“Técnicas de Muestreo y Estadísticas Oficiales”

Facultad de Estudios Estadísticos de la UCM

Profesorado
Amador Pacheco J., Cintas del Río M. R.

Técnicas de Muestreo y Estadísticas Oficiales

I.- Aspectos generales:

Población y muestra, marco, tipos de muestreo, falta de respuesta.

II.- Muestreo probabilístico, conceptos básicos:

Unidades de muestreo, probabilidades de inclusión, métodos de selección de unidades, estimadores y errores de muestreo.

III.- Muestreo de unidades elementales con probabilidades iguales:

Muestreo aleatorio simple, muestreo estratificado y estimador de razón.

IV.- Muestreo de conglomerados sin submuestreo y muestreo de conglomerados con submuestreo:

Selección de conglomerados con probabilidades iguales o desiguales.

V.- Errores de muestreo y métodos de estimación:

Conglomerados últimos, semimuestras reiteradas, Jackknife, Bootstrap.

VI.- Determinación de tamaños muestrales.

VII.- Diseño muestral en las encuestas de hogares y económicas.

CAPÍTULO I. Aspectos Generales

CONTENIDOS

I.1.- Introducción

- I.1.1.-Población y Muestra
- I.1.2.- Tipos de muestreo
- I.1.3.- Encuestas por muestreo

I.2.- Esquema general del diseño muestral

- I.2.1.- Ámbito de estudio
- I.2.2.- Marco
- I.2.3.- Evaluación de la calidad de los datos
- I.2.4.- Falta de respuesta

I.1.- Introducción

Operación estadística:

Proceso por el cual se obtiene información estadística

Necesidades de los usuarios →

- ✓ Viabilidad y Objetivos
- ✓ Metodología a seguir
- ✓ Posibles fuentes de datos
- ✓ Costes
- ✓ Otras especificaciones

Ley de la función estadística pública

Regula la actividad estadística para fines estatales y encomienda al **INE** la realización de las operaciones estadísticas de interés nacional: Censos demográficos y económicos, cuentas nacionales, Indicadores económicos y sociales, estadísticas demográficas y sociales,….

Tipos de operaciones estadísticas

CENSOS: Investigaciones de tipo exhaustivo

ENCUESTAS POR MUESTREO: Proceso por el cual se obtienen conclusiones de la población a partir de la información proporcionada por una parte de ella

I.1.1.- Población y Muestra

Objetivo de la Inferencia Estadística:

Extraer conclusiones de la población, observando los datos en una muestra

I.1.1.- Población y Muestra

Población objetivo (Universo)

Conjunto de unidades del que se requiere información.

Unidad de investigación (elemento)

Unidad sobre la que se realiza la medición.

Unidad de observación	Información
Hogares	Gasto medio en alimentación
Personas	Renta per cápita mayores de 16
Empresas	Volumen de ventas

¿Valores verdaderos? **Valores observados**

Sesgos

Errores de tipo sistemático que se cometen en las observaciones (distancia entre valores verdaderos y valores observados).

I.1.1.- Población y Muestra

Muestra

Subconjunto de la población obtenido con el fin de investigar algunas características de la misma.
Los datos obtenidos a partir de ella se denominan **estimaciones**.

Unidad de muestreo

Es la unidad que se utiliza en la selección de la muestra.
No tienen por qué coincidir con las unidades de investigación.
Ejemplo: Estamos interesados en estudiar a los individuos (unidad de investigación) pero sólo se dispone de una lista de viviendas (unidad de muestreo)

Marco de muestreo

Conjunto de unidades de muestreo. En una situación ideal el marco de muestreo debe coincidir con la población objetivo.

I.1.1.- Población y Muestra

Muestras distintas proporcionan valores distintos de las estimaciones

Error de muestreo

Medida de la variabilidad de las estimaciones en torno a su media.

Precisión + Sesgo = **Acuracidad**

Objetivo:

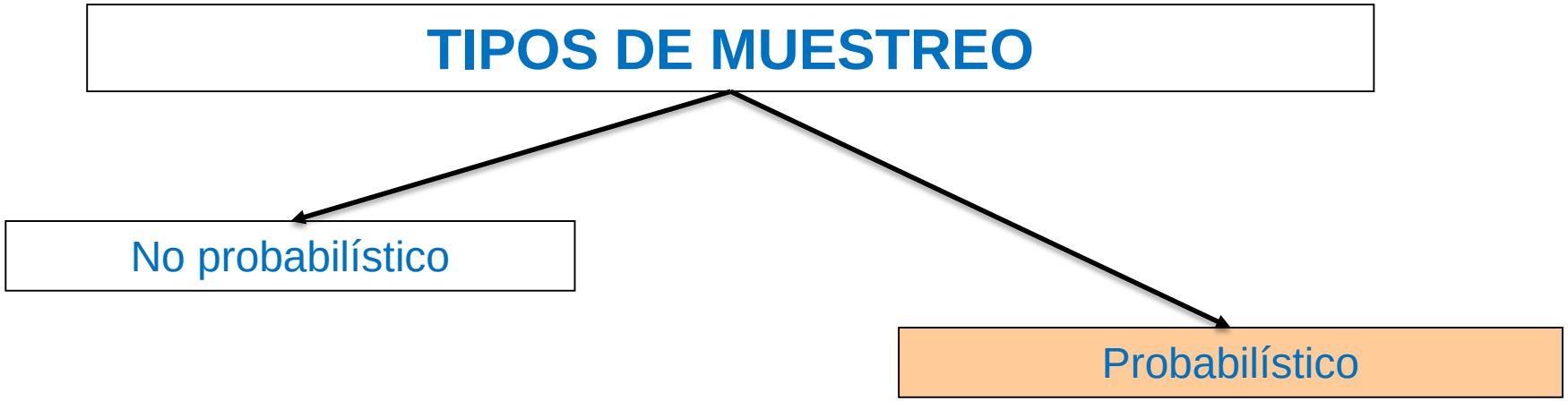
Conseguir muestras **representativas**

Cada unidad muestreada representa las características de una cantidad conocida de unidades de la población

I.1.2.- Tipos de Muestreo

Diseño Muestral o Plan de muestreo

Es el procedimiento por el cual se seleccionan una o más muestras.



I.1.2.- Muestreo NO Probabilístico

✓ La selección de la muestra no está sometida a criterios probabilísticos.

✓ Suele presentar grandes sesgos y es poco fiable; no garantiza la representatividad de la muestra, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos y, por tanto, no permite realizar estimaciones inferenciales sobre la población.

✓ Tiene utilidad en estudios exploratorios en los que el muestreo probabilístico resulta excesivamente costoso, o cuando es difícil enumerar o precisar el universo objeto de estudio o no existen registros de datos.

Ejemplos

-**Por cuotas**: la muestra se selecciona en un número proporcional al de los que cumplen una característica de la población. Por ejemplo, si se conoce que el 20% de la población lo constituyen mujeres y el 80% hombres, se respeta esta proporción en la muestra. (Encuestas de opinión).

-**Intencional u opinático:** el "investigador" escoge la muestra teniendo en cuenta su propia opinión, norma o criterio (selección de unidades tipo).

I.1.2.- Muestreo PROBABILÍSTICO

- Está basado en el Cálculo de Probabilidades y la Estadística Matemática

- Se conoce a priori la probabilidad que tiene cada una de las posibles muestras de ser seleccionada.

- Permite el **cálculo de los errores de muestreo.**

- Es el tipo de muestreo utilizado en los Institutos Nacionales de Estadística para las encuestas oficiales.

I.1.2.- Ejemplos de Muestreo PROBABILÍSTICO

MUESTREO ALEATORIO SIMPLE

La selección de los elementos se hace en una sola etapa, directamente con o sin reemplazamiento. En la práctica equivale a censar o utilizar un censo de la población objeto de estudio, para extraer a continuación, al azar, los elementos que van a formar parte de la muestra.

En este tipo de muestreo todos los elementos tienen la misma probabilidad de ser elegidos para formar parte de la muestra (las muestras son equiprobables).

***Ej.*:** Se quiere extraer una muestra de 1500 elementos de una población formada por los médicos en activo de España. En el Colegio Oficial de Médicos nos proporcionan una lista con todos los médicos que ejercen en España y seleccionamos la muestra utilizando un sorteo por medio de números aleatorios o un bombo, o cualquier otro procedimiento que garantice la aleatoriedad.

I.1.2.- Tipos de Muestreo: PROBABILÍSTICO

MUESTREO ALEATORIO ESTRATIFICADO

- Se divide la población en subpoblaciones denominadas *estratos*.
- Se estratifica de acuerdo a que los elementos de esa subpoblación sean lo más parecidos posible, dado que las poblaciones homogéneas permiten extraer muestras más pequeñas sin que eso implique pérdida de información.
- Seguidamente, se selecciona una muestra de cada estrato, de manera independiente y mediante muestreo aleatorio simple

En este tipo de muestreo cada unidad seleccionada pertenece a un único estrato.

***Ej.*:** De la población de médicos seleccionamos la muestra por especialidades.

Ejemplo

Supongamos un centro escolar con 600 alumnos que se reparten de la siguiente manera:

100 alumnos de 1º de ESO
100 alumnos de 2º de ESO
100 alumnos de 3º de ESO
120 alumnos de 4º de ESO
100 alumnos de 1º de Bachillerato
80 alumnos de 2º de Bachillerato

La dirección necesita conocer urgentemente la opinión del alumnado sobre un tema concreto y no puede (o no quiere) preguntarlo a todo el mundo. Se decide recabar las opiniones de 60 de los chicos.

"Malas maneras" de hacerlo

- Preguntar a 60 alumnos de 2º de Bachillerato
- Preguntar a 60 chicas
- Recoger la opinión de los 60 primeros que se presenten de forma voluntaria

Muestreo **No** Probabilístico

Ejemplo

100 alumnos de 1º de ESO
100 alumnos de 2º de ESO
100 alumnos de 3º de ESO
120 alumnos de 4º de ESO
100 alumnos de 1º de Bachillerato
80 alumnos de 2º de Bachillerato

Formas correctas de hacerlo

- Tomar una lista numerada de todos los alumnos del centro y realizar un sorteo de 60 números y preguntar a quién corresponden esos números.

- Tomar la lista numerada, ordenarla por curso y sortear 10 alumnos de cada curso de ESO, 10 de 1º de Bachillerato y 8 de 2º de Bachillerato.

Muestreo Probabilístico

I.1.3.- Encuestas por Muestreo

VENTAJAS DEL MUESTREO FRENTE AL CENSO

Menor coste asociado, tanto temporal como económico

Resultados rápidos

Mayor facilidad a la hora de controlar el error ajeno al muestreo

Menos limitaciones en las características a investigar

I.1.3.- Encuestas por Muestreo

Aún así, el censo es necesario……

Proporciona una gran cantidad de información a un nivel muy elevado de desagregación

Complementa a las encuestas por muestreo con información necesaria para:

- Preparación de las bases de muestreo (MARCOS)
- Procesos de ESTRATIFICACIÓN
- Procesos de ESTIMACIÓN

I.1.3.- Encuestas por Muestreo: Etapas

Planteamiento de objetivos: Claridad y delimitación de lo que se pretende analizar. Definir claramente la población y que elementos pertenecen a ella. Definir el marco, características a estimar, medidas de precisión que se van a utilizar y un modelo de tablas de resultados.

Plan de muestreo: Describir las distintas técnicas a utilizar para seleccionar la muestra, estimadores y nivele de precisión.

Trabajo de campo: Elaboración de encuestas, establecimiento del método de recogida de datos, selección y formación de entrevistadores. Prueba del muestreo mediante una encuesta piloto que permita detectar problemas del proceso.

Tratamiento de la información: Operaciones de depuración e imputación a que se someten los datos con el fin de obtener un fichero de datos completo y consistente (tratamiento de la no respuesta).

Difusión de resultados: Presentación de los resultados y elaboración del informe final, en el que se incluirá una descripción de la metodología, conceptos, variables y clasificaciones utilizadas.

I.2.1.- Ámbito de estudio

Ámbito poblacional: Se refiere a la población objeto de estudio

Ej: El ámbito poblacional en la encuesta industrial es el conjunto de empresas con una o más personas remuneradas y cuya actividad principal está incluida en las secciones B a E de la CNAE-09

Ámbito geográfico: Es el territorio abarcado por el objetivo de la encuesta

Ejs: Provincias, Comunidades autónomas o Total nacional

Ámbito temporal: Tiene un doble aspecto: el de referencia de la encuesta y el de referencia de la toma de datos

I.2.2.- Marco

- Lo constituye toda la información útil disponible sobre las unidades de la población objeto de estudio (listas, ficheros planos, etc.), en cualquier etapa del diseño muestral.
- El marco, en sentido estricto, es la **lista de unidades de muestreo** y debe ser un fiel reflejo de la población objetivo: la construcción de un marco muestral lo más perfecto posible es importante a fin de que exista una correspondencia biunívoca entre las unidades muestrales poblacionales y las listas físicas que lo conforman.
- Se pueden utilizar **marcos de áreas o marcos de listas** (muestreos en una o varias etapas).
- La formación del marco puede afectar de manera importante al coste de la encuesta. Generalmente se recurre a formar marcos a partir de otras fuentes ya existentes.
- Existe, también, **información complementaria** que puede utilizarse para la mejora del diseño muestral, tanto en los procesos de estratificación, estimación, ajuste de falta de respuesta, etc..

I.2.3.- Evaluación de la calidad de los datos

Procedimiento: Medición de los principales tipos de error

ERRORES DE MUESTREO

Debidos a la estimación de las características poblacionales a partir de la muestra

Se calculan mediante procedimientos directos e indirectos.

Permiten obtener el intervalo de confianza que contiene al verdadero valor del parámetro con una probabilidad prefijada.

ERRORES AJENOS AL MUESTREO

Se presentan en cualquiera de las etapas del proceso. Introducen sesgos en las estimaciones difíciles de cuantificar.

Ej: La falta de respuesta

Su evaluación es generalmente costosa y difícil de llevar a la práctica

I.2.4.- La falta de respuesta

Efecto inmediato: Tamaño muestral obtenido inferior al tamaño muestral efectivo

$(u_1,....,u_n)$

$(y_1,....,y_k)$

Falta de respuesta total:

Una o más unidades muestrales no pueden ser observadas

Falta de respuesta parcial:

En una o más unidades sólo son observadas h<k variables

Incidencias que dan lugar a la falta de respuesta

• Unidades **no encuestables**: son unidades seleccionadas para la muestra que no pertenecen a la población objetivo. Son debidas a errores en los marcos.

• Unidades **ausentes o no contactadas**: son aquellas que pertenecen a la población objetivo pero con las que no se ha podido contactar. Su existencia depende, en cierta medida, de la organización del trabajo de campo.

• Unidades **negativas a contestar**: son aquellas que rechazan colaborar. Pueden ser negativas en el contacto inicial o posteriormente.

• Unidades **incapaces de contestar**: son aquellas que por causas diversas (enfermedad, desconocimiento del idioma) no colaboran.

I.2.4.- La falta de respuesta

La existencia de estos tipos de unidades da lugar a…

- **Presencia de sesgos** en las estimaciones, por no ser aleatoria la muestra de unidades que no responden.
- **Incremento de la varianza**, por producir disminución en el tamaño de la muestra.
- **Incremento del coste**. Es necesario incrementar la muestra para mantener los niveles de precisión exigidos.

Posibles soluciones

- **Incremento del tamaño muestral teórico** para que el tamaño muestral efectivo cumpla las expectativas. No elimina los sesgos.
- **Uso de información auxiliar**, que permita reducir el sesgo aplicando técnicas de calibrado.

CAPÍTULO II. Muestreo Probabilístico: conceptos básicos

CONTENIDOS

II.1.- Unidades de muestreo

II.2.- Probabilidades de inclusión

II.3.- Concepto de estimador y error de muestreo

II.4.- Métodos de selección de unidades

II.5.- Estimadores lineales insesgados

II.6.- Tipos de muestreo

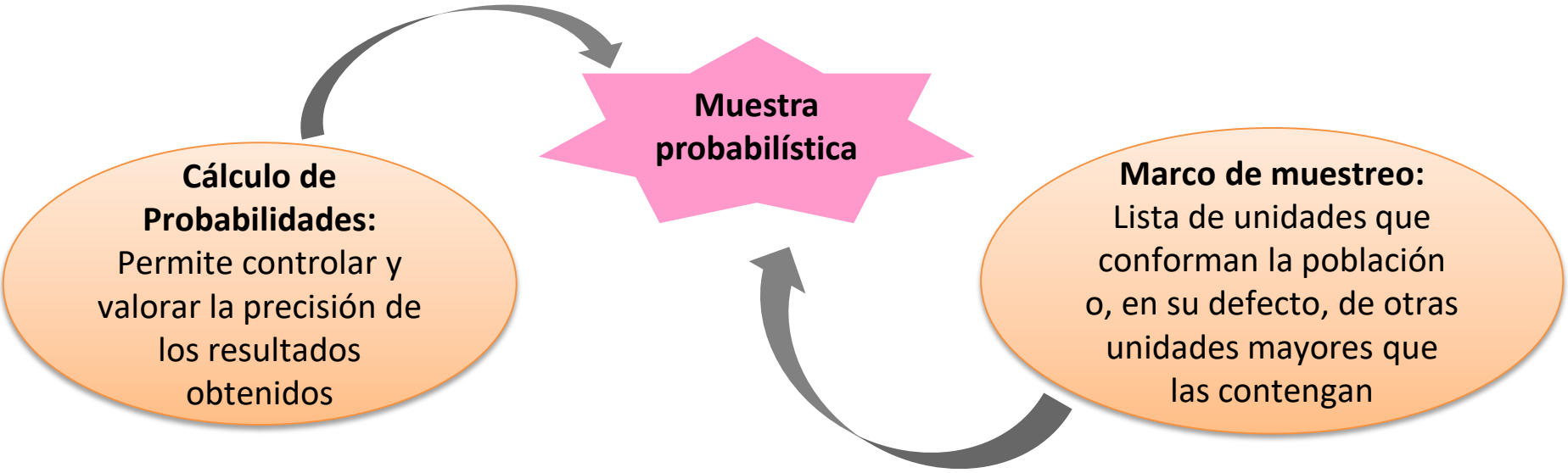
II.1.- Unidades de Muestreo

POBLACIÓN A ESTUDIO HOMOGÉNEA → **Muestra representativa**

Problema: La población a muestrear NO ES HOMOGÉNEA

Ejs: poblaciones humanas, viviendas, explotaciones agrícolas, industrias…

¿Cómo conseguir muestras representativas?



Unidad de muestreo

Es la unidad que se utiliza en la selección de la muestra (no tienen por qué coincidir con las unidades de investigación).

Marco de muestreo

Conjunto de unidades de muestreo. En una situación ideal el marco de muestreo debe coincidir con la población objetivo.

Unidades elementales:

Son las unidades de las que tratamos de obtener información

Son las unidades últimas en el proceso de selección

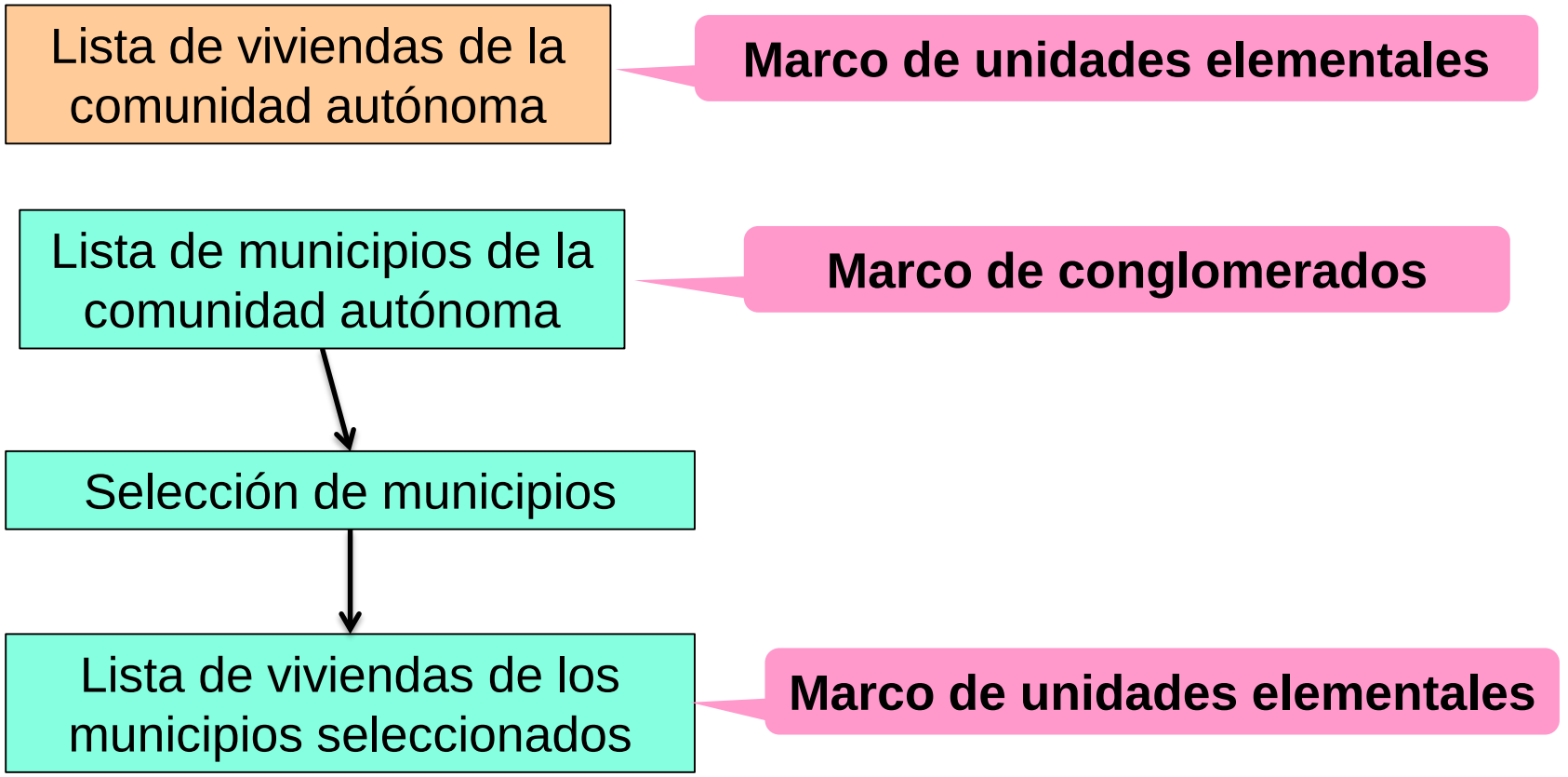
Coinciden con las unidades de investigación

Conglomerados:

Grupos de unidades elementales

Ejemplo 2.1

Supongamos que estamos interesados en estudiar el precio de alquiler de la vivienda en una determinada comunidad autónoma. Podemos seleccionar la muestra probabilística de viviendas a partir de:



Ejemplo 2.2

En un fichero se dispone información de trabajadores que alguna vez han recibido algún curso de formación. Se desea conocer la edad media de estos trabajadores

POBLACIÓN: personas recogidas en el fichero

UNIDAD DE OBSERVACIÓN: cada una de estas personas

UNIDAD DE MUESTREO: cada una de estas personas

MUESTRA: subconjunto de ellas elegidas al azar

MEDICIÓN : edad del trabajador

Ejemplo 2.3

Se desea conocer la proporción de niños de cierta ciudad inmunizados del sarampión

POBLACIÓN: niños escolarizados

UNIDAD ELEMENTAL: cada uno de estos niños

UNIDAD DE MUESTREO: colegios

MUESTRA: niños de los colegios elegidos al azar

MEDICIÓN : estado inmunológico del niño:

1: Sí está inmunizado

0: No está inmunizado

NOTACIÓN Y DEFINICIONES

U=Población card(U)=N N=Tamaño poblacional (Finito)

Unidades del marco $U = \{u_1, u_2, \dots, u_N\}$

X=característica a estudiar $\{X_1, X_2 \dots, X_N\}$

S= Espacio muestral (**Conjunto de todas las posibles muestras conocido**)

Muestra= cualquier subconjunto de U

Nº de posibles muestras $S_U = 2^N$ $S \subset S_U$

$s \in S$ $s = \{u_1, u_2, \dots, u_n\}$

n=Tamaño muestral

Cada muestra posible *s* tiene asignada una **probabilidad *p(s)* conocida de selección**

Ejemplo 2.4

Supongamos que el marco tiene tres unidades $U = \{u_1, u_2, u_3\}$
Seleccionamos una muestra de tamaño 2 con probabilidades iguales, sin unidades repetidas y considerando que el orden en que seleccionamos las unidades no es importante.

