Centro de Investigación en Matemáticas A.C

Caracterización de las secciones electorales por su representatividad en el voto

T E S I S

que para obtener el grado de

Maestría en Ciencias con Especialidad en Probabilidad y Estadística

presenta:

Luis Fernando Contreras Cruz

Director de Tesis:
Dr. José Elías Rodríguez Muñoz

Guanajuato, Gto., México,15 Junio del 2006.

Caracterización de las secciones electorales por su representatividad en el voto

Luis Fernando Contreras Cruz

15 Junio 2006

Agradecimientos

Expreso mi sincero agradecimiento al Centro de Investigación en Matemáticas A. C. (CIMAT) por el apoyo económico que recibí para la realización de mis estudios de maestría y terminación de tesis.

Quiero agradecer a mi asesor, el Dr. José Elías Rodríguez Muñoz por su apoyo brindado durante la realización de esta tesis. De igual manera, agradezco a mis sinodales por sus acertados comentarios y sugerencias, a los profesores, Dra. Eloísa Díaz-Francés Murguía y Dr. Rogelio Ramos Quiroga.

Un agradecimiento muy especial a mis padres, Bélgica y Fernando, por todo el apoyo que me han brindado durante mi existencia. A mis hermanos, Carlos Arturo y Marcos, por echarme siempre los ánimos.

A mis compañeros de maestría, Henry, Eliud y Mauricio, por los momentos que hemos pasado juntos durante este trayecto.

Contenido

Capítulo 1

Introducción

En este trabajo haremos referencia constantemente a eventos electorales. Para los fines de esta tesis, un evento electoral es el acto de un conjunto de personas para elegir un candidato a ocupar un puesto de elección popular por medio del voto. Ejemplos de eventos electorales en nuestro país son: elección de gobernador, elección de presidente de la república, elección de diputados federales, etc.

La culminación de un evento electoral se dá cuando se conocen los resultados de las votaciones, ya sea conociendo al ganador del evento o la distribución del voto entre los contendientes. Antes de esta culminación, casi siempre es de interés para los candidatos o para el público en general tener una predicción de quién será el ganador del evento electoral o una estimación de la distribución del voto en un evento electoral futuro, como en el caso de los diputados en nuestro país. En este trabajo solamente se estimará la distribución del voto en un evento electoral futuro.

Existen varias formas de hacer dichas predicciones; por ejemplo, por medio de la consulta de un panel de expertos o seleccionando una muestra de votantes potenciales y preguntándoles su intención del voto o ambas cosas. Se pueden encontrar algunas propuestas que utilizan información de una muestra para la predicción de un ganador de un evento electoral en *Bernardo(1984)* y *Yu y Lam(1997).*

En este mismo sentido, la forma de seleccionar una muestra de votantes es una parte

importante para la predicción de un ganador de un evento electoral; ya sea por medio del estudio de la intención del voto, una encuesta de salida o un conteo rápido.

Es deseable entonces que la forma de seleccionar una muestra de votantes garantice la obtención de la mayor cantidad de información de la intención del voto de dichos votantes. Por esto, en este trabajo se plantea una forma de seleccionar muestras de votantes que proporcionen esta mayor cantidad de información de la distribución del voto. Estos votantes serán seleccionados como conglomerados de votantes e identificados dichos conglomerados como secciones electorales.

Es conveniente precisar que el diseño de muestreo utilizado en el trabajo de *Yu y Lam (1997)* es el muestreo aleatorio simple con reemplazo; diseño que no utiliza información sobre el comportamiento del voto. Sin embargo, en el trabajo de *Bernardo (1984)* si se utiliza información sobre dicho comportamiento para seleccionar votantes en una muestra. Esta información corresponde a un evento electoral inmediato pasado y la selección de votantes, correspondería aquí, a las 20 secciones más representativas con probabilidad uno.

En el presente trabajo de tesis también utilizaremos información sobre la distribución del voto pero utilizando toda la información histórica disponible y confiable. Adicionalmente, la metodología propuesta permitirá que cualquier sección electoral tenga posibilidad de estar en la muestra. Con la precisión, como se hará hincapié más adelante, de que las secciones que proporcionen mayor información de la distribución del voto tendrán mayor probabilidad de estar en la muestra. La propiedad de que toda sección electoral tenga una probabilidad positiva de estar en la muestra es para reflejar la posibilidad de que el comportamiento del voto de cada una de estas secciones puede cambiar de un evento electoral a otro.

Para llevar a buen termino el anterior objetivo, primero se identificarán secciones electorales representativas. Para identificar tales secciones representativas nos basaremos en los resultados históricos de eventos electorales para una región determinada; como puede ser en nuestro país: un municipio, un distrito electoral, un estado o todo el país. En cada evento electoral se definirá para cada sección una medida de representatividad de la distribución del voto. Con los valores de esta medida para los diferentes eventos electorales se producirá

una medida combinada de representatividad de la distribución del voto de las secciones electorales, esta será una de las contribuciones de este trabajo. Esta parte se presentará en el Capítulo 4.

En segundo lugar, se utilizará esta medida combinada de representatividad para seleccionar muestras de secciones electorales en eventos electorales futuros. En el mismo Capítulo 4 se propone la forma de utilizar dicha medida de representatividad. Adicionalmente en este capítulo, se propone una forma eficiente de estimar los porcentajes de votos para cada uno de los candidatos en la contienda. Aunado a lo anterior, también se proporciona una forma de estimar la varianza del estimador del porcentaje de votos.

Con el propósito de evaluar la efectividad de la metodología a proponer en esta tesis, se presentará en el Capítulo 5 un estudio por simulación donde se aplica dicha metodología. Para la construcción de la medida combinada de representatividad de las secciones electorales, se utilizarán los resultados de las votaciones del año 2000 en la elección federal de diputados, senadores y presidente. Específicamente se utilizarán los resultados electorales de los estados de Aguascalientes y Zacatecas. La medida de representatividad resultante de las secciones electorales en cada uno de los anteriores estados se utilizará para evaluar la efectividad de la metodología a proponer en la reproducción de la distribución del voto de las elecciones federales del 2003 de diputados.

Con el objetivo de tener un mejor entendimiento de las herramientas estadísticas que se utilizarán en la propuesta metodológica, se presentará en el Capítulo 2 una revisión básica sobre la teoría de diseños de muestreo y técnicas estadísticas empleadas. Adicionalmente, en el Capítulo 3 se revisará el material sobre teoría de información; herramienta básica para definir la medida de representatividad de una sección electoral.

Finalmente, en el Capítulo 6 se presentarán algunos comentarios finales y posibles trabajos futuros de investigación.

Capítulo 2

Fundamentos de Muestreo y técnicas estadísticas

En ocasiones en *Estadística* queremos estudiar una población finita de interés. Cuando la población es demasiado grande, es imposible estudiar las características individuales de los elementos de la población debido a tiempo y aspectos económicos. Entonces se recurre a estudiar solamente una *muestra* de esa *población*; donde la muestra seleccionada debe reflejar las características de interés de la población. A partir de la muestra, se realizan inferencias para estudiar a toda la población.

En este capítulo se revisarán algunos métodos de selección y análisis de muestras de poblaciones finitas. Estos métodos serán utilizados en capítulos posteriores. Adicionalmente, se hará una breve revisión sobre componentes principales, herramienta utilizada también más adelante.

2.1 Población, muestra y selección de muestras

Consideremos una población constituida de N elementos etiquetados como $k = 1, 2, \ldots, N$. Denotaremos a la población finita como $U = \{1, \ldots, k, \ldots, N\}$, asumiendo que N no es necesariamente conocido. La característica de intéres para cada elemento de la población

se representará por $y_U = \{y_1, y_2, \ldots, y_N\}$. Para el objetivo del presente capítulo, $y_j \in \mathbb{R}$, $j = 1, 2, \ldots, N$. En ocasiones es conveniente modelar la característica de interés con el conjunto de variables aleatorias $Y_U = \{Y_1, Y_2, \ldots, Y_N\}$, es decir, el valor de la característica y_U se conceptúa como un posible valor de la variable aleatoria Y_U.

El espacio de probabilidad sobre el cual está definido el conjunto de variables aleatorias Y_U o la función de distribución conjunta del mismo, se denomina modelo de las características de interés.

Una *muestra* es cualquier subconjunto s de U. Por ejemplo una muestra s de tamaño n es de la forma $s = \{k_1, \ldots, k_n\}$, donde k_j representa al $k_j - ésimo$ elemento de la población U.

Para seleccionar los elementos de la población en la muestra es necesario un *procedimiento de selección*. Un procedimiento de selección es un conjunto de experimentos cuyo objetivo es seleccionar elementos de la población para integrar la muestra.

2.2 Diseño de muestreo

Si S es un conjunto de muestras de U, entonces un *diseño de muestreo* es la probabilidad P de obtener cada una de las muestras en S, siempre que $\sum_{s \in S} P(\{s\}) = 1$. Usualmente, P es la función de probabilidad resultante del *procedimiento de selección* de la muestra.

Un conjunto de utilidad práctica de muestras S es el conjunto $\{s \in S : P(\{s\}) > 0\}$. Si S tiene esta última propiedad, a este conjunto se le denomina el conjunto de posibles muestras de U bajo el diseño de muestreo P. Con un conjunto de muestras S de este tipo y con el diseño de muestreo P es posible formar el espacio de probabilidad $(S, 2^S, P)$. Más adelante se mostrará que este espacio de probabilidad es el usado para hacer inferencia, siempre que dicha inferencia utilice únicamente la aleatoriedad descrita por tal espacio.

Varios procedimientos de selección de la muestra pueden producir un mismo diseño de muestreo P. Dado un diseño de muestreo P, el conjunto de experimentos del procedimiento de selección debe ser tal que la probabilidad resultante de seleccionar la muestra coincida

con el diseño de muestreo P.

En ocasiones un diseño de muestreo se conoce por su procedimiento de selección y no por su probabilidad de seleccionar una muestra, por ejemplo, en el "*muestreo aleatorio simple sin reemplazo*".

Para ilustrar lo anterior, veamos dos procedimientos de selección para el muestreo aleatorio simple sin reemplazo(SR).

Procedimiento 1 de selección.

1.- Seleccionar con la misma probabilidad, $1/N$, un primer elemento de los N elementos de la población y no regresarlo;

2.- Seleccionar con la misma probabilidad, $1/(N-1)$, un segundo elemento de los $N-1$ elementos restantes y no regresarlo;

$\vdots$

n.- Seleccionar con la misma probabilidad, $1/(N-n+1)$, un $n-ésimo$ elemento de los $N-n+1$ elementos que quedan después de las primeras $n-1$ selecciones.

Este procedimiento de selección produce muestras s de tamaño n. La probabilidad de obtener cada muestra s es

$$P(\{s\}) = \frac{1}{\binom{N}{n}}.$$

El conjunto de posibles muestras es $S = \{s \subset U : s \text{ tiene } n \text{ elementos distintos}\}$.

Otro procedimiento de selección en un muestreo aleatorio simple sin reemplazo es el siguiente.

Procedimiento 2 de selección.

1.- Para cada elemento de la población generar un valor de una distribución $uniforme$ $(0,1)$;

2.- Ordenar en forma ascendente los elementos de acuerdo a los números generados;

3.- Seleccionar los primeros n elementos para integrar la muestra.

Observese que los dos procedimientos de selección descritos anteriormente producen el mismo diseño de muestreo; ver página 67 de *Särndal*, *Swensson* y *Wretman (1992)*.

2.3 Probabilidades de inclusión

Por la estructura del espacio de probabilidad asociada a un diseño, cualquier función del conjunto de muestras S a los números reales $\mathbb{R}$ es una variable aleatoria. En especial las funciones

$$\lambda_k(s) = \begin{cases} \alpha & \text{si} \quad k \in s,\, \alpha \neq 0 \\ 0 & \text{si} \qquad k \notin s \end{cases}, \tag{2.1}$$

para todo $k \in U$ y toda $s \in S$, son variables aleatorias.

La importancia de este conjunto de variables aleatorias en la teoría de muestreo se mostrará más adelante.

Ejemplo 1.- *Variables Indicadoras.* Sea I_k la variable aleatoria que nos indica si la unidad k está en la muestra s. Esta variable se define como

$$I_k(s) = \begin{cases} 1 & \text{si} \quad k \in s \\ 0 & \text{si} \quad k \notin s \end{cases},$$

para todo $k \in U$ y toda $s \in S$.

Ahora analicemos algunas propiedades de estas variables indicadoras. Notemos que cada variable indicadora I_k es una variable aleatoria Bernoulli con probabilidad de éxito

$$P(I_k = 1) = P(\{s \in S : s \ni k\}) = \sum_{s \ni k} P(\{s\}) := \pi_k.$$

A π_k se le conoce como la *probabilidad de inclusión de primer orden* del elemento k. A $1/\pi_k$ se le denomina *factor de expansión* y se utiliza en la estimación de parámetros poblacionales.

Dado que I_k es una variable aleatoria Bernoulli, entonces

$$E(I_k) = \pi_k, \tag{2.2}$$

y

$$Var(I_k) = \pi_k(1 - \pi_k),$$

$k = 1, \ldots, N$.

Ahora veamos como calcular la probabilidad de que la muestra contenga dos elementos, es decir,

$$\begin{aligned} P(\{j,k\} \subset s) &= P(I_j I_k = 1) \\ &= P(\{s \in S : s \supset \{j,k\}\}) \\ &= \sum_{s \supset \{j,k\}} P(\{s\}) := \pi_{jk}. \end{aligned}$$

A π_{jk} se le conoce como la *probabilidad de inclusión de segundo orden* de los elementos j y k. También es posible mostrar que

$$\begin{aligned} E(I_j I_k) &= \pi_{jk} \\ Cov(I_j, I_k) &= \pi_{jk} - \pi_j \pi_k. \end{aligned}$$

Veamos un ejemplo en donde apliquemos los conceptos de probabilidades de inclusión dados anteriormente, cuando utilizamos SR para la selección de las muestras.

Ejemplo 2.- El conjunto de posibles muestras para el SR es de la forma

$$S_{SR} = \{s \subset U : s \text{ tiene } n \text{ elementos distintos}\},$$

y la probabilidad de seleccionar cada muestra, es decir, el diseño de muestreo, es

$$P_{SR}(\{s\}) = \frac{1}{\binom{N}{n}},$$

para toda $s \in S_{SR}$.

De esta forma la probabilidad de inclusión de primer orden para cada elemento k es

$$\pi_k = \sum_{s \ni k} P(\{s\}) = \sum_{s \ni k} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

ya que existen $\binom{N-1}{n-1}$ muestras de tamaño n que contienen al elemento k.

De forma similar se obtiene que la probabilidad de inclusión de segundo orden de los elementos j y k es

$$\pi_{jk} = \sum_{s \supset \{j,k\}} P(\{s\}) = \sum_{s \supset \{j,k\}} \frac{1}{\binom{N}{n}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.$$

Cuando las muestras son de tamaño variable es útil tener la siguiente variable aleatoria, que nos proporciona el tamaño de la muestra. Esta variable se expresa como

$$\mathbf{n} = \sum_{k \in U} I_k.$$

Entonces para una muestra s seleccionada, su tamaño de muestra es

$$\mathbf{n}(s) = \sum_{k \in U} I_k(s) = \sum_{k \in s} 1,$$

donde $\mathbf{n}(s)$ es el número de elementos diferentes en s. También se tienen algunas propiedades de $\mathbf{n}$,

$$E(\mathbf{n}) = \sum_{k \in U} E(I_k) = \sum_{k \in U} \pi_k;$$

es decir, la suma de las probabilidades de inclusión de primer orden proporciona el tamaño

de muestra esperado. La expresión de la varianza del tamaño de la muestra es

$$Var(\mathbf{n}) = \sum_{k \in U} Var(I_k) + \sum_{j \neq k \in U} \sum Cov(I_j, I_k) = \sum_{k \in U} \pi_k(1 - \pi_k) + \sum_{j \neq k \in U} \sum (\pi_{jk} - \pi_j \pi_k).$$

Si la muestra s es de tamaño fijo n, entonces $\mathbf{n} = n$ con probabilidad 1 y

$$n = \sum_{k \in U} \pi_k.$$

2.4 Parámetros poblacionales y sus estimadores

En la primera sección denotamos a la característica de interés de cada elemento de la población por $y_U = \{y_1, y_2, \ldots, y_N\}$. Cualquier función de todas las y_U se le denomina *parámetro poblacional*. Si la característica de interés se modela con las variables aleatorias $Y_U = \{Y_1, Y_2, \ldots, Y_N\}$ cuya función de distribución conjunta es G_θ, entonces a θ se le denomina parámetro del modelo.

Las siguientes cantidades son ejemplos de parámetros poblacionales:

Total poblacional:

$$t = \sum_{k \in U} y_k,$$

Media poblacional:

$$\mu = \frac{1}{N} \sum_{k \in U} y_k,$$

Varianza poblacional:

$$\sigma^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu)^2 .$$

Si para cada elemento de la población se estudian dos características, digamos

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\},$$

un parámetro poblacional puede ser

$$\theta = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k}.$$

Por otro lado, un estimador del parámetro poblacional θ es de la forma

$$\widehat{\theta}\left(y_1, y_2, \ldots, y_N, \lambda_1, \lambda_2, \ldots, \lambda_N\right),$$

donde las $\lambda_i^{'}s$ están definidas por la ecuación (2.1). En particular es función de la forma $\widehat{\theta}(y_1, y_2, \ldots, y_N, I_1, I_2, \ldots, I_N)$. Si el estimador del parámetro poblacional θ es de la forma

$$\widehat{\theta}(y_1, y_2, \ldots, y_N, \lambda_1, \lambda_2, \ldots, \lambda_N),$$

la inferencia se dice basada en diseños, es decir, las propiedades estadísticas de $\widehat{\theta}$ se calculan con base al diseño de muestreo P. Si el estimador fuera de la forma $\widehat{\theta}(Y_1, Y_2, \ldots, Y_N, \lambda_1, \lambda_2, \ldots, \lambda_N)$ y sus propiedades estadísticas se calculan con base al modelo G_θ únicamente, entonces la inferencia se dice basada en modelos. Además, el parámetro poblacional $\theta(Y_1, Y_2, \ldots, Y_N)$, antes visto como $\theta(y_1, y_2, \ldots, y_N)$, es formalmente un estadístico y

$$\widehat{\theta}(Y_1, Y_2, \ldots, Y_N, \lambda_1, \lambda_2, \ldots, \lambda_N)$$

es formalmente un predictor de este estadístico.

