PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
Escuela de Posgrado
Maestría en Estadística

TÉCNICAS DE MUESTREO

Notas de clase

Luis Hilmar Valdivieso Serrano

2018

Presentación

Estas notas de clase han sido redactadas como material de apoyo para los estudiantes de la maestría en Estadística de la Pontificia Universidad Católica del Perú y ofrecen una introducción al estudio de las principales técnicas de muestreo probabilístico.

Si bien existen en la literatura varios textos clásicos de muestreo como el de Cochran (1977), Mendenhall et al. (2007) y Lohr (2000) y más avanzados como el de Tillé (2006) y Lumley (2010), falta todavía en mi humilde opinión un texto de nivel intermedio que integre estos enfoques y que a su vez incluya más aplicaciones a datos reales de dominio público. Estas notas pretenden cubrir tal vacío presentando no sólo las técnicas de muestreo probabilístico clásicos, sino también tópicos de muestreo complejo y una implementación computacional que actúe transversalmente a lo largo de los diferentes temas del curso. Para ello haremos uso principalmente de los paquetes survey y sampling escritos en el software libre R. Información sobre estos se puede consultar, respectivamente, en los enlaces

`http://cran.r-project.org/web/packages/survey/survey.pdf`

`https://cran.r-project.org/web/packages/sampling/sampling.pdf`

o en los textos de Lumley (2010) y Tillé (2006). Otra excelente referencia en el espíritu de estas notas y que incluye al paquete PracTools en R, es Valliant et al. (2013).

Las notas han sido divididas en 5 capítulos. En el primero damos una introducción a algunos conceptos básicos de Estadística, poniendo énfasis en la diferencia que existe entre los enfoques basados en el modelo y en el diseño. El capítulo 2 presenta la teoría del muestreo aleatorio simple (MAS) introduciéndose aquí no sólo los conceptos teóricos pertinentes sino también su implementación computacional y aplicación a datos reales. En el tercer capítulo definimos el muestreo aleatorio estratificado como el agregado de un MAS aplicado a subconjuntos relativamente homogéneos de la población a los cuales denominaremos estratos. El capítulo cuatro aborda el muestreo por conglomerados, el cual es quizás el esquema clásico más utilizado para grandes poblaciones. A diferencia del diseño anterior, este esquema resulta ser más eficiente cuando los subconjuntos de la población (que denominaremos conglomerados) muestran una marcada heterogeneidad en su interior pero gran similitud entre ellos. Un tema central y unificador en este capítulo será el estudio de los estimadores de Horvitz-Thompson para totales en diseños de conglomerados de una o

más etapas con probabilidades de selección no siempre constantes. De este se derivan casi todos los esquemas anteriores, como por citar el de conglomerados de una etapa y su caso particular el muestreo sistemático. El último capítulo está dedicado al estudio de muestras complejas. Estas se originan cuando debido a la configuración y tamaño de la población en estudio se hace necesario el restringir o combinar dos o más técnicas ya sea que cada selección se haga con igual probabilidad o no. Aquí nos interesará no sólo obtener estimaciones puntuales de los parámetros de interés al expandir apropiadamente la muestra a la población, sino fundamentalmente estimar la variabilidad de las estimaciones. Para ello discutiremos diversas técnicas como la linealización y el remuestreo y nos apoyaremos, al igual que en los capítulos anteriores, en los paquetes survey y sampling de R. Este capítulo bridará también una introducción al análisis estadístico bajo muestras complejas. Como ilustración veremos aquí el análisis de datos categóricos, el de regresión y los contrastes de hipótesis para una, dos o más poblaciones. El capítulo incluye algunos diseños muestrales y su correspondiente análisis para las bases de datos introducidas en el curso.

El texto se complementa con diversos ejercicios propuestos y algunas sugerencias y/o respuestas a estos en un anexo final. Tales ejercicios son de nivel teórico y práctico y hacen uso, muchos de ellos, de bases de datos de dominio público tanto locales como foráneas.

Dr. Luis Valdivieso

Índice general

Capítulo 1

Introducción

1.1. Enfoques basados en el diseño y el modelo

Supongamos que una entidad bancaria este interesada en conocer el ahorro medio mensual de las familias de un distrito. Sea y la variable (estadística) que asigna a cada familia del distrito su respectivo ahorro mensual en soles. Naturalmente, si se realiza un censo en esta población en el cual se pregunte y averigue (con fortuna) los ahorros de las N familias del distrito, uno obtendra N números $y_1, y_2, \ldots, y_N$ y consecuentemente el ahorro medio de interés:

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Desafortunadamente, como es común, al banco no le es factible hacer el censo por lo que alternativamente planifica realizar un muestreo probabilístico seleccionando al azar, y por simplicidad con reemplazamiento, una por una a las familias del padrón de la municipalidad hasta un número $n < N$. Note que según este esquema una familia cualesquiera tiene una probabilidad de ser escogida de $\frac{n}{N}$. Al término del estudio el Banco obtendrá la muestra

$$Y_1, Y_2, \ldots, Y_n, \tag{1.1}$$

donde Y_i denota al valor (aleatorio) que podría tomar la variable estadística y en la i-ésima selección de la muestra. Realizada las observaciones, el ahorro medio mensual de las familias del distrito podrá estimarse mediante:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{j_i},$$

siendo y_{j_i} el valor observado de la variable aleatoria Y_i, y $j_i \in \{1, 2, \ldots, N\}$. Note aquí, que la aleatoriedad es introducida por el esquema de selección en el diseño de la muestra. Así, podríamos escribir indistintamente a la variable aleatoria correspondiente a la estimación

anterior como

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{ó} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i\delta_i, \tag{1.2}$$

siendo δ_i una variable aleatoria con distribución binomial de parámetros n y probabilidad $\frac{1}{N}$ que denota al número de veces o frecuencia en que la $i-$ésima familia en el distrito es seleccionada en la muestra.

Estadísticamente (1.2) es un buen estimador de μ_N. Por citar, su valor esperado o media

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{N} y_i E(\delta_i) = \frac{1}{n}\sum_{i=1}^{N} y_i \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu_N$$

es precisamente el parámetro que buscamos; es decir, $\bar{Y}$ es un estimador insesgado de μ_N.

El enfoque hasta aquí utilizado se denomina un enfoque basado en el diseño. Un lector perspicaz podría preguntarse, porque este difiere al clásico de inferencia en el cual uno puede simplemente asumir una distribución o superpoblación para el ahorro Y de las familias del distrito, digamos normal con media μ y varianza σ^2, y por tanto estimar μ (que es la cantidad que el Banco quiere) al tomarse una muestra aleatoria $Y_1, Y_2, \ldots, Y_n$ de Y y considerarse el estimador

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

La respuesta a esta interrogante no es tan directa. El enfoque clásico comentado, llamado enfoque basado en el modelo, difiere al del diseño en el sentido que los parámetros poblacionales μ y μ_N son por naturaleza distintos a menos que la población sea infinita y el modelo este bien especificado. En efecto, uno puede integrar ambos enfoques pensando que si la población fuese hipotéticamente grande ($N \to \infty$), entonces la distribución empírica de los números $y_1, y_2, \ldots, y_N$ (piense por simplicidad en el polígono de frecuencias relativas del histograma de estos datos) debería de converger (si el modelo es correcto) a la curva normal. Luego podríamos pensar en la colección dada por (1.1) como una muestra aleatoria de la variable aleatoria Y.

Observe que en un modelo basado en el diseño, a diferencia que en de su par basado en el modelo, la distribución de Y es irrelevante a menos, como precisamos, uno tenga interés y tenga sentido el analizar cuestiones asintóticas. Desde un punto de vista práctico el enfoque basado en el diseño nos será útil para estudiar poblaciones finitas; mientras que el enfoque basado en el modelo será útil para el estudio de poblaciones infinitas o muy grandes.

1.2. Estimadores puntuales y por intervalos

Al margen del enfoque o diseño muestral utilizado, existen tres características primordiales que uno debe de tomar en cuenta en un estudio inferencial. Ellos son, el tamaño de la

muestra que se utilizará, el nivel de confianza y el error de estimación. Todos estos conceptos están íntimamente ligados a la teoría de la estimación puntual y por intervalos, puntos que revisaremos brevemente antes de presentar los principales tipos de muestreo probabilísticos.

Sea X una variable aleatoria (v.a) cuya distribución depende de un parámetro poblacional desconocido θ. A esto lo denotaremos por $X \sim \theta$. Dada una muestra aleatoria (m.a) de tamaño n de X; vale decir, una colección $X_1, X_2, \ldots, X_n$ de n v.a's independientes y con la misma distribución que X, estaremos interesados en obtener un estimador $\hat{\theta}_n = g(X_1, X_2, \ldots, X_n)$ de θ. Por definición, este estimador puede ser cualquier estadística (función de la m.a), pero es claro que nos interesarán estimadores buenos en el sentido que de observarse la muestra, podamos garantizar que el valor observado de $\hat{\theta}_n$, al que llamaremos una estimación, se ubique cerca a θ. Dado que no conocemos θ, esta cercanía debe evaluarse a través de métodos probabilísticos. En general un buen estimador, $\hat{\theta}_n$ de θ, debe de verificar en lo posible las siguientes tres propiedades básicas:

- $\hat{\theta}_n$ debe de ser un estimador insesgado; i.e, $E(\hat{\theta}_n) = \theta$.
- $\hat{\theta}_n$ debe de ser eficiente; i.e, debe de tener varianza pequeña, por lo usual mínima bajo una clase de estimadores insesgados.
- $\hat{\theta}_n$ debe de ser consistente; i.e, $\hat{\theta}_n \xrightarrow{P} \theta$, conforme $n \to \infty$.

El problema básico con la estimación puntual, es que esta no nos provee de un indicador de que cuan cerca o lejos esté la estimación $\hat{\theta}_n$ de θ. Por tal motivo, surge la llamada estimación por intervalos.

Un intervalo de confianza (IC) al $100(1-\alpha)\,\%$ para un parámetro poblacional θ de una v.a. X es un intervalo con estadísticas L_1 y L_2 en sus extremos ($IC = [L_1, L_2]$) tal que

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

Una técnica para obtener IC's es utilizar alguna estadística o variable pivote que tenga distribución conocida y que dependa tan solo de θ como valor desconocido. Por ejemplo, si deseamos estimar la media de una v.a. $X \sim N(\mu, \sigma^2)$ con varianza conocida; podríamos utilizar como variable pivote la estadística

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Tomando luego en la distribución normal estándar dos puntos cuyas áreas en las colas sean iguales a $\frac{\alpha}{2}$ (¿porqué?), obtendremos el siguiente intervalo de confianza al $100(1-\alpha)\,\%$ para μ:

$$IC = [\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \ .$$

Vale destacar que gracias al teorema del límite central (TLC) este IC es aún válido para la media de cualquier distribución, siempre que n sea lo suficientemente grande y se tenga, de no conocerse σ, una estimación de esta desviación estándar.

Otro parámetro recurrente en diversas aplicaciones lo constituye la proporción p de elementos en la población que comparten cierta característica. A fin de obtener un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para p, tomemos al azar n elementos de la población física y consideremos las v.a's X_i definidas como 1 si es que en la i-ésima selección se encuentra un elemento con la característica buscada y 0 en caso contrario. Note que los elementos de esta muestra sólo podrán garantizarse distintos, si es que la muestra es tomada sin reemplazamiento. Esto ocasiona que las variables $X_1, \ldots, X_n$ no sean independientes; sin embargo, si el tamaño N de la población es grande o infinito, se podría garantizar una casi independencia. En la práctica si N es grande estas variables son consideradas independientes, por lo que la distribución de $X = \sum_{i=1}^{n} X_i$, que representa al número de elementos en la muestra que comparten la característica buscada, puede asumirse que tiene aproximadamente una distribución binomial de parámetros n y p. Más aún, si n es grande, podremos utilizar la aproximación de la distribución binomial por la normal y usar:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1),$$

con $\bar{p} = \frac{X}{n}$, como variable pivote para la construcción del IC para p. En efecto, tomando simétricamente valores $-z_{1-\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ en la tabla normal estándar, podemos afirmar que:

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

Con el fin de despejar p en esta expresión, podemos considerar la probabilidad equivalente:

$$P(|\frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}|^2 \leq z^2_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

ó

$$P(p^2(1 + \frac{z^2_{1-\frac{\alpha}{2}}}{n}) - p(2\bar{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{n}) + \bar{p}^2 \leq 0) = 1 - \alpha.$$

Esta probabilidad, puede escribirse como:

$$P((p - p_1)(p - p_2) \leq 0) = 1 - \alpha,$$

donde p_1 y p_2 constituyen las raíces de la ecuación cuadrática asociada a la inecuación anterior. Si ahora en la fórmula del discriminante de la ecuación cuadrática despreciamos al término $\frac{z^2_{1-\frac{\alpha}{2}}}{n}$, por ser pequeño al ser n grande, obtendremos el $IC = [p_1, p_2]$ al $100(1-\alpha)\,\%$ para p siguiente:

$$IC = [\bar{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\ ,\ \bar{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}].$$

Este es conocido como el intervalo de Wald para p, a diferencia del primero (sin la aproximación) llamado de Wilson.

1.3. Distribuciones importantes asociadas al muestreo

Aparte de la muy conocida distribución normal, requeriremos en el curso las formas tanto univariadas como multivariadas de las distribuciones binomial e hipergeométrica. Estas las asociaremos luego al contexto de los muestreos con y sin reemplazamiento, respectivamente.

1.3.1. La distribución binomial

Consideremos un experimento aleatorio sencillo, llamado de Bernoulli, que tiene solo dos posibles resultados: E (de éxito) y F (de fracaso). Sea $p = P(E)$ la probabilidad de que ocurra un éxito. Si repetimos este experimento n veces de manera independiente y definimos la variable aleatoria

$$X = \text{ Número de éxitos en los n experimentos independientes de Bernoulli,}$$

entonces diremos que X es una v.a. con distribución binomial de parámetros n y p y la denotaremos por $X \sim B(n,p)$.

Si $X \sim B(n,p)$, su función de probabilidad esta dada por:

$$P_X(x) = P(X = x) = \begin{cases} C_x^n p^x (1-p)^{n-x} & \text{, si } x = 0,1,2,...,n. \\ 0 & \text{, en otro caso.} \end{cases}$$

Es fácil probar que la media y varianza de esta v.a vienen dadas respectivamente por $E(X) = np$ y $V(X) = np(1-p)$.

1.3.2. La distribución multinomial

Esta es la extensión multivariada de la distribución anterior. Para describirla, consideremos ahora un experimento aleatorio cuyos resultados pueden caer en cualquiera de k categorías excluyentes y exhaustivas $C_1, C_2, \ldots, C_k$ con probabilidades respectivas $p_1, p_2, \ldots, p_k$ tales que $\sum_{i=1}^k p_i = 1$. Si este experimento se repite de manera independiente n veces y se definen las variables aleatorias:

$$X_i = \text{número de veces en que ocurre la categoría } C_i, \quad i = 1,2,\ldots,k,$$

entonces el vector aleatorio $(X_1, X_2, \ldots, X_k)$ se dice que tiene distribución multinomial de parámetros n, $p_1, p_2, \ldots, p_k$, y se le denota por $(X_1, X_2, \ldots, X_k) \sim Mul(n, p_1, p_2, \ldots, p_k)$. La función de probabilidad (conjunta) de este vector viene dada por:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} & \text{, si } (x_1, x_2, \ldots, x_k) \in R \\ 0 & \text{, en caso contrario} \end{cases}$$

donde $R = \{(n_1, n_2, \ldots, n_k) \in \{0, 1, \ldots, n\}^k \ / \ \sum_{i=1}^k n_i = n\}$ denota rango del vector. Así, es posible deducir las siguientes propiedades de esta distribución:

Proposición 1.1 *a)* $X_i \sim B(n, p_i), \quad \forall i = 1, 2, \ldots, k.$

b) $Cov(X_i, X_j) = -np_ip_j, \quad \forall i \neq j \in \{1, 2, \ldots, k\}.$

Más aún se tienen los siguientes intervalos de confianza asintóticos para esta distribución

Proposición 1.2 *a) Si n es grande, un intervalo de confianza al* $100(1-\alpha)\,\%$ *para* p_i *es:*

$$[\bar{p}_i - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n}}, \bar{p}_i + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n}}] \ ,$$

donde $\bar{p}_i$ *denota la proporción de veces que ocurre la categoría* C_i *en los n experimentos.*

b) Si n es grande, un intervalo de confianza al $100(1-\alpha)\,\%$ *para* $p_i - p_j$*, con* $i \neq j$*, es:*

$$[\bar{p}_i - \bar{p}_j - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i) + \bar{p}_j(1-\bar{p}_j) + 2\bar{p}_i\bar{p}_j}{n}}, \ \bar{p}_i - \bar{p}_j + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i) + \bar{p}_j(1-\bar{p}_j) + 2\bar{p}_i\bar{p}_j}{n}}] \ ,$$

donde $\bar{p}_i$ *y* $\bar{p}_j$ *denotan, respectivamente, a la proporción de veces que ocurre la categoría* C_i *y la categoría* C_j *en los n experimentos.*

Observe que el IC último difiere del clásico para la diferencia entre dos proporciones en poblaciones independientes por la presencia del término $2\bar{p}_i\bar{p}_j$, el cual no aparece en el IC clásico.