¿Cuántas muestras posibles tenemos?

Muestras posibles
$s_1=(u_1,u_2)$
$s_2=(u_1,u_3)$
$s_3=(u_2,u_3)$

Todas las muestras tienen probabilidad 1/3 de ser seleccionadas

Este tipo de muestreo es probabilístico

Y en la práctica ¿Cómo lo hacemos?

Sortear un nº entre 1 y 3 para elegir la muestra

Es necesario construir todas las muestras posibles para numerarlas e identificar la que corresponde al nº que salga en el sorteo.

Impracticable en poblaciones grandes

Solución:

Sortear con probabilidades iguales las unidades de la población que aún no han sido elegidas antes de cada extracción

$p(s_1) = p(u_1,u_2)+p(u_2,u_1) = (1/3)(1/2)+(1/3)(1/2) = 1/3$
$p(s_2) = p(u_1,u_3)+p(u_3,u_1) = (1/3)(1/2)+(1/3)(1/2) = 1/3$
$p(s_3) = p(u_2,u_3)+p(u_3,u_2) = (1/3)(1/2)+(1/3)(1/2) = 1/3$

II.2.- Probabilidades de Inclusión

Probabilidad de inclusión de primer orden π_i

Probabilidad de cada unidad del marco de aparecer en una muestra.
(Es conocida puesto que cada muestra s tiene una probabilidad conocida de ser elegida)

$$\pi_i = p(unidad\ i\ en\ la\ muestra) = \sum_{u_i \epsilon s} p(s)$$

$$\pi_i > 0 \quad \forall i = 1, \dots, N$$

Probabilidad de inclusión de segundo orden π_{ij}

Probabilidad que tiene el par (u_i, u_j)de aparecer en una muestra.

$$\pi_{ij} = p(unidad\ i\ \text{y}\ unidad\ \text{j}\ \ en\ la\ muestra) = \sum_{\substack{u_i \epsilon s \\ u_j \epsilon s}} p(s)$$

$$\pi_{ij} = \pi_{ji}$$

Ejemplo 2.4 (cont.)

Supongamos que el marco tiene tres unidades $U = \{u_1, u_2, u_3\}$
Seleccionamos una muestra de tamaño 2 con probabilidades iguales, sin unidades repetidas y considerando que el orden en que seleccionamos las unidades no es importante.

$$\pi_1 = p(s_1) + p(s_2) = 2/3$$
$$\pi_2 = p(s_1) + p(s_3) = 2/3$$
$$\pi_3 = p(s_2) + p(s_3) = 2/3$$

Probabilidades de inclusión de primer orden

$$\pi_{12} = p(s_1) = 1/3$$
$$\pi_{13} = p(s_2) = 1/3$$
$$\pi_{23} = p(s_3) = 1/3$$

Probabilidades de inclusión de segundo orden

II.3.- Concepto de estimador

Las unidades poseen muchas características: edad, nº de hermanos, si es zurdo o no, estado civil...

Características Numéricas:

Edad

Nº hermanos

Características categóricas:

Zurdo, diestro, ambidiestro

Estado civil

$\{X_1, X_2, \ldots, X_N\}$ la característica (numérica o categórica) de todos los elementos de la población

Interesa conocer la edad media, la proporción de casados, el porcentaje de zurdos...

PARÁMETROS O CARACTERÍSTICAS POBLACIONALES

Estimador

Expresión matemática que permite inferir las características de la población a partir de la muestra.

El valor que toma el estimador en una determinada muestra se conoce como **estimación**.

Ejemplo 2.4 (cont.)

Supongamos ahora que conocemos los valores de una determinada característica X en las N=3 unidades de nuestra población:

$$X_1 = X_2 = 2 \quad X_3 = 5$$

Valor medio poblacional $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ (Cantidad a estimar)

Valor medio muestral $\bar{x} = \frac{1}{n}\sum_{i \in s} X_i$ (Estimador)

Muestras posibles	Valores	Valor medio muestral ($\overline{x}$)	p(s)
$s_1=(u_1,u_2)$	(2,2)	2	1/3
$s_2=(u_1,u_3)$	(2,5)	7/2=3,5	1/3
$s_3=(u_2,u_3)$	(2,5)	7/2=3,5	1/3

Valor de la estimación si la muestra seleccionada es s_3

Variable aleatoria discreta que toma dos valores

Ninguno de ellos coincide con el valor poblacional a estimar:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{2+2+5}{3} = 3 \neq$$

Distribución del estimador en el muestreo

$\overline{x}$	probabilidad
2	1/3
3,5	2/3

II.3.- Concepto de Error de Muestreo

Siempre que utilizamos una muestra para estimar datos de una población, cometemos un error.
Este error es propio del muestreo y no existe en los censos

Como hemos visto, el estimador es una v.a. cuyos valores particulares son las estimaciones, por lo que tiene sentido hablar de:

Esperanza del estimador

$$E[\overline{x}] = \sum_{s} p(s)\overline{x}$$

Media ponderada de todas las estimaciones posibles, donde el peso es la probabilidad de que aparezca ese valor particular.
En nuestro ejemplo:

$$E[\bar{x}] = 2\frac{1}{3} + 3{,}5\frac{2}{3} = 3 = \overline{X}$$

Varianza del estimador

$$V[\bar{x}] = \sum_{s} p(s)(\bar{x} - E[\bar{x}])^2$$

Medida del ***grado de dispersión de las estimaciones alrededor de su media o esperanza***. Se calcula como la media ponderada de las desviaciones al cuadrado entre las estimaciones y su media.

En nuestro ejemplo:

$$V[\bar{x}] = (2-3)^2 \frac{1}{3} + (3{,}5-3)^2 \frac{2}{3} = 0{,}5$$

En general, si mediante un estimador concreto pretendemos estimar un valor poblacional, puede ocurrir que:

E(estimador) = Valor poblacional

Decimos que el estimador es **insesgado** para ese valor.

En nuestro ejemplo la media muestral ha resultado insesgada para la media poblacional

E(estimador) ≠ Valor poblacional

Decimos que el estimador es **sesgado** para ese valor.

Podemos calcular el **sesgo** del estimador como la diferencia entre la esperanza del estimador y el valor poblacional que trata de estimar:

$$B(\overline{x}) = E(\overline{x}) - \overline{X}$$

Error cuadrático medio del estimador

$$ECM(\overline{x}) = \sum_{s} p(s)(\overline{x} - \overline{X})^2$$

Medida del ***grado de dispersión de las estimaciones alrededor del valor poblacional a estimar***. Se calcula como la media ponderada de las desviaciones al cuadrado entre las estimaciones y el valor poblacional.

Se puede desglosar en dos componentes:

$$ECM(\overline{x}) = V(\overline{x}) + B(\overline{x})^2$$

Si el estimador es insesgado, el ECM del estimador viene dado sólo por la varianza

Error de muestreo del estimador

$$EM(\overline{x}) = +\sqrt{V(\overline{x})}$$

Se calcula como la raíz cuadrada de la varianza del estimador. De esta forma el error se expresa en las mismas unidades de los datos.

En nuestro ejemplo: $\sqrt{0,5} = 0,707$

Error relativo de muestreo del estimador

$$ERM(\overline{x}) = \frac{+\sqrt{V(\overline{x})}}{\overline{X}}$$

En nuestro ejemplo: $0,707/3 = 0,2357 \quad (23,57\%)$

OBSERVACIÓN

En la práctica el error de un estimador insesgado, tanto absoluto como relativo, se estima a partir de los datos proporcionados por la muestra

$$\widehat{EM}(\overline{x}) = +\sqrt{\widehat{V}(\overline{x})}$$

$$\widehat{ERM}(\overline{x}) = \frac{+\sqrt{\widehat{V}(\overline{x})}}{\overline{x}}$$

Ejemplo 2.5

En una ciudad hay seis colegios y se quiere estimar el nº medio de alumnos por colegio. Si se conociera que el nº de alumnos que tiene cada uno es

Colegio	1	2	3	4	5	6
Nº alumnos	59	28	90	44	36	57

$$\overline{X} = \frac{59+28+90+44+36+57}{6} = 52.33$$

X_i: nº alumnos en el colegio *i*

Supongamos que se toma una muestra de **dos** colegios. Con los valores de esos dos colegios, se estimaría calculando la $\overline{X}$

MEDIA MUESTRAL

$$\overline{x} = \frac{x_1 + x_2}{2}$$

Si se eligiera la muestra...	La estimación del nº medio de alumnos sería...
1,2	(59+28)/2= 43.5
1,3	(59+90)/2= 74.5
1,4	(59+44)/2= 51.5
1,5	(59+36)/2= 47.5
1,6	(59+57)/2= 58
2,3	(28+90)/2= 59
2,4	(28+44)/2= 36
2,5	(28+36)/2= 32
2,6	(28+57)/2= 42.5
3,4	(90+44)/2= 67
3,5	(90+36)/2= 63
3,6	(90+57)/2= 73.5
4,5	(44+36)/2= 40
4,6	(44+57)/2= 50.5
5,6	(36+57)/2= 46.5

¿Cuántas muestras hay?

Como en la práctica sólo se toma una muestra y la estimación es muy distinta dependiendo de cuál sea ésta, será necesario dar, junto a la estimación, un valor de la variación o dispersión que dé idea de su validez o exactitud

VARIANZA DEL ESTIMADOR

y

ERROR DE MUESTREO

$$V(\bar{x})$$

$$EM(\bar{x}) = +\sqrt{V(\bar{x})}$$

Sus valores van a depender de la varianza poblacional y, por lo tanto, también serán desconocidos. Así habrá que estimarlos a partir de la muestra.

En el ejemplo...

$$V(\bar{x}) = \frac{(43.5-52.3)^2 + \ldots + (46.5-52.3)^2}{15} = 160.889$$

$$EM(\bar{x}) = +\sqrt{160.889} = 12.68 \quad alumnos$$

Para poder calcular estos valores se debe disponer de todas las muestras, cosa que no ocurre. Se estimarán a partir de una única muestra. El estimador dependerá del tipo de muestreo que se utilice.

$$Media\ de\ las\ \bar{x} = E(\bar{x}) = \frac{43.5+74.5+\ldots+50.5+46.5}{15} = \frac{785}{15} = 52.3 = \bar{X}$$

El valor medio de las estimaciones es el parámetro que se quiere estimar: ESTIMACIÓN INSESGADA

En el ejemplo...

Supongamos que eligiendo una de las 15 muestras al azar, se hubiera obtenido la muestra nº 10, formada por los colegios 3 y 4.

Estimación del nº medio de alumnos por colegio:

$$\bar{x} = \frac{90+44}{2} = 67$$

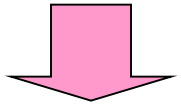
Estimación de la varianza: En este tipo de muestreo la expresión del estimador es

$$\hat{V}(\bar{x}) = (1-\frac{n}{N})\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n(n-1)} = (1-\frac{2}{6})\frac{(90-67)^2+(44-67)^2}{2(2-1)} = \frac{1058}{3} = 352.6$$

Estimación del error de muestreo

$$\hat{EM} = +\sqrt{352.6} = 18.77$$

Será importante dar la estimación del parámetro con un margen de error: **error máximo admisible**. Esto se hará calculando un intervalo de confianza que contendrá al parámetro con una seguridad o confianza establecida



INTERVALO DE CONFIANZA:

(Estimador±error máximo admisible)

Este error máximo admisible dependerá del:

- Error de muestreo
- La confianza o seguridad que fijemos

Intervalo de confianza para el valor estimado

$$IC(Parámetro) = \left[Estimador - 2\widehat{EM}\ , Estimador + 2\widehat{EM}\ \right]$$

El intervalo así construido ***en base a la muestra seleccionada*** contiene el verdadero valor con una probabilidad de más del 95% (si el estimador se distribuye como una normal y la muestra es suficientemente grande).

Cuanto menos amplitud tenga el intervalo, mejor será la estimación

En el ejemplo de los colegios…

elijamos un procedimiento arbitrario para calcular un intervalo de confianza para el nº medio de alumnos por colegio.

IC=(Estimador-2ÊM, Estimador+2ÊM)

Dependiendo de la muestra, se obtendrá un intervalo u otro

¿Qué confianza tiene este tipo de intervalo?

Si se eligiera la muestra	La estimación del nº medio de alumnos sería $\bar{x} = \frac{x_1 + x_2}{2}$	La estimación del error de muestreo sería $E\hat{M} = \sqrt{(1-\frac{2}{6})\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2}{2(2-1)}}$	El intervalo de confianza sería $(\bar{x}-2E\hat{M}, \bar{x}+2\hat{EM})$
1,2	43.5	12.65	(18.2,68.8)
1,3	74.5	12.65	(49.2,99.8)
1,4	51.5	6.12	(39.2,63.7)
1,5	47.5	9.38	(28.7,66.2)
1,6	58	0.81	(56.38,59.62)
2,3	59	25.31	(8.38,109.62)
2,4	36	6.53	(22.94,49.06)
2,5	32	3.26	(25.48,38.52)
2,6	42.5	11.83	(18.84,66.16)
3,4	67	18.77	(29.46,104.54)
3,5	63	22.04	(18.92,107.08)
3,6	73.5	13.47	(46.56,100.44)
4,5	40	3.26	(33.48,46.52)
4,6	50.5	5.30	(39.9,61.1)
5,6	46.5	8.57	(29.36,63.64)

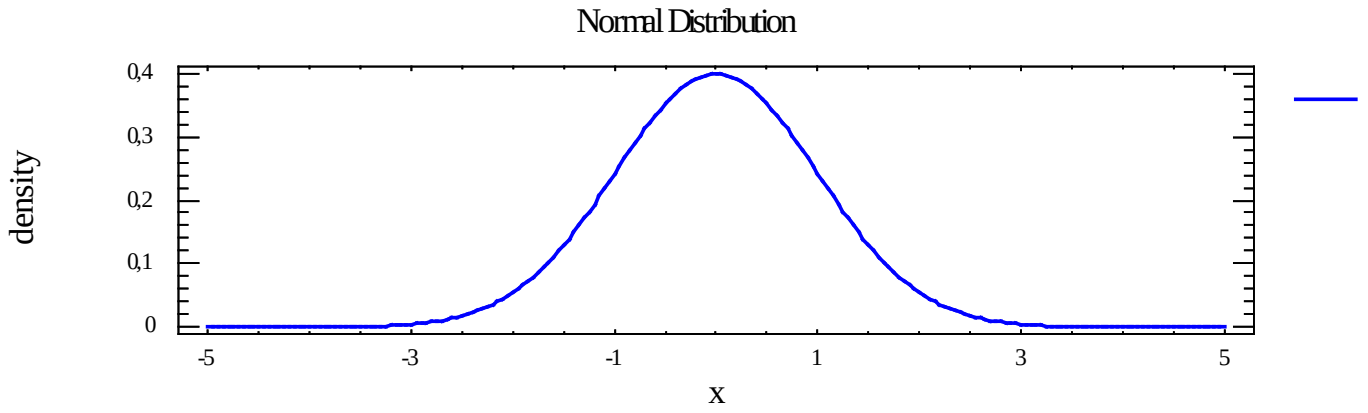
En 11 de los 15 intervalos estaría el verdadero nº medio de alumnos 52.3

El intervalo (Estimador-2ÊM,Estimador+2ÊM)

es de una confianza del 11/15%=73.3%

Como en la práctica sólo se toma una muestra, el intervalo con una confianza fija dependerá de la distribución de todas las posibles estimaciones.

Si dibujáramos todas las posibles medias obtenidas con todas las posibles muestras, veríamos que se ajustan a una curva en forma de campana: Es la llamada campana de Gauss o distribución normal.

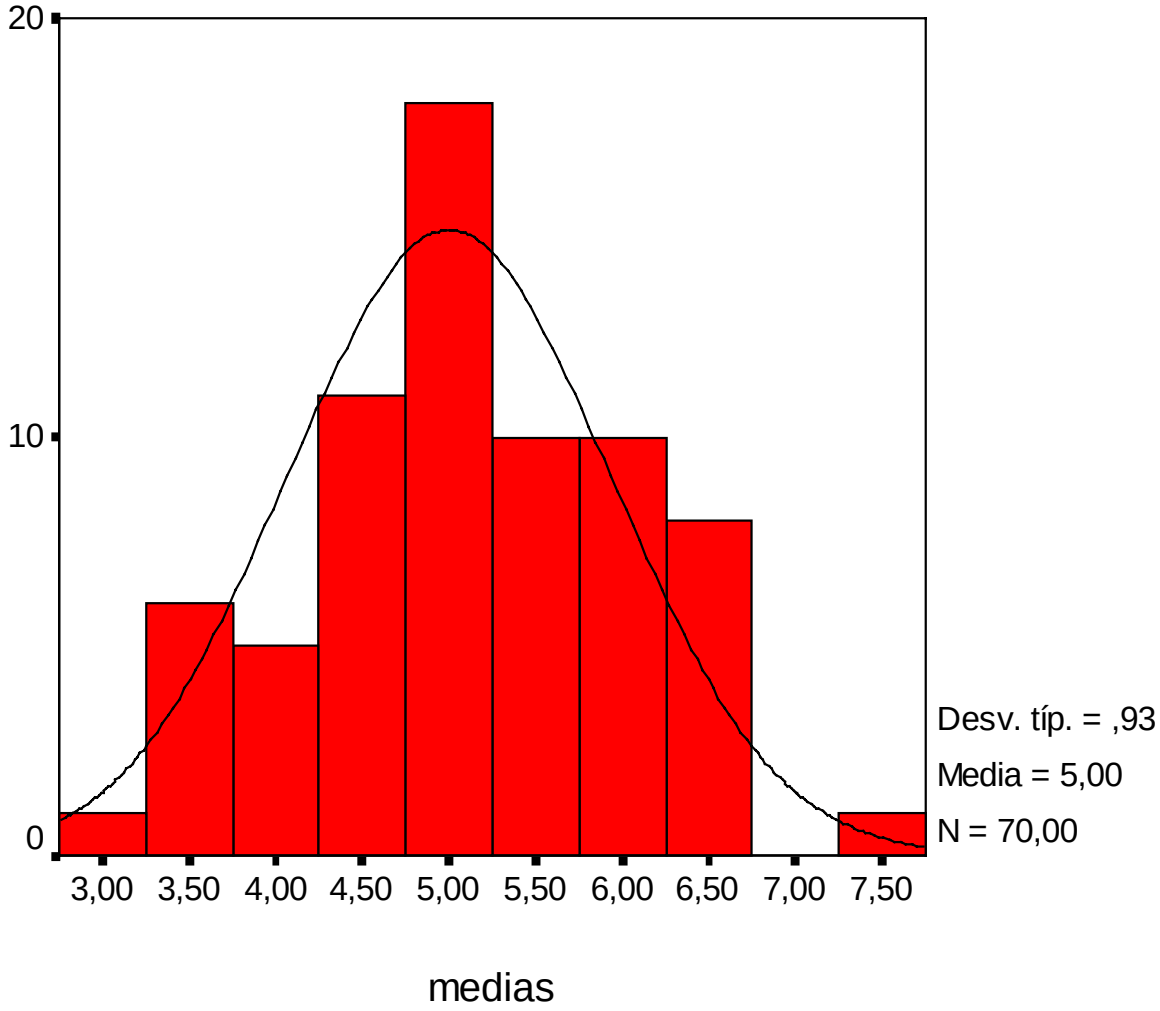


Ejemplo 2.6

Partimos de una población de 8 unidades con los siguientes valores de X

Unidad	1	2	3	4	5	6	7	8
X	1	2	4	4	7	7	7	8

Si construimos todas las muestras de cuatro unidades (70 muestras), calculamos las medias muestrales y las dibujamos, el aspecto del gráfico es el siguiente



La media muestral se ajusta bien a una distribución normal

Se pueden calcular intervalos de confianza de un (1-α)% de confianza, basándose en valores de esta distribución normal a los que denotamos con $z_{\alpha/2}$

$$(estimador - z_{\alpha/2}\hat{EM}(estimador), estimador + z_{\alpha/2}\hat{EM}(estimador))$$

Valores z para construir intervalos con distintos niveles de confianza

Confianza (%)	50	80	90	95	99
Valor z	0.67	1.28	1.64	1.96	2.58

II.4.- Métodos de Selección de Unidades

MUESTREO SIN REEMPLAZAMIENTO (SR)

- Las unidades seleccionadas no se reponen a la población y por lo tanto no pueden ser de nuevo seleccionadas en extracciones sucesivas. ***Las muestras así obtenidas tienen todos sus elementos distintos*****.**
- La estructura de la población no es constante a lo largo del proceso de selección de la muestra lo que hace que las probabilidades de selección de las unidades varíen.
- Las extracciones sucesivas hasta completar la muestra no son independientes.
- Es necesario únicamente conocer las probabilidades de inclusión, tanto de primer como de segundo orden, para obtener los estimadores (este cálculo puede ser complejo dependiendo del tamaño de la muestra y del diseño muestral utilizado).
- El estimador usual en este tipo de muestreo es el de **Horvitz-Thompson**.

MUESTREO CON REEMPLAZAMIENTO (CR)

- Las unidades seleccionadas se reponen a la población y por lo tanto pueden de nuevo ser seleccionadas en extracciones sucesivas. **Pueden existir muestras con elementos repetidos.**
- La estructura de la población permanece constante a lo largo del proceso de selección de la muestra lo que hace que las probabilidades de selección de las unidades no varíen.
- Las extracciones sucesivas hasta completar la muestra son independientes.
- Es necesario conocer las probabilidades de selección tanto de primer como de segundo orden (los cálculos son mucho más sencillos que en el caso SR).
- El estimador usual en este tipo de muestreo es el de **Hansen-Hurwitz**.