Ejemplo 3.- Si utilizamos SR, un posible estimador del total poblacional t es

$$\widehat{t} = \frac{N}{n} \sum_{k \in U} y_k I_k.$$

Para una muestra s seleccionada, la estimación de t es

$$\widehat{t}(s) = \frac{N}{n} \sum_{k \in s} y_k.$$

2.5 Estimador de Horvitz-Thompson

Supóngase que el parámetro poblacional de interés es el total poblacional t. El estimador de Horvitz-Thompson (**HT**) $\widehat{t}_\pi$ del total poblacional t es de la forma

$$\widehat{t}_\pi = \sum_{k \in U} y_k \frac{I_k}{\pi_k},$$

siempre que $\pi_k > 0$ para todo $k \in U$ y π_k fue definida por la ecuación (2.2). Para una muestra s seleccionada, la estimación del total poblacional t es

$$\widehat{t}_\pi(s) = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

Tenemos las siguientes propiedades del estimador de Horvitz-Thompson:

Proposición 2.5.1.-

a.- El estimador es insesgado, es decir, $E(\widehat{t}_\pi) = t$.

b.- Su varianza (un parámetro poblacional) es

$$Var(\widehat{t}_\pi) = \sum_{k \in U} \frac{y_k^2}{\pi_k} + \sum_{j \in U} \sum_{k \neq j \in U} \frac{\pi_{jk}}{\pi_j \pi_k} y_j y_k - t^2. \tag{2.3}$$

c.- Un estimador insesgado de la varianza es

$$\widehat{Var(\widehat{t}_\pi)} = \sum_{k \in U} \left(\frac{1}{\pi_k} - 1 \right) y_k^2 \frac{I_k}{\pi_k} + \sum_{j \in U} \sum_{k \neq j \in U} (\pi_{jk} - \pi_j \pi_k) \frac{y_j y_k}{\pi_j \pi_k} \frac{I_j I_k}{\pi_{jk}},$$

suponiendo que $\pi_{jk} > 0$ para todo $k \neq j \in U$. Para una muestra s seleccionada, la estimación

de $Var(\widehat{t_\pi})$ es

$$\widehat{Var(\widehat{t_\pi})}(s) = \sum_{k\in s}\left(\frac{1}{\pi_k}-1\right)\frac{y_k^2}{\pi_k} + \sum_{j\in s}\sum_{k\neq j\in s}\left(\frac{1}{\pi_j\pi_k}-\frac{1}{\pi_{jk}}\right)y_j y_k.$$

d.- Si el tamaño de muestra es fijo, entonces la varianza en la parte (b) se puede expresar como

$$Var(\widehat{t_\pi}) = \sum_{i\in U}\sum_{i>j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \tag{2.4}$$

Esta cantidad se conoce como la expresión de *Sen-Yates-Grundy* de la varianza del estimador de **HT**. Además, un estimador insesgado de la anterior cantidad es

$$\widehat{Var(\widehat{t_\pi})} = \sum_{i\in U}\sum_{i>j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{I_iI_j}{\pi_{ij}}. \tag{2.5}$$

Para una muestra s seleccionada, la estimación correspondiente es

$$\widehat{Var(\widehat{t_\pi})}(s) = \sum_{i\in s}\sum_{i>j\in s}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{1}{\pi_{ij}}. \tag{2.6}$$

Demostración.- Ver *páginas 42-46 de Särndal, Swensson* y *Wretman (1992).*

2.6 Muestreo para unidades poblacionales

A continuación se mostrarán algunos procedimientos de selección de muestras. Estos procedimientos son comúnmente utilizados en la práctica, con excepción del muestreo Bernoulli que tiene un interés teórico mas que práctico.

2.6.1 Muestreo Bernoulli

El *muestreo Bernoulli* es un diseño de muestreo que produce muestras de tamaño variable. Una forma de seleccionar una muestra por este diseño es realizar N experimentos **Bernoulli** independientes y con probabilidad de éxito π. Si el $i-ésimo$ experimento es un éxito,

entonces el individuo i de la población es seleccionado en la muestra.

Este diseño tiene un valor didáctico y sirve para ilustrar un diseño de muestreo que produce muestras de tamaño variable. También, este diseño se puede utilizar para modelar el número de individuos que sí responden a un o varios ítem de un cuestionario.

Las muestras pueden ser de tamaño cero hasta el tamaño de la población. Si π es la probabilidad de éxito de los experimentos Bernoulli con los cuales se seleccionan a los individuos en la muestra, entonces el diseño de muestreo es

$$P(\{s\}) = \pi^{\mathbf{n}(s)}(1-\pi)^{N-\mathbf{n}(s)},$$

donde $\mathbf{n}$ representa a la variable aleatoria tamaño de la muestra. Es posible mostrar que las variables indicadoras $I_1, \ldots, I_N$ son variables aleatorias independientes Bernoulli(π). Entonces la variable aleatoria $\mathbf{n}$ se distribuye como una $Binomial(N, \pi)$.

De acuerdo a lo comentado arriba tenemos que las probabilidades de inclusión de primer y segundo orden son

$$\pi_j = E(I_j) = \pi;$$

para todo $j \in U$, y

$$\pi_{ij} = E(I_i I_j) = \pi^2;$$

para todos $i \neq j \in U$.

Por consiguiente, el estimador de **HT** del total poblacional t es

$$\widehat{t}_{\pi B} = \frac{1}{\pi} \sum_{j \in U} y_j I_j.$$

Para una muestra s seleccionada la estimación resultante es

$$\widehat{t}_{\pi B}(s) = \frac{1}{\pi} \sum_{j \in s} y_j.$$

Ahora, la varianza del estimador de **HT** del total poblacional t es

$$Var(\widehat{t}_{\pi B}) = \left(\frac{1}{\pi} - 1\right) \sum_{j \in U} y_j^2,$$

ya que de la ecuación (2.3), tenemos que

$$Var(\widehat{t}_{\pi B}) = \sum_{k \in U} \frac{y_k^2}{\pi} + \sum_{j \in U} \sum_{k \neq j \in U} \frac{\pi^2}{\pi\pi} y_j y_k - \left(\sum_{k \in U} y_k\right)^2 = \left(\frac{1}{\pi} - 1\right) \sum_{k \in U} y_k^2.$$

Un estimador de la varianza es

$$\widehat{Var(\widehat{t}_{\pi B})} = \left(\frac{1}{\pi} - 1\right) \sum_{j \in U} y_j^2 \frac{I_j}{\pi}.$$

Para una muestra s seleccionada la estimación resultante es:

$$\widehat{Var(\widehat{t}_{\pi B})}(s) = \left(\frac{1}{\pi} - 1\right) \frac{1}{\pi} \sum_{j \in s} y_j^2.$$

2.6.2 Muestreo aleatorio simple sin reemplazo

El *muestreo aleatorio simple sin reemplazo*, *SR*, es uno de los diseños de muestreo más conocidos. Este diseño produce muestras de tamaño n fijo. También este diseño fué mencionado en la *sección 2.2* y se recomienda utilizarlo cuando la población es homogénea con respecto a la característica que se desea estudiar. Además este diseño presupone que se tiene una lista de los individuos de la población y la cual se utilizará en la etapa de selección de la muestra. En ocasiones *SR* se utiliza por conveniencia, cuando no se tiene información que permita la elección de un diseño alternativo.

La probabilidad de seleccionar una muestra por *SR* es

$$P(\{s\}) = \frac{1}{\binom{N}{n}},$$

para toda muestra s de tamaño fijo n. Ya vimos en el ejemplo 2 de la *sección 2.3* que las probabilidades de inclusión de primer y segundo orden son

$$\pi_j = \frac{n}{N},$$

para todo $j \in U$ y

$$\pi_{jk} = \frac{n(n-1)}{N(N-1)},$$

para cualesquiera $\{j, k\} \subset U$.

Entonces el estimador de **HT** del total poblacional t queda de la siguiente manera

$$\widehat{t}_{\pi SR} = \frac{N}{n} \sum_{j \in U} y_j I_j.$$

Para una muestra s seleccionada, la estimación resultante es:

$$\widehat{t}_{\pi SR}(s) = \frac{N}{n} \sum_{j \in s} y_j.$$

La varianza del estimador de **HT** es de la manera siguiente:

$$Var(\widehat{t}_{\pi SR}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sigma^2, \tag{2.7}$$

antes de desarrollar la deducción de (2.7), obsérvese que en este caso tenemos muestras de

tamaño fijo n, por tanto la ecuación (2.3) se reduce a la ecuación (2.4). Por consiguiente

$$\begin{aligned}
Var(\widehat{t}_{\pi SR}) &= \sum_{i\in U}\sum_{i>j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2 \\
&= \frac{N-n}{n(N-1)}\sum_{i\in U}\sum_{i>j\in U}(y_i-y_j)^2 \\
&= \frac{N-n}{2n(N-1)}\{\sum_{i\in U}\sum_{j\in U}(y_i-y_j)^2-\sum_{i\in U}(y_i-y_i)^2\} \\
&= \frac{N-n}{2n(N-1)}\{2\sum_{i\in U}\sum_{j\in U}y_i^2-2\sum_{i\in U}\sum_{j\in U}y_iy_j\} \\
&= \frac{N-n}{n(N-1)}\{\sum_{i\in U}Ny_i^2-N^2\mu^2\} \\
&= \frac{N(N-n)}{n(N-1)}\sum_{i\in U}(y_i-\mu)^2 \\
&= \frac{N(N-n)}{n(N-1)}\sum_{i\in U}(y_i-\mu)^2 \\
&= \frac{N(N-n)}{n}\sigma^2 \\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2.
\end{aligned}$$

Un estimador de la varianza anterior es:

$$\widehat{Var(\widehat{t}_{\pi SR})}=\left(\frac{1}{n}-\frac{1}{N}\right)\frac{N}{N-1}\sum_{j\in U}\sum_{j<k\in U}(y_j-y_k)^2\frac{I_jI_k}{\pi_{jk}},$$

observese que la ecuación anterior se puede deducir de la ecuación (2.5). Para una muestra s seleccionada, la estimación resultante es:

$$\widehat{Var(\widehat{t}_{\pi SR})}(s)=N^2\left(\frac{1}{n}-\frac{1}{N}\right)\widehat{\sigma}_s^2, \tag{2.8}$$

donde

$$\widehat{\sigma}_s^2=\frac{1}{n-1}\sum_{j\in s}(y_j-\mu_s)^2$$

y

$$\mu_s = \frac{1}{n}\sum_{j\in s} y_j.$$

La expresión (2.8) se obtiene de forma similar a como se obtuvo la ecuación (2.7).

2.6.3 Muestreo con probabilidad proporcional al tamaño

En ocasiones se conoce el valor de una variable para cada uno de los individuos de la población, denotemos estos valores por $\{x_1, \ldots, x_N\}$. También en ocasiones, estos valores son tales que $x_k \propto y_k$ (aproximadamente) para todo $k \in U$. Si este es el caso, se acostumbra a denotar a x_k como el tamaño de la $k - ésima$ unidad poblacional.

Con lo anterior, el término *muestreo con probabilidad proporcional al tamaño*, PPT, sirve para denotar a cualquier diseño de muestreo donde las probabilidades de inclusión de primer orden son proporcionales al tamaño de la unidad poblacional, donde $\pi_k \propto x_k$.

Obsérvese que si se utiliza un diseño que produce muestras de tamaño fijo y si se utiliza el estimador de HT para estimar el total, entonces la varianza de dicho estimador se puede expresar como (ver ecuación 2.4)

$$Var(\widehat{t_\pi}) = \sum_{i\in U}\sum_{i>j\in U} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

De aquí, si π_k es aproximadamente proporcional a x_k, y por tanto a y_k, entonces el estimador tendrá una varianza pequeña. Es posible obtener esta propiedad sobre la precisión del estimador cuando se tienen muestras de tamaño variable. De lo anterior se desprende, que la combinación de un diseño con probabilidad proporcional al tamaño y el estimador de HT del total, producen estimaciones con una alta precisión.

Por lo tanto, es conveniente utilizar el muestreo PPT cuando se tiene la información disponible sobre lo que se ha denominado tamaño de la unidad poblacional.

Diseño de muestreo

En lugar de proporcionar la probabilidad de selección de la muestra, se describirá al

menos un procedimiento de selección de la muestra.

Procedimiento de selección:

1.− Ordenar a los individuos de la población en forma descendente con respecto a su valor del tamaño de la unidad;

2.− Para el elemento $k = 1$, generar un valor ε_1 de una *Uniforme*$(0, 1)$. Si

$$\varepsilon_1 \leq \frac{nx_1}{t_x},$$

entonces seleccionar al elemento $k = 1$ en la muestra, en caso contario no. Aquí $t_x = \sum_{j \in U} x_j$;

3.− Para $k = 2, 3, \ldots$, generar un valor ε_k de una *Uniforme*$(0, 1)$. Si

$$\varepsilon_k \leq \frac{(n - n_k)\, x_k}{t_{xk}},$$

entonces seleccionar al elemento k en la muestra, en caso contrario no. Aquí n_k es el número de elementos seleccionados en la muestra hasta el momento $k-1$ y $t_{xk} = x_k + x_{k+1} + \cdots + x_N$.

4.− El proceso descrito en los pasos 1 al 3 termina cuando $n_k = n$ o $k = k^*$, lo que ocurra primero, donde $k^* = \min\{k_0, N - n + 1\}$ y k_0 es el mínimo k para el cual $\frac{nx_k}{t_{xk}} > 1$;

5.− Si $n_{k^*} < n$, el proceso de selección descrito en los pasos 1 al 3 no produjo una muestra del tamaño deseado. Los $n - n_{k^*}$ elementos faltantes en la muestra son seleccionados de los restantes $N - k^* + 1$ de la población por muestreo SR.

Probabilidades de Inclusión

Es posible mostrar que las probabilidades de inclusión de primer orden son:

$$\pi_k = \begin{cases} \frac{nx_k}{t_x}, & k = 1, 2, \ldots, k^* - 1 \\ \frac{n\overline{x}_{k^*}}{t_x}, & k = k^*, 2, \ldots, N \end{cases}$$

donde

$$\overline{x}_{k^*} = \frac{t_{xk^*}}{N - k^* + 1}.$$

El diseño de muestreo descrito aquí produce probabilidades de inclusión de primer orden proporcionales al tamaño de la unidad, excepto para las últimas unidades de tamaño más pequeño.

Se puede mostrar también que las probabilidades de inclusión de segundo orden son:

$$\pi_{jk} = \begin{cases} \frac{n(n-1)}{t_x} g_j x_j x_k & \text{para} \quad 1 \leq j < k < k^* \\ \frac{n(n-1)}{t_x} g_j x_j \overline{x}_{k^*} & \text{para} \quad 1 \leq j < k^* \leq k \leq N \\ \frac{n(n-1)}{t_x} g_{k^*-1} \frac{t_{k^*} - x_{k^*-1}}{t_{k^*} - \overline{x}_{k^*}} \left(\overline{x}_{k^*}\right)^2 & \text{para} \quad k^* \leq j < k \leq N, \end{cases}$$

donde $g_1 = \frac{1}{t_{x2}}$ y para $k = 2, 3, \ldots, k^* - 1$

$$\begin{aligned} g_k &= \left(1 - \frac{x_1}{t_{x,2}}\right)\left(1 - \frac{x_2}{t_{x,3}}\right) \cdots \left(1 - \frac{x_{k-1}}{t_{x,k}}\right) / t_{x,k+1} \\ &= g_{k-1} \frac{t_k - x_{k-1}}{t_{k+1}}. \end{aligned}$$

Veamos un ejemplo ilustrando el muestreo PPT.

Ejemplo 4.- Supongamos que se desea seleccionar una muestra de tamaño $n = 2$ de una población de tamaño $N = 5$ por muestreo PPT y con:

k	x_k	t_{xk}	$\frac{nx_k}{t_{xk}}$
1	40	100	$\frac{80}{100}$
2	25	60	$\frac{50}{60}$
3	20	35	> 1
4	10	15	> 1
5	5	5	> 1

Aquí $k_0 = 3$, $k^* = \min\{k_0, N - n + 1\} = 3$,

$$\pi_1 = \frac{80}{100},\ \pi_2 = \frac{50}{100},\ \pi_3 = \pi_4 = \pi_5 = \frac{7}{30}.$$

Además

$g_1 = \frac{1}{60}$;
$g_2 = g_1 \frac{t_{x2}-x_1}{t_{x3}} = \frac{1}{60}\frac{60-40}{35} = \frac{1}{105}$;
$\overline{x}_3 = \frac{35}{3}$;
$\pi_{1,2} = \frac{2}{100}\frac{1}{60}(40)(25) = \frac{1}{3}$;
$\pi_{1,3} = \pi_{1,4} = \pi_{1,5} = \frac{2}{100}\frac{1}{60}(40)\frac{35}{3} = \frac{7}{45}$;
$\pi_{2,3} = \pi_{2,4} = \pi_{2,5} = \frac{2}{100}\frac{1}{105}(25)\frac{35}{3} = \frac{1}{18}$;
$\pi_{3,4} = \pi_{3,5} = \pi_{4,5} = \frac{2}{100}\frac{1}{105}\frac{35-25}{35-35/3}\left(\frac{35}{3}\right)^2 = \frac{1}{90}$.

2.6.4 Muestreo estratificado

En *muestreo estratificado*, la población es dividida en subpoblaciones no traslapadas llamados *estratos*. Un muestreo de probabilidad es seleccionado en cada estrato. Las selecciones en los estratos diferentes son independientes. El muestreo estratificado es un método poderoso y flexible que es ampliamente usado en la práctica.

Algunas de las razones por las cuales es conveniente utilizar muestreo estratificado son:

i).- Supongamos que se desea obtener estimaciones por separado de algunas de las subpoblaciones. Cada subpoblación de estudio puede tratarse como un estrato si la pertenencia a cada una de éstas está especificada en el marco de muestreo;

ii).- Por razones administrativas, diferentes regiones geográficas pueden tratarse como estratos.

Por variable de estratificación entendemos como la característica o las características usadas para subdividir la población en estratos. Un diseño de muestreo y un tamaño de muestra puede ser especificado en cada estrato. Frecuentemente el mismo tipo de diseño de muestreo es aplicado en todos los estratos. Un estimador puede ser especificado para cada estrato. A menudo esta elección es también hecha uniformemente para todos los estratos.

Por una estratificación de una población finita $U = \{1, \ldots, k, \ldots, N\}$ entendemos como una partición de U en H subpoblaciones. Un *estrato* es un subconjunto de la población tal que la pertenencia de un elemento al estrato es conocida. Los estratos de una población U se representarán por $U_1, \ldots, U_H$. Estos subconjuntos son tales que $U_j \cap U_k = \emptyset$ para toda

$j \neq k$ y $U_1 \cup U_2 \cup \cdots \cup U_H = U$. Por *muestreo estratificado* entendemos que una muestra s_h es seleccionada de U_h acorde a un diseño $P_h(\cdot)$ $(h = 1, \ldots, H)$ y que la selección en un estrato es independiente de las selecciones en todos los otros estratos. En esta familia de diseños se seleccionan independientemente una muestra en cada estrato y posiblemente con diseños diferentes.

La muestra total que resulta, denotado como s, estará compuesta como

$$s = \bigcup_{h=1}^{H} s_h,$$

y por la característica de la independencia,

$$P(\{s\}) = P_1(s_1)P_2(\{s_2\}) \cdots P_H(\{s_H\}).$$

Las probabilidades de inclusión de primer y segundo orden están definidos de la manera siguiente:

$$\pi_k = P(\{k \in s\}) = P_h(k \in s_h),$$

para $k \in U_h$. Las probabilidades de inclusión de segundo orden están dados por

$$\pi_{kl} = P(\{k, l\} \subset s) = P_h(\{i, j\} \in s_h) = \pi_k \pi_l$$

si k y l pertenecen a diferentes estratos U_i y U_j.

El número de elementos en el estrato h, es llamado el tamaño del estrato h, denotado por N_h, el cual suponemos conocido. Ya que los estratos forman una partición de U, tenemos que $N = \sum_{h=1}^{H} N_h$. Además, la población total puede estar descompuesta como

$$t = \sum_{k \in U} y_k = \sum_{h=1}^{H} t_h = \sum_{h=1}^{H} N_h \mu_{U_h},$$

donde, $t_h = \sum_{k \in U_h} y_k$ es el total del estrato h y μ_{U_h} es la media del estrato h. Sea $W_h = \frac{N_h}{N}$ el tamaño relativo del estrato U_h, entonces la media poblacional tiene la descomposición

$\mu_U = \sum_{h=1}^{H} W_h \mu_{U_h}$.

En muestreo estratificado, el estimador de la población total $t = \sum_{k \in U} y_k$ puede ser escrita como

$$\widehat{t}_\pi = \sum_{h=1}^{H} \widehat{t}_{h\pi},$$

donde $\widehat{t}_{h\pi}$ es el estimador de $t_h = \sum_{k \in U_h} y_k$. La varianza puede ser escrita como

$$Var_{ST}(\widehat{t}_\pi) = \sum_{h=1}^{H} Var_{ST}(\widehat{t}_{h\pi}).$$

Un estimador insesgado de la varianza esta dado por

$$\widehat{Var_{ST}(\widehat{t}_\pi)} = \sum_{h=1}^{H} \widehat{Var_{ST}(\widehat{t}_{h\pi})},$$

suponiendo que existe un estimador insesgado de la varianza $\widehat{Var_{ST}(\widehat{t}_{h\pi})}$ para cualquier h.