Vale comentar que las variables aleatorias δ_i definidas en (1.2), que denotan al número de veces en que el $i-$ésimo elemento de la población física de tamaño N es seleccionado en una muestra al azar y con reemplazamiento de tamaño n, son todas v.a's con distribución $B(n, \frac{1}{N})$. De manera más general, si se tuviera interés, por ejemplo, en las frecuencias de selección de los elementos $i \neq j$ de la población, entonces no es difícil verificar que

$$(\delta_i, \delta_j, \delta_0) \sim Mul(n, \frac{1}{N}, \frac{1}{N}, 1 - \frac{2}{N}),$$

donde δ_0 denota a la frecuencia de selecciones de otras unidades distintas a i y j. Note que estas v.a's no son independientes, desde que por ejemplo:

$$P(\delta_j = y \mid \delta_i = x) = \frac{P(\delta_i = x, \delta_j = y, \delta_0 = n - x - y)}{P(\delta_i = x)} = C_y^{n-x}(\frac{1}{N-1})^y(1 - \frac{1}{N-1})^{n-y}$$

$$\neq C_y^n(\frac{1}{N})^y(1 - \frac{1}{N})^{n-y} = P(\delta_j = y), \ \forall x, y \in \{0, 1, \ldots, n\} \text{ con } x + y \leq n.$$

1.3.3. La distribución hipergeométrica

Considérese una población de N elementos, M de los cuales son de tipo A, y supongamos se extraen al azar y sin reemplazamiento una muestra de n elementos de esta población. Si definimos

$$X = \text{ Número de elementos de tipo } A \text{ en la muestra,}$$

entonces se dice que X es una v.a. con distribución hipergeométrica de parámetros N,M y n y se le denota por $X \sim H(N, M, n)$. La función de probabilidad de esta variable viene dada por:

$$P_X(x) = P(X = x) = \begin{cases} \frac{C_x^M C_{n-x}^{N-M}}{C_n^N} & \text{, si } x = 0, 1, 2, ..., n. \\ 0 & \text{, en otro caso.} \end{cases},$$

donde se hace la convención que $\binom{b}{a} = 0$, si $a > b$. Se comprueba que la media y varianza de esta distribución vienen dadas respectivamente por

$$E(X) = n\frac{M}{N} \quad \text{y} \quad V(X) = n\frac{M}{N}(1 - \frac{M}{N})(\frac{N-n}{N-1}).$$

1.3.4. La distribución hipergeométrica multivariada

Esta es la extensión multivariada de la distribución anterior. Aquí en lugar de estar la población de tamaño N dividida en sólo dos clases (A y A^c), ella se particiona en k clases a las que denotaremos por $C_1, C_2, \ldots, C_k$. Cada clase C_i posee M_i elementos de tal manera que $N = M_1 + M_2 + \ldots + M_k$. Si seleccionamos ahora al azar y sin reemplazamiento n elementos de esta población y definimos las variables aleatorias

$$X_i = \text{número de elementos de la clase } C_i \text{ seleccionados en la muestra, } \quad i = 1, 2, \ldots, k,$$

entonces el vector aleatorio $(X_1, X_2, \ldots, X_k)$ se dice que tiene distribución Hipergeométrica multivariada de parámetros n, $M_1, M_2, \ldots, M_k$, y se le denota por $(X_1, X_2, \ldots, X_k) \sim Hmul(n, M_1, M_2, \ldots, M_k)$. La función de probabilidad (conjunta) de este vector viene dada por:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{C_{x_1}^{M_1} C_{x_2}^{M_2} \ldots C_{x_k}^{M_k}}{C_n^N},$$

donde algunas de las combinatorias $\binom{b}{a} = 0$ arriba son nulas si $a > b$. Esta distribución tiene las siguientes propiedades

Proposición 1.3 *a)* $X_i \sim H(N, M_i, n), \quad \forall i = 1, 2, \ldots, k.$

b) $Cov(X_i, X_j) = -\frac{nM_iM_j}{N^2}(\frac{N-n}{N-1}), \quad \forall i \neq j \in \{1, 2, \ldots, k\}.$

c) Si la muestra fuera con reemplazamiento, entonces

$$(X_1, X_2, \ldots, X_k) \sim Mul(n, \frac{M_1}{N}, \frac{M_2}{N}, \ldots, \frac{M_k}{N}).$$

Note finalmente que las v.a's δ_i discutidas en (1.2) tienen una naturaleza completamente distinta si la muestra es tomada sin reemplazamiento. En efecto, si esta fuera la situación y se tuviera interés en la selección por decir de las unidades $i \neq j$ de la población física, entonces para la distribución conjunta del vector $(\delta_i, \delta_j, \delta_0)$, que denota respectivamente a las frecuencias de selección de las unidades i, j u otras en la muestra, se cumpliría que

$$(\delta_i, \delta_j, \delta_0) \sim Hmul(n, 1, 1, N-2).$$

Aprecie que las v.a's δ_i y δ_j de este vector están ahora restringidas a tomar sólo dos valores (0 ó 1) y no son independientes desde que por citar:

$$P(\delta_j = 1 \mid \delta_i = 1) = \frac{P(\delta_i = 1, \delta_j = 1, \delta_0 = n-2)}{P(\delta_i = 1)} = \frac{n-1}{N-1} \neq \frac{n}{N} = P(\delta_j = 1),$$

ya que marginalmente $\delta_j \sim H(N, 1, n)$.

1.4. Esperanza y varianza condicionada

Antes de culminar este capítulo introductorio, será de gran utilidad recordar una propiedad bastante recurrente en diversas aplicaciones. Ella está referida al cálculo indirecto de la media y varianza de una v.a. Y mediante el condicionamiento de esta variable a un vector o variable aleatoria $\mathbf{X}$.

Proposición 1.4 *Si Y es una v.a. con varianza finita y $\mathbf{X}$ un vector aleatorio, entonces*

$$E(Y) = E(E(Y|\mathbf{X}))$$

y

$$V(Y) = E(V(Y|\mathbf{X})) + V(E(Y|\mathbf{X})).$$

Ejemplo 1.1 *La producción diaria de una fábrica, que es de 200 artículos, contiene 12 artículos con un defecto de tipo A y 8 artículos con un defecto de tipo B. Todos estos artículos son empacados para venta distribuyéndolos al azar en 4 cajas de 50 artículos cada uno. Suponga que usted adquiere 20 artículos, que son seleccionados al azar y sin reemplazamiento de la primera de estas cajas, y sabe que cada artículo bueno le reportará de venderlo una utilidad de 25 soles; mientras que cada artículo con defectos de tipo A y B le reportará una pérdida de 5 y 10 soles, respectivamente.*
a) ¿ Con qué probabilidad usted obtendrá una utilidad de 500 soles de lograr vender los 20 artículos?
b) Halle el valor esperado y la desviación estándar de la utilidad que usted obtendrá de vender los 20 artículos.

Solución: a) La utilidad pedida se obtendrá si todos los artículos adquiridos son buenos y ello entonces dependerá de saber cuantos artículos buenos, con defectos de tipo A y defectos de tipo B contiene la primera caja de fábrica. Si denotamos por $\mathbf{X} = (X_1, X_2, X_3)$ al vector aleatorio que realiza este último conteo es claro apreciar que $\mathbf{X} \sim HMul(50, 180, 12, 8)$, pues distribuir la producción en las cuatro cajas equivaldrá a seleccionar secuencialmente al azar y sin reemplazamiento 50 unidades de la producción para colocarlas en las 4 cajas. Sea, por otro lado, $\mathbf{Y} = (Y_1, Y_2, Y_3)$ el vector aleatorio que nos indica el número de artículos buenos, con defectos de tipo A y defectos de tipo B, respectivamente, que nos tocarán al adquirir los 20 artículos de la primera caja. De conocerse el valor de $\mathbf{X}$, la distribución de este vector sería también hipergeométrica multivariada y en particular la distribución de Y_1 hipergeométrica. Aquí se nos pide $P(Y_1 = 20)$ la cual se podría obtener en base al teorema de probabilidad total como

$$P(Y_1 = 20) = \sum_{x=30}^{50} P(Y_1 = 20 \mid X_1 = x)P(X_1 = x) = \sum_{x=30}^{50} \frac{C_{20}^{x} C_{0}^{50-x}}{C_{20}^{50}} \times \frac{C_{x}^{180} C_{50-x}^{20}}{C_{50}^{200}}$$

En R esto nos da

```
x = 30:50
aux = choose(x,20)*choose(180,x)*choose(20,50-x)/(choose(50,20)*choose(200,50))
sum(aux)

## [1] 0.108542
```

b) La utilidad que se obtendrá por vender los 20 artículos viene dada por

$$U = 25Y_1 - 5Y_2 - 10Y_2.$$

Dada la dependencia de esta a la distribución de defectos en la primera caja; más explícitamente, dado que

$$\mathbf{Y}|\mathbf{X} \sim HMul(20, X_1, X_2, X_3),$$

será conveniente hacer uso de la proposición 1.4. Por tanto,

$$E(U) = E(E(U \mid \mathbf{X})) = E(25E(Y_1|\mathbf{X}) - 5E(Y_2|\mathbf{X}) - 10E(Y_3|\mathbf{X})),$$

donde $Y_i|\mathbf{X} \sim H(50, X_i, 20)$. Así, la utilidad esperada en soles será de

$$E(U) = E(25\frac{20X_1}{50} - 5\frac{20X_2}{50} - 10\frac{20X_3}{50}) = E(10X_1 - 2X_2 - 4X_3) = 436.$$

Para aplicar la propiedad 1.4 en el cálculo de la varianza notemos, por lo previamente desarrollado, que $E(U \mid \mathbf{X}) = 10X_1 - 2X_2 - 4X_3$. Así

$$V(E(U \mid \mathbf{X})) = 100V(X_1)+4V(X_2)+16V(X_3)-40Cov(X_1, X_2)-80Cov(X_1, X_3)+16Cov(X_2, X_3)$$

$= 100(3.39196)+4(2.125628)+16(1.447236)+40(2.035176)+80(1.356784)-16(0.09045226) = 559.3568.$

De otro lado, la varianza condicional de U dado $\mathbf{X}$ es

$$V(U \mid \mathbf{X}) = 625V(Y_1 \mid \mathbf{X}) + 25V(Y_2 \mid \mathbf{X}) + 100V(Y_3 \mid \mathbf{X})$$

$$-250Cov(Y_1, Y_2 \mid \mathbf{X}) - 500Cov(Y_1, Y_3 \mid \mathbf{X}) + 100Cov(Y_2, Y_3 \mid \mathbf{X})$$

$$= 625 \times 20\frac{X_1}{50}(1 - \frac{X_1}{50})\frac{30}{49} + 25 \times 20\frac{X_2}{50}(1 - \frac{X_2}{50})\frac{30}{49} + 100 \times 20\frac{X_3}{50}(1 - \frac{X_3}{50})\frac{30}{49}$$

$$+250 \times 20\frac{X_1X_2}{50^2}\frac{30}{49} + 500 \times 20\frac{X_1X_3}{50^2}\frac{30}{49} - 100 \times 20\frac{X_2X_3}{50^2}\frac{30}{49}$$

$$= \frac{6}{1,225}(31,250X_1-625X_1^2+1,250X_2-25X_2^2+5,000X_3-100X_3^2+250X_1X_2+500X_1X_3-100X_2X_3)$$

y su esperanza viene dada por

$$E(V(U \mid \mathbf{X})) = \frac{6}{1,225}(31,250(45) - 625(2,028.392) + 1,250(3) - 25(11.12563) + 5,000(2)$$

$$-100(5.447236) + 250(132.9648) + 500(88.64322) - 100(5.909548)) = 1,118.713$$

Por tanto la varianza de U es $V(U) = 1,118.713 + 559.3568 = 1,678.07$ y la desviación estándar de las utilidades será de 40.96425 soles.

□

Capítulo 2

Muestreo aleatorio simple

En un muestreo aleatorio simple (MAS) toda muestra de tamaño n tiene la misma probabilidad de ser seleccionada, lo cual corresponde teóricamente a la noción de muestra aleatoria dada en la sección anterior si la población es infinita. En la práctica las poblaciones son finitas, digamos con N elementos. Aquí veremos como tomar en cuenta este hecho y nos interesará encontrar tamaños de muestra y errores de estimación para tres de los parámetros más frecuentemente referidos en un estudio inferencial, la media poblacional μ, el total poblacional τ y la proporción de elementos p de la población que comparten alguna característica particular. Para ser más precisos enfatizaremos sobre todo el primero y último de estos parámetros, pues el análisis para el total poblacional

$$\tau = N\mu \quad \text{ó} \quad \tau = Np$$

es directamente deducible de los de μ y p.

2.1. Muestreo con y sin reemplazamiento

Existen dos esquemas de muestreo aleatorio simple importantes: el muestreo aleatorio simple con reemplazamiento, que lo denotaremos en adelante por MASc, y el muestreo aleatorio simple sin reemplazamiento, que lo denotaremos en adelante por MASs. Con base en un enfoque basado en el diseño, consideremos primero la siguiente población física $\mathcal{P}$ de tamaño N a cuyos elementos los identificaremos por simplicidad con los números naturales positivos. Estos que pudieran ser sujetos, eventos, materiales, escuelas, países, etc. los llamaremos unidades.

$$\mathcal{P} = \{1, 2, \ldots, N\}.$$

Sobre estas unidades mediremos una variable estadística y para generar la población estadística $\mathcal{P}_y$ constituida por todos los valores de y en $\mathcal{P}$; es decir,

$$\mathcal{P}_y = \{y_1, y_2, \ldots, y_N\},$$

siendo y_i el valor de y para la unidad i. Note que algunos de estos valores pueden repetirse, cosa que no ocurre en $\mathcal{P}$. Sea $n < N$ el tamaño de muestra a seleccionarse.

En un esquema MASc, las unidades se seleccionan al azar una a una de la población, con la peculiaridad de que estos son repuestos o reemplazados en cada etapa de selección. Así, una unidad cualesquiera $j \in \mathcal{P}$ podría ser elegida en más de una oportunidad. De otro lado, en el esquema MASs las unidades seleccionadas no se reponen y por tanto una unidad cualesquiera $j \in \mathcal{P}$ podría ser elegida en a lo más una oportunidad. Note en este caso que seleccionar las unidades una a una hasta completar la muestra equivale a seleccionar toda la muestra de una sola vez. La ventaja del diseño MASc es que las variables aleatorias definidas en (1.1) y asociadas a los valores de y en las unidades seleccionados, son variables independientes. En efecto, esto se sigue desde que para cualquier par de selecciones $j < k$ y cualquier par de elementos $y_p, y_q \in \mathcal{P}_y$ de la población estadística:

$$P(Y_j = y_p, Y_k = y_q) = P(Y_k = y_q \mid Y_j = y_p)P(Y_j = y_p) = P(Y_k = y_q)P(Y_j = y_p).$$

En un MASs, de otro lado, lo anterior no siempre se cumple, ya que por ejemplo

$$P(Y_2 = y_q \mid Y_1 = y_p) = \frac{1}{N-1} \neq \frac{1}{N} = P(Y_2 = y_q)$$

en el que caso que los elementos de la población estadística sean todos distintos.

Si bien la falta de independencia en un MASs puede acarrear problemas técnicos, este es en la práctica el esquema más utilizado al garantizar siempre selecciones distintas en $\mathcal{P}$.

Enfaticemos ahora el estudio y propiedades de dos de los estimadores más recurrentes en el muestreo, la media y varianza muestrales

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i\delta_i \quad \text{y} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2\delta_i,$$

donde recordemos que δ_i es una variable aleatoria que cuenta el número de veces que la unidad i de $\mathcal{P}$ es seleccionada en la muestra.

Tanto en el MASc como en el MASs, estas estadísticas constituyen los estimadores naturales de la media poblacional

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

y varianza poblacional

$$\sigma_N^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_N)^2 \quad \text{ó} \quad \sigma_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_N)^2.$$

De aquí en adelante convendremos en denotar a las variables aleatorias con letras mayúsculas (con la excepción de los δ_i) y con letras minúsculas a las no aleatorias.

