

M U E S T R E O

$$\hat{t}_{Hajek} = N \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$$

Dr. Emilio López Escobar

emilio@numerika.mx

Curso para el:



**Instituto Nacional para la
Evaluación de la Educación**

Ciudad de México
Octubre de **2018**

Índice general

I	Información sobre el curso	VIII
	Contenido general del curso	IX
	Objetivo del curso	X
	Descripción del curso	X
	Conocimientos previos que son necesarios	X
	Referencias bibliográficas del curso	XI
	Software estadístico	XIII
	Calendarización del curso	XIV
	Recursos y material del curso	XV
	Evaluación del curso	XVI
II	Fundamentos de muestreo	1
1.	Introducción	2
	1.1. Creciente importancia del muestreo	3
	1.2. El gran supuesto de la teoría estadística estándar	4
	1.3. Comentarios sobre la enseñanza del muestreo en México y bibliografía del curso	5
2.	El objetivo del muestreo y el marco muestral	8
	2.1. El objetivo del muestreo	9
	2.2. Sobre inferir o generalizar...	10
	2.2.1. Siempre inferimos, siempre generalizamos...	10

2.2.2. ¿Inferir es aprender?...	10
2.3. Inferir o generalizar sobre U a partir de s	11
2.3.1. Un ejemplo equivocado...	11
2.4. Sobre los 3 grandes enfoques teóricos del muestreo	13
2.4.1. 'Design-based approach'	13
2.4.2. 'Model-based approach'	14
2.4.3. 'Model-assisted approach'	15
2.5. Marco muestral	16
2.6. Radiografía general de una encuesta por muestreo	18
2.7. Algunos comentarios	20
2.7.1. Incorporación de técnica a el objetivo del muestreo	21
2.7.2. Un ejemplo sobre el marco muestral (de Lohr, 1999)	22
3. Muestreo probabilístico y extracción de la muestra	23
3.1. Muestreando probabilísticamente	24
3.1.1. Muestreo en 1 etapa	24
3.2. Muestreando en más de 1 etapa	27
3.2.1. Muestreo en 3 etapas	27
3.2.2. Ventaja de las muestras probabilísticas	30
3.2.3. Muestreo en 2 etapas	30
4. Estimación a partir de muestras probabilísticas	32
4.1. Población, muestra y selección	33
4.2. La función diseño de muestreo	35
4.3. Probabilidades e indicadores de inclusión	37
4.3.1. Las indicadores de inclusión muestral	37
4.3.2. Las probabilidades de inclusión	37
4.3.3. Comentarios sobre las probabilidades de inclusión	39
4.3.4. Estadísticos bajo el diseño muestral	41
4.4. Muestreo Bernoulli (BE)	44
4.5. Muestreo Aleatorio Simple (SI)	46
5. Estimadores y sus propiedades estadísticas básicas	49
5.1. Estimadores comunes	50
5.2. Distribución muestral de un estimador	52
5.3. Los Estimadores π y sus propiedades	62
5.4. El estimador π bajo el diseño BE	71

5.5. El estimador π bajo el diseño SI	73
5.6. El efecto de diseño	75
6. ¿Qué tamaño de muestra utilizar?	78
6.1. Tamaño de muestra para una media bajo muestreo SI asumiendo normalidad	81
6.2. Tamaño de muestra para una media bajo muestreo SI sin asumir normalidad	83
6.2.1. Utilizando el coeficiente de variación	83
6.2.2. Utilizando la desigualdad de Tchebychev	84
6.3. Tamaño de muestra para una proporción bajo muestreo SI asumiendo normalidad	85
6.4. Tamaño de muestra para una proporción bajo muestreo SI sin asumir normalidad	87
6.5. ¿Cuándo se puede considerar a N grande?	87
6.6. El efecto del diseño: ajuste del tamaño de muestra	89
6.7. Ajuste del tamaño de muestra por la tasa de respuesta	90
6.8. Comentarios finales sobre el tamaño de muestra	91
7. Estratificación	92
7.1. Introducción a la estratificación	93
7.1.1. ¿Cómo se ve la estratificación en otros textos y cómo la trataremos?	93
7.1.2. ¿De qué se trata la estratificación?	93
7.1.3. Utilidad y usos de la estratificación	93
7.1.4. ¿Estratificar o no estratificar?	95
7.1.5. La peor de las situaciones	95
7.1.6. Concepción equivocada y muy usada al estratificar	95
7.2. ¿Hay una buena estratificación?	96
7.3. El número de estratos	97
7.4. El tamaño de muestra asociado a la población a partir del tamaño de muestra asociado a los dominios de estimación	98
7.5. ¿Muestreo PPT o mejor estratificar?	99
7.6. Notación y uso de la estratificación	100
7.6.1. El diseño de muestreo aleatorio simple estratificado, STSI	102
7.6.2. Sobre la estimación de un total y una media con estratificación: un error común	104
7.7. Afijación, asignación o distribución de muestra en estratos	105
7.7.1. Una función de costos	107

7.7.2. Distribución Óptima	109
7.7.3. Distribución de Neyman	109
7.7.4. Distribución proporcional	110
7.7.5. Distribuciones alternativas	110
8. Conglomeración	112
8.1. Introducción a la conglomeración	113
8.1.1. ¿Cómo se ve la conglomeración en otros textos y cómo la trataremos?	113
8.1.2. ¿Qué problemas soluciona o qué facilita la conglomeración? Su utilidad...	114
8.1.3. ¿En qué consiste el muestreo por conglomerados?	116
8.1.4. ¿En qué consiste el muestreo en dos etapas?	117
8.1.5. ¿En qué consiste el muestreo multi-etápico?	118
8.2. Estimación de totales y medias con conglomeración	118
8.3. Muestreo de conglomerados unietápico	120
8.4. Muestreo de conglomerados unietápico aleatorio simple	125
8.5. Muestreo bietápico	126
8.5.1. Muestreo bietápico de elementos	129
8.5.2. Muestreo bietápico de elementos: diseño auto-ponderado	133
8.6. Post-Estratificación, ajuste o calibración de factores de expansión	136
9. Muestreo Poisson y Muestreo Poisson Condicional	138
9.1. Muestreo Poisson (PO)	139
9.2. El estimador de Hájek	143
9.3. Muestreo Poisson Condicional (CPO, CPS)	144
10. Estimación de Parámetros No-Lineales	146
10.1. Parámetros no-lineales	147
11. Estimación de Varianza (Introducción / Linealización)	155
11.1. El problema de la estimación de varianza	156
11.2. La complicación del problema de estimar la varianza	157
11.3. Método de Linealización de Taylor	158
11.3.1. Enfoque tradicional: Woodruff (1971); Robinson & Särndal (1983); Deville (1999)	159
11.4. Estimación de una Razón	164

11.4.1. Cota Superior del Sesgo de \hat{R}	165
11.5. Estimación de Varianza por Linealización de Taylor para una Razón	167
11.6. Estimación asistida por modelos lineales: El Estimador de Razón de un Total	170
11.7. Estimación de una media	174
11.8. Estimación de un total utilizando Hájek	176
11.8.1. Enfoque moderno: Demnati & Rao (2004)	177
11.8.2. Enfoque moderno alternativo: Graf (2011)	179
12. Utilizando R con lo hasta ahora visto	180
12.1. Estimadores Puntuales	181
12.2. Algunos Estimadores de Varianza de tales Estimadores Puntuales	184
12.2.1. Para el estimador puntual de Narain (1951); Horvitz-Thompson (1952) para un total	184
12.2.2. Para el estimador puntual de Narain (1951); Horvitz-Thompson (1952) para una media	185
12.2.3. Para el estimador puntual de una razón	185
13. Estimación de Varianza (Remuestreo)	186
13.1. Introducción	187
13.2. Sobre la estimación de varianza (revisita)	187
13.3. Propiedades de los estimadores de varianza	190
13.4. Grupos Aleatorios Independientes	192
13.5. Grupos Aleatorios No-Independientes	195
13.6. Jackknife de Quenouille (1956); Tukey (1958)	195
13.7. Bootstrap de Efron (1979)	197
13.7.1. Fortalezas y Debilidades del Jackknife y del Bootstrap	198
13.8. Jackknife Generalizado para Funciones de Medias de Campbell (1980); Berger & Skinner (2005)	199
13.9. Jackknife Generalizado para Funciones de Medias de Berger (2007)	200
13.10. Jackknife para Muestras Bi-Etápicas Autoponderadas de Escobar & Ber- ger (2013)	200
13.11. Jackknife Generalizado Post-Expansión para Funciones de Totales de de Escobar & Berger (2013)	201
III Apéndices	202
Relación entre distribuciones de probabilidad	203

Varianzas hipotéticas de algunas distribuciones (Kish, 1965)	204
Teorema Central del Límite, Velocidad de convergencia a una Normal, Aproximaciones a la varianza de una variable, Desigualdad de Tchebychev (Mendez, Eslava & Romero, 2004)	205
Ejercicios de Muestreo	210

IV Sesiones prácticas en R 215

INFORMACIÓN SOBRE EL CURSO

CONTENIDO GENERAL DEL CURSO

- ▶ Introducción al muestreo.
- ▶ El enfoque particularizado vs. generalizado en el muestreo.
- ▶ Los 3 enfoques filosóficos del muestreo contemporáneo.
- ▶ Teoría de muestreo bajo el enfoque generalizado de los estimadores π o de Narain-Horvitz-Thompson (probabilidades arbitrarias). La noción de factor de expansión. El estimador de Narain (1951); Horvitz-Thompson (1952). El estimador de Hájek (1971). Ventajas, desventajas, desempeño práctico y vicios comunes.
- ▶ Muestreo aleatorio simple. Muestreo Bernoulli
- ▶ Distribución muestral de un estimador. Propiedades de los estimadores. La varianza del estimador, errores estándares. Calidad de estimaciones. Coeficiente de variación. Calidad de un esquema de muestreo específico. El efecto de diseño.
- ▶ Cálculo de tamaño de muestra. Estimación en dominios o subpoblaciones.
- ▶ Estratificación. Métodos de asignación (afijación) de muestra.
- ▶ Conglomeración. Muestreo en dos etapas. Introducción al muestreo en más de dos etapas.
- ▶ Introducción al muestreo con probabilidades proporcionales al tamaño. Ventajas, desventajas y precauciones.
- ▶ Sobre diseños de muestreo autoponderados, post-estratificación y consecuencias de suponer muestreo aleatorio simple en la estimación cuando éste no fue utilizado en la extracción de la muestra.
- ▶ Planteamiento de problemas prácticos y comunes de muestreo complejo.
- ▶ Estimación de varianza. Linealización de Taylor (Método Delta). Enfoque tradicional de Woodruff (1971); Robinson & Särndal (1983); Deville (1999). Enfoque moderno de Demnati & Rao (2004) y otros enfoques, Graf (2011).
- ▶ El Jackknife de Quenouille (1956); Tukey (1958). El Bootstrap de Efron (1979). Otros métodos recientes. Ventajas, desventajas, desempeño práctico y vicios comunes.
- ▶ Introducción al muestreo complejo con paquetes estadísticos (SPSS o de preferencia R).

OBJETIVO DEL CURSO

Conocer la **base teórica** y las **principales aplicaciones** de métodos estadísticos en muestreo para poblaciones finitas.

Se hará especial énfasis en la teoría de muestreo contemporánea bajo una perspectiva unificada y generalizada. Se discutirán ejemplos y casos.

DESCRIPCIÓN DEL CURSO

Los temas serán presentados y motivados por el instructor. Se discutirán tópicos teóricos y ejemplos prácticos en clase. El aprendizaje se reforzará y se recargará en ejercicios prácticos, lecturas y tareas fuera del salón de clases.

CONOCIMIENTOS PREVIOS QUE SON NECESARIOS

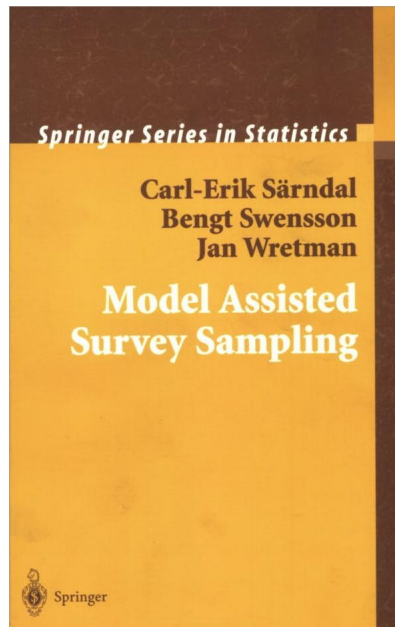
Es deseable que los alumnos cuenten con los siguientes conocimientos previos:

- ▶ Álgebra (conjuntos, doble sumas, conocimientos de **conteo**),
- ▶ Cálculo de probabilidades (distribuciones de probabilidad básicas, cálculo de probabilidades, función de densidad Bernoulli y Normal),
- ▶ Estadística descriptiva - suficiente como para obviarlo,
- ▶ Inferencia estadística (**deseable** - estimación puntual, intervalos de confianza, pruebas de hipótesis, pruebas de significancia),
- ▶ R - suficiente para ejecutar cosas simples.

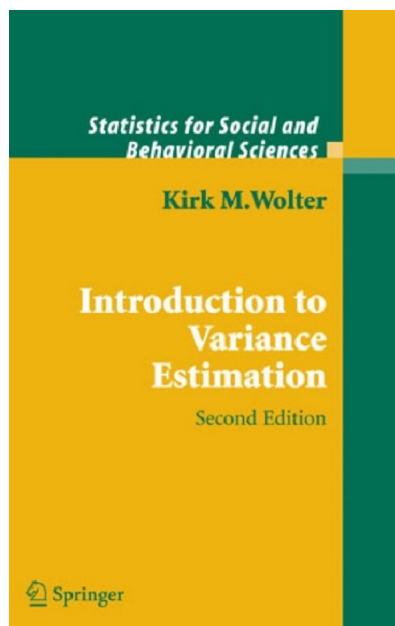
REFERENCIAS BIBLIOGRÁFICAS DEL CURSO

Las referencias base del curso son:

- ▶ Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.



- ▶ Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd Ed. Springer.



Y también nos apoyaremos en pasajes o ejemplos de:

- ▶ Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons.
- ▶ Lehtonen, R. & Pahkinen, E.J. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. 2nd Ed. John Wiley & Sons.
- ▶ Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
- ▶ Méndez, I., Eslava, G. & Romero, P. (2004). *Conceptos Básicos de Muestreo*. Monografía 27, Vol. 12, IIMAS-UNAM.
- ▶ Pfeffermann, D. & Rao, C.R. (eds.) (2009). *Handbook of Statistics 29A. Sample Surveys: Designs, Methods and Applications*. North-Holland.
- ▶ Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.
- ▶ Tillé, Y. (2006). *Sampling Algorithms*. Springer.

Referencias adicionales de muestreo «tradicionales» (**enfoque particularizado**):

- ▶ Deming(1950) (Algo ilustrativo pero desactualizado).
- ▶ Kish(1965) (Un clásico - Muy **bueno** en sus consejos y resolución de **problemas prácticos** - Notación y enfoque desactualizado.).
- ▶ Raj(1968). (desactualizado).
- ▶ Kish(1972) (Traducción al Español difícil de encontrar a la venta).
- ▶ Cochran(1977) (Un clásico - Anticipa varios **problemas teóricos** serios de muestreo - Notación y enfoque desactualizado).
- ▶ Sukhatme(1984).
- ▶ Kish(1987) (Varios detalles útiles para investigación).

Características generales de la bibliografía base que utilizaremos, Särndal et al. (1992):

- ▶ Ofrece un enfoque o perspectiva **unificada** del muestreo.
- ▶ Rico en conceptos estadísticos pero a la vez no es de alto nivel matemático.
- ▶ **Y algo importante para este curso, su planteamiento es el mismo que utiliza cualquier software contemporáneo especializado de muestreo.**

SOFTWARE ESTADÍSTICO

Utilizaremos primordialmente R.

Es gratuito. Está en la Comprehensive R Archive Network (CRAN):

<http://www.r-project.org/>

<http://cran.r-project.org/>

Este será el paquete estadístico básico.



¿Por qué R? Por que es el mejor. Para acabar pronto... terminarán utilizando R en algún momento. Empiecen desde ahora. Vamos de la mano.

Si hay tiempo, podríamos utilizar también software comercial de amplia distribución como **SPSS** de IBM o cualquier otro. Sólo si hay tiempo.

Son libres de utilizar el software que prefieran. Por supuesto, se sugiere utilizar R.

CALENDARIZACIÓN DEL CURSO

El curso está compuesto de:

- ▶ Número de sesiones efectivas: **15**
- ▶ Número de semanas del curso: **3**

- ▶ Calendario:

OCTUBRE 2018							
	D	L	M	M	J	V	S
		1	2	3	4	5	6
1	7	8	9	10	11	12	13
		1	2	3	4	5	
2	14	15	16	17	18	19	20
		6	7	8	9	10	
3	21	22	23	24	25	26	27
		11	12	13	14	15	
	28	29	30	31			

- ▶ Lugar de impartición del curso:

Las sesiones serán en instalaciones del INEE, en la Ciudad de México.

RECURSOS Y MATERIAL DEL CURSO

Todo el material que utilicemos en el curso:

- ▶ Notas,
- ▶ Datos,
- ▶ Material extra,
- ▶ Scripts de R,
- ▶ etc.

se encontrarán disponible en la siguiente liga (con contraseña):

<http://www.Info-Emilio.NET/teaching/inee-muestreo-e2y1>

La contraseña para acceder se las haré llegar por otro medio.

EVALUACIÓN DEL CURSO

Asistencia			10 %
Tarea	Equipo (3 o 4)	sencilla, con mucho tiempo, dudas en clase	50 %
Examen	Individual	breve, ultima clase, (en parte) tomado de la tarea	40 %
Notas: Adicionalmente, habrán tareas opcionales que valdrán (en total) 5 o 10 % más, dependiendo de cuántas son. Traten de hacerlas para ayudarse.			

Notar que la tarea junto con la asistencia y las opcionales pueden llegar a sumar un 60 % o más de su calificación. De modo que el curso es sencillo de aprobar involucrándose.

Recuerden utilizar tiempo de estudio adicional a su clase... (lectura, repaso, ejercicios).

No será suficiente 'venir a ver' la clase. Se trata de que se involucren.

FUNDAMENTOS DE MUESTREO

CAPÍTULO 1

Introducción

Algo de motivación al curso...

