parasit2
113. Presencia de parásito 3	Parasit3
114. Presencia de parásito 4	Parasit4
115. Presencia de parásito 5	Parasit5
116. Presencia de parásito 6	Parasit6
117. Presencia de parásito 7	Parasit7
118. Número de parásitos por persona	N_parasi
119. Número de parásitos patógenos por persona	N_patog
120. Número de parásitos comensales por persona	N_comens
Funciones Vitales y Antropometría	
121. Peso	Peso

Nombre de la variable/pregunta	Nombre del campo en la BD
122. Talla	Talla
123. Frecuencia cardiaca	FC
124. Recuencia respíratoria	FR
125. Temperatura	Temp
126. Presión Arterial sistólica	PASist
127. Presión Arterial diastólica	PADiast
128. Saturación de Oxígeno	SatOxig
Toma y entrega de Muestras y encuesta de exposiciones previas a la toma de muestras	
129. Toma de muestra de sangre	Sangre
130. Toma de muestra de Orina Simple	Orinas
131. Entrega muestra de Orina de 24 horas	orina24
132. Entrega de muestra de Heces 1	heces1
133. Entrega de muestra de Heces 2	heces2
134. Prepara alimentos en olla de barro	sangre1
135. Consume alimentos en utensilios de barro	sangre2
136. Consumo de atún en la semana	sangre3
Nombre de la variable/pregunta	**Nombre del campo en la BD**
137. Pintado de pared en la última semana	sangre4
138. Trabaja en CMA	sangre5
139. Consumo de truchas o ranas	orina1
140. Toma de pastillas	orina2
141. Consumo de atún en la semana	orina3
142. Pintado de pared en la semana	orina4
143. Fumador	orina5
Exámenes de Laboratorio: Sangre, Bioquímico y de Orina Simple	
144. Niveles de Beta 2 Microglobulina (Orina)	b2glob
145. Valores de Hematies	hematies
146. Valores de Leucocitos	leucocit
147. Valores de Plaquetas	plaquets
148. Valores de Hemoglobina	hb
149. Valor de Hemoglobina ajustada	hb_ajust
150. Altitud sobre el nivel del mar en que vive el paciente	altitud
151. Valores de Diagnóstico anemia	Dx_anem
152. Valores de Hematocrito	hto
153. Valores de Volumen Corpuscular Medio	vcm
154. Valores de Hemoglobina Corpuscular Media	hcm
155. Valores de Concentración media hemoglobina corpuscular	ccmh
156. Porcentaje de Eosinofilos	peosinof
157. Porcentaje de Basofilos	pbasofil
158. Porcentaje de Mielocitos	pmieloci
159. Porcentaje de Metamielocitos	pmetmiel
160. Porcentaje de Abastonados	pabaston
161. Porcentaje de Segmentados	psegment
162. Porcentaje de Linfocitos	plinfoci
163. Porcentaje de Monocitos	pmonocit
164. Valores de Eosinofilos	eosinofil

