

Técnicas de Muestreo I

Patricia Isabel Romero Mares

Departamento de Probabilidad y Estadística
IIMAS UNAM

agosto 2023

“El pensamiento estadístico será algún día tan necesario para el ciudadano competente como la habilidad de leer y escribir”.

H. G. Wells

Introducción

- En las encuestas por muestreo, el principal objetivo es **estimar** características de la población usando los datos de una **muestra**.
- Mahalanobis (1965, p45) resumió las ventajas de las encuestas por muestreo:
 - “... encuestas por muestreo a grandes escalas, cuando se realizan de la manera apropiada con un diseño muestral satisfactorio, pueden proporcionar, **rápidamente** y a un **menor costo**, información con suficiente **precisión** para fines prácticos y con la posibilidad de **evaluar el margen de incertidumbre** con una base objetiva”.

Mahalanobis, P.C.(1965). Statistics as a key technology. *The American Statistician*, 19, 43-46.

Introducción

- ¿qué es una muestra?

Es una parte de una población de interés. Un **subconjunto** de ésta.

- ¿qué es la población de interés?

Es un conjunto **finito** de objetos (elementos o unidades muestrales) identificables con ubicación en **tiempo y espacio**.

- muestreo en la vida diaria

Utilizamos muestreo, por ejemplo, al cocinar, al comprar, al comer.

- objetivos del muestreo

Las técnicas del muestreo se utilizan para conocer las características generales de la población de interés, al estudiar solo una parte de ésta.

Introducción

- ¿dónde se usa?
 - Encuestas de opinión
 - Ratings de televisión
 - Industria. Control de calidad
 - Laboratorios. Estudios en sangre
 - Encuestas electorales
 - Encuestas de INEGI. (Ingreso-Gasto, Empleo, Turismo, etc.)
 - Estudios de mercado

Introducción

- ¿por qué utilizar la información de una muestra?
 - menor **Costo**
 - mayor **Confiabilidad** en la información recabada
 - si tenemos pruebas destructivas, queremos destruir pocos elementos
 - mayor **Rapidez** en reunir la información

Introducción

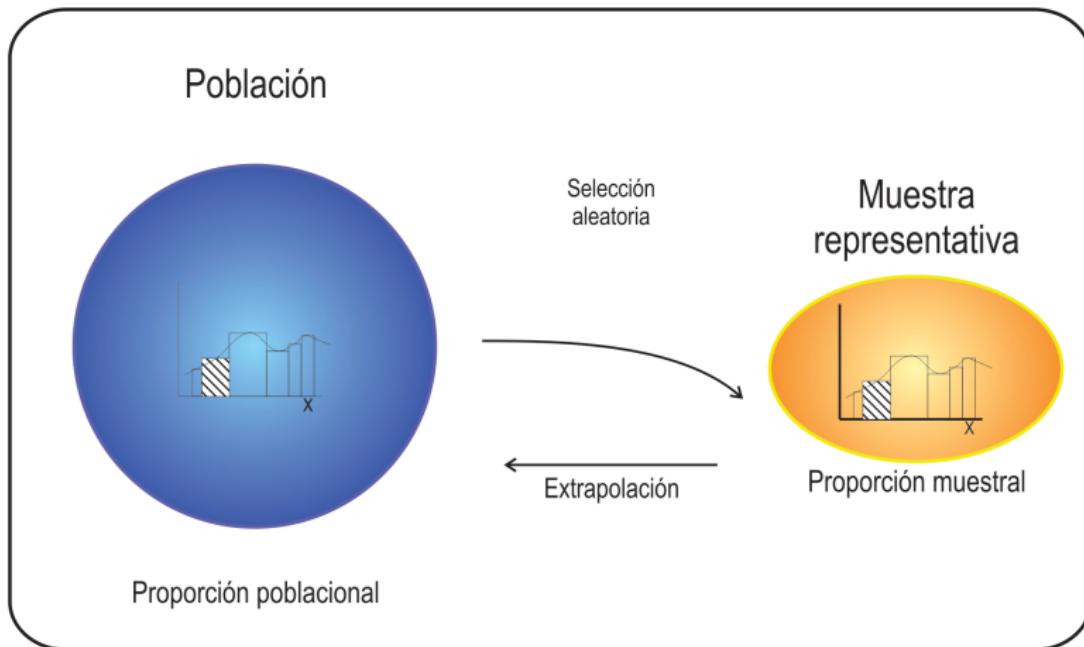
- Objetivos del muestreo.