Antes de analizar algunas propiedades de los estimadores $\bar{Y}$ y S^2, es útil recordar que el vector aleatorio de frecuencias de conteo por unidad de la muestra $(\delta_1, \delta_2, \ldots, \delta_N)$ tiene una distribución multinomial o hipergeométrica multivariada, dependiendo de si el esquema es un MASc o un MASs, respectivamente. Más aún, por lo visto en (1.2) tanto la media como la varianza muestral podrían escribirse alternativamente como:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

y

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

donde $Y_1, Y_2, \ldots, Y_n$ denotan a los valores que secuencialmente la variable estadística en estudio y podría tomar en cada selección de la muestra. La proposición siguiente nos brinda algunas propiedades de estas últimas variables aleatorias.

Proposición 2.1

a) En un MASc, las v.a's $Y_1, Y_2, \ldots, Y_n$ son independientes e idénticamente distribuidas con media $E(Y_1) = \mu_N$ y varianza $V(Y_1) = \sigma_N^2$.

b) En un MASs, las v.a's $Y_1, Y_2, \ldots, Y_n$ son idénticamente distribuidas con media $E(Y_1) = \mu_N$, varianza $V(Y_1) = \sigma_N^2$ y se cumple que $Cov(Y_i, Y_j) = -\frac{1}{N}\sigma_{N-1}^2$, $\forall i \neq j$.

Demostración: Supongamos, sin pérdida de generalidad, que todos los elementos en $\mathcal{P}_y$ son distintos.
a) La independencia ya fue vista antes. Que las v.a's $Y_1, Y_2, \ldots, Y_n$ tengan la misma distribución de media μ_N y varianza σ_N^2 es por otro lado consecuencia directa de que la distribución de cualesquiera de estas variables, digamos Y_i, viene definida por la función de probabilidad

$$P_{Y_i}(y) = P(Y_i = y) = \begin{cases} \frac{1}{N} & , \text{ si } y = y_1, y_2, y_3, \ldots, y_N \\ 0 & , \text{ en otro caso.} \end{cases} \tag{2.1}$$

b) Claramente, como la selección es secuencial, Y_1 tiene la distribución (2.1). Más aún, condicionando y trabajando inductivamente se puede probar que la distribución de cualesquiera de las variables $Y_1, Y_2, \ldots, Y_n$, digamos Y_i tiene la función de probabilidad dada en (2.1). Por citar, para cualquier $j \in \mathcal{P}$:

$$P(Y_2 = y_j) = \sum_{i=1}^{N} P(Y_2 = y_j \mid Y_1 = y_i)P(Y_1 = y_i)$$

$$= \sum_{\substack{i=1 \\ i\neq j}}^{N} P(Y_2 = y_j \mid Y_1 = y_i)\frac{1}{N} = \sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{1}{N-1}\frac{1}{N} = \frac{1}{N}.$$

Otra manera de ver lo anterior y que nos servirá también para las otras afirmaciones es notando que la distribución conjunta del vector $(Y_1, Y_2, \ldots, Y_n)$ viene dada por:

$$P(Y_1 = y_{j1}, Y_2 = y_{j2}, \ldots, Y_n = y_{jn})$$

$$= P(Y_n = y_{jn} \mid Y_1 = y_{j1}, \ldots, Y_{n-1} = y_{j(n-1)}) \ldots P(Y_2 = y_{j2} \mid Y_1 = y_{j1}) P(Y_1 = y_{j1})$$

$$= \frac{1}{N-n+1} \times \frac{1}{N-n+2} \times \ldots \times \frac{1}{N-1} \times \frac{1}{N}$$

cualesquiera sea $k \in \{1, 2 \ldots, n\}$ e $y_{jk} \in \mathcal{P}_y$. De esta distribución conjunta se pueden hallar distintas marginales, como la de la v.a Y_i, la cual se obtiene sumando esta sobre todos los valores de las demás variables. Estas sumas contienen $(N-1)(N-2)\ldots(N-n+1)$ términos, por lo que su resultado nos dará $\frac{1}{N}$, que es precisamente la misma distribución que en el caso MASc. Por tal razón las Y_i's tienen la misma media y varianzas anteriores. Podemos también, por otro lado, hallar la distribución conjunta del vector (Y_i, Y_j) con $i \neq j$. Esta viene dada por la suma de la distribución conjunta sobre todos los valores de las demás $n-2$ variables que no contengan a los valores donde se evalúan Y_i e Y_j. Estas sumas, como no es difícil ver, contienen $(N-2)(N-3)\ldots(N-n+1)$ términos, de aquí que se tenga que

$$P(Y_i = y_p, Y_j = y_q) = \frac{(N-2)(N-3)\ldots(N-n+1)}{(N-n+1)(N-n+2)\ldots(N-1)N} = \frac{1}{N(N-1)}, \forall p \neq q \in \mathcal{P}.$$

Consecuentemente,

$$Cov(Y_i, Y_j) = E((Y_i - \mu_N)(Y_j - \mu_N)) = \sum_{p=1}^{N}\sum_{q=1}^{N}(y_p - \mu_N)(y_q - \mu_N)P(Y_i = y_p, Y_j = y_q)$$

$$= \sum_{\substack{p=1 \\ q \neq p}}^{N}\sum_{q=1}^{N}(y_p - \mu_N)(y_q - \mu_N)\frac{1}{N(N-1)} = \frac{1}{N(N-1)}\sum_{p=1}^{N}(y_p - \mu_N)(\sum_{q=1}^{N}(y_q - \mu_N) - (y_p - \mu_N))$$

$$= \frac{1}{N(N-1)}((\sum_{p=1}^{N}(y_p - \mu_N))^2 - \sum_{p=1}^{N}(y_p - \mu_N)^2) = -\frac{1}{N}\sigma^2_{N-1}.$$

□

Ejemplo 2.1 *Considere una población de sujetos* $\mathcal{P} = \{1, 2, 3, 4, 5, 6, 7\}$ *y la población estadística* $\mathcal{P}_y = \{12, 32, 18, 37, 22, 18, 28\}$ *asociada a sus edades en años y. Suponga ahora que se toma un MAS con* $n = 3$. *Halle la distribución muestral de la media y varianza para esta muestra y verifique que estos son estimadores insesgados. Realice esto para los dos esquemas de muestreo estudiados.*

Solución: La media y varianza poblacionales de y vienen dadas por:

$$\mu_7 = 23.9,\ \sigma_6^2 = 78.1 \ \text{ y } \ \sigma_7^2 = 68.4$$

En un MASc tenemos un total de $7^3 = 343$ muestras posibles mientras que un MASs un total de $C_3^7 = 35$. Nosotros desarrollaremos aquí el caso de un MASs, quedando como ejercicio para el lector el otro esquema. Como ayuda utilizaremos el paquete combinat de R. Dado que en este problema precisamos obtener la distribución muestral de la media y varianza muestrales apelaremos al uso del comando combn obteniendo para cada posible muestra tanto su media, varianza y probabilidad de selección. El código respectivo se muestra seguidamente y los resultados se resumen en los Cuadros 2.1, 2.2 y 2.3.

```
library(combinat)
options(digits=3)
ypop = c(12, 32, 18, 37, 22, 18, 28)
samplesMASs = t(as.matrix(combn(ypop,3)))
ybar = apply(samplesMASs,1,mean)
s2 = apply(samplesMASs,1,var)
probs = rep(1/length(ybar), length(ybar))
bsamplesMASs = cbind(samplesMASs,ybar,s2,probs)
pp1 = aggregate(bsamplesMASs[,6],by = list(bsamplesMASs[,4]),sum)
colnames(pp1) = c("Media muestral","Probabilidad")
pp2 = aggregate(bsamplesMASs[,6],by = list(bsamplesMASs[,5]),sum)
colnames(pp2) = c("Varianza muestral","Probabilidad")
```

Vale comentar que si la muestra fuese con reemplazamiento, podríamos encontrar los índices de todas las posibles muestras con el comando expand.grid(rep(list(1:7), 3)).

De acuerdo a las tablas mostradas, los valores esperados de la media y varianza muestrales vendrán dados respectivamente por

```
c(sum(pp1[,1]*pp1[,2]),sum(pp2[,1]*pp2[,2]))

## [1] 23.9 78.1
```

mientras que la varianza de la media muestral es

```
sum(((pp1[,1] - sum(pp1[,1]*pp1[,2]))^2)*pp1[,2])

## [1] 14.9
```

Esto nos indica que la media muestral $\bar{Y}$ es efectivamente un estimador insesgado de μ_7; mientras que la varianza muestral S^2 es un estimador insesgado de σ_6^2.

Muestra				Mediam	Varm	Probs	Muestra				Mediam	Varm	Probs
1	12	32	18	20.7	105.3	0.0286	19	32	18	28	26	52	0.0286
2	12	32	37	27	175	0.0286	20	32	37	22	30.3	58.3	0.0286
3	12	32	22	22	100	0.0286	21	32	37	18	29	97	0.0286
4	12	32	18	20.7	105.3	0.0286	22	32	37	28	32.3	20.3	0.0286
5	12	32	28	24	112	0.0286	23	32	22	18	24	52	0.0286
6	12	18	37	22.3	170.3	0.0286	24	32	22	28	27.3	25.3	0.0286
7	12	18	22	17.3	25.3	0.0286	25	32	18	28	26	52	0.0286
8	12	18	18	16	12	0.0286	26	18	37	22	25.7	100.3	0.0286
9	12	18	28	19.3	65.3	0.0286	27	18	37	18	24.3	120.3	0.0286
10	12	37	22	23.7	158.3	0.0286	28	18	37	28	27.7	90.3	0.0286
11	12	37	18	22.3	170.3	0.0286	29	18	22	18	19.3	5.3	0.0286
12	12	37	28	25.7	160.3	0.0286	30	18	22	28	22.7	25.3	0.0286
13	12	22	18	17.3	25.3	0.0286	31	18	18	28	21.3	33.3	0.0286
14	12	22	28	20.7	65.3	0.0286	32	37	22	18	25.7	100.3	0.0286
15	12	18	28	19.3	65.3	0.0286	33	37	22	28	29	57	0.0286
16	32	18	37	29	97	0.0286	34	37	18	28	27.7	90.3	0.0286
17	32	18	22	24	52	0.0286	35	22	18	28	22.7	25.3	0.0286
18	32	18	18	22.7	65.3	0.0286							

Cuadro 2.1: Probabilidades, medias y varianzas de todas las posibles muestras en un MASs para el ejemplo 2.1.

□

Como el ejemplo anterior lo sugiere tenemos las siguientes propiedades en un MAS.

Proposición 2.2 *La media muestral $\bar{Y}$ es un estimador insesgado de la media poblacional μ_N y se tiene que:*

a) $V(\bar{Y}) = \frac{\sigma_N^2}{n}$ *en un MASc.*

b) $V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n}$ *en un MASs.*

La demostración de la proposición anterior es directa y puede deducirse de la demostración del siguiente resultado de suma importancia.

Proposición 2.3

a) La media muestral es el MELI (mejor estimador lineal e insesgado) de la media poblacional.

b) La varianza muestral es un estimador insesgado de σ_N^2 para un MASc y de σ_{N-1}^2 para un MASs.

Demostración: Siendo la demostración de esta proposición directa en el caso MASc, la dejaremos como ejercicio. Nosotros centraremos nuestra atención en el caso MASs.

	Media muestral	Probabilidad
1	16.000	0.029
2	17.333	0.057
3	19.333	0.086
4	20.667	0.086
5	21.333	0.029
6	22.000	0.029
7	22.333	0.057
8	22.667	0.086
9	23.667	0.029
10	24.000	0.086
11	24.333	0.029
12	25.667	0.086
13	26.000	0.057
14	27.000	0.029
15	27.333	0.029
16	27.667	0.057
17	29.000	0.086
18	30.333	0.029
19	32.333	0.029

Cuadro 2.2: Distribución de la media muestral para el ejemplo 2.1.

	Varianza muestral	Probabilidad
1	5.333	0.029
2	12.000	0.029
3	20.333	0.029
4	25.333	0.143
5	33.333	0.029
6	52.000	0.114
7	57.000	0.029
8	58.333	0.029
9	65.333	0.114
10	90.333	0.057
11	97.000	0.057
12	100.000	0.029
13	100.333	0.057
14	105.333	0.057
15	112.000	0.029
16	120.333	0.029
17	158.333	0.029
18	160.333	0.029
19	170.333	0.057
20	175.000	0.029

Cuadro 2.3: Distribución de la varianza muestral para el ejemplo 2.1.

a) Sea $\hat{\mu}_N$ un estimador lineal arbitrario de la media poblacional; es decir, un estimador de la forma $\hat{\mu}_N = \sum_{i=1}^{n} c_i Y_i$, donde las constantes c_i que la definen son arbitrarias. Para que este sea un estimador insesgado se debe de satisfacer

$$\mu_N = E(\hat{\mu}_N) = E(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i E(Y_i) = \mu_N \sum_{i=1}^{n} c_i;$$

es decir, las constantes c_i deben de sumar 1. De otro lado, la varianza de este estimador lineal viene dado por:

$$V(\hat{\mu}_N) = \sum_{i=1}^{n} c_i^2 V(Y_i) + \sum_{\substack{i=1 \\ j \neq i}}^{n} \sum_{j=1}^{n} c_i c_j Cov(Y_i, Y_j)$$

o más explícitamente, de lo visto en la proposición 2.1, por

$$V(\hat{\mu}_N) = \sigma_N^2 \sum_{i=1}^{n} c_i^2 - \frac{1}{N}\sigma_{N-1}^2 \sum_{\substack{i=1 \\ j \neq i}}^{n} \sum_{j=1}^{n} c_i c_j = \frac{N-1}{N}\sigma_{N-1}^2 \sum_{i=1}^{n} c_i^2 - \frac{1}{N}\sigma_{N-1}^2 (\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j - \sum_{i=1}^{n} c_i^2)$$

$$= \sigma_{N-1}^2 (\sum_{i=1}^{n} c_i^2 - \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j). \quad (*)$$

Por tanto el MELI de μ_N se obtendrá al hallar las constantes c_i's que resuelvan el siguiente problema de optimización

$$\min_{s.a\ \sum_{i=1}^n c_i=1} \sum_{i=1}^n c_i^2 - \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^n c_i c_j$$

Dada la convexidad de la función objetivo, bastará considerar las condiciones de primer orden del lagrangiano de esta función, el cual viene dado por:

$$l = \sum_{i=1}^n c_i^2 - \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^n c_i c_j + \lambda(1 - \sum_{i=1}^n c_i)$$

De las derivadas parciales con respecto a c_k se obtiene que:

$$0 = \frac{\partial l}{\partial c_k} = 2c_k - \frac{2}{N}\sum_{i=1}^n c_i - \lambda,$$

de donde $c_k = \frac{1}{N} + \frac{\lambda}{2}$. De la condición de insesgamiento, el multiplicador de Lagrange óptimo resulta ser $\lambda = \frac{2}{n}(1 - \frac{n}{N})$, el cual reemplazándolo en la expresión previa nos da:

$$c_k = \frac{1}{N} + \frac{1}{n}(1 - \frac{n}{N}) = \frac{1}{n}.$$

Consecuentemente el MELI de μ_N es $\bar{Y}$. Más aún, la varianza de este estimador es por (*)

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n}.$$

b) Puesto que $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1}(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2)$, se tiene que en un MASs:

$$E(S^2) = \frac{1}{n-1}(\sum_{i=1}^n E(Y_i^2) - nE(\bar{Y}^2)) = \frac{1}{n-1}(\sum_{i=1}^n (V(Y_i) + E(Y_i)^2) - n(V(\bar{Y}) + E(\bar{Y})^2))$$

$$= \frac{1}{n-1}(\sum_{i=1}^n (\sigma_N^2 + \mu_N^2) - n((1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n} + \mu_N^2))$$

$$= \frac{1}{n-1}(n(\frac{N-1}{N}\sigma_{N-1}^2 + \mu_N^2) - n((1 - \frac{n}{N})\frac{\sigma_{N-1}^2}{n} + \mu_N^2)) = \sigma_{N-1}^2.$$

□

2.2. Tamaños de muestra y errores de estimación

Los intervalos de confianza obtenidos en el capítulo anterior se basan en el clásico teorema del límite central, el cual asume una muestra aleatoria de la variable en estudio. Desafortunadamente en un MASs, que es a la larga el esquema de muestreo más utilizado, esta suposición

no es correcta dada la no independencia entre sus componentes. Para subsanar este problema tenemos aquí dos caminos que dependerán de la naturaleza del tamaño de la muestra. Cuando esta es fija y el tamaño de la población $N \to \infty$, el esquema MASs converge a un MASc. De otro lado, si $n \to \infty$ deberíamos tambien consentir que $N \to \infty$. Si denotamos por μ_N y σ^2_{N-1} a la media y varianza de las correspondientes superpoblaciones, Hajek (1960) propuso el siguiente teorema del límite central. Si $\frac{n}{N} \to \tau \in]0,1[$ y $\text{máx}_{1\leq i\leq N} \frac{Y_i-\mu_N}{\sum_{i=1}^N (Y_i-\mu_N)^2} \to 0$ conforme $n \to \infty$ y $N \to \infty$ ó $N\,\text{máx}_{1\leq i\leq N} \frac{Y_i-\mu_N}{\sum_{i=1}^N (Y_i-\mu_N)^2}$ es acotado en el límite cuando $N \to \infty$, entonces

$$Z = \frac{\bar{Y} - \mu_N}{\sqrt{1-\frac{n}{N}}\frac{\sigma_{N-1}}{\sqrt{n}}} \xrightarrow{\mathcal{D}} N(0,1).$$

conforme n y $N-n$ tiendan a infinito.