Muestreo aleatorio simple sin reemplazo estratificado En este diseño se selecciona una muestra por *SR* en cada estrato. Este diseño lo denotaremos por *STSR*. Las siguientes propiedades de los estimadores para *STSR* se pueden demostrar usando las técnicas utilizadas en la *sección 2.6.2*. Entonces el estimador de HT del total poblacional t es:

$$\widehat{t}_{\pi STSR} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \sum_{j \in U_h} y_j I_j \right).$$

Para una muestra $s = s_1 \cup s_2 \cup \cdots \cup s_H$ seleccionada, la estimación resultante es:

$$\widehat{t}_{\pi STSR}(s) = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \sum_{j \in s_h} y_j \right).$$

Ahora, la varianza del estimador de HT del total poblacional t es:

$$Var(\widehat{t}_{\pi STSR}) = \sum_{h=1}^{H} \left[N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_h^2 \right].$$

Un estimador de la varianza anterior es:

$$\widehat{Var(\widehat{t}_{\pi STSR})} = \sum_{h=1}^{H} \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) \frac{N_h}{N_h - 1} \sum_{j \in U_h} \sum_{j<k \in U_h} (y_j - y_k)^2 \frac{I_j I_k}{\pi_{jk}} \right].$$

Para una muestra $s = s_1 \cup s_2 \cup \cdots \cup s_H$ seleccionada, la estimación resultante es:

$$\widehat{Var(\widehat{t}_{\pi STSR})}(s) = \sum_{h=1}^{H} \left[N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \sigma_{s_h}^2 \right],$$

donde

$$\sigma_{s_h}^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_j - \mu_{s_h})^2$$

y

$$\mu_{s_h} = \frac{1}{n_h} \sum_{j \in s_h} y_j.$$

2.6.5 Muestreo de conglomerados en una sola etapa

En ocasiones no se tiene en el marco de muestreo, la lista de los individuos de la población. Sin embargo, sí se tiene una lista de subconjuntos de la población. Como no se tiene la lista de los individuos de la población, no es posible saber que elementos pertenecen a que subconjunto y muy posiblemente tampoco se conozca cuantos elementos contiene cada subconjunto. A estos subconjuntos así descritos se les denomina *conglomerados*.

En estudios socioeconómicos de una población, en general no se dispone de la lista de individuos de estudio pero sí se tiene por ejemplo una lista de viviendas de estos individuos. En estudios electorales no se tiene la lista de los votantes pero si una lista de las secciones electorales. Estos son ejemplos para ilustrar posibles conglomerados (viviendas, secciones electorales, etc.).

El muestreo de conglomerados es una familia de diseños; en los cuales los conglomerados son las unidades de muestreo.

Una razón para utilizar muestreo de conglomerados es por conveniencia si existe necesidad para utilizar este diseño. Si se tiene una lista de los individuos de la población, es mejor utilizar ésta en lugar de conglomerados. Se utiliza un diseño de muestreo que selecciona conglomerados porque no se tiene mayor información en el marco de muestreo.

Supóngase que la población está compuesta por N^I conglomerados; de tal forma que $U = U_1^I \cup \cdots U_{N^I}^I$ y $N = N_1^I + \cdots + N_{N^I}^I$, donde U_j^I es el $j - ésimo$ conglomerado y N_j^I su respectivo tamaño.

La forma de seleccionar una muestra de conglomerados es como sigue:

$a.-$ Una muestra s^Ide n^I conglomerados es seleccionada con el diseño de muestreo $P^I(\cdot)$.

$b.-$ Todos los individuos de la población en los conglomerados seleccionados forman la muestra.

Así, la muestra de individuos de la población es

$$s = \bigcup_{j \in s^I} U_j^I,$$

cuyo tamaño es $n = \sum_{j \in s^I} N_j^I$.

Las probabilidades de inclusión de primer y segundo orden inducidas por el diseño $P^I(\cdot)$ son $\pi_k = \pi_j^I$ para todo $k \in U_j^I$ y donde π_j^I es la probabilidad de inclusión de primer orden para el $j - ésimo$ conglomerado. Ahora, las probabilidades de inclusión de segundo orden son

$$\pi_{kl} = \begin{cases} \pi_j^I & \text{si} \quad k, l \in U_j^I \\ \pi_{ij}^I & \text{si} \quad k \in U_i^I \text{ y } l \in U_j^I, \end{cases}$$

donde π_{ij}^I es la probabilidad de inclusión de segundo orden para los conglomerados i e j.

Es conveniente introducir la notación simplificada

$$t_j^I = \sum_{k \in U_j^I} y_k$$

para el total del $j - ésimo$ conglomerado. La población total t puede ser expresada como

$$t = \sum_{j=1}^{N^I} t_j^I.$$

En el muestreo de conglomerados, el estimador de la población total t puede ser escrito como

$$\widehat{t}_{\pi C} = \sum_{j=1}^{N^I} t_j^I \frac{I_j^I}{\pi_j^I},$$

donde I_j^I es la indicadora de que el $j - ésimo$ conglomerado está en la muestra. Para una muestra $s = \bigcup_{j \in s^I} U_j^I$ seleccionada, la estimación correspondiente es

$$\widehat{t}_{\pi C}(s) = \sum_{j \in s^I} \frac{t_j^I}{\pi_j^I}.$$

La varianza es dada por

$$Var(\widehat{t}_{\pi C}) = \sum_{j=1}^{N^I} \frac{\left(t_j^I\right)^2}{\pi_j^I} + \sum_{i=1}^{N^I} \sum_{i \neq j=1}^{N^I} \frac{\pi_{ij}^I}{\pi_i^I \pi_j^I} t_i^I t_j^I - t^2. \tag{2.9}$$

Un estimador insesgado de la varianza anterior es

$$\widehat{Var(\widehat{t}_{\pi C})} = \sum_{j=1}^{N^I} \left(\frac{1}{\pi_j^I} - 1\right) \left(t_j^I\right)^2 \frac{I_j^I}{\pi_j^I} + \sum_{i=1}^{N^I} \sum_{i \neq j=1}^{N^I} \left(\pi_{ij}^I - \pi_i^I \pi_j^I\right) \frac{t_i^I t_j^I}{\pi_i^I \pi_j^I} \frac{I_i^I I_j^I}{\pi_{ij}^I}.$$

Para una muestra $s = \bigcup_{j \in s^I} U_j^I$ seleccionada, la estimación correspondiente es

$$\widehat{Var(\widehat{t}_{\pi C})}\,(s) = \sum_{j \in s^I} \left(\frac{1}{\pi_j^I} - 1 \right) \frac{\left(t_j^I\right)^2}{\pi_j^I} + \sum_{i \in s^I} \sum_{j \neq i \in s^I} \left(\frac{1}{\pi_i^I \pi_j^I} - \frac{1}{\pi_{ij}^I} \right) t_i^I t_j^I.$$

Si el tamaño de muestra (de conglomerados) es fijo, entonces la varianza $Var(\widehat{t}_{\pi})$ de la expresión (2.9) se reduce a

$$Var(\widehat{t}_{\pi C}) = -\frac{1}{2} \sum_{i=1}^{N^I} \sum_{j=1}^{N^I} (\pi_{ij}^I - \pi_i^I \pi_j^I) \left(\frac{t_i^I}{\pi_i^I} - \frac{t_j^I}{\pi_j^I} \right)^2.$$

Un estimador insesgado de la varianza anterior para una muestra $s = \bigcup_{j \in s^I} U_j^I$ seleccionada es

$$\widehat{Var(\widehat{t}_{\pi C})}\,(s) = -\frac{1}{2} \sum_{i \in s^I} \sum_{j \in s^I} \frac{(\pi_{ij}^I - \pi_i^I \pi_j^I)}{\pi_{ij}^I} \left(\frac{t_i^I}{\pi_i^I} - \frac{t_j^I}{\pi_j^I} \right)^2.$$

El resultado anterior nos lleva a algunas interesantes conclusiones acerca de la eficiencia de muestreo de conglomerados. De la ultima ecuación, vemos que si todos los $\frac{t_i^I}{\pi_i^I}$ son iguales, entonces $Var(\widehat{t}_{\pi C}) = 0$. Así, si podemos elegir $\pi_i^I \propto t_i^I$, entonces el muestreo de conglomerados será altamente eficiente. Si los tamaños de conglomerados N_i^I son conocidos en la etapa de planeamiento, uno puede elegir un diseño con $\pi_i^I \propto N_i^I$. Ya que $t_i^I = N_i^I \mu_{U_i^I}^I = \sum_{k \in U_I^I} y_k$, esto es una buena elección si hay poca variación entre la media de los conglomerados $\mu_{U_i^I}^I$. Si todos los $\mu_{U_i^I}^I$son iguales, en efecto tendríamos $Var(\widehat{t}_{\pi C}) = 0$.

Muestreo de conglomerados bajo SR Esto es, una muestra s^I de tamaño fijo n^I es tomada por SR de los N^I conglomerados en U^I, y todos los elementos en los conglomerados seleccionados son observados. Por el resultado presentado anteriormente, el estimador de la población total es dado por

$$\widehat{t}_{\pi CSR} = N^I \mu_{s^I}^I,$$

donde, $\mu^I_{s^I} = \frac{\sum_{i\in s^I} t^I_i}{n^I}$ es la media de los conglomerados totales t^I_i en s^I. La varianza puede ser escrita como

$$Var(\widehat{t}_{\pi CSR}) = \left(N^I\right)^2 \frac{1-f^I}{n^I}\left(\sigma^I_{tU^I}\right)^2,$$

donde, $f^I = \frac{n^I}{N^I}$ es la fracción de muestreo del conglomerado y

$$\left(\sigma^I_{tU^I}\right)^2 = \frac{1}{N^I-1}\sum_{i\in U^I}(t^I_i - \mu^I_{U^I_I})^2,$$

con $\mu^I_{U^I_I} = \frac{\sum_{i\in U^I} t^I_i}{N^I}$. El estimador de la varianza insesgado es

$$\widehat{Var(\widehat{t}_{\pi CSR})} = \left(N^I\right)^2 \frac{1-f^I}{n^I}\sigma^2_{ts^I},$$

donde,

$$\sigma^2_{ts^I} = \frac{1}{n^I-1}\sum_{i\in s^I}(t^I_i - \mu^I_{s^I})^2.$$

Muestreo estratificado con SR de conglomerados en cada estrato, *STSRC* En este caso, la población esta dividida en H estratos; $U = U_1 \cup U_2 \cup \cdots U_H$. Ahora el $h-ésimo$ estrato está compuesto de N^I_h conglomerados; $U_h = U^I_{h1} \cup U^I_{h2} \cup \cdots U^I_{hN^I_h}$. El total poblacional se puede expresar como

$$t = \sum_{h=1}^{H} t_h,$$

donde, $t_h = \sum_{j=1}^{N^I_h} t^I_{hj}$ y $t^I_{hj} = \sum_{k\in U^I_{hj}} y_k$.

Ahora, el estimador de HT del total poblacional t es:

$$\widehat{t}_{\pi STSRC} = \sum_{h=1}^{H}\left(\frac{N^I_h}{n^I_h}\sum_{j=1}^{N^I_h} t^I_{hj} I^I_{hj}\right).$$

Para una muestra $s = s_1 \cup s_2 \cup \cdots \cup s_H$ seleccionada; donde $s_h = \bigcup_{j\in s^I_h} U^I_{hj}$ y s^I_h representa a

la muestra de conglomerados seleccionados con el diseño P_h^I, la estimación resultante es:

$$\widehat{t}_{\pi STSRC}(s) = \sum_{h=1}^{H} \left(\frac{N_h^I}{n_h^I} \sum_{j \in s_h^I} t_{hj}^I \right).$$

La varianza del estimador de HT del total poblacional t es:

$$Var(\widehat{t}_{\pi STSRC}) = \sum_{h=1}^{H} \left(\left(N_h^I\right)^2 \left(\frac{1}{n_h^I} - \frac{1}{N_h^I} \right) \left(\sigma_h^I\right)^2 \right)$$

donde,

$$\left(\sigma_h^I\right)^2 = \frac{1}{N_h^I - 1} \sum_{j \in U_h^I} (t_{hj}^I - \mu_h^I)^2$$

y

$$\mu_h^I = \frac{1}{N_h^I} \sum_{j \in U_h^I} t_{hj}^I.$$

Un estimador de la varianza anterior es:

$$\widehat{Var(\widehat{t}_{\pi STSRC})} = \sum_{h=1}^{H} \left[\left(N_h^I\right)^2 \left(\frac{1}{n_h^I} - \frac{1}{N_h^I} \right) \frac{1}{n_h^I(n_h^I - 1)} \sum_{j \in U_h^I} \sum_{j<k \in U_h^I} \left(t_{hj}^I - t_{hk}^I\right)^2 I_{hj}^I I_{hk}^I \right].$$

Para una muestra $s = s_1 \cup s_2 \cup \cdots \cup s_H$ seleccionada, la estimación resultante es:

$$\widehat{Var(\widehat{t}_{\pi STSRC})}(s) = \sum_{h=1}^{H} \left(\left(N_h^I\right)^2 \left(\frac{1}{n_h^I} - \frac{1}{N_h^I} \right) \sigma_{s_h^I}^2 \right),$$

donde,

$$\sigma_{s_h^I}^2 = \frac{1}{n_h^I - 1} \sum_{j \in s_h^I} (t_{hj}^I - \mu_{s_h}^I)^2$$

y

$$\mu_{s_h}^I = \frac{1}{n_h^I} \sum_{j \in s_h^I} t_{hj}^I.$$

En la siguiente sección estudiaremos técnicas para conocer las propiedades más impor-

tantes de estimadores de razón, las cuales serán utilizadas en capítulos posteriores.

2.7 Estimador de razón

En esta sección se investigarán algunas propiedades importantes, como la varianza y el sesgo de los estimadores que se mostrarán en el *Capítulo 4*. Estos estimadores son una razón de dos estimadores de totales insesgados. En general, la razón no es un estimador insesgado. Entonces aplicaremos algunas técnicas útiles para encontrar una expresión aproximada de las varianzas de los estimadores propuestos. Los estimadores propuestos son de la siguiente forma:

$$\widehat{\theta} = \frac{\sum\limits_{k \in s} \frac{y_k}{\pi_k}}{\sum\limits_{k \in s} \frac{x_k}{\pi_k}} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}}.$$

El procedimiento que conduce del parámetro θ al estimador $\widehat{\theta}$ es un ejemplo simple de un principio general importante para la estimación de un parámetro poblacional θ que puede ser expresado como una función de varios totales poblacionales, $t_1, t_2, \ldots, t_q$, el cuál se expresa de la manera siguiente

$$\theta = f\left(t_1, \ldots, t_j, \ldots, t_q\right),$$

donde

$$t_j = \sum_{k \in U} y_{jk},$$

y $y_{1k}, \ldots, y_{qk}$ son valores para el $k - \acute{e}simo$ elemento de las variables de estudios $y_1, \ldots, y_q$. En la función $f\left(\cdot, \ldots, \cdot\right)$, reemplazamos cada total desconocido t_j por su correspondiente estimador,

$$\widehat{t}_{j\pi} = \sum_{k \in s} \frac{y_{jk}}{\pi_k}.$$

El estimador resultante de θ es

$$\widehat{\theta} = f\left(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{j\pi}, \ldots, \widehat{t}_{q\pi}\right).$$

Por *aproximación de Taylor de primer orden*, $\widehat{\theta}$ puede entonces ser aproximada por una función lineal, la cuál es más fácil de trabajar. El procedimiento general es dado en la siguiente subsección.

2.7.1 Técnica de linealización de Taylor para la estimación de la varianza

Ahora examinamos un problema que a menudo ocurre cuando se estima un parámetro poblacional θ. Supongamos que este parámetro puede ser expresado como una función de q totales poblacionales, $t_1, \ldots, t_q$

$$\theta = f\left(t_1, \ldots, t_q\right)$$

donde $t_j = \sum_{k \in U} y_{jk}$ $(j = 1, \ldots, q)$. Supongáse además que el vector $\left(y_{1k}, \ldots, y_{jk}, \ldots, y_{qk}\right)^T$ puede ser observado para $k \in s$, el cual nos permite formar los estimadores

$$\widehat{t}_{j\pi} = \sum_{k \in s} \frac{y_{jk}}{\pi_k}; \;\; j = 1, \ldots, q.$$

Por lo mencionado anteriormente, un estimador de θ es entonces

$$\widehat{\theta} = f\left(\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}\right).$$

La técnica de aproximación al estimador no lineal $\widehat{\theta}$ por un pseudoestimador, denotado por $\widehat{\theta}_0$, el cual es una función lineal de $\widehat{t}_{1\pi}, \ldots, \widehat{t}_{q\pi}$, así será fácil de calcular dicha aproximación. Si la aproximación es adecuada, $\widehat{\theta}_0$ será muy parecida a $\widehat{\theta}$ y se puede encontrar fácilmente la varianza $Var\left(\widehat{\theta}_0\right)$ como una aproximación de $Var\left(\widehat{\theta}\right)$. Un estimador de la varianza $\widehat{Var\left(\widehat{\theta}\right)}$ es también fácilmente encontrado.

La técnica para encontrar $\widehat{\theta}_0$ consiste de la aproximación de Taylor de primer orden de la función f, expandiendo la serie alrededor del punto $(t_1, \ldots, t_q)$, y despreciando los términos

restantes. En otros términos,

$$\widehat{\theta} \approx \widehat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j \left(\widehat{t}_{j\pi} - t_j\right),$$

donde

$$a_j = \frac{\partial \widehat{\theta}}{\partial \widehat{t}_{j\pi}} \Big|_{\left(\widehat{t}_{1\pi},\ldots,\widehat{t}_{q\pi}\right)=(t_1,\ldots,t_q)}.$$

Definamos $u_k = \sum_{j=1}^{q} a_j y_{jk}$, entonces la aproximación a la varianza de $\widehat{\theta}$ es obtenida como

$$\begin{aligned} AV(\widehat{\theta}) &= Var\left(\widehat{\theta}_0\right) = Var(\sum_{j=1}^{q} a_j \widehat{t}_{j\pi}) \\ &= Var\left(\sum_{k\in s} \frac{u_k}{\pi_k}\right) \\ &= \sum_{k\in U}\sum_{l\in U} \left(\pi_{kl} - \pi_k\pi_l\right) \frac{u_k}{\pi_k}\frac{u_l}{\pi_l}. \end{aligned}$$

Ahora, la estimación de u_k es de la siguiente manera

$$\widehat{u}_k = \sum_{j=1}^{q} \widehat{a}_j y_{jk},$$

para todo $k \in s$. Por consiguiente, una estimación de la varianza $Var\left(\widehat{\theta}\right)$ es dada por

$$\widehat{Var(\widehat{\theta})} = \sum_{k\in s}\sum_{l\in s} \frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_{kl}} \frac{\widehat{u}_k}{\pi_k^2}\frac{\widehat{u}_l}{\pi_l^2}.$$

Como es usual, cuando el diseño es de tamaño fijo, una alternativa para el estimador de la varianza es dado por

$$\widehat{Var(\widehat{\theta})} = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s} \frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_{kl}} \left(\frac{\widehat{u}_k}{\pi_k^2} - \frac{\widehat{u}_l}{\pi_l^2}\right)^2.$$

2.7.2 Estimación de una razón de totales poblacionales

Ahora regresamos al problema de estimar una razón entre dos totales poblacionales desconocidos,

$$\theta = \frac{t_y}{t_x} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}.$$

Si los dos totales poblacionales desconocidos son estimados, respectivamente, por $\widehat{t}_{y\pi} = \sum_{k\in s} \frac{y_k}{\pi_k}$ y $\widehat{t}_{x\pi} = \sum_{k\in s} \frac{x_k}{\pi_k}$, entonces el estimador no-lineal de θ es

$$\widehat{\theta} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}}.$$

Ahora aplicaremos la técnica de linealización de Taylor descrita en la *sección 2.7.1* para encontrar la varianza aproximada de $\widehat{\theta}$ y su estimador de la varianza. El estimador $\widehat{\theta}$ es una función de dos variables aleatorias $\widehat{t}_{y\pi}$ y $\widehat{t}_{x\pi}$,

$$\widehat{\theta} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}} = f\left(\widehat{t}_{y\pi}, \widehat{t}_{x\pi}\right).$$