Seleccionar “**buenas**” muestras de un tamaño “**apropiado**”, considerando la información que tenemos de la población que estamos estudiando y el presupuesto con que contamos.

- ¿qué es una “buena” muestra?

Es una **muestra representativa** de la población, es decir, que las variables de interés en la muestra presenten una distribución semejante a las de la población.

Introducción



Introducción

- ¿qué es una tamaño de muestra “apropiado”?

Depende de:

- la **variabilidad** que tiene, en la población, la característica que queremos estudiar
- la **precisión** con que queremos hacer la inferencia
- el **presupuesto** con que se cuenta
- el **tamaño** de la población

Definición de conceptos

Población Objetivo. Conjunto de elementos identificables con ubicación en tiempo y espacio. La población se define al especificar qué elementos son (a veces también cuáles no son) y qué características deben tener.

Ejemplo de una población no completamente especificada, ¿qué le faltaría?

- personas mayores de 18 años que han vivido los últimos 6 meses en la Ciudad de México

Los **elementos** de la población pueden ser personas, familias, hospitales, etc.

Definición de conceptos

Población muestreada. Es la población de donde se extrae la muestra.

En una encuesta ideal la población muestreada será idéntica a la población objetivo.

Unidad de muestreo. Es la unidad donde realizamos la muestra, la que se selecciona.

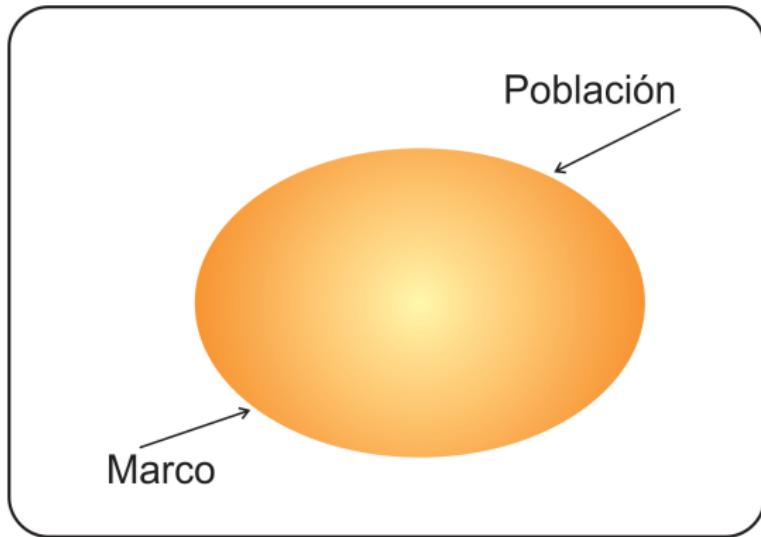
Unidad de observación. Es el objeto (elemento) sobre el cual se realiza la medición.

Muchas veces son iguales la unidad de muestreo y la unidad de observación.

Definición de conceptos

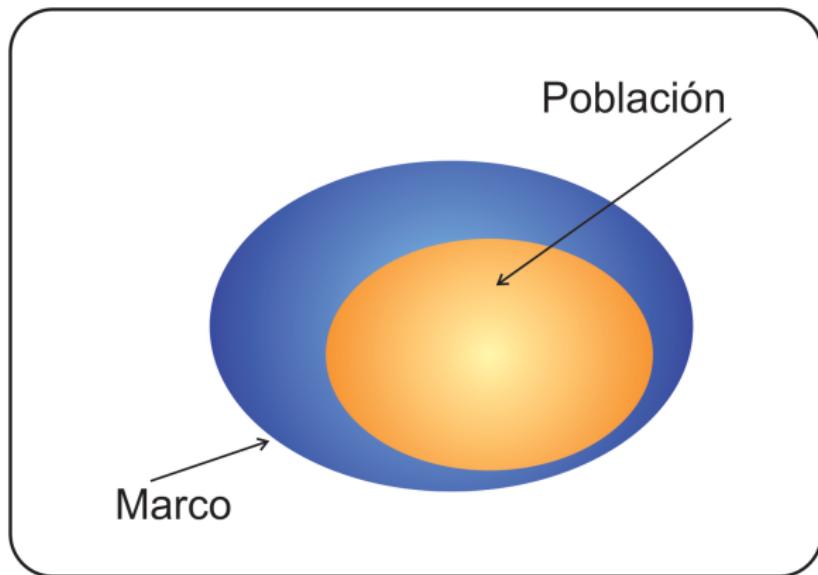
Marco de muestreo. Es el medio físico que identifica a las unidades de muestreo de la población.