1.1 CRECIENTE IMPORTANCIA DEL MUESTREO

- ▶ **Creciente importancia de la consideración** estadística de los *efectos* de la obtención de información (datos) **a través de encuestas por muestreo**.
- ▶ Es cada vez más **difícil asumir** que los datos disponibles son los “datos verdaderos”, o que tienen **cierta estructura estocástica** que permite utilizar teoría **estadística estándar**. ¿Más conciencia del origen de los datos?
- ▶ El **muestreo ocupa un renovado interés** en el inventario de conocimientos que debe poseer, afinar y en general tomar en cuenta cualquier persona que realice trabajo estadístico.
- ▶ Viendo como están ahora las cosas en el mercado de las encuestas en México, es claro que hay mucha **oportunidad de mejora** (oportunidades de trabajo o de negocio).
- ▶ La **sustitución de censos por encuestas**... Francia, Alemania, Holanda, Reino Unido (próximamente)... El caso de México.
- ▶ Es una rama de la Estadística bastante **práctica**, muy **realista**...

1.2 EL GRAN SUPUESTO DE LA TEORÍA ESTADÍSTICA ESTÁNDAR

- ▶ El acostumbrado supuesto de independencia e igualdad de distribución en las observaciones (uniformidad), que subyace en la teoría de la mayoría de los modelos estadísticos (econométricos, etc.); no aplica con la misma frecuencia y naturalidad en la práctica cotidiana.
- ▶ Comúnmente denotado *v.a.i.i.d.* (*variables aleatorias independientes idénticamente distribuidas* -muestreo aleatorio simple, *m.a.s.*; muestreo Bernoulli).
- ▶ En la práctica estadística mundana (médicos, abogados, politólogos, etc.) se suele violar constantemente tal supuesto.
- ▶ Piensen en lo blando de los datos de origen no-exacto como por ejemplo de las ciencias sociales. ¿Tendrán la estructura estocástica ideal? ¿Valdrán los supuestos de cursos de estadística estándar?
- ▶ Por supuesto, ojo, no quiere decir que toda la teoría recargada en esos supuestos no funcione sino que se reconoce el problema matemático (abierto en su mayoría) de la necesidad de **adaptar** la teoría estándar.
- ▶ Este tipo de adaptación es **cada vez más fácil de incorporar a la práctica y cada vez más cercano a disciplinas no cuantitativas**, módulos o bibliotecas.
- ▶ Por ejemplo, el paquete para disciplinas no cuantitativas, *SPSS*® (Statistical Package for the Social Sciences). Con el módulo 'Muestras Complejas'.
- ▶ Otras opciones para implementar tales adaptaciones:
 - STATA (la serie de comandos *svy*)
 - WesVar (de Westat)
 - SUDAAN (SURvey DATA ANalysis) - SAS
 - R** (paquetes de muestreo, e.g. *sampling*, *survey*, *samplingVarEst*, etc.)

1.3 COMENTARIOS SOBRE LA ENSEÑANZA DEL MUESTREO EN MÉXICO Y BIBLIOGRAFÍA DEL CURSO

¿Por qué tan **árida** la enseñanza de muestreo?

¿Por qué **no gusta** o porqué suele ser **difícil de enseñar**?

Un breve listado de la bibliografía que comúnmente se ha utilizado para la enseñanza del muestreo en México es:

- ▶ Deming(1950) (Ilustrativo pero ya muy desactualizado).
- ▶ Kish(1965) (Un clásico - Muy **bueno** en lo que atañe a consejos y resolución de **problemas prácticos** - No obstante, algo anticuado - Mejor consultarlo después del Särndal).
- ▶ Raj(1968).
- ▶ Kish(1972) (Traducción al Español difícil de encontrar a la venta).
- ▶ Cochran(1977) (Un clásico - Anticipa varios **problemas teóricos** serios a los que se enfrentaría un muestrista - No obstante, algo desactualizado - Mejor consultarlo después del Särndal).
- ▶ Sukhatme(1984).
- ▶ Kish(1987) (Varios detalles importantes para cuestiones de investigación en muestreo).
- ▶ Méndez, Eslava & Romero(2004)(**Ayuda** mucho a tener una visión rápida y sencilla sobre conceptos básicos - Mejor consultarlo después del Särndal y sólo si tienen dudas en el 'arranque').

Todos estos textos de muestreo tienen un **enfoque particularizado**:

- ▶ Para cada uno de los esquemas de selección de muestras hay una expresión matemática de estimación **específica**.

- ▶ Si la forma de seleccionar individuos que se está utilizando **no corresponde** a alguno de estos esquemas, entonces **no hay una expresión directa** o es **muy complicado** derivar la expresión matemática que verdaderamente correspondería.
- ▶ Se tendría, entonces, que hacer supuestos (práctica común y que **puede acarrear grandes errores** - el truco está, como veremos más adelante, en poder dimensionar las consecuencias de las decisiones tomadas en el ejercicio de muestrear-).
- ▶ O bien, se tendría que ajustar el esquema de selección a los listados en este tipo de bibliografías para que 'encaje' (**en la práctica muchas veces es imposible**).
- ▶ Es decir, en general tratan al muestreo como un problema de estimación por **separado** según el esquema de selección de individuos utilizado. Y es prácticamente imposible listar todas las formas posibles de realizar una selección.
- ▶ A esto se debe quizás la *tradicional* **aridez** de la enseñanza del muestreo o el desinterés en él.

La bibliografía que utilizaremos como base en este curso, el 'yellow-book':

▶ Särndal, Swensson, & Wretman (1992). *Model Assisted Survey Sampling*, por el contrario, tiene las siguientes características:

- ▶ Ofrece un enfoque o perspectiva **unificada** del problema de inferir sobre una población a partir de una muestra.
- ▶ Presenta una forma o **enfoque general** de abordar el problema de estimación.
- ▶ Esta forma de enseñar muestreo (como se menciona en el texto) ha sido probada y utilizada con éxito. Recientemente ya en casi todas las instituciones educativas modernas en sus disciplinas cuantitativas.

- ▶ Adicionalmente, este enfoque es seguido en la forma reciente de presentación de artículos de investigación en revistas técnicas relacionadas con muestreo.
- ▶ Es rico en conceptos estadísticos pero a la vez no es de alto nivel matemático (al menos los capítulos que tocaremos).
- ▶ Como ya se mencionó, los retos más grandes del muestreo o las complicaciones generales vienen en la implementación práctica (y esto se acentúa más en nuestro país - no hay bancos de datos confiables o no son accesibles).
- ▶ **Algo importante para este curso, es que el planteamiento de este libro es el mismo que utiliza cualquier software especializado de muestreo** en sus entrañas y en la interfase del usuario.
- ▶ En otras palabras, si no entendemos el muestreo de la forma en que se plantea en esta bibliografía, **no será posible** la correcta interacción con los paquetes estadísticos especializados en muestreo.
- ▶ No sabremos lo que el software nos solicita o cómo pedirle algo específico.

CAPÍTULO 2

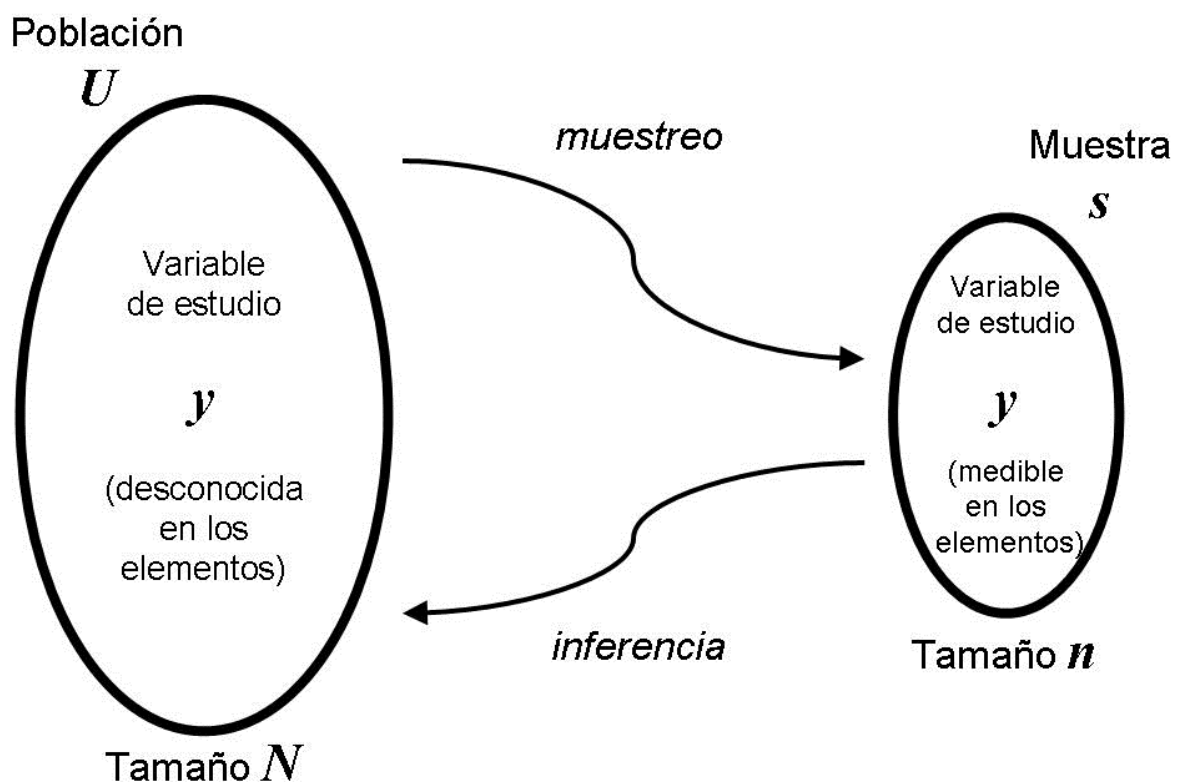
El objetivo del muestreo y el marco muestral

2.1 EL OBJETIVO DEL MUESTREO

Hay una **población finita (conjunto de elementos)** de la cual nos interesa conocer alguna(s) característica(s).

Nos aproximaremos a esta población mediante una **muestra (subconjunto de elementos)**.

Se trata entonces de **inferir sobre ciertas propiedades de una población a partir de la información parcial** de ésta.



2.2 SOBRE INFERIR O GENERALIZAR...

2.2.1 SIEMPRE INFERIMOS, SIEMPRE GENERALIZAMOS...

- ▶ Notemos que **siempre** estamos infiriendo.
- ▶ No podemos estar experimentando **exhaustivamente** todas las cosas o vivencias.
- ▶ Probamos algo y entonces **decidimos**, inferimos **sobre el resto**.
- ▶ Ejemplo: Enología.... ¿Otros ejemplos?
- ▶ Es más, como tenemos memoria, siempre estamos en este proceso de **inducción** donde generalizamos a partir de **información parcial**.

2.2.2 ¿INFERIR ES APRENDER?...

- ▶ Ejemplo: Opiniones formadas.... ¿Otros ejemplos?

2.3 INFERIR O GENERALIZAR SOBRE U A PARTIR DE s

2.3.1 UN EJEMPLO EQUIVOCADO...

Tomemos un ejemplo que aparece en Lohr (1999).

En el libro *Mujeres y amor: Una revolución cultural en progreso* por Shere Hite (1987) se encuentran los siguientes resultados (bastante citados de hecho):

- ▶ 84 % de las mujeres están “no satisfechas emocionalmente con su relación sentimental” (p. 804)
- ▶ 70 % de todas las mujeres “casadas 5 o más años tienen relaciones sexuales fuera de sus matrimonios” (p. 856)
- ▶ 95 % de las mujeres “reportan formas de abuso emocional o psicológico de parte de hombres con quienes están en una relación amorosa” (p. 810)
- ▶ 84 % de las mujeres reportan formas de desdén o indiferencia por parte de los hombres en su relación amorosa (p. 809)

Este libro ha sido citado y criticado bastante en los Estados Unidos por periódicos y revistas. ¿Por qué ha sido tan criticado? ¿Será información que ofende?

Por supuesto que no. El estudio de Hite discute verdaderos temas de interés, no obstante su error es **generalizar a todas** las mujeres por el sólo hecho de haber o no participado en su encuesta.

Estas características hacen que tal generalización no sea posible:

- ▶ La muestra fue auto-seleccionada. Esto es, las mujeres que recibían el cuestionario por correo decidían si estarían en muestra o no. Hite mandó 100,000 cuestionarios y sólo le regresaron 4.5 %
- ▶ Los cuestionarios se los hicieron llegar mediante asociaciones profesionales de

mujeres, grupos de trabajo, iglesias, etc. y dejaron fuera a todas las demás mujeres que no acudían a tales lugares

- ▶ La encuesta tiene 127 preguntas abiertas y varias preguntas tenían varias partes ¿Quién tendería a contestar tales preguntas?
- ▶ Muchas preguntas son vagas y usan palabras como “amor”. El concepto de “amor” tiene muchas interpretaciones - sin criterios de interpretación válidos o comparables.
- ▶ Muchas de las preguntas son tendenciosas. Esto es, que sugieren al entrevistado qué respuesta dar. Por ejemplo: “¿Tu esposo/amante te ve como igual? ¿O hay veces en que parece que él te trata como alguien inferior? ¿O no te deja tomar decisiones? ¿O actúa superior? (p. 795)”

Posteriormente Hite escribe: “¿Es posible que una investigación no basada en la probabilidad o en una muestra aleatoria permita generalizar sus resultados a la gran población? Si el estudio es lo suficientemente grande y la muestra lo suficientemente amplia y si una generaliza con cuidado, sí. (p. 778)”

Por supuesto que para un estadístico muestrista la respuesta es no. La muestra final no representa a las mujeres de los Estados Unidos y los estadísticos obtenidos sólo describen a las mujeres que decidieron responder.

Entonces, por ejemplo...

- ▶ ¿son válidos los sondeos por Internet?
- ▶ ¿son válidos los cuestionarios por correo electrónico a empleados de una empresa?
- ▶ ¿son válidas las generalizaciones que se hacen a partir de este tipo de sondeos?
- ▶ La respuesta es sí son válidos, lo que puede no ser válido son las generalizaciones que se hagan.
- ▶ Entonces, en lo que nos tenemos que fijar es no sólo en el ‘instrumento’ o cuestionario, sino quiénes contestan, qué se infiere o generaliza. Notar que el tamaño de muestra en tal validez no es importante, su importancia viene después.
- ▶ Regularmente en Marketing, en Opinión Pública, o en la práctica en general, es en esta inferencia en donde más se abusa.

2.4 SOBRE LOS 3 GRANDES ENFOQUES TEÓRICOS DEL MUESTREO

- ▶ El objetivo o problema de muestreo puede resolverse de varias formas.
- ▶ Hay 3 principales enfoques o perspectivas, dependiendo de **dónde se encuentra (o se asume) está la estructura estocástica del problema**.

2.4.1 'DESIGN-BASED APPROACH'

- ▶ Lo importante: ¿Cómo fue extraída la muestra?
- ▶ El muestrista puede elegir cómo, lo crucial será que considere este cómo a la hora de estimar.
- ▶ Otros nombres: 'muestreo' a secas por colegas no expertos, 'muestreo basado en diseño', 'muestreo con enfoque aleatorizado', 'muestreo directo'.

- ▶ Fortalezas:

- ▶ Objetividad. Si se hace de manera documentada, nadie puede cuestionar la objetividad de la muestra, o el que haya o no sido seleccionada de acuerdo a un diseño de muestreo.

No se confundan. Notar que la objetividad no está ligada a la arbitrariedad del diseño de muestreo que elija el muestrista (e.g. una vez de acuerdo todos en cómo se extraerá la muestra no hay subjetividad).

La aparente confusión es un argumento mal utilizado para atacar este enfoque.

- ▶ Exactitud (insesgamiento). Y conforme se aumente el tamaño de muestra se tenderá al verdadero valor.

- ▶ Debilidades:

- ▶ Tamaños de muestra grandes. Para obtener buenas estimaciones se requieren tamaños de muestra considerables o de plano muy grandes.
- ▶ Elevados costos. Por el tamaño de muestra grande necesita de más recursos económicos.

2.4.2 'MODEL-BASED APPROACH'

- ▶ Asume la existencia de una super-población U^* que 'generó' a la población U que tenemos enfrente a través de un modelo.
- ▶ Lo importante: ¿el modelo?
- ▶ El muestrista tiene que elegir el modelo que impondrá. El modelo determina qué partes son aleatorias y qué parte no lo son, también la estructura estocástica de la parte aleatoria.
- ▶ Otros nombres: 'muestreo basado en modelos', 'muestreo con enfoque de super-población' (áreas pequeñas, etc.).
- ▶ Notar que el modelo se impone subjetivamente. Se impone un modelo a algo que no se conoce.
- ▶ Fortalezas:
 - ▶ Precisión (estabilidad de las estimaciones).
 - ▶ Se pueden manejar tamaños de muestra muy pequeños o de plano tamaño de muestra cero.
 - ▶ Encuestas muy económicas.
- ▶ Debilidades:
 - ▶ No insesgamiento. Ni siquiera aumentando el tamaño de muestra te puedes quitar el sesgo.
 - ▶ Subjetividad. (e.g. aunque todos estemos de acuerdo en el modelo, no es cierto, porque no conocemos la super-población).
- ▶ Ojo, no estoy diciendo que este enfoque sea equivocado. Claramente tiene sus ventajas (principalmente económicas y muy fuertes). El problema es la subjetividad que puede echar abajo todo. Como siempre que se utilizan modelos, no hay forma alguna de saber si son ciertos.

2.4.3 'MODEL-ASSISTED APPROACH'

- ▶ Lo importante: La información auxiliar disponible y los recursos computacionales.
- ▶ En palabras llanas, combina los dos anteriores.
- ▶ Otros nombres: 'muestreo modelo asistido', 'estimación GREG'.

- ▶ Fortalezas:
 - ▶ Robustez: 'Siempre jala'. Aunque el modelo está mal especificado se obtienen buenas estimaciones porque automáticamente se le da más peso a la parte design-based. Si el modelo está muy bien especificado (resultó ser muy realista) automáticamente el método da más peso a la parte model-based.
 - ▶ Objetividad.
 - ▶ Exactitud (insesgamiento).
 - ▶ Precisión (estabilidad de las estimaciones).

- ▶ Debilidades:
 - ▶ Para que de verdad funcione y mejore al design-based, lo necesario para dar estimaciones (los g -weights) son a nivel máximo de desagregación (individuo - observación). Los g -weights dependen de las probabilidades de inclusión de los individuos y de un parámetro de variabilidad por individuo.
 - ▶ Elevados costos informáticos (información, cómputo, etc.).