Este teorema del límite central, nos permite entonces construir, utilizando como variable pivote a la v.a Z, un intervalo de confianza aproximado al $100(1-\alpha)\,\%$ para la media poblacional μ. Este, suprimiéndose el subíndice $N-1$ para la varianza, toma para un tamaño de muestra y población suficientemente grandes, la forma:

$$IC = [\bar{Y} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\ ,\ \bar{Y} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}] = [\bar{Y} - z_{1-\frac{\alpha}{2}}SE(\bar{Y})\ ,\ \bar{Y} + z_{1-\frac{\alpha}{2}}SE(\bar{Y})\ ],$$

donde a $SE(\bar{Y})$, que es la raiz de la varianza asintótica de $\bar{Y}$, se le denomina el error estándar de estimación de $\bar{Y}$. Observe que este IC para μ difiere del clásico para poblaciones infinitas sólo por el factor $\sqrt{1-\frac{n}{N}}$. Note además que si $N \to \infty$, este factor tiende a 1 y por tanto uno obtiene el clásico IC para μ.

De manera similar, es posible realizar un estudio inferencial para poblaciones finitas con una proporción poblacional p ya que este es un caso particular de media cuando la variable Y es dicotómica. En este caso la variable pivote Z normal toma la forma:

$$Z = \frac{\bar{p} - p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}},$$

con $\bar{p}$ igual a la proporción muestral. Así, tomándose simétricamente valores $-z_{1-\frac{\alpha}{2}}$ y $z_{1-\frac{\alpha}{2}}$ en la tabla normal estándar, podemos escribir:

$$P(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{p} - p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}} \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha.$$

Con el fin de despejar p en esta expresión, podemos considerar la probabilidad equivalente:

$$P(|\frac{\bar{p} - p}{\sqrt{1-\frac{n}{N}}\sqrt{\frac{Np(1-p)}{n(N-1)}}}|^2 \leq z^2_{1-\frac{\alpha}{2}}) = 1-\alpha$$

ó

$$P(p^2(1+a) - p(2\bar{p}+a) + \bar{p}^2 \leq 0) = 1-\alpha,$$

donde $a = z^2_{1-\frac{\alpha}{2}} \frac{N-n}{n(N-1)}$. Esta probabilidad, puede escribirse como:

$$P((p-p_1)(p-p_2) \leq 0) = 1-\alpha,$$

siendo p_1 y p_2 las raíces de la ecuación asociada a la inecuación cuadrática anterior. Consecuentemente $[p_1, p_2]$ constituye un IC tipo Wilson al $100(1-\alpha)\,\%$ para p. Si ahora en el IC anterior despreciamos el término $\frac{z^2_{1-\frac{\alpha}{2}}}{n}$, por ser este pequeño cuando n es grande, obtendremos el $IC = [p_1, p_2]$ al $100(1-\alpha)\,\%$ para p tipo Wald siguiente:

$$IC = [\bar{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}} \ , \ \bar{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}}].$$

Si bien en el curso utilizaremos por simplicidad este último IC, hay que tener la precaución de que si la verdadera proporción es extrema (cercana a 0 o 1), este IC tipo Wald no presenta en general una adecuada cobertura. En tales situaciones una opción más recomendable sería usar el IC tipo Wilson. Tal problema de cobertura puede ilustrarse a través del siguiente estudio de simulación, donde hemos graficado la proporción de cuántos de los 1,000 IC's, generados a través 1,000 MASs de tamaño 30 de una población de tamaño 400, contienen al verdadero parámetro p.

```
IC<-function(x,alpha,n,N,tipo){
# tipo = 1: Wald, tipo 2 = Wilson
        pbar = x/n
        z= qnorm(1-alpha/2)
        a = (z^2)*(N-n)/(n*(N-1))
        aux = a
        if(tipo==1) aux = 0
        e = 4*a*pbar + aux^2 - 4*a*pbar^2
        L1 = (2*pbar + aux - sqrt(e))/(2*(1+aux))
        L2 = (2*pbar + aux + sqrt(e))/(2*(1+aux))
        c(L1,L2)}
# Estudio de simulación:
cover <- function(n,N,p,alpha,tipo) {
        nsim = 1000
        count = 0
        for (i in 1:nsim) {
                x = rhyper(1,N*p,N*(1-p),n)
                if(tipo==1){ci = IC(x,alpha,n,N,1)}
                else {ci = IC(x,alpha,n,N,2)}
                if(p >= ci[1] & p <= ci[2]) {count = count + 1}
        }
```

```
        cover = count/nsim
        cover}
p = seq(0.005,0.995,by=0.01)
np = length(p)
cc1 = 0
cc2 = 0
N = 400
n = 30
for(j in 1:np){cc1[j] = cover(n,N,p[j],0.05,1)}
for(j in 1:np){cc2[j] = cover(n,N,p[j],0.05,2)}
```

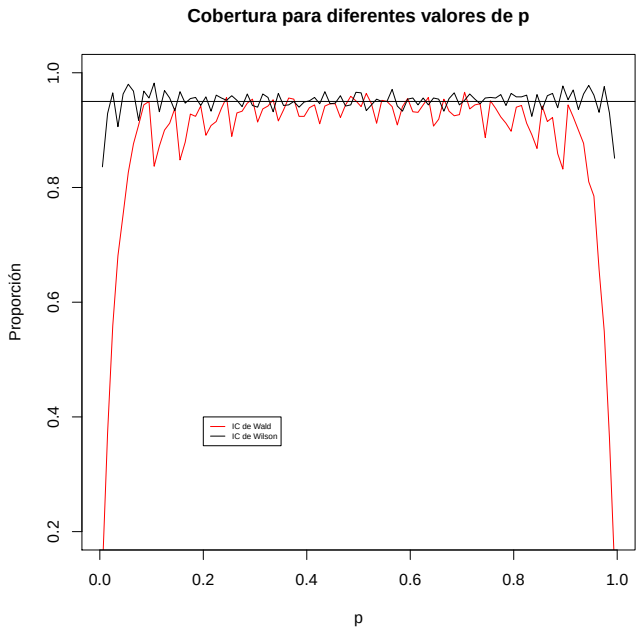


Figura I: Simulación de la cobertura de los IC de Wald y Wilson al 95 % sobre una proporción.

Establecidas las fórmulas de los IC aproximados al $100(1-\alpha)\,\%$ para cualquier media y proporción poblacional, nos interesará ahora hallar el tamaño de muestra n que uno debería de considerar para poder garantizar a un nivel de confianza del $100(1-\alpha)\,\%$ un error máximo de estimación e, donde por error de estimación entenderemos a la diferencia en valor absoluto $|\hat{\theta}_n - \theta|$ entre el parámetro y su estimador. Esto se obtiene directamente de los IC obtenidos. En efecto, si queremos estimar μ, su IC correspondiente al $100(1-\alpha)\,\%$ puede reescribirse como:

$$P(|\bar{Y} - \mu| \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}) = 1 - \alpha.$$

Luego, según lo convenido, se debe tener que:

$$e = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}},$$

de donde despejando obtenemos la siguiente fórmula para el tamaño de muestra:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \sigma^2 N}{z^2_{1-\frac{\alpha}{2}} \sigma^2 + e^2 N}.$$

Note que si $N \to \infty$:

$$n = \frac{(z_{1-\frac{\alpha}{2}} \sigma)^2}{e^2}.$$

De manera similar, podemos deducir la siguiente fórmula del tamaño de muestra n para la estimación de p con un error máximo de estimación de e y un nivel de confianza del $100(1-\alpha)\,\%$:

$$n = \frac{(z^2_{1-\frac{\alpha}{2}} \bar{p}(1-\bar{p}))N}{z^2_{1-\frac{\alpha}{2}} \bar{p}(1-\bar{p}) + e^2(N-1)}$$

y si $N \to \infty$:

$$n = \frac{z^2_{1-\frac{\alpha}{2}} \bar{p}(1-\bar{p})}{e^2}.$$

Vale agregar que la consideración de tamaños de muestra en base a los errores máximos de estimación prefijados, también llamados errores absolutos e, no es universal. En la literatura es también común encontrar la consideración del coeficiente de variación o de los errores relativos. Recordemos que el coeficiente de variación poblacional (CV) de una variable estadística y se define como el cociente entre la desviación estándar y la media de esta variable, siendo este cociente usualmente expresado en porcentajes. La adimensionalidad de este indicador, facilita claramente la determinación de valores objetivos sin que interese la escala en que uno mida la variable. Una regla práctica (a tomarla con precaución) nos dice que un estimador no es confiable si su CV supera al 30 %; contrariamente estimadores con un CV del 10 % o menos se suelen catalogar como confiables. Otra cantidad citada en el cálculo del tamaño de muestra es el error relativo, el cual se define como

$$e_r = z_{1-\frac{\alpha}{2}} CV(\hat{\theta}),$$

siendo $\hat{\theta}$ el estimador de interés para θ. Para su interpretación, basta notar que si $\hat{\theta}$ es un estimador insesgado y la muestra es suficientemente grande tendremos que aproximadamente con una confianza del $100(1-\alpha)\,\%$:

$$P(|\hat{\theta} - \theta| \leq z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\theta})}) = 1 - \alpha$$

ó

$$P(|\frac{\hat{\theta} - \theta}{\theta}| \leq z_{1-\frac{\alpha}{2}} \frac{\sqrt{V(\hat{\theta})}}{E(\hat{\theta})}) = P(|\frac{\hat{\theta} - \theta}{\theta}| \leq z_{1-\frac{\alpha}{2}} CV(\hat{\theta})) = P(|\frac{\hat{\theta} - \theta}{\theta}| \leq e_r) = 1 - \alpha.$$

Así, todas las fórmulas dadas en esta sección sobre n se satisfacen si en lugar de especificarse e, uno especificará un error relativo e_r ó un coeficiente de variación CV_0 para el estimador de interés a través de la siguiente relación

$$e = \theta e_r = \theta z_{1-\frac{\alpha}{2}} CV_0.$$

2.3. Estimaciones previas

Un aspecto problemático en las fórmulas previas lo constituyen tanto σ como $\bar{p}$, ya que el primero es en general un parámetro poblacional no conocido y el otro no puede calcularse sin la muestra. En la práctica se tienen las siguientes alternativas para solucionar este problema:

- Estimar estas cantidades mediante un muestreo piloto (es decir, con una réplica previa, pero en escala menor del muestreo final).
- Estimar estas por cantidades similares de otros estudios semejantes.
- Estimar σ por
$$\hat{\sigma} = \frac{Rango}{6},$$
donde *Rango* denota el ancho del intervalo que estimamos contenga a todos los posibles valores de la variable Y. Esto se justifica en base a la desigualdad de Chebyshev, la cual recordemos nos dice que la probabilidad de que Y se encuentre en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$, siendo μ la media de Y, es muy cercano a 1 (concretamente de al menos 0.89).
- Tomar $\bar{p} = \frac{1}{2}$. Esta es una regla conservadora, que simplemente asigna el valor de $\bar{p}$ que maximiza el tamaño de muestra de tal manera que uno pueda siempre garantizar, al margen del verdadero $\bar{p}$, un error de estimación de a lo más e.

Ejemplo 2.2 *La facultad de Ingeniería de una Universidad cuenta con 1,200 alumnos y esta interesada en realizar una encuesta con el fin de determinar, entre otras cosas, el número de sus alumnos que tienen una PC en su casa. El coordinador de la facultad desea estimar este total con un error máximo no mayor a los 30 alumnos y una confianza del 95 % ¿ A cuantós alumnos de la facultad se les debería aplicar la encuesta?*

Solución: Se desea estimar $\tau =$ número los alumnos de la facultad que poseen un PC en su casa con un margen de error no mayor a los 30 alumnos y un nivel de confianza del 95 %. Dado que la población de alumnos en la facultad es finita ($N = 1,200$) y $\tau = Np$, donde p denota a la proporción de alumnos de la facultad que poseen un PC en su casa, el problema equivale a estimar p con un margen de error no mayor a $e = \frac{30}{1,200} = 0.025$ y un nivel de

confianza del 95 %. Dado que $\bar{p}$ se desconoce, tomaremos la regla conservadora $\bar{p} = \frac{1}{2}$. Con ello el tamaño de muestra requerido será de

$$n = \frac{z^2_{0.975} 0.5^2 1,200}{z^2_{0.975} 0.5^2 + 0.025^2 1,199} = 674.0536 \equiv 675 \text{ alumnos.}$$

Vale observar que de no haberse tomado en cuenta el tamaño de la población ($N \to \infty$) uno hubiese obtenido con el máximo error de estimación en la proporción 0.025 pre-establecida, un tamaño de muestra de $n = 1,537$ alumnos, lo cual ciertamente no tiene sentido. □

Observaciones

- Dado que los tamaños de muestra se basaron en el estudio de un sólo parámetro, es lógico preguntarse que pasaría si en una investigación existen varios parámetros de interés. En tal caso se suguiere ubicar, según los objetivos del estudio, cuáles son estos parámetros de relevancia. Hecho esto, uno puede obtener tantos tamaños de muestra como parámetros de interés tenga y luego tomar el mayor valor de estos. Tal estrategia garantiza que en todos los casos relevantes uno obtenga a lo más los errores de estimación pre-establecidos.

- Los tamaños de muestra calculados deben siempre aproximarse por exceso a un número entero, de lo contrario no satisfaceríamos el requerimiento del máximo error pre-establecido. De otro lado, es importante en la práctica inflar estos tamaños por no respuesta. La información de tasas de no respuesta en estudios previos, pilotos o similares es en muchas situaciones fáciles de obtener.

- En este curso hemos priorizado el muestreo bajo el contexto en que nos interesa estimar ciertos parámetros poblacionales. Sin embargo, en algunas aplicaciones el estudio es comparativo o correlacional y más que estimar los parámetros con una precisión determinada nos podría interesar, por ejemplo, poder detectar ciertas diferencias entre las medias o proporciones de las poblaciones a comparar o realizar por decir un análisis de regresión. El análisis estadístico con un MASs, o más genéricamente con una muestra compleja, lo introduciremos en el capítulo 5.

2.4. Uso de software estadístico

2.4.1. Selección de muestras

Antes de mostrar un paquete especializado para el análisis de muestras probabilísticas es importante recordar cómo extraer muestras aleatorias simples, ya sea que estas sean con o sin reemplazamiento. En el primer caso, la extracción es directa y se realiza utilizando la

función de distribución empírica asociada a la selección de las unidades de la población física $\mathcal{P} = \{1, 2, \ldots, N\}$:

$$\hat{F}(i) = \frac{i}{N}.$$

Aquí basta generar n números aleatorios de una distribución uniforme en el intervalo $[0, 1]$:

$$u_1, u_2, \ldots u_n,$$

y obtener las n unidades $i_1, i_2, \ldots, i_n$ seleccionadas en $\mathcal{P}$ mediante

$$i_k = \text{mín}\{i \in \mathcal{P} \ / \hat{F}(i) \geq u_k\}, \quad \forall\ k = 1, 2, \ldots, n.$$

La muestra aleatoria simple con reemplazamiento (en $\mathcal{P}_y$) estará luego constituida por

$$y_{i_1}, y_{i_2}, \ldots, y_{i_n}.$$

En un MASs, el procedimiento anterior no es tan sencillo, pues la no restitución de los elementos previamente tomados modifica la función de distribución empírica asociada a la selección de los elementos de la población física, la cual se va también modificando. Aquí uno debe proceder secuencialmente empezando por generar un número aleatorio $u_1 \in [0, 1]$ y obteniéndo como primer elemento de la muestra a y_{i_1}, donde $i_1 = \text{mín}\{i \in \mathcal{P} \ / \hat{F}(i) \geq u_1\}$. Una vez seleccionado el k-ésimo elemento, y_{i_k}, uno procederá a generar un número aleatorio, $u_{k+1} \in [0, 1]$ y obtener

$$i_{k+1} = \text{mín}\{i \in \mathcal{P} \setminus \{i_1, i_2, \ldots, i_k\} \ / \hat{F}(i) = \frac{1}{N-k} \geq u_k\}, \quad \forall\ k = 1, 2, \ldots, n.$$

El elemento $k + 1$ de la muestra será entonces $y_{i_{k+1}}$.