En la figura la población objetivo es igual a la población muestreada.



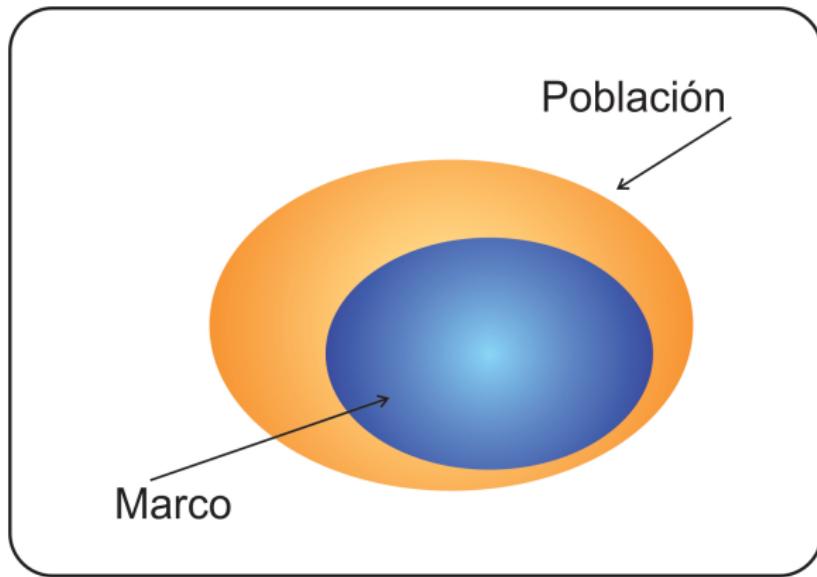
Marco de muestreo

En este caso se desechan las unidades que no son parte de la población.



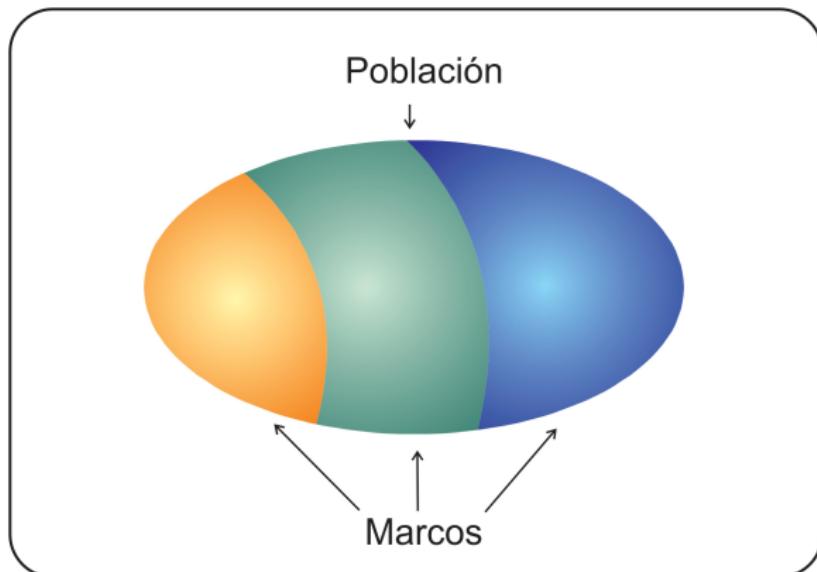
Marco de muestreo

No se puede usar este marco. Se puede redefinir la población a que coincide con el marco o complementar el marco con otro(s).