Número de muestras posibles según el método de selección de unidades

	Sin orden	Con orden
Muestreo sin reposición (SR)	$C_{N,n} = \binom{N}{n}$	$V_{N,n} = N \cdot n! = \frac{N!}{(N-n)!}$
Muestreo con reposición (CR)	$CR_{N,n} = \binom{N+n-1}{n}$	$VR_{N,n} = N^n$

II.5.- Estimadores Lineales Insesgados

CARACTERÍSTICAS CUANTITATIVAS
(se pueden medir y a cada unidad se le asigna un número)

	Valor Poblacional	Estimador
Total	$X = X_1 + \cdots + X_N = \sum_{i=1}^{N} X_i$	$\hat{X} = \sum_{i=1}^{n} w_i x_i$
Media	$\bar{X} = \frac{X_1 + \cdots + X_N}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i$	$\hat{\bar{X}} = \frac{\hat{X}}{N}$
Razón	$R = \frac{X}{Y} = \frac{\bar{X}}{\bar{Y}}$	$\hat{R} = \frac{\hat{X}}{\hat{Y}} = \frac{\hat{\bar{X}}}{\hat{\bar{Y}}}$

¡Ojo! No es lineal ni insesgado

CARACTERÍSTICAS CUALITATIVAS
(a cada unidad se le asigna una cualidad)

Para cada unidad de la población se define:

$$X_i = \begin{cases} 1 & si\ u_i\ posee\ la\ cualidad\ de\ interés \\ 0 & en\ caso\ contrario \end{cases}$$

	Valor Poblacional	Estimador
Total de Clase	$A = X_1 + \cdots + X_N = \sum_{i=1}^{N} X_i$	$\widehat{A} = \sum_{i=1}^{n} w_i x_i$
Proporción	$P = \frac{A}{N} = \overline{X}$	$\widehat{P} = \frac{\widehat{A}}{N} = \widehat{\overline{X}}$
Tasa	$T = \frac{A}{A'} = \frac{P}{P'}$	$\widehat{T} = \frac{\widehat{A}}{\widehat{A}'} = \frac{\widehat{P}}{\widehat{P}'}$

¡Ojo! No es lineal ni insesgado

NOTACIÓN Y DEFINICIONES

PARÁMETROS POBLACIONALES (X cuantitativa)

Media poblacional $\mu = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Total poblacional $X = \sum_{i=1}^{N} X_i = N\bar{X}$

Varianza poblacional $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$

Cuasivarianza poblacional $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$

Coeficiente de variación $CV(X) = \frac{\sigma}{\bar{X}}$

Razón poblacional $R = \frac{\bar{X}}{\bar{Y}} = \frac{X}{Y}$ con $Y:\{Y_1, \ldots, Y_N\}$

Covarianza poblacional $\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$

Cuasicovarianza poblacional $S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$

Coeficiente de correlación $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{S_{XY}}{S_X S_Y}$

NOTACIÓN Y DEFINICIONES

PARÁMETROS POBLACIONALES (C atributo)

Asociado a C se define $X_i = \begin{cases} 1 & si\ u_i \epsilon C \\ 0 & si\ u_i \notin C \end{cases} \qquad C = \{X_1, \dots, X_N\}$

Proporción poblacional
(Proporción de unidades de la población que poseen la característica C)

$$\mathrm{P} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Total de clase poblacional
(Total de unidades en la población que pertenecen a C)

$$A = \sum_{i=1}^{N} X_i = \mathrm{X}$$

Varianza poblacional

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^2 = \frac{1}{\mathrm{N}}\left(\sum \mathrm{X_i^2} - \mathrm{N}\bar{\mathrm{X}}^2\right) = \frac{1}{\mathrm{N}}\left(\sum \mathrm{X_i} - \mathrm{NP}^2\right)$$
$$= \mathrm{P} - \mathrm{P}^2 = \mathrm{PQ}$$

Cuasivarianza poblacional $\qquad S^2 = \frac{N}{N-1}\mathrm{PQ}$

II.6.- Tipos de Muestreo

	TIPOS DE MUESTREO	
Método de selección de unidades	SR CR	
Unidades de muestreo utilizadas	Muestreo de unidades elementales	
	Muestreo de Conglomerados	• Sin submuestreo • Con submuestreo
Información auxiliar	NO	
	SÍ	• Mejora de la selección de la muestra • Mejora del estimador

CAPÍTULO III. Muestreo de unidades elementales con probabilidades iguales

CONTENIDOS

III.1.- Muestreo aleatorio simple

III.1.1.- Definición de muestreo aleatorio simple (M.A.S.), con reposición y sin reposición.

III.1.2.- Estimación del Total y la Media poblacionales
Errores de muestreo asociados y su estimación.
Intervalos de confianza.

III.1.3.- Estimación de la Proporción y el Total de clase poblacionales
Errores de muestreo asociados y su estimación.
Intervalos de confianza.

III.1.4.- Muestreo sistemático: relación con el M.A.S.

CAPÍTULO III. Muestreo de unidades elementales con probabilidades iguales

CONTENIDOS

III.2.- Muestreo estratificado

III.2.1.- Definición del muestreo aleatorio estratificado
III.2.2.- Tipos de afijación
III.2.3.- Estimación del total y la media poblacionales
III.2.4.- Estimación de la proporción y el total de clase poblacionales
III.2.5.- Errores de muestreo según tipo de afijación
III.2.6.- Estimaciones de los errores de muestreo

CAPÍTULO III. Muestreo de unidades elementales con probabilidades iguales

CONTENIDOS

III.3.- Estimador de Razón

III.3.1.- Definición
III.3.2.- Estimador de razón bajo M.A.S.
III.3.3.- Estimador de razón bajo M.A.E.

III.1.1.- Muestreo Aleatorio Simple

- El diseño aleatorio simple es el diseño muestral más sencillo, siendo utilizado a menudo como diseño base con el que comparar otros diseños más complejos.

- Es de tamaño fijo y exige que los elementos del marco poblacional estén perfectamente identificados, lo que hace que frecuentemente se utilice junto a otro tipo de técnicas.

- Distinguiremos dos situaciones: sin reposición (**m.a.s.**) y con reposición (**m.a.s.r.**)

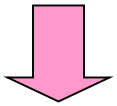
III.1.1.- Muestreo Aleatorio Simple

M.A.S. sin reposición En el muestreo en poblaciones finitas, cuando la muestra se obtiene unidad a unidad, sin reposición de éstas a la población después de cada selección, se dice que el muestreo es aleatorio simple sin reposición.

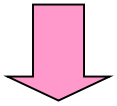
M.A.S. con reposición La muestra se obtiene unidad a unidad, pero la unidad seleccionada se repone en la población antes de la siguiente selección. La población siempre mantiene la misma estructura.

¿Cómo se extrae una muestra aleatoria simple?

Asignar un número de 1 a N a cada unidad de la población



Seleccionar n números, de los N, mediante algún proceso aleatorio.



Las unidades de la población correspondientes a esos números forman la muestra

Las muestras que constan de las mismas unidades se consideran idénticas, independientemente del orden.

Ejemplo 3.1

Se desea seleccionar una muestra de 2 casas de una población de cinco casas

1º Se numeran las casas del 1 al 5

2º Elegimos dos números aleatorios del 1 al 5

3º Las muestras posibles serán las correspondientes a los elementos poblacionales con números:

M.A.S. sin reposición: **(1,2) (1,3) (1,4) (1,5) (2,3) (2,4) (2,5) (3,4) (3,5) (4,5)**

M.A.S. con reposición: **(1,1) (1,2) (1,3) (1,4) (1,5) (2,2) (2,3) (2,4) (2,5) (3,3) (3,4) (3,5) (4,4) (4,5) (5,5)**

m.a.s.(N,n)

Diseño en el que se seleccionan las unidades de la población **sin reponer los elementos observados**, de forma que:

- Todas las unidades tienen la misma probabilidad de selección.
- Todas las muestras son de tamaño fijo y equiprobables.
- Dos muestras que consten de las mismas unidades se consideran idénticas, es decir el orden de extracción no es importante.

Partimos de una población *U* formada por *N* unidades elementales

$$U = \{u_1, \dots, u_N\}$$

El número de muestras posibles de tamaño n es $\binom{N}{n}$

Para cada posible muestra ***s*** de tamaño ***n:*** $s = \{u_1, \dots, u_n\}$

la probabilidad de ser seleccionada es $p(s) = \dfrac{1}{\binom{N}{n}}$

Probabilidades de inclusión

De primer orden:

$$\boldsymbol{\pi_i} = P(u_i \in s) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!\,(N-n)!}}{\frac{N!}{n!\,(N-n)!}} = \frac{n}{N}$$

De segundo orden:

$$\boldsymbol{\pi_{ij}} = P(u_i, u_j \in s) = \frac{\binom{N-1}{n-1}\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

m.a.s.r.(N,n)

Diseño en el que se seleccionan las unidades de la población de forma análoga a la del diseño m.a.s., salvo que **una vez observada la unidad ésta es devuelta a la población, de esta forma la composición de la población permanece constante en cada extracción.**

En las muestras puede haber elementos repetidos.

Partimos de una población *U* formada por *N* unidades elementales

$$U = \{u_1, \dots, u_N\}$$

El número de muestras posibles de tamaño n dependerá de si se considera el orden de extracción de estas unidades o no.

Si el orden importa, tendremos N^n posibles muestras, cada una con probabilidad:

$$p(s) = \frac{1}{N^n}$$

Probabilidades de inclusión

De primer orden:

$$\boldsymbol{\pi_i} = P(u_i \in s) = \frac{nVR_{N,n-1}}{VR_{N,n}} = \frac{nN^{n-1}}{N^n} = \frac{n}{N}$$

III.1.2.- Estimación del Total y la Media

$$U = \{u_1, \ldots, u_N\}$$

Característica ***cuantitativa o numérica*** a estudiar (edad, duración,..)

$\{X_1, \cdots, X_N\}$ Valores desconocidos de la característica en las unidades poblacionales

Seleccionamos una muestra aleatoria sin reposición, **m.a.s.(N,n)**

(Enumeramos las unidades de 1 a N y, a continuación, se extraen n números aleatorios entre 1 y N. En cada extracción, el proceso debe otorgar la misma oportunidad de selección a todos y cada uno de los números que no hayan salido)

III.1.2.- Estimación del Total y la Media

Objetivo:

Desde los resultados encontrados en la muestra se desea encontrar puntualmente o mediante un intervalo el valor del ***Total Poblacional***

$$X = X_1 + \cdots + X_N = \sum_{i=1}^{N} X_i$$

Estimador lineal insesgado para el Total Poblacional

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} x_i$$

$\frac{N}{n} \equiv$ ***Factor de elevación***

Total de unidades de la población que están representadas por una unidad de la muestra

$f = \frac{n}{N} \equiv$ ***Fracción de muestreo***

Porcentaje de la población total representado por la muestra

Ejemplo 3.2

N=1000 n=10

Factor de elevación=**100**

Fracción de muestreo= **0,01**

Cada unidad de la muestra representa a 100 unidades poblacionales, es decir, la muestra representa un 1% de la población total

Varianza del estimador del Total Poblacional

$$V(\hat{X}) = N^2 \left(\frac{1-f}{n}\right) S_X^2$$

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$ **Cuasivarianza poblacional**

A mayor homogeneidad de la característica estudiada en la población, menor será la varianza del estimador y por tanto menor error en la estimación.

Caso extremo: mismo valor de la característica para todas las unidades poblacionales (cuasivarianza cero y error de muestreo nulo)

A mayor tamaño de la muestra, menor varianza y por tanto menor error en la estimación.

Caso extremo: los censos donde n=N, por tanto f=1 y error de muestreo nulo

En la práctica, una vez seleccionada la muestra, **sólo conocemos los valores de la característica estudiada en las unidades muestrales** y no en todas las unidades. Por tanto, no podemos calcular el valor exacto de la cuasivarianza ni, en consecuencia, el de la varianza del estimador.

¿Qué hacemos?

Estimar el valor de la cuasivarianza a partir de la información proporcionada por la propia muestra

Estimador de la Varianza

$$\widehat{V}(\widehat{X}) = N^2 \left(\frac{1-f}{n}\right) \widehat{S}_X^2$$

$$\widehat{S}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ejemplo 3.3

Consideremos una población de 100 familias de las que se quiere conocer el gasto total en alimentación. Seleccionamos una muestra de 10 familias mediante m.a.s. que proporcionan los gastos siguientes en euros:

400, 400, 260, 450, 580, 600, 500, 420, 700, 200

Estimación del gasto total

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n} x_i = \frac{100}{10}(400 + 400 + \cdots + 200) = 45100$$

Factor de elevación

(100/10)=10

Cada unidad de la muestra representa a 10 unidades de la población

(El gasto realizado por cada unidad muestral se multiplica por 10)

¿Qué fiabilidad tiene esta estimación como posible valor del gasto total?

Varianza estimada

$$\widehat{V}(\widehat{X}) = 100^2\left(1 - \frac{10}{100}\right)\frac{23210}{10} = 20889000$$

Error de muestreo

$$\sqrt{20889000} = 4570{,}44$$

Error relativo de muestreo

$$\frac{\sqrt{20889000}}{45100} = 0{,}101 \approx 10\%$$

Estimación del gasto total mediante IC

Intervalo de confianza para el total poblacional

$$IC(X) = \left[\hat{X} - 2\sqrt{\hat{V}(\hat{X})}, \hat{X} + 2\sqrt{\hat{V}(\hat{X})}\right]$$

Intervalo de confianza para el gasto total poblacional al 95%

$$IC(X) = (45100 \pm 2 \times 4570{,}44) = (35959{,}12; 54240{,}88)$$

El verdadero valor del gasto total se sitúa en este rango de valores con una confianza de más del 95%. (De cada 100 construcciones de este intervalo, correspondientes a otras tantas muestras seleccionadas, tenemos la certeza de que ***X*** estará en 95 de ellas)

Si la selección de la muestra se hubiera realizado con reposición , **m.a.s.r.(N,n)**

Estimador lineal insesgado para el Total Poblacional

$$\hat{X} = \frac{N}{n}\sum_{i=1}^{n} x_i$$

Estimador de la Varianza

$$\hat{V}(\hat{X}) = N^2 \frac{\hat{S}_x^2}{n}$$

$S^2 = N\sigma^2/(N-1)$: cuasivarianza poblacional

$f=n/N$: fracción de muestreo.

$(1-f)$: coeficiente de corrección por finitud.

Poblaciones grandes:

Si $f \leq 0{,}05$ (Se muestrea a lo sumo el 5% de la población)

$$(1-f) \approx 1$$

$$S^2 \approx \sigma^2$$

Los errores de muestreo coinciden en el muestreo con reposición y sin reposición

Objetivo:

Desde los resultados encontrados en la muestra se desea encontrar puntualmente o mediante un intervalo el valor de la ***Media Poblacional***

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

	m.a.s.(N,n)	m.a.s.r.(N,n)
Estimador	$\widehat{\overline{X}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\widehat{X}}{N}$	
Varianza del estimador	$V(\bar{x}) = \frac{1-f}{n} S_x^2$	$V(\overline{x}) = \frac{\sigma_x^2}{n}$
Estimador de la varianza	$\widehat{V}(\bar{x}) = \frac{1-f}{n} \widehat{S}_x^2$	$\widehat{V}(\bar{x}) = \frac{\widehat{S}_x^2}{n}$

Ejemplo 3.3 (cont.)

Estimar el gasto medio en alimentación de las 100 familias a partir de la información de la m.a.s. de tamaño 10

Estimación del gasto medio

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(400 + 400 + \cdots + 200)}{10} = 451$$

Error de muestreo

$$\sqrt{2088,9} = 45,7$$

Varianza estimada

$$\hat{V}(\bar{x}) = \left(1 - \frac{10}{100}\right)\frac{23210}{10} = 2088,9$$

Error relativo de muestreo

$$\frac{\sqrt{2088,9}}{451} = 0,101$$

Estimación del gasto medio mediante IC

Intervalo de confianza para la Media

$$IC(\overline{X}) = \left[\overline{x} - 2\sqrt{\hat{V}(\overline{x})}\,,\ \overline{x} + 2\sqrt{\hat{V}(\overline{x})}\right]$$

IC para el gasto medio poblacional al 95%

$$IC(\overline{X}) = (451 \pm 2 \times 45,7) = (359,6; 542,4)$$

Obsérvese que los valores de los extremos de este intervalo coinciden con los del obtenido para el ***total poblacional*** divididos por ***N=100***)

III.1.3.- Estimación de la Proporción y el Total de Clase

$$U = \{u_1, \ldots, u_N\}$$

Característica ***cualitativa*** a estudiar (sexo, población activa,…)

Para cada unidad de la población se define:

$$X_i = \begin{cases} 1 & \text{si } u_i \text{ posee la cualidad de interés} \\ 0 & \text{caso contrario} \end{cases}$$

Seleccionamos una muestra aleatoria sin reposición, **m.a.s.(N,n)**

(Enumeramos las unidades de 1 a N y, a continuación, se extraen n números aleatorios entre 1 y N. En cada extracción, el proceso debe otorgar la misma oportunidad de selección a todos y cada uno de los números que no hayan salido)

III.1.3.- Estimación de la Proporción y el Total de Clase

Objetivo:

Desde los resultados encontrados en la muestra se desea encontrar puntualmente o mediante un intervalo el valor de la ***Proporción Poblacional***

$$P = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

	m.a.s.(N,n)	m.a.s.r.(N,n)
Estimador	$p = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	
Varianza del estimador	$V(p) = \frac{(1-f)}{n}\frac{NPQ}{N-1} = \frac{N-n}{N-1}\frac{PQ}{n}$	$V(p) = \frac{PQ}{n}$
Estimador de la varianza	$\hat{V}(p) = \frac{(1-f)}{(n-1)}pq = \frac{N-n}{N}\frac{pq}{(n-1)}$	$\hat{V}(p) = \frac{pq}{n-1}$

$$IC_{0.95}(P) = \left(p - 2\sqrt{\hat{V}(p)} - \frac{1}{2n}, p + 2\sqrt{\hat{V}(p)} + \frac{1}{2n}\right)$$

$$Q = 1 - P \qquad q = 1 - p$$

Ejemplo 3.4

Estimar el valor de la proporción de individuos de sexo masculino en la Comunidad de Madrid a partir de una muestra de 1049 personas observadas de la CAM, en la cual 550 eran hombres y 499 mujeres.

Estimación de la proporción

$$p = \frac{550}{1049} = 0,524$$

Varianza estimada

$$\hat{V}(p) = \frac{pq}{n-1} = \frac{0,524 \times (1-0,524)}{(1049-1)} = 0,00023$$

Error de muestreo

$$\sqrt{\hat{V}(p)} = \sqrt{0,00023} = 0,0154$$

Obsérvese que, en este caso, la fracción de muestreo es muy pequeña por lo que se puede aplicar la expresión CR

Error relativo de muestreo

$$\frac{\sqrt{0,00023}}{0,524} = 0,0294 \approx 2,9\%$$

Estimación de la proporción mediante IC

Intervalo de confianza para la Proporción

$$IC_{0.95}(P) = \left(p - 2\sqrt{\hat{V}(p)}, p + 2\sqrt{\hat{V}(p)}\right)$$

IC para la proporción poblacional al 95%

$$IC(P) = \left(0,524 \pm \left(2\sqrt{0,00023}\right)\right) = (0,493; 0,555)$$

Efecto de *P* en los errores de muestreo

En la tabla siguiente se muestra la función *PQ* y su raíz cuadrada, para distintas proporciones poblacionales

P	0	0,1	0, 2	0,3	0,4	**0,5**	0,6	0,7	0,8	0,9	1
PQ	0	0,09	0,16	0,21	0,24	**0,25**	0,24	0,21	0,16	0,09	0
$\sqrt{PQ}$	0	0,3	0,4	0,46	0,49	**0,50**	0,49	0,46	0,4	0,3	0

Los errores de muestreo son máximos cuando la población está igualmente dividida entre las dos clases, es decir, ***P*=*Q*=0,5** (o 50%, si hablamos en porcentaje)

Objetivo:

Desde los resultados encontrados en la muestra se desea encontrar puntualmente o mediante un intervalo el valor del ***Total de clase Poblacional***

$$A = \sum_{i=1}^{N} X_i = X$$

	m.a.s.(N,n)	m.a.s.r.(N,n)
Estimador	$\widehat{A} = Np$	
Varianza del estimador	$V(\widehat{A}) = N^2 V(p)$	
Estimador de la varianza	$\widehat{V}(\widehat{A}) = N^2 \widehat{V}(p)$	

$$IC_{0.95}(A) = \left(\hat{A} \pm \left(2\sqrt{\hat{V}(\hat{A}) + \frac{N}{2n}} \right) \right)$$

Ejemplo 3.4 (cont.)

Si la población de la CAM es de 6500000 habitantes, estimar el número total de individuos de sexo masculino en la Comunidad de Madrid a partir de una muestra de 1049 personas observadas de la CAM, en la cual 550 eran hombres y 499 mujeres.

Estimación del total de hombres

$$A = Np = 6500000 * 0,524 = 3408007,63 \approx 3408008$$

Varianza estimada

$$\hat{V}(A) = 1,0055E + 10$$

Error de muestreo

$$\sqrt{\hat{V}(A)} = 100274,11$$

Error relativo de muestreo

$$\frac{\sqrt{\hat{V}(A)}}{A} = \frac{100274,11}{3408008} = 0,0294 \approx 2,9\%$$

Estimación del total de clase mediante IC

Intervalo de confianza para el Total de clase

$$IC_{0.95}(A) = \left(\hat{A} \pm 2\sqrt{\hat{V}(\hat{A})}\right)$$

IC para el gasto medio poblacional al 95%

$$IC\ (A) = \left(3408008 \pm \left(2\ \sqrt{1{,}0055E+10}\right)\right)$$
$$= (3207459; 3608556)$$

Obsérvese que los valores de los extremos de este intervalo coinciden (aproximadamente, salvo redondeos) con los del obtenido para la ***proporción multiplicados*** por ***N***=6500000

III.1.4.- Muestreo Sistemático: relación con el M.A.S.

Definición: Consiste en tomar las unidades poblacionales, que formarán la muestra, de *k* en *k,* a partir de una elegida al azar entre las *k* primeras.

Ejemplo: De 15000 especialistas deseamos entrevistar a 100.

k=15000/100 = 150

Seleccionamos al azar un nº entre 1 y 150 → {j}

Muestra sistemática: $\{u_j, u_{j+150}, u_{j+300}, \ldots\ldots u_{j+14850}\}$

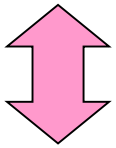
Por ejemplo si j=3, muestra sistemática

$\{u_3, u_{153}, u_{303}, u_{453}, \ldots\ldots, u_{14703}, u_{14853}\}$

III.1.4.- Muestreo Sistemático: relación con el M.A.S.

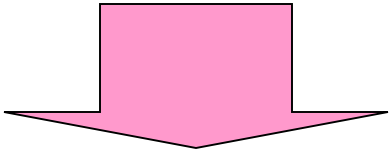
Si el orden de las unidades en la población es más o menos aleatorio, entonces

Muestra sistemática



Similar

Muestra aleatoria simple (sin reposición)



Los métodos del M.A.S. pueden utilizarse en el análisis dada una muestra sistemática.

III.2.1- Muestreo Aleatorio Estratificado

Si la población está formada por unidades elementales que no son homogéneas respecto a la característica que se estudia no es aconsejable utilizar el MAS

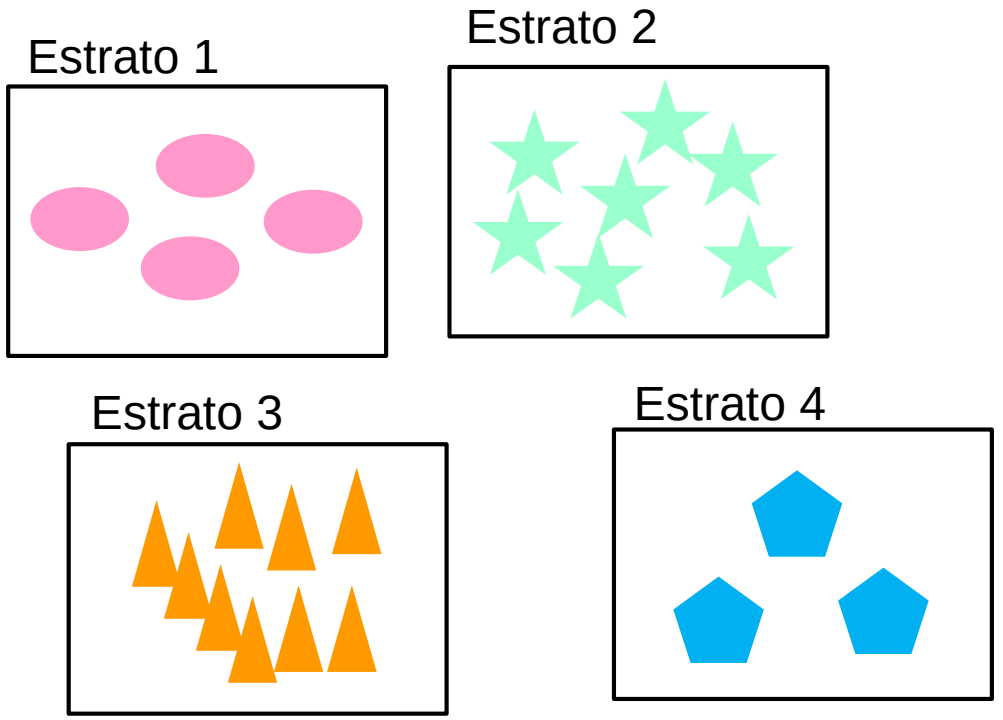
Ejemplos

- Si X: nº de calorías consumidas diariamente, se sabe que, en general, los hombres consumen más que las mujeres
- Si X: nº de horas de estudio semanales de un alumno de la ESO, es de suponer que los de cuarto tienen que dedicar más horas al estudio que los de primero.

- En el primer caso, la población estaría dividida en dos subconjuntos de unidades elementales homogéneas: hombres y mujeres

- En el segundo caso, la población estaría dividida en cuatro subconjuntos de alumnos homogéneos: los de 1º, los de 2º, los de 3º y los de 4º.

A cada uno de los subconjuntos de unidades elementales homogéneas de la población, se le llama ESTRATO



En esta situación es mejor tomar una m.a.s. dentro de cada estrato, y todas juntas, formar la muestra global de toda la población.