Como se aprecia el procedimiento dado puede resultar engorroso, sobre todo si la muestra es con reemplazamiento. Afortunadamente en R se dispone del comando sample, el cual nos permite extraer este tipo de muestras de manera mucho más directa. La sintaxis de este comando es

```
m = sample(x, size, replace, prob)
```

donde x denota al vector con los elementos de la población estadística a escoger o simplemente es su tamaño N, size es el tamaño de muestra, replace es TRUE o FALSE, dependiendo si la muestra es con reemplazamiento o sin reemplazamiento, respectivamente, (argumento opcional que por defecto es sin reemplazamiento) y prob es un vector con las probabilidades de selección para cada elemento en x (argumento opcional que por defecto asume un muestreo con probabilidades iguales).

Si por ejemplo escribimos en R

```
m = sample(80, 10)
```

m será un vector, cuyas componentes corresponderán a los elementos seleccionados en $\mathcal{P} = \{1, 2, \ldots, 80\}$, mediante un MASs de tamaño 10. Otros comandos que nos permiten tomar un MASs o un MASc son respectivamente srswor y srswr, los cuales se encuentran en el paquete sampling.

2.4.2. El paquete survey

Existen en la literatura diferentes software estadísticos que pueden utilizarse para el análisis de muestras complejas. Información sobre ellos puede encontrarse por ejemplo en:

`http://www.hcp.med.harvard.edu/statistics/survey-soft/`

Nosotros haremos uso, aparte del siempre útil Excel y de ciertas rutinas de R, de los paquetes survey y sampling en R. Del segundo nos ocuparemos en los posteriores capítulos. En cuanto el primero tiene esencialmente dos propósitos principales:

- Enlazar la data al diseño de metadata (pesos, probabilidades de selección, unidades primarias, identificadores de estratos, etc) con el fin de poder realizar los ajustes que sean necesarios al diseño de manera confiable y automática. Esto se hace con las funciones **svydesign** y **svrepdesign** que crean objetos conteniendo no sólo la base de datos sino también la información del diseño. Así por ejemplo, uno podría extraer un subconjunto de la data y preservar su diseño aplicado a este subconjunto.
- Proveer de estimaciones válidas de la varianza para los estadísticos calculados sobre estos objetos.

El primer paso para realizar un análisis con el paquete survey consiste en crear un objeto diseño apropiado que contenga la data y la metada necesaria. Esto se hace con la función **svydesign** ó **svrepdesign** en caso se den pesos de replicación. Las funciones de análisis usualmente toman como argumento el objeto diseño y una fórmula modelo que especifica las variables a ser usadas. Los nombres de las funciones de análisis para los objetos creados con **svydesign** y **svrepdesign** comienzan con **svy** y **svr**, respectivamente.

2.4.3. La base de datos api

Como introducción al uso del paquete survey en R, consideraremos un MAS para la población contenida en la base de datos api. Una descripción de esta base de datos junto con información de las 37 variables en ella consideradas puede encontrarse en

`http://cran.fhcrc.org/web/packages/survey/survey.pdf`

Como resumen vale comentar que el estado de California exige anualmente una evaluación de sus escuelas públicas. En tal sentido el departamento de educación de este estado registra anualmente el índice api (de Academic Performance Index) que mide cuán bien va una escuela en términos de rendimiento. El archivo api contiene este índice junto con información demográfica de todas las 6,194 escuelas públicas de California con al menos 100 alumnos por escuela.

Para acceder a la base de datos y al uso del paquete survey (que debe ser instalado con antelación) escribamos:

```
library(survey)
data(api)
apipop[1:2,]

##              cds stype         name        sname snum                dname
## 1 01611190130229     H Alameda High Alameda High    1 Alameda City Unified
## 2 01611190132878     H Encinal High Encinal High    2 Alameda City Unified
##   dnum   cname cnum flag pcttest api00 api99 target growth sch.wide
## 1    6 Alameda    1   NA      96   731   693      5     38      Yes
## 2    6 Alameda    1   NA      99   622   589     11     33      Yes
##   comp.imp both awards meals ell yr.rnd mobility acs.k3 acs.46 acs.core
## 1      Yes  Yes    Yes    14  16   <NA>        9     NA     NA       25
## 2       No   No     No    20  18   <NA>       13     NA     NA       27
##   pct.resp not.hsg hsg some.col col.grad grad.sch avg.ed full emer enroll
## 1       91       6  16       22       38       18   3.45   85   16   1278
## 2       84      11  20       29       31        9   3.06   90   10   1113
##   api.stu
## 1    1090
## 2     840
```

Aquí mostramos los dos primeros registros de la base de datos api (que está en apipop). Consideremos ahora un MASs de escuelas públicas de tamaño 100, donde hemos fijado la semilla aleatoria para que usted pueda replicar los mismos resultados aquí obtenidos.

```
set.seed(12345)
N = dim(apipop)[1]
index1 = sample(N,100)
sample1 = apipop[index1,]
```

Por razones, que comentaremos luego, será también interesante agregar a esta data dos nuevas variables fpc y pw. La primera indicará simplemente el tamaño de la población (6,194) y la otra los pesos $pw = \frac{6{,}194}{100} = 61.94$ de muestreo. Ello se hace con

```
aux = data.frame(fpc = rep(N,100), pw = rep(61.94,100))
sample1 = cbind(sample1,aux)
```

Definamos ahora un objeto diseño apropiado que contenga la data y metada necesaria. Esto se hace con

```
diseñoMASs = svydesign(id=~1,fpc=~fpc,data = sample1)
```

El argumento id es para indicar los niveles de conglomerados, los cuales en este caso no existen y es por ello que colocamos id=$\sim$ 1. El argumento fpc (de factor de corrección para poblaciones finitas) nos da el tamaño de la población con lo cual implícitamente asumimos que se deben de aplicar las formulaciones de corrección para poblaciones finitas. La notación $\sim$ indica que la variable fpc está ya definida en la muestra sample1. Si el argumento fpc se omite, entonces deben de indicarse las probabilidades de selección o pesos de muestreo. Tanto id como fpc, aparte de los valores por defecto, conforman la metadata del diseño.

Otro diseño que se podría aplicar a este mismo ejemplo es por citar un MASc, para lo cual deberíamos formalmente de tomar la muestra aleatoria con reemplazamiento mediante:

```
sample2 = apipop[sample(N,100, replace=TRUE),]
sample2 = cbind(sample2,aux)
```

El objeto diseño correspondiente sería:

```
diseñoMASc = svydesign(id=~1,weights =~pw,data = sample2)
```

De pedirse información obtendríamos:

```
diseñoMASc

## Independent Sampling design (with replacement)
## svydesign(id = ~1, weights = ~pw, data = sample2)
```

Supongamos ahora que estemos interesados en estimar ciertos parámetros poblacionales, como por ejemplo el número total de alumnos matriculados, la proporción por tipo de escuelas y las medias y diferencia de medias del api entre los años 1999 y 2000. Esto, con el diseño MASs se puede hacer mediante:

```
svytotal(~enroll,diseñoMASs)

##        total SE
## enroll    NA NA

svymean(~stype, diseñoMASs)
```

```
##         mean   SE
## stypeE 0.68 0.05
## stypeH 0.20 0.04
## stypeM 0.12 0.03

means1 = svymean(~api00+api99,diseñoMASs)
means1

##       mean   SE
## api00  652 12.6
## api99  628 12.9

svycontrast(means1,c(api00=1,api99=-1))

##          contrast   SE
## contrast     24.5 2.96
```

Con un MASc, lo anterior se convierte en:

```
svytotal(~enroll,diseñoMASc)

##          total     SE
## enroll 3885868 240788
```

El hecho que se obtenga el último resultado es porque existe en la muestra con reemplazamiento algún o algunos casos perdidos. Esto puede corregirse eliminando tales casos mediante

```
svytotal(~enroll,diseñoMASc,na.rm=T)

##          total     SE
## enroll 3885868 240788
```

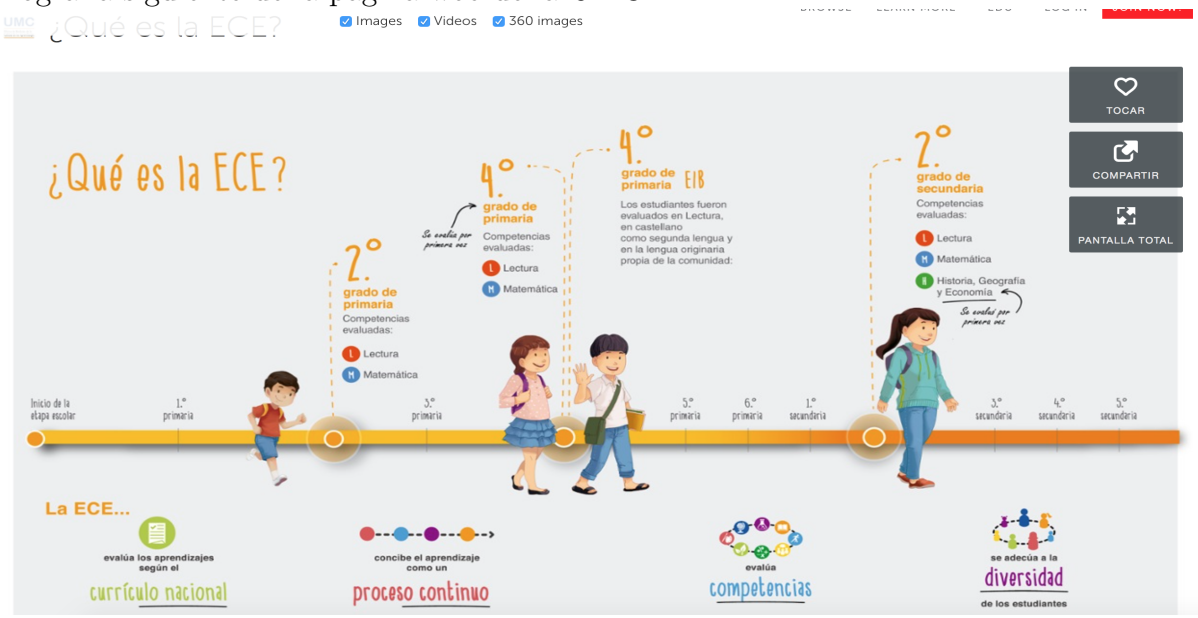
Tenemos también

```
svymean(~stype, diseñoMASc)

##         mean   SE
## stypeE 0.66 0.05
## stypeH 0.14 0.03
## stypeM 0.20 0.04
```

```
means1 = svymean(~api00+api99,diseñoMASc)
means1

##       mean   SE
## api00  693 11.5
## api99  661 12.2

svycontrast(means1,c(api00=1,api99=-1))

##          contrast   SE
## contrast     32.1 3.21
```

2.4.4. La evaluación censal de estudiantes 2016

La unidad de medición de la calidad de los aprendizajes (UMC) del Ministerio de Educación, ha publicado recientemente en el 2017 los resultados de la última evaluación censal de estudiantes (ECE) 2016. Ella consta de una serie de documentos, los que se ilustran en la infografía siguiente de la página web de la UMC.



En la misma página web se encuentran todas las bases de datos descritas en formato SPSS. Se presentan allí también reportes de los resultados e inclusive una sintaxis en R para el análisis de los resultados de la evaluación muestral para el segundo grado de primaria. Nosotros consideraremos inicialmente como nuestra población objetivo a los resultados censales de la dirección regional de Amazonas para el segundo grado de secundaria. Más adelante trabajaremos con una población mayor, llegando incluso, a través de un diseño muestral, a analizar los datos a nivel nacional. Las variables de interés en estas bases de datos son los

puntajes de evaluación en las áreas de Matemáticas, Comprensión lectora e Historia, Geografía y Economía (todas en una escala Rasch normalizada a un puntaje promedio de 500 puntos). De particular interés para el ministerio son también los niveles de logro, los cuales se obtienen al categorizar los puntajes anteriores en 4 niveles. Luego de una breve edición de la base de datos en SPSS, importaremos ella a R bajo el nombre ce2s16Am.RData mediante el paquete foreign. Los comandos y una descripción de la base de datos se presentan seguidamente:

```
library(foreign)
ce2s16Am = read.spss("ce2s16Am.sav", use.value.labels=TRUE)
ce2s16Am = as.data.frame(ce2s16Am)
save(ce2s16Am,file='ce2s16Am.RData')
head(ce2s16Am)

##   Colegio Seccion        Nom_ugel Departamento       Provincia
## 1   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
## 2   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
## 3   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
## 4   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
## 5   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
## 6   00201      01 Chachapoyas     AMAZONAS     CHACHAPOYAS
##          Distrito GestionE   Area   Sexo          Grupo_L          Grupo_M
## 1 CHACHAPOYAS      Estatal Urbana  Mujer Previo al inicio Previo al inicio
## 2 CHACHAPOYAS      Estatal Urbana Hombre       En proceso    Satisfactorio
## 3 CHACHAPOYAS      Estatal Urbana Hombre        En inicio        En inicio
## 4 CHACHAPOYAS      Estatal Urbana  Mujer        En inicio Previo al inicio
## 5 CHACHAPOYAS      Estatal Urbana Hombre    Satisfactorio    Satisfactorio
## 6 CHACHAPOYAS      Estatal Urbana Hombre        En inicio        En inicio
##          Grupo_HGE M500_L M500_M M500_H
## 1 Previo al inicio    497    478    399
## 2    Satisfactorio    624    682    685
## 3       En proceso    552    543    588
## 4 Previo al inicio    571    509    411
## 5    Satisfactorio    680    660    748
## 6        En inicio    575    546    487
```

Note que a diferencia de la base de datos api, las unidades en esta base son lo alumnos y no los colegios.

Supongamos ahora que nuestro interés sea estimar el rendimiento medio de los alumnos tanto en Matemáticas (M), Comprensión lectora (L) e Historia, Geografía y Economía (H),

con un margen de error no mayor a 5 puntos y un nivel de confianza del 95 %. Para encontrar el tamaño de muestra, requeriremos de estimaciones de la varianza para estos puntajes, las cuales las podríamos obtener de la ECE 2015 o a través de un estudio piloto. Si optamos por esta última estrategia con una muestra piloto de 30 colegios, la selección correspondiente así como la estimación de las varianzas requeridas se hará como sigue.

```
library(survey)
set.seed(12345)
N = dim(ce2s16Am)[1]
index1 = sample(N,30)
mp16Am = ce2s16Am[index1,]
dismp = svydesign(id=~1,fpc=rep(N,30),data=mp16Am)
sigmae2_L = coef(svymean(~M500_M,dismp,na.rm=T))
sigmae2_M = coef(svymean(~M500_L,dismp,na.rm=T))
sigmae2_H = coef(svymean(~M500_H,dismp,na.rm=T))
```

Dado que tenemos 3 variables optaremos, como comentamos, en seleccionar el mayor tamaño de muestra proveniente de estas variables con un redondeo por exceso.

```
d = 25*N/(qnorm(0.975)^2)
n1 = N*sigmae2_L/(d + sigmae2_L)
n2 = N*sigmae2_M/(d + sigmae2_M)
n3 = N*sigmae2_H/(d + sigmae2_H)
n = ceiling(max(n1,n2,n3))
```

El tamaño de muestra requerido será por tanto igual a

```
n

## [1] 86
```

alumnos.