2.5 MARCO MUESTRAL

Para extraer una muestra de la población se requiere de algo que denominamos **marco muestral**, **marco de muestreo** o simplemente **marco**.

Este es una lista que me permitirá:

1. Identificar los individuos de mi población y proporcionarme información adicional útil para un mejor uso del muestreo.
¿Cuántos individuos hay en la población, cómo está dividida, etc.?
2. Acceder a los individuos o poder establecer contacto con ellos.
¿Dónde están, teléfono, dirección, coordenadas, etc..?

En el peor de los casos si no existe una lista, un marco muestral puede ser:

- ▶ un mapa geográfico,
- ▶ una delimitación en el plano cartesiano,
- ▶ el *boot* de un disco duro,
- ▶ el directorio de un CD,
- ▶ el directorio telefónico,
- ▶ el listado nominal electoral,
- ▶ el padrón de un partido político, etc.

Lo importante es que el marco me **esquematice** a la población de interés.

Problemas o imperfecciones en el marco muestral:

- ▶ Incompleto (No cobertura).
- ▶ Muy general (Muy grueso, sin detalle ni info. adicional para muestrear).
- ▶ Desactualizado.
- ▶ Inexistente (el clásico problema en México).

Mucha de la **labor de muestreo tiene que ver con la construcción de un buen marco muestral**.

Importante: los errores de marco pueden ser **indetectables** en la lectura de resultados de una muestra si en su construcción fueron obviados detalles, huecos, etc. Pueden llegar a ser grandes **errores arrastrados**. Un marco muestral equivocado puede ser un gran problema.

En muestreos más complejos se requerirá que el marco proporcione **información adicional** para la obtención de estimaciones más precisas y esquemas de selección más económicos.

La bibliografía base del curso, Särndal *et al.* (1992), habla más sobre marcos muestrales y también aquella bibliografía clásica como el Kish (1965) que toca el tema de manera muy completa y hasta con sugerencias ante complicaciones.

2.6 RADIOGRAFÍA GENERAL DE UNA ENCUESTA POR MUESTREO

Para ir familiarizándonos más con el problema al que da respuesta el muestreo (inferir sobre una población a partir de un subconjunto de individuos) y con la nomenclatura (sinónimos) de lo que utilizaremos, consideremos el siguiente listado **muy** sintético del proceso de una encuesta (una aplicación muy natural del muestreo, ojo, pero no la única).

1. Una encuesta tiene que ver con un conjunto de **elementos** denominado **población finita**.
2. Se dispone de una regla o listado que define de manera **inequívoca** a los elementos que pertenecen a la población; a tal regla se le denomina **marco muestral**.
3. El objetivo de la encuesta es proveer de información sobre la población finita o sobre subpoblaciones de especial interés, por ejemplo, *hombres y mujeres* como dos subpoblaciones; tales subpoblaciones son denominadas **dominios de estudio** o simplemente **dominios**.
4. Se tiene asociado un valor de una o más **variables de estudio** para cada elemento de la población. El objetivo de la encuesta es obtener información sobre **características poblacionales** o **parámetros**.
5. Los parámetros son funciones de los valores de las variables de estudio. Estos, son desconocidos y pueden ser medidas cuantitativas de interés para la investigación en curso, por ejemplo, el ingreso total, el ingreso medio, número de desempleados; para la población entera o para dominios específicos.
6. En la mayoría de las encuestas, la observación y el acceso a los elementos individuales (en ocasiones denominados **unidades de análisis**) de la población es establecido a partir de un **marco muestral**. Este asocia a los elementos de la población con las **unidades muestrales** contenidas en el marco.

7. Una **muestra** (un subconjunto) de elementos se selecciona de la población. Esto se lleva a cabo seleccionando unidades muestrales de un marco.
8. Una muestra es una **muestra probabilística** si fue obtenida mediante un mecanismo aleatorio y con ciertos lineamientos.
9. Se realiza una **observación** de los elementos muestrales, esto es que, para cada elemento de la muestra, se hace una **medición** de las variables de estudio y sus valores son registrados. Las mediciones son acorde a un **plan de medición** bien definido.
10. Los valores registrados de las variables son utilizados para el cálculo de **estimaciones (puntuales)** de los parámetros poblacionales de interés (totales, medias, medianas, razones, coeficientes de regresión, etc.). Luego se realizan estimaciones de la **precisión** de las estimaciones (los errores). Por último, se publican los resultados.

En una encuesta por muestreo, el ejercicio de observación se limita a un subconjunto de la población. Un tipo especial de encuesta es aquella en donde toda la población es observada; denominada **censo o enumeración completa**.

- ▶ ¿Un censo significa automáticamente la estimación de un parámetro sin errores?
- ▶ ¿Qué se suele hacer con los censos en lo que toca al gran número de variables de estudio?
- ▶ ¿Cuál es la tendencia cada vez más creciente en primer mundo con respecto a los censos?

2.7 ALGUNOS COMENTARIOS

Considerando los ejemplos anteriores y lo hasta ahora visto, notar lo siguiente:

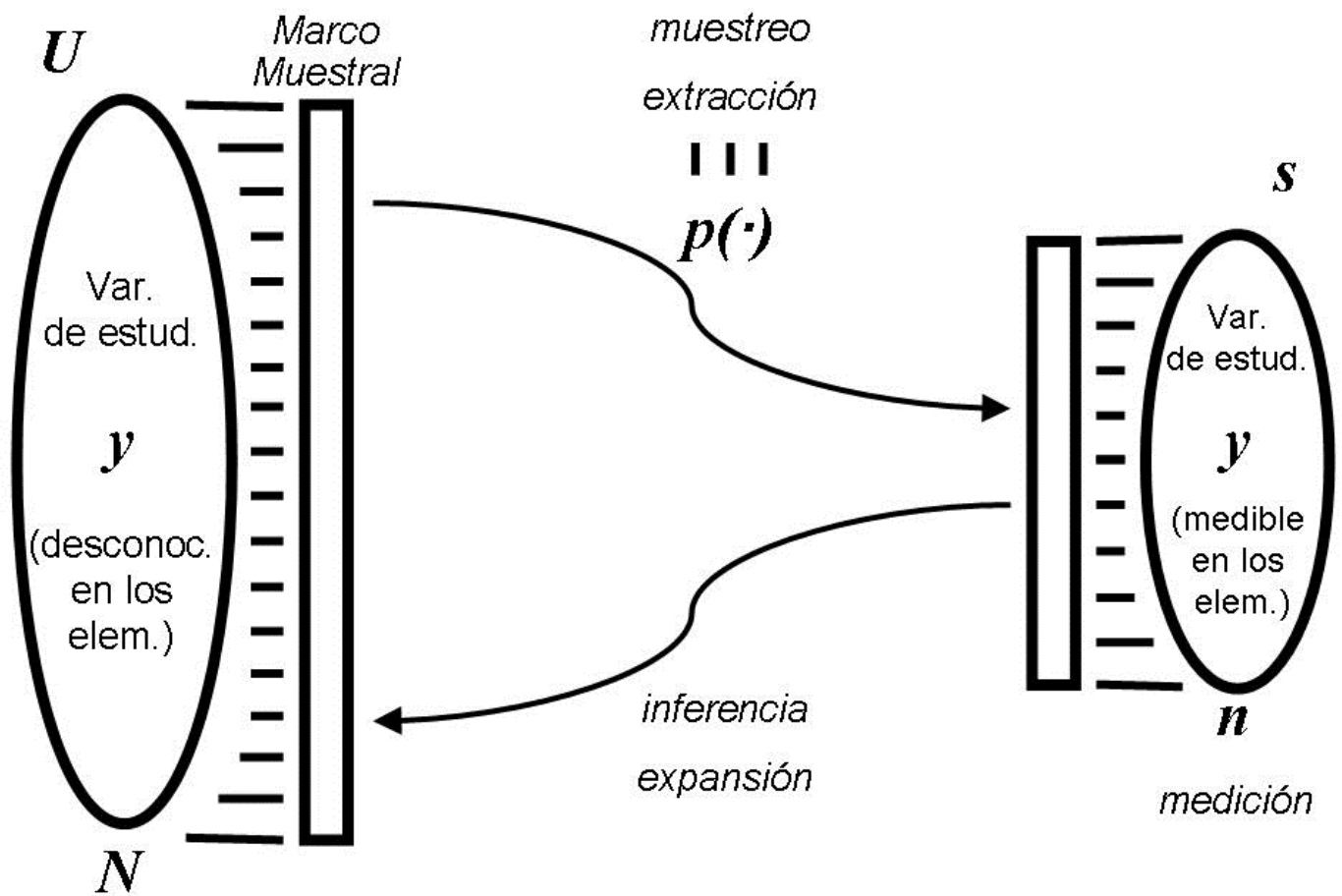
1. La complejidad de una encuesta por muestreo puede variar mucho.
2. Aunque una encuesta involucra observaciones individuales de los elementos de la población, el propósito de la encuesta **no es utilizar esos datos a nivel individual** sino la obtención de estadísticos **resumen para la población o subgrupos específicos**.
3. En la misma encuesta pueden haber **muchas variables** de estudio, **muchos dominios** de estudio, **muchos parámetros** de interés y quizás muchos tipos de estos.
4. Una muestra es cualquier subconjunto de la población. Puede o no ser extraída mediante un mecanismo aleatorio. **Nosotros nos concentraremos en aquellas probabilísticas**.

Un ejemplo de aquellas no probabilísticas son aquellas en las que un experto en la materia del estudio ligada a la encuesta decide la selección de los individuos de modo que la muestra “represente” las características de la población de estudio.

En general, **sólo en circunstancias “afortunadas” una selección no probabilística arrojaría estimaciones adecuadas**.

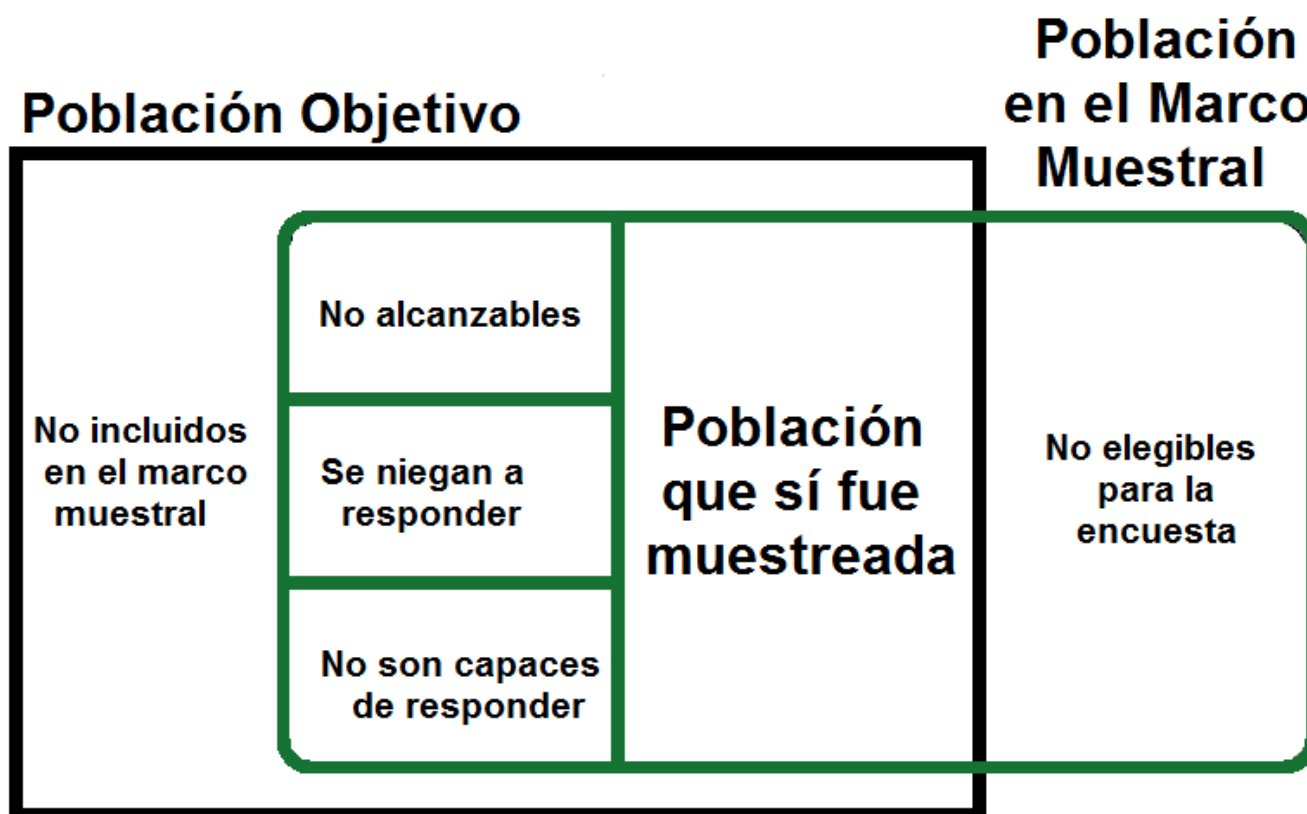
5. La correcta medición y registro de información puede ser difícil y en ocasiones imposible. Respuestas falsas, no respuesta, rechazo a responder. Todos estos **errores no muestrales** pueden llegar a ser considerables.

2.7.1 INCORPORACIÓN DE TÉCNICA A EL OBJETIVO DEL MUESTREO



2.7.2 UN EJEMPLO SOBRE EL MARCO MUESTRAL (DE LOHR, 1999)

- ▶ Población objetivo y población muestreada en una **encuesta telefónica** de posibles **votantes en una elección**.
- ▶ **No todos los hogares tienen teléfono**, de modo que cierta cantidad de personas de la población objetivo de posibles votantes no estarán asociados a los números telefónicos del marco muestral.
- ▶ En algunos hogares con teléfono, los residentes no están empadronados para votar y entonces **no son elegibles** para la encuesta.
- ▶ Algunas personas que sí son elegibles y que están en el marco muestral **no responden** debido a varias razones: No pueden contestar, no quieren contestar, o son incapaces de contestar.



CAPÍTULO 3

Muestreo probabilístico y extracción de la muestra

Ahora...

¿Cómo es la extracción? ¿Cómo se extrae la muestra?

Respuesta: Mediante **muestreo probabilístico**.

3.1 MUESTREANDO PROBABILÍSTICAMENTE

Éste es una forma de selección de muestras que **satisface ciertas condiciones**. Si no, entonces no se le puede llamar probabilístico.

3.1.1 MUESTREO EN 1 ETAPA

Para el caso en el que se hace una **selección directa de elementos** de la población, es decir, muestreo en una etapa; tales condiciones son las siguientes:

1. Es posible definir a $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$, el conjunto de todas las muestras posibles del esquema de selección.
2. Se tiene una probabilidad conocida de selección $p(s)$ asociada con cada posible muestra $s \in \mathcal{S}$.
3. El esquema de selección $p(\cdot)$, aunque está definido para s , 'hereda' a cada elemento k en la población una probabilidad de ser seleccionado $\pi_k \neq 0$.
4. Se selecciona una muestra s mediante un 'mecanismo aleatorio' que permita que cada s posible tenga exactamente la probabilidad $p(s)$ de ser seleccionada.

Nótese que 1, 2 y 4 **tienen que ver con muestras** o probabilidades de obtener éstas; mientras que 3 **tiene que ver con elementos** de la población.

Se le denomina **muestra probabilística** a una muestra obtenida bajo estas condiciones.

La función $p(\cdot)$ define una distribución de probabilidad sobre $\mathcal{S} = \{s_1, s_2, \dots, s_M\}$, el conjunto de todas las muestras posibles.

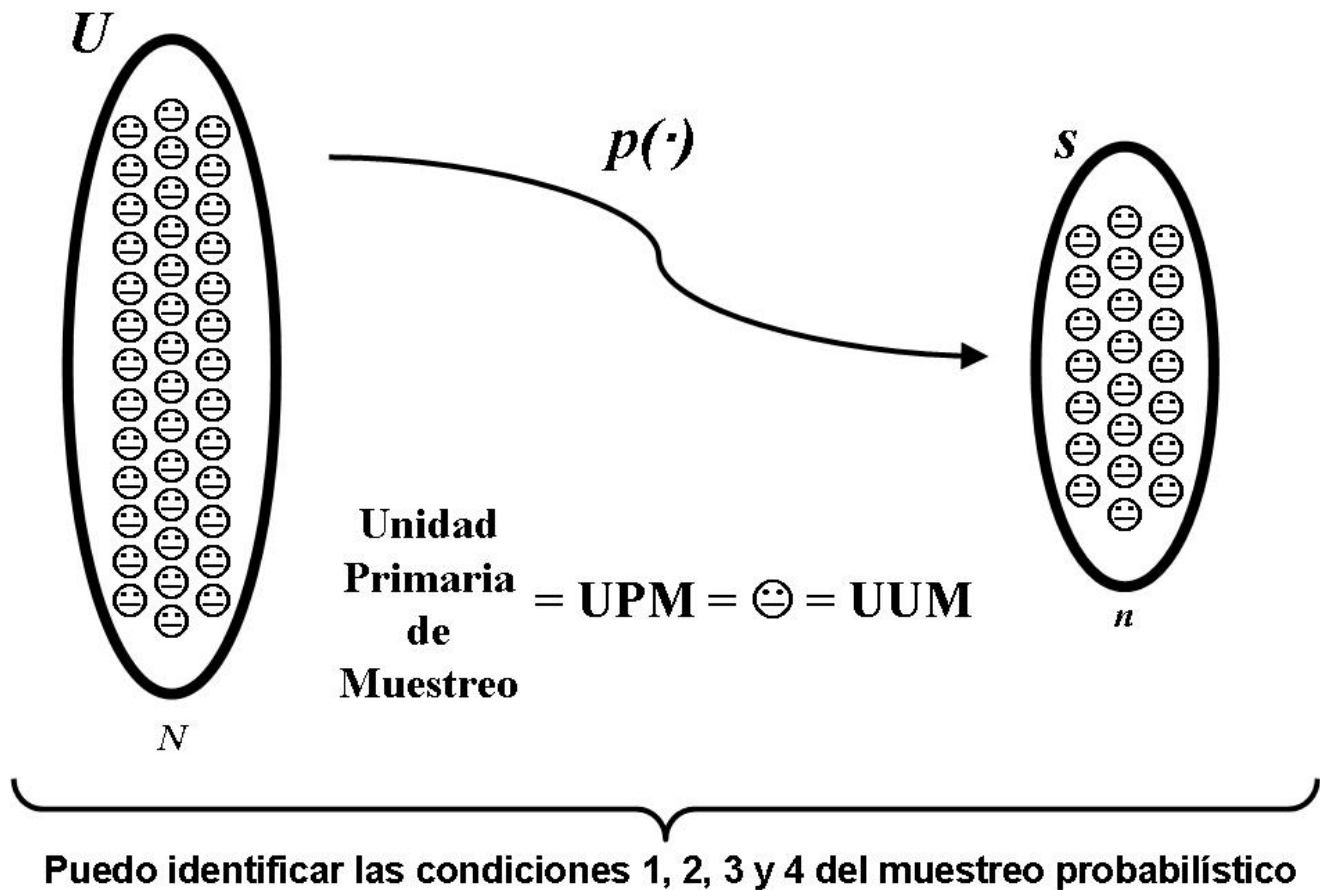
A la función $p(\cdot)$ se le denomina **función diseño de muestreo** o simplemente **función diseño**. Es la que “matematiza” la forma en que previamente establecimos será seleccionada la muestra.