MUESTREO ALEATORIO ESTRATIFICADO

Todos los estratos estarán representados en la muestra, y ésta representará mejor a la población ya que está asegurado que tendrá unidades de todos los estratos

Además...

Se pueden dar estimaciones en cada uno de los estratos

Se consigue una mayor precisión en la estimación con el mismo tamaño de muestra

Se pueden reducir los costes del muestreo

¿QUÉ TIPO DE SUBGRUPOS SUELEN SER LOS ESTRATOS?

Suelen ser grupos "naturales"o subpoblaciones de unidades elementales

Áreas geográficas: Países, provincias, municipios, distritos…

Características biológicas: el mismo sexo, el mismo grupo de edad…

Cercanía temporal: días de la semana, años, meses...

Pero también, se pueden formar a partir de una variable de estratificación que deberá estar relacionada con la característica a estimar

¿QUÉ TIPO DE SUBGRUPOS SUELEN SER LOS ESTRATOS?

Por ejemplo:

Se quiere estimar el volumen total de ventas de productos navideños. Tomamos como unidades elementales los establecimientos que venden este tipo de artículos. Éstos pueden clasificarse según su tamaño (variable de estratificación). Esta variable está muy relacionada con las ventas

En general, los estratos deben estar formados por unidades sobre las cuales varíen poco las mediciones X de la variable de interés

PARÁMETROS POBLACIONALES

Además de los parámetros definidos sobre toda la población, se podrán definir los parámetros referidos a cada estrato

La población está formada por *L* estratos $E_1,\ldots,E_L$. Cada unidad pertenece a uno y sólo un estrato

El estrato *h* tiene N_h unidades. Así $N_1+\ldots+N_L=N$

Se llamará peso del estrato *h* a la proporción de unidades que pertenecen a ese estrato $W_h=N_h/N$

Para indicar en qué estrato está la unidad, denotaremos con X_{hi}, al valor de la característica o variable *X* sobre la unidad *i* que está en el estrato *h*

1. MEDIA POBLACIONAL del estrato *h*

$$\overline{X}_h = \frac{X_{h1} + X_{h2} + \ldots + X_{hN_h}}{N_h} = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h}$$

2. TOTAL POBLACIONAL del estrato *h*

$$X_h = \sum_{i=1}^{N_h} X_{hi}$$

- Nº MEDIO DE CALORÍAS CONSUMIDAS POR LOS HOMBRES

- Nº TOTAL DE FOTOCOPIAS HECHAS EN EL DEPARTAMENTO DE VENTAS

3. PROPORCIÓN O PORCENTAJE DE UNA CLASE en el estrato *h*

$$P_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h}$$

4. TOTAL DE CLASE en el estrato *h*

$$A_h = \sum_{i=1}^{N_h} X_{hi}$$

$$X_{hi} = \begin{cases} 1 & \textit{elemento i del estrato h está en la clase} \\ 0 & \textit{elemento i del estrato h no está en la clase} \end{cases}$$

- PROPORCIÓN DE PARADOS ENTRE LAS MUJERES

- Nº TOTAL DE EMPLEADOS DEL DEPARTAMENTO DE VENTAS CON MÁS DE UN CURSO DE FORMACIÓN

5. CUASIVARIANZA POBLACIONAL en el estrato *h*

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2}{N_h - 1}$$

6. CUASIDESVIACIÓN TÍPICA POBLACIONAL en el estrato *h*

$$S_h = +\sqrt{S_h^2}$$

¿Existe alguna relación entre los parámetros de los estratos y los referentes a toda la población?

RELACIÓN ENTRE LOS PARÁMETROS POBLACIONALES

$$\overline{X} = \sum_{h=1}^{L} W_h \overline{X}_h$$

$$X = \sum_{h=1}^{L} X_h$$

$$P = \sum_{h=1}^{L} W_h P_h$$

$$A = \sum_{h=1}^{L} A_h$$

¿Cuántas unidades hay que muestrear en cada estrato?

Se llama afijación de la muestra al reparto o distribución del tamaño de la muestra *n* entre los distintos estratos.

n_h : nº de unidades muestreadas en el estrato *h*

$$n_1 + n_2 + \ldots + n_L = n$$

$f_h = \frac{n_h}{N_h}$ ***Fracción de muestreo* en el estrato *h***

N_h/n_h ***Factor de elevación en el estrato h***

El reparto se podrá hacer teniendo en cuenta:

- **El peso o tamaños de los estratos: W_h, N_h**
- **La variabilidad de los estratos: S^2_h**
- **El coste de muestrear en cada estrato: c_h**

En el **muestreo aleatorio estratificado**, **M.A.E.**, seleccionamos una **m.a.s.(n_h) sr** en cada estrato de manera que las muestras sean independientes

Si n_h se eligen proporcionalmente al tamaño del estrato

→ Afijación proporcional

Si n_h se eligen proporcionalmente al tamaño del estrato y a la variabilidad de la característica dentro de cada estrato

→ Afijación óptima de Neyman

Si n_h se eligen proporcionalmente al tamaño del estrato, a la variabilidad de la característica dentro de cada estrato y teniendo en cuenta el coste de muestreo en cada estrato

→ Afijación óptima de costes variables

1. AFIJACIÓN PROPORCIONAL

$$n_h = nW_h \qquad f_h = f = \frac{n}{N}$$

2. AFIJACIÓN ÓPTIMA DE NEYMAN

$$n_h = n\frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j}$$

Los valores de las cuasivarianzas en los estratos se estiman partir de una muestra piloto, a partir de experiencias anteriores o a partir de la variable de estratificación

3. AFIJACIÓN ÓPTIMA CON COSTES VARIABLES

La afijación óptima con costes variables consiste en repartir de manera que, fijado un error de muestreo se tenga un coste mínimo, o fijado un coste C se obtenga un error mínimo. Si se tiene una función de coste $C=c_1n_1+...+c_Ln_L$,

$$n_h = n\frac{N_hS_h/\sqrt{c_h}}{\sum_{j=1}^{L} N_jS_j/\sqrt{c_j}}$$

$$n_h = C\frac{N_hS_h/\sqrt{c_h}}{\sum_{j=1}^{L} N_jS_j\sqrt{c_j}}$$

Los valores de las cuasivarianzas en los estratos se estiman partir de una muestra piloto o a partir de experiencias anteriores

Ejemplo 3.5

En cierto país ha habido 260000 bajas laborales en el último año y se quiere estimar el nº medio de días de una baja. De todas ellas tiene 150000 informatizadas mientras que del resto, sólo dispone de los partes de baja.

Se dispone de 10000 euros para realizar el muestreo y se estima que el muestreo y el procesamiento de una baja ya informatizada es de 0.32 euros, mientras que si hay que informatizarla previamente, el coste se eleva a 0.98 euros. Se supone también que la cuasidesviación típica en el primer grupo es la mitad que en el segundo

Con ese presupuesto, ¿cuántas bajas laborales se deberían elegir de cada grupo?

X_i: nº de días que ha durado la baja laboral i

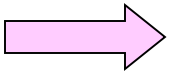
Estrato 1: bajas informatizadas

Estrato 2: bajas no informatizadas

$N_1=150000$ $N_2=110000$ $C=10000$

$c_1=0.32$ $c_2=0.98$ $S_1=S_2/2$

Como el precio de muestreo depende del estrato



Afijación óptima para costes variables con C fijado

$$n_h = C\frac{N_h S_h / \sqrt{c_h}}{\sum_{j=1}^{L} N_j S_j \sqrt{c_j}}$$

N_1=150000 N_2=110000 C=10000

c_1=0.32 c_2=0.98 S_1=S_2/2

$$n_1 = 10000\frac{150000 \cdot S_1 / \sqrt{0.32}}{150000 \cdot S_1\sqrt{0.32} + 110000 \cdot 2S_1\sqrt{0.98}} = 8761.6$$

$$n_2 = 10000\frac{110000 \cdot S_2 / \sqrt{0.98}}{150000 \cdot (S_2/2)\sqrt{0.32} + 110000 \cdot S_2\sqrt{0.98}} = 7343.12$$

Debería muestrear 8762 informatizados y 7343 no informatizados, es decir un total de 16105 bajas

III.2.3- Estimación del Total y la Media

Las estimaciones de los parámetros sobre toda población vendrán expresadas en términos de las estimaciones de los parámetros de cada estrato

$$\hat{\bar{X}} = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$\hat{X} = N\hat{\bar{X}}$$

donde $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ es la media muestral en el estrato *h*

Ejemplo 3.6

En una población de 1500 familias se quiere estimar los gastos medios mensuales en transporte por familia.

Estos gastos varían considerablemente dependiendo de los ingresos que tenga cada familia. Así se estratifican las familias en función de esta variable “ingresos”

Estrato 1: familias con ingresos bajos (<900 euros): 517 familias

Estrato 2: familias con ingresos medios (900-2000 euros): 633 familias

Estrato 3: familias con ingresos altos (más de 2000 euros): 350 familias

Se muestrean un total de 30 familias: 10 en el primer estrato, 13 en el segundo y 7 en el tercero, y se obtienen las siguientes medias en cada uno de ellos

$$\bar{x}_1 = 95 \quad \bar{x}_2 = 150 \quad \bar{x}_3 = 280$$

Así, el gasto medio mensual en transporte por familia sería:

$$\hat{\bar{X}} = 95 \cdot \frac{517}{1500} + 150 \cdot \frac{633}{1500} + 280 \frac{350}{1500} = 161.37 \; euros$$

III.2.4- Estimación de la Proporción y el Total de Clase

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{P}_h$$

$$\hat{A} = N\hat{P}$$

donde $\hat{P}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$ es la proporción muestral en el estrato *h*

$$x_{hi} = \begin{cases} 1 & \textit{elemento i del estrato h pertenece a la clase} \\ 0 & \textit{elemento i del estrato h no pertenece a la clase} \end{cases}$$

Con la afijación proporcional $\hat{P}$ es la proporción muestral

III.2.5- Errores de Muestreo

Las expresiones generales de las varianzas de los estimadores de la media, total, proporción y total de clase poblacionales vienen en función de las varianzas de las estimaciones en cada estrato

$$V(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2 V(\bar{x}_h) \qquad V(\hat{X}) = N^2 V(\hat{\bar{X}}) = \sum_{h=1}^{L} V(\hat{X}_h)$$

$$V(\hat{P}) = \sum_{h=1}^{L} W_h^2 V(\hat{P}_h) \qquad V(\hat{A}) = N^2 V(\hat{P}) = \sum_{h=1}^{L} V(\hat{A}_h)$$

Suponiendo un m.a.s. sin reposición en cada estrato,

$$V(\bar{x}_h) = \frac{1-f_h}{n_h} S_h^2 \qquad V(\hat{X}_h) = N_h^2 \frac{1-f_h}{n_h} S_h^2$$

Además, $V(\hat{P}_h) = V(\bar{x}_h) \qquad V(\hat{A}_h) = V(\hat{X}_h)$

siendo $S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h)$

En general,

$$V(\hat{\bar{X}}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N} = V(\hat{P})$$

$$V(\hat{X}) = \sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} N_h S_h^2 = V(\hat{A})$$

siendo $S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h)$ en el caso de atributos

Los errores de muestreo son la raíz cuadrada de estas varianzas

1. CON AFIJACIÓN PROPORCIONAL

$$n_h = nW_h$$

$$V_{PROP}(\hat{\bar{X}}) = (\frac{1-f}{n})\sum_{h=1}^{L} W_h S_h^2 = V_{PROP}(\hat{P})$$

$$V_{PROP}(\hat{X}) = N^2(\frac{1-f}{n})\sum_{h=1}^{L} W_h S_h^2 = V_{PROP}(\hat{A})$$

siendo $S_h^2 = \frac{N_h}{N_h - 1} P_h(1-P_h)$ en el caso de atributos

Los errores de muestreo son la raíz cuadrada de estas varianzas

2. CON AFIJACIÓN ÓPTIMA DE NEYMAN

$$n_h = n\frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j}$$

$$V_{NEY}(\hat{\bar{X}}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 = V_{NEY}(\hat{P})$$

$$V_{NEY}(\hat{X}) = \frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - N\sum_{h=1}^{L} W_h S_h^2 = V_{NEY}(\hat{A})$$

siendo $S_h^2 = \frac{N_h}{N_h - 1} P_h(1 - P_h)$ en el caso de atributos

Los errores de muestreo son la raíz cuadrada de estas varianzas

3. CON AFIJACIÓN ÓPTIMA CON COSTES VARIABLES

$$n_h = n\frac{N_h S_h / \sqrt{c_h}}{\sum_{j=1}^{L} N_j S_j / \sqrt{c_j}}$$

Los errores de muestreo son la raíz cuadrada de estas varianzas

$$V_{COST}(\hat{\bar{X}}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 = V_{COST}(\hat{P})$$

$$V_{COST}(\hat{X}) = \frac{N^2}{n}\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right) - N\sum_{h=1}^{L} W_h S_h^2 = V_{COST}(\hat{A})$$

siendo $S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h)$ en el caso de atributos

AFIJACIÓN QUE REQUIERE MÁS DEL 100% DEL MUESTREO

Ejemplo 3.7

Supongamos la situación siguiente en donde se tiene la población dividida en tres estratos y se quiere tomar una muestra de tamaño 140.

Estrato	N_h	S_h
1	100	50
2	110	10
3	120	5

¿Qué hacer?

Utilizando la afijación de Neyman

$$n_h = n\frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j}$$

n_1=104

n_2= 23

n_3=13

AFIJACIÓN QUE REQUIERE MÁS DEL 100% DEL MUESTREO

Ejemplo 3.7 (cont.)

Estrato	N_h	S_h
1	100	50
2	110	10
3	120	5

Se toma $n_1=N_1=100$ y se calculan los otros dos tamaños muestrales a partir de la expresión:

$$\tilde{n}_h = (n - N_1)\frac{N_h S_h}{\sum_{j=2}^{3} N_j S_j} \quad h = 2,3$$

$$\tilde{n}_2 = (140-100)\frac{110\cdot 10}{110\cdot 10+120\cdot 5} = 40\frac{1100}{1700} = 25.88$$

$$\tilde{n}_3 = (140-100)\frac{120\cdot 5}{110\cdot 10+120\cdot 5} = 40\frac{600}{1700} = 14.11$$

AFIJACIÓN QUE REQUIERE MÁS DEL 100% DEL MUESTREO

Ejemplo (cont.)

Así: $\tilde{n}_1 = N_1 = 100 \quad \tilde{n}_2 = 26 \quad \tilde{n}_3 = 14$

La varianza del estimador de la media queda modificada teniendo la expresión

$$V_{NEY}(\hat{\bar{X}}) = \frac{1}{n'}\left(\sum_{h=2}^{3} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=2}^{3} W_h S_h^2$$

$$n' = \tilde{n}_2 + \tilde{n}_3$$

III.2.6.- Estimación de los Errores de Muestreo

En general,

$$\hat{V}(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\bar{x}_h) \Rightarrow \hat{EM}(\hat{\bar{X}}) = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\bar{x}_h)}$$

$$\hat{V}(\hat{X}) = N^2 \hat{V}(\hat{\bar{X}}) = \sum_{h=1}^{L} V(\hat{X}_h) \Rightarrow \hat{EM}(\hat{X}) = \sqrt{\sum_{h=1}^{L} \hat{V}(\hat{X}_h)}$$

$$\hat{V}(\hat{P}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\hat{P}_h) \Rightarrow \hat{EM}(\hat{P}) = \sqrt{\sum_{h=1}^{L} W_h^2 \hat{V}(\hat{P}_h)}$$

$$\hat{V}(\hat{A}) = N^2 \hat{V}(\hat{P}) = \sum_{h=1}^{L} V(\hat{A}_h) \Rightarrow \hat{EM}(\hat{A}) = \sqrt{\sum_{h=1}^{L} \hat{V}(\hat{A}_h)}$$

Suponiendo un m.a.s. sin reposición en cada estrato,

$$\hat{V}(\hat{\bar{X}}) = \sum_{h=1}^{L} \frac{W_h^2 \hat{S}_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h \hat{S}_h^2}{N} = \hat{V}(\hat{P})$$

$$E\hat{M}(\hat{\bar{X}}) = \sqrt{\sum_{h=1}^{L} \frac{W_h^2 \hat{S}_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h \hat{S}_h^2}{N}} = E\hat{M}(\hat{P})$$

siendo

$$\hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (X_{hi} - \bar{x}_h)^2$$

y en el caso de atributos

$$\hat{S}_h^2 = \frac{n_h}{n_h - 1} \hat{P}_h (1 - \hat{P}_h)$$

$$\hat{V}(\hat{X}) = \sum_{h=1}^{L} \frac{N_h^2 \hat{S}_h^2}{n_h} - \sum_{h=1}^{L} N_h \hat{S}_h^2$$

$$E\hat{M}(\hat{X}) = \sqrt{\sum_{h=1}^{L} \frac{N_h^2 \hat{S}_h^2}{n_h} - \sum_{h=1}^{L} N_h \hat{S}_h^2} = E\hat{M}(\hat{A})$$

siendo $\hat{S}_h^2 = \frac{n_h}{n_h - 1} \hat{P}_h (1 - \hat{P}_h)$ en el caso de atributos

Además,

$$E\hat{M}(\hat{X}) = N \cdot E\hat{M}(\hat{\overline{X}})$$

$$E\hat{M}(\hat{A}) = N \cdot E\hat{M}(\hat{P})$$

1. CON LA AFIJACIÓN PROPORCIONAL

$$n_h = nW_h$$

$$\hat{EM}_{PROP}(\hat{\bar{X}}) = \sqrt{\frac{(1-f)}{n}\sum_{h=1}^{L} W_h \hat{S}_h^2} = \hat{EM}_{PROP}(\hat{P})$$

siendo $\hat{S}_h^2 = \frac{n_h}{n_h - 1}\hat{P}_h(1-\hat{P}_h)$ en el caso de atributos

$$\hat{EM}_{PROP}(\hat{X}) = N \cdot \hat{EM}_{PROP}(\hat{\bar{X}})$$

$$\hat{EM}_{PROP}(\hat{A}) = N \cdot \hat{EM}_{PROP}(\hat{P})$$

2. CON LA AFIJACIÓN ÓPTIMA DE NEYMAN

$$n_h = n \frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j}$$

$$\hat{EM}_{NEY}(\hat{\bar{X}}) = \sqrt{\frac{1}{n}\left(\sum_{h=1}^{L} W_h \hat{S}_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h \hat{S}_h^2} = \hat{EM}_{NEY}(\hat{P})$$

siendo $\hat{S}_h^2 = \frac{n_h}{n_h - 1}\hat{P}_h(1-\hat{P}_h)$ en el caso de atributos

$$\hat{EM}_{NEY}(\hat{X}) = N \cdot \hat{EM}_{NEY}(\hat{\bar{X}})$$

$$\hat{EM}_{NEY}(\hat{A}) = N \cdot \hat{EM}_{NEY}(\hat{P})$$

3. CON AFIJACIÓN ÓPTIMA CON COSTES VARIABLES

$$n_h = n \frac{N_h S_h / \sqrt{c_h}}{\sum_{j=1}^{L} N_j S_j / \sqrt{c_j}}$$

$$\hat{EM}_{COST}(\hat{\bar{X}}) = \sqrt{\frac{1}{n}\left(\sum_{h=1}^{L} W_h \hat{S}_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \hat{S}_h / \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h \hat{S}_h^2} = \hat{EM}_{COST}(\hat{P})$$

siendo $\hat{S}_h^2 = \frac{n_h}{n_h - 1}\hat{P}_h(1-\hat{P}_h)$ en el caso de atributos

$$\hat{EM}_{COST}(\hat{X}) = N \cdot \hat{EM}_{COST}(\hat{\bar{X}})$$

$$\hat{EM}_{COST}(\hat{A}) = N \cdot \hat{EM}_{COST}(\hat{P})$$

Ejemplo 3.8

Una cadena de almacenes está interesada en estimar la proporción de cuentas por cobrar cuyo importe es despreciable. Esta cadena consta de cuatro almacenes, que considera como estratos. Utiliza la afijación proporcional y los datos se recogen en la siguiente tabla.

	Almacén			
	1	*2*	*3*	*4*
N	65	42	93	25
n	14	9	23	6
Nº cuentas despreciables	4	2	8	1

Estimar la proporción de cuentas impagadas despreciables y dar su error de estimación

Las estimaciones de las proporciones en cada estrato son:

$$\hat{P}_1 = \frac{4}{14} = 0.28 \quad \hat{P}_2 = \frac{2}{9} = 0.22 \quad \hat{P}_3 = \frac{8}{23} = 0.34 \quad \hat{P}_4 = \frac{1}{6} = 0.16$$

A partir de ellas

$$\hat{P} = \sum_{h=1}^{4} W_h \hat{P}_h = \frac{65}{225} \cdot \frac{4}{14} + \frac{42}{225} \cdot \frac{2}{9} + \frac{93}{225} \cdot \frac{8}{23} + \frac{25}{225} \cdot \frac{1}{6} = 0.28$$

El porcentaje de cuentas impagadas despreciables es del 28%

El error de la estimación será:

$$E\hat{M}_{PROP}(\hat{P}) = \sqrt{\frac{(1-f)}{n}\sum_{h=1}^{L} W_h \hat{S}_h^2} = \sqrt{\frac{(1-f)}{n}\sum_{h=1}^{L} W_h \frac{n_h}{n_h - 1}\hat{P}_h\hat{Q}_h}$$

$$E\hat{M}_{PROP}(\hat{P}) = \sqrt{\frac{(1-0.2)}{52}\left(\frac{65}{225}\frac{14}{14-1}\frac{4}{14}\frac{10}{14} + \ldots + \frac{25}{225}\frac{6}{6-1}\frac{1}{6}\frac{5}{6}\right)} = 0.056$$

La proporción de facturas impagadas despreciables está entre 0.17 y 0.378 con una seguridad del 95%

Ejemplo 3.9

En cierta región se quiere estimar el nº total de Ha dedicadas al cultivo del cereal. Se decide estratificar las fincas según su tamaño. Se muestrean 240 fincas clasificadas en uno de los cuatro estratos definidos.

Ha.	*0-200*	*201-400*	*401-600*	*+600*
N	86	72	52	30
n	14	12	9	5
Total muestral	887	2197	3065	2362
$\hat{S}_h$	32.73	95.07	129.59	269.02

Estimar el total de Ha dedicadas al cultivo del cereal y estimar su error de muestreo

Ya que se utiliza la afijación proporcional

$$\hat{X} = N\hat{\bar{X}} = 240\frac{887+2197+3065+2362}{40} = 240\frac{8511}{40} = 51066$$

$$\hat{V}_{PROP}(\hat{X}) = N^2\hat{V}_{PROP}(\hat{\bar{X}}) = N^2\frac{(1-f)}{n}\sum_{h=1}^{L} W_h\hat{S}_h^2$$

$$\hat{V}_{PROP}(\hat{X}) = 240^2\frac{(1-0.16)}{40}\left(\frac{86}{240}32.73^2 + \frac{72}{240}95.07^2 + \frac{52}{240}129.59^2 + \frac{30}{240}269.02^2\right)$$

$$\hat{EM}_{PROP}(\hat{X}) = +\sqrt{19088012} = 4368.9$$

Error máximo admisible= 1.96·4368.9=8563

Estimamos que el nº Ha. dedicadas al cereal es de 51066 con un error de ±8563 Ha.

Por último…..

La selección de muestras independientes en estratos tiene varias ***ventajas***:

- **Mejora la representatividad de la muestra**, en lo que se refiere a las variables utilizadas en la estratificación
- **Mejora la precisión del estimador global**, si la estratificación construye agrupaciones homogéneas de las unidades elementales
- **Permite un reparto óptimo de la muestra por estratos** en cuanto a precisión y coste

El único ***inconveniente*** es la **necesidad de información auxiliar** disponible en el marco

III.3.1.- Estimador de Razón

Si el parámetro de interés es una **razón:**

Salario cobrado por trabajador

Tasa de paro

debemos observar dos características ***cuantitativas o numéricas*** en cada unidad de la muestra

X:

Cantidad de euros pagada por la empresa en concepto de salarios

Y:

Total de trabajadores de la empresa

III.3.1.- Estimador de Razón

$U = \{u_1, \dots, u_N\}$ → Seleccionamos una muestra aleatoria sin reposición, **m.a.s.(N,n)**

$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$

Valores conocidos de las características en las unidades muestrales

Objetivo:

Desde los resultados encontrados en la muestra se desea encontrar el valor de la ***Razón Poblacional***

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

III.3.2.- Estimador de Razón bajo M.A.S.