Las derivadas parciales son

$$\frac{\partial\widehat{\theta}}{\partial\widehat{t}_{y\pi}} = \frac{1}{\widehat{t}_{x\pi}}; \quad \frac{\partial\widehat{\theta}}{\partial\widehat{t}_{x\pi}} = -\frac{\widehat{t}_{y\pi}}{\widehat{t}^2_{x\pi}},$$

evaluando el punto (t_y, t_x) en las derivadas anteriores se tiene que

$$\begin{aligned} a_1 &= \frac{\partial\widehat{\theta}}{\partial\widehat{t}_{y\pi}}\Big|_{(t_y,t_x)} = \frac{1}{t_x} \\ a_2 &= \frac{\partial\widehat{\theta}}{\partial\widehat{t}_{x\pi}}\Big|_{(t_y,t_x)} = -\frac{t_y}{t_x^2} = -\frac{\theta}{t_x} \end{aligned}$$

Entonces,

$$u_k = a_1 y_k + a_2 x_k = \frac{1}{t_x}\left(y_k - \theta x_k\right),$$

y

$$\widehat{u}_k = \frac{1}{\widehat{t}_{x\pi}} \left(y_k - \widehat{\theta} x_k \right)$$

y se tiene el siguiente resultado(ver *páginas 177-179 de Särndal, Swensson* y *Wretman (1992)*):

Teorema 2.7.1.-

Usando la técnica de linealización de Taylor, la estadística de razón $\widehat{\theta} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}}$ es aproximada como

$$\widehat{\theta} \approx \widehat{\theta}_0 = \theta + \frac{1}{t_x} \sum_{k \in s} \frac{y_k - \theta x_k}{\pi_k}. \tag{2.10}$$

El estimador $\widehat{\theta}$ es aproximadamente insesgado para θ, con la aproximación a la varianza, dada por

$$AV\left(\widehat{\theta}\right) = Var\left(\widehat{\theta}_0\right) = \frac{1}{t_x^2} \sum_{k \in U} \sum_{l \in U} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{y_k - \theta x_k}{\pi_k} \frac{y_l - \theta x_l}{\pi_l}. \tag{2.11}$$

El estimador de la varianza es

$$\widehat{Var\left(\widehat{\theta}\right)} = \frac{1}{\widehat{t}_{x\pi}^2} \sum_{k \in s} \sum_{l \in s} \frac{\left(\pi_{kl} - \pi_k \pi_l\right)}{\pi_{kl}} \frac{y_k - \widehat{\theta} x_k}{\pi_k} \frac{y_l - \widehat{\theta} x_l}{\pi_l}. \tag{2.12}$$

Las siguientes expresiones son algunas veces útiles,

$$\widehat{\theta}_0 = \theta + \frac{1}{t_x} \left(\widehat{t}_{y\pi} - \theta \widehat{t}_{x\pi} \right), \tag{2.13}$$

$$AV\left(\widehat{\theta}\right) = Var\left(\widehat{\theta}_0\right) = \frac{1}{t_x^2} \left[Var\left(\widehat{t}_{y\pi}\right) + \theta^2 Var\left(\widehat{t}_{x\pi}\right) - 2\theta Cov\left(\widehat{t}_{y\pi}, \widehat{t}_{x\pi}\right) \right], \tag{2.14}$$

$$\widehat{Var\left(\widehat{\theta}\right)} = \frac{1}{\widehat{t}_{x\pi}^2} \left[\widehat{Var\left(\widehat{t}_{y\pi}\right)} + \widehat{\theta}^2 \widehat{Var\left(\widehat{t}_{x\pi}\right)} - 2\widehat{\theta} \widehat{Cov\left(\widehat{t}_{y\pi}, \widehat{t}_{x\pi}\right)} \right]. \tag{2.15}$$

Bajo la aproximación mostrada en (2.10), se tiene que

$$E\left(\widehat{\theta}\right) \approx E\left(\widehat{\theta}_0\right) = \theta,$$

en otras palabras, el sesgo de $\widehat{\theta}$, aunque es distinto a cero, es aproximado por cero.

Veamos una aplicación de los resultados vistos anteriormente cuando se utiliza un diseño *SR*.

Considerese un diseño *SR*, con tamaño de muestra $n = fN$. Entonces se tiene que $\widehat{t}_{y\pi} = N\sum_{k\in s}\frac{y_k}{n} = N\overline{y}_s$, $\widehat{t}_{x\pi} = N\overline{x}_s$ y $\widehat{\theta} = \frac{\overline{y}_s}{\overline{x}_s}$. La aproximación lineal dada por (2.10) o equivalentemente por (2.13) es

$$\widehat{\theta}_0 = \theta + \frac{1}{\overline{x}_U}\frac{1}{n}\sum_{k\in s}(y_k - \theta x_k) = \theta + \frac{\overline{y}_s - \theta\overline{x}_s}{\overline{x}_U}.$$

La aproximación a la varianza dado por (2.11) o (2.14) es

$$\begin{aligned} AV\left(\widehat{\theta}\right) &= \frac{1}{\overline{x}_U^2}\frac{1-f}{n}\frac{1}{N-1}\sum_{k\in U}(y_k - \theta x_k)^2 \\ &= \frac{1}{\overline{x}_U^2}\frac{1-f}{n}\left(S_{yU}^2 + \theta^2 S_{xU}^2 - 2\theta S_{yxU}\right), \end{aligned}$$

donde S_{yxU} es la covarianza poblacional. De (2.12) o (2.15) se tiene que

$$\begin{aligned} \widehat{Var\left(\widehat{\theta}\right)} &= \frac{1}{\overline{x}_s^2}\frac{1-f}{n}\frac{1}{n-1}\sum_{k\in s}\left(y_k - \widehat{\theta}x_k\right)^2 \\ &= \frac{1}{\overline{x}_s^2}\frac{1-f}{n}\left(S_{ys}^2 + \widehat{\theta}^2 S_{xs}^2 - 2\widehat{\theta}S_{yxs}\right), \end{aligned}$$

donde

$$S_{ys}^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y}_s)^2,$$

$$S_{yxs} = \frac{1}{n-1}\sum_{k\in s}(y_k - \overline{y}_s)(x_k - \overline{x}_s),$$

y S_{xs}^2 es análogo a S_{ys}^2.

2.7.3 Distribución aproximada de un estimador

Dada la dificultad para obtener la función de distribución de un estimador $\widehat{\theta}$ del parámetro poblacional θ, se puede optar por obtener una aproximación a dicha distribución.

Para cierto tipo de estimadores y diseños de muestreo es posible obtener que

$$\frac{\widehat{\theta} - \theta}{\sqrt{\widehat{Var}\left(\widehat{\theta}\right)}},$$

tiene una distribución aproximada *Normal* $(0, 1)$, ver *página 64 de Thompson, M.E. (1997).* Esto es, cuando el tamaño de la población $N \to \infty$ y el tamaño de la muestra $n \to \infty$, entonces la distribución del cociente anterior tiende a una normal estándar.

Con base en lo anterior, podemos construir intervalos de confianza para θ de la forma

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{Var}\left(\widehat{\theta}\right)},$$

donde $z_{1-\alpha/2}$ es el cuantil de orden $1-\alpha/2$ de una normal estándar. En otras palabras, dado el comportamiento de la distribución de $\widehat{\theta}$, se espera que aproximadamente $100\,(1-\alpha)\,\%$ de los intervalos así construidos contengan al valor del parámetro poblacional θ (tengan una probabilidad de cobertura aproximada de $100\,(1-\alpha)\,\%$).

2.8 Componentes Principales

Este tema no es propio de muestreo, pero lo incluimos aquí porque será una técnica más que utilizaremos en capítulos posteriores.

Para examinar la relación entre un conjunto de k variables correlacionadas, puede ser útil transformar el conjunto de variables originales a un nuevo conjunto de variables no correlacionadas llamadas *componentes principales*. Entonces dado un conjunto de k variables

correlacionadas $\{X_1, X_2, \ldots, X_k\}$ se desea encontrar una transformación ortogonal de las anteriores variables para producir un nuevo conjunto de k variables $\{Y_1, Y_2, \ldots, Y_k\}$ que no están correlacionadas de tal forma que Y_1 se encuentre en la dirección de mayor variabilidad de los datos originales, Y_2 se encuentre en la dirección de segunda mayor variabilidad y que sea ortogonal a Y_1 y así sucesivamente. Las variables $Y_1, Y_2, \ldots, Y_k$ deben satisfacer que

$$Var(Y_1) \succeq Var(Y_2) \succeq \cdots \succeq Var(Y_k)$$

y

$$Cov(Y_i, Y_j) = 0, \text{ para todo } i \neq j.$$

A las variables $\{Y_1, Y_2, \ldots, Y_k\}$ se les llama *Componentes Principales.*

La ortogonalidad de la transformación nos permite que las distancias euclideanas entre los individuos a los que se miden estas k variables se conserve. Así, los componentes principales reducen la dimensión k a uno de dimensión menor, digamos p, ya que si las variables $\{X_1, X_2, \ldots, X_k\}$ están muy correlacionadas, se observará que la nube de puntos formada por los n individuos a los que se miden estas variables, que está en $\mathbb{R}^k$, se localiza en un subespacio $\mathbb{R}^p$. De tal manera que la suma de las varianzas de los primeros p componentes principales representen una proporción cercana a la suma total de las varianzas de las variables originales.

El primer componente principal es una combinación lineal de las variables $X_1, X_2, \ldots, X_k$,

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1k}X_k,$$

sujeto a la condición de que $a_{11}^2 + a_{12}^2 + \cdots + a_{1k}^2 = 1$. El segundo componente principal,

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2k}X_k,$$

sujeto a las condiciones

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2k}^2 = 1 \text{ y } Cov(Y_1, Y_2) = 0.$$

De igual manera se construyen los demás componentes principales.

En el análisis de los componentes principales implica encontrar los eigenvalores de la matriz de covarianza de las variables $X_1, X_2, \ldots, X_k$. La matriz es simétrica y tiene la siguiente forma

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix}$$

donde $c_{ii} = Var(X_i)$ y $c_{ij} = Cov(X_i, X_j)$.

Las *varianzas* de los componentes principales son los *autovalores* de la matriz C. Hay k *autovalores* de tal forma que son ordenados como $\lambda_1 \succeq \lambda_2 \succeq \cdots \succeq \lambda_k \succeq 0$, entonces λ_i corresponde al $i - ésimo$ componente principal

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ik}X_k.$$

En particular, $Var(Y_i) = \lambda_i$ y las constantes a_{i1}, a_{i2}, ..., a_{ik} son los elementos del correspondiente *autovector*.

Una propiedad importante de los autovalores es que

$$\lambda_1 + \lambda_2 + \cdots + \lambda_k = c_{11} + c_{22} + \cdots + c_{kk},$$

es decir, la suma de la varianza de los componentes principales es igual a la suma de las varianzas de las variables originales.

El porcentaje acumulado de varianza (PAV), se define como,

$$PAV = \left(\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{k} \lambda_i} \right) \text{x}100\%,$$

es la cantidad de varianza total acumulada hasta el $p - ésimo$ componente principal.

Finalmente, el análisis de componentes principales es una técnica matemática el cual no requiere el uso y especificación de un modelo estadístico para explicar la estructura de error. Para mayores detalles sobre los componentes principales, ver *Chatfield y Collins (1980)*, y *Manly (1986)*.

Capítulo 3

Elementos básicos de Teoría de Información

La teoría de información se origina en la teoría de comunicación establecida por Claude Shannon, ver *página viii de Cover and Thomas (1991).* En este capítulo expondremos los conceptos básicos de teoría de información desde el punto de vista probabilistíco. En particular, en el *Capítulo 4* trabajaremos con la medida de *Entropía Relativa*, la cuál caracterizará a las secciones electorales representativas. Se dirá que una sección electoral es representativa con respecto a la población, si su distribución en el voto es parecida a la distribución del voto poblacional, por tanto la sección electoral nos aporta mucha información de la distribución del voto global o poblacional.

3.1 Definición de Información

Antes de dar una definición formal de *Información* veamos una motivación del significado de ésta. Vamos a introducir este concepto partiendo de su idea intuitiva, para ello analizaremos el siguiente ejemplo: supongamos que tenemos una bolsa con 9 esferas negras y una blanca. ¿Cuánta información obtenemos si alguien nos dice que ha sacado una esfera blanca de la bolsa?. ¿Y cuánta obtenemos si después saca otra y nos dice que es negra?. La respuesta

a la primera pregunta es que nos aporta bastante información, puesto que estábamos casi seguros de que la esfera tenía que salir negra. Análogamente si hubiera salido negra diríamos que ese suceso no nos extraña, es decir, nos suministra poca información. Podemos fijarnos en la cantidad de información como una medida de la disminución de incertidumbre acerca de un suceso. Entonces si un evento tiene una alta probabilidad de ocurrir entonces nos dará poca información, si este sucede, y un evento con baja probabilidad de ocurrir nos da mucha información cuando ocurre.

Consideremos el espacio de probabilidad $(\Omega, \mathcal{F}, P)$, donde consideraremos que $\#(\Omega) = n$. Entonces la información contenida por el evento $E \in \mathcal{F}$ la denotaremos por $I(E)$. Por la discusión dada arriba, es claro que I debería ser una función decreciente de $P(E)$, esto es, si E,$F \in \mathcal{F}$ con $P(E) \leq P(F)$, entonces $I(E) \geq I(F)$. Para ver que forma tiene I, supongáse que seleccionamos aleatoriamente una carta de un mazo de 52 cartas y consideramos los siguientes eventos:

(i) la carta es un corazón (E_1),

(ii) la carta es un 7 (E_2),

(iii) la carta es el 7 de corazones $(E_1 \cap E_2)$.

Tenemos que $P(E_1) = \frac{1}{4}$, $P(E_2) = \frac{1}{13}$, $P(E_1 \cap E_2) = \frac{1}{52}$. Entonces

(a) $I(E_1 \cap E_2) \geq I(E_2) \geq I(E_1)$.

Nuestra intuición nos dice que la cantidad de información $I(E_1 \cap E_2)$ es la suma de la información contenida en E_1 y en E_2, si E_1 y E_2 son eventos independientes, como efectivamente lo son, entonces

(b) $I(E_1 \cap E_2) = I(E_1) + I(E_2)$.

Una última condición es que no se puede tener una información negativa de un evento, esto es

(c) $I(E) \geq 0$ para todo $E \in \mathcal{F}$.

Veamos en el siguiente teorema, el cual dice que sólo existe una función que satiface (a), (b) y (c). Como se puede observar la información depende de la probabilidad del evento E, es decir, I está en función de $P(E)$, entonces se usará la siguiente notación $I(P(E))$ o

simplemente $I(p)$.

Teorema 3.1.1.-

Cualquier función de la forma

$$I\left(E\right) = -K\log_a\left(P(E)\right),$$

donde a y K son constantes positivas, satisface (a), (b) y (c).

Demostración.-

Para (a) se observa que $P(E_1 \cap E_2) \leq P(E_2) \leq P(E_1)$. Entonces

$$-K\log_a(P(E_1 \cap E_2)) \geq -K\log_a(P(E_2)) \geq -K\log_a(P(E_1)),$$

por tanto,

$$I\left(E_1 \cap E_2\right) \geq I\left(E_2\right) \geq I\left(E_1\right).$$

Para (b) se tiene que $P(E_1 \cap E_2) = P(E_1)P(E_2)$. Entonces

$$\begin{aligned}\log_a P(E_1 \cap E_2) &= \log_a\left[P(E_1)P(E_2)\right] \\ &= \log_a P(E_1) + \log_a P(E_2).\end{aligned}$$

por tanto,

$$I\left(E_1 \cap E_2\right) = I\left(E_1\right) + I\left(E_2\right).$$

Por último, para (c) es fácil ver que $I\left(E\right) = -K\log_a\left(P(E)\right) \geq 0$, ya que $\log_a\left(P(E)\right) \leq 0$, para cualquier $0 < P(E) < 1$.

Algunos autores tienen un distinto enfoque de Información; por ejemplo, *ver páginas 3-6 de Kullback (1968)*, porque define la información como el logaritmo de un cociente de dos densidades de probabilidades y dice que es la información en una muestra observada para discriminar a favor de la hipótesis nula contra la hipótesis alternativa; estás son hipótesis de que la muestra observada proviene de dos distintos modelos estadísticos.

En todo este capítulo, haremos la siguiente elección $K = 1$ y $a = e$. Con la discusión anterior podemos dar una definición formal de información de un evento.

Definición.- Consideremos el espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Entonces la información contenida por el evento $E \in \mathcal{F}$ la definiremos como

$$I(E) = -\ln(P(E)).$$

Observemos que la información contenida en un evento depende sólo sobre sus probabilidades. Sea X una variable aleatoria discreta que toma valores $\{x_1, x_2, \ldots, x_n\}$ con probabilidades $\{p_1, p_2, \ldots, p_n\}$. En este caso escribiremos la información como

$$I(p_j) = I(X = x_j), 1 \leq j \leq n.$$

3.2 Entropía

Para cada valor que toma la variable aleatoria se tiene la información para cada uno de los valores. Ahora es de interés tener un promedio de esas informaciones, el promedio de las informaciones es llamada *entropía*.

Dada una variable aleatoria discreta X que toma valores $\{x_1, x_2, \ldots, x_n\}$, donde tenemos las informaciones para cada evento, $I(p_1)$, $I(p_2)$, ..., $I(p_n)$. Entonces definimos la *entropía* $H(X)$ como la información media, esto es,

$$H(X) = E(I(X)) = -\sum_{j=1}^{n} p_j \ln(p_j).$$

La entropía es una medida de incertidumbre y se puede demostrar que es la única medida de incertidumbre, *ver página 95 de Applebaum (1996).*

Antes de mencionar algunas propiedades generales de entropía, necesitamos mencionar una importante desigualdad.

Lema 3.2.1.- $\ln(x) \leq x - 1$ con igualdad sí y sólo sí $x = 1$.

Demostración.-

Observemos que la función $\ln(x)$ tiene derivada $1/x$. Entonces la tangente a $\ln(x)$ en $x = 1$ es la línea $y = x - 1$. Además, ya que la función $\ln(x)$ es concáva hacia abajo, tenemos para $x > 0$ que

$$\ln(x) \leq x - 1,$$

con igualdad sí y sólo sí $x = 1$.

Teorema 3.2.2.-

Sea X una variable aleatoria discreta, entonces

(a) $H(X) \geq 0$ y $H(X) = 0$ sí y sólo sí X toma uno de sus valores con certeza,

(b) $H(X) \leq \ln(n)$ con igualdad sí y sólo sí X se distribuye uniformemente.

Demostración.-

(a) La no negatividad de $H(X)$ se debe a que $-p_j \ln(p_j) \geq 0$ para todo p_j. Ahora supongamos que $H(X) = 0$; entonces cada $p_j \ln(p_j) = 0$, por lo tanto debemos tener para algún k $(1 \leq k \leq n)$ que $p_j = 0$ $(j \neq k)$ y $p_k = 1$.

(b) Primero supongamos que $p_j > 0$ $(1 \leq j \leq n)$. Entonces

$$\begin{aligned}
H(X) - \ln(n) &= -\left(\sum_{j=1}^{n} p_j \ln(p_j) + \ln(n)\right) \\
&= -\left(\sum_{j=1}^{n} p_j \left(\ln(p_j) + \ln(n)\right)\right) \\
&= -\left(\sum_{j=1}^{n} p_j \ln(p_j n)\right) \\
&= \left(\sum_{j=1}^{n} p_j \ln\left(\frac{1}{p_j n}\right)\right) \\
\text{Lema 3.2.1} \quad &\leq \left(\sum_{j=1}^{n} p_j \left(\frac{1}{p_j n} - 1\right)\right) \\
&= \left(\sum_{j=1}^{n} \left(\frac{1}{n} - p_j\right)\right) \\
&= 0.
\end{aligned}$$

Por el Lema 3.2.1 tenemos la igualdad sí y sólo sí $\frac{1}{p_j n} - 1 = 0$, esto es, cada $p_j = \frac{1}{n}$.