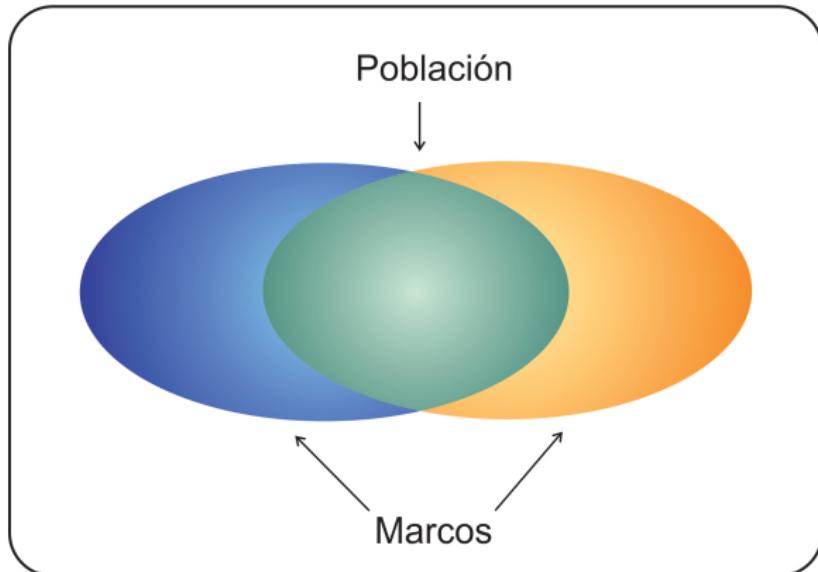
Marco de muestreo

Obliga a usar muestreo con estratos.



Marco de muestreo

Eliminar las unidades que se repiten en alguno de los dos marcos.



Definición de conceptos

Una **muestra** es un conjunto de unidades de la población seleccionadas del marco.

Las formas de tomar una muestra:

No probabilística

- **A juicio.** Se usa la experiencia del investigador para seleccionar la muestra.
- **Cuotas.** Se selecciona la muestra para que cumpla con cuotas de ciertas variables como sexo, edad.
- La muestra no probabilística puede resultar en una muestra sesgada, no representativa.
- No hay forma de estimar el error.

Probabilística

Todos los elementos de la población tienen una probabilidad conocida y mayor que cero de ser seleccionados.
(OJO. No se dice que deba ser igual probabilidad)

- Hay forma de estimar el error
- Se tiene apoyo de herramientas de probabilidad

Fuentes de error

1. Error de muestreo
2. Errores que no son de muestreo

Error de muestreo. Es el error de estimación

$$|\hat{\theta} - \theta|$$

Se controla con el diseño.

Se debe a que tenemos una muestra solamente y no **toda** la población.

Errores que no son de muestreo.

- No respuesta. Puede introducir sesgo a la estimación.
- Información falsa
 - Encuestas de salida en elecciones. Veracidad de la información.
 - Preguntas sensitivas (hay métodos). Veracidad de la información. Aumento de No respuesta.
 - Preguntas mal redactadas.
 - Términos mal definidos.
- Sustitución arbitraria de los elementos de la muestra.
Ejemplo de la leche.

Fuentes de error

Los errores que no son de muestreo se pueden controlar poniendo especial atención a la construcción del cuestionario y a los detalles en el trabajo de campo a través de una buena supervisión.

Pasos para realizar una encuesta por muestreo

1. Establecimiento de objetivos.
2. Definición de la población objetivo.
3. Construcción del Marco de muestreo.
4. Diseño de la muestra. ¿Cómo se va a seleccionar la muestra?.
5. Método de medición:
 - Entrevistas personales (entrevistador).
 - Entrevistas telefónicas.
 - Cuestionarios de autollenado.
 - Por correo (electrónico, postal).
 - Por internet.
 - Observación directa.

pasos encuesta por muestreo

6. Instrumento de medición.

Diseño del cuestionario

- Orden de las preguntas.
- Redacción de las preguntas.
- Omitir dobles negaciones.
- Preguntas sensitivas.
- ¿Preguntas abiertas o cerradas?.
- Definición de términos y conceptos (lealtad, amor, justicia).