La probabilidad mencionada en el punto 3 es denominada la **probabilidad de inclusión** (en la muestra) de los elementos en la población.

El proceso de aleatorización del punto 4 regularmente puede llevarse a cabo mediante un algoritmo *fácil* o ya integrado a algún software estadístico.

3.1.1.1 ESQUEMA DE MUESTREO EN 1 ETAPA

Muestreo Probabilístico en 1 Etapa



3.2 MUESTREANDO EN MÁS DE 1 ETAPA

- ▶ La selección de una muestra regularmente se lleva a cabo en dos o más etapas.
- ▶ Esto quiere decir que se seleccionan conglomerados de elementos en la etapa inicial por ejemplo y posteriormente se seleccionan individuos o elementos dentro de los conglomerados seleccionados.
- ▶ Esto puede suceder en una o más etapas de muestreo (**submuestreo**); los elementos tal cual son muestreados entonces hasta la última etapa.
- ▶ **Importante:** En un diseño de muestreo probabilístico de más de 1 etapa se tienen que cumplir las condiciones 1-4 en cada etapa.

3.2.1 MUESTREO EN 3 ETAPAS

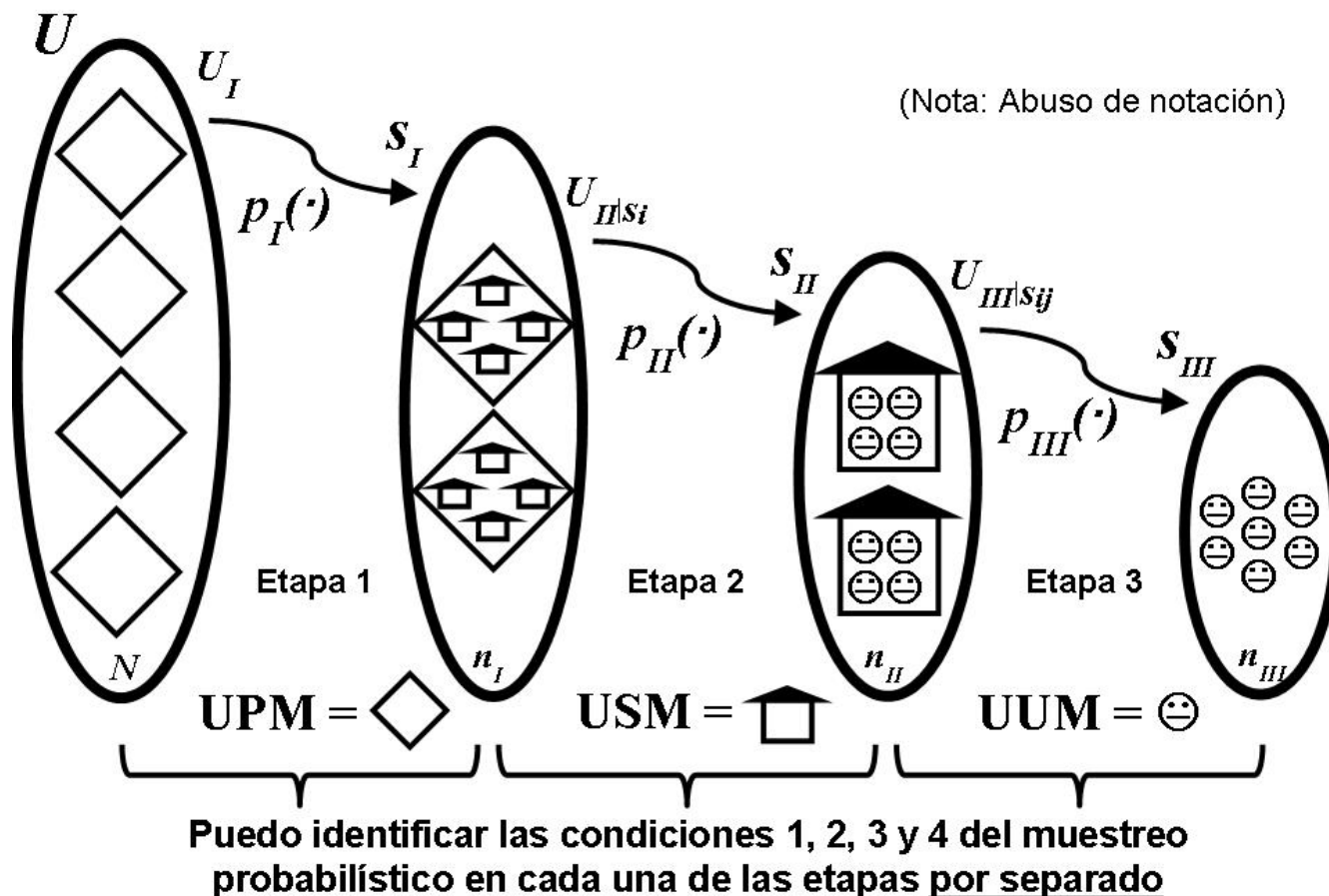
- ▶ Para ello necesito un marco muestral organizado en 3 niveles.
- ▶ Por ejemplo, de la siguiente forma:
 1. Manzanas (representado por rombos) compuesta de viviendas
 2. Viviendas que estan conformadas por individuos
 3. Individuos
- ▶ La población U de individuos está organizada de modo que tengo una población U_I de manzanas, una población U_{II} de viviendas y una población U_{III} de individuos.

La siguiente tabla ejemplifica esta estructura anidada en los datos.

Id U	Id U_I	Id U_{II}	Id U_{III}
Id Único Individuos	Id Manzanas	Id Viviendas	Id Individuos
1	1	1	1
2	1	1	2
3	1	1	3
4	1	2	1
5	1	2	2
6	1	3	1
7	1	3	2
8	2	1	1
9	2	1	2
10	2	1	3
11	2	1	4
12	2	2	1
13	2	2	2
14	3	1	1
15	3	1	2
16	3	1	3
17	3	1	4
18	3	2	1
19	3	2	2
20	3	3	1
21	3	3	2
22	3	3	3
23	3	3	4
24	4	1	1
25	4	1	2
26	4	1	3

3.2.1.1 ESQUEMA DE MUESTREO EN 3 ETAPAS

Muestreo Probabilístico en 3 Etapas



- ▶ Entonces, finalmente, se deberá tener una *probabilidad de inclusión de ser seleccionado para cada uno de los elementos de la población* sin importar el número de etapas del esquema de muestreo.
- ▶ Esto lo veremos más adelante, y se denominan las *probabilidades de inclusión de individuos (elementos) de una población en muestra*.
- ▶ Hay que tener cuidado en no confundir estas con la *probabilidad de selección de una muestra*.

3.2.2 VENTAJA DE LAS MUESTRAS PROBABILÍSTICAS

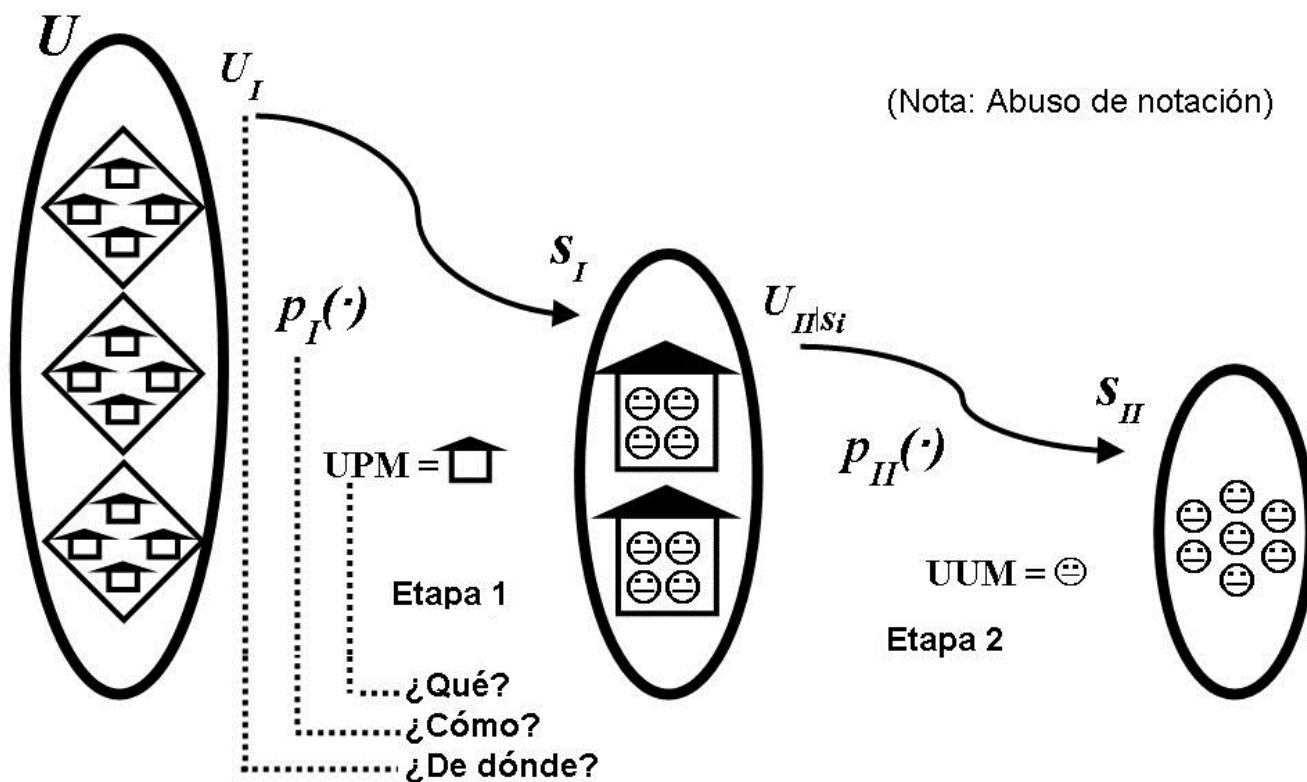
- ▶ La ventaja principal que tienen las muestras probabilísticas sobre las demás es que **permiten el uso de la teoría estadística para inferir sobre la población** de la que fueron tomadas. Con esto se tiene la **capacidad de producir medidas de error y de precisión en términos probabilísticos**.
- ▶ Por último, el muestreo probabilístico garantiza la eliminación de cualquier subjetividad **en el proceso de selección** de elementos en una muestra.
- ▶ Esa subjetividad ausente, es lo que coloquialmente algunos llaman sesgo. No obstante esta palabra tiene otras connotaciones estadísticas.
- ▶ Es por ello que las muestras obtenidas mediante muestreo probabilístico son *objetivas* y por lo tanto gozan de mayor aceptación.

3.2.3 MUESTREO EN 2 ETAPAS

- ▶ Con lo único que se sabe del ejemplo anterior de 3 etapas...
- ▶ Rápidamente... ¿Cómo podría mejorar el diseño de muestreo anterior?
- ▶ ¿Más etapas implica un mejor diseño?
- ▶ ¿Más etapas implica un diseño más económico?
- ▶ ¿Hay respuesta absoluta a estas preguntas?

3.2.3.1 ESQUEMA DE MUESTREO EN 2 ETAPAS

Muestreo Prob. en 2 Etapas (Mejorando el anterior de 3 Etapas)



CAPÍTULO 4

Estimación a partir de muestras probabilísticas

4.1 POBLACIÓN, MUESTRA Y SELECCIÓN

Considérese la **población**, U , un conjunto finito de N elementos etiquetados $k = 1, \dots, N$,

$$(4.1) \quad \{u_1, \dots, u_k, \dots, u_N\}$$

Por simplicidad, representemos al elemento k -ésimo, u_k , únicamente por su etiqueta k . De modo que:

$$(4.2) \quad U \stackrel{def}{=} \{1, \dots, k, \dots, N\}$$

Por lo pronto, tomaremos como conocido a N , que representará el **tamaño de la población**.

Ahora, considérese a y la **variable de estudio**, y sea $y_k, k \in U$ el valor de la variable y para el k -ésimo elemento de la población U . Sabemos que y_k **existe pero la desconocemos**.

Supóngase que interesa el **total poblacional** t de la variable y ,

$$(4.3) \quad t = \sum_{k \in U} y_k \stackrel{def}{=} \sum_U y_k$$

o de la **media poblacional** \bar{y}_U de la variable y ,

$$(4.4) \quad \bar{y}_U = t/N = \sum_U y_k / N$$

- Nótese que cuando y toma únicamente los valores 0 y 1 tendríamos que \bar{y}_U es una **proporción**.

- ▶ Entonces, como **una proporción es una media y la media es un total dividido entre la constante N** , plantearemos todo en términos del problema de estimar al total t .
- ▶ Esto, de nuevo es otra generalización del libro base del curso que antes no se efectuaba en libros tradicionales.
- ▶ Para la estimación de t a partir de una **muestra probabilística** s , subconjunto de elementos de la población U seleccionados mediante un mecanismo aleatorio, tendremos que observar los valores que toma $y_k, k \in s$; es decir, los valores de y únicamente para aquellos elementos que fueron seleccionados en la muestra probabilística.
- ▶ Esto es, se generarán estimaciones de t con la información que contengan las $y_k, k \in s$.

4.2 LA FUNCIÓN DISEÑO DE MUESTREO

- ▶ Ya tenemos definida nuestra población U de tamaño N , le extraeremos una muestra probabilística s mediante un **esquema aleatorio de selección**.
- ▶ De modo que es posible (aunque no siempre sencillo) determinar la probabilidad de selección $p(s)$ de la muestra específica s .
- ▶ Asumimos que existe la función $p(\cdot)$ tal que $p(s)$ indica la probabilidad de seleccionar s bajo el esquema utilizado.
- ▶ A la función $p(\cdot)$ la denominaremos **función diseño de muestreo**.
- ▶ Es fundamental pues determinará las propiedades estadísticas esenciales de las *cantidades aleatorias* calculadas a partir de la muestra
- ▶ Por ejemplo: la distribución muestral, el valor esperado y la varianza de la media muestral, la mediana muestral y la varianza muestral.
- ▶ Estas *cantidades aleatorias* vendrían siendo lo que en cursos como Inferencia Estadística se denominan estimadores, i.e. funciones con variabilidad pues dependen de un conjunto aleatorio o sucesión de variables aleatorias.
- ▶ Para un diseño dado $p(\cdot)$, se puede entonces considerar cualquier muestra s como la realización de la variable aleatoria (o *output* del evento aleatorio) S , cuya distribución de probabilidad queda explicitada mediante la función $p(\cdot)$.
- ▶ Sea \mathcal{S} el conjunto de todas las muestras s posibles. Entonces, \mathcal{S} es un conjunto de 2^N subconjuntos de U , si incluimos al conjunto vacío y también al conjunto U mismo; i.e. con un mismo diseño muestral se tienen un total de 2^N muestras posibles incluyendo a la muestra vacía y a la muestra censal.

- Entonces tenemos que:

$$(4.5) \quad \Pr \{S = s\} = p(s)$$

para cualquier $s \in \mathcal{S}$. Como $p(s)$ es una distribución de probabilidad sobre \mathcal{S} , tenemos

$$(4.6) \quad i. \quad p(s) \geq 0, \quad \forall s \in \mathcal{S}$$

$$(4.7) \quad ii. \quad \sum_{s \in \mathcal{S}} p(s) = 1$$

- Nótese que muchas de las 2^N muestras contenidas en \mathcal{S} pueden tener de hecho probabilidad cero. El subconjunto de \mathcal{S} compuesto de aquellas s cuyas $p(s)$ son estrictamente mayores que cero constituyen el conjunto de muestras verdaderamente posibles. Ellas serán las únicas que podrán ser extraídas.
- El **tamaño de muestra**, n_s , es el número de elementos en s , es decir la cardinalidad del conjunto s .
- n_s no es necesariamente el mismo para todas las muestras posibles, esto dependería del diseño de muestreo utilizado.
- El diseño de muestreo $p(\cdot)$, determina las propiedades estadísticas de las cantidades calculadas a partir de la muestra. No obstante, $p(\cdot)$ es principalmente un **constructo teórico** matemático, no práctico *per se* pero fundamental para el desarrollo de la teoría estadística que sostiene el muestreo probabilístico.
- Es de extrema importancia la elección del diseño de muestreo y a su vez la simultanea elección de un esquema de selección que haga posible la implementación del diseño. Ambos **tienen que estar íntimamente relacionados**.
- En otras palabras, **la realidad de mi forma de extraer muestras tiene que estar perfectamente compaginada con la teoría que asumo** para la extracción y/o proceso de inferencia.

4.3 PROBABILIDADES E INDICADORAS DE INCLUSIÓN

Supóngase que determinado diseño de muestreo $p(s)$ ha **quedado establecido**, i.e. que se tiene una forma matemática para $p(s)$.

4.3.1 LAS INDICADORAS DE INCLUSIÓN MUESTRAL

Entonces, **la inclusión de un elemento** determinado k **en una muestra** es un **evento aleatorio** indicado por la variable aleatoria I_k , denominada la **indicadora de inclusión muestral** del elemento k , definida de la forma siguiente,

$$(4.8) \quad I_k = \begin{cases} 1 & \text{si } k \in S \\ 0 & \text{en otro caso} \end{cases}$$

Nótese que $I_k = I_k(S)$ es una función de la variable aleatoria S .