Ahora supongase que $p_k = 0$ para algún k; entonces tenemos que $-p_k \ln(p_k) = 0 < \frac{1}{n} - p_k$ y resultado sigue siendo válido.

3.3 Entropía conjunta y condicional

Si X y Y son dos variables aleatorias discretas definidas sobre $(\Omega, \mathcal{F}, P)$. Definimos la entropía conjunta de X, Y como

$$H(X,Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \ln(p_{jk}), \tag{3.1}$$

donde p_{jk} es la probabilidad conjunta de X y Y. Claramente, $H(X,Y)$ es una medida de la incertidumbre combinada de X y Y. Notemos que $H(X,Y) = H(Y,X)$.

Para explorar la relación entre dependencia y entropía más cuidadosamente necesitaremos otro concepto, llamado *entropía condicional*. Denotaremos como $p_{j/k}$ la probabilidad condicional de que $X = j$ dado que $Y = k$. Entonces definimos la *entropía condicional* de X dado que $Y = k$ como

$$H_k(X) = -\sum_{j=1}^{m} p_{j/k} \ln\left(p_{j/k}\right),$$

donde se entiende que $p_{j/k} > 0$. Ahora la entropía condicional de X dado Y la definiremos como

$$H_Y(X) = E(H.(X)) = -\sum_{k=1}^{m} p_k H_k(X).$$

A continuación veremos algunos resultados referente a entropía condicional, que son relativamente sencillos de demostrar(*ver página 99-101 de Applebaum(1996)*).

Lema 3.3.1.-

$$H_Y(X) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \ln\left(p_{j/k}\right).$$

Lema 3.3.2.-

Si X y Y son independientes, entonces

$$H_Y(X) = H(X).$$

Teorema 3.3.3.-

$$H(X,Y) = H(Y) + H_Y(X).$$

Demostración.- Como se menciona en la ecuación (3.1), tenemos que

$$\begin{aligned} H(X,Y) &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \ln\left(p_{j/k} p_k\right) \\ &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \ln\left(p_{j/k}\right) - \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \ln\left(p_k\right) \\ &= H_Y(X) + H(Y). \end{aligned}$$

Del teorema anterior se desprende que $H(X,Y) = H(X) + H_X(Y)$.

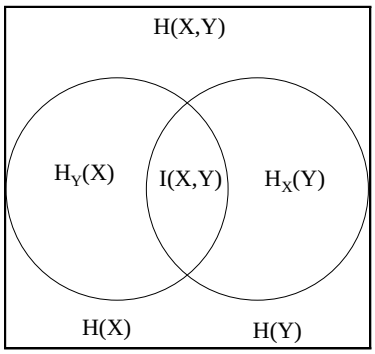
Corolario 3.3.4.- Si X y Y son independientes, entonces

$$H(X,Y) = H(X) + H(Y).$$

Se puede observar que el resultado del anterior corolario es similar a lo que sucede a la información conjunta de eventos independientes, *ver la sección 3.1.*

3.3.1 Información mutua

Ahora $H_X(Y)$ es una medida de la información que contiene Y pero la cual no está contenida en X. Por lo tanto la información contenida en Y, la cual esta contenida en X, es $H(Y) - H_X(Y)$, el cual se muestra en el siguiente diagrama.



La *información mutua* de X y Y y denotada como $I(X,Y)$, se define como:

$$I(X,Y) = H(Y) - H_X(Y).$$

A continuación tenemos algunas propiedades de la información mutua en el siguiente teorema(*ver páginas 101-102 de Applebaum (1996)*):

Teorema 3.3.5.-

(a) $I(X,Y) = \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln(\frac{p_{jk}}{p_j q_k})$.

(b) $I(X,Y) = I(Y,X)$.

(c) Si X y Y son independientes, entonces $I(X,Y) = 0$.

Demostración.-

(a) Tenemos que $H(Y) = -\sum_{k=1}^{m} q_k \ln(q_k) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln(q_k)$, entonces

$$\begin{aligned} I(X,Y) &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln(q_k) + \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln\left(p_{k/j}\right) \\ &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln(q_k) + \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln\left(\frac{p_{jk}}{p_j}\right) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk}\ln\left(\frac{p_{jk}}{p_j q_k}\right). \end{aligned}$$

(b) Es inmediato de (a) debido a que $p_{jk} = p_{kj}$ y $p_j p_k = p_k p_j$.

(c) Es una consecuencia del Lema 3.3.2.

3.4 Entropía relativa

Sean X y Y dos variables aleatorias discretas tales que sus posibles valores son $\{x_1, x_2, \ldots, x_n\}$ y $\{y_1, y_2, \ldots, y_n\}$, con probabilidades $\{p_1, p_2, \ldots, p_n\}$ y $\{q_1, q_2, \ldots, q_n\}$, respectivamente. Definiremos la *entropía relativa* $D(X, Y)$ de X y Y como

$$D(X, Y) = \sum_{j=1}^{n} p_j \ln\left(\frac{p_j}{q_j}\right).$$

La entropía relativa compara dos distribuciones de probabilidad. La definición no está en función de la dependencia entre las variables. Sin embargo, si existiera dependencia entre las variables y la medida de entropía fuera pequeña podríamos utilizar una variable para estudiar indirectamente a la otra. Está última interpretación es la que utilizaremos más adelante, cuando se use entropía relativa.

A continuación veamos algunas propiedades de la *entropía relativa*.

Teorema 3.4.1.- La entropía relativa es no negativa, es decir,

$$D(X, Y) \geq 0.$$

y con igualdad sí y sólo sí X y Y son identicamente distribuidas.

Teorema 3.4.2.- Si Y tiene una distribución uniforme, entonces

$$D(X, Y) = \ln(n) - H(X).$$

Demostración.-

Si Y tiene una distribución uniforme, entonces $q_j = \frac{1}{n}$ para todo $j = 1, \ldots, n$. Por lo

tanto,

$$\begin{aligned} D(X,Y) &= \sum_{j=1}^{n} p_j \ln\left(\frac{p_j}{q_j}\right) \\ &= \sum_{j=1}^{n} p_j \ln(np_j) \\ &= \sum_{j=1}^{n} p_j \ln(n) + \sum_{j=1}^{n} p_j \ln(p_j) \\ &= \ln(n) - H(X). \end{aligned}$$

Como se mencionó anteriormente, la medida de entropía relativa D nos será de utilidad en el *Capítulo 4* para medir la representatividad de una sección electoral en un evento electoral pasado con respecto al comportamiento global de la región(País, Estado, Municipio, etc.). Se dirá que una sección electoral es representativa con respecto a la población, si su distribución en el voto es parecida a la distribución del voto poblacional, por consiguiente, su medida de entropía relativa será cercana a cero.

Capítulo 4

Metodología Propuesta

En esta parte se pretende dar una medida de representatividad de las secciones electorales. Lo anterior se hará utilizando los resultados históricos de las votaciones. Además se presentará la manera de utilizar esta medida en la selección de una muestra de secciones. Esto es, proponer una metodología que proporcione muestras de secciones representativas para efectos de predicción de los porcentajes de votos por partido en un evento electoral futuro. Adicionalmente se pretende también proporcionar la metodología de estimación estadística que explote las características de secciones electorales representativas. La anterior metodología de muestreo y estimación estadística puede utilizarse en conteos rápidos, encuestas de salida o en estudios de intención del voto.

En cada evento electoral podemos dar una medida de representatividad de cada sección electoral. Esta medida se obtiene de la entropía relativa que resulte de comparar la distribución del voto de dicha sección con la distribución poblacional.

Si se tienen varios eventos electorales, entonces se tendrán tantas medidas de entropías como eventos. Si este es el caso, será conveniente combinar dichas medidas en una sola para facilitar su utilización. La forma como proponemos combinar estas medidas de información es calculando sus respectivos componentes principales(*ver sección 2.8*). Si las correlaciones entre las entropías son positivas, los coeficientes del primer componente principal serán positivos(*ver página 241 de Bibby, Kent* y *Mardia(1995)*). Además si la variabilidad captada

por este primer componente es alta(70% o más), entonces podemos utilizar éste como la medida combinada de representatividad del voto para cada sección electoral.

Si el primer componente principal no capta la mayoría de la variabilidad, entonces este comportamiento es indicativo que se necesitarán dos o más componentes para resumir satisfactoriamente la información proporcionada por las entropías relativas de cada evento electoral. Aquí proponemos separar los grupos eventos que influyan más en cada componente y analizar cada grupo por separado. Por ejemplo, si los eventos electorales que influyen en un componente principal corresponden a elecciones locales y nos interesa estudiar una elección local futura, entonces sugerimos utilizar únicamente los resultados de dichas elecciones para definir una medida de representatividad para cada sección electoral. Por otro lado, si alguna correlación entre entropías relativas de diferentes eventos es negativa, entonces dicho comportamiento será indicativo de que la mayoría de las secciones cambiaron su comportamiento en el voto. Comportamiento observado únicamente a través de la distribución del voto. En este caso también sugerimos agrupar eventos que tengan correlaciones positivas entre si y valorar su utilización según el tipo de evento electoral futuro a analizar.

Cabe aclarar que la metodología propuesta aquí se utilizará solamente cuando las correlaciones entre las entropías de diferentes eventos electorales sean positivas. Para los demás casos mencionados arriba y antes de utilizar mencionada metodología, debe hacerse un análisis previo de los posibles agrupamientos de eventos que convenga hacer según el evento electoral futuro a analizar.

4.1 Secciones representativas

Supóngase que en un evento electoral pasado las proporciones de votos de r partidos políticos que contendieron en la j-ésima sección son $q_{j1}, \ldots, q_{jr}$. Por otra parte, las proporciones de votos de los r partidos políticos en toda la población(Municipio, Estado o País) son $q_1, \ldots, q_r$. A partir de aquí se pueden construir las entropías relativas(*ver sección 3.4*) en todas las

secciones electorales, es decir

$$D_j = \sum_{k=1}^{r} q_k \ln\left(\frac{q_k}{q_{jk}}\right).$$

Esta cantidad se puede interpretar como la cantidad de información que proporciona la j-ésima sección sobre la distribución del voto con respecto al comportamiento poblacional(global). Consideremos un ejemplo simple para describir a una sección representativa. Supóngase que contendieron los partidos A, B y C, donde las proporciones de votos de los 3 partidos son dados en la siguiente tabla.

Partidos	A	B	C	D_j
Población	0.50	0.30	0.20	
Sección Representativa	0.49	0.31	0.20	0.00026
Sección No Representativa	0.25	0.65	0.10	0.25324

De la anterior tabla, la sección es representativa ya que la distribución en el voto es parecida a la distribución poblacional y se tiene una medida de entropía cercana a cero. Por otra parte, la sección es no representativa ya que la distribución en el voto no es parecida a la distribución poblacional y por tanto se tiene una medida de entropía mayor que en la sección representativa.

Para cada evento electoral pasado, se tienen las medidas de información para cada sección electoral. Ahora veremos como utilizar las informaciones de todos los eventos electorales pasados para dar una medida de representatividad a cada sección electoral.

4.2 Combinación de las entropías relativas

Para cada evento electoral pasado se tienen las informaciones(*Entropía Relativa*) en cada sección de la población bajo estudio. Ahora, con base en las medidas de entropías, considérese una combinación lineal de ellas para medir la representatividad de cada sección. Se usa la metodología de componentes principales (ver *sección 2.8*) de tal forma que las corre-

laciones entre las variables(Entropías relativas) sea alta y cercana a uno, con esto podemos usar el primer componente principal que concentre "satisfactoriamente" la información de cada evento electoral pasado. Se hace notar que para cada evento electoral, las secciones generan una medida de entropía relativa(Información), es decir, se está tomando en cuenta la información que generan las secciones en cada evento electoral pasado. Por ejemplo, una sección electoral puede ser representativa solamente en el evento electoral de Presidente pero no es representativa en los otros eventos electorales. Entonces una sección será más representativa si su distribución en el voto es similar al comportamiento global en todos los eventos electorales.

Supongamos que tenemos N *secciones electorales* y M *eventos electorales pasados*, entonces *proponemos* representar a las secciones electorales con los eventos electorales pasados de la siguiente manera:

$$\begin{array}{cccccc} \text{Secciones} & 1 & 2 & \cdots & M \\ 1 & D_{11} & D_{12} & \cdots & D_{1M} \\ 2 & D_{21} & D_{22} & \cdots & D_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ N & D_{N1} & D_{N2} & \cdots & D_{NM} \end{array}$$

donde D_{ij} es la entropía relativa para la $i-ésima$ sección en el $j-ésimo$ evento electoral. Aquí estamos considerando a las secciones como los "individuos" y las entropías relativas de los eventos electorales pasados como las variables. Entonces a partir de la información obtenida de todos los eventos electorales pasados en cada sección se puede tener una medida combinada. Esta medida combinada se puede construir usando los componentes principales, y se expresan de la siguiente manera:

$$C_j = a_{j1}D_1 + a_{j2}D_2 + \cdots + a_{jM}D_M,$$

para algunas constantes a_{jl} y $D_k = [D_{1k}, D_{2k}, \ldots, D_{Nk}]^T$; *ver sección 2.8 y capítulo 4 de*

Chatfield y Collins (1980). A partir de esto, se elegirán los primeros componentes principales que capten la mayor variabilidad de los datos originales. El primer componente principal será la medida combinada de representatividad para cada sección electoral en un evento electoral futuro, porque es el componente principal con mayor varianza. En general se da que los primeros 2 o 3 componentes principales suelen acumular más de un 80% de la varianza para este tipo de datos.

4.3 Selección de muestras y estimaciones

Con base en el primer componente principal, ya que siempre tiene la mayor varianza, se propone una selección de muestras de secciones para eventos electorales futuros de acuerdo al valor del primer componente principal. Como se mencionó anteriormente en este capítulo, un valor del primer componente principal cercano a cero para una sección electoral dice que la correspondiente sección es representativa en la distribución del voto global, entonces al momento de seleccionar secciones en la muestra de un evento electoral futuro estarán con mayor posibilidad las secciones más representativas. Lo importante es que se están tomando en cuenta todas las secciones en la población, ya que las secciones pueden ser representativas en algunos eventos electorales.

Bernardo(1984) realizó una predicción de votos para grupos políticos, basado en un análisis bayesiano, en la ciudad de Valencia en España, tomando una muestra pequeña de *polling stations*. Para caracterizar a las *polling stations* representativas utilizó la entropía relativa tomando solamente 20 *polling stations* de las 1774 *estaciones* existentes y se basó en el evento de Presidente. En este trabajo nosotros *proponemos* una manera se seleccionar una muestra secciones en un evento electoral futuro usando resultados de eventos electorales históricos, con base en el primer componente principal.

Se utilizan dos diseños de muestreo para la selección de secciones y para estimar porcentajes para los partidos que contendrán para un evento electoral futuro. Los dos diseños de muestreo usados son, Muestreo Aleatorio Simple sin Reeemplazo (SR) y Muestreo Aleatorio

Simple sin Reemplazo Estratificado ($STSR$).

Antes de empezar a describir los dos diseños de muestreo, veamos qué papel juegan los votos de los años 2000 y 2003. Los resultados de votos del año 2000, se utilizan para caracterizar a las secciones electorales representativas y para hacer selección de secciones, basados en las entropías relativas. Los resultados de las votaciones del año 2003, se utilizarán para evidenciar la utilidad de la metodología propuesta, su utilidad para elegir secciones electorales que proporcionen adecuadas estimaciones de la distribución del voto en un evento futuro.

Se empieza describiendo el diseño de Muestreo Aleatorio Simple sin Reeemplazo que se utilizó para seleccionar secciones.

El parámetro de interés que se desea estimar, el porcentaje de votos para el partido j para un evento electoral futuro es

$$q_j = \frac{1}{N}\sum_{k\in U} N_k q_{jk}, \tag{4.1}$$

donde N es el número de votos, N_k es el número de votos en la $k-ésima$ sección, q_{jk} es el porcentaje de votos para el partido j en la $k-ésima$ sección. Así $N_k q_{jk}$ es el número de votos para el partido j en la $k-ésima$ sección.

En este caso las secciones son seleccionadas usando *SR,* donde las secciones tienen la misma probabilidad de ser seleccionadas en la muestra. Se utiliza un tamaño de muestra n tomados de la población de tamaño N. Entonces a partir de las secciones seleccionadas en la muestra y con base en el evento electoral de *Diputados* del año *2003*, se pueden estimar las proporciones de votos(*ver sección 2.6.2*) para los partidos que contienden en un evento electoral futuro, como se hace comúnmente a traves de un estimador de razón:

$$\widehat{q}_{jSR} = \frac{\sum_{k\in U} y_{kj}\frac{I_k}{\pi_k}}{\sum_{k\in U} x_k\frac{I_k}{\pi_k}} = \frac{\frac{N}{n}\sum_{k\in U} y_{kj}I_k}{\frac{N}{n}\sum_{k\in U} x_k I_k} = \frac{\sum_{k\in U} y_{kj}I_k}{\sum_{k\in U} x_k I_k},$$

donde y_{kj} es el total de votos para el partido j en la $k-ésima$ sección y x_k es el total de votos en la $k-ésima$ sección. Para una muestra s seleccionada, por SR se tiene que

$$\widehat{q}_{jSR}(s) = \frac{\sum_{k\in s} y_{kj}}{\sum_{k\in s} x_k}.$$

Como son conocidas las proporciones de votos para los partidos en el evento electoral de Diputados 2003, se pueden calcular las entropías relativas de la distribución de votos estimadas con respecto a la distribución de votos de Diputados 2003. Supongáse que las proporciones de votos de r partidos políticos que contienden en la secciones son $\widehat{q}_1, \ldots, \widehat{q}_r$, las cuales son las estimaciones que se obtuvieron anteriormente bajo SR. Por otra parte, las proporciones de votos de los mismos r partidos políticos que contienden en toda la población son $p_1, \ldots, p_r$. Por consiguiente, se pueden construir las entropías relativas para cada muestra seleccionada s bajo SR, es decir, se tiene que

$$D_{SR}(s) = \sum_{k=1}^{r} p_k \ln\left(\frac{p_k}{\widehat{q}_{kSR}}\right).$$

Lo anterior se calcula para ver si la muestra de secciones seleccionada s bajo SR reproduce la distribución de votos de Diputados 2003, por medio del porcentaje de votos estimados para los partidos que contienden en un evento futuro.

Notése que en el diseño de muestreo SR no se está tomando en cuenta la información sobre secciones representativas, es decir, las secciones tienen la misma probabilidad de estar en la muestra seleccionada. Ahora, en el siguiente diseño de muestreo, las secciones representativas tienen mayor probabilidad de estar en la muestra seleccionada.

Ahora se describe como se seleccionan las muestras de secciones electorales bajo $STSR$.

Primero se ordenan las secciones de acuerdo al valor del primer componente principal. La ordenación se hizó en forma ascendente, en donde los primeros valores nos caracterizan a las secciones más representativas y los últimos nos caracterizan a las secciones menos

representativas. A partir de la ordenación de las secciones con su correspondiente valor del componente principal se divide a esta población en H *estratos* de tamaños iguales, de tal manera que en los primeros estratos estén las secciones más representativas, en los segundos estratos las secciones de mediana representatividad y en los últimos estratos las secciones menos representativas. Es conveniente recordar que el tamaño de la población de las secciones es N de tal manera que los tamaños de los estratos N_1, N_2,..., N_H son iguales. Para cada estrato se hace selección de secciones usando *SR*. Como se quiere que en las muestras de secciones estén las más representativas, entonces se proponen los siguientes tamaños de muestras para los estratos, n_1, n_2, ..., n_H, de tal manera que $n_1 > n_2 > \cdots > n_H$. Como se puede notar, el tamaño de muestra en el estrato de las secciones más representativas es mayor al de los otros estratos y en el estrato de las secciones menos representativas el tamaño de muestra es pequeño. Lo anterior se debe a que en la selección de muestras de secciones, se está dando mayor prioridad de estar en la muestra seleccionada a las secciones más representativas y menos prioridad a las secciones menos representativas.