Nombre de la variable/pregunta	Nombre del campo en la BD
165. Valores de Basofilos	basofil
166. Valores de Mielocitos	mielocit
167. Valores de Metamielocitos	metamiel
168. Valores de Abastonados	abaston
169. Valores de Segmentados	segment
170. Valores de Linfocitos	linfocit
171. Valores de Monocitos	monocit
172. Valores de serología de Hidatidosis, Elisa	hidatid
173. Valores de Inmunoglobulina E (Ig E)	IgE
174. Color de orina	ori_colo
175. Aspecto de orina	ori_aspe
176. Ph de orina	ori_ph
177. Densidad de orina	ori_dens
178. Glucosa en orina	ori_gluc
179. Proteinas en orina	ori_prot
180. Cuerpos Cetonicos en orina	ori_ccet
181. Urobilinogeno en orina	ori_urob
182. Pigmentos Biliares en orina	ori_pigm
183. Sangre en orina	ori_sang
184. Nitritos en orina	ori_nitr
185. Celulas Epiteliales en orina	ori_cepi
186. Leucocitos en orina	ori_leuc
Nombre de la variable/pregunta	**Nombre del campo en la BD**
187. Hematies en orina	ori_hema
188. Cilindros en orina	ori_cili
189. Cristales en orina	ori_cris
190. Filamento Mucoide en orina	ori_film
191. Germenes en orina	ori_germ
192. Levaduras en orina	ori_leva
193. Tricomonas en orina	ori_tric
194. Espermatozoides en orina	ori_espr
195. Valores de Proteina Total	protot
196. Valores de Albumina	albumin
197. Valores de Globulinas	globulin
198. Valores de Relacion Albumina / Globulina	r_alb_gl
199. Valores de Bilirrubina Total	bili_tot
200. Valores de Bilirrubina Directa	bili_dir
201. Valores de Bilirrubina Indirecta	bili_ind
202. Valores de Trasaminasa Tgo (Asat)	TGO
203. Valores de Trasaminasa Tgp (Alat)	TGP
204. Valores de Fosfatasa Alcalina	FA
205. Valores de Gamma Glutamil Transpeptidasa	gam_glut
206. Diagnóstico de Beta 2 Microglobulina (Orina)	dx_b2glob
207. Recuento de Hematíes	dx_hematies
208. Recuento de Leucocitos	dx_leucocit
209. Recuento de Plaquetas	dx_plaquets

Nombre de la variable/pregunta	Nombre del campo en la BD
210. Recuento de Eosinofilos	dx_eosinofil
211. Recuento de Segmentados	dx_segment
212. Recuento de Linfocitos	dx_linfocit
213. Recuento de Monocitos	dx_monocit
214. Serología de Hidatidosis, Elisa	dx_hidatid
215. Niveles de Inmunoglobulina E (Ig E)	dx_IgE
216. Diagnóstico de Proteina Total	dx_proto
217. Diagnóstico de Albumina	dx_album
218. Diagnóstico de Globulinas	dx_globu
219. Diagnóstico relación Albumina/Globulinas	dx_albgl
220. Estado de Bilirrubina Total	dx_bilto
221. Estado de Bilirrubina Directa	dx_bildi
222. Estado de Bilirrubina Indirecta	dx_bilin
223. Diagnóstico de Trasaminasa Tgo (Asat)	dx_TGO
224. Diagnóstico de Trasaminasa Tgp (Alat)	dx_TGP
225. Diagnóstico de Fosfatasa Alcalina	dx_FA
226. Diagnóstico de Gamma Glutamil Transpeptidasa	dx_gamag
Datos de laboratorio interpretados por los médicos (interpretación cualitativa)	
227. Estado de leucocitos en sangre (cualitativo)	leuc_med
228. Estado de plaquetas (cualitativo)	trom_med
229. Estado de linfocitos (cualitativo)	linf_med
230. Presencia de desviación izquierda (cualitativo)	dizq_med
231. Presencia de eosinofilia (cualitativo)	eosi_med
232. Estado de proteínas totales (cualitativo)	prot_med
233. Estado de albúmina (cualitativo)	albu_med
234. Presencia de disproteinemia	dispro_m
Nombre de la variable/pregunta	**Nombre del campo en la BD**
235. Estado de biblirrubinas totales (cualitativo)	bilito_m
236. Estado de Transaminasas (cualitativo)	tgo_p_m
237. Estado de Fosfatasa Alcalina (cualitativo)	fa_med
238. Estado de Gamaglutamil transferasa (cualitativo)	gama_med
239. Estado de IgE (cualitativo)	ige_med
240. Estado de Hidatidosis (cualitativo)	hidat_me
241. Estado de Beta 2 globulina (cualitativo)	beta2_me
242. Presencia de proteinuria (cualitativo)	proori_m
243. Presencia de glucosuria (cualitativo)	glcosu_m
244. Presencia de cuerpos cetónicos (cualitativo)	ceton_me
245. Presencia de Hematuria (cualitativo)	hturia_m
246. Presencia de nitritos (cualitativo)	nitri_me
247. Presencia de leucocituria (cualitativo)	lcuria_m