4.3.2 LAS PROBABILIDADES DE INCLUSIÓN

De modo que la probabilidad de que el elemento k está en muestra es π_k donde,

$$(4.9) \quad \pi_k = Pr\{k \in S\} = Pr\{I_k = 1\} = \sum_{s \ni k} p(s)$$

Y la probabilidad de que los elementos k y l están simultáneamente en muestra,

$$(4.10) \quad \pi_{kl} = \pi_{lk} = Pr\{k \& l \in S\} = Pr\{I_k I_l = 1\} = \sum_{s \ni k \& l} p(s)$$

También, tenemos que,

$$(4.11) \quad \pi_{kk} = Pr \{I_k^2 = 1\} = Pr \{I_k = 1\} = \pi_k, \quad \forall k = 1, \dots, N$$

Formalmente para evitar abusos de notación, en la ecuación (4.9) lo escrito como $\{k \in S\}$ debe ser interpretado como el evento aleatorio $\{S \ni k\}$, el cual es el evento *una muestra en cuya realización contiene al elemento k* .

Entonces, dado $p(\cdot)$, se tienen asociados N valores,

$$(4.12) \quad \pi_1, \dots, \pi_k, \dots, \pi_N$$

denominadas las **probabilidades de inclusión de primer orden**. También están asociados $N(N-1)/2$ valores,

$$(4.13) \quad \pi_{12}, \pi_{13}, \dots, \pi_{kl}, \dots, \pi_{N-1,N}$$

denominadas las **probabilidades de inclusión de segundo orden**.

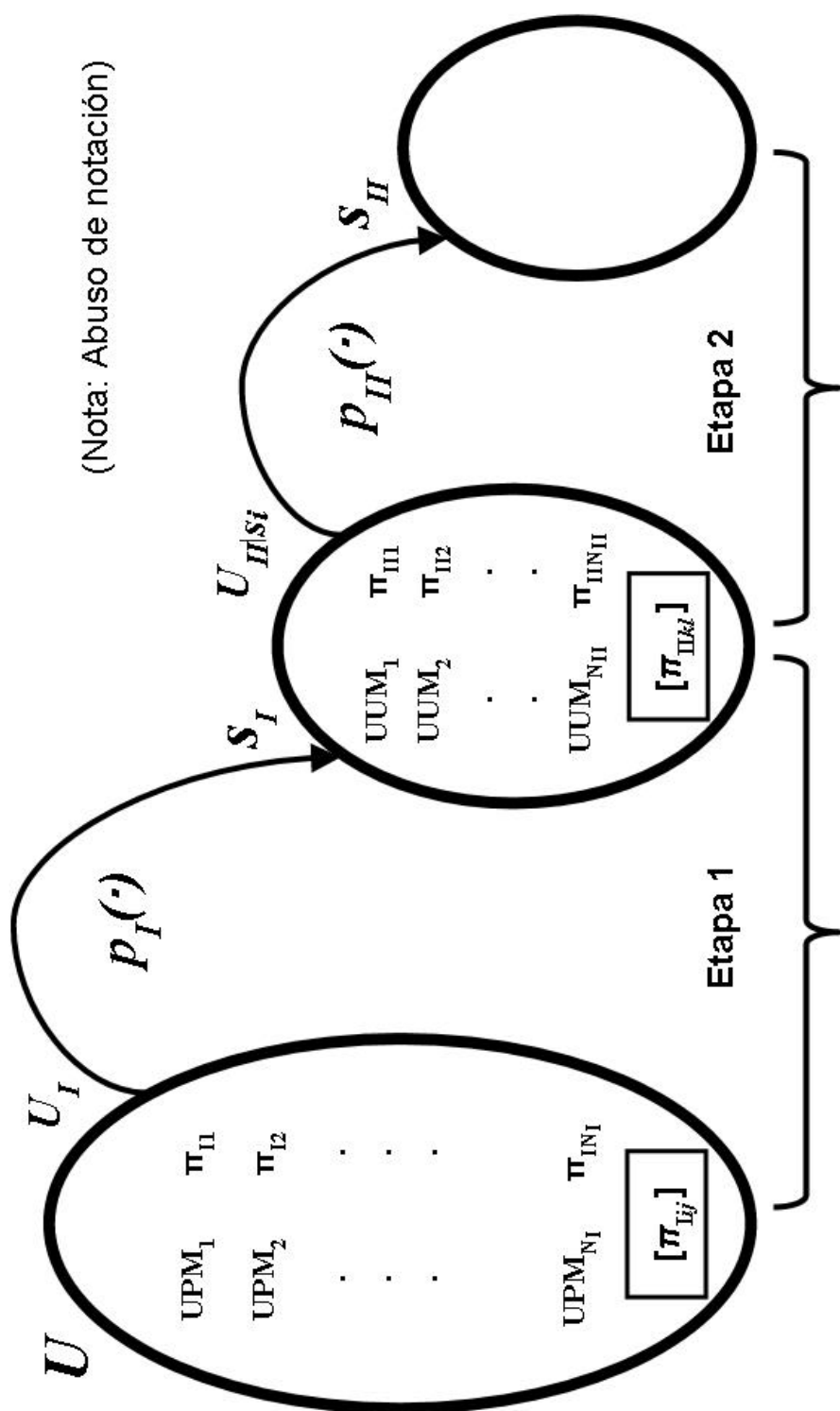
- ▶ Desde luego, así le podemos seguir con probabilidades de inclusión de tercer orden, etc... partiendo de la función diseño $p(\cdot)$, pero no tiene caso pues no serán necesarias para este curso y tampoco son contempladas o mejor dicho necesarias para los diseños de muestreo comúnmente usados.
- ▶ Esto verdaderamente implicaría innecesarias complicaciones ya que juegan un papel mucho menos importante como podremos apreciar más adelante.
- ▶ Usualmente el diseño de muestreo se escoge en función de la facilidad para el cálculo de las probabilidades de inclusión de primero y segundo orden.
- ▶ También se busca un compromiso entre tal facilidad y factibilidad en la realidad.
- ▶ Por otro lado, $p(\cdot)$ pueda llegar a ser complicada pero eso no afecta tanto.

- ▶ Existe la posibilidad de alcanzar uno de los objetivos principales, la obtención del valor esperado y la varianza de ciertas cantidades calculadas a partir de la muestra, esto último únicamente a partir de las π_k y las π_{kl} solamente.

4.3.3 COMENTARIOS SOBRE LAS PROBABILIDADES DE INCLUSIÓN

- ▶ Formalmente, hemos visto en la sección 3.1 en el punto 3, que para que una muestra sea considerada una muestra probabilística, se tendría que cumplir que $\pi_k > 0, \forall k \in U$.
- ▶ No obstante, **en la práctica** a veces se le asigna probabilidad cero a algunos individuos en la población de modo que estos nunca salgan en muestra.
- ▶ Esta práctica (previa a la extracción de la muestra) tiene como objeto eliminar de posibles muestras a individuos que se sabe previamente que la información que pudiesen aportar no es importante. Desde luego, esta es una práctica **delicada** porque varias expresiones tienen estos valores como denominador.
- ▶ En el muestreo directo de individuos (es decir, una sola etapa de muestreo), todas las $\pi_k, k = 1, \dots, N$ son (**y deben ser**) normalmente **conocidas antes de la extracción de la muestra**.
- ▶ En diseños de muestreo más complejos esto no es posible o resulta muy complicado. Sin embargo, en muestreo en varias etapas, conocer todas las π_k y las π_{kl} no es indispensable pues basta con el conocimiento *a priori* de las probabilidades de inclusión para las unidades de muestreo al momento de la extracción en cada etapa.
- ▶ En otras palabras, basta con conocerlas previo a muestrear en cada etapa. Así lo podemos apreciar en el siguiente gráfico.

Probs. de Inclusión (en un Muestreo Probabilístico en 2 Etapas)



No es indispensable conocer todas las probabilidades de inclusión de todas las etapas. Este conocimiento puede ser etapa por etapa. Lo mismo aplica para el marco muestral.

4.3.4 ESTADÍSTICOS BAJO EL DISEÑO MUESTRAL

En la teoría general de estadística el término **estadístico** se refiere a una función que toma valores reales cuyo valor puede **variar** de acuerdo a las diferentes realizaciones de determinado experimento.

En muestreo, nosotros queremos examinar cómo un estadístico varía de una realización de una muestra s a otra dentro del conjunto aleatorio S . En otras palabras, es la variación del estadístico, muestra a muestra lo que nos interesa como muestristas.

Si $Q(S)$ es una función real del conjunto aleatorio S , esta función tomará valores una vez que se tenga la realización s de S y se tengan recolectados los datos de los elementos que componen a s .

En la práctica cuando una muestra es extraída, exactamente una realización s del conjunto aleatorio S ha ocurrido. Una vez que s se realizó, asumimos que es posible observar y medir determinadas variables de interés, por ejemplo y y z , para cada elemento k en s . Entonces, por ejemplo para el caso del estadístico $Q(S) = \sum_S y_k / \sum_S z_k$, podemos, después de la medición, calcular el valor de la realización con el estadístico $Q(s) = \sum_s y_k / \sum_s z_k$.

¡Importante!. Nótese que en este resumido ejemplo y y z son variables en el sentido de tomar posibles valores diferentes y_k y z_k para los elementos k contenidos en s . No obstante, y y z **no serán tratados como variables aleatorias**.

¡Importante!. La naturaleza aleatoria del estadístico $Q(S)$ **recae solamente** del hecho de que **el conjunto S es aleatorio**.

Es muy importante que esto quede claro. En otras palabras, **la aleatoriedad reside en cuál muestra fue extraída y no en los posibles valores de las variables de interés en los elementos de la muestra**. De modo que consideraremos que los valores de las variables de interés son dados (**fijos**) en los elementos, **no son aleatorios pero sí son**

desconocidos; la incertidumbre que manejaremos por medio de la estadística vendrá entonces de la muestra que utilizaremos y no de lo que medimos en los elementos que la componen.

Como el estadístico $Q(S)$ es una variable aleatoria, ésta tiene varias propiedades estadísticas.

Definición 4.3.4.1 *La esperanza y la varianza del estadístico $Q = Q(S)$ se definen, respectivamente, por las siguientes expresiones,*

$$(4.14) \quad E(Q) = \sum_{s \in S} p(s)Q(s)$$

$$(4.15) \quad V(Q) = E\{[Q - E(Q)]^2\}$$

$$(4.16) \quad = \sum_{s \in S} p(s)[Q(s) - E(Q)]^2$$

La covarianza entre dos estadísticos $Q_1 = Q_1(S)$ y $Q_2 = Q_2(S)$ se define por,

$$(4.17) \quad C(Q_1, Q_2) = E\{[Q_1 - E(Q_1)][Q_2 - E(Q_2)]\}$$

$$(4.18) \quad = \sum_{s \in S} p(s)[Q_1 - E(Q_1)][Q_2 - E(Q_2)].$$

Nótese (de nueva cuenta) que estas definiciones hacen referencia a la **variación sobre todas las muestras posibles** que pueden ser obtenidas bajo el diseño de muestreo dado, $p(s)$.

Para hacer énfasis en lo anterior, algunos libros de muestreo utilizan los términos esperanza diseño, varianza diseño y covarianza diseño. Aquí no utilizaremos la palabra *diseño* (como apellido) en estos estadísticos dado que no hay riesgo de mala interpretación.

Los estimadores que serán de nuestro interés pueden (y gracias a esto son más fáciles

de manejar) ser expresados como funciones de las indicadores de inclusión muestral definidas en la ecuación (4.8). Es por lo tanto importante describir las propiedades básicas de los estadísticos $I_k = I_k(S)$, para $k = 1, \dots, N$.

Resultado 4.3.1.1 *Para un diseño de muestreo $p(s)$ arbitrario, y para $k, l = 1, \dots, N$,*

$$(4.19) \quad E(I_k) = \pi_k$$

$$(4.20) \quad V(I_k) = \pi_k(1 - \pi_k)$$

$$(4.21) \quad C(I_k, I_l) = \pi_{kl} - \pi_k \pi_l \stackrel{\text{def}}{=} \Delta_{kl}$$

Demostración. Primero, nótese que $I_k = I_k(S)$ es una variable aleatoria Bernoulli, entonces $E(I_k) = \Pr\{I_k = 1\} = \pi_k$, esto por la ecuación (4.9). Luego como $E(I_k^2) = E(I_k) = \pi_k$, se sigue que $V(I_k) = E(I_k^2) - \pi_k^2 = \pi_k(1 - \pi_k)$. Ahora, $I_k I_l$ también es una Bernoulli que toma el valor 1 si y sólo si ambas k y l pertenecen a s . Entonces, por la ecuación (4.10), $E(I_k I_l) = \Pr\{I_k I_l = 1\} = \pi_{kl}$. Y por lo tanto finalmente se tiene que $C(I_k, I_l) = E(I_k I_l) - E(I_k)E(I_l) = \pi_{kl} - \pi_k \pi_l$ \square

Dependiendo del diseño, $C(I_k, I_l)$ puede ser positiva, negativa o cero. Nótese que si $k = l$,

$$(4.22) \quad V(I_k) = \Delta_{kk}$$

4.4 MUESTREO BERNOULLI (BE)

- ▶ N elementos en un marco muestral con cierto orden, que no nos interesa.
- ▶ De antemano, se fija π constante, $0 < \pi < 1$, i.e. $\pi_k = \pi, \forall k \in U$
- ▶ Sean $\varepsilon_1, \dots, \varepsilon_N$ un conjunto de N realizaciones independientes de una variable aleatoria $Unif(0, 1)$.
- ▶ La selección o no del elemento k -ésimo se decide de la siguiente forma:
Si $\varepsilon_k < \pi$, entonces k es seleccionado, de otro modo no. $k = 1, \dots, N$.
- ▶ Entonces, la probabilidad de seleccionar al individuo k -ésimo es:

$$Pr\{\varepsilon_k < \pi\} = \pi, \quad \forall k \in U.$$

- ▶ Y tenemos que para $k \neq \ell$ los eventos $\{k \in s\}$ y $\{\ell \in s\}$ son independiente.
- ▶ El número de elementos seleccionados $n_s = \#(s) = \sum_U I_k$, se distribuye $Bin(N, \pi)$. Es decir, n_s **no es fijo**, es una variable aleatoria.

$$Pr\{n_s = n\} = \binom{N}{n} \pi^n (1 - \pi)^{N-n}, \quad n = 1, \dots, N.$$

- ▶ De modo que:

$$E_{BE}[n_s] = N\pi \quad y \quad V_{BE}(n_s) = N\pi(1 - \pi)$$

- ▶ Y entonces tenemos que:

$$p(s) = \pi^{n_s} (1 - \pi)^{N-n_s}$$

- ▶ Notar que no tenemos que conocer N para determinar las π 's.
- ▶ Notar que **el tamaño de muestra es aleatorio**, pero sabemos como se comporta.

- ▶ Ojo: Esto no es un modelo impuesto. **Predefinimos** que así sería la selección de individuos, con una probabilidad fija π .
- ▶ ¿En qué casos es útil este diseño de muestreo?
- ▶ ¿Algún ejemplo real?



4.5 MUESTREO ALEATORIO SIMPLE (SI)

- ▶ Queremos seleccionar específicamente n elementos de una población de N sin reemplazo y donde cada selección sea con igual probabilidad.
- ▶ Lo más fácil es imaginarlo como si seleccionáramos n elementos de una urna con N elementos. Elemento que fue seleccionado se separa y se siguen extrayendo elementos de la urna hasta alcanzar una muestra de tamaño n .
- ▶ Hay varias formas de llevar a cabo este esquema de selección. El más sencillo es un procedimiento 'basado en extracciones', tal cual como se mencionó, con una urna o con 'papelitos':
 1. Seleccionar con igual probabilidad $1/N$ al primer elemento de entre N posibles y apartarlo.
 2. Seleccionar con igual probabilidad $1/(N - 1)$ al segundo elemento de entre los restantes $N - 1$ y apartarlo.
 - ⋮
 - n . Seleccionar con igual probabilidad $1/(N - n + 1)$ al n -ésimo elemento de entre los restantes $N - n + 1$ después de $n - 1$ extracciones y apartarlo.
- ▶ ¿Otra forma? ¿Se les ocurre otra forma?
- ▶ Otra forma es 'siguiéndose':
 1. Seleccionar con igual probabilidad $1/N$ al primer elemento de entre N posibles y reemplazarlo (devolverlo a la urna).
 2. Repetir el paso anterior ν veces hasta obtener n elementos distintos, $Pr\{\nu \geq n\} = 1$.
- ▶ ¿Otra forma? ¿Se les ocurre otra forma?
- ▶ Otra forma es, grosso modo, convirtiendo el primer esquema en un esquema 'secuencial de lista' (Fan, Muller & Rezucha, 1962).

1. Se generan $\varepsilon_1, \varepsilon_2, \dots$ realizaciones $Unif(0, 1)$ independientes. Seleccionar el primer elemento si $\varepsilon_1 < n/N$, si no, no.
2. Para los siguientes elementos $k = 2, 3, \dots$, sea n_k el número de elementos que hemos seleccionado entre los primeros $k-1$ elementos en la lista de la población (marco). Si

$$\varepsilon_k < \frac{n - n_k}{N - k + 1}$$

se elige el elemento k -ésimo, si no, no.

3. El procedimiento termina cuando $n_k = n$.

► ¿Otro? Sí, uno muy fácil que yo llamo 'con hojita de Excel'. Pizarrón.

1. Se generan $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ realizaciones $Unif(0, 1)$ independientes.
2. Ordenar la población acorde con estas variables generadas. Y elegir los primeros n elementos.

► Éste último tiene la particularidad de generar tantas muestras SI como yo quiera y que además no se traslapen ('negatively coordinated samples').

► ¿Desventajas de estos esquemas? ¿Alternativas?

► De modo que, bajo SI tenemos que:

$$p(s) = \begin{cases} 1/\binom{N}{n} & \text{si } \#(s) = n, \\ 0 & \text{en otro caso.} \end{cases}$$

- Y usando la definiciones que vimos, podemos calcular π_k y $\pi_{k\ell}$.
- Tenemos exactamente $\binom{N-1}{n-1}$ muestras s que tienen al elemento k -ésimo, y $\binom{N-2}{n-2}$ muestras s que tienen a los elementos k y ℓ -ésimo ($k \neq \ell$).
- Dado que todas las muestras de tamaño n tienen la misma probabilidad:

$$\pi_k = \binom{N-1}{n-1} / \binom{N}{n} = \frac{n}{N}, \quad k = 1, \dots, N$$

y

$$\pi_{k\ell} = \binom{N-2}{n-2} / \binom{N}{n} = \frac{n(n-1)}{N(N-1)}, \quad k \neq \ell = 1, \dots, N$$

- ▶ Notar que aquí $n_s = n$ es fijo. Por cómo definimos que íbamos a seleccionar.
- ▶ ¿Cómo ven los textos tradicionales al muestreo aleatorio simple?