Estimador para la Razón Poblacional bajo M.A.S. (N,n)

$$\hat{R} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \frac{\hat{X}}{\hat{Y}} = \frac{\overline{x}}{\overline{y}}$$

¡Ojo! : El estimador de Razón es sesgado

Situaciones en las que el sesgo es despreciable:

- El tamaño de la muestra es grande
- La relación entre las dos características cuantitativas es una recta que pasa por el origen

Ej: Es el caso de los salarios y trabajadores de una empresa (cuántos más trabajadores hay más salarios se pagan, y si no hay trabajadores no se pagan salarios)

III.3.2.- Estimador de Razón bajo M.A.S.

$$\widehat{ERM}(\bar{y}) = \frac{\sqrt{\hat{V}(\bar{y})}}{\bar{y}} = \frac{1}{\bar{y}}\sqrt{(1-f)\frac{\hat{S}_Y^2}{n}}$$

Regla práctica:

Si $\widehat{ERM}(\bar{y}) < 0{,}1 \Rightarrow$ ***el sesgo es despreciable***

Factores que pueden reducir el sesgo:

- Relación de proporcionalidad entre X e Y
- Bajo coeficiente de variación de la variable Y: $\frac{S_Y}{\bar{Y}}$
- Tamaño muestral suficientemente alto

Para el cálculo de la varianza se utiliza el método de linealización de Taylor , **válido cuando el sesgo es despreciable**, que proporciona una expresión aproximada

Como en ocasiones anteriores, no es posible calcular la varianza del estimador que se estima en base a la propia muestra

Estimador de la Varianza

$$\hat{V}(\hat{R}) \approx \frac{(1-f)}{\bar{y}^2 n(n-1)} \sum_{i=1}^{n} (x_i - \hat{R} y_i)^2$$

$$= \frac{(1-f)}{n\bar{y}^2} \left(\hat{S}_X^2 + \hat{R}^2 \hat{S}_Y^2 - 2\hat{R}\hat{S}_{XY} \right)$$

III.3.2.- Estimador de Razón bajo M.A.S.

Hay veces en que se desea estimar el total poblacional (X) de una característica cuantitativa y se dispone del total poblacional (Y) de otra característica cuantitativa correlada positivamente con la anterior (Ej. Estimar el total de salarios y disponer del total de trabajadores).

Entonces, los parámetros poblacionales total y media muestral en función de la razón poblacional quedan como:

$$X = R \cdot Y \qquad \overline{X} = R \cdot \overline{Y}$$

$$\hat{X}_R = \hat{R} \cdot Y \qquad \widehat{\overline{X}}_R = \hat{R} \cdot \overline{Y}$$

Observación:

La ganancia en precisión con el estimador del total por el método de la razón frente el estimador usual del total es tanto mayor si la correlación entre ambas variables cualitativas es alta y positiva

(Si la correlación es negativa no debe aplicarse este método)

III.3.3.- Estimador de Razón bajo M.A.E.

Si en el marco las unidades están agrupadas en L estratos y se obtiene una muestra estratificada, hay dos formas de obtener **el estimador del total X por el método de la razón bajo muestreo aleatorio estratificado**:

- **SEPARADO**
- **COMBINADO**

III.3.3.- Estimador de Razón bajo M.A.E.

SEPARADO

• Obtenemos la estimación separada del total por razón en cada estrato h:

$$\hat{X}_{Rh} = \hat{R}_h \cdot Y_h$$

• Efectuamos la suma para obtener el estimador final:

$$\hat{X}_{RS} = \sum_{h=1}^{L} \hat{X}_{Rh} = \sum_{h=1}^{L} \hat{R}_h Y_h$$

III.3.3.- Estimador de Razón bajo M.A.E.

COMBINADO

• Obtenemos la estimación de R como cociente de los estimadores insesgados del numerador y del denominador bajo muestreo estratificado:

$$\hat{R}_{st} = \frac{\hat{X}_{st}}{\hat{Y}_{st}}$$

• Se multiplica la suma anterior por el total Y para obtener el estimador final:

$$\hat{X}_{RC} = \hat{R}_{st} Y$$

III.3.3.- Estimador de Razón bajo M.A.E.

¿Cuál elegimos?

- El **estimador separado** requiere una información auxiliar más desagregada y tiene el riesgo de acumular el sesgo a lo largo de los estratos, si éste existiese y fuera siempre del mismo signo, positivo o negativo.
- Sin embargo, el estimador separado permite dar estimaciones separadas para cada estrato.
- El estimador separado tiene menor variabilidad al suponer que la razón no permanece constante de un estrato a otro.

Si disponemos de la información auxiliar necesaria y no haya riesgo de sesgos acumulados → Preferible el estimador separado

CAPÍTULO IV. Muestreo de conglomerados

CONTENIDOS

IV.1.- Muestreo de conglomerados sin submuestreo.

- IV.1.1.- Conglomerados de igual tamaño: Estimación de parámetros poblacionales.
- IV.1.2.- Conglomerados de tamaño desigual, selección de conglomerados con igual probabilidad: Estimación de parámetros poblacionales.
- IV.1.3.- Conglomerados de tamaño desigual, selección de conglomerados con probabilidades desiguales.

IV.2.- Muestreo de conglomerados con submuestreo

- IV.2.1.- Conglomerados de igual tamaño
- IV.2.2.- Conglomerado de tamaño desigual

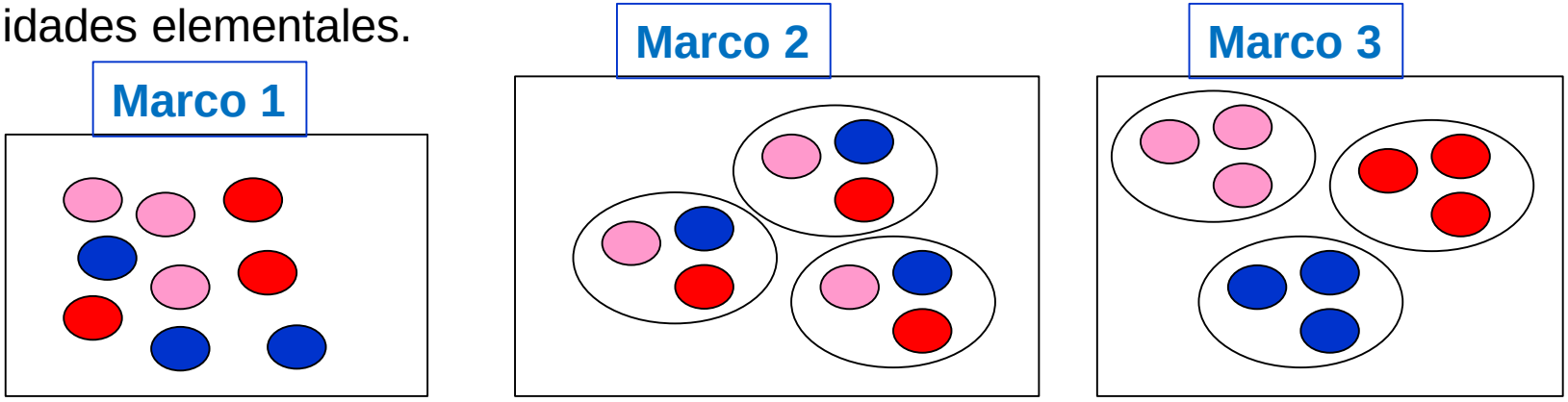
IV.- Muestreo de conglomerados

Cuando el marco disponible no contiene a las unidades de la población a investigar sino unidades mayores que son grupos de ellas (conglomerados), entonces no podemos utilizar ninguno de los muestreos vistos anteriormente y hemos de llevar a cabo un **muestreo de conglomerados.**

Conglomerado: unidad de muestreo formada por un grupo de unidades elementales.

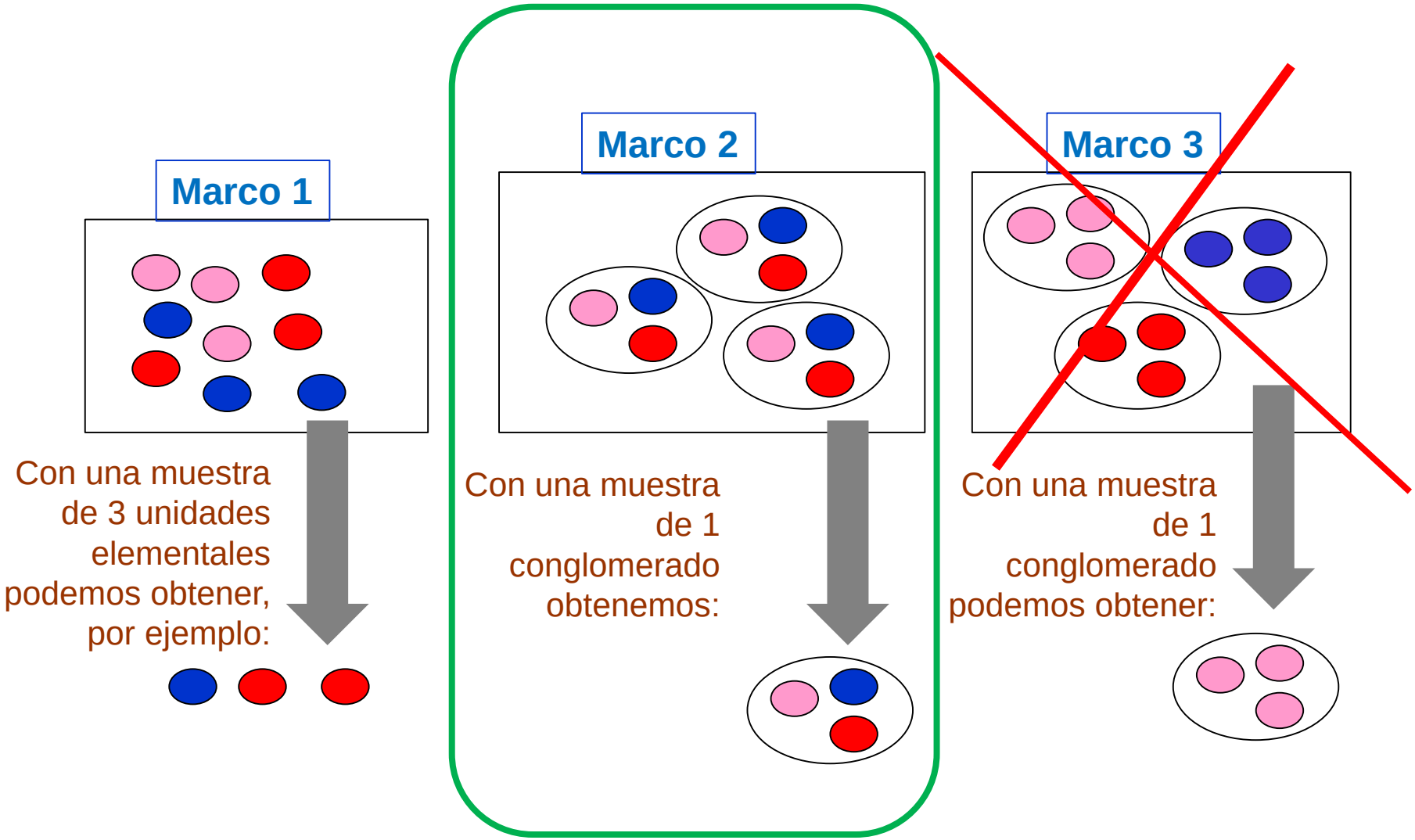
Ejemplo 4.1 Una lista de hospitales para investigar a personas hospitalizadas en cardiología nos obliga a seleccionar **conglomerados** (**hospitales**) para llegar a la población objetivo de estudio (personas hospitalizadas en cardiología).

En ocasiones podemos elegir entre el muestreo de conglomerados y el de unidades elementales.



IV.- Muestreo de conglomerados

¿Qué marco es preferible?



IV.- Muestreo de conglomerados

Dependiendo de si investigamos los conglomerados muestreados total o parcialmente, hablamos de:

Muestreo de conglomerados sin submuestreo o monoetápico (MCM): Se investigan todas las unidades elementales que componen los conglomerados muestreados.

Muestreo de conglomerados con submuestreo o bietápico (MCB): Se seleccionan aleatoriamente unidades elementales dentro de los conglomerados muestreados.

Dependiendo del número de unidades elementales que componen los conglomerados, hablamos de:

Conglomerados de igual tamaño: Todos los conglomerados están formados por el mismo número de unidades elementales.

Conglomerados de distinto tamaño: El número de unidades elementales que forman los conglomerados difieren.

IV.1.- Muestreo de conglomerados sin submuestreo

Muestreo de conglomerados sin submuestreo o monoetápico (MCM): Se investigan todas las unidades elementales que componen los conglomerados muestreados.

Notación:

Tamaños:	
N	Nº de conglomerados en la población
M_i	Nº de unidades elementales en el conglomerado i-ésimo
$M = \sum_{i=1}^{N} M_i$	Tamaño poblacional, e.d., nº de unidades elementales en la población
$\overline{M} = \frac{M}{N}$	Tamaño medio del conglomerado, e.d., número medio de unidades elementales por conglomerado
n	Nº de conglomerados muestreados

Ejemplo 4.1 Se quieren analizar diferentes características en personas hospitalizadas en el servicio de cardiología de los hospitales de una región y disponemos de una lista de los 500 hospitales de dicha región de los que seleccionamos 10.

Tamaños:	
$N = 500$	Nº de hospitales
M_i	Nº personas hospitalizadas en cardiología del hospital i-ésimo
M	Nº total de personas hospitalizas en cardiología en la región
$\overline{M} = \frac{M}{N}$	Número medio de personas hospitalizadas en cardiología por hospital
$n = 10$	Nº de hospitales muestreados

Notación característica numérica:

Características numéricas	
X_{ij}	Valor que toma la característica en estudio en la unidad elemental j-ésima del conglomerado i-ésimo.
$X_i = \sum_{j=1}^{M_i} X_{ij}$	Total del conglomerado i-ésimo
$\overline{X}_i = \frac{X_i}{M_i}$	Media del conglomerado i-ésimo
$X = \sum_{i=1}^{N} X_i$	Total poblacional
$\overline{\overline{X}} = \frac{X}{M}$	Media por unidad elemental, poblacional
$\overline{X} = \frac{X}{N}$	Media por conglomerado, poblacional
$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (X_{ij} - \overline{X}_i)^2$	Cuasivarianza del conglomerado i-ésimo
$\sigma_i^2 = \frac{(M_i - 1)}{M_i} S_i^2$	Varianza del conglomerado iésimo
$S^2 = \frac{1}{M - 1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (X_{ij} - \overline{\overline{X}})^2$	Cuasivarianza poblacional
$\sigma^2 = \frac{(M - 1)}{M} S^2$	Varianza poblacional

IV.1.- Muestreo de conglomerados sin submuestreo

Ejemplo 4.1. Una de las características a estudiar es el número de días que la persona permanece hospitalizada en el servicio de cardiología.

Característica numérica: Número de días de hospitalización	
X_{ij}	Número de días hospitalizada la persona j en el hospital i
$X_i = \sum_{j=1}^{M_i} X_{ij}$	Número total de días que las personas han estado hospitalizadas en cardiología en el hospital i
$\overline{X}_i = \frac{X_i}{M_i}$	Número medio de días de hospitalización, por persona, en el hospital i
$X = \sum_{i=1}^{N} X_i$	Total de días que las personas han estado hospitalizada en cardiología en la región
$\overline{\overline{X}} = \frac{X}{M}$	Número medio de días de hospitalización, por persona, en la región.
$\overline{X} = \frac{X}{N}$	Número medio de días de hospitalización, por hospital, en la región.

Notación característica cualitativa C:

Característica cualitativa	
A_i	Total de clase del conglomerado i-ésimo, e.d., total de unidades elementales que poseen la característica en el conglomerado i.
$P_i = \frac{A_i}{M_i}$	Proporción de unid. elem. del conglomerado i-que poseen la característica
$Q_i = 1 - P_i$	Proporción de unid. elem. del conglomerado i-que no poseen la característica
$A = \sum_{i=1}^{N} A_i$	Total de clase poblacional
$P = \frac{A}{M}$	Proporción de unidades elementales en la población que poseen la característica
$\overline{A} = \frac{A}{N}$	Número medio de unidades elementales por conglomerado que poseen la característica
$\sigma_i^2 = P_i Q_i$	Varianza del conglomerado i-ésimo
$\sigma^2 = PQ, \quad (Q = 1 - P)$	Varianza poblacional

(Si definimos $X_{ij} = \begin{cases} 1, & \text{si la unidad j del conglomerado i posee C,} \\ 0, & \text{si la unidad j del conglomerado i no posee C,} \end{cases}$ entonces $X_i = A_i, \overline{X}_i = P_i, X = A, \overline{\overline{X}} = P$).

Ejemplo 4.1. Otra de las características a estudiar es si la persona hospitalizada en el servicio de cardiología fallece en el hospital

Característica cualitativa	
A_i	Total de personas hospitalizadas en cardiología del hospital i que fallecen
$P_i = \frac{A_i}{M_i}$	Proporción de personas hospitalizadas en cardiología del hospital i que fallecen
$Q_i = 1 - P_i$	Proporción de personas hospitalizadas en cardiología del hospital i que no fallecen
$A = \sum_{i=1}^{N} A_i$	Total de personas hospitalizadas en cardiología que fallecen
$P = \frac{A}{M}$	Proporción de personas hospitalizadas en cardiología que fallecen
$\overline{A} = \frac{A}{N}$	Número medio de personas que fallecen en cardiología, por hospital.

IV.1.1.- Conglomerados de igual tamaño

Si todos los conglomerados están formados por el mismo número de unidades elementales, es decir, tienen el mismo tamaño, $M_i = \overline{M}$, entonces:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 + \frac{1}{N}\sum_{i=1}^{N}(\overline{X}_i - \overline{\overline{X}})^2$$

Varianza dentro de los conglomerados

σ_d^2

σ_c^2 Varianza entre los conglomerados

$$S_d^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2 = \frac{\overline{M}}{\overline{M}-1}\sigma_d^2$$

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\overline{X}_i - \overline{\overline{X}})^2 = \frac{N}{N-1}\sigma_c^2$$

Miden la variabilidad dentro de los conglomerados

Miden la variabilidad entre los conglomerados

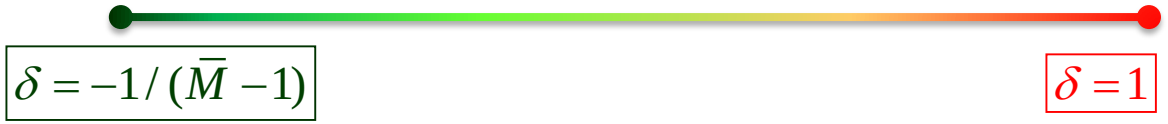
Cuanto mayor sea la variabilidad dentro de los conglomerados y menor la variabilidad entre conglomerados, mejor será el muestreo de conglomerados

IV.1.1.- Conglomerados de igual tamaño

Para medir la homogeneidad de los conglomerados tenemos el coeficiente de correlacción intraclase:

$$\delta = \frac{\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}\sum_{\substack{k=1\\k\neq j}}^{\bar{M}}(X_{ij}-\overline{\overline{X}})(X_{ik}-\overline{\overline{X}})}{(\bar{M}-1)\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(X_{ij}-\overline{\overline{X}})^2} = \frac{(\bar{M}-1)\sigma_c^2-\sigma_d^2}{(\bar{M}-1)\sigma^2} = \frac{\bar{M}(N-1)S_c^2-NS_d^2}{(\bar{M}N-1)S^2}$$

- Si los conglomerados son totalmente homogéneos dentro, este coeficiente toma el valor 1.
- Si los conglomerado son totalmente homogéneos entre ellos, este coeficiente toma el valor $-1/(\bar{M}-1)$.



IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Estimadores insesgados:

Características cuantitativas:

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\overline{X}} = \frac{\widehat{X}}{N} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{N\overline{M}} = \frac{1}{n}\sum_{i=1}^{n} \overline{X}_i.$$

Características cualitativas:

$$\widehat{A} = \frac{N}{n}\sum_{i=1}^{n} A_i, \quad \widehat{\overline{A}} = \frac{\widehat{A}}{N} = \frac{1}{n}\sum_{i=1}^{n} A_i, \quad \widehat{P} = \frac{\widehat{A}}{N\overline{M}} = \frac{1}{n}\sum_{i=1}^{n} P_i.$$

IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Varianzas de los estimadores:

Características cuantitativas:

Sin reposición	Con reposición
$V\left(\widehat{\overline{X}}\right)=\frac{(1-f)}{n}S_c^2, \quad f=\frac{n}{N}.$	$V\left(\widehat{\overline{X}}\right)=\frac{\sigma_c^2}{n}.$

$$V\left(\widehat{X}\right)=(N\overline{M})^2 V\left(\widehat{\overline{X}}\right); \quad V(\widehat{X})=\overline{M}^2 V\left(\widehat{\overline{X}}\right).$$

Características cualitativas:

Sin reposición	Con reposición
$V\left(\widehat{P}\right)=\frac{(1-f)}{n}S_c^2, \quad f=\frac{n}{N}.$	$V\left(\widehat{P}\right)=\frac{\sigma_c^2}{n}.$

$$V\left(\widehat{A}\right)=(N\overline{M})^2 V\left(\widehat{P}\right); \quad V(\widehat{A})=\overline{M}^2 V\left(\widehat{P}\right).$$

Los errores de muestreo se obtienen mediante la raíz cuadrada de la varianza

IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Comparación con el muestreo aleatorio simple de unidades elementales

Con reposición:

$$MCM(n):\quad V_{MCM}\left(\hat{\bar{X}}\right)=\frac{\sigma_c^2}{n}=\frac{\sigma^2}{n\bar{M}}(1+(\bar{M}-1)\delta)$$

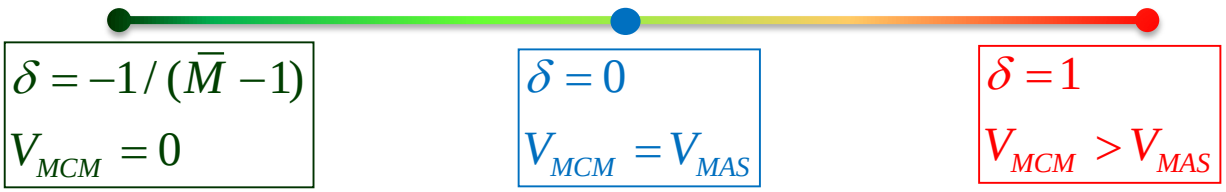
$$MAS(n\bar{M}):\quad V_{MAS}\left(\hat{\bar{X}}\right)=\frac{\sigma^2}{n\bar{M}}$$

Sin reposición:

$$MCM(n):\quad V_{MCM}\left(\hat{\bar{X}}\right)=\frac{(1-f)S_c^2}{n}\simeq\frac{(1-f)S^2}{n\bar{M}}(1+(\bar{M}-1)\delta)$$

$$MAS(n\bar{M}):\quad V_{MAS}\left(\hat{\bar{X}}\right)=\frac{(1-f)S^2}{n\bar{M}}$$

$$V_{MCM}=(1+(\bar{M}-1)\delta)V_{MAS}$$



IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Estimación de las varianzas de los estimadores:

Características cuantitativas:

$$\hat{V}\left(\hat{\bar{\bar{X}}}\right) = \frac{(1-f)}{n}\hat{S}_c^2 = \frac{\hat{S}_c^2}{n}, \quad \hat{S}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}_i - \hat{\bar{\bar{X}}})^2.$$

Sin reposición

Con reposición

$$\hat{V}\left(\hat{X}\right) = (N\bar{M})^2\hat{V}\left(\hat{\bar{\bar{X}}}\right); \quad \hat{V}(\hat{\bar{X}}) = \bar{M}^2\hat{V}\left(\hat{\bar{\bar{X}}}\right).$$

Características cualitativas:

$$\hat{V}\left(\hat{P}\right) = \frac{(1-f)}{n}\hat{S}_c^2 = \frac{\hat{S}_c^2}{n}, \quad \hat{S}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(P_i - \hat{P})^2.$$

Sin reposición

Con reposición

$$\hat{V}\left(\hat{A}\right) = (N\bar{M})^2\hat{V}\left(\hat{P}\right); \quad \hat{V}(\hat{\bar{A}}) = \bar{M}^2\hat{V}\left(\hat{P}\right).$$

Los errores de muestreo estimados se obtienen mediante la raíz cuadrada de la varianza estimada

IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Estimación del coeficiente de correlación intraclase:

Con reposición:

$$\hat{\delta} = \frac{(\bar{M}-1)\hat{S}_c^2 - \hat{\sigma}_d^2}{(\bar{M}-1)\hat{\sigma}^2}, \quad \hat{\sigma}_d^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2, \quad \hat{\sigma}^2 = \hat{S}_c^2 + \hat{\sigma}_d^2$$

Sin reposición:

$$\hat{\delta} = \frac{\bar{M}(N-1)\hat{S}_c^2 - N\hat{S}_d^2}{(N\bar{M}-1)\hat{S}^2}, \quad \hat{S}_d^2 = \frac{1}{n}\sum_{i=1}^{n}S_i^2, \quad \hat{S}^2 = \frac{1}{N\bar{M}-1}\left(N(\bar{M}-1)\hat{S}_d^2 + \bar{M}(N-1)\hat{S}_c^2\right).$$

IV.1.1.- Conglomerados de igual tamaño. Estimación de parámetros

Ejemplo 4.2. Disponemos un listado de 200 parejas de las que seleccionamos una m.a.s. de 20. Para cada una de las parejas seleccionadas preguntamos a ambas personas si trabajan o no y cuales son sus ingresos mensuales. Los datos siguientes muestran las respuestas obtenidas, donde la variable trabaja toma el valor 1 si la persona trabaja y los ingresos vienen dados en miles de €.