Nuevamente, el parámetro de interés que se desea estimar es el porcentaje de votos para el partido j para un evento electoral futuro. Para una población estratificada, el parámetro en la ecuación (4.1) se transforma en

$$q_j = \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in U_h} N_k q_{jk}.$$

Entonces a partir de las secciones seleccionadas en la muestra y con base en el evento electoral de *Diputados* del año *2003*, proponemos estimar las proporciones de votos para los partidos que contienden en un evento electoral futuro de la siguiente manera:

$$\widehat{q}_{jSTSR} = \frac{1}{n^I}\sum_{h=1}^{H}\left(\sum_{k\in U_h}\frac{y_{hkj}}{x_{hk}}I_k\right) = \frac{1}{n^I}\sum_{h=1}^{H}\left(n_h\sum_{k\in U_h}\frac{q_{hkj}}{n_h}I_k\right),$$

donde y_{hkj} es el total de votos para el partido j en la $k-ésima$ sección para el estrato h y x_{hk} es el total de votos en la $k-ésima$ sección en el estrato h, es decir, $x_{hk} = y_{hk1}+\ldots+y_{hkr}$,

y $n^I = n_1 + \ldots + n_h$.

Como se puede notar de la última igualdad, el estimador anterior está en función de los tamaños de muestras en los estratos, n_h, y se está proporcionando un peso n_h a las estimaciones de los porcentajes de votos para el partido j en la $k - ésima$ sección en los diferentes estratos, de tal manera que el porcentaje será más preciso en el estrato en donde estén las secciones más representativas. Luego la estimación del porcentaje de votos del partido j para un evento futuro, se divide entre la suma de los tamaños de muestra de los estratos.

Ahora, para una muestra s seleccionada, por $STSR$ se tiene que

$$\widehat{q}_{jSTSR}(s) = \frac{1}{n^I} \sum_{h=1}^{H} \left(\sum_{k \in s_h} \frac{y_{hkj}}{x_{hk}} \right).$$

Como son conocidas las proporciones de votos para los partidos en el evento electoral de Diputados 2003, se pueden calcular también las entropías relativas para ver el efecto que tienen las estimaciones de los porcentajes de votos para el partido j, tomando en cuenta la información de las secciones representativas. Supongáse que las proporciones de votos de r partidos políticos que contienden en las secciones son $\widehat{q}_1, \ldots, \widehat{q}_r$, las cuales son las estimaciones que se obtuvieron anteriormente bajo $STSR$. Por otra parte, las proporciones de votos de los mismos r partidos políticos que contienden en toda la población son $q_1, \ldots, q_r$. A partir de aquí se pueden construir las entropías relativas para cada muestra seleccionada s bajo $STSR$, es decir, se tiene que

$$D_{STSR}(s) = \sum_{j=1}^{r} q_j \ln \left(\frac{q_j}{\widehat{q}_{jSTSR}} \right).$$

Cabe mencionar que no se utilizó *muestreo proporcional al tamaño* ya que las probabilidades de inclusión de primer orden π_k no son proporcionales al número de votos para el partido j en la $k - ésima$ sección, y_{kj}. Si utilizaramos este diseño con el estimador de $H\text{-}T$ tendríamos inadecuadas estimaciones del porcentaje de votos para el partido j para un

evento electoral futuro. Esas estimaciones de los porcentajes de votos nos llevaría a tener mayor variabilidad de éstas, ver *sección 2.6.3*.

4.4 Propiedades de los estimadores

En esta parte del capítulo presentaremos las varianzas de los estimadores propuestos, bajo el diseño SR y STSR, así como los respectivos estimadores de las varianzas.

4.4.1 Varianza de los estimadores propuestos

Para el diseño *SR*, veamos como obtener una expresión del estimador de la varianza del estimador del porcentaje de votos para el partido j en un evento futuro, es decir, daremos una expresión para $\widehat{Var(\widehat{q}_{jSR})}$. Recordemos que el respectivo estimador del cociente es

$$\widehat{q}_{jSR} = \frac{\sum_{k\in U} y_{kj} I_k}{\sum_{k\in U} x_k I_k} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{x\pi}}.$$

La respectiva varianza aproximada(*ver sección 2.7.2*) es

$$\begin{aligned} Var(\widehat{q}_{jSR}) &= \frac{1}{t_x^2}\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{k\in U}(y_{kj}-q_j x_k)^2\right\} \\ &= \frac{1}{\mu_x^2}\left(\frac{1}{n}-\frac{1}{N}\right)\left\{\sigma_y^2+q_j^2\sigma_x^2-2q_j\sigma_{yx}\right\}. \end{aligned}$$

Un estimador de la varianza de $\widehat{q}_{jSR}$ es

$$\widehat{Var(\widehat{q}_{jSR})} = \frac{1}{\widehat{\mu}_{x\pi}^2}\left(\frac{1}{n}-\frac{1}{N}\right)\left\{\widehat{\sigma}_y^2+\widehat{q}_{jSR}^2\widehat{\sigma}_x^2-2\widehat{q}_{jSR}\widehat{\sigma}_{yx}\right\},$$

donde,

$$\widehat{\sigma}_y^2 = \frac{1}{n-1}\left(\sum_{k\in U} y_{kj}^2 I_k - n\widehat{\mu}_{y\pi SR}^2\right),$$

$\widehat{\sigma}_x^2$ es análogo a $\widehat{\sigma}_y^2$ y el estimador de la covarianza es

$$\widehat{\sigma}_{yx}^2 = \frac{1}{n-1}\left(\sum_{k\in U} y_{kj}x_k I_k - n\widehat{\mu}_{y\pi SR}\widehat{\mu}_{x\pi SR}\right).$$

Por otra parte, para el diseño STSR, veamos como obtener una expresión del estimador de la varianza del estimador del porcentaje de votos para el partido j en un evnto futuro, es decir, daremos una expresión para $\widehat{Var(\widehat{q}_{jSTSR})}$. Por los resultados comentados en la *sección 2.6.4*, tenemos que

$$\begin{aligned}
Var\left(\widehat{q}_{jSTSR}\right) &= Var\left(\frac{1}{n^I}\sum_{h=1}^{H}\left(\sum_{k\in U_h} q_{hkj}I_k\right)\right) \\
&= Var\left(\frac{1}{n^I}\sum_{h=1}^{H}\left(\sum_{k\in U_h} (\pi_k q_{hkj})\frac{I_k}{\pi_k}\right)\right) \\
&= \frac{1}{(n^I)^2}Var\left(\sum_{h=1}^{H}\left(\sum_{k\in U_h} (\pi_k q_{hkj})\frac{I_k}{\pi_k}\right)\right) \\
&= \frac{1}{(n^I)^2}\sum_{h=1}^{H}\left[n_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{1}{N_h-1}\sum_{k\in U_h}(q_{hkj}-\overline{q_{hj}})^2\right].
\end{aligned}$$

Un estimador de la varianza anterior es:

$$\widehat{Var(\widehat{q}_{jSTSR})} = \frac{1}{(n^I)^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\frac{n_h^2}{N_h(N_h-1)}\sum_{l\in U_h}\sum_{l<k\in U_h}(q_{kjSTSR}-q_{ljSTSR})^2\frac{I_l I_k}{\pi_{lk}}\right].$$

Para una muestra $s = s_1 \cup s_2 \cup \cdots \cup s_H$ seleccionada, la estimación resultante es:

$$\widehat{Var(\widehat{q}_{jSTSR})}(s) = \frac{1}{(n^I)^2}\sum_{h=1}^{H}\left[n_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\sigma_{s_h}^2\right],$$

donde

$$\sigma_{s_h}^2 = \frac{1}{n_h-1}\sum_{k\in s_h}(q_{hkjSTSR}-\overline{q_{hjSTSR}})^2$$

y

$$\overline{q_{hjSTSR}} = \frac{1}{n_h} \sum_{k \in s_h} q_{hkjSTSR}.$$

A partir de las estimaciones propuestas de los estimadores de porcentajes de votos, bajo los dos diseños de muestreo, se pueden construir intervalos de confianza para el verdadero valor del parámetro de la siguiente manera, *ver sección 2.7.3*:

$$\widehat{q}_j(s) \pm z_{1-\alpha/2} \sqrt{\widehat{Var(\widehat{q}_j)}(s)}$$

donde $z_{1-\alpha/2}$ es el cuantil de orden $1 - \alpha/2$ de una normal estándar, ver *página 64 de Thompson, M.E. (1997).* Entonces podríamos comparar los intervalos de confianza bajo los dos diseños de muestreo, esperando que los intervalos bajo el diseño *STSR* sean más cortos y por consiguiente, se tendrían intervalos de confianza para el verdadero valor del parámetro más precisos.

Capítulo 5

Aplicación

Como una aplicación de la metodología presentada en el capítulo 4, analizaremos dos poblaciones estatales del país, los estados de Aguascalientes y Zacatecas. Los resultados de las votaciones del año 2003, se utilizarán para evidenciar la utilidad de la metodología propuesta, su utilidad para elegir secciones electorales que proporcionen estimaciones precisas de la distribución del voto en un evento electoral futuro.

5.1 Caso Aguascalientes

El estado de Aguascalientes se componía en el año 2000 de 486 secciones electorales para los tres eventos federales, presidente, diputados y senadores. Para el año 2003, se conservaron las mismas secciones electorales pero solamente ocurrió el evento electoral de diputados. Los resultados de los eventos electorales federales de los años 2000 y 2003 se resumen en las siguientes tablas.

Eventos						
Presidente	54.86%	34.47%	7.12%	0.60%	0.38%	2.57%
Diputados	52.47%	34.72%	8.39%	1.08%	0.70%	2.64%
Senadores	50.12%	38.28%	7.26%	1.17%	0.68%	2.49%

Tabla 1 Porcentajes de votos de los grupos políticos en las elecciones federales del año 2000 en los tres eventos electorales en el estado de Aguascalientes.

Eventos					
Diputados	44.03%	42.00%	7.05%	3.00%	1.21%
Eventos					
Diputados	0.25%	0.70%	1.09%	0.25%	0.42%

Tabla 2 Porcentajes de votos de los grupos políticos en las elecciones federales del año 2003 en el evento electoral de Diputados en el estado de Aguascalientes.

5.1.1 Secciones representativas

Como se comentó anteriormente en el año 2000 se tuvieron 3 eventos electorales federales, donde el número de secciones electorales fué de 486. Los partidos en la contienda se consideraron de la siguiente forma: PAN, PRI, PRD y los demás partidos. Entonces en un evento electoral pasado(2000) las proporciones de votos de los 4 grupos políticos que contendieron en la $j-ésima$ sección son $q_{j1},\ldots,q_{j4}$, donde $j=1,2,\ldots,486$. Por otra parte, las proporciones de votos de los 4 grupos políticos en todo el estado son $q_1,\ldots,q_4$(ver Tabla1). A partir de aquí se pueden construir las entropías relativas(*ver sección 3.4*) en todas las secciones

electorales, es decir, se tiene que

$$D_j = \sum_{k=1}^{4} q_k \ln\left(\frac{q_k}{q_{jk}}\right),$$

esta cantidad es la información que proporciona la $j - ésima$ sección de la distribución del voto con respecto al comportamiento en el estado y para cada uno de dichos eventos.

Se hace notar que para cada evento electoral del año 2000, las secciones generan una entropía relativa(Información), es decir, se está tomando en cuenta la información que generan las secciones en cada evento electoral. Por ejemplo, una sección puede ser representativa solamente en el evento electoral de Presidente pero no es representativa en los otros eventos electorales. Entonces una sección será más representativa si su distribución en el voto es similar en todos los eventos electorales pasados. Por ejemplo, en el estado de Aguascalientes se tuvieron las siguientes seccciones electorales, las cuales fueron representativas en algunos o en los tres eventos electorales del año 2000, lo cual se ejemplifica en la siguiente tabla.

Sección	Presidente	Diputados	Senadores	Valor del primer componente
180	R(0.000311678)	R(0.005848474)	R(0.002262204)	0.00495
259	R(0.000646625)	R(0.003037039)	R(0.001514973)	0.00304
104	R(0.000746755)	R(0.000124874)	R(0.00144777)	0.00130
39	R(0.000922329)	R(0.003985781)	NR(0.010515635)	0.00869
270	R(0.001508589)	R(0.000612657)	R(0.000216275)	0.00136
459	R(0.001964198)	R(0.002565982)	R(0.002362055)	0.00398
92	NR(0.285473894)	NR(0.270397476)	NR(0.223785639)	0.45135
119	NR(0.450580614)	NR(0.449146692)	NR(0.411858012)	0.75794
29	NR(0.419869825)	NR(0.380587027)	NR(0.544938544)	0.77132
70	NR(0.344297695)	NR(0.346226469)	NR(0.320632366)	0.58422

Tabla 3 Secciones representativas y no representativas en los tres eventos electorales, asociados con su entropia relativa y el valor del primer componente

De la Tabla 3 se observa que las secciones 180, 259, 104, 270 y 459 son representativas en los tres eventos electorales del año 2000 asociadas con su entropía relativa y el valor del primer componente principal; sin embargo, la sección 39 es representativa en los eventos electorales de Presidente y Diputados. Por otra parte, las secciones 92, 119, 29 y 70 no son representativas en ninguno de los tres eventos electorales y sus valores del primer componente principal están alejados de cero.

En la siguiente sección veremos una forma de combinar toda la información de las secciones electorales en los tres eventos electorales para tener una medida combinada de información de la sección para un evento electoral a futuro.

5.1.2 Combinación de las entropías relativas

Para cada evento electoral pasado del año 2000 se tienen las informaciones(Entropía Relativa) en cada sección de la población bajo estudio. Ahora, con base en las medidas de entropía, considérese una combinación lineal de ellas para medir la representatividad de cada sección. Se usa la metodología de componentes principales como se describió en la sección 2.8.

Ahora se analiza un ejemplo, en el estado de *Aguascalientes* para el año 2000 se tuvieron tres eventos electorales a nivel federal, Presidentes, Diputados y Senadores, los datos fueron tomados en la página de internet del Instituto Federal Electoral(*http://www.ife.org.mx*).

Basado en la información de datos de los tres eventos electorales, y aplicando la metodología de Componentes Principales, se encontró que la primera componente principal capta el 95% de la variabilidad de los datos originales, entonces consideramos suficiente trabajar con el primer componente principal para un ánalisis subsecuente.

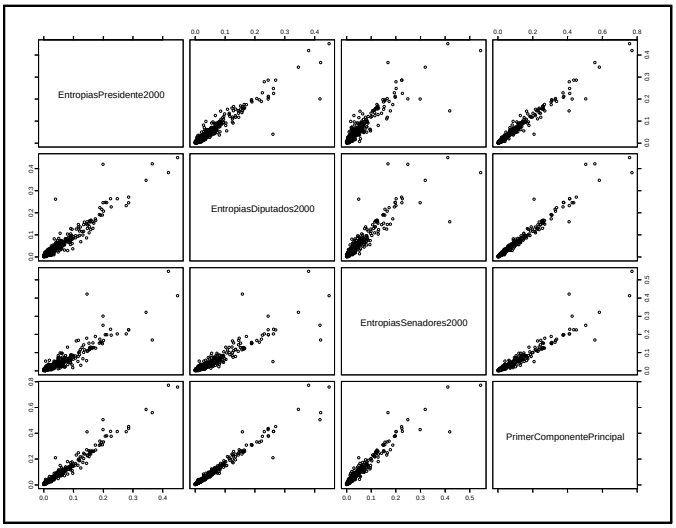


Figura 1.- Gráficas de dispersión entre las entropias relativa y el primer componente principal

Como se observa en la Figura 1, hay una relación aproximadamente lineal entre las entropías relativas de los tres eventos electorales con su primer componente principal. En particular, entre más representativa sea la sección electoral más pequeño es el correspondiente valor del primer componente principal.

En la Tabla 4, tenemos las correlaciones entre las entropías de los eventos electorales del año 2000, obsérvese que las correlaciones son positivas y cercanas a uno, lo cual nos indica que hay una alta correlación positiva entre las 3 variables y nos asegura usar solamente el primer componente principal como nuestra medida de representatividad para cada sección electoral.

Eventos	EntropiasPresidente2000	EntropiasDiputados2000	EntropiasSenadores2000
EntropiasPresidente2000	1	0.9513010863	0.9107620848
EntropiasDiputados2000	0.9513010863	1	0.9064263241
EntropiasSenadores2000	0.9107620848	0.9064263241	1

Tabla 4 Matriz de correlación de las entropias relativas de los tres eventos electorales del año 2000 en el estado de Aguascalientes

Ahora, en la Tabla 5 se muestran los autovalores(varianza de los componentes) y se puede ver que el primer componente acumula aproximadamente el 95% de la varianza total, como se menciono anteriormente. El primer autovector, que es el coeficiente del primer componente principal nos da una idea de los pesos que está asignando a cada variable(Presidente, Diputados y Senadores), ya que las entropías del evento electoral diputados tiene un coeficiente de 0.605, y los coeficientes de las entropías de los otros eventos electorales es menor.

Autovalores	0.01016401889	0.000361027638	0.0001795559186
Proporción de Varianza Acumulada	0.9494998942	0.98322628800	1.00000000000
Primer Autovector	0.580	0.605	0.546
Segundo Autovector	-0.325	-0.443	0.836
Tercer Autovector	0.747	-0.662	0

Tabla 5

De la Figura 2, se observa que el primer componente principal está proporcionando información de las tres variables(Entropías Relativas), las secciones representativas tienen un valor cercano a cero del primer componente.

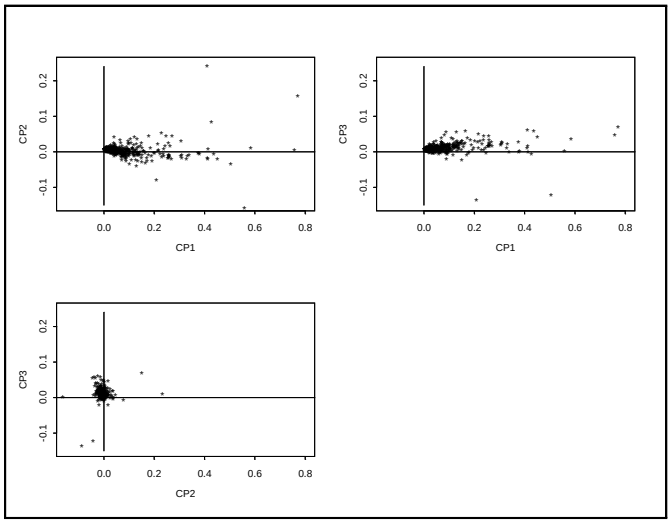


Figura 2.- Gráficas de dispersión de los tres componentes principales en el estado de Aguascalientes

5.1.3 Selección de muestras y estimaciones

Con base en el primer componente principal se propone una selección de muestras de secciones, es decir, cada sección electoral tiene su correspondiente valor del primer componente principal, como se mencionó anteriormente en este capítulo, un valor del primer componente principal cercano a cero para una sección electoral dice que la correspondiente sección es representativa en la distribución del voto del estado, entonces al momento de seleccionar secciones en la muestra estarán con mayor posibilidad las secciones más representativas. Tengáse presente que se están tomando en cuenta todas las secciones en la población, ya que las secciones pueden ser representativas en algunos eventos electorales.

En este caso las secciones fueron seleccionadas usando *SR,* donde aquí las secciones tienen la misma probabilidad de ser seleccionadas en la muestra. Se utilizó un tamaño de muestra $n = 49$, es decir, aproximadamente el 10% del tamaño de la población $N = 486$, se toma como referencia, ya que se considera que para un tamaño de muestra mayor las estimaciones son más precisas y para un tamaño de muestra menor las estimaciones son menos precisas.