7. Prueba piloto. Sirve para probar el cuestionario y el trabajo de campo, y estimar varianzas para el cálculo de tamaño de muestra.

pasos encuesta por muestreo

8. Organización del trabajo de campo.

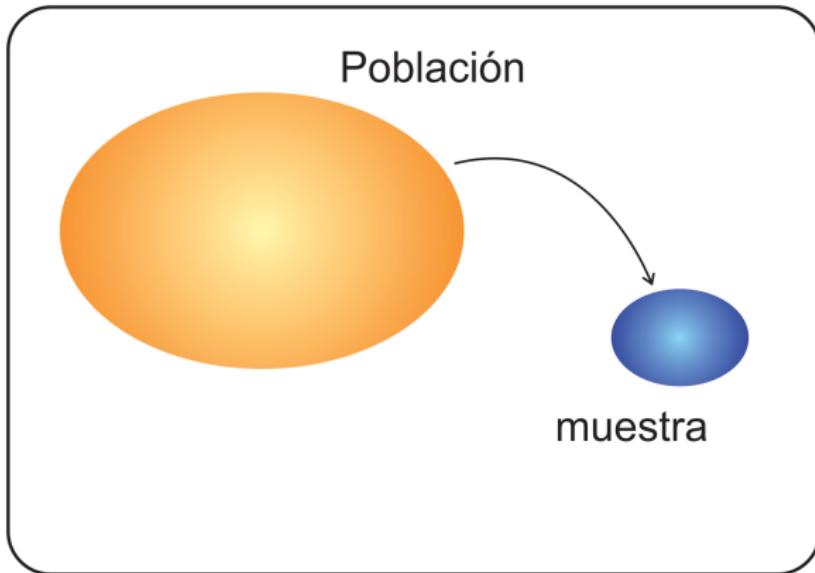
- Supervisores
- Encuestadores
- Logística

9. Organización del manejo de la información.

- ¿Qué tipo de análisis se van a hacer?
- Tablas
- Control de la calidad de la información

10. Análisis de datos y reporte final.

Objetivos del muestreo



Objetivos del muestreo

Estimar características generales de la población bajo estudio, tales como promedios, totales o porcentajes.

Esta estimación se hace con los valores observados de algunas variables en una muestra.

Valor de la variable de interés en la Población (fijas y desconocidas)

$$X_1, X_2, \dots, X_N$$

Valor de la variable de interés en la muestra (conocidas)

$$x_1, x_2, \dots, x_n$$

Otras definiciones

Estadístico(a). Es una función de la muestra que no tiene involucrados parámetros desconocidos.

Estimador. Es un estadístico que se construye para estimar un parámetro de la población (su valor varía de muestra a muestra).

Estimación. Es el valor que toma el estimador una vez observados los valores de la muestra.

Distribución muestral. Es la función de distribución de un estimador.

Ejemplo

Se tiene una población de 6 personas a las cuales se les mide cierta característica Y .

U_i	U_1	U_2	U_3	U_4	U_5	U_6
Y_i	A	B	C	D	E	F
	0	1	2	3	4	5

El promedio de la característica en toda la población es

$$\bar{Y} = \frac{15}{6} = 2.5$$

Ejemplo

Suponga que con una muestra de tamaño 2 se desea estimar este promedio. Se selecciona esta muestra aleatoria de tal manera que cualquier muestra de tamaño 2 tenga la misma probabilidad de ser seleccionada.

Cuántas muestras posibles hay?

$$\binom{6}{2} = \frac{6!}{2!4!} = \frac{30}{2} = 15$$

ejemplo

15 muestras posibles

muestra	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
elementos	A	A	A	A	A	B	B	B	B	C	C	C	D	D	E
	B	C	D	E	F	C	D	E	F	D	E	F	E	F	F
valores	0	0	0	0	0	1	1	1	1	2	2	2	3	3	4
	1	2	3	4	5	2	3	4	5	3	4	5	4	5	5
\bar{y}	0.5	1	1.5	2	2.5	1.5	2	2.5	3	2.5	3	3.5	3.5	4	4.5

ejemplo

El procedimiento de selección implica que cualquiera de estas muestras tiene la misma probabilidad de ser seleccionada, es decir, no se favorece la selección de unas de estas muestras sobre otras.