Persona	Pareja	Trabaja	Ingreso
1	1	1	3
2	1	0	0
3	2	0	0.6
4	2	1	1.9
5	3	1	0.8
6	3	1	2.2
7	4	1	1.3
8	4	0	0
9	5	0	1
10	5	1	3.2
11	6	1	1.8
12	6	1	2
13	7	1	0.9
14	7	1	2.2
15	8	1	1.9
16	8	1	1.5

a) Estima el porcentaje de personas que trabajan y su error de muestreo. Obtén un intervalo de confianza al 95%.
b) Estima el ingreso medio por persona y su error de muestreo.
c) Estima el ingreso medio por pareja y su error de muestreo.
d) Estima el porcentaje de parejas con ingresos superiores o iguales a 3000€ y su error de muestreo.

Ejemplo 4.2.

a) Estima el porcentaje de personas que trabajan y su error de muestreo. Obtén un intervalo de confianza al 95%.

$$\hat{P} = 0.825, \quad \hat{EM}(\hat{P}) = \sqrt{\frac{(1-0.1)0.05986842}{20}} = 0.0519,$$

$$IC_{0.95}(P) = (0.825 \mp 2 \cdot 0.0519) = (0.825 \mp 0.104) = (0.721; 0.929)$$

❖ Estimamos que el 82.5% de las personas trabajan, con un error de muestreo de 5.19%.

❖ Estimamos que el 82.5% de las personas trabajan, con un error de más-menos 10.4%, con una confianza del 95%.

b) Estima el ingreso medio por persona y su error de muestreo.

X_{ij} = ingreso de la persona j en la pareja i

$$\hat{\bar{X}} = 1.5375, \quad \hat{EM}(\hat{P}) = \sqrt{\frac{(1-0.1)0.2313}{20}} = 0.1020$$

❖ Estimamos que el ingreso medio por persona es de 1537.5€ con un error de muestreo de 102€.

Pareja i	Ai	Xi	Pi	$\bar{X}_i$
1	1	3	0.5	1.5
2	1	2.5	0.5	1.25
3	2	3	1	1.5
4	1	1.3	0.5	0.65
5	1	4.2	0.5	2.1
6	2	3.8	1	1.9
7	2	3.1	1	1.55
8	2	3.4	1	1.7
9	2	3.9	1	1.95
10	2	1.9	1	0.95
11	2	2.8	1	1.4
12	2	2.9	1	1.45
13	1	3.8	0.5	1.9
14	2	3.3	1	1.65
15	2	2.8	1	1.4
16	1	1.9	0.5	0.95
17	2	5.6	1	2.8
18	2	2.2	1	1.1
19	1	2.4	0.5	1.2
20	2	3.7	1	1.85
Media	1.65	3.075	0.825	1.5375

Ejemplo 4.2.

c) Estima el ingreso medio por pareja y su error de muestreo.

$$\widehat{\overline{X}} = 3.075, \quad \hat{EM}(\widehat{\overline{X}}) = \sqrt{\frac{(1-0.1)0.9251}{20}} = 2\hat{EM}(\widehat{\overline{X}}) = 0.204$$

❖ Estimamos que el ingreso medio por pareja es de 3075€ con un error de muestreo de 204€.

d) Estima el porcentaje de parejas con ingresos superiores o iguales a 3000€ y su error de muestreo.

$$P' = \text{proporción de parejas con ingresos } \geq 3000€$$

$$\widehat{P}' = 0.55, \quad \hat{EM}(\hat{P}) = \sqrt{\frac{(1-0.1)0.55 \cdot 0.45}{19}} = 0.1083$$

❖ Estimamos que el 55% de las parejas tienen ingresos superiores o iguales a 3000€ con un error de muestreo de 10.83%

Pareja i	Ai	Xi	Pi	$\overline{X}_i$
1	1	3	0.5	1.5
2	1	2.5	0.5	1.25
3	2	3	1	1.5
4	1	1.3	0.5	0.65
5	1	4.2	0.5	2.1
6	2	3.8	1	1.9
7	2	3.1	1	1.55
8	2	3.4	1	1.7
9	2	3.9	1	1.95
10	2	1.9	1	0.95
11	2	2.8	1	1.4
12	2	2.9	1	1.45
13	1	3.8	0.5	1.9
14	2	3.3	1	1.65
15	2	2.8	1	1.4
16	1	1.9	0.5	0.95
17	2	5.6	1	2.8
18	2	2.2	1	1.1
19	1	2.4	0.5	1.2
20	2	3.7	1	1.85
Media	1.65	3.075	0.825	1.5375

IV.1.2.- Conglomerados de tamaño desigual.

Si la variación en los tamaños es muy grande, el hecho de incluir unos conglomerados u otros en la muestra puede hacer variar mucho el valor del estimador o, dicho de otra forma, el efecto de la variación del tamaño de los conglomerados sobre la varianza del estimador puede llegar a ser importante.

Distintas maneras de paliar este efecto cuando se sospecha que puede ser grave:

- **Estratificación de los conglomerados por tamaño.**
  Se realiza estimación separada por estratos formados por conglomerados de tamaño similar. Requiere conocer a priori los tamaños de todos los conglomerados, lo que en la práctica puede ser difícil, pero es posible aproximar esos tamaños con variables auxiliares, si es solamente para la formación de estratos.
- **Estimación de razón a tamaño.**
  Es uno de los métodos más eficaces si no se puede estratificar. Normalmente, el total por conglomerado está relacionado de manera proporcional con el tamaño, por lo que la estimación de razón está plenamente justificada.
- **Selección de conglomerados con probabilidades desiguales.**
  Asignando probabilidades mayores a los conglomerados de mayor tamaño. Requiere conocer a priori el tamaño de todos los conglomerados o una variable auxiliar muy correlacionada con el tamaño. Este método permite evitar el riesgo que existe en el MAS de que los conglomerados grandes (más importantes en términos relativos para el investigador) queden fuera de la muestra, o que puedan estar en la muestra conglomerados muy pequeños con escasa representatividad.

IV.1.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\widehat{X}=\frac{N}{n}\sum_{i=1}^{n}X_i,\quad \widehat{\overline{X}}=\frac{\widehat{X}}{N}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \widehat{\overline{\overline{X}}}=\frac{\widehat{X}}{M}.$$

$$\widehat{A}=\frac{N}{n}\sum_{i=1}^{n}A_i,\quad \widehat{\overline{A}}=\frac{\widehat{A}}{N}=\frac{1}{n}\sum_{i=1}^{n}A_i,\quad \widehat{P}=\frac{\widehat{A}}{M}.$$

Necesitamos conocer M

Varianzas

Sin reposición	Con reposición
$V\left(\widehat{X}\right)=\frac{N^2(1-f)}{n(N-1)}\sum_{i=1}^{N}(X_i-\overline{X})^2.$	$V\left(\widehat{X}\right)=\frac{N^2}{nN}\sum_{i=1}^{N}(X_i-\overline{X})^2.$

$$V\left(\widehat{\overline{X}}\right)=\frac{V\left(\widehat{X}\right)}{N^2};\quad V\left(\widehat{\overline{\overline{X}}}\right)=\frac{V\left(\widehat{X}\right)}{M^2}.$$

Sin reposición	Con reposición
$V\left(\widehat{A}\right)=\frac{N^2(1-f)}{n(N-1)}\sum_{i=1}^{N}(A_i-\overline{A})^2.$	$V\left(\widehat{A}\right)=\frac{N^2}{nN}\sum_{i=1}^{N}(A_i-\overline{A})^2.$

$$V(\widehat{\overline{A}})=\frac{V\left(\widehat{A}\right)}{N^2};\quad V\left(\widehat{P}\right)=\frac{V\left(\widehat{A}\right)}{M^2}.$$

IV.1.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\overline{X}} = \frac{\widehat{X}}{N} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{M}.$$

$$\widehat{A} = \frac{N}{n}\sum_{i=1}^{n} A_i, \quad \widehat{\overline{A}} = \frac{\widehat{A}}{N} = \frac{1}{n}\sum_{i=1}^{n} A_i, \quad \widehat{P} = \frac{\widehat{A}}{M}.$$

Estimación de varianzas

Sin reposición	Con reposición
$\widehat{V}\left(\widehat{X}\right) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n}(X_i - \widehat{\overline{X}})^2.$	$\widehat{V}\left(\widehat{X}\right) = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(X_i - \widehat{\overline{X}})^2.$

$$\widehat{V}\left(\widehat{\overline{X}}\right) = \frac{\widehat{V}\left(\widehat{X}\right)}{N^2}; \quad \widehat{V}\left(\widehat{\overline{\overline{X}}}\right) = \frac{\widehat{V}\left(\widehat{X}\right)}{M^2}.$$

Sin reposición	Con reposición
$\widehat{V}\left(\widehat{A}\right) = \frac{N^2(1-f)}{n(n-1)}\sum_{i=1}^{n}(A_i - \widehat{\overline{A}})^2.$	$\widehat{V}\left(\widehat{A}\right) = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(A_i - \widehat{\overline{A}})^2.$

$$\widehat{V}(\widehat{\overline{A}}) = \frac{\widehat{V}\left(\widehat{A}\right)}{N^2}; \quad \widehat{V}\left(\widehat{P}\right) = \frac{\widehat{V}\left(\widehat{A}\right)}{M^2}.$$

IV.1.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores de razón al tamaño:

$$\hat{\bar{X}}_R = \frac{\hat{X}}{\hat{M}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}, \quad \hat{X}_R = \hat{\bar{X}}_R \cdot M$$

$$\hat{P}_R = \frac{\hat{A}}{\hat{M}} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i}, \quad \hat{A}_R = \hat{P}_R \cdot M$$

Estos estimadores son sesgados

Varianzas

Suponiendo que el sesgo es despreciable:

Sin reposición	Con reposición
$V\left(\hat{\bar{X}}_R\right) = \frac{(1-f)}{n\bar{M}^2(N-1)}\sum_{i=1}^{N}(X_i - \bar{\bar{X}}M_i)^2.$	$V\left(\hat{\bar{X}}_R\right) = \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(X_i - \bar{\bar{X}}M_i)^2.$

$$V\left(\hat{X}_R\right) = M^2 V\left(\hat{\bar{X}}_R\right).$$

Sin reposición	Con reposición
$V\left(\hat{P}_R\right) = \frac{(1-f)}{n\bar{M}^2(N-1)}\sum_{i=1}^{N}(A_i - P\cdot M_i)^2.$	$V\left(\hat{P}_R\right) = \frac{1}{n\bar{M}^2 N}\sum_{i=1}^{N}(A_i - P\cdot M_i)^2.$

$$V\left(\hat{A}_R\right) = M^2 \cdot V\left(\hat{P}_R\right)$$

IV.1.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores de razón al tamaño:

Suponiendo que el sesgo es despreciable:

Estimación de varianzas

Sin reposición	Con reposición
$\hat{V}\left(\hat{\bar{\bar{X}}}_R\right)=\frac{(1-f)}{n\hat{\bar{\bar{M}}}^2(n-1)}\sum_{i=1}^{n}(X_i-\hat{\bar{\bar{X}}}_R M_i)^2.$	$\hat{V}\left(\hat{\bar{\bar{X}}}_R\right)=\frac{1}{n\hat{\bar{\bar{M}}}^2(n-1)}\sum_{i=1}^{n}(X_i-\hat{\bar{\bar{X}}}_R M_i)^2.$

$$\hat{V}\left(\hat{X}_R\right)=M^2\hat{V}\left(\hat{\bar{\bar{X}}}_R\right).$$

Sin reposición	Con reposición
$\hat{V}\left(\hat{P}_R\right)=\frac{(1-f)}{n\hat{\bar{\bar{M}}}^2(n-1)}\sum_{i=1}^{n}(A_i-\hat{P}_R M_i)^2.$	$\hat{V}\left(\hat{P}_R\right)=\frac{1}{n\hat{\bar{\bar{M}}}^2(n-1)}\sum_{i=1}^{n}(A_i-\hat{P}_R M_i)^2.$

$$\hat{V}\left(\hat{A}_R\right)=M^2\cdot\hat{V}\left(\hat{P}_R\right)$$

IV.1.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Ejemplo 4.3. Una empresa con 50 grandes superficies comerciales quiere tener información acerca de la falta de los empleados al trabajo por enfermedad. Selecciona una m.a.s. de 5 superficies y recoge información sobre el nº de empleados, el nº de horas perdidas por enfermedad en un mes y el nº de empleados "enfermos", entendiendo por empleado "enfermo" aquel que ha faltado al menos un día en el mes por enfermedad. Los datos obtenidos son:

Superficie	Nº empleados	Nº total horas perdidas	Nº empleados enfermos
1	160	184	3
2	150	80	2
3	120	80	4
4	140	120	2
5	150	48	6

Estima los siguientes parámetros junto con sus errores de muestreo:.

a) Total de horas perdidas por enfermedad en un mes.
b) Porcentaje de empleados enfermos en un mes.
c) Número medio de empleados enfermos por superficie.
d) Número medio de horas perdidas por empleado.
e) Número medio de horas perdidas por empleado enfermo.

Si la empresa sabe que tiene un total de 7300 empleados, ¿modificarías algunas de las estimaciones previas?

Ejemplo 4.3.

a) Total de horas perdidas por enfermedad en un mes.

X_{ij} = nº horas perdidas, por enfermedad, el empleado j de la superficie i.

$$\hat{X} = \frac{50}{5}512 = 5120, \quad \hat{EM}(\hat{X}) = \sqrt{\frac{50^2(1-0.1)2732.8}{5}} = 1109$$

Superficie	Nº empleados	Nº total horas perdidas	Nº empleados enfermos
1	160	184	3
2	150	80	2
3	120	80	4
4	140	120	2
5	150	48	6
Total	720	512	17

❖ Estimamos que se pierden un total de 5120 horas al mes por enfermedad en la empresa, con un error de muestreo de 1109 horas.

b) Porcentaje de empleados enfermos en un mes.

$$\widehat{P}_R = \frac{17}{720} = 0.0236, \quad \hat{EM}(\widehat{P}_R) = \sqrt{\frac{(1-0.1)3.0227}{5 \cdot 144^2}} = 0.0051$$

❖ Estimamos que el 2.36% de los empleados ha estado enfermo en un mes, con un error de muestreo de 0.51%

c) Número medio de empleados enfermos por superficie.

$$\hat{A} = \frac{50}{5}17 = 170, \quad \hat{EM}(\hat{A}) = \sqrt{\frac{50^2(1-0.1)2.8}{5}} = 35.4965,$$

$$\hat{\bar{A}} = \frac{\hat{A}}{50} = 3.4, \quad \hat{EM}(\hat{\bar{A}}) = \frac{\hat{EM}(\hat{A})}{50} = 0.71$$

❖ Estimamos que el número medio de empleados enfermos por superficie es 3.4 al mes con un error de muestreo de 0.71.

Ejemplo 4.3.

d) Número medio de horas perdidas por empleado.

$$\hat{\bar{X}}_R = \frac{512}{720} = 0.71, \quad \hat{EM}(\hat{\bar{X}}_R) = \sqrt{\frac{(1-0.1)2382.62}{5 \cdot 144^2}} = 0.1438$$

Superficie	Nº empleados	Nº total horas perdidas	Nº empleados enfermos
1	160	184	3
2	150	80	2
3	120	80	4
4	140	120	2
5	150	48	6
Total	720	512	17

❖ Estimamos que cada empleado pierde, en media, 0.71 horas por enfermedad al mes, con un error de muestreo de 0.1438 horas.

e) Número medio de horas perdidas por empleado enfermo.

$$\hat{R} = \frac{\hat{X}}{\hat{A}} = 30.12, \quad \hat{EM}(\hat{R}) = \sqrt{\frac{(1-0.1)7995.24}{5 \cdot 3.4^2}} = 11.16$$

❖ Estimamos que cada empleado enfermo pierde, en media, 30.12 horas al mes, con un error de muestreo de 11.16 horas.

Si la empresa sabe que tiene un total de 7300 empleados, ¿modificarías algunas de las estimaciones previas?

· Los apartados a)-d) pueden resolverse mediante estimadores insesgados o de razón.

$$\widehat{ERM}(\hat{M}) = \frac{\widehat{EM}(\hat{M})}{\hat{M}} = 0.045 < 0.1$$

Sesgo despreciable

Ejemplo 4.3.

Si la empresa sabe que tiene un total de 7300 empleados, ¿modificarías algunas de las estimaciones previas?

Los apartados a)-d) pueden resolverse mediante estimadores insesgados o de razón.

$$\widehat{ERM}(\widehat{M}) = \frac{\widehat{EM}(\widehat{M})}{\widehat{M}} = 0.045 < 0.1$$

Sesgo despreciable

Comparando los errores relativos de muestreo de los estimadores insesgados y de razón:

$$\widehat{ERM}(\widehat{X}) = 0.2166, \quad \widehat{ERM}(\widehat{X}_R) = 0.2022.$$

$$\widehat{ERM}(\widehat{A}) = 0.2088, \quad \widehat{ERM}(\widehat{A}_R) = 0.2169.$$

Similares resultados para estimadores insesgados y de razón, algo mejores los de razón para la característica cuantitativa y los insesgados para la cualitativa.

Modificamos las estimaciones de los apartados a) y b):

a) Total de horas perdidas por enfermedad en un mes: $\widehat{X}_R = 5191, \quad \widehat{EM}(\widehat{X}) = 1049$

b) Porcentaje de empleados enfermos en un mes: $\widehat{P} = 0.0233, \quad \widehat{EM}(\widehat{P}) = 0.0049$

IV.1.3.- Conglomerados de tamaño desigual. Selección de conglomerados con diferentes probabilidades

- **Estimadores de Hansen y Hurwitz (selección con reposición)**
- **Estimadores de Horvitz y Thompson (selección sin reposición)**

Estimadores de Hansen y Hurwitz (selección con reposición)

$$\widehat{X}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{p_i}.$$

$$V\left(\widehat{X}_{HH}\right) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{p_i} - X\right)^2 p_i.$$

$$\widehat{V}\left(\widehat{X}_{HH}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{X_i}{p_i} - \widehat{X}\right)^2.$$

✓ p_i es la probabilidad de seleccionar el conglomerado i.

✓ Si los conglomerados se seleccionan con probabilidad proporcional a su tamaño $p_i = M_i / M$.

IV.1.3.- Conglomerados de tamaño desigual. Selección de conglomerados con diferentes probabilidades

- **Estimadores de Hansen y Hurwitz (selección con reposición)**
- **Estimadores de Horvitz y Thompson (selección sin reposición)**

Estimadores de Horvitz y Thompson (selección sin reposición)

$$\widehat{X}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}.$$

$$V\left(\widehat{X}_{HT}\right) = \sum_{i=1}^{N} X_i^2 \frac{(1-\pi_i)}{\pi_i} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} X_i X_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}.$$

$$\widehat{V}_1\left(\widehat{X}_{HT}\right) = \sum_{i=1}^{n} X_i^2 \frac{(1-\pi_i)}{\pi_i^2} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} X_i X_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}}.$$

$$\widehat{V}_2\left(\widehat{X}_{HT}\right) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j}\right)^2.$$

$\pi_i \equiv$ probabilidad de inclusión de primer orden.

$\pi_{ij} \equiv$ probabilidad de inclusión de segundo orden

IV.1.3.- Conglomerados de tamaño desigual. Selección de conglomerados con diferentes probabilidades

Estimadores de Hansen y Hurwitz (selección con reposición)

Para ilustrar la selección con probabilidades diferentes veamos un ejemplo en el que teóricamente conocemos toda la población.

Ejemplo 4.4. Una ciudad tiene tiendas de 4 cadenas alimenticias (A, B, C y D) variando el número de tiendas por cadena entre 10 y 100. Queremos estimar el total de ventas del último mes, en las cuatro cadenas, con una muestra de dos cadenas.
Supongamos que conocemos las ventas del últimos mes en cada cadena (*observemos que es sólo un ejemplo ilustrativo porque si realmente conociésemos las ventas de todas las cadenas no necesitaríamos estimar el total)*

Cadena	Tiendas	Ventas (en millones)
A	10	11
B	20	20
C	30	24
D	100	245

Seleccionemos la muestra de 2 cadenas **con reposición** (método de la probabilidad acumulada y método de Lahiri)

Estimadores de Hansen y Hurwitz (selección con reposición)

Ejemplo 4.4.

Seleccionamos la muestra de 2 cadenas **con reposición** por el método de la probabilidad acumulada: seleccionamos 2 números aleatorios ente 1 y 160, supongamos que los números seleccionados han sido **10** y **52,** que corresponden a las cadenas A y C.

Cadena	Tiendas	Tiendas Acumul
A	10	10
B	20	30
C	30	60
D	100	160
Total	160	

$$\widehat{X}_{HH} = \frac{1}{2}\left(\frac{11}{0.0625} + \frac{24}{0.1875}\right) = 152.$$

$$\widehat{V}\left(\widehat{X}_{HH}\right) = \frac{1}{2}\left(\left(\frac{11}{0.0625} - 152\right)^2 + \left(\frac{24}{0.1875} - 152\right)^2\right) = 575.$$

$$\widehat{EM}\left(\widehat{X}_{HH}\right) = 24.$$

Estimamos que el total de ventas en el mes han sido de 152 millones de euros con un error de muestreo de 24 millones.

Estimador de Horvitz y Thompson (selección sin reposición)

Ejemplo 4.4.

Seleccionamos la muestra de 2 cadenas **sin reposición** por el método de la probabilidad acumulada: seleccionamos 1 número aleatorio ente 1 y 160, supongamos que ha sido **59,** que corresponden a la cadena C. Seleccionamos un número aleatorio entre 1 y 130, supongamos que ha sido **103,** que corresponde a la cadena D.