Entonces a partir de las secciones seleccionadas en la muestra y con base en el evento electoral de *Diputados* del año *2003* en el estado de Aguacalientes, se pueden estimar las proporciones de votos para los partidos que contienden en el evento electoral de Diputados 2003, el cual es de la siguiente manera:

$$\widehat{q}_{jSR} = \frac{\sum_{k \in U} y_{kj} I_k}{\sum_{k \in U} x_k I_k},$$

donde y_{kj} es el total de votos para el partido j en la $k - ésima$ sección y x_k es el total de votos en la $k - ésima$ sección. Para una muestra s seleccionada, por SR se tiene que

$$\widehat{q}_{jSR}(s) = \frac{\sum_{k \in s} y_{kj}}{\sum_{k \in s} x_k}.$$

Como son conocidas las proporciones de votos para los partidos en el evento electoral de Diputados 2003, se pueden calcular las entropías relativas de la siguiente manera. Supongáse que las proporciones de votos de los 4 grupos políticos que contienden en la secciones son $\widehat{q}_1, \ldots, \widehat{q}_4$, las cuales son las estimaciones que se obtuvieron anteriormente bajo SR. Por otra parte, las proporciones de votos de los mismos 4 grupos políticos que contienden en todo el estado son $p_1, \ldots, p_4$. A partir de aquí se pueden construir las entropías relativas para cada muestra seleccionada s bajo SR, es decir, se tiene que

$$D_{SR}(s) = \sum_{k=1}^{4} p_k \ln\left(\frac{p_k}{\widehat{q}_{kSR}}\right). \tag{5.1}$$

Lo anterior se calcula para ver si la muestra de secciones seleccionada s bajo SR reproduce la distribución de votos de Diputados 2003, por medio del porcentaje de votos estimados para los partidos que contienden en el evento diputados 2003.

Aquí no se está tomando en cuenta la información sobre secciones representativas, es decir, las secciones tienen la misma probabilidad de estar en la muestra seleccionada, pero

en el siguiente diseño de muestreo, las secciones representativas tienen mayor probabilidad de estar en la muestra seleccionada.

Para el diseño *SRST*, primero se ordenan las secciones de acuerdo al primer componente principal. La ordenación se hizó en forma ascendente, en donde los primeros valores nos caracterizan a las secciones más representativas y los últimos nos caracterizan a las secciones menos representativas. A partir de la ordenación de las secciones con su correspondiente valor del componente principal se divide a esta población en *3 estratos* iguales, de tal manera que en el primer estrato estén las secciones más representativas, en el segundo estrato las secciones de mediana representatividad y en el tercer estrato las secciones menos representativas. El tamaño de la población de las secciones es $N = 486$, de tal manera que los tamaños de los estratos son $N_1 = 162$, $N_2 = 162$, y $N_3 = 162$. Para cada estrato se hace selección de secciones usando *SR*. Como se quiere que en las muestras de secciones estén las más representativas, entonces se proponen los siguientes tamaños de muestras para los estratos, $n_1 = 25$, $n_2 = 16$ y $n_3 = 8$, en una relación $1 : 2 : 3$, está relación es una forma fácil y práctica de asignar tamaños de muestras en los diferentes estratos de acuerdo a la representatividad de las secciones. Como se puede notar, el tamaño de muestra en el estrato de las secciones representativas es mayor a los otros dos estratos y en el estrato de las secciones menos representativas el tamaño de muestra es pequeño. Lo anterior se debe a que en la selección de muestras de secciones, se está dando mayor prioridad a las secciones más representativas y menos prioridad a las secciones menos representativas.

En la Figura 3 del histograma se observa que los valores del primer componente están acumulados cerca de cero y se indican los puntos de corte de tal manera que se formen las particiones y se tengan los 3 estratos de secciones electorales.

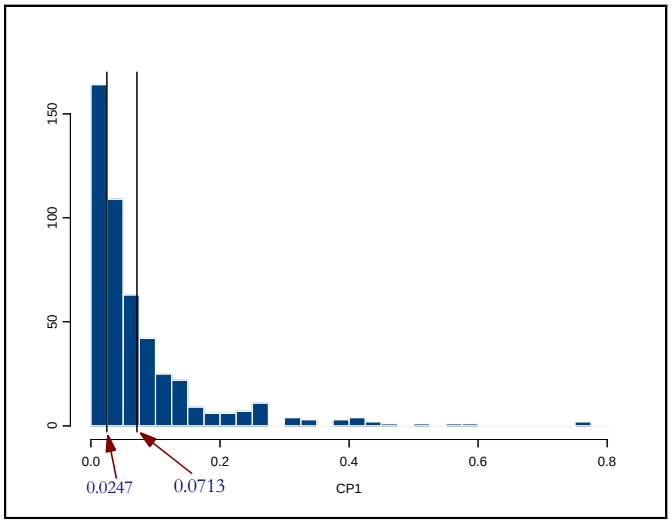


Figura 3.- Histograma de los valores del primer componente

A partir de las secciones seleccionadas en la muestra y con base en el evento electoral de *Diputados* del año *2003* en el estado de Aguascalientes, se pueden estimar las proporciones de votos para los los partidos que contienden en el evento electoral de Diputados 2003, el cual es de la siguiente manera:

$$\widehat{q}_{jSTSR} = \frac{1}{n_1 + n_2 + n_3} \sum_{h=1}^{3} \left(\sum_{k \in U_h} \frac{y_{hkj}}{x_{hk}} I_{hk} \right),$$

donde y_{hkj} es el total de votos para el partido j en la $k - ésima$ sección para el estrato h y x_{hk} es el total de votos en la $k-ésima$ sección en el estrato h, es decir, $x_{hk} = y_{hk1} + \ldots + y_{hk4}$. Ahora, para una muestra s seleccionada, por $STSR$ se tiene que

$$\widehat{q}_{jSTSR}(s) = \frac{1}{n_1 + n_2 + n_3} \sum_{h=1}^{3} \left(\sum_{k \in U_h} \frac{y_{hkj}}{x_{hk}} \right).$$

Ahora, de igual manera que en el diseño SR se puede construir la entropía relativa de la distribución del voto estimado con respecto a la distribución del estado del evento diputados del 2003. Para este diseño, $SRST$, se está incorporando la información de las secciones

representativas.

La entropía relativa tiene la forma siguiente:

$$D_{STSR}(s) = \sum_{j=1}^{4} q_j \ln\left(\frac{q_j}{\widehat{q}_{jSTSR}}\right), \tag{5.2}$$

donde las proporciones de votos de los 4 grupos políticos que contienden en la secciones son $\widehat{q}_1, \ldots, \widehat{q}_4$, las cuales son las estimaciones que se obtuvieron anteriormente bajo *STSR*. Por otra parte, las proporciones de votos de los mismos 4 grupos políticos que contienden en todo el estado son $q_1, \ldots, q_4$.

Ahora veánse los resultados encontrados para el estado de Aguascalientes. Se hace la comparación de las entropías relativas de las ecuaciones 5.1 y 5.2 bajo los diseños de muestreo utilizados, *SR* y *STSR*, para el evento electoral de *Diputados 2003*. Como se observa en la Figura 4, la cual es una gráfica cuantil-cuantil, las entropías relativas en el caso *STSR*(eje horizontal), son más pequeñas en comparación con el diseño *SR*(eje vertical), que era de esperar ya que en el diseño *STSR* está incorporando la información de las secciones representativas, lo cuál no se hace en el diseño *SR*.

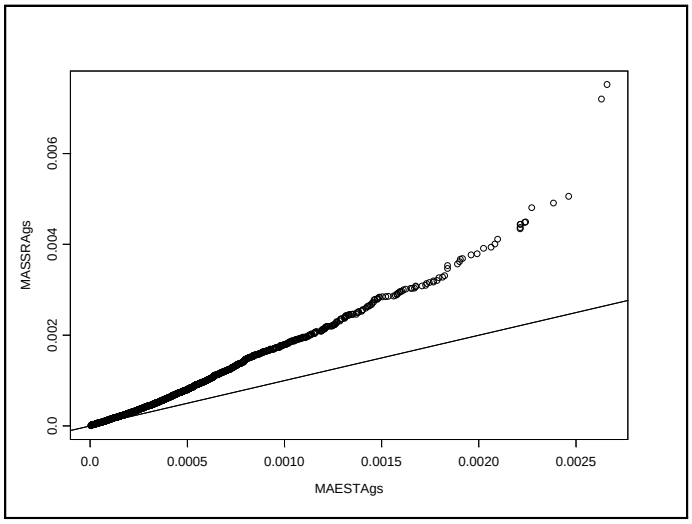


Figura 4.- Entropías Relativas bajo SR y STSR con base en el evento electoral Diputados 2003

Lo anterior proporciona evidencia empírica que el diseño de muestreo y estimador propuestos, diseño y estimador que utilizan información sobre la distribución del voto en las secciones, son mejores que el muestreo aleatorio simple sin reemplazo con su respectivo estimador de razón, diseño y estimador que no utilizan información sobre la distribución del voto en cada sección.

Ahora, veamos que tan precisas son las estimaciones de las proporciones de votos para los dos grupos políticos encabezados por el PAN y PRI que obtuvieron mayor porcentaje de votos en el evento de diputados del 2003, bajo los dos diseños de muestreo mencionados anteriormente. Como se observa en el diagrama de cajas en la Figura 5, las estimaciones de las proporciones para los dos partidos son parecidas, pero hay más variabilidad en las estimaciones bajo el diseño *SR*, lo cuál quiere decir también que las estimaciones de cada proporción que se proponen bajo esta metodología son más precisas. Esto es, no sólo la estimación de la distribución del voto es más precisa sino también las estimaciones de los porcentajes del voto por separado. Lo primero nos puede llevar a dar una adecuada esti-

mación de la camara de diputados, por ejemplo, y lo segundo nos puede servir para dar una predicción adecuada del ganador de un evento electoral.

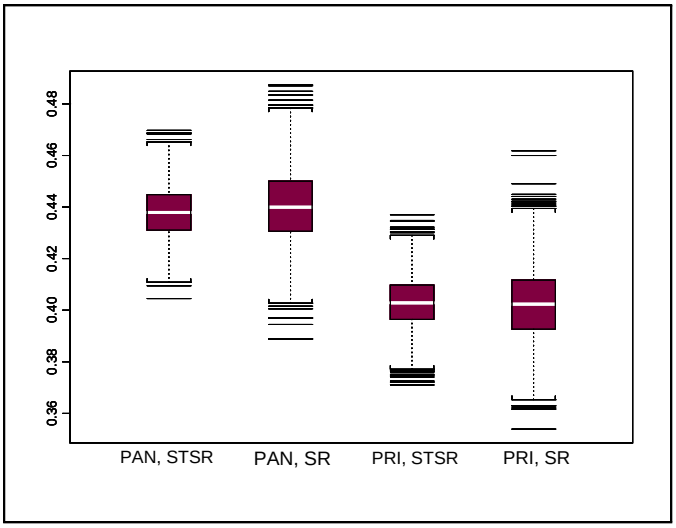


Figura 5.- Diagramas de Cajas para las proporciones de votos para el PAN y PRI bajo SR y STSR

5.1.4 Intervalos de Confianza

Ahora veremos que tan eficiente es el estimador de la varianza propuesto bajo los dos diseños de muestreo utilizados. Para ello construiremos intervalos de confianza del 95% y veremos cual es la probabilidad de cobertura para el verdadero valor del parámetro q_1, donde $q_1 = 0.4403$ es el porcentaje de votos que obtuvó el PAN para elección de Diputados Federales del 2003 en el estado de Aguascalientes.

En la Tabla 6 se muestran los resultados encontrados para el estado de Aguascalientes, referente a los intervalos de confianza.

Diseño	Cobertura	LongitudPromedioIntervalos
SR	93.80%	0.053976
STSR	93.85%	0.044167

Tabla 6

Con los dos estimadores de la varianza propuestos en la *sección 4.4* y usando un nivel de significancia, $\alpha = 0.05$, los intervalos aleatorios fueron simulados por separado para cada muestra seleccionada usando los dos diseños de muestreo, *SR*, *STSR*. Los resultados fueron los siguientes: con 2000 simulaciones, las probabilidades de cobertura fueron de 0.9380 bajo el diseño *SR* y de 0.9385 bajo el diseño *STSR*. También calculamos las longitudes promedios de los intervalos de confianzas, bajo el diseño *SR* el promedio es de 0.053976 y para el diseño *STSR* es de 0.044167.

Como se puede observar de la Tabla 6, la probabilidades de cobertura son muy cercanas a la nominal de 0.95 para los dos diseños de muestreo, *ver Thompson(1997)*. Las longitudes en promedio para el diseño *STSR* son más pequeñas en comparación con el diseño *SR*, lo que significa, que tenemos estimaciones más precisas bajo el diseño *STSR*, ya que se está incorporando la representatividad de las secciones.

Ahora, describamos el proceso de simulación utilizado. Para la entidad federativa considerada y datos por secciones electorales del evento diputados federales del 2003 y cada diseño de muestreo utilizado, el proceso de simulación se replicó 2000 veces y se utilizó el programa estadístico S-PLUS 2000. El proceso de simulación en cada ciclo:

1.− se selecciona una muestra de secciones,

2.− se calculan los porcentajes de votos para cada partido,

3.− se calcula la entropía relativa de la distribución del voto estimada(paso 2) con respecto a la distribución de voto del estado(ver Tabla 2),

4.− se calcula la varianza del porcentaje de votos del grupo político que obtuvó mayor porcentaje en el evento diputados 2003,

5.− se construye el intervalo de confianza para el porcentaje de votos del grupo político que obtuvó mayor porcentaje en el evento diputados 2003.

Por otra parte, en la Figura 6 se muestran los histogramas de los intervalos de confianza simulados bajo los dos diseños de muestreo y la recta vertical es el verdadero valor del parámetro q_1, esto nos proporciona una manera gráfica de visualizar el comportamiento de las coberturas de los intervalos de confianza simulados bajo los dos diseños de muestreo.

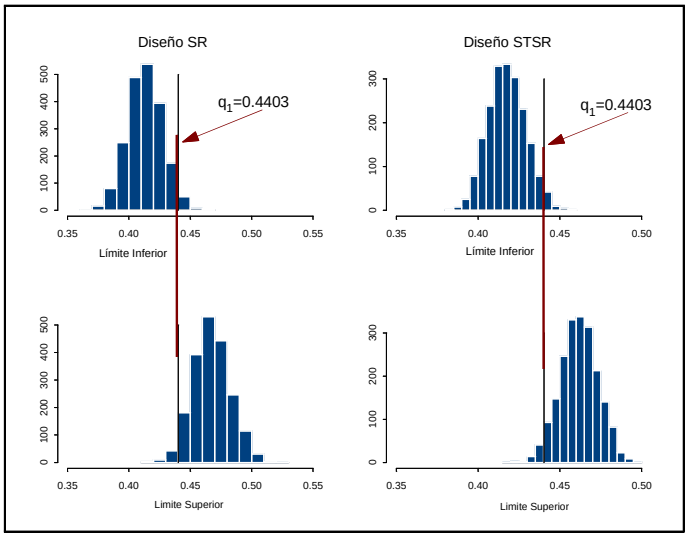


Figura 6.- Histogramas de los Intervalos de Confianza bajo los diseños SR y STSR

Ahora veamos el comportamiento de las longitudes de los intervalos de confianza bajo los dos diseños de muestreo. Para visualizar esto, se utiliza un diagrama de cajas(ver Figura 7).

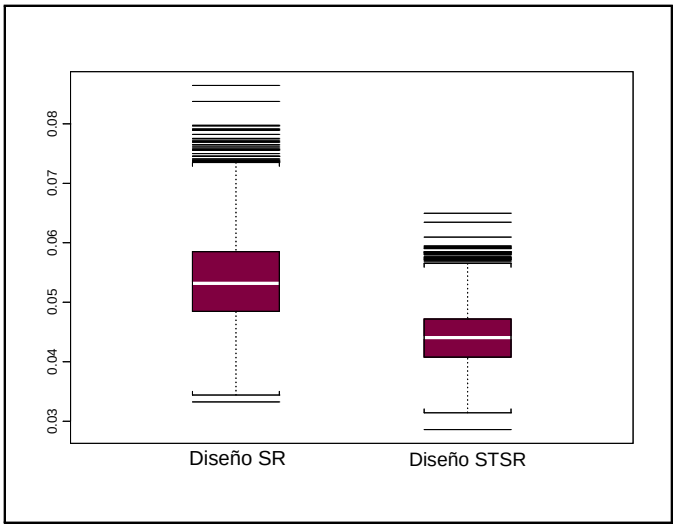


Figura 7.- Longitudes de los Intervalos de Confianza bajo SR y STSR

Como se observa en la Figura 7, hay más variabilidad en las longitudes de los intervalos de confianza bajo el diseño *SR*, lo cual no pasa bajo el diseño *STSR*, ya que los valores de las longitudes están más concentrados alrededor de la mediana. Con esto se confirma que la metodología presentada reproduce satisfactoriamente la proporción del voto para Diputados 2003 en el estado de Aguascalientes, usando el diseño *STSR*.

5.2 Caso Zacatecas

Analizaremos el estado de Zacatecas ya que la distribución en el voto del año 2000 en los tres eventos electorales no se conservó en el 2003 para el evento electoral diputados 2003; lo cuál no ocurrió para el estado de Aguascalientes, ya que no hubó alternancia en el poder.

El estado de Zacatecas se componía en el año 2000 de 1875 secciones electorales para los tres eventos federales, presidente, diputados y senadores. Para el año 2003, se conservaron las mismas secciones electorales pero solamente ocurrió el evento electoral de diputados. Los

resultados de los eventos electorales federales de los años 2000 y 2003 se resumen en las siguientes tablas.

Eventos	PAN	PRD	Alianza México (PT)	PCD	PARM	Partido Social
Presidente	34.26%	39.81%	23.68%	0.59%	0.40%	1.26%
Diputados	24.35%	39.36%	33.70%	0.91%	0.56%	1.12%
Senadores	23.71%	44.34%	29.80%	0.78%	0.53%	0.84%

Tabla 7 Porcentajes de votos de los grupos políticos en las elecciones federales del año 2000 en los tres eventos electorales en el estado de Zacatecas

Eventos	PAN	PRD	PRD	PT	Convergencia	
Diputados	12.80%	29.18%	46.81%	5.49%	3.31%	1.33%
Eventos	PSN	PAS	México Posible	PLM	Fuerza Ciudadana	
Diputados	0.09%	0.27%	0.40%	0.19%	0.13%	

Tabla 8 Porcentajes de votos de los grupos políticos en las elecciones federales del año 2003 en el evento electoral de Diputados en el estado de Zacatecas.

En las siguientes secciones, se muestran los resultados obtenidos para el estado de Zacatecas. Los resultados se obtuvieron de manera similar que para el caso Aguascalientes.

5.2.1 Secciones representativas

Por ejemplo, en el estado de Zacatecas se tuvieron las siguientes seccciones electorales, las cuales fueron representativas en algunos o en todos los tres eventos electorales del año 2000,

el cual se muestra en la siguiente tabla.

Sección	Presidente	Diputados	Senadores	Valor del primer componente
149	R(0.000599202)	R(0.001333517)	R(0.007233316)	0.00489
1151	R(0.001354051)	R(0.008199153)	R(0.000689974)	0.00617
1787	R(0.001505623)	R(0.002320344)	R(0.002516078)	0.00362
10	R(0.024931755)	NR(0.1575557)	R(0.019789428)	0.12131
2	R(0.007753743)	R(0.005045936)	R(0.003542828)	0.00957
1	R(0.010066148)	R(0.017898346)	R(0.005120836)	0.01962
52	NR(0.099436478)	R(0.062475283)	R(0.058330268)	0.12805
18	NR(0.487636843)	NR(0.535890591)	NR(0.365111156)	0.80861
19	NR(0.841343161)	NR(0.602458193)	NR(0.526033596)	1.14521
23	NR(0.969194524)	NR(0.658392965)	NR(0.724351121)	1.35836

Tabla 9 Secciones representativas y no representativas en los tres eventos electorales, asociados con su entropia relativa y el valor del primer componente

De la Tabla 9, se observa que las secciones 149, 1151, 1787, 1 y 2 fueron representativas en los tres eventos electorales del año 2000 en el estado de Zacatecas, de acuerdo al valor de su entropía relativa y el valor del primer componente; sin embargo, las secciones 18, 19 y 23 no fueron representativas en ningún evento electoral, ya que sus valores del primer componente están alejados de cero. La sección 10 fué representativa en los eventos electorales Presidente y Senadores, pero la sección 52 fué representativa en los eventos electorales Diputados y Senadores.