CAPÍTULO 5

Estimadores y sus propiedades estadísticas básicas

5.1 ESTIMADORES COMUNES

- ▶ Vimos en general estadísticos bajo el diseño muestral...
- ▶ La gran mayoría de los estadísticos que utilizaremos son estimadores.
- ▶ Un **estimador** es un estadístico **pensado para la producción de valores cercanos a un** valor poblacional de interés que desconocemos, que denominaremos **parámetro** y denotaremos por θ .
- ▶ Si, por ejemplo, sólo hay una variable de estudio y , se puede pensar a θ como una función de y_1, \dots, y_N , los N valores de y en la población.

$$\theta = \theta(y_1, \dots, y_N)$$

- ▶ Un ejemplo de parámetro podría ser el **total poblacional** t de y ,

$$\begin{aligned}\theta &= t \\ &= \sum_{k \in U} y_k \\ &\stackrel{def}{=} \sum_U y_k\end{aligned}$$

Otro, la **media poblacional** \bar{y}_U de y ,

$$\begin{aligned}\theta &= \bar{y}_U \\ &= \frac{t}{N} \\ &= \frac{\sum_U y_k}{N}\end{aligned}$$

- ▶ Otro ejemplo de parámetro que es función de dos variables de estudio y y z ,

sería la **razón de los totales poblacionales** de y y z ,

$$\theta = \frac{\sum_U y_k}{\sum_U z_k}$$

Denotaremos al estimador de θ como,

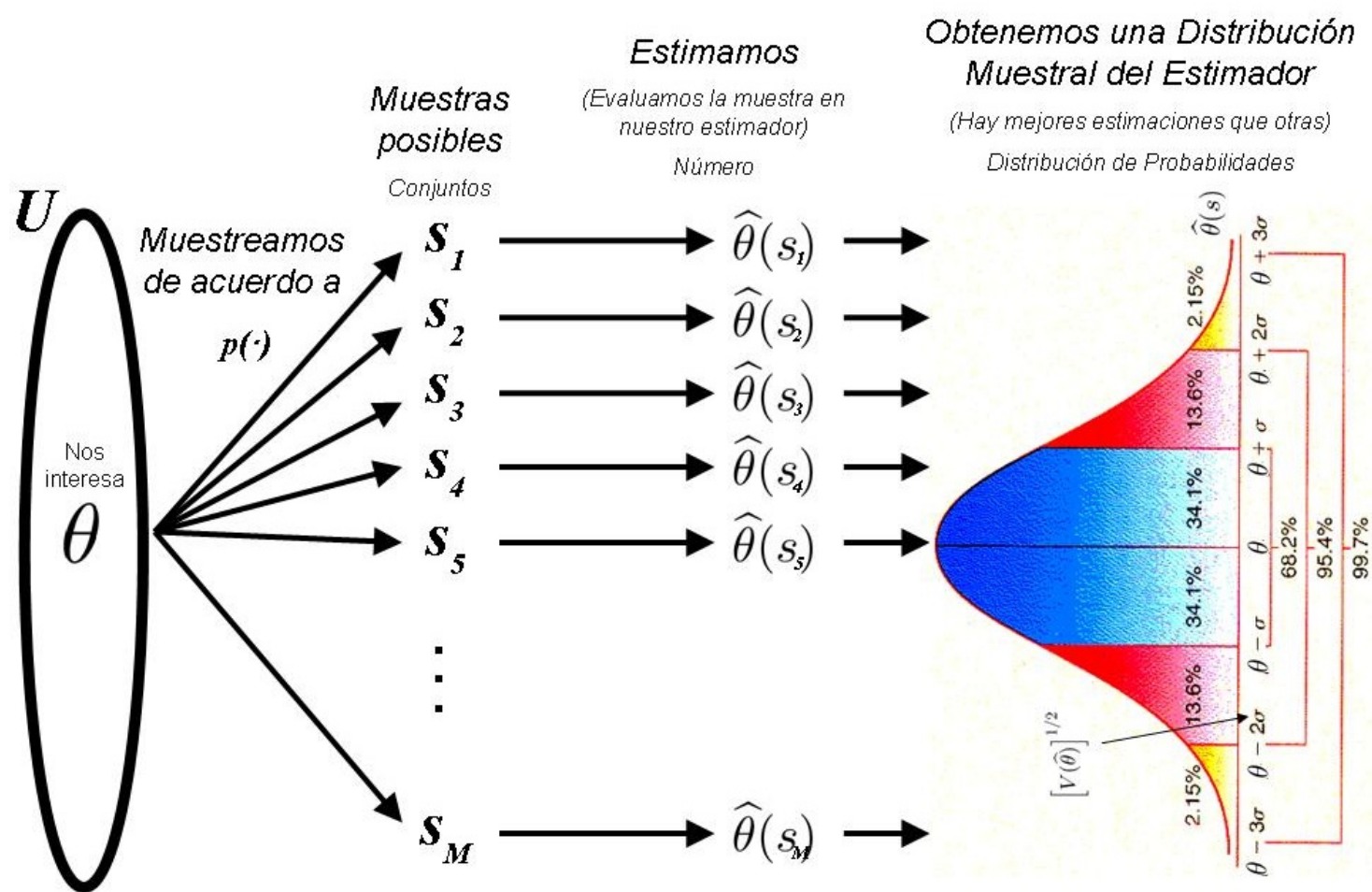
$$\hat{\theta} = \hat{\theta}(S)$$

Si s es una realización del conjunto aleatorio S , entonces podemos calcular $\hat{\theta}$ a partir de la(s) variable(s) de estudio asociadas a los elementos $k \in s$.

5.2 DISTRIBUCIÓN MUESTRAL DE UN ESTIMADOR

Como ya se dijo, para nosotros es de interés describir la variación muestra a muestra del estimador $\hat{\theta}$ que utilicemos.

Un estimador que varíe poco alrededor del valor desconocido del parámetro θ es intuitivamente mejor que otro que varíe mucho.



Esta descripción del **comportamiento muestra a muestra** de $\hat{\theta}$ la logramos mediante la **distribución muestral del estimador $\hat{\theta}$** .

En ella se describen todos los valores posibles del estimador junto con la probabilidad

correspondiente para cada uno de esos valores, todo esto bajo el diseño de muestreo $p(s)$ en uso.

Ejemplo de la Distribución Muestral: Las Letras (A,B,C,D,E,F,G,H).

 k u_k y_k θ $\hat{\theta}$ N n $\#(\mathcal{S})$ i s_i $\hat{\theta}(s_i)$

Frecuencias relativas

Distribución muestral de $\hat{\theta}$

En teoría, dado el diseño, el estimador y las mediciones de la variable de interés; **habría de ser posible** la obtención de la distribución muestral del estimador.

No obstante, **puede ser complicado debido al gran número de muestras posibles** que se traducirían en un gran número de valores del estimador.

Sin embargo, es posible tener, de manera teórica a partir de la Definición 4.3.4.1, medidas resumen (usualmente desconocidas) que describen importantes aspectos de la distribución muestral de un estimador.

La **esperanza** de $\hat{\theta}$ está dada por,

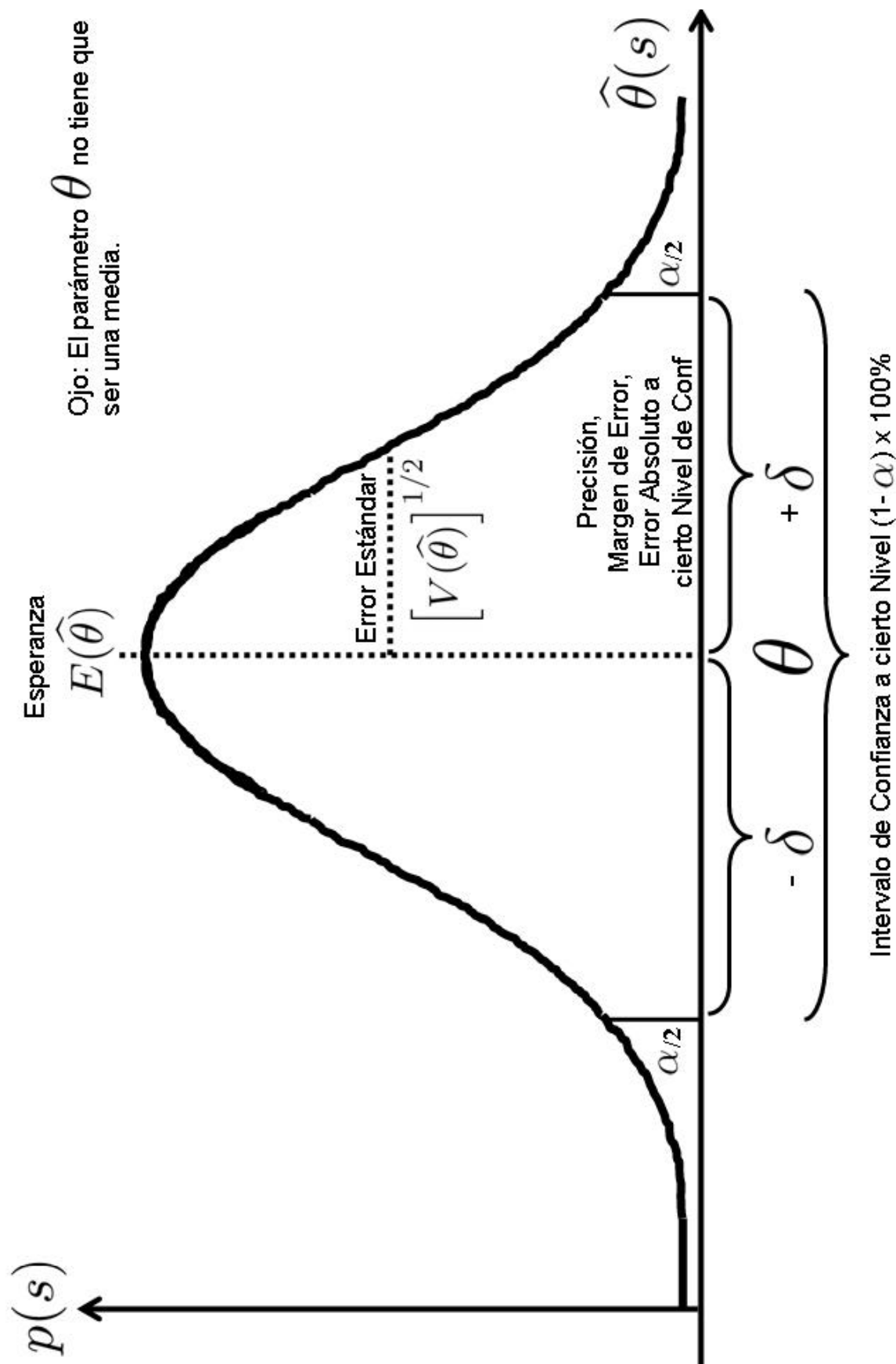
$$E(\hat{\theta}) = \sum_{s \in \mathcal{S}} p(s) \hat{\theta}(s)$$

Mientras que la **varianza** está dada por,

$$V(\hat{\theta}) = \sum_{s \in \mathcal{S}} p(s) \left[\hat{\theta}(s) - E(\hat{\theta}) \right]^2$$

Distribución Muestral de un Estimador.

(Mismo estimador diferentes muestras)



Hay dos medidas importantes de la calidad de un estimador $\hat{\theta}$, son el sesgo y el error cuadrático medio. El **sesgo** de $\hat{\theta}$ se define como,

$$(5.1) \quad B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Un estimador $\hat{\theta}$ se dice que es **insesgado** de θ si:

$$(5.2) \quad B(\hat{\theta}) = 0, \quad \forall \mathbf{y} = (y_1, \dots, y_N)' \in \mathbb{R}^N$$

El **error cuadrático medio** de $\hat{\theta}$ se define como,

$$(5.3) \quad MSE(\hat{\theta}) = E [\hat{\theta} - \theta]^2$$

$$(5.4) \quad = \sum_{s \in \mathcal{S}} p(s) [\hat{\theta}(s) - \theta]^2$$

$$(5.5) \quad = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

Y, por supuesto, si el estimador $\hat{\theta}$ es insesgado para θ , entonces por la ecuación (5.5), $MSE(\hat{\theta}) = V(\hat{\theta})$.

(Es muy importante que esto quede claro, es un error muy común.) **Nótese la diferencia entre una estimación y un estimador.** Una estimación $\hat{\theta}(s)$ es un número, es producido por un estimador $\hat{\theta} = \hat{\theta}(S)$, una función.

$\hat{\theta}(s)$ es un número que puede ser calculado una vez que hay una realización s del conjunto aleatorio S y ha sido observado y la(s) variable(s) de estudio ha(n) sido medida(s) para los elementos $k \in s$.

En adelante, ignoraremos la diferencia tipográfica entre S , el conjunto aleatorio y s la realización de S . Por simplicidad designaremos a ambos con la notación s .

En palabras, un estimador es insesgado si el promedio ponderado (sobre todas las muestras posibles utilizando las probabilidades $p(s)$ como pesos) es igual al valor del parámetro desconocido.

Los estimadores que son de mayor interés al muestreo son aquellos que son insesgados o aproximadamente insesgados.

Estos últimos son aquellos en donde el sesgo es muy pequeño. ¿Qué tan pequeño? Se puede relativizar tal sesgo con lo que se está midiendo (coeficiente de variación). También, es posible calcular tal sesgo. El muestrista decidirá si lo considera grande o pequeño.

Una nota, formalmente hablando. **No existen estimaciones insesgadas** pues las estimaciones (como ya se dijo) son números, valores constantes. **Los que pueden o no ser insesgados son únicamente los estimadores.** No obstante, en la práctica, cuando se habla coloquialmente de una estimación insesgada se está hablando de una estimación proveniente de un estimador insesgado.

Un muestrista en la práctica **tendrá que decidir entre varios posibles estimadores para un mismo parámetro.** Buscará utilizar aquel cuya distribución muestral está altamente concentrada, poco dispersa alrededor de θ .

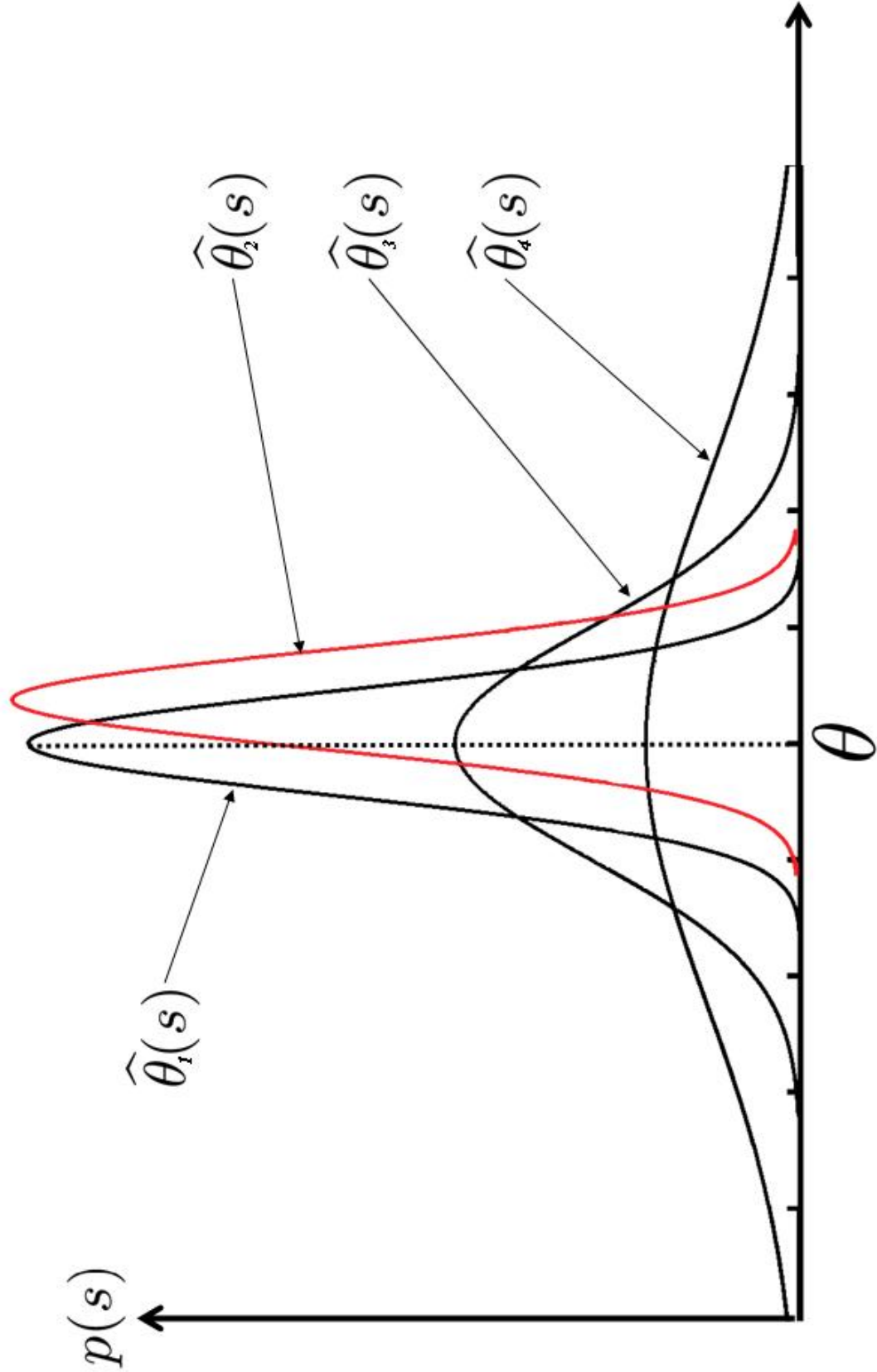
No obstante, **aún cuando la distribución muestral está altamente concentrada alrededor de θ siempre existirá una pequeña posibilidad de que** nuestra muestra en particular

haya sido *desafortunada (mala)*, de tal manera que **la estimación caiga en una de las colas de la distribución, muy lejos de θ . Tendrán que vivir con esta posibilidad.**

¿Entonces qué puede uno controlar como muestrista?