Cadena	Tiendas	T. Acum. 1ª selec	T. Acum. 2ª selec
A	10	10	10
B	20	30	30
C	30	60	
D	100	160	130
Total	160		

Estimamos que el total de ventas en el mes ha sido de 316.67 millones.

$$\pi_3 = p_3\left(1+\frac{p_1}{1-p_1}+\frac{p_2}{1-p_2}+\frac{p_4}{1-p_4}\right)=0.5393,$$

$$\pi_4 = p_4\left(1+\frac{p_1}{1-p_1}+\frac{p_2}{1-p_2}+\frac{p_3}{1-p_3}\right)=0.9002,$$

$$\pi_{34} = \frac{p_3 p_4}{1-p_3}+\frac{p_3 p_4}{1-p_4}=0.4567,$$

$$\widehat{X}_{HT} = \frac{24}{0.5393}+\frac{245}{0.9002}=316.67$$

$$\widehat{V}_1\left(\widehat{X}_{HT}\right)=6782.817, \quad \widehat{EM}_1\left(\widehat{X}_{HT}\right)=82.36$$

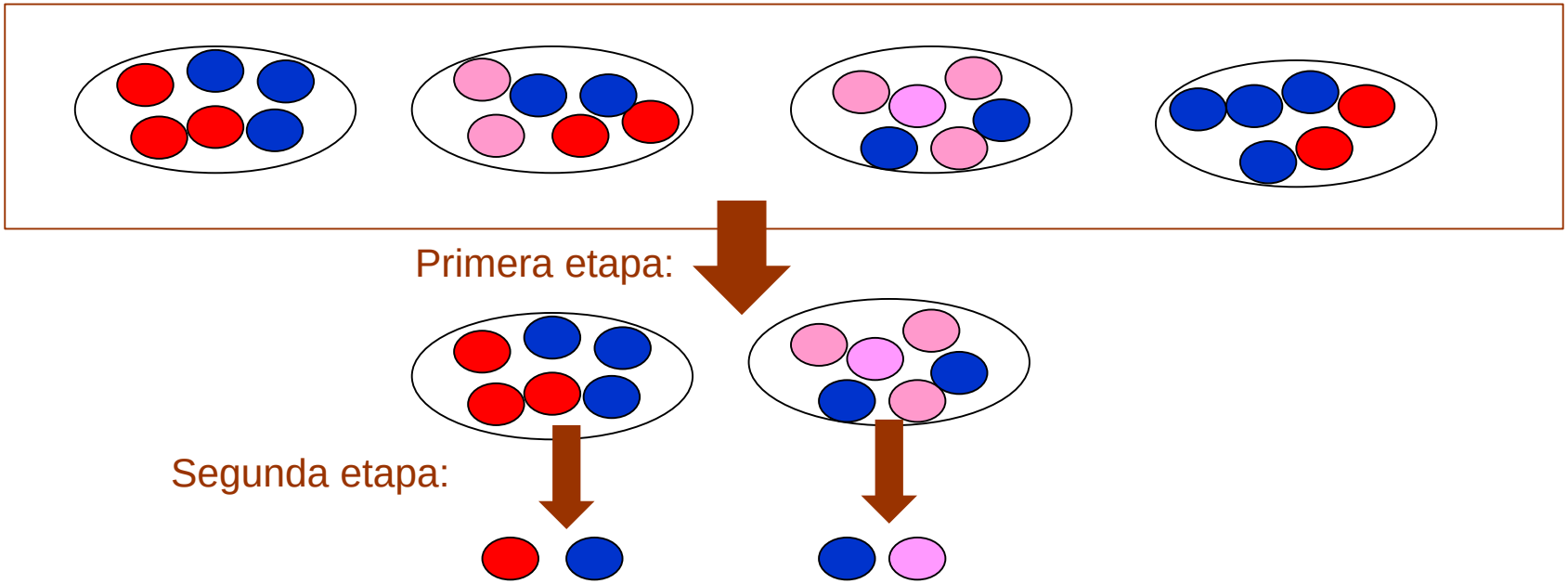
$$\widehat{V}_2\left(\widehat{X}_{HT}\right)=3259.784, \quad \widehat{EM}_2\left(\widehat{X}_{HT}\right)=57.09$$

IV.2.- Muestreo de conglomerados con submuestreo

Muestreo de conglomerados con submuestreo o bietápico (MCB): Cuando las unidades elementales que componen los conglomerados son homogéneas se selecciona parte de las unidades elementales de cada conglomerado seleccionado. Tenemos dos etapas en el muestreo:

Primera etapa: selección aleatoria de conglomerados.

Segunda etapa: selección aleatoria de unidades elementales dentro de cada conglomerado seleccionado en primera etapa



IV.2.- Muestreo de conglomerados con submuestreo

Muestreo de conglomerados con submuestreo o bietápico (MCB): Se selecciona parte de las unidades elementales que componen los conglomerados muestreados. Tenemos dos etapas en el muestreo:

Primera etapa: selección aleatoria de conglomerados.

Segunda etapa: selección aleatoria de unidades elementales dentro de cada conglomerado seleccionado en primera etapa

Notación:

Tamaños:		
Primera etapa	N	Nº de conglomerados en la población
	n	Nº de conglomerados muestreados
	$f_1 = \frac{n}{N}$	Fracción de muestreo de primera etapa
Segunda etapa	M_i	Nº de unidades elementales en el conglomerado i-ésimo
	m_i	Nº de unidades elementales seleccionadas en el conglomerado i-ésimo
	$f_{2i} = \frac{m_i}{M_i}$	Fracción de muestreo de segunda etapa para el conglomerado i-ésimo

Ejemplo 4.4 Se quieren analizar diferentes características en personas hospitalizadas en el servicio de cardiología de los hospitales de una región. Disponemos de una lista de los 500 hospitales de dicha región de los que seleccionamos 10 y en cada uno de éstos seleccionamos el 50% de las personas hospitalizadas.

Tamaños:	
$N = 500$	Nº de hospitales
M_i	Nº personas hospitalizadas en cardiología del hospital i-ésimo
M	Nº total de personas hospitalizas en cardiología en la región
$\overline{M} = \frac{M}{N}$	Número medio de personas hospitalizadas en cardiología por hospital
$n = 10$	Nº de hospitales muestreados
$f_1 = 0.02$	Proporción de hospitales muestreados
$f_{2i} = 0.5$	Proporción de personas hospitalizadas muestreadas en cada hospital
$m_i = 0.5 \cdot M_i$	Nº de personas muestreadas en cada hospital

La notación para los parámetros poblacionales es la misma que la dada en el MCM

Notación muestral, característica cuantitativa:

Característica cuantitativa	
$x_i = \sum_{j=1}^{m_i} X_{ij}$	Total muestral del conglomerado i-ésimo
$\overline{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{ij}$	Media muestral del conglomerado i-ésimo
$\hat{S}_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(X_{ij} - \overline{x}_i \right)^2$	Cuasivarianza muestral del coglomerado i-esimo

Notación muestal, característica cualitativa C:

Característica cualitativa	
a_i	Total de unidades elementales de la muestra del conglomerado i que poseen la característica.
$p_i = \frac{a_i}{m_i}$	Proporción muestral de unidades elementales del conglomerado i que poseen la característica
$q_i = 1 - p_i$	Proporción muestral de unidades elementales del conglomerado i que no poseen la característica
$p_i q_i$	Varianza muestral del conglomerado i-ésimo
$\frac{m_i}{m_i - 1} p_i q_i$	Cuasivarianza muestral del conglomerado i-ésimo

IV.2.1.- Conglomerados de igual tamaño. Estimación de parámetros

$M_i = \overline{M}$

Primera etapa: Selección de conglomerados mediante MAS $f_1 = \frac{n}{N}$

Segunda etapa: Selección de unidades elementales dentro de cada conglomerado mediante MAS.

$f_2 = \frac{\overline{m}}{\overline{M}}$, $\overline{m} \equiv$ nº de unidades muestreadas dentro de cada conglomerado.

Estimadores insesgados:

Características cuantitativas:

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n}\hat{X}_i = \frac{N}{n}\sum_{i=1}^{n}\left(\frac{\overline{M}}{\overline{m}}\sum_{j=1}^{\overline{m}}X_{ij}\right) = \frac{N}{n}\sum_{i=1}^{n}\overline{M}\overline{x}_i,$$

$$\widehat{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n}\overline{x}_i = \frac{\hat{X}}{N\overline{M}},$$

$$\hat{\overline{X}} = \frac{\hat{X}}{N} = \overline{M}\,\widehat{\overline{X}}.$$

Características cualitativas:

$$\hat{A} = \frac{N}{n}\sum_{i=1}^{n}\hat{A}_i = \frac{N}{n}\sum_{i=1}^{n}\overline{M}p_i,$$

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n}p_i = \frac{\hat{A}}{N\overline{M}},$$

$$\hat{\overline{A}} = \frac{\hat{A}}{N} = \overline{M}\,\hat{P}.$$

IV.2.1.- Conglomerados de igual tamaño. Estimación de parámetros

Varianzas de los estimadores:

Características cuantitativas:

Sin reposición	Con reposición (en ambas etapas)
$V\left(\hat{\bar{\bar{X}}}\right)=\frac{(1-f_1)}{n}S_c^2+\frac{(1-f_2)}{n\bar{m}}S_d^2.$	$V\left(\hat{\bar{\bar{X}}}\right)=\frac{\sigma_c^2}{n}+\frac{\sigma_d^2}{n\bar{m}}.$

$$V\left(\hat{X}\right)=(N\bar{M})^2V\left(\hat{\bar{\bar{X}}}\right);\quad V(\hat{\bar{X}})=\bar{M}^2V\left(\hat{\bar{\bar{X}}}\right).$$

Características cualitativas:

Sin reposición	Con reposición (en ambas etapas
$V\left(\hat{P}\right)=\frac{(1-f_1)}{n}S_c^2+\frac{(1-f_2)}{n\bar{m}}S_d^2.$	$V\left(\hat{P}\right)=\frac{\sigma_c^2}{n}+\frac{\sigma_d^2}{n\bar{m}}.$

$$V\left(\hat{A}\right)=(N\bar{M})^2V\left(\hat{P}\right);\quad V(\hat{\bar{A}})=\bar{M}^2V\left(\hat{P}\right).$$

Los errores de muestreo se obtienen mediante la raíz cuadrada de la varianza

IV.2.1.- Conglomerados de igual tamaño. Estimación de parámetros

Estimación de las varianzas de los estimadores:

Características cuantitativas:

$$\hat{V}\left(\hat{\bar{\bar{X}}}\right)=\frac{(1-f_1)}{n}s_c^2+\frac{f_1(1-f_2)}{n\bar{m}}s_d^2,\quad s_c^2=\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_i-\hat{\bar{\bar{X}}})^2,\quad s_d^2=\frac{1}{n}\sum_{i=1}^{n}\hat{S}_i^2.$$

Sin reposición

$$\hat{V}\left(\hat{X}\right)=(N\bar{M})^2\hat{V}\left(\hat{\bar{\bar{X}}}\right);\quad \hat{V}(\hat{\bar{X}})=\bar{M}^2\hat{V}\left(\hat{\bar{\bar{X}}}\right).$$

Características cualitativas:

$$\hat{V}\left(\hat{P}\right)=\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(p_i-\hat{P})^2+\frac{f_1(1-f_2)}{n^2(\bar{m}-1)}\sum_{i=1}^{\bar{m}}p_iq_i.$$

Sin reposición

$$\hat{V}\left(\hat{A}\right)=(N\bar{M})^2\hat{V}\left(\hat{P}\right);\quad \hat{V}(\hat{\bar{A}})=\bar{M}^2\hat{V}\left(\hat{P}\right).$$

Los errores de muestreo estimados se obtienen mediante la raíz cuadrada de la varianza estimada

IV.2.1.- Conglomerados de igual tamaño. Estimación de parámetros

Ejemplo 4.5. Una ciudad tiene 30 centros de salud y cada uno de ellos tiene 8 médicos de familia. Se selecciona una muestra aleatoria simple de 3 centros y en cada uno se seleccionan 2 médicos mediante MAS. Se obtiene información sobre el número de pacientes vistos por el médico en un día y el número de pacientes derivados a un especialistas. Los datos obtenidos son:.

C.Salud	Medico	Nº pac. Vistos (Xij)	Nº pac. deriv (Yij)
1	1	44	6
	2	18	6
2	1	42	3
	2	10	2
3	1	16	5
	2	32	14

a) Estima el total de pacientes vistos en la ciudad en un día y su error de muestreo.
b) Estima el total de pacientes derivados al especialista en un día y su error de muestreo.
c) Estima el número medio de pacientes vistos por un médico al día y su error de muestreo.
d) Estima el número medio de pacientes derivados a un especialista por centro.

Ejemplo 4.5.

C.Salud	Medico	Nº pac. vistos(Xij)	Nº pac. deriv (Yij)	FactElev*Xij	FactElev*Yij
1	1	44	6	1760	240
	2	18	6	720	240
2	1	42	3	1680	120
	2	10	2	400	80
3	1	16	5	640	200
	2	32	14	1280	560
			Total	6480	1440

a) Estima el total de pacientes vistos en la ciudad en un día y su error de muestreo.

X_{ij} = Nº de pacientes atendidos por el médico j del centro i

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n}\hat{X}_i = \frac{N}{n}\sum_{i=1}^{N}\left(\frac{\bar{M}}{\bar{m}}\sum_{j=1}^{\bar{m}}X_{ij}\right) = 6480$$

C.Salud	Med_X	Cuasiv_X	Med_Y	Cuasiv_Y
1	31	338	6	0
2	26	512	2.5	0.5
3	24	128	9.5	40.5
	Xij	**Yij**		
Media	27	6		
Cuasiv entre	13	12.25		

$$\hat{V}\left(\widehat{X}\right) = (N\bar{M})^2\hat{V}\left(\widehat{\bar{X}}\right) = 459360 \Rightarrow \widehat{EM}\left(\widehat{X}\right) = 677.76$$

$$\hat{V}\left(\widehat{\bar{X}}\right) = \frac{0.9\cdot 13}{3} + \frac{0.1\cdot 0.75\cdot 326}{3\cdot 2} = 7.975$$

❖ Estimamos que en la ciudad se atienden un total de 6480 pacientes al día con un error de muestreo de 678 pacientes.

Ejemplo 4.5.

C.Salud	Medico	Nº pac. vistos(Xij)	Nº pac. deriv (Yij)	FactElev*Xij	FactElev*Yij
1	1	44	6	1760	240
	2	18	6	720	240
2	1	42	3	1680	120
	2	10	2	400	80
3	1	16	5	640	200
	2	32	14	1280	560
			Total	6480	1440

b) Estima el total de pacientes derivados al especialista en un día y su error de muestreo.

Y_{ij} = Nº de pacientes derivados a un especialista por el médico j del centro i

C.Salud	Med_X	Cuasiv_X	Med_Y	Cuasiv_Y
1	31	338	6	0
2	26	512	2.5	0.5
3	24	128	9.5	40.5
	Xij	**Yij**		
Media	27	6		
Cuasiv entre	13	12.25		
Cuasiv dentro	326	13.666667		

$$\hat{Y} = 1440$$

$$\hat{V}\left(\hat{Y}\right) = 221520 \rightarrow \widehat{EM}\left(\hat{Y}\right) = 470.659$$

$$\hat{V}\left(\hat{\hat{Y}}\right) = \frac{0.9 \cdot 12.25}{3} + \frac{0.1 \cdot 0.75 \cdot 13.67}{3 \cdot 2} = 3.8458$$

❖ Estimamos que en la ciudad se derivan a un especialista un total de 1440 pacientes al día con un error de muestreo de 471 pacientes.

Ejemplo 4.5.

c) Estima número medio de pacientes vistos por un médico al día y su error de muestreo.

$$\hat{\bar{X}} = \frac{\hat{X}}{N\hat{M}} = \frac{6480}{30 \cdot 8} = 27$$

$$\hat{V}\left(\hat{\bar{X}}\right) = 7.975 \Rightarrow \widehat{EM}\left(\hat{\bar{X}}\right) = 2.824$$

❖ Estimamos que el número medio de pacientes vistos por un médico, al día, es 27, con un error de muestreo de 2.824.

d) Estima el número medio de pacientes derivados a un especialista por centro.

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{1440}{30} = 48$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{\hat{V}\left(\hat{Y}\right)}{N^2} = \frac{221520}{30^2} = 246.13 \Rightarrow \widehat{EM}\left(\hat{\bar{Y}}\right) = 15.69$$

❖ Estimamos que, al día, el número medio de pacientes derivados a un especialista es de 48 pacientes por centro, con un error de muestreo de 15.69.

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n}\hat{X}_i = \frac{N}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}X_{ij}\right), \quad \widehat{\overline{X}} = \frac{\widehat{X}}{N}, \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{M}.$$

$$\widehat{A} = \frac{N}{n}\sum_{i=1}^{n}\hat{A}_i = \frac{N}{n}\sum_{i=1}^{n}M_i p_i, \quad \widehat{\overline{A}} = \frac{\widehat{A}}{N}, \quad \widehat{P} = \frac{\widehat{A}}{M}.$$

❖ **Muestras autoponderadas ≡** el factor de elevación es el mismo para todos los elementos de la muestra.

❖ Si la fracción de muestreo en segunda etapa es constante entonces tenemos muestras autoponderadas.

$$\text{Si } f_{2i} = f_2 = \frac{m}{M}, \text{ entonces } \widehat{X} = \frac{N}{n}\sum_{i=1}^{n}\hat{X}_i = \frac{N}{n}\frac{M}{m}\sum_{i=1}^{n}\sum_{j=1}^{m_i}X_{ij}$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\widehat{\overline{X}} = \frac{N}{n}\sum_{i=1}^{n}\hat{\overline{X}}_i = \frac{N}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}X_{ij}\right), \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{N}, \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{M}.$$

Varianzas

$$V\left(\widehat{\overline{X}}\right) = \frac{N^2(1-f_1)}{n(N-1)}\sum_{i=1}^{N}(\overline{X}_i - \overline{\overline{X}})^2 + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{21})}{m_i}S_i^2 = \frac{N^2}{n(N-1)}\sum_{i=1}^{N}(X_i - \overline{\overline{X}})^2$$

Sin reposición en las dos etapas

Con reposición en la primera etapa

$$V\left(\widehat{\overline{\overline{X}}}\right) = \frac{V\left(\widehat{X}\right)}{N^2}; \quad V\left(\widehat{\overline{\overline{X}}}\right) = \frac{V\left(\widehat{X}\right)}{M^2}.$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\hat{A} = \frac{N}{n}\sum_{i=1}^{n}\hat{A}_i = \frac{N}{n}\sum_{i=1}^{n}M_i p_i, \quad \hat{\bar{A}} = \frac{\hat{A}}{N}, \quad \hat{P} = \frac{\hat{A}}{M}.$$

Varianzas

$$V\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n(N-1)}\sum_{i=1}^{N}(A_i-\bar{A})^2 + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{21})}{m_i}\left(\frac{M_i P_i Q_i}{M_i-1}\right) = \frac{N^2}{n(N-1)}\sum_{i=1}^{N}(A_i-\bar{A})^2$$

Sin reposición en las dos etapas

Con reposición en la primera etapa

$$V(\hat{\bar{A}}) = \frac{V\left(\hat{A}\right)}{N^2}; \quad V\left(\hat{P}\right) = \frac{V\left(\hat{A}\right)}{M^2}.$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\widehat{X} = \frac{N}{n}\sum_{i=1}^{n}\hat{X}_i = \frac{N}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}X_{ij}\right), \quad \widehat{\overline{X}} = \frac{\widehat{X}}{N}, \quad \widehat{\overline{\overline{X}}} = \frac{\widehat{X}}{M}.$$

Estimación de varianzas

$$\hat{V}\left(\widehat{X}\right) = \frac{N^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{X}_i - \widehat{\overline{X}})^2 + \frac{N}{n}\sum_{i=1}^{n}M_i^2(1-f_{2i})\frac{\hat{S}_i^2}{m_i}$$

$$\hat{V}\left(\widehat{\overline{X}}\right) = \frac{\hat{V}\left(\widehat{X}\right)}{N^2}; \quad V\left(\widehat{\overline{\overline{X}}}\right) = \frac{\hat{V}\left(\widehat{X}\right)}{M^2}.$$

❖ **Si las muestras son autoponderadas:**

$$\hat{V}\left(\widehat{\overline{\overline{X}}}\right) = \frac{(1-f_1)}{n}s_c^2 + \frac{f_1(1-f_2)}{n\overline{M}f_2}s_d^2,$$

$$s_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\overline{M}}\overline{x}_i - \widehat{\overline{\overline{X}}}_i\right)^2, \quad s_d^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\overline{M}}\hat{S}_i^2.$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores insesgados:

$$\hat{A} = \frac{N}{n}\sum_{i=1}^{n}\hat{A}_i = \frac{N}{n}\sum_{i=1}^{n}M_i p_i, \quad \hat{\bar{A}} = \frac{\hat{A}}{N}, \quad \hat{P} = \frac{\hat{A}}{M}.$$

Estimación de varianzas

$$\hat{V}\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{A}_i - \hat{\bar{A}})^2 + \frac{N}{n}\sum_{i=1}^{n}M_i^2(1-f_{2i})\frac{p_i q_i}{m_i - 1}$$

$$\hat{V}(\hat{\bar{A}}) = \frac{\hat{V}\left(\hat{A}\right)}{N^2}; \quad \hat{V}\left(\hat{P}\right) = \frac{\hat{V}\left(\hat{A}\right)}{M^2}.$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Estimadores de razón al tamaño:

$$\hat{\bar{X}}_R = \frac{\hat{X}}{\hat{M}} = \frac{\sum_{i=1}^{n} \hat{X}_i}{\sum_{i=1}^{n} M_i}$$

$$\hat{P}_R = \frac{\hat{A}}{\hat{M}} = \frac{\sum_{i=1}^{n} \hat{A}_i}{\sum_{i=1}^{n} M_i}$$

Estos estimadores son sesgados

Estimación de varianzas

$$\hat{V}\left(\hat{\bar{X}}_R\right) = \frac{1}{\hat{\bar{M}}^2}\left(\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{X}_i - M_i\hat{\bar{X}}_R)^2 + \frac{1}{Nn}\sum_{i=1}^{n} M_i^2(1-f_{2i})\frac{\hat{S}_i^2}{m_i}\right)$$

$$\hat{V}\left(\hat{P}_R\right) = \frac{1}{\hat{\bar{M}}^2}\left(\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{A}_i - M_i\hat{P}_R)^2 + \frac{1}{Nn}\sum_{i=1}^{n} M_i^2(1-f_{2i})\frac{p_i q_i}{m_i - 1}\right)$$

IV.2.2.- Conglomerados de tamaño desigual. Selección de conglomerados con igual probabilidad

Ejemplo 4.6. Se extrae una muestra aleatoria simple de 3 hospitales de una población con 10 hospitales. Dentro de cada uno se extrae una submuestra del 10% de las admisiones, obteniéndose los siguientes datos:

Hospital	Total de ingresados	Total de ingresados muestreados	Total de ingresados en situación crítica
1	4290	429	47
2	640	64	17
3	2150	215	24

a) Estima el total de pacientes ingresados y su error de muestreo.
b) Estima el total de pacientes ingresados en situación crítica y su error de muestreo.
c) Estima número medio de pacientes ingresados por hospital y su error de muestreo.
d) Estima el número medio de pacientes ingresados en situación crítica, por hospital, y su error de muestreo.
e) Estima la proporción de pacientes ingresados en situación crítica.

Ejemplo 4.6. :

N	n	f1	f2
10	3	0.3	0.1

Hospital	Mi	mi	ai
1	4290	429	47
2	640	64	17
3	2150	215	24

a) Estima el total de pacientes ingresados y su error de muestreo.

$$\hat{M} = \frac{N}{n}\sum_{i=1}^{n} M_i = \frac{10}{3}7080 = 23600$$

$$\hat{V}(\hat{M}) = \frac{N^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}\left(M_i - \hat{\bar{M}}\right) = \frac{10^2(1-0.3)}{3}3363700 = 78486333,$$

$$\widehat{EM}(\hat{M}) = \sqrt{78486333} = 8859$$

❖ Estimamos que han ingresado un total de 23600 pacientes con un error de muestreo de 8859 pacientes.

Ejemplo 4.6. :

b) Estima el total de pacientes ingresados en situación crítica y su error de muestreo.

$$\hat{A} = \frac{N}{n}\sum_{i=1}^{n} A_i = \frac{10}{3}880 = 2933.$$

$$\hat{V}\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{A}_i - \hat{\bar{A}})^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2(1-f_{21})\frac{p_i q_i}{m_i - 1} = 574777.8 + 22815.5 = 597593$$

$$\widehat{EM}\left(\hat{A}\right) = \sqrt{\hat{V}\left(\hat{A}\right)} = 773.04.$$

❖ Estimamos que el total de pacientes ingresados en situación crítica es 2933 pacientes y que su error de muestreo es de 773 pacientes, o equivalentemente podemos decir que estimamos el total de pacientes ingresados en situación crítica en 2933 con un error de, más menos, 1546 pacientes, con una confianza del 95%.

c) Estima el número medio de pacientes ingresados por hospital y su error de muestreo.

$$\hat{\bar{M}} = \frac{\hat{M}}{N} = \frac{23600}{10} = 2360.$$

$$\widehat{EM}(\hat{\bar{M}}) = \frac{\widehat{EM}(\hat{M})}{N} = 885.9$$

❖ Estimamos que el número medio de pacientes ingresados por hospital es de 2360 pacientes con un error de muestreo de 885.9 pacientes.