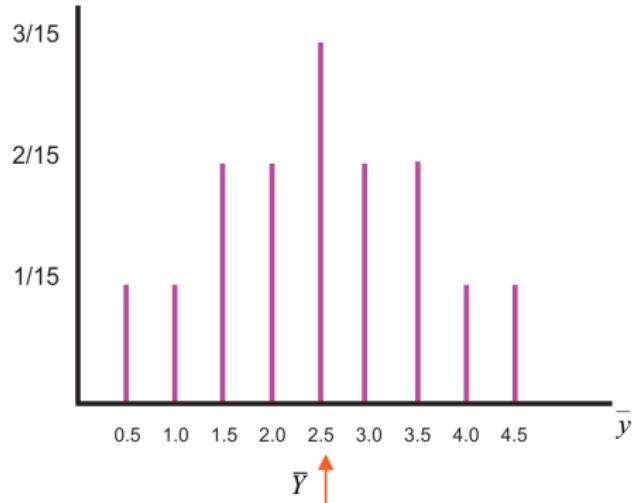
$$P(\text{ cualquier muestra }) = \frac{1}{15}$$

$$P(\text{ A en muestra }) = \frac{5}{15} = \frac{1}{3} = P(\text{ B en muestra }) = \text{etc.}$$

Distribución muestral

valor del promedio muestral	frecuencia (No. de muestras con este promedio)	frecuencia relativa
0.5	1	$\frac{1}{15}$
1	1	$\frac{1}{15}$
1.5	2	$\frac{2}{15}$
2	2	$\frac{2}{15}$
2.5	3	$\frac{3}{15}$
3	2	$\frac{2}{15}$
3.5	2	$\frac{2}{15}$
4	1	$\frac{1}{15}$
4.5	1	$\frac{1}{15}$

Ejemplo de distribución muestral



Propiedades deseables de un estimador

Como vimos con la función de distribución muestral del estimador "promedio muestral", los valores que puede tomar varían de muestra a muestra.

Una propiedad deseable de este estimador es que la esperanza del estimador sea el parámetro, en otras palabras que sea un estimador **insesgado**.

Definición de Esperanza. Sea $X \sim p_X(x)$

$$E(X) = \sum_x xp(x)$$

propiedades de un estimador

En el ejemplo:

valor de \bar{y}	probabilidad
0.5	$\frac{1}{15}$
1	$\frac{1}{15}$
1.5	$\frac{2}{15}$
2	$\frac{2}{15}$
2.5	$\frac{3}{15}$
3	$\frac{2}{15}$
3.5	$\frac{2}{15}$
4	$\frac{1}{15}$
4.5	$\frac{1}{15}$

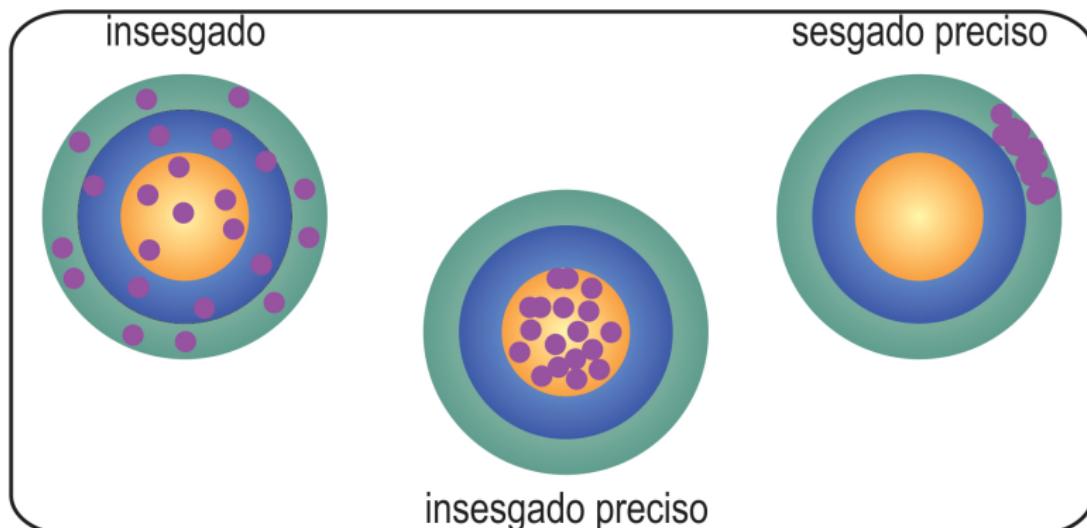
$$E(\bar{y}) = \frac{1}{15} [0.5 + 1 + 2(1.5) + 2(2) + 3(2.5) + 2(3) + 2(3.5) + 4 + 4.5]$$

$$E(\bar{y}) = \frac{1}{15} (37.5) = 2.5 = \bar{Y}$$

propiedades de un estimador

Pedir que el estimador sea insesgado no es suficiente. Otra propiedad que se pide es que tenga varianza mínima, es decir, que su distribución muestral esté muy concentrada en su media.