La toma de muestra, definición del diseño y las estimaciones de los rendimientos y proporciones de logro se muestran a continuación

```
#library(survey)
set.seed(12345)
index = sample(N,n)
m16Am = ce2s16Am[index,]
disem = svydesign(id=~1,fpc=rep(N,n),data=m16Am)
```

```
mean_L = svymean(~M500_L,disem,na.rm=T)
mean_M = svymean(~M500_M,disem,na.rm=T)
mean_H = svymean(~M500_H,disem,na.rm=T)
meanp_L = svymean(~Grupo_L,disem,na.rm=T)
meanp_M = svymean(~Grupo_M,disem,na.rm=T)
meanp_H = svymean(~Grupo_HGE,disem,na.rm=T)
mean_L

##        mean   SE
## M500_L  550 8.27

mean_M

##        mean   SE
## M500_M  540 9.86

mean_H

##        mean   SE
## M500_H  481 10.5

pr = rbind(meanp_L,meanp_M,meanp_H)
colnames(pr) = c("Previo al inicio","Inicio","En proceso","Satisfactorio")
pr

##         Previo al inicio Inicio En proceso Satisfactorio
## meanp_L            0.247  0.506      0.106        0.1412
## meanp_M            0.477  0.267      0.163        0.0930
## meanp_H            0.321  0.202      0.381        0.0952
```

2.4.5. El censo nacional de población penitenciaria 2016

El censo nacional de población penitenciaria, 2016; realizado por primera vez en el país por el Instituto Nacional de Estadística e Informática (INEI), generó información estadística cuantitativa y cualitativa actualizada sobre la problemática penitenciaria en el Perú, permitiendo contar con indicadores respecto a las condiciones sociales y familiares de la persona privada de libertad, tipicidad del delito, condiciones de vida en los establecimientos penitenciarios y rol de las instituciones. El censo busco formular estrategias de asistencia post penitenciaria, reeducación, formación laboral y rehabilitación, que permitan una adecuada política penitenciaria de reinserción de las personas privadas de libertad a la sociedad así co-

mo de implementar políticas públicas de seguridad y mantenimiento de los establecimientos penitenciarios del país. La base de datos de este censo es de libre disponibilidad y se puede encontrar en la página web del INEI:

`http://iinei.inei.gob.pe/microdatos/`

La versión de esta base de datos, que utilizaremos a lo largo del curso, se encuentra en la intranet del curso con el nombre BasR.sav. Ella está en formato SPSS y contiene todos los 76,180 registros de personas privadas de libertad en el país consignadas en el censo y la gran mayoría de preguntas de la encuesta, última que también se encuentra en la página web del INEI. Para utilizar la base de datos en R, debemos primero instalar el paquete foreign y luego invocar los comandos:

```
library(foreign)
#cp16b <- read.spss(file.choose(), use.value.labels=TRUE)
cp16b <- read.spss("BasR.sav", use.value.labels=TRUE)
cp16 = as.data.frame(cp16b)
cp16_labels <- attr(cp16b, "variable.labels")
cp16_cat <- attr(cp16b, "label.table")
save(cp16,file='cp16.RData')
```

La base de datos a utilizar es cp16; mientras que los archivos `cp16_labels` y `cp16_cat` contienen información de respectivamente las etiquetas y categorías de las variables seleccionadas. Como se aprecia, la base de datos cp16 ha sido también grabada para uso futuro en el formato de R. Esta base tiene como seguidamente se aprecia 189 variables, de las cuales mostramos las primeras 18.

```
head(cp16[,1:18])
```

```
##   ID      PDEP      PPROV       PDIS        PCP
## 1  3 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 2 19 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 3 24 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 4 26 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 5 39 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
## 6 40 CAJAMARCA CAJAMARCA  CAJAMARCA  CAJAMARCA
##                          OFICINA_R EST_PENIT PABELLON GENERO     E_CIVIL
## 1 Oficina Regional Norte Chiclayo Cajamarca        4  Mujer   Casado(a)
## 2 Oficina Regional Norte Chiclayo Cajamarca       NA  Mujer    Viudo(a)
## 3 Oficina Regional Norte Chiclayo Cajamarca       NA Hombre   Casado(a)
```

```
## 4 Oficina Regional Norte Chiclayo Cajamarca       NA Hombre    Viudo(a)
## 5 Oficina Regional Norte Chiclayo Cajamarca        3 Hombre   Casado(a)
## 6 Oficina Regional Norte Chiclayo Cajamarca        7 Hombre Conviviente
##     RELIGION EDAD    NACIONALIDAD        PAIS_NAC     DEP_NAC  DEP_URES
## 1   Católica   39 PERUANO         PERU                   LIMA      LIMA
## 2     Mormón   49 PERUANO         PERU                   LIMA      LIMA
## 3    Ninguna   25 PERUANO         ESTADOS UNIDOS           NA        NA
## 4       Otra   26 PERUANO         PERU                 CUSCO      LIMA
## 5 Evangélica   49 PERUANO         PERU              CAJAMARCA CAJAMARCA
## 6    Ninguna   40 PERUANO         PERU            LA LIBERTAD CAJAMARCA
##                CP_URES                          DEL_GENERICO_CD
## 1 CIUDAD DE DIOS                    DELITOS CONTRA EL PATRIMONIO
## 2 BARRIO OBRERO INDUST              DELITOS CONTRA EL PATRIMONIO
## 3                                   DELITOS CONTRA EL PATRIMONIO
## 4 VILLA EL SALVADOR                 DELITOS CONTRA EL PATRIMONIO
## 5 LA COLPA             DELITOS CONTRA LA ADMINISTRACION PUBLICA
## 6 CAJAMARCA                         DELITOS CONTRA EL PATRIMONIO
```

La distribución de frecuencias del número de internos, condición de género (CG) y capacidad de los establecimientos penitenciarios en cada oficina regional y departamento se muestran en el Cuadro 2.4.

Como una primera aproximación al análisis de esta base de datos consideraremos un MASs, cuyo objetivo será poder estimar cualquier proporción poblacional con un margen de error no mayor a 0.03 y una confiabilidad del 95 %. Para ello el tamaño de muestra requerido estará dado por:

$$n = \frac{1.96^2 0.5(1-0.5)76180}{1.96^2 0.5(1-0.5) + 0{,}03^2 76179} = 1,052.383$$

que redondeando nos da un valor de 1,053 internos. Si bien usaremos este número, vale comentar que ello es asumiendo de que todos responderán a la encuesta. En encuestas similares para la región se han encontrado tasas de no respuesta de entre el 21 % y 22 %. Una práctica que comentaremos en el último capítulo es el de inflar este número ante la posibilidad de no respuesta. Ello nos sugeriría encuestar a 1,285 internos. Para efectos de este ejercicio tomaremos sólo 1,053 ya que que en nuestro caso es posible acceder a toda la información. Tomada la muestra, estimemos por citar la edad promedio de los internos, la proporción de internos sentenciados y la proporción de ellos que tienen un abogado. Los códigos siguientes nos permitarán hacer todo ello.

```
set.seed(12345)
load('cp16.RData')
```

OFICINA REGIONAL	DEPARTAMENTO	E.PENITENCIARIO	NUMERO DE INTERNOS	CG	Capacidad
Norte Chiclayo	CAJAMARCA	Cajamarca	1389	Mix	888
		Chota	131	H	65
		Jaen	377	Mix	50
		San Ignacio	79	H	150
	LA LIBERTAD	Pacasmayo	11	M	72
		Trujillo	4471	H	1518
		Mujeres de Trujillo	283	M	160
	LAMBAYEQUE	Chiclayo	3163	Mix	1143
	PIURA	Piura	3098	H	1370
		Sullana	94	M	50
	TUMBES	Tumbes	860	Mix	384
Lima	ANCASH	Huaraz	1014	Mix	350
		Chimbote	2321	Mix	920
	CALLAO	Callao	3201	H	572
		Base Naval Callao	7	H	8
	ICA	Chincha	1331	H	1152
		Ica	3943	Mix	1464
	LIMA	Cañete	1982	H	768
		Huaral	3164	H	823
		Huacho	1738	Mix	644
		Ancon	2289	H	1620
		Modelo Ancon II	1462	Mix	2200
		Anexo Mujeres Chorrillos	309	M	288
		Mujeres de Chorrillos	810	M	450
		Virgen de Fatima	339	M	548
		Virgen de la Merced	13	H	42
		Lurigancho	9602	H	3204
		Miguel Castro Castro	4359	H	1142
		Barbadillo	1	H	1
Sur Arequipa	AREQUIPA	Arequipa	1971	H	667
		Mujeres de Arequipa	151	M	67
		Camana	262	H	78
	TACNA	Tacna	830	H	222
		Mujeres de Tacna	110	M	40
		Challapalca	162	H	214
Centro Huancayo	AYACUCHO	Ayacucho	2438	Mix	644
		Huanta	101	H	42
	HUANCAVELICA	Huancavelica	200	H	60
	JUNIN	Chanchamayo	572	Mix	120
		Huancayo	1972	H	680
		Mujeres de Concepción	31	M	105
		Jauja	104	M	85
		Satipo	164	H	50
		Tarma	84	H	48
		Oroya	114	Mix	64
Oriente Huanuco (Pucallpa)	HUANUCO	Huanuco	2554	Mix	1074
	PASCO	Cerro Pasco	195	Mix	96
	UCAYALI	Pucallpa	2053	Mix	788
Sur Oriente Cusco	APURIMAC	Abancay	256	Mix	90
		Andahuaylas	354	Mix	248
	CUSCO	Cusco	2288	H	800
		Mujeres del cusco	137	M	62
		Quillabamba	347	Mix	80
	MADRE DE DIOS	Pto. Maldonado	712	H	590
Nor Oriente San Martín	AMAZONAS	Chachapoyas	629	Mix	288
		Bagua Grande	230	Mix	60
	LORETO	Yurimaguas	157	Mix	286
		Iquitos	1025	H	600
		Mujeres de Iquitos	64	M	78
	SAN MARTIN	Juanjui	686	Mix	654
		Moyobamba	588	Mix	544
		Sananguillo	548	H	636
		Tarapoto	463	H	180
Altiplano Puno	PUNO	Lampa	136	M	44
		Puno	582	H	778
		Juliaca	1069	Mix	420

Cuadro 2.4: Distribución de frecuencias del número de internos, condición de género (CG) y capacidad de los establecimientos penitenciarios en cada oficina regional y departamento del Perú

```
N = dim(cp16)[1]
index = sample(N,1053)
sample = cp16[index,]
aux = data.frame(fpc = rep(N,1053), pw = rep(72.34568,1053))
sample = cbind(sample,aux)
diseMASs = svydesign(id=~1,fpc=~fpc,data = sample)
svymean(~EDAD, diseMASs,na.rm=T)

##      mean   SE
## EDAD 35.8 0.35

svymean(~SITUACION_JURIDICA,diseMASs,na.rm=T)

##                                 mean   SE
## SITUACION_JURIDICAProcesado    0.222 0.01
## SITUACION_JURIDICASentenciado 0.778 0.01

svymean(~ABOGADO,diseMASs,na.rm=T)

##           mean   SE
## ABOGADOSí 0.53 0.02
## ABOGADONo 0.47 0.02
```

Otro análisis de interés podría ser analizar si existe relación entre si el interno consumía drogas o no y el tipo de delito que ha cometido. Antes de analizar ello será conveniente recodificar la tipicidad del delito a los delitos más comúnes, creando la variable DGEN. Como la prueba indica y se visualiza en el gráfico de barras agrupadas no encontramos evidencia de una asociación entre el consumo de drogas y la tipificación del delito.

```
DGEN = cp16$DEL_GENERICO_CD
levels(DGEN)[c(1,2,3,4,5,7,8,9,10,11,14,16,17,18,19)] = "OTROS"
DGEN = DGEN[index]
DGEN = factor(DGEN,levels(DGEN)[c(2,3,4,5,1)])
chisq.test(DGEN,sample$DROGAS)

##
##  Pearson's Chi-squared test
##
## data:  DGEN and sample$DROGAS
## X-squared = 3, df = 4, p-value = 0.6

tab = table(sample$DROGAS,DGEN)
```

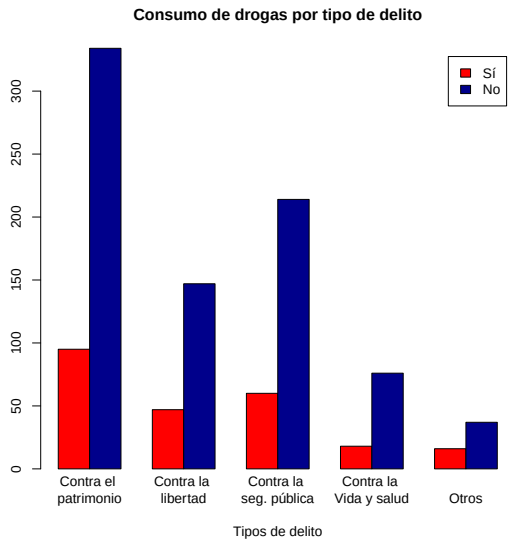


Figura II: Frecuencias de consumo de drogas por tipo de delito

2.4.6. La población peruana con DNI 2016

En este último ejemplo consideraremos a la población peruana que se encuentra en el registro nacional de identificación y estado civil (RENIEC) al 31 de Diciembre del 2016 y que por tanto cuenta con su documento nacional de identidad (DNI), el cual otorga derecho a sufragio a partir de los 18 años. La información pública de la RENIEC incluye el lugar de residencia, edad, sexo y condición de extranjería de la persona. Ella puede ser obtenida fácilmente en formato Excel o SPSS de la página web de esta institución. Una mirada a la base de datos

```
library(foreign)
reniec16 = read.spss("Reniec2016.sav", use.value.labels=TRUE)
```

```
## re-encoding from latin-9
```

```
reniec16 = as.data.frame(reniec16)
head(reniec16)
```

```
##   Departamento  Provincia   Distrito   Sexo Edad Cantidad
## 1   Amazonas   Chachapoya Chachapoya Hombre    0      342
## 2   Amazonas   Chachapoya Chachapoya  Mujer    0      339
## 3   Amazonas   Chachapoya Chachapoya Hombre    1      312
## 4   Amazonas   Chachapoya Chachapoya  Mujer    1      350
## 5   Amazonas   Chachapoya Chachapoya Hombre    2      347
## 6   Amazonas   Chachapoya Chachapoya  Mujer    2      300
```

revela que la última variable, Cantidad, contiene la frecuencia de casos que comparten las demás variables. Como ilustración, al 2016 se tenían 342 varones registrados en el distrito de Chachapoyas, provincia de Chachapoyas y departamento de Amazonas que áun no habían cumplido el año de edad. Esta variable por tanto, es una variable de ponderación para toda la base de datos, con lo cual esta contendrá a nivel nacional un total de

```
N = sum(reniec16$Cantidad)
N

## [1] 33949373
```

registros.

En este ejemplo estaremos interesados mediante un MASs de estimar cualquier proporción de interés con un margen de error no mayor a 0.02 y una confianza del 95 %. Esto por ejemplo, podría darse para fines de una encuesta de opinión electoral u otros, sólo que para acceder a la vivienda específica del entrevistado habría que gestionarse la obtención de su dirección u otra información pertinente. Si tomamos la regla conservadora de $\bar{p} = 0.5$, entonces el tamaño de muestra requerido será de

$$n = \frac{(1.96^2 0.5(1-0.5))N}{1.96^2 0.5(1-0.5) + 0.02^2(N-1)} = 2,400.83$$

2,401 personas.

Antes de tomar la muestra requeriremos expandir la base de datos de individuos en base a la variable Cantidad. Esto puede hacerse con el siguiente comando en R, en el cual hemos generado la base de datos expandida reniec16x.RData

```
reniec16x = reniec16[rep(1:nrow(reniec16),reniec16$Cantidad),]
save(reniec16x,file='reniec16x.RData')
```

Si tomamos el MASs planificado, obtendremos la siguiente base de datos muestral

```
load('reniec16x.RData')
set.seed(12345)
indexp = sample(N,2401)
sampleDNI = reniec16x[indexp,]
head(sampleDNI[,1:5])

##              Departamento       Provincia        Distrito   Sexo Edad
## 265566.476 Lima            Huaral          Huaral           Mujer   20
## 322308.77  Puno            El Collao       Ilave          Hombre   20
## 277589.270 Loreto          Mariscal Ramón  San Pablo        Mujer    8
```

```
## 328189     San Martín      Lamas           San Roque de Cu Hombre   42
## 238581.706 Lima            Lima            Carabayllo       Mujer   57
## 124381.125 Cajamarca       San Ignacio     San Ignacio      Hombre   44
```

Si bien considerar un MASs es teóricamente posible y ha sido en este y los anteriores ejemplos bastante simple, este no es ciertamente un diseño recomendable para poblaciones tan grandes como las aquí consideradas. En nuestros ejemplos contamos en todos los casos con una base de datos poblacional, situación que raramente se presenta en la práctica. En la realidad frecuentemente el marco muestral está desactualizado, pobremente definido o es inexistente y, por otro lado, la muestra aleatoria simple resulta estar tan geográficamente dispersa que los costos y logística resultan inmanejables. En esta muestra por citar apreciemos el lugar de residencia de las 6 primeras personas seleccionadas. Si la encuesta objetivo es de opinión e incluso tuvieramos sus direcciones, tiene poco sentido el tratar de ubicar a estas personas tan alejadas unas de otras debido al costo que ello implicaría y al aparato logístico que se tendría que implementar para garantizar la supervisión y calidad del trabajo de campo. En los capítulos siguientes exploraremos diseños mucho más eficientes para los fines buscados.

Para terminar, obtengamos la estimación de la proporción de mujeres y de personas con derecho a votar (con 18 o más años de edad) en esta población.

```
diseDNI = svydesign(id=~1,fpc=rep(N,nrow(sampleDNI)),data=sampleDNI)
svymean(~Sexo,diseDNI)

##              mean   SE
## SexoHombre 0.502 0.01
## SexoMujer  0.498 0.01

svymean(~Edad>=18,diseDNI)

##                   mean   SE
## Edad >= 18FALSE 0.302 0.01
## Edad >= 18TRUE  0.698 0.01
```

2.5. Ejercicios

1.- Considere n unidades, n_1 de los cuales son indistinguibles de tipo A_1, n_2 indistinguibles de tipo A_2 y así sucesivamente hasta n_k indistinguibles de tipo A_k.
a) Muestre que el número de maneras de ordenar estas n unidades viene dado por

$$\frac{n!}{n_1!n_2!\ldots n_k!}.$$

b) Para una entrevista de trabajo se han presentado 12 estadísticos, 3 de la UNI, 4 de San Marcos, 3 de la Agraria y 2 de la PUCP. Si ellos son llamados para la entrevista en un orden aleatorio ¿con qué probabilidad los tres primeros entrevistados serán de San Marcos?