Ejemplo 4.6. :

d) Estima el número medio de pacientes ingresados en situación crítica, por hospital, y su error de muestreo.

$$\hat{\bar{A}} = \frac{\hat{A}}{N} = \frac{2933}{10} = 293.3.$$

$$\widehat{EM}\left(\hat{\bar{A}}\right) = \frac{\widehat{EM}\left(\hat{A}\right)}{N} = \frac{773.04}{10} = 77.304$$

❖ Estimamos que el número medio de pacientes ingresados en situación crítica por hospital 293.3 con un error de muestreo de 77.304 pacientes.

e) Estima la proporción de pacientes ingresados en situación crítica.

$$\hat{P}_R = \frac{\hat{A}}{\hat{M}} = \frac{2933}{23600} = 0.1243$$

$$\hat{V}\left(\hat{P}_R\right) = \frac{1}{\hat{\bar{M}}^2}\left(\frac{(1-f_1)}{n(n-1)}\sum_{i=1}^{n}(\hat{A}_i - M_i\hat{P}_R)^2 + \frac{1}{Nn}\sum_{i=1}^{n}M_i^2(1-f_{2i})\frac{p_i q_i}{m_i - 1}\right) = 0.0003116;$$

$$\widehat{EM}\left(\hat{P}_R\right) = \sqrt{0.0003116} = 0.01765$$

❖ Estimamos que el 12.43% de los pacientes ingresados lo hacen en situación crítica

CAPÍTULO V. Errores de muestreo y métodos de estimación

CONTENIDOS

V.1.- Conglomerados últimos.

V.2.- Semimuestras reiteradas.

V.3.- Jackknife.

V.4.- Bootstrap.

Recordemos que:

✓ **El error de muestreo** es la raíz cuadrada de la varianza del estimador insesgado (o con sesgo despreciable) y depende de la forma del estimador y del procedimiento para seleccionar la muestra.

✓ **La varianza del estimador** depende de parámetros poblacionales desconocidos y, por tanto, hemos de estimarla.

✓ **Un estimador de la varianza** depende del tipo de muestreo llevado a cabo y, en ocasiones, resulta difícil aplicar la fórmula de esta estimación o incluso es complicado obtener dicha fórmula. Esto ocurre en diseños complejos como, por ejemplo, los muestreos multietápicos con estratificación en las unidades de primera etapa.

Los métodos indirectos son una alternativa para estimar la varianza. Tienen fórmulas más sencillas y son aproximadamente insesgados para muestras grandes.

Entre otros métodos indirectos están: método de los **conglomerados últimos, semimuestras reiteradas, Jackniffe** y **Bootstrap.**

V.1- Conglomerados últimos

En el muestreo polietápico el término **conglomerado último** representa el conjunto de unidades de última etapa seleccionadas en una unidad primaria (por ejemplo el conjunto de viviendas seleccionadas en un municipio seleccionado en una primera etapa).

Con cada conglomerado último obtenemos $\hat{\theta}_i$, una estimación insesgada para el parámetro poblacional θ objeto de estudio, de manera que el estimador insesgado $\hat{\theta}$ construido con la muestra completa verifica que $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$. Así **el estimador de la varianza** por este método es $\hat{V}\left(\hat{\theta}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta})^2$ siendo *n* el número de conglomerados últimos seleccionados.

V.2- Semimuestras reiteradas

Partimos de $\hat{\theta}$ estimador insesgado del parámetro poblacional θ basado en la muestra completa de tamaño n. La idea es seleccionar de dicha muestra completa una submuestra de tamaño n/2 (supuesto n es par) que llamamos **semimuestra** y repetirlo K veces de forma independiente. De esta forma obtenemos K semimuestras y construimos K estimadores que verifiquen las condiciones siguientes:

- El estimador $\hat{\theta}_r$ obtenido con la r-ésima semimuestra debe ser insesgado si la semimuestra fuera considerada como una muestra, $E(\hat{\theta}_r) = \theta$, y por otro lado si consideramos la muestra como población y la semimuestra como muestra el estimador también debe ser insesgado, $E_2(\hat{\theta}_r) = \hat{\theta}$ donde la segunda esperanza es considerando la muestra como población.
- Por otra parte, se supone que $V(\hat{\theta}_r) = 2V(\hat{\theta})$ lo cual es en general será cierto debido a la construcción de las semimuestra.

V.2- Semimuestras reiteradas

Entonces **el estimador de la varianza** viene dado por la expresión:

$$\hat{V}(\hat{\theta}) = \frac{1}{K}\sum_{r=1}^{K}(\hat{\theta}_r - \hat{\theta})^2$$

Este es el método usado en la **encuesta de población activa** (EPA):

Se usan 40 reiteraciones. Primero se agrupan todas las secciones de cada estrato por pares, después se asigna aleatoriamente la primera sección de cada par a 20 reiteraciones y la otra a las otras 20. De esta forma cada reiteración queda formada por el 50% de la muestra (semimuestra) y cada sección aparece en la mitad de las reiteraciones.

V.3- Jackniffe

La idea es la siguiente: tenemos la muestra con n elementos y para el parámetro θ tenemos el estimador $\hat{\theta}$ cuya varianza queremos estimar. Llamamos $\hat{\theta}_{(j)}$ al estimador basado en **la muestra jackknife** de tamaño n-1 que resulta de eliminar la unidad j en la muestra completa y que se calcula de la misma manera que $\hat{\theta}$. Definimos para cada $j=1 \ldots, n$ el pseudovalor $\tilde{\theta}_j = n\hat{\theta} - (n-1)\hat{\theta}_{(j)}$. Entonces **el estimador Jackknife de la varianza** es

$$\hat{V}_{JK}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{j=1}^{n} (\tilde{\theta}_j - \hat{\theta}_{JK})^2 = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}_{(j)} - \hat{\theta}_{(.)})^2$$

donde $\hat{\theta}_{JK} = \frac{1}{n}\sum_{j=1}^{n} \tilde{\theta}_j$ y $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{j=1}^{n} \hat{\theta}_{(j)}$.

V.3- Jackniffe

En el caso de **muestreo multietápico** se eliminan en cada ocasión todas las unidades que componen las unidades de primera etapa (conglomerado).

Este método se ha usado en en INE en varias encuestas dirigidas a los hogares como la **Encuesta Nacional de Salud 2006 o la Encuesta sobre Participación Adulta en las Actividades de Aprendizaje 2007-**

V.4- Bootstrap

La idea es extraer **una muestra bootstrap** de la muestra original con reemplazamiento, probabilidades iguales e igual tamaño que la original, y obtener el estimador para cada muestra bootstrap de la misma forma que el estimador sobre la muestra original. Repetimos el proceso B veces de forma independiente y obtenemos B estimadores independientes cuya distribución imita a la del estimador de la muestra original. **El estimador bootstrap de la varianza** es:

$$\hat{V}_{BOOT}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}_\cdot^*)^2 \qquad \text{donde} \qquad \hat{\theta}_\cdot^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

En el INE se ha utilizado para el cálculo de los errores de muestreo de los indicadores de exclusión social obtenidos de la Encuesta de Condiciones de Vida.

CAPÍTULO VI. Determinación de tamaños muestrales

CONTENIDOS

VI.1.- Determinación de tamaño muestral en muestreo aleatorio simple.

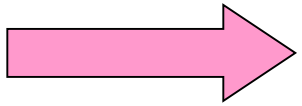
VI.2.- Determinación de tamaño muestral en muestreo aleatorio estratificado.

VI.3.- Determinación de tamaño muestral en muestreo de conglomerados.

VI.- Determinación de tamaños muestrales

Establecida la característica o características a estimar y el grado de confianza y de precisión requeridos, hay que decidir cuál va a ser el tamaño de la muestra que va a utilizarse, de modo que los resultados no sean excesivamente costosos o imprecisos.

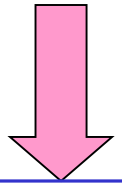
Pasos para determinar el tamaño muestral en todo tipo de muestreo



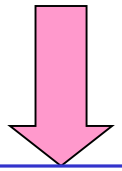
VI.- Determinación de tamaños muestrales

Pasos para determinar el tamaño muestral :

1º Fijar la precisión y la confianza deseadas



2º Determinar la ecuación que relacione el tamaño n con la precisión y confianza fijadas



3º Estimar todas las cantidades poblacionales desconocidas y despejar n

VI.1.- Tamaño muestral en MAS

1º Fijar la precisión y la confianza deseadas

Parámetro a estimar: $\overline{X}$
Error máximo admisible: e
Nivel de confianza: 1-α

$$\Pr\left\{\left|\hat{\overline{X}} - \overline{X}\right| \leq e\right\} = 1-\alpha$$

Parámetro a estimar: X
Error máximo admisible: e
Nivel de confianza: 1-α

$$\Pr\left\{\left|\hat{X} - X\right| \leq e\right\} = 1-\alpha$$

Parámetro a estimar: X ó $\overline{X}$
Error relativo: r
Nivel de confianza: 1-α

$$\Pr\left\{\left|\frac{\hat{X} - X}{X}\right| \leq r\right\} = 1-\alpha$$

VI.1.- Tamaño muestral en MAS

2º Determinar la ecuación que relacione el tamaño n con la precisión y confianza fijadas

Parámetro a estimar: $\overline{X}$
Error máximo admisible: e
Nivel de confianza: 0.95

$$e = 2\sqrt{S^2\left(\frac{1}{n} - \frac{1}{N}\right)}$$

Parámetro a estimar: X
Error máximo admisible: e
Nivel de confianza: 0.95

$$e = 2N\sqrt{S^2\left(\frac{1}{n} - \frac{1}{N}\right)}$$

Parámetro a estimar: X ó $\overline{X}$
Error relativo: r
Nivel de confianza: 0.95

$$r = 2\frac{S}{\overline{X}}\sqrt{\frac{1}{n} - \frac{1}{N}}$$

VI.1.- Tamaño muestral en MAS

2º Determinar la ecuación que relacione el tamaño n con la precisión y confianza fijadas

Parámetro a estimar: $\overline{X}$
Error máximo admisible: e
Nivel de confianza: 0.95

Tamaño aproximado

$$n_0 = \frac{4S^2}{e^2}$$

Tamaño exacto

$$n = \frac{n_0}{1 + {}^{n_0}/_{N}}$$

Parámetro a estimar: X
Error máximo admisible: e
Nivel de confianza: 0.95

Tamaño aproximado

$$n_0 = \frac{4N^2S^2}{e^2}$$

S^2 se estima a través de muestra piloto, de información anterior o conjeturas

VI.1.- Tamaño muestral en MAS

2º Determinar la ecuación que relacione el tamaño n con la precisión y confianza fijadas

Parámetro a estimar: X ó $\overline{X}$
Error relativo: r
Nivel de confianza: 1-α

Tamaño aproximado

$$n_0 = \frac{4\left(S/\overline{X}\right)^2}{r^2}$$

Tamaño exacto

$$n = \frac{n_0}{1 + n_0/N}$$

$\left(S/\overline{X}\right)^2$ estimación a través de muestra piloto, de información anterior o conjeturas sobre la población

VI.1.- Tamaño muestral en MAS

Ejemplo 6.1.- Se va a extraer una m.a.s de farmacias, de las 2500 existente en una ciudad, con el fin de estimar el precio medio de un determinado medicamento. Se quiere que la estimación que se haga no difiera en más de un 10% del precio real, con una confianza del 95%. En una encuesta telefónica a 20 farmacias de otra ciudad se obtuvo un precio medio de 7 euros, con una cuasidesviación típica de 1,4 euros.

Solución: Tomamos de la muestra telefónica los datos para estimar

$$\left(\frac{S}{\overline{X}}\right)^2 = \left(\frac{1,4}{7}\right)^2 = 0,04$$

Por tanto: $$n_0 = \frac{4.0,04}{0,1^2} = 16$$

Como n_0/N=0,0064 <0,05, tomamos $n=n_0$

Extraeremos una m.a.s. de 16 farmacias.

VI.1.- Tamaño muestral en MAS

Característica cualitativa C:

Parámetro a estimar ***P***
Error máximo admisible **e**
Nivel de confianza **0.95**

$$P\left\{\left|\hat{P}-P\right| \leq e\right\}=1-\alpha$$

Tamaño aproximado

$$n_0 = 4\frac{PQ}{e^2}$$

Tamaño exacto

$$n = \frac{n_0}{1+{}^{(n_0-1)}\!/_{N}}$$

PQ = 0,25 (caso más desfavorable)

o PQ se estima a partir de una muestra piloto o de experiencias previas

VI.1.- Tamaño muestral en MAS

Característica cualitativa C:

Parámetro a estimar ***A***
Error máximo admisible **e**
Nivel de confianza **0.95**

$$P\left\{\left|\hat{A}-A\right|\leq e\right\}=1-\alpha$$

Tamaño aproximado

$$n_0=\frac{4N^2PQ}{e^2}$$

Tamaño exacto

$$n=\frac{n_0}{1+{}^{(n_0-1)}\!/_{N}}$$

PQ = 0,25 (caso más desfavorable)

o PQ se estima a partir de una muestra piloto o de experiencias previas

Característica cualitativa C:

Parámetro a estimar ***P*** o ***A***
Error relativo **r**
Nivel de confianza **0.95**

$$P\left\{\left|\frac{\hat{A}-A}{A}\right| \le r\right\} = P\left\{\left|\frac{\hat{P}-P}{P}\right| \le r\right\} = 1-\alpha$$

Tamaño aproximado

$$n_0 = 4\frac{Q/P}{r^2}$$

Tamaño exacto

$$n = \frac{n_0}{1+{}^{(n_0-1)}\!/_{N}}$$

$$\frac{Q}{P} \simeq \frac{(1-\hat{P})}{\hat{P}}$$ de una muestra piloto o experiencia previa

Ejemplo 6.2. Un antropólogo desea estimar el porcentaje de habitantes de una isla que pertenecen al grupo sanguíneo O, con un margen de error de ±5%, aceptando el riesgo de una posibilidad en 20 de obtener una muestra poco afortunada. ¿Qué tamaño de muestra debería tomar?

Solución: $P = 0.5$ (situación más desfavorable)
e = 0.05 $1-\alpha = 0.95$

$$n = 4\frac{0.25}{0.05^2} = 400$$

Suponiendo que sólo hay 3200 isleños, entonces

$$n = \frac{400}{1+399/3200} = 356$$

Extraeremos una m.a.s. de 356 isleños

VI.2.- Tamaño muestral en MAE

Vamos a determinar el tamaño de muestra para estimar la media, total, proporción o total de clase poblacionales, fijado el nivel de confianza "**1-α**" y, dependiendo de que se fije el error máximo admisible "***e***" o el error relativo "***r***".

En el caso de fijar el error máximo admisible "***e***" para estimar el total o el total de clase, utilizaremos las expresiones dadas a continuación (correspondientes a la determinación de "***n***" para estimar la media o proporción, con error máximo admisible "***e***") sustituyendo "***e***" por "***e / N***".

Afijación proporcional:

Tamaño aproximado

Tamaño exacto

Error máximo admisible ***e*** → $n_0 = \frac{4}{e^2}\sum_{h=1}^{L} W_h S_h^2$ → $n = \frac{n_0}{1 + n_0/N}$

Error relativo ***r*** → $n_0 = \frac{4}{r^2} \cdot \frac{\sum_{h=1}^{L} W_h S_h^2}{\bar{X}^2}$ → $n = \frac{n_0}{1 + n_0/N}$

$S_h^2 = \frac{N_h . P_h . Q_h}{N_h - 1}$ en el caso de características cualitativas o atributos

$S_h^2, \bar{X}$ se estiman de una muestra piloto o experiencias previas

Afijación óptima de Neyman:

Tamaño aproximado Tamaño exacto

Error máximo admisible ***e*** → $n_0 = \frac{4}{e^2}\left(\sum_{h=1}^{L} W_h S_h\right)^2$ → $n = \frac{n_0}{1 + \frac{4}{e^2 N}\sum_{h=1}^{L} W_h S_h^2}$

Error relativo ***r*** → $n_0 = \frac{4}{r^2} \cdot \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{\bar{X}^2}$ → $n = \frac{n_0}{1 + \frac{4}{r^2 N \bar{X}^2}\sum_{h=1}^{L} W_h S_h^2}$

$S_h^2 = \frac{N_h . P_h . Q_h}{N_h - 1}$ en el caso de características cualitativas o atributos

$S_h^2, \bar{X}$ se estiman de una muestra piloto o experiencias previas

VI.2.- Tamaño muestral en MAE

Ejemplo 6.3. Una escuela desea estimar la calificación media que se puede obtener en un examen de comprensión de lectura. Los estudiante se agrupan en tres estratos, los que aprenden rápido en el estrato I y los que aprenden lento en el III. Hay 56 estudiante en el estrato I, 80 en el II y 64 en el III. Se extrae una m.a.e. de 50 estudiantes asignada proporcionalmente. El examen se aplica a los estudiantes muestreados, obteniéndose los resultados siguientes:

Est. I: 80 68 72 85 90 62 61 92 85 87 91 81 79 83
Est.II: 85 48 53 65 49 72 53 68 71 59 82 75 73 78 69 81 59 52 61 42
Est. III: 42 36 65 43 53 61 42 39 32 31 29 19 14 31 30 32

a) Estima la calificación media y da un límite para el error.

b) Si al final del curso se va a volver a realizar el estudio, ¿cuántos estudiantes debemos muestrear para que la calificación media estimada no difiera de la real en ± 4 puntos?

Ejemplo 6.3.

a) Estima la calificación media y da un límite para el error.

$$\bar{x}_1 = 79.7143 \quad \bar{x}_2 = 64.75 \quad \bar{x}_3 = 37.4375$$

$$\hat{\bar{X}} = \sum_{h=1}^{L} W_h \bar{x}_h = 0.28 \cdot 79.71 + 0.4 \cdot 64.75 + 0.32 \cdot 37.4375 = 60.2$$

$$\widehat{EM}(\hat{\bar{X}}) = \sqrt{V(\hat{\bar{X}})} = \sqrt{\frac{(1-f)}{n} \sum_{h=1}^{L} W_h \hat{S}_h^2} = \sqrt{\frac{(1-0.25)}{50} 152.28} = 1.51$$

Estimamos que la calificación media es de 60.2 con un error de ± 3 puntos

b) Si al final del curso se va a volver a realizar el estudio, ¿cuántos estudiantes debemos muestrear para que la calificación media estimada no difiera de la real en ±4 puntos?

Si usamos afijación proporcional:

$$n_0 = \frac{4}{e^2} \sum_{h=1}^{L} W_h S_h^2 = 38, \quad n = \frac{n_0}{1 + n_0 / N} = 32.$$

Deberíamos muestrear 32 estudiantes, de los cuales 9 deberían ser del estrato I, 13 del estrato II y 10 del estrato III.

Ejemplo 6.3.

b) Si al final del curso se va a volver a realizar el estudio, ¿cuántos estudiantes debemos muestrear para que la calificación media estimada no difiera de la real en ±4 puntos?

Si usamos afijación óptima de Neyman:

$$n_0 = \frac{4}{e^2}\left(\sum_{h=1}^{L} W_h S_h\right)^2 = 37.625, \quad n = \frac{n_0}{1 + \frac{4}{e^2 N}\sum_{h=1}^{L} W_h S_h^2} = 32.$$

Deberíamos muestrear 32 estudiantes, de los cuales 8 deberían ser del estrato I, 13 del estrato II y 11 del estrato III.

VI.3.- Tamaño muestral en muestreo de conglomerados monoetápico

Conglomerados de igual tamaño:

Fijado un error máximo admisible e y un nivel de confianza 0.95

Parámetro a estimar	Tamaño aproximado	Tamaño exacto
Media o proporción poblacional:	$n_0 = \frac{4S_c^2}{e^2}$	$n = \frac{n_0}{1 + n_0/N}$
Media por conglomerado:	$n_0 = \frac{4N^2S_c^2}{e^2}$	
Total poblacional:	$n_0 = \frac{4M^2S_c^2}{e^2}$	

S_c^2 se estima a partir de una muestra piloto o de experiencia previa

VI.3.- Tamaño muestral en muestreo de conglomerados monoetápico

Ejemplo 6.4 Cierto tipo de placas de circuitos tiene 12 microcircuitos por placa. Durante la inspección de control de calidad de 10 de esas placas, el nº de microcircuitos defectuosos por placa fue:

2, 0, 1, 3, 2, 0, 0, 1, 3, 4

a. Estima el porcentaje de microcircuitos defectuosos y dar un límite para su error.

Solución: Consideramos ***M*** suficientemente grande.

$$\hat{P} = \frac{(2+0+..+4)}{12.10} = 0{,}133$$

Estimamos que el 13,3% de los microcircuitos son defectuosos.

Estimemos su error de muestreo, para ello previamente calculamos la variabilidad entre conglomerados (placas)

VI.3.- Tamaño muestral en muestreo de conglomerados monoetápico

$$E\hat{M}(\hat{P}) = \sqrt{\frac{\hat{S}_c^2}{n}} = \sqrt{\frac{0.01419753}{10}} = 0.0377; \quad 2.E\hat{M}(\hat{P}) = 2.0,0377 = 0,075$$

El error máximo admisible es 7,5%. Por lo tanto, estimamos que hay un 13,3% de microcircuitos defectuosos con un error de $\pm$7,5%

b) ¿Cuántas placas más habría que inspeccionar para que, con una confianza del 95%, el porcentaje estimado de microcircuitos defectuosos no difiera del real en $\pm$ 5%?

$$n_0 = \frac{4.\hat{S}_c^2}{e^2} = \frac{2^2 \cdot 0.01419753}{0,05^2} = 22.7$$

Tendríamos que inspeccionar 13 placas más para totalizar una muestra de 23 placas, que es la que necesitamos para conseguir la precisión fijada con la confianza del 95%.

CAPÍTULO VII. Diseños muestrales en encuestas de hogares y económicas

CONTENIDOS

VII.1.- Encuestas de hogares:

Encuesta de población activa (EPA)

http://www.ine.es/inebaseDYN/epa30308/docs/epa05_disenc.pdf

Encuesta de presupuestos familiares (EPF)

http://www.ine.es/metodologia/t25/t2530p458.pdf

Encuesta de condiciones de vida (ECV)

http://www.ine.es/daco/daco42/condivi/ecv_metodo.pdf

CAPÍTULO VII. Diseños muestrales en encuestas de hogares y económicas

CONTENIDOS

VII.2.- Encuestas económicas:

Encuesta industrial de empresas

http://www.ine.es/daco/daco42/encindem/metoeiae.Pdf

Encuesta sobre el uso de Tecnologías de la Información y de las Comunicaciones y del Comercio Electrónico en las Empresas (ETICCE)

http://www.ine.es/daco/daco42/comele/meto cor.pdf

Referencias

•William.G. Cochran. Técnicas de Muestreo. Compañía Editorial Continental, 1996.

•*Francisco Ramón Fernández García, José Antonio Mayor Gallego. Muestreo en poblaciones finitas: Curso Básico. EUB, 1995.*

•*Francisco Ramón Fernández García, José Antonio Mayor Gallego Ejercicios y prácticas de muestreo en poblaciones finitas. EUB 1995.*

•*Julio Miras. Elementos de muestreo para poblaciones finitas. INE, 1995.*

•*José Luis Sánchez-Crespo Rodríguez. Curso intensivo de muestreo en poblaciones finitas. INE. 1984.*

•*José Luis Sánchez-Crespo, Javier de Parada. Ejercicios y problemas resueltos de muestreo en poblaciones finitas. INE. 1990.*

•*Richard L. Cheaffer, William Mendenhall, Lyman Ott. Elementos de muestreo. Thomson, 2007.*

• *"*Curso *sobre Diseño Muestral de las encuestas de población y económicas"* Escuela de Estadística de las Administraciones Públicas. INE