Gráfica tomada del libro de Sharon L. Lohr



Primera ley de los grandes números

Sean X_1, X_2, \dots, X_n $n \geq 1$ variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.), tales que $X_i = \{0, 1\}$

$$E(X_i) = p; \quad V(X_i) = p(1-p)$$

Sea $S_n = X_1 + X_2 + \dots + X_n$.

Se dice que S_n puede tomar valores $0, 1, \dots, n$ y tiene distribución binomial con media y varianza dados por:

$$E(S_n) = np; \quad V(S_n) = np(1-p).$$

Entonces,

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{S_n}{n} - p \right| < c \right] = 1 \quad \forall c > 0.$$

Segunda ley de los grandes números

Sea $X_i \ i \geq 1$, una secuencia de v.a.i.i.d. con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$.

Sea $S_n = X_1 + X_2 + \dots + X_n$ y $\bar{X} = \frac{S_n}{n}$, entonces

$$\lim_{n \rightarrow \infty} P \left[|\bar{X} - \mu| < c \right] = 1 \quad \forall c > 0.$$

Teorema Central del Límite

Sea X_i $i \geq 1$, una secuencia de v.a.i.i.d. con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$.

Sea $S_n = X_1 + X_2 + \dots + X_n$ y $\bar{X} = \frac{S_n}{n}$ y sean a y b con $a < b$ dos números cualquiera, entonces

$$\lim_{n \rightarrow \infty} P \left[a < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

Es decir, \bar{X} tiende a tener una distribución $N(\mu, \sigma^2/n)$.

Teorema Central del Límite

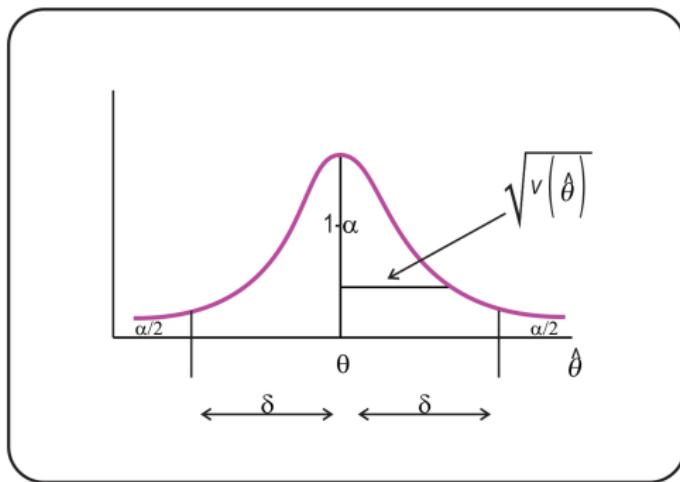
Para que se alcance una distribución parecida a la normal en el conjunto de posibles valores del promedio muestral, se requiere que n sea grande.

Sin embargo, la rapidez de acercamiento a la normal (velocidad de convergencia) también depende de la forma de la distribución de la variable de interés en la población.

Teorema Central del Límite

En general, en la población se tendrá un parámetro θ , que al tomar muchas muestras posibles con un diseño de muestra específico y una forma de estimador dada, produce muchos valores de $\hat{\theta}$.

Por el Teorema Central del Límite:



Teorema Central del Límite

$$\begin{aligned}E(\hat{\theta}) &= \theta \\V(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta})]^2 = E[\hat{\theta} - \theta]^2 \\P[\theta - \delta \leq \hat{\theta} \leq \theta + \delta] &= 1 - \alpha\end{aligned}$$

equivalente a:

$$P[|\hat{\theta} - \theta| \leq \delta] = 1 - \alpha$$

En palabras, la probabilidad de una discrepancia de a lo más δ entre θ y $\hat{\theta}$ es $1 - \alpha$.

A δ se le conoce como **precisión** del muestreo o **error de estimación**, y a $1 - \alpha$ como **confianza**.