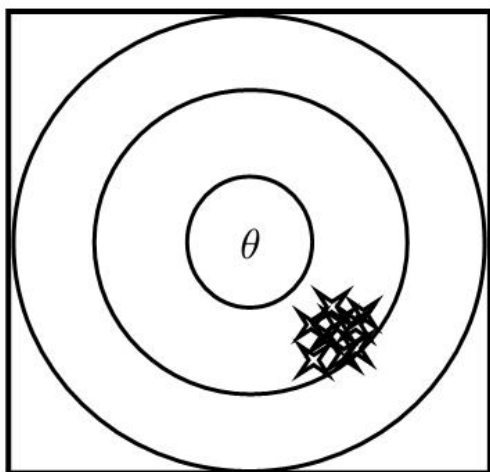
Las Distribuciones Muestrales. Mejora de Estimaciones.

(Diversos Estimadores. ¿Importa mucho que mi estimador sea incesgado? Depende...)

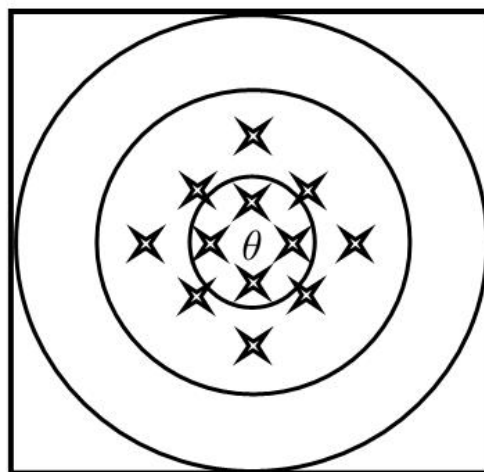


Las Distribuciones Muestrales.

Estimadores Insegados y No Insegados. Argumento Bayesiano (Ejemplo del Arquero)



*Arquero Bayesiano
(estimador Bayesiano)*



*Arquero Frecuentista
(estimador Frecuentista)*

A la raíz cuadrada de la varianza del estimador $\left[V(\hat{\theta})\right]^{1/2}$ se le denomina el **error estándar** del estimador $\hat{\theta}$. Al cociente del error estándar del estimador y la esperanza del estimador, $CV(\hat{\theta}) = \left[V(\hat{\theta})\right]^{1/2} / E(\hat{\theta})$ se le denomina el **error estándar relativo** o el **coeficiente de variación** del estimador.

En la práctica, se desconoce a $V(\hat{\theta})$. Esto porque tendría que conocer todos los valores posibles que toma el estimador de muestra en muestra y para ello necesitaría conocer la variable de interés en toda la población.

Por lo tanto, tal varianza se estima a partir de los datos disponibles de la muestra mediante el estimador $\hat{V}(\hat{\theta})$.

Pero este estimador, $\widehat{V}(\widehat{\theta})$, nos dice poca información de manera directa, pues está en unidades al cuadrado de las unidades en las que está el estimador $\widehat{\theta}$, de modo que se acostumbra tomar su raíz cuadrada, el **error estándar estimado**, $[\widehat{V}(\widehat{\theta})]^{1/2}$ y también se calcula el **coeficiente de variación estimado**, (normalmente expresado en porcentaje) que se define de la siguiente manera,

$$(5.6) \quad cve(\widehat{\theta}) = \frac{[\widehat{V}(\widehat{\theta})]^{1/2}}{\widehat{\theta}}$$

Nota. En la práctica suele llamarse coloquialmente al cve como el coeficiente de variación, aunque esto no es correcto si observamos las dos definiciones anteriores. No obstante, no hay confusión pues es evidente que si uno está trabajando con datos muestrales, no es posible el cálculo del coeficiente de variación de acuerdo a la definición de la expresión específica y por lo tanto se utiliza la expresión (5.6) que finalmente tiene la misma intención o utilidad.

¿Para qué nos sirve el cve?

¿Por qué no lo utilizan en México?

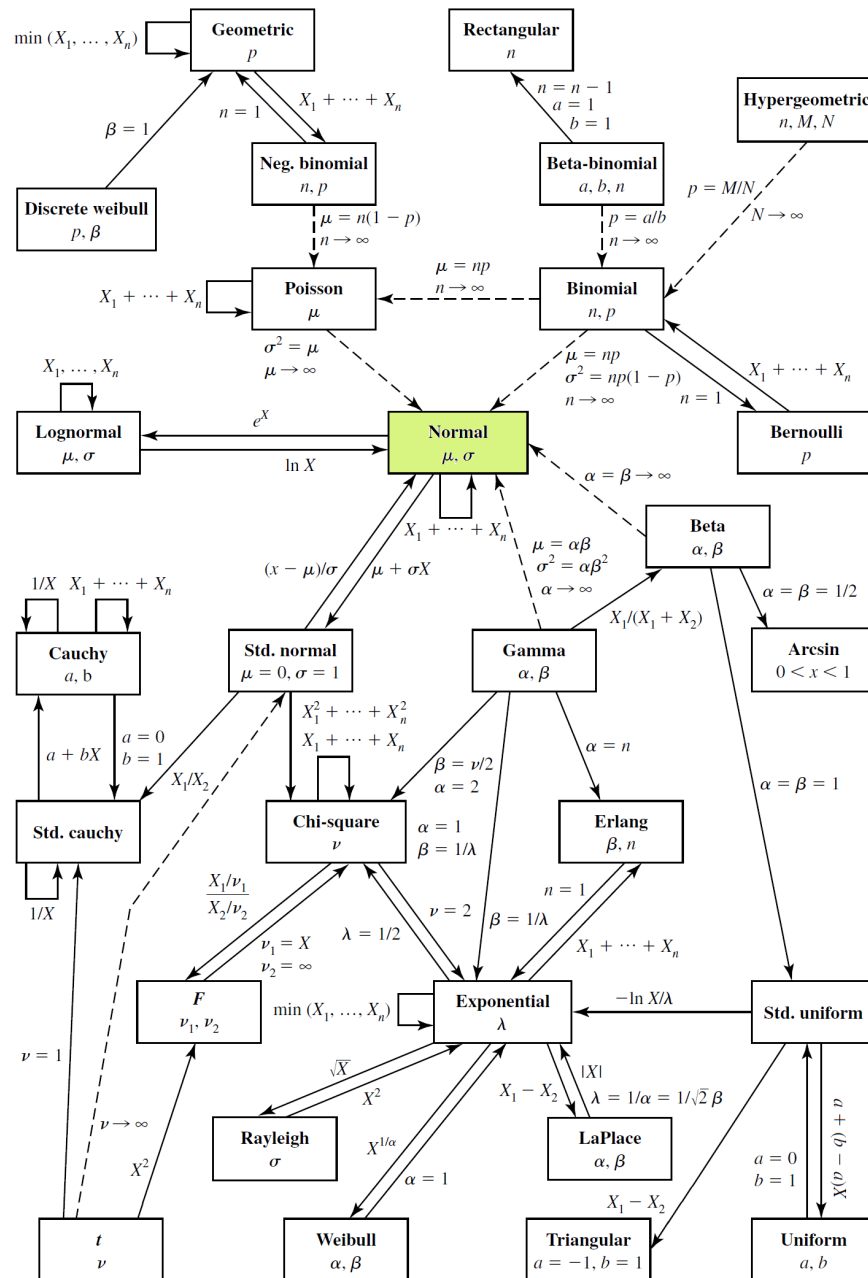
¿Tiene sentido que al muestrear de la misma forma, con el mismo tamaño de muestra y medir lo mismo, se tengan mejores o peores estimaciones que otras?

Ejemplo de los Millones de Dólares

¿Entonces, cuáles son los niveles aceptables o utilizados para el cve?

APÉNDICES

RELACIÓN ENTRE DISTRIBUCIONES DE PROBABILIDAD



Fuente: Leemis, L. M. (1986). Relationships among common univariate distributions. *Am. Stat.*, **40**, pp. 143–6.

VARIANZAS HIPOTÉTICAS DE ALGUNAS DISTRIBUCIONES (KISH, 1965)


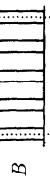
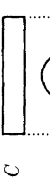


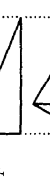
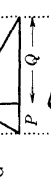



	Mean	Variance	Special Cases
	P	PQ	$P = \frac{1}{2} \quad \sigma^2 = \frac{1}{4} \text{ max}$
	$\frac{1}{2}$	$\frac{1}{12} + \frac{1}{6K}$	$K \rightarrow \infty \quad \sigma^2 \rightarrow \frac{1}{12} \text{ min}$
	$\frac{1}{2}$	$\frac{1}{12}$	
	$\frac{1}{2}$	$\frac{1}{16}$	
	$\frac{1}{2}$	$\frac{1}{36}$	
	$\frac{1}{3}$	$\frac{1}{18}$	
	$\frac{(1+P)/3}{2} + \frac{(1-R)(1+P)}{3}$	$\frac{(1-PQ)/18}{(1-R)(1-PQ)/18} + \frac{R}{12} + \frac{R(1-R)(1-2P)^2}{36}$	$P = \frac{1}{2} \quad \sigma^2 = \frac{1}{24} \text{ min}$
	$\frac{Q}{2}$	$\frac{Q}{12} (1+3P)$	$P = \frac{1}{3} \quad \sigma^2 = \frac{1}{9} \text{ max}$
	$\frac{Q+PL}{3}$	$\frac{Q+PL^2+2PQ(1-L^2)}{18}$	$L = \frac{2Q}{1+2Q} \quad \sigma^2 = \frac{3Q}{18(1+2Q)} \text{ min}$
	for $L = 0: Q/3$	$\frac{Q}{18} (1+2P)$	$P = \frac{1}{4} \quad \sigma^2 = \frac{1}{16} \text{ max}$

TABLE 8.2.II Variances of Several Finite Distributions

To facilitate comparisons, the distributions are presented with unit areas and unit width; if the width is changed to K , the variance is changed by K^2 . Irregularities and discreteness of actual distributions would tend to increase the variances of smooth distributions. For example, $\sigma^2 = \frac{1}{12}$ of C is increased to $\sigma^2 = \frac{1}{12} + \frac{1}{6K}$, for the discrete rectangular of $K + 1$ points, shown with $K = 4$ in B . With stratification the variances may be decreased. Other distributions may be obtained by combining simpler forms. Thus, G , H , I , and J were obtained from C and F .

**TEOREMA CENTRAL DEL LÍMITE,
VELOCIDAD DE CONVERGENCIA A UNA NORMAL,
APROXIMACIONES A LA VARIANZA DE UNA VARIABLE,
DESIGUALDAD DE TCHEBYCHEV
(MENDEZ, ESLAVA & ROMERO, 2004)**

$$= \frac{\frac{10}{10} \sum Y_i}{160} + \frac{\frac{50}{10} \sum Y_i}{160} + \frac{\frac{100}{10} \sum Y_i}{160} = \frac{\sum_{i=1}^n W_i Y_i}{N}$$

Para estimar el total, cada elemento de la muestra se multiplica por su **factor de expansión**, w_i ; los elementos del primer estrato se multiplican por 1, los del segundo por 5, y los del tercero por 10. Si se quiere el promedio, además se divide entre $N=160$.

1.3 Teorema central del límite

Un teorema fundamental en estadística es el **Teorema central del límite**. De manera laxa, dice que los promedios de **muchas muestras** probabilísticas de una población tienden, al aumentar el tamaño de muestra n , a tener distribución normal, a pesar de que la variable que se mide no tenga distribución normal en la población; se ejemplifica en la Figura 1.6. Una definición más formal de este teorema se encuentra en la sección 2.5 de este texto.

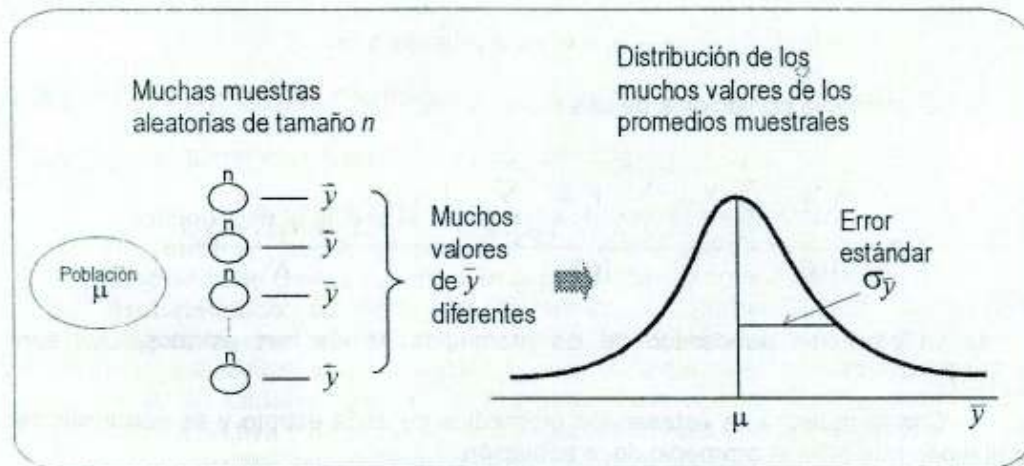


Figura 1.6 Teorema central del límite

Para que se alcance una distribución parecida a la normal en el conjunto de posibles promedios muestrales se requiere que n sea grande. Sin embargo, la rapidez de acercamiento a la normal (velocidad de convergencia) también depende de la forma de la distribución de la variable en la población. En la Figura 1.7 se consignan tamaños mínimos de muestra para una "buena" cercanía a la normal, según la forma de la distribución poblacional. Esto se ha establecido empíricamente en estudios de simulación.

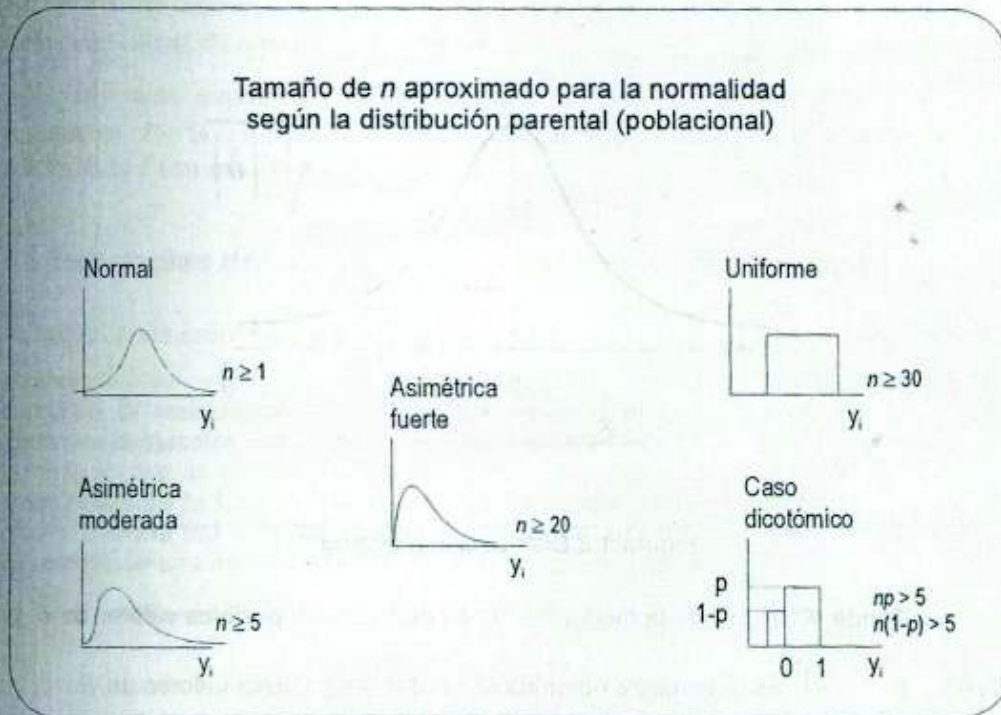


Figura 1.7 Tamaño de muestra

En general, en el trabajo de muestreo, en la población se tendrán parámetros θ , que al tomar muchas posibles muestras con un diseño de muestra específico y una forma dada de estimador, produce muchos valores de $\hat{\theta}$. El Teorema central del límite opera con muestras grandes, ver Figura 1.8.

Forma	Nombre	Varianza = σ_y^2
	Uniforme	$h^2/12$
	Triangular simétrica	$h^2/24$
	Triangular asimétrica	$h^2/18$
		$h^2/8$
	Elipse	$h^2/16$
	Normal	$h^2/36$
a ← h → b		

TABLA 3.1 Varianza de distribución en función de forma y amplitud

Con un conocimiento más o menos profundo del fenómeno estudiado (el que determina $U(U_i)=Y_i$ y el tipo de unidades U_i) se puede determinar h y la forma de la distribución de los valores de Y y con ellos obtener σ_y^2 que se usará posteriormente para fijar n . Kish (1965, pág. 262) presenta una ampliación de esta tabla.

Recordemos que:

$$V(X) = E[X - E(X)]^2 = E(X^2) - E^2(X)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$$

donde $f(x)$ es la función de densidad.

$$n = \frac{1}{\frac{\delta^2}{z_{\alpha/2}^2 S_y^2} + \frac{1}{N}} \doteq \frac{z_{\alpha/2}^2 S_y^2}{\delta^2} \quad (4.4)$$

Si $\alpha = 0.05$ entonces:

$$n \doteq \frac{(1.96)^2 S_y^2}{\delta^2}$$

Se puede usar $n' = \frac{z_{\alpha/2}^2 S_y^2}{\delta^2}$ como una primera aproximación y luego

corregir usando
$$n = \frac{n'}{1 + \frac{n'}{N}}$$

Si no se puede suponer normalidad de la distribución del estimador, se recurre a la desigualdad de Tchebycheff.