2.- Juan, Pepe, Rosa, Luis y María son los únicos participantes de un sorteo donde se seleccionarán al azar 4 nombres, repartiéndose al que sea elegido en cada ocasión 50 soles.
a) Si Juan desea ganar algún monto ¿qué es lo que más le convendría: que la selección se haga con o sin reemplazamiento?
b) Si la selección se hace con reemplazamiento ¿qué probabilidad hay de que Juan gane 50 soles y Rosa 100 soles? ¿Es esta probabilidad la misma a que Juan gane 200 soles?
c) Bajo reemplazamiento ¿con qué probabilidad sólo Rosa y Luis ganaran dinero?
d) Halle, bajo reemplazamiento, el monto que esperará obtener Juan en el sorteo.

3.- Considere una pequeña población conformada por 6 personas, a las que se les ha medido el nivel de hemoglobina, encontrándose en gramos por decilitro las siguientes mediciones

$$13.9, 11.5, 16.7, 14.4, 14.6, 15.1$$

Mediante un MASc y un MASs de tamaño $n = 3$:
a) Halle la probabilidad de que la media del nivel de hemoglobina de las tres personas seleccionadas supere los 14 gramos por decilitro.
b) Suponga que para estimar el nivel promedio de hemoglobina en estas 6 personas se propone la mediana de los valores observados en la muestra ¿sería este un estimador insesgado? ¿tiene este una menor varianza que la media muestral?
c) Usando los números aleatorios 0.018, 0.310 y 0.549, tome las muestras requeridas y estime la media del nivel de hemoglobina de las 6 personas.

4.- Dos encuestadoras han realizado MASs de tamaños 20 y 10; respectivamente, para una población de 50 personas. Puesto que el muestreo fue hecho en fechas distintas, ninguna supo que la otra había realizado la encuesta. Halle la función de probabilidad, valor esperado y varianza del número de personas en esta población que fueron entrevistadas por ambas encuestadoras.

5.- a) Si $X \sim B(N, p)$ e $Y|_{X=x} \sim H(N, x, n)$, muestre que $Y \sim B(n, p)$.
b) Suponga que en un estudio sobre la prevalencia de una enfermedad (proporción p de personas que la padecen) se ha diseñado un MASc de tamaño 420. Al ser consultado, un estadístico manifiesta que tal tamaño es demasiado grande, pues conocer si las personas tienen o no la enfermedad pasará por una prueba cara y de logística complicada. Dado que se han enviado ya cartas a las personas seleccionadas, el estadístico sugiere tomar más bien un MASs de tamaño 80 de la población de los 420 inicialmente seleccionados. Si se acepta la sugerencia del estadístico y si p fuera 0.1 ¿con qué probabilidad se encontrará en la muestra a más 5 personas que padezcan la enfermedad?

6.- Una manera de estimar el tamaño N de una población consiste en usar métodos de captura-recaptura. Estos consisten en primero seleccionar al azar m elementos de la población para reponerlos a ella luego de marcarlos. Seguidamente se tienen dos estrategias. El método directo consiste en seleccionar al azar y sin reemplazamiento otra muestra de n elementos de la población para registrar luego el número de elementos marcados X que se encuentren en ella. El segundo método, llamado muestreo inverso, consiste en seleccionar al azar y con reemplazamiento (podría también analizar el caso sin reemplazamiento) sistemáticamente elementos de la población hasta ubicar r elementos marcados. Con ello se tienen los siguientes dos estimadores de N:

$$\hat{N}_1 = \frac{nm}{X} \quad \text{y} \quad \hat{N}_2 = \frac{mY}{r},$$

donde Y denota al número de intentos hasta obtener la cuota de r elementos marcados.
a) Usando una expansión de Taylor adecuada, muestre que aproximadamente se cumple que $E(\hat{N}_1) = N + \frac{2N(N-m)(N-n)}{nm(N-1)}$ y

$$V(\hat{N}_1) = \frac{N^2(N-m)(N-n)}{nm(N-1)}.$$

b) Como se aprecia en a) $\hat{N}_1$ es no sólo es un estimador sesgado de N sino que presenta una gran varianza si la muestra correspondiente contiene muy pocos elementos marcados. Muestre que contrariamente $\hat{N}_2$ es un estimador insesgado de N y que tiene una varianza igual a

$$V(\hat{N}_2) = \frac{N(N-m)}{r}.$$

Pruebe además que

$$\hat{V}(\hat{N}_2) = \frac{m^2 Y(Y-r)}{r^2(r+1)}.$$

es un estimador insesgado de la varianza última ¿Qué desventaja cree que pudiera tener este método con respecto al muestreo directo?
c) Suponga que para estimar el número de personas N que pertenecen a un gran consorcio se han seleccionado al azar a 20 de sus trabajadores, registrándoles y colocándoseles un sello en su DNI. Tiempo después la dirección de recursos humanos tomó un MASs de 100 trabajadores, encontrándo que 4 de ellos tenían el sello en el DNI. Por su parte usted optó más bien por seleccionar secuencialmente al azar y con reemplazamiento trabajadores del consorcio hasta ubicar a 5 con el sello en el DNI, realizando un total de 127 registros. Obtenga las estimaciones correspondientes de N, junto con sus intervalo de confianza asintóticos al 95 %. Comente.

7.- a) Demuestre que en un MASc, la media muestral es el MELI de la media poblacional.
b) Demuestre la fórmula de la covarianza entre las componentes de una distribución hipergeométrica multivariada y utilice ella para obtener la varianza de la distribución hipergeométrica.

8.- Considere una población finita de tamaño N en la cual se desea estudiar una variable estadística y, la cual toma un valor muy pequeño para el primer elemento del marco muestral y_1 y un valor muy grande para el último elemento del marco muestral y_N. Con el propósito de estimar la media de y para esta población, μ, se ha propuesto, en base a un MASs de tamaño n, el estimador:

$$\bar{Y}_c = \begin{cases} \bar{Y} + c, & \text{si } y_1 \text{ pertence a la muestra e } y_N \text{ no pertence a la muestra} \\ \bar{Y} - c, & \text{si } y_1 \text{ no pertence a la muestra e } y_N \text{ pertence a la muestra} \\ \bar{Y}, & \text{en otro caso,} \end{cases}$$

donde c es una constante positiva.

a) ¿ Es $\bar{Y}_c$ un estimador insesgado de μ ?

b) Halle la varianza de $\bar{Y}_c$.

c) ¿ Existen valores de c que hagan que $\bar{Y}_c$, tenga menor varianza que $\bar{Y}$? ¿Contradice esto a que $\bar{Y}$ sea el MELI de μ ?

9.- Suponga que se desea estimar, con un error no mayor al 3 % y una confianza del 95 % la prevalencia de una rara enfermedad al interior de una pequeña comunidad integrada por 500 habitantes. Se espera que la proporción de personas de la comunidad que tengan esta enfermedad es pequeña, lo cual se ha evidenciado en una muestra piloto realizada a 30 de sus habitantes en la que se encontró que sólo 2 de ellos tenían la enfermedad.

a) Halle, con la fórmula estándar, el tamaño de muestra para este estudio.

b) Puesto que la proporción a estimar es extrema utilice mas bien el IC de Wilson para obtener el tamaño de muestra para este estudio. Comente la diferencia encontrada con a) e indique cuál de los tamaños de muestra utilizaría usted para el estudio.

10.- Una ciudad cuenta con 720 fábricas, de las cuales 10, 20 y 8 pertenecen respectivamente a los consorcios A,B y C. El ministerio de trabajo desea hacer un estudio de salud ocupacional en las fábricas de la ciudad. Dado que muchos de los indicadores a estudiarse son proporciones, el ministerio desea tomar un MASs de tamaño n de tal manera que pueda estimar cualquier proporción con un margen de error no mayor a 0.1 y un nivel de confianza del 95 %

a) ¿Cuál debería el tamaño de muestra a tomarse?

b) ¿Con qué probabilidad se seleccionará en la muestra del tamaño tomado en a) a alguna de las fábricas del consorcio B?

c) Suponga que tomada la muestra en a), y dadas las características especiales de los tres consorcios en mención, el ministerio ordena que de ser seleccionada cualquier fábrica de algunos de los consorcios, se seleccione igualmente a todas las fábricas del consorcio seleccionado ¿Cuál sería el tamaño de muestra esperado final que se obtendría a través de este procedimiento?

11.- En cierta área de una ciudad, que contiene 14,848 residencias, se desea estimar el número medio de personas μ por residencia. Si en un MASs con tamaño 30 se obtuvieron las siguientes cantidades de personas por residencia:

$$5, 6, 3, 3, 2, 3, 3, 3, 4, 4, 3, 2, 7, 4, 3, 5, 4, 4, 3, 3, 4, 3, 3, 1, 2, 4, 3, 4, 2, 4.$$

a) Estime μ, junto con su intervalo de confianza al 95 %.
b) Suponga que se desea estimar el número medio anterior con el doble de precisión que la brindada por la muestra anterior ¿cuál debería ser el tamaño de muestra para lograr ello?

12.- Su distrito, que cuenta con N viviendas, participará en una encuesta con un MASs de tamaño n. Suponga que existe una probabilidad constante q de que una vivienda no responda a esta encuesta. Para prevenir ello el supervisor ha decidido, de ser necesario, seleccionar al azar y sin reemplazamiento durante un segundo día un número igual al número de viviendas sin respuesta en el primer día de entre las viviendas aún no seleccionadas.
a) ¿Con qué probabilidad su vivienda será encuestada el primer día?
b) Si en el primer día su vivienda no es seleccionada y no hubo respuestas en M viviendas ¿con qué probabilidad su vivienda será seleccionada en el segundo día?
c) Si sus padres residen en otra vivienda de su distrito ¿qué probabilidad existe de que su vivienda y la de sus padres sean seleccionadas?
d) ¿Con qué probabilidad no será posible completar el tamaño de muestra que ha sido planificado para la encuesta?
e) Obtenga d) si $q = 0.06$ y $n = 100$.

13.- Para realizar una encuesta de opinión se ha diseñado un MASs de tamaño 100 para una población de 150,000 habitantes, entre las que se encuentra usted y un amigo suyo.
a) ¿Con qué probabilidad usted integrará la muestra?
b) Suponga ahora que 5 muestras como las anteriores son secuencialmente seleccionadas de esta población mediante un MASs ¿qué probabilidad existe de que ni a usted ni a su amigo se les pida su opinión? Asuma que los encuestadores de cada una de las 5 muestras no toman en cuenta el registro de si una persona fué o no seleccionada en otra de las muestras.
c) ¿Con qué probabilidad le pedirán en b) dos veces su opinión?

14.- Considere una población finita de N personas y supongamos se extrae de ella una muestra aleatoria simple con reemplazamiento de tamaño $n = 5$.
a) Halle la función de probabilidad de la variable aleatoria X que denota al número de personas distintas que contendrá la muestra.
b) Suponga que extraida la muestra anterior es de interés estimar el total τ de una variable y, para lo cual usted multiplicará por una constante C la suma de todos los valores de y en la muestra que correspondan sólo a personas distintas ¿cuál sería el valor de C que haga que este sea un estimador insesgado del total? Halle también la varianza de este estimador.

15.- Replique el estudio realizado con la base de datos ECE 2016, pero ahora para la DRE de Lima Metropolitana. Dado que, a diferencia de la base de datos de Amazonas, esta incluye un indicador de nivel socio-económico (cuya metodología de construcción por componentes principales puede encontrarse en la página web de la UMC), se le pide también indicar mediante un MASs, pueda indicar si existe o no asociación significativa para esta DRE, entre el nivel socio-económico y los niveles de logro. Use un nivel de significación de $\alpha = 0.05$.

16.- En este capítulo vimos que S^2 es un estimador insesgado de la varianza poblacional σ_N^2 en un MASc y de σ_{N-1}^2 en un MASs ¿pero qué hay de su varianza?
a) Muestre que

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (Y_i - Y_j)^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - y_j)^2 \delta_i \delta_j.$$

b) Muestre, usando la fórmula anterior, que S^2 es efectivamente un estimador insesgado de su correspondiente varianza poblacional.
c) Cho y Cho (2008) han derivado fórmulas para la varianza de S^2, tanto en un esquema MASc como en un MASs. Estas vienen dadas respectivamente por:

$$V_{MASc}(S^2) = \frac{1}{n}(\mu_4 - (\frac{n-3}{n-1})\sigma_N^4) \quad y$$

$$V_{MASs}(S^2) = C((Nn - N - n - 1)\mu_4 - (\frac{N^2 n - 3n - 3N^2 + 6N - 3}{N-1})\sigma_N^4),$$

donde $C = \frac{N(N-n)}{n(n-1)(N-1)(N-2)(N-3)}$ y $\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_N)^4$ es el cuarto momento centrado poblacional. Muestre que conforme $N \to \infty$, $V_{MASs}(S^2)$ converge a $V_{MASc}(S^2)$.
d) Suponga que para determinar el tamaño de muestra en un MASs, que permita estimar la media de una variable y en una población de tamaño N, usted dispone de dos encuestas similares previas independientes de tamaños n_1 y n_2 realizadas mediante un MASs sobre esta misma población. A efectos de estimar la varianza poblacional de y, usted decide tomar como estimador a $\hat{\sigma}^2 = c_1 S_1^2 + c_2 S_2^2$, siendo c_1 y c_2 constantes a determinar y S_1^2 y S_2^2 las varianzas muestrales de y para las encuestas previas ¿Cómo seleccionaría estas constantes para que su estimador sea insesgado y tenga además la menor varianza entre todos los estimadores de este tipo?

17.- En un país se ha diseñado una encuesta con el fin de medir la tasa de desempleo, que se cree está en torno al 10 % de la población activa civil. La población activa civil se define como la población de 14 años o más, y constituye cerca del 65 % de la población total del país que fué estimada en el último censo en 2.3 millones de habitantes. Si queremos estimar la tasa de desempleo con un margen de error no mayor al 1 % para un nivel de confianza del 95 %, mediante un MASs ¿cuál sería el tamaño de muestra requerido?