5.2.2 Combinación de las entropías relativas

Basado en la información de datos por secciones de los tres eventos electorales del año 2000 en el estado de Zacatecas, y aplicando la metodología de los Componentes Principales, se observa que la primera componente principal capta el 89% de la variabilidad de los datos

originales, entonces consideramos suficiente trabajar con el primer componente principal para un ánalisis posterior.

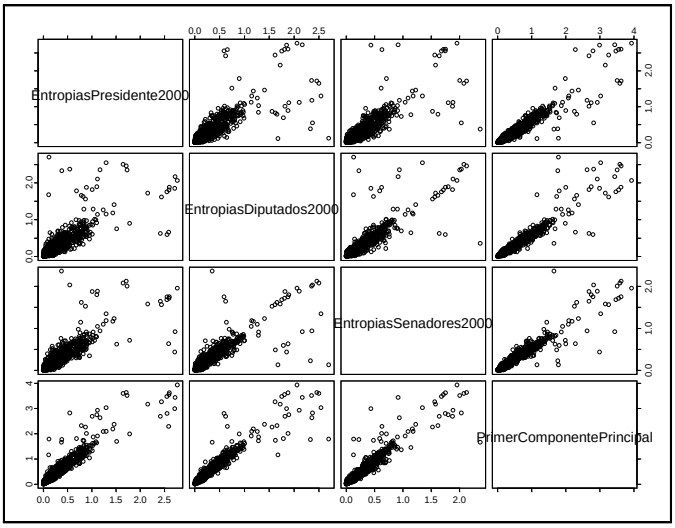


Figura 8.- Gráficas de dispersión entre las entropias y el primer componente principal

De la gráfica de dispersión, ver Figura 8, se observa que hay una relación aproximadamente lineal entre las entropías relativas de los tres eventos electorales con su primer componente principal y entre más representativa sea la sección electoral más pequeño es el correspondiente valor del primer componente principal.

Ahora, la Tabla 10 nos muestra las correlaciones entre las entropías de los eventos electorales del año 2000 en el estado de Zacatecas y nos indica que hay una alta correlación positiva entre las 3 variables y entonces podemos utilizar el primer componente como nuestra medida de representatividad..

Eventos	EntropiasPresidente2000	EntropiasDiputados2000	EntropiasSenadores2000
EntropiasPresidente2000	1	0.8074094150	0.8470846909
EntropiasDiputados2000	0.8074094150	1	0.8640134841
EntropiasSenadores2000	0.8470846909	0.8640134841	1

Tabla 10 Matriz de correlación de las entropias relativas de los tres eventos electorales del año 2000 en el estado de Zacatecas

Por otra parte, la Tabla 11 muestra los autovalores(varianzas) y se observa que el primer componente acumula aproximadamente el 89% de la varianza total, como se mencionó anteriormente. El primer autovector nos da una idea de los pesos que está asignando a cada variable, ya que las entropías del evento electoral diputados tiene un coeficiente de 0.610, y los coeficientes de las entropías de los otros eventos electorales es menor.

Autovalores	0.2044036614	0.01631777783	0.008501769404
Proporción de Varianza Acumulada	0.8917232361	0.96291052091	1.00000000000
Primer Autovector	0.603	0.610	0.514
Segundo Autovector	-0.740	0.669	0
Tercer Autovector	0.298	0.426	-0.854

Tabla 11

De igual manera que para el caso Aguascalientes, de la Figura 9, se observa que el primer componente está proporcionando información de las tres variables, las secciones representativas tienen un valor cercano a cero del primer componente.

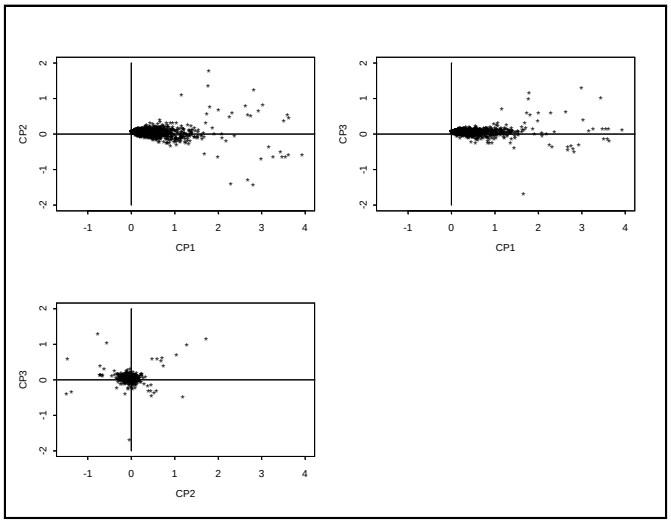


Figura 9.- Gráficas de dispersión de los tres componentes principales en el estado de Zacatecas

5.2.3 Selección de muestras y estimaciones

Para la selección de muestras de secciones usando el diseño *SR*, se utiliza un tamaño de muestra $n = 187$ de la población de tamaño $N = 1875$.

Bajo el diseño *STSR*, la población es estratificada en 3 estratos de tamaños iguales, es decir, $N_1 = 625$, $N_2 = 625$, y $N_3 = 625$. Para cada estrato se hace selección de secciones usando *SR*. Como se quiere que en las muestras de secciones estén las más representativas, entonces se proponen los siguientes tamaños de muestras para los estratos, $n_1 = 94$, $n_2 = 62$ y $n_3 = 31$.

Para el estado de Zacatecas se observa en la Figura 10, que los valores del primer componente están acumulados cerca de cero y se indican los puntos de corte de tal manera que se formen las particiones en los 3 estratos.

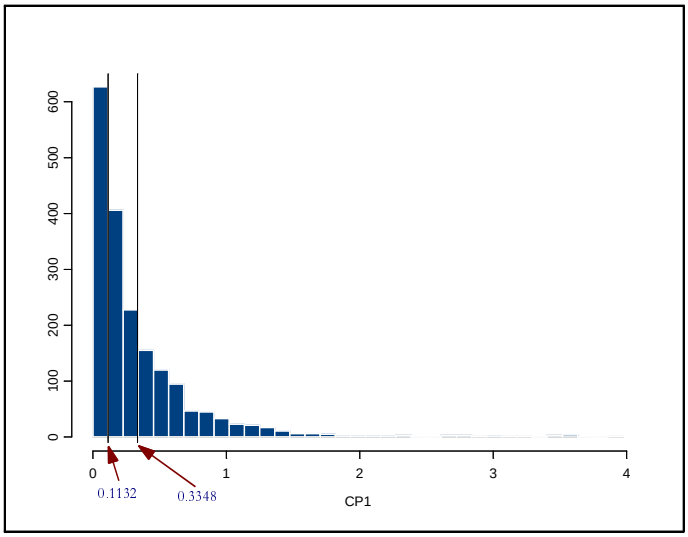


Figura 10.- Histograma de los valores del primer componente

Ahora veánse los resultados encontrados para el estado de Zacatecas. Se hace la comparación de las entropías relativas bajo los diseños de muestreos utilizados, *SR* y *STSR*, para el evento electoral de *Diputados 2003*. Como se observa en la gráfica cuantil-cuantil, ver Figura 11, las entropías relativas en el caso *STSR*(eje horizontal), son más pequeñas en comparación con el diseño *SR*(eje vertical), que era de esperar ya que en el diseño *STSR* está incorporando la información de las secciones representativas, lo cuál no se hace en el diseño *SR*.

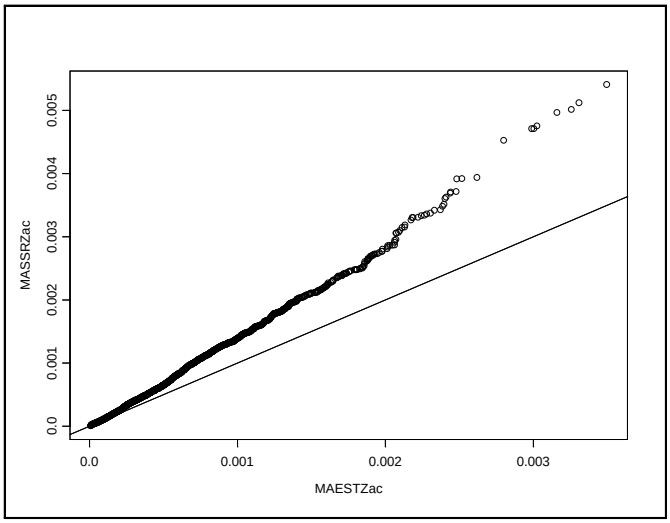


Figura 11.- Entropías Relativas bajo SR y STSR con base en el evento electoral Diputados 2003

De igual manera que para el caso Aguascalientes, lo anterior muestra que el muestreo estratificado es más adecuado para reproducir la distribución del voto, ya que utiliza la información de secciones representativas en un evento electoral pasado.

Por otra parte, veamos que tan precisas son las estimaciones de las proporciones de votos para los grupos políticos encabezados por el PRI y PRD, bajo los dos diseños de muestreo mencionados anteriormente. Como se observa en el diagrama de cajas, ver Figura 12, las estimaciones de las proporciones para los dos partidos son parecidas, pero hay más variabilidad en las estimaciones bajo el diseño *SR*, lo cual quiere decir también que las estimaciones de cada proporción que se proponen bajo esta metodología son más precisas, aún teniendo alternancia en el poder.

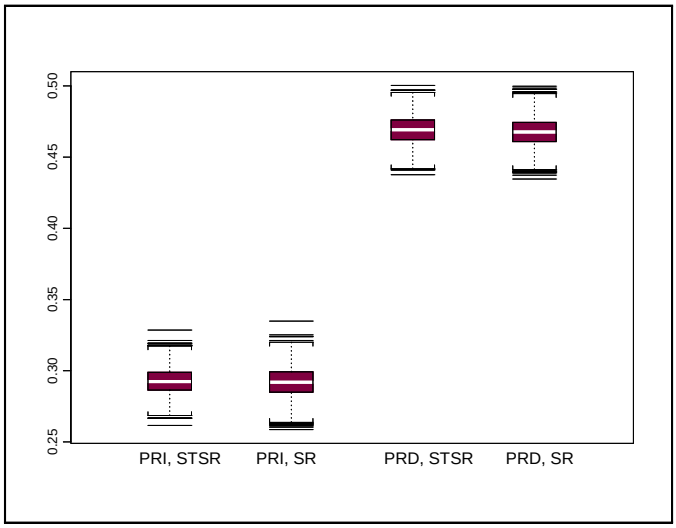


Figura 12.- Diagramas de cajas para las proporciones de votos para el PRD y PRI bajo SR y STSR.

5.2.4 Intervalos de Confianza

Ahora veremos que tan eficiente es el estimador de la varianza propuesto bajo los dos diseños de muestreo utilizados. Para ello construiremos intervalos de confianza del 95% y veremos cual es la probabilidad de cobertura para el verdadero valor del parámetro q_3, donde $q_3 = 0.4681$ es el porcentaje de votos que obtuvó el PRD para elección de Diputados Federales del 2003 en el estado de Zacatecas.

En la Tabla 12 se muestran los resultados encontrados para el estado de Zacatecas, referente a los intervalos de confianza.

Diseño	Cobertura	LongitudPromedioIntervalos
SR	94.55%	0.03899
STSR	94.85%	0.03745

Tabla 12

Con los dos estimadores de la varianza propuestos y usando un nivel de significancia, $\alpha = 0.05$, los intervalos aleatorios fueron simulados por separado para cada muestra seleccionada usando los dos diseños de muestreo, *SR*, *STSR*.

De igual manera como para el caso Aguascalientes, ver Tabla 12, la probabilidades de cobertura son muy cercanas a la nominal de 0.95 para los dos diseños de muestreo, *ver Thompson (1997)*. Las longitudes en promedio para el diseño *STSR* son más pequeñas en comparación con el diseño *SR*, lo que significa, que tenemos estimaciones más precisas bajo el diseño *STSR*, ya que se está incorporando la representatividad de las secciones e independiente de la alternancia en el poder.

Por otra parte, en la Figura 13 se muestran las gráficas de los intervalos de confianzas simulados bajo los dos diseños de muestreo y el punto que está sobre la recta vertical es el verdadero valor del parámetro q_3 y nos da una idea gráfica de las coberturas de los intervalos simulados.

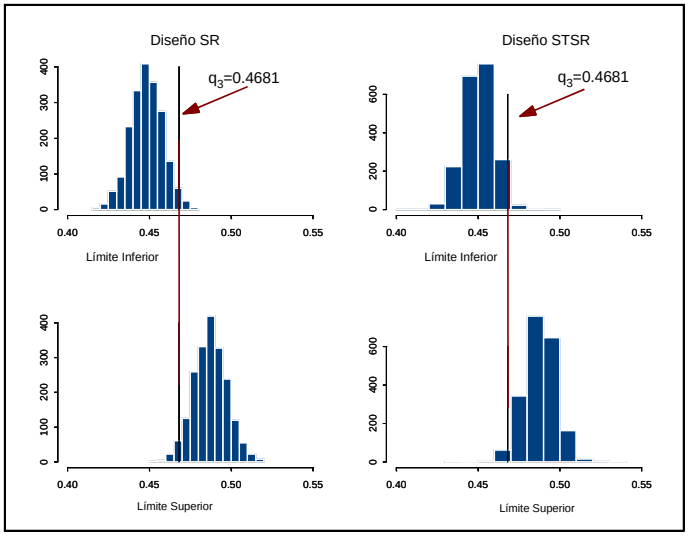


Figura 13.- Histogramas de los Intervalos de Confianza bajo los diseños SR y STSR

Finalmente, veamos el comportamiento de las longitudes de los intervalos de confianza

bajo los dos diseños de muestreo. Para visualizar esto, se utiliza un diagrama de cajas(ver Figura 14).

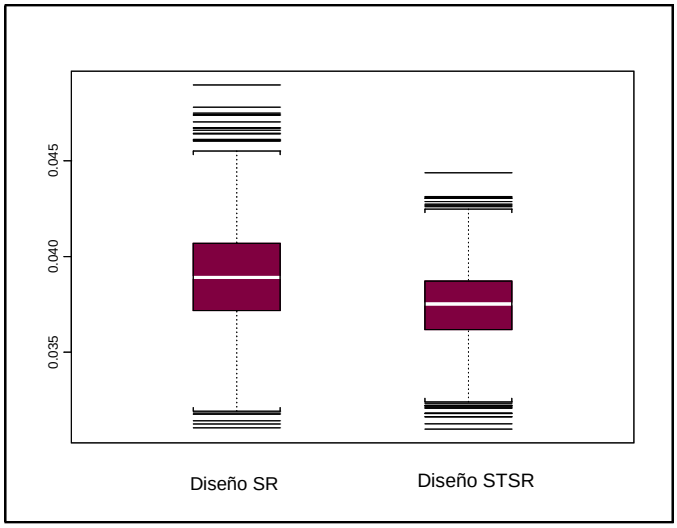


Figura 14.- Longitudes de los Intervalos de Confianza bajo SR y STSR

Como se observa en la Figura 14, hay más variabilidad en las longitudes de los intervalos de confianza bajo el diseño *SR*, lo cual no pasa bajo el diseño *STSR*, ya que los valores de las longitudes están más concentrados alrededor de la mediana. Con esto se confirma que la metodología presentada reproduce satisfactoriamente la proporción del voto para Diputados 2003 en el estado de Zacatecas, usando el diseño *STSR*, aún teniendo alternancia en el poder para el año 2003.

Capítulo 6

Comentarios finales

En este trabajo se propuso la medida de entropía relativa como medida de representatividad de una sección electoral. Además, con base en la anterior medida se combinaron las medidas de entropías relativas correspondientes a diferentes eventos electorales para producir una sóla medida de representatividad a través del uso del primer componente principal. Adicionalmente, la forma de seleccionar una muestra de secciones electorales aprovechando la anterior medida de representatividad, fue otra de las contribuciones.

Como se mencionó en el capítulo 4, si el primer componente principal no capta la mayoría de la variabilidad, proponemos separar los grupos de eventos que influyan más en cada componente y analizar cada grupo por separado. Por ejemplo, si los eventos electorales que influyen en un componente principal corresponden a elecciones locales y nos interesa estudiar una elección local futura, entonces sugerimos utilizar únicamente los resultados de dichas elecciones para definir una medida de representatividad para cada sección electoral. También sugerimos agrupar eventos que tengan correlaciones positivas entre si y valorar su utilización según el tipo de evento electoral futuro a analizar.

Adicional a la forma de seleccionar muestras, se propuso una forma de estimar los porcentajes de votos con los datos de una muestra de secciones electorales, y no se usó el estimador que comúnmente se usa, el estimador de Horvitz-Thompson, ya que sus estimaciones no son eficientes para los porcentajes estimados. Por último, también se presentó una

forma de estimar la varianza del anterior estimador.

Para las dos poblaciones analizadas con la metodología propuesta hay que tener en cuenta que una gran ventaja de la metodología que aquí se propone es que no se ve afectada por la alternancia de poderes en los eventos electorales. En el caso de Aguascalientes, el PAN tenía mayoría de votos en los tres eventos electorales del año 2000 y se conservó para el evento electoral de diputados del año 2003. En caso contrario, en el estado de Zacatecas, el PRI tenía mayoría de votos en los tres eventos eventos electorales del año 2000 y el PRD tenía mayoría de votos para el evento electoral de diputados del año 2003. Esto nos dice que la metodología planteada parece independiente de la alternancia en el poder.

Los resultados de esta tesis han sugerido otros posibles trabajos futuros de investigación. Por ejemplo, aprovechar la medida de representatividad para agrupar las secciones electorales y posteriormente se pueden estudiar las características de los votantes, en este grupo de secciones electorales. También cómo utilizar la metodología desarrollada en este trabajo para predecir el ganador de un evento electoral a futuro. Hay trabajos acerca de las predicciones de un ganador para un evento electoral futuro, por ejemplo *Bernardo (1984)* propuso una metodología bayesiana para la predicción del ganador y *Yu y Lam (1997)* propusieron una metodología utilizando conceptos de probabilidad de selección correcta y establecen un subconjunto en donde se encuentra el ganador o los ganadores de un evento electoral. En este mismo sentido se podría utilizar la metodología de selección correcta, metodología que aparecen en *Bechhofer, et.al. (1995)* y *Gibbons, et.al. (1999)*. La investigación futura en este sentido será como aprovechar la metodología aquí propuesta para predecir un ganador de un evento electoral y comparar con las propuestas antes mencionadas.

Referencias

1. APPLEBAUM, D. (1996). *Probability and Information: an integrated approach.* Cambridge University Press.

2. BERNARDO, J.M. (1984). Monitoring the 1982 Spanish Socialist victory: A Bayesian analysis. *Journal American Statistical Association.* Vol. 79, No.387, 510-515.

3. BECHHOFER, R.E., D.M. GOLDSMAN y T.J. SANTNER . (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons.* Wiley Series in Probability and Statistics.

4. BIBBY, J.M., J.T. KENT y K.V. MARDIA (1995). *Multivariate Analysis.* Academic Press.

5. CHATFIELD, C. y J. COLLINS (1980). *Introduction to Multivariate Analysis.* Chapman and Hall.

6. COVER, THOMAS M. y JOY A. THOMAS (1991). Elements of Information Theory. New York: John Wiley.

7. GIBBONS, J.D., D.M. OLKIN y T.J. SOBEL . (1999). *Selecting and Ordering Populations: A New Statistical Methodology.* SIAM.

8. INSTITUTO FEDERAL ELECTORAL. Página Web *http://www.ife.org.mx.*

9. KULLBACK, SOLOMON. (1968). *Information Theory and Statistics.* Dover Publications.

10. MANLY, BRYAN F. J. (1986). *Multivariate Statistical Methods: A primer.* Chapman and Hall.

11. SARNDAL, C.E., B. SWENSSON y J. WRETMAN (1992). *Model Assisted Survey Sampling*. Springer Verlag.

12. THOMPSON, M.E. (1997). *Theory of Sample Surveys*. Chapman and Hall.

13. YU, P.L.H. y K. LAM (1997). How to predict election winners from a poll. *Journal of Applied Statistics*. Vol. 24, No. 1, 11-23.