4.2.1 Desigualdad de Tchebycheff

Sea U una variable aleatoria con cualquier distribución y $E(U) = \mu_U, V(U) = \sigma_U^2$

$$\Rightarrow P[|U - \mu_U| \geq \lambda \sigma_U] \leq \frac{1}{\lambda^2}$$

$$\Rightarrow P[|U - \mu_U| \leq \lambda \sigma_U] \geq 1 - \frac{1}{\lambda^2}$$

$$\Rightarrow P[U - \lambda \sigma_U \leq \mu_U \leq U + \lambda \sigma_U] \geq 1 - \frac{1}{\lambda^2}$$

$$\Rightarrow P[\bar{y} - \lambda \sqrt{V(\bar{y})} \leq \bar{Y} \leq \bar{y} + \lambda \sqrt{V(\bar{y})}] \geq 1 - \frac{1}{\lambda^2}$$

$$\lambda = 2 \quad 1 - \frac{1}{\lambda^2} = .75$$

$$\lambda = 3 \quad 1 - \frac{1}{\lambda^2} = .889$$

$$\lambda = 4.4 \quad 1 - \frac{1}{\lambda^2} = .95$$

$$\delta = 4.4 \sqrt{V(\bar{y})}$$

EJERCICIOS DE MUESTREO

1. (Pregunta abierta) ¿Cuál es el objetivo principal del muestreo, es decir, en qué situaciones se usa o qué pregunta ayuda a responder el muestreo?
2. (Pregunta abierta) ¿Cómo podemos relacionar las siguientes ideas en una sola oración: variabilidad, muestreo, obtención y recolección de datos, estimación, inferencia, población, responder preguntas, precisión, términos probabilísticos, control, medición, parte de la estadística? Es decir, haga una oración que contenga todas las palabras y que a la vez no esté diciendo algo equivocado.
3. (Pregunta abierta) ¿Qué diferencia tienen los libros tradicionales de muestreo y el libro de Särndal que estamos utilizando?
4. (Pregunta abierta) ¿Qué relación hay entre el software de muestreo en general y el Särndal?
5. (Pregunta abierta) ¿Cuál es la principal desventaja de un enfoque particularizado del muestreo en la práctica, en la oficina, en la realidad?
6. (Pregunta abierta) Comente en sus palabras cuál sería el procedimiento general o esqueleto del proceso que involucra una encuesta. Como si lo estuviera platicando o explicando a un político o a un joven sin contacto previo con el muestreo.
7. (Pregunta abierta) Proporcione 3 ejemplos sobre el uso del muestreo diferente a una encuesta electoral o de opinión pública. Es decir, se necesitan ejemplos en donde no se trate de una encuesta. En donde no se necesite un cuestionario tal cual como ordinariamente se hace en una encuesta de opinión. De preferencia de ejemplos diferentes a los comentados en clase.
8. (Pregunta abierta) Es importante definir bien todos los elementos o detalles involucrados dentro de un ejercicio de muestreo de poblaciones finitas ¿Qué relación tiene esto con el ejercicio de inferir?
9. (Pregunta abierta) ¿Qué es un marco muestral y para qué me sirve dentro de la teoría de muestreo?
10. (Pregunta abierta) ¿Por qué es importante tener un marco muestral de buena calidad y actualizado?
11. (Pregunta abierta) ¿En qué casos tengo problemas con mi marco muestral, cuáles son los típicos problemas que pueden presentarse?
12. (Pregunta abierta) ¿Una encuesta me sirve para responder preguntas de un individuo en particular. Sí o no? Explique ampliamente.
13. (Pregunta abierta) ¿Todos los errores en una encuesta tienen que ver con muestreo. Sí o no? Explique ampliamente.
14. (Pregunta abierta) Explique de manera simple las ventajas y desventajas de un enfoque de muestreo basado en diseño.
15. (Pregunta abierta) Pensando en un enfoque de muestreo basado en modelos, explique ¿por qué es posible tener tamaños de muestra muy pequeños en este "approach"?
16. (Pregunta abierta) Explique ¿cómo es posible que el enfoque basado en diseño pueda utilizar diseños de muestreo (o probabilidades de inclusión) arbitrarias y a la vez no se considera un enfoque subjetivo?
17. (Pregunta abierta) ¿Qué es el muestreo probabilístico?
18. (Pregunta abierta) Comente por qué no es posible determinar que una muestra es probabilística si sólo se observa la muestra extraída.
19. (Pregunta abierta) ¿Qué son las probabilidades de inclusión?
20. (Pregunta abierta) ¿Qué es el diseño de muestreo?
21. (Pregunta abierta) ¿Cuál es la diferencia entre $p(s)$ y π_k ?
22. (Pregunta abierta) ¿Para qué me sirve determinar $p(s)$ y π_k en todo este asunto del muestreo que vemos en el curso. Qué importancia tiene cada uno en la teoría vista?
23. (Pregunta abierta) ¿Es posible (¿y por qué?) utilizar técnicas de muestreo que hemos visto con muestras no probabilísticas?
24. (Pregunta abierta) ¿Qué es un parámetro (en la teoría de muestreo)?
25. (Pregunta abierta) ¿Un parámetro tiene variabilidad. Sí, no, por qué?
26. (Pregunta abierta) ¿Y la variable de estudio, es una variable aleatoria. Sí, no, por qué?
27. (Pregunta abierta) ¿Un estimador de un parámetro tiene variabilidad. Sí, no, por qué?

28. (Pregunta abierta) Explique cómo es eso de que un estimador estima un parámetro. ¿Qué es un estimador? ¿Cómo funciona con "peras y manzanas"? ¿Qué quiero de un estimador y cómo me aseguro de que eso que quiero suceda? Explíquelo a un niño preguntón.
29. (Pregunta abierta) ¿De donde viene la variabilidad en el muestreo bajo el enfoque basado en diseño?
30. (Pregunta abierta) ¿La variabilidad en el muestreo basado en diseño la puedo controlar o mínimo describir? ¿Para qué me interesa controlarla o describirla? ¿Cómo? ¿Mediante qué? Explique.
31. (Pregunta abierta) ¿Cuál es la diferencia entre un estimador y una estimación?
32. (Pregunta abierta) ¿Qué es la distribución muestral? ¿Qué me dice? ¿Es fácil obtenerla siempre. Sí, no, por qué? En caso de que no, ¿Qué puedo hacer entonces?
33. (Pregunta abierta) ¿Por qué nos importa estimar en todo momento la media y la varianza de un estimador? ¿Cómo se conecta con el concepto de la distribución muestral?
34. (Pregunta abierta) ¿Qué tiene que ver con la calidad del diseño de muestreo que utilicemos el cálculo o estimación de la varianza?
35. (Pregunta abierta) ¿Cómo se relaciona en general un total, una media y una proporción?
36. (Pregunta abierta) Si la calidad de un estimador, una de las características de las que depende es el sesgo de éste, ¿Qué significa que un estimador sea insesgado formalmente hablando? ¿Y que significa en palabras coloquiales como las entendería para un político o cliente comercial?
37. (Pregunta abierta) ¿Es lo mismo hablar del sesgo de un estimador que de que una muestra tiene sesgo, como habla coloquialmente la gente ajena a técnicas de muestreo? Sí, no, explique ampliamente.
38. (Pregunta abierta) ¿Por qué formalmente hablando no existe una estimación insesgada?
39. (Pregunta abierta) ¿Explique cómo se construye una distribución muestral de un estimador? Explique como para un chavito de preparatoria.
40. (Pregunta abierta) Hasta lo que hemos visto, si se quisieran mejorar las estimaciones. ¿En qué elementos tengo control (es decir, no depende del azar) y qué cosa usted podría alterar o mejorar?
41. (Pregunta abierta) ¿En poblaciones finitas, es posible determinar todas las muestras posibles? ¿Sirve de algo eso en la práctica, necesito listarlas todas?
42. (Pregunta abierta) ¿Para qué nos sirve el coeficiente de variación estimado? Explique su utilidad práctica a un subalterno que estudió matemáticas.
43. (Pregunta abierta) ¿Cómo explicarle a un político o a un niño en términos coloquiales en realidad qué hace el coeficiente de variación? Ejemplifique si lo considera pertinente.
44. (Pregunta abierta) En palabras, sin fórmulas ni notación matemática... ¿De qué se trata el uso de los estimadores π o de Narain-Horvitz-Thompson? ¿Cuál es la idea intuitiva que hay detrás? Explique ampliamente de manera simple. Ejemplifique si lo considera pertinente.
45. (Pregunta abierta) ¿Qué restricciones hay en las probabilidades de inclusión para poder utilizar los estimadores de Narain-Horvitz-Thompson? ¿Qué restricciones tengo para establecerlas?
46. (Pregunta abierta) ¿Qué es la fracción de muestreo y qué información me da si la tengo términos porcentuales?
47. (Pregunta abierta) Explique ¿qué significa estratificar en términos prácticos y en términos matemáticos?
48. (Pregunta abierta) ¿Por qué se recomienda estratificar como una técnica útil para mejorar estimaciones? ¿Cómo convencería a su jefe ignorante en muestreo sin tanto tecnicismo?
49. (Pregunta practica abierta) Suponga que tiene un marco muestral de 40mil registros. Usted sabe de antemano que la variable Z , disponible en su marco, es "ideal" para utilizarse como variable de estratificación. Desafortunadamente, no todos los registros en su marco tienen registros de esa variable. Aproximadamente un 20 % de su marco muestral no presenta información sobre tal variable. ¿Cuál es la mejor alternativa que usted sugeriría? Discuta ampliamente las otras alternativas y por qué lo que propone es mejor. Convenza al jefe que estudió medicina.
50. (Verdadero o Falso con justificación) La función diseño de muestreo es la que determina las propiedades estadísticas del estadístico que estoy utilizando como estimador.
51. (Verdadero o Falso con justificación) En muestreo directo de elementos, es decir en 1 etapa, y bajo un diseño Si se requiere forzosamente tener el marco muestral completo que identifique a los elementos de la población.
52. (Verdadero o Falso) Si se incorporan más etapas al diseño de muestreo regularmente se aumenta la varianza del estima-

dor.

53. (Verdadero o Falso) La ventaja principal de las muestras probabilísticas sobre las no probabilísticas es que no hay errores no muestrales.
54. Para mejorar la precisión en un diseño de muestreo de varias etapas se sugiere tratar de aumentar el tamaño de muestra de las unidades primarias de muestreo, es decir el número de elementos a muestrear en la primera etapa. Muchas veces esto tiene que hacerse disminuyendo el número de unidades últimas de muestreo para no afectar el tamaño de muestra global.
55. (Verdadero o Falso con justificación) Es posible obtener muestras insesgadas incluso bajo diseños de muestreo diferentes al SI.
56. (Verdadero o Falso con justificación) El tamaño de muestra se determina mayormente por el tamaño de la población objetivo.
57. (Verdadero o Falso con justificación) En un muestreo SI. Si censamos se obtiene una varianza del estimador igual a cero y también la estimación de la varianza del estimador es igual a cero.
58. (Verdadero o Falso con justificación) Una proporción es una media de variables continuas.
59. (Verdadero o Falso) En el muestreo aleatorio simple, todas las muestras tienen la misma probabilidad de ser extraídas.
60. (Verdadero o Falso) En el muestreo aleatorio simple estratificado, todos los elementos de la población tienen la misma probabilidad de ser seleccionados.
61. (Verdadero o Falso) En el muestreo aleatorio simple, todos los elementos de la población tienen la misma probabilidad de ser seleccionados.
62. (Verdadero o Falso con justificación) Para mejorar la precisión en un diseño de muestreo se sugiere aumentar el tamaño de muestra.
63. (Verdadero o Falso con justificación) Siempre que tenga un nivel de precisión en los dominios de estimación, al combinar las estimaciones para dar una estimación global, el nivel de precisión de la estimación global es mejor que el de la estimación por dominios.
64. (Verdadero o Falso con justificación) Para estimar proporciones se pueden usar prácticamente las mismas expresiones matemáticas que para estimar medias.
65. (Verdadero o Falso con justificación) El muestreo polietápico, es decir en más de dos etapas de muestreo requiere forzosamente de un marco muestral completo que identifique a todas las unidades últimas de muestreo.
66. (Verdadero o Falso con justificación) El deff teórico para cualquier estimador del diseño SI es igual a cero siempre. Esto por su definición.
67. (Verdadero o Falso con justificación) Siempre que utilizamos conglomeración se aumenta la precisión en mis estimaciones.
68. (Verdadero o Falso con justificación) Según la teoría vista en el curso. El esquema real de muestreo puede ser diferente a mi función diseño de muestreo al momento de estimar. Se vale y es correcto.
69. (Verdadero o Falso con justificación) Siempre que se quiera mejorar la precisión en un diseño de muestreo en varias etapas se sugiere reducir el número de etapas, es decir dejar de conglomerar para algunas etapas.
70. (Verdadero o Falso con justificación) El coeficiente de variación (teórico, no estimado) puede tener valores iguales a cero si censo.
71. (Verdadero o Falso con justificación) El error estándar y la desviación estándar no son lo mismo.
72. (Verdadero o Falso) Si muestreamos bajo el enfoque basado en modelos lo estocástico o variabilidad está en el componente aleatorio del modelo.
73. (Verdadero o Falso con justificación) No se pueden calcular errores de estimación con muestreo no probabilístico. Por eso no tiene sentido calcular un tamaño de muestra.
74. (Verdadero o Falso con justificación) Para calcular un tamaño de muestra a cierta precisión y confianza necesito siempre el supuesto de Normalidad.
75. (Verdadero o Falso con justificación) Una manera de estimar a N , el tamaño de la población, es sumando los factores de expansión de los individuos caídos en muestra.
76. (Verdadero o Falso) La probabilidad de inclusión conjunta para el par de elementos (k, k) , es igual a la probabilidad de inclusión de primer orden del elemento k .

77. (Verdadero o Falso con justificación) Es posible tener probabilidades de inclusión de primer orden igual a n/N y tener un diseño de muestreo $p(\cdot)$ distinto del muestreo SI.
78. (Verdadero o Falso con justificación) No se puede estimar puntualmente una proporción si no se conocen sus probabilidades π_{kl} .
79. (Verdadero o Falso con justificación) Con las expresiones que vimos en clase, no es posible calcular la varianza con un tamaño de muestra menor a 2.
80. (Verdadero o Falso con justificación) Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo aleatorio simple.
81. (Verdadero o Falso con justificación) El tamaño de muestra se determina mayormente por el tamaño de la población objetivo.
82. (Verdadero o Falso con justificación) Es conservador que la estimación de varianza de un estimador tenga un sesgo negativo a uno positivo. Es decir, es conservador obtener errores estándares ligeramente sub-estimados.
83. (Verdadero o Falso con justificación) En las expresiones de estimación puntual de Narain-Horvitz-Thompson las probabilidades de inclusión pueden ser arbitrarias sin restricción.
84. (Verdadero o Falso con justificación) Los errores no muestrales siempre son pequeños en comparación a los errores muestrales.
85. (Verdadero o Falso con justificación) Al incorporar más etapas al diseño de muestreo se puede perder el insesgamiento del estimador puntual lineal.
86. (Verdadero o Falso con justificación) Siempre que la población es mucho más grande, la muestra tiene que ser mucho más grande.
87. (Verdadero o Falso con justificación) Siempre que se quiera mejorar la precisión en una etapa específica de muestreo se sugiere disminuir el número de unidades muestrales correspondientes a esa etapa.
88. (Verdadero o Falso con justificación) Se necesitan al menos tanta cantidad de estratos como cantidad de dominios de estudio tengo planeados.
89. (Verdadero o Falso) Si censamos una población de elementos tenemos una fracción de muestreo de 1.
90. (Verdadero o Falso con justificación) De acuerdo a la teoría vista en el curso. El total de elementos en mi población a los que les asigno probabilidad $\pi_k = 1$ no puede ser mayor al tamaño de muestra n .
91. (Verdadero o Falso con justificación) Si sumamos las probabilidades de inclusión de los elementos en toda mi población obtenemos exactamente el valor n .
92. (Verdadero o Falso con justificación) Cuando usamos muestreo aleatorio simple no podemos asumir el gran supuesto estadístico de tener observaciones independientes idénticamente distribuidas.
93. (Verdadero o Falso con justificación) Siempre que la población es más chica mejora la precisión de mis cálculos.
94. (Verdadero o Falso con justificación) Por su definición, Δ_{kl} es la correlación de las indicadoras de inclusión muestral de los elementos k y l .
95. (Verdadero o Falso con justificación) Un parámetro tiene variabilidad y esta se mide por la varianza de éste, pero para calcular su varianza se requiere de toda la información de la población.
96. (Verdadero o Falso con justificación) Cuando alcanzo cierto error estándar en mis estimaciones globales, si quiero dar resultados por sub-poblaciones, dominios o cruces, estos tendrán un error estándar más grande.
97. (Verdadero o Falso con justificación) Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo aleatorio simple.
98. (Verdadero o Falso con justificación) Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo Bernoulli.
99. (Verdadero o Falso con justificación) Siempre que se quiera mejorar la precisión en un diseño de muestreo se sugiere estratificar.
100. (Verdadero o Falso con justificación) La varianza del estimador de un parámetro en un muestreo estratificado aleatorio simple es casi siempre menor que la varianza si no hay estratos y se utilizó un muestreo aleatorio simple.
101. (Verdadero o Falso con justificación) Siempre se disminuye la varianza del estimador si se aumenta el tamaño de muestra en un diseño SI.

102. (Verdadero o Falso con justificación) De acuerdo al curso. No es posible asignar probabilidades de inclusión 1 a algunos elementos en el marco muestral porque no estaríamos haciendo muestreo probabilístico.
103. (Verdadero o Falso con justificación) Si estratificamos un diseño de muestreo (sin importar si es un diseño de muestreo de más de una etapa), ésta puede hacer perder al estimador lineal su insesgamiento.
104. (Verdadero o Falso con justificación) No es posible tener tamaño de muestra 1 en un estrato, aun cuando su tamaño poblacional sea 1.
105. (Verdadero o Falso con justificación) No existen restricciones en el tamaño de muestra asignado a los estratos cuando se incorpora una estratificación al diseño de muestreo utilizado.
106. (Verdadero o Falso con justificación) En un muestreo en varias etapas. No es posible utilizar la muestra de la etapa anterior como población para extraer muestras en la etapa siguiente.
107. (Ejercicio algebraico) Vimos en clase (y usted demostró como tarea opcional) que:
Resultado 4.3.1.1 Para un diseño de muestreo $p(s)$ arbitrario, y para $k, l = 1, \dots, N$,

$$\begin{aligned} E(I_k) &= \pi_k \\ V(I_k) &= \pi_k(1 - \pi_k) \\ C(I_k, I_l) &= \pi_{kl} - \pi_k \pi_l \stackrel{def}{=} \Delta_{kl} \end{aligned}$$

Sea n_s el tamaño de muestra para cualquier diseño de muestreo, tenemos que éste puede expresarse en términos de las indicadoras de inclusión muestral I_k como: $n_s = \sum_U I_k$.

(a) Calcule: $E(n_s)$

(b) Sabiendo que:

$$V\left(\sum_U I_k\right) = \sum_{k \in U} \sum_{\ell \in U} C(I_k, I_\ell)$$

Complete la expresión para $V(n_s)$, rellenando las siguientes expresiones:

$$\begin{aligned} V(n_s) &= \sum_U \pi_k(1 - \pi_k) + \\ &= \quad \quad \quad - \left(\sum_U \pi_k\right)^2 + \end{aligned}$$

SESIONES PRÁCTICAS EN R