18.- Una zona rural de 3,000 viviendas ha sido elegida para tomarse una encuesta con un MASs de 100 viviendas. Un interés de la encuesta es estimar el consumo total mensual de agua para los hogares que cuentan con servicio de agua y desague. Se asume que antes de tomarse la muestra no es posible identificar en la región si una vivienda de la zona posee o no estos servicios, pero que si se sabe el 70 % de las viviendas los tienen.
a) En general, dada una población estadística $\mathcal{P}_y = \{y_1, y_2, \ldots, y_N\}$ y un MASs de tamaño n en ella muestre que para cierto subconjunto de esta población (dominio)

$$\hat{\tau}_d = \frac{N}{n}\sum_{i=1}^{N} y_i\gamma_i\delta_i \quad \text{ó} \quad \hat{\tau}_d = \frac{N}{n}\sum_{i=1}^{n} Y_i\gamma_i$$

donde Y_i es el valor de y para la $i-$ésima unidad seleccionada en la muestra y γ_i es una variable indicadora (no aleatoria) que vale respectivamente 1 o 0 si la $i-$ésima unidad pertenece o no al dominio, es un estimador insesgado del total τ_d de y para el dominio.
b) Halle la varianza de $\hat{\tau}_d$.
c) Considere la variable y^* que toma el valor y para los elementos del dominio d y 0 en caso contrario y sea σ^2 la varianza de $\mathcal{P}_{y^*}$. Si σ_d^2 es la varianza de y para los elementos del dominio, muestre que aproximadamente $\sigma^2 = \frac{1}{N-1}((N_d-1)\sigma_d^2 + q_d N_d \mu_d^2)$, donde N_d es el tamaño del dominio, μ_d la media de y en el dominio, p_d la proporción de unidades del dominio en la poblaciøn y $p_d = 1 - q_d$.
d) Muestre, en base a c), que si se desea estimar τ_d con un máximo error de estimación e y una confianza del $100(1-\alpha)\,\%$, el tamaño de muestra apropiado viene dado por

$$n = \frac{((N_d-1)\sigma_d^2 + q_d N_d \mu_d^2) z_{1-\frac{\alpha}{2}}^2 N^2}{((N_d-1)\sigma_d^2 + q_d N_d \mu_d^2) z_{1-\frac{\alpha}{2}}^2 N + e^2(N-1)}.$$

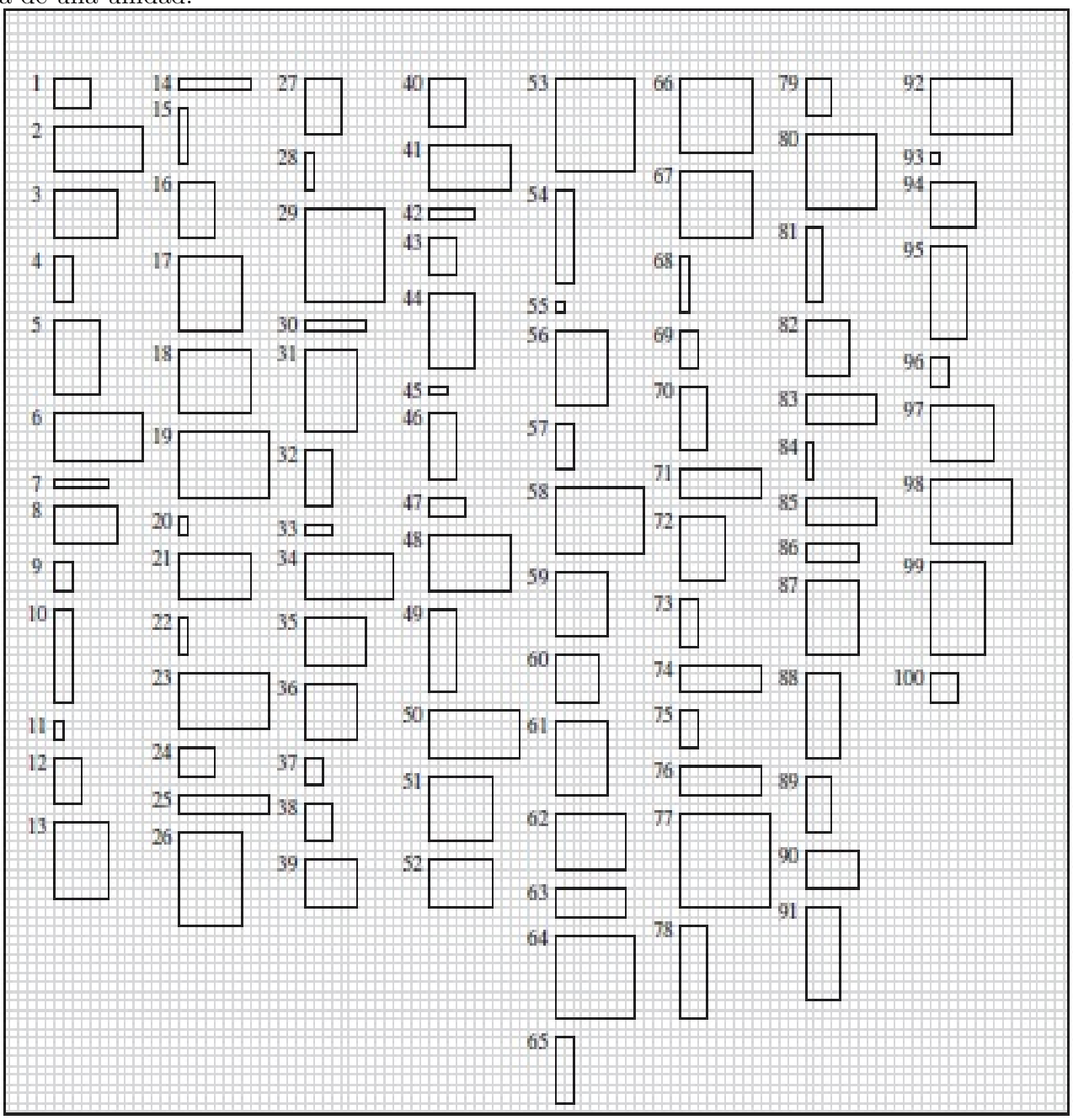
e) Muestre que el tamaño de muestra anterior, en caso se desee obtener un coeficiente de variación de a lo más CV_0 para el total estimado, puede reescribirse como:

$$n = \frac{((N_d-1)CV_d^2 + q_d N_d}{CV_d^2(p_d - \frac{1}{N}) + p_d q_d + (N-1)p_d^2 CV_0^2},$$

donde $CV_d^2 = \frac{\sigma^2}{\mu_d^2}$ denota al cuadrado del coeficiente de variación de y en el dominio.
f) Obtenga el tamaño de muestra que se necesitaría en una encuesta futura para medir el consumo de agua en la zona rural, si es que se deseara estimar τ_d con un margen de error no mayor al millón de litros con una confianza del 95 %. Suponga que en la encuesta se encontró que 60 hogares contaban con servicios de agua y desague y en promedio ellos consumieron en el mes 5,100 litros con una desviación estándar de 380 litros ¿Qué estimación de τ_d le da esta encuesta?

19.- A fin de obtener un MASs que corresponda al 20 % de una población de tamaño 100, un alumno propone el siguiente esquema. Simular 100 números aleatorios en el intervalo $[0, 1]$ y tomar como muestra a las unidades $i \in \mathcal{P} = \{1, 2, \ldots, 100\}$ cuyos correspondientes números aleatorios sean menores o iguales a 0.2 ¿es correcto este esquema de selección? Justifique.

20.- En esta actividad sugerida por Gnanadesikan (1997) se tiene la siguiente figura que contiene 100 rectángulos. El objetivo es estimar el área total de todos los rectángulos tomando una muestra de 10 rectángulos, donde se asume que cada cuadradito de la grilla tiene un área de una unidad.



a) Tome un MASs de 10 rectángulos y obtenga un intervalo de confianza al 95 % para el área total.

b) Replique a) pero con un MASc.

c) Compare su intervalo obtenido con el de sus compañeros e indique el porcentaje de estos que contienen a la verdadera área que es de 3,079 unidades.

21.- Usando la base de datos api, obtenga el tamaño de muestra que se requeriría para estimar el índice api del 2000 de tal manera que se tenga para este un CV del 3 %. Tomada la muestra estime también el total de matriculados y la proporción de colegios por tipo de escuela. Compare finalmente los verdaderos valores (que en un estudio real se desconocen) con las estimaciones encontradas.

22.- Mediante un MASs piloto de tamaño n_1 se ha calculado que el tamaño final de muestra a tomarse para estimar la media de una variable y con un máximo error de estimación de e y una confianza de $100(1-\alpha)\,\%$ es n. Un colega suguiere que en vez de registrarse las n observaciones bastaría tomarse un MASs de tamaño $n-n_1$ de la población no muestreada, pues argumenta que la muestra piloto ya recabo información de y y juntando esta con la última completaría el tamaño n requerido ¿estaría usted de acuerdo con su colega? Justifique.

23.- Suponga que es de interés para usted estimar la media de los costos por crédito que cobran las maestrías de la PUCP y la proporción de estas maestrías que tienen en su plana docente un 50 % o más de Doctores. Para ello decide usted utilizar un MASs.
a) Halle el tamaño de muestra que le permita obtener tales estimaciones con un margen de error no mayor a los 50 soles para los costos y no mayor al 5 % para el porcentaje buscado con un nivel de confianza del 95 %.
b) Tome la muestra requerida en a) y genere la base de datos correspondiente.
c) Obtenga las estimaciones buscadas junto con sus errores estándar de estimación estimados.

24.- Suponga que para un MASs de tamaño n sobre una población de tamaño N se tiene interés en estudiar dos variables estadísticas x e y.
a) Muestre que la covarianza entre las medias muestrales de estas variables viene dada por:

$$Cov(\bar{X},\bar{Y}) = (1-\frac{n}{N})\frac{\sigma_{xy}}{n},$$

donde

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$$

es la covarianza poblacional entre x e y y μ_x y μ_y las medias poblacionales de x e y, respectivamente.
b) Proponga algún estimador insesgado para esta covarianza.

25.- La Internet Movie Database (IMDb) es una base de datos en línea que almacena información relacionada con películas, personal de equipo de producción (incluyendo directores y productores), actores, series de televisión, programas de televisión, videojuegos, actores de doblaje y, más recientemente, personajes ficticios que aparecen en los medios de entretenimiento visual. Recibe más de 100 millones de usuarios únicos al mes y cuenta con una versión móvil. Una de sus secciones, "The IMDb Top 250" es destinada a ser un listado de las 250 películas con mejor calificación, basado en calificaciones de los usuarios registrados del sitio

web. En esta sección cada película aparece con una estrella y un ranking de a lo más 10 puntos. Debajo de este ranking uno puede acceder a las calificaciones otorgadas por los usuarios en forma de un histograma. La intención de este miniproyecto es estimar, con un margen de error de a lo más 0.1 puntos y un nivel de confianza del 95 %, la desviación estándar media (como medida de controversia) de los rankings asignados a estas 250 películas.
a) Halle el tamaño de muestra necesario para este estudio.
b) Tome la muestra respectiva y reporte la estimación pedida junto con su intervalo de confianza o error estándar de estimación estimado.
c) Según sus resultados ¿podría decir que *Whiplash: Música y obsesión (2014)* es una película de calificación controversial?

26.- El sector salud está interesado en saber la estatura promedio de los habitantes de una región particular que cuenta con 700 habitantes. De los datos de los registros de las clínicas de salud de la región, se realizó un MASs con 35 registros de esta población, obteniéndose:

Obs.	Estatura (mts)	Género	Obs.	Estatura (mts)	Género	Obs.	Estatura (mts)	Género
1	1.65	Hombre	13	1.75	Hombre	25	1.53	Mujer
2	1.80	Hombre	14	1.68	Hombre	26	1.65	Mujer
3	1.84	Hombre	15	1.78	Hombre	27	1.70	Mujer
4	1.83	Hombre	16	1.80	Hombre	28	1.70	Mujer
5	1.73	Hombre	17	1.73	Hombre	29	1.58	Mujer
6	1.83	Hombre	18	1.83	Hombre	30	1.75	Mujer
7	1.80	Hombre	19	1.85	Hombre	31	1.70	Mujer
8	1.85	Hombre	20	1.65	Hombre	32	1.73	Mujer
9	1.80	Hombre	21	1.78	Hombre	33	1.73	Mujer
10	1.78	Hombre	22	1.75	Hombre	34	1.57	Mujer
11	1.85	Hombre	23	1.75	Hombre	35	1.70	Mujer
12	1.80	Hombre	24	1.88	Hombre			

a) Estime la media y varianza de las estaturas en esta población, así como también la proporción de mujeres en ella. Puede hacerlo manualmente o con R.
b) ¿Cuál es el error máximo de estimación que se está asumiendo en la estimación de la estaura media para un nivel de confianza del 95 %?
c) Si se hubiese tenido interés en estimar la estatura media de esta población con un margen de error (o error máximo de estimación) de un centímetros a un nivel de confianza del 95 % ¿será suficiente el tamaño de muestra tomado en el estudio?
d) Si en un estudio futuro se deseará estimar la estatura media de esta población de tal manera que se tenga un CV no mayor al 0.5 % ¿cuál sería el tamaño de muestra? ¿Es aquí necesario fijar el nivel de confianza?

27.- Consideremos la siguiente base de datos, que llamaremos Province91, tomada del texto de Lehtonen y Pahkinen (2004). Ella contiene información censal de las 32 municipalidades de una de las 14 provincias (Finlandia central) en las que se dividía el país de Finlandia a finales del año 1991. En ella se registran para cada municipalidad una variable de estratificación (Stratum con 1=Urbano y 2 = Rural), de conglomeración (Cluster formado al juntar 4 municipalidades geográficamente vecinas), de población (POP91), de fuerza laboral o población económicamente activa (LAB), del número de personas desempleadas (UE91) y del número de hogares en base al censo de 1985 (HOU85). La base de datos es:

Stratum	Cluster	Id	Municipality	POP91	LAB91	UE91	HOU85
1	1	1	Jyväskylä	67200	33786	4123	26881
1	2	2	Jämsä	12907	6016	666	4663
1	2	3	Jämsänkoski	8118	3818	528	3019
1	2	4	Keuruu	12707	5919	760	4896
1	3	5	Saarijärvi	10774	4930	721	3730
1	5	6	Suolahti	6159	3022	457	2389
1	3	7	Äänekoski	11595	5823	767	4264
2	5	8	Hankasalmi	6080	2594	391	2179
2	6	9	Joutsa	4594	2069	194	1823
2	7	10	Jyväskmlk	29349	13727	1623	9230
2	4	11	Kannonkoski	1919	821	153	726
2	4	12	Karstula	5594	2521	341	1868
2	8	13	Kinnula	2324	927	129	675
2	8	14	Kivijärvi	1972	819	128	634
2	3	15	Konginkangas	1636	675	142	556
2	5	16	Konnevesi	3453	1557	201	1215
2	1	17	Korpilahti	5181	2144	239	1793
2	2	18	Kuhmoinen	3357	1448	187	1463
2	4	19	Kyyjärvi	1977	831	94	672
2	5	20	Laukaa	16042	7218	874	4952
2	6	21	Leivonmäki	1370	573	61	545
2	6	22	Luhanka	1153	522	54	435
2	7	23	Multia	2375	1059	119	925
2	1	24	Muurame	6830	3024	296	1853
2	7	25	Petäjävesi	3800	1737	262	1352
2	8	26	Pihtipudas	5654	2543	331	1946
2	4	27	Pylkönmäki	1266	545	98	473
2	3	28	Sumiainen	1426	617	79	485
2	1	29	Säynätsalo	3628	1615	166	1226
2	6	30	Toivakka	2499	1084	127	834
2	7	31	Uurainen	3004	1330	219	932
2	8	32	Viitasaari	8641	4011	568	3119

Usando la librería survey de R, realice tanto un MASc como un MASc de tamaño $n = 8$, para estimar la población total de la provincia y el porcentaje o tasa de desempleo en ella. Reporte en ambos casos los errores estándar de estimación. Compare sus estimaciones con las obtenidas en el texto de Lehtonen y Pahkinen (2004).

28.- En el conteo rápido de votos realizado a 1,600 urnas seleccionadas al azar de una gran población se obtuvo que 812 votaron por el candidato opositor, 480 lo hicieron por el candidato de gobierno, 50 votaron en blanco y el resto fueron votos inválidos. Al 95 % de confianza
a) ¿Cuál sería el máximo error de estimación que se cometería en esta encuesta al estimar la proporción de ciudadanos que votan por el candidato opositor?
b) Mediante un intervalo de confianza, indique si podría asegurar, con la confianza dada, de que el candidato opositor ganará las elecciones. Para ello se requiere el 50 % de votos válidos más uno.

29.- En una investigación para estudiar la relación entre la propensión al consumo de alcohol por parte de adolescentes varones y variables como el control parental, regulación emocional y madurez social, se desea tomar un MASs para sólo el distrito de San Miguel. Puesto que la propensión se medirá mediante una proporcion, es de interés estimar esta con un margen de error no mayor a 0.07 y un nivel de confianza del 95 %. Usando en lo posible el paquete survey de R.
a) Halle el tamaño de muestra requerido para este estudio. Para ello y para crear su marco muestral puede hacer uso de la página web del ministerio de educación:

`http://escale.minedu.gob.pe/web/inicio/padron-de-iiee`

la cual contiene información de todos los colegios del país en base al censo nacional escolar del 2016.
b) Tome la muestra anterior y estime en base a ella el total de alumnos matriculados el 2016 en los colegios de varones de San Miguel, así como la proporción de estudiantes que pertenecen a un colegio de gestión privada. En ambos casos obtenga el error de estimación estimado de los estimadores correspondientes.
c) ¿Cree usted que el diseño MASs empleado sea apropiado para los fines de este estudio? Indique si no fuera el caso, qué dificultades acarrea este diseño.

30.- En la subsección 2.4.3 obtuvimos el error estándar de estimación para la diferencia de medias del índice de rendimiento api para los años 1999 y 2000.
a) Tome en esta base de datos un MASs de tamaño $n = 500$ y estime con la libreria survey la diferencia de medias del índice api para estos años.
b) Obtenga, con la librería survey, un intervalo de confianza al 95 % para la diferencia anterior.
c) Con la misma muestra tomada en a) obtenga el IC en b) pero ahora sin usar el paquete survey.

Bibliografía

Cho, E. y Cho, M. (2008). The variance of sample variance from a finite population, *Survey Research Methods Section, American Statistical Association*, Denver, CO.

Cochran, W. (1977). *Sampling techniques*, Wiley Series in Probability and Statistics.

Gnanadesikan, R. (1997). *Statistical data analysis of multivariate observations*, Wiley.

Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population, *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei* **5**: 361–374.

Lehtonen, R. y Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, Jhon Wiley Sons, Ltd.

Lohr, S. (2000). *Muestreo: Diseño y Análisis*, Internacional Thomson editores.

Lumley, T. (2010). *Complex surveys*, Wiley Series in Survey Methodology.

Mendenhall, W., Scheaffer, R. y Ott, L. (2007). *Elementos de muestreo*, Thomson editores.

Tillé, I. (2006). *Sampling Algorithms*, Springer.

Valliant, R., Dever, J. y Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*, Springer.