Técnicas de Muestreo – Métodos

- Muestreo aleatorio:

  a) unidad muestral elemental:

  a.1) muestreo aleatorio simple

  a.2) muestreo (seudo)aleatorio sistemático

  a.3) muestreo aleatorio estratificado

  b) unidad muestral grupo:

  b.1) muestreo por áreas y conglomerados

  b.2) muestreo por etapas

- Muestreo no aleatorio y semialeatorio (en general, no "científico"; no estudia precisión):

  - por cuotas

  - opinático o de intención

Teorema Central del Límite

- TCL: indica que, en condiciones muy generales, la distribución de la suma de v.a. tiende a una distribución normal cuando la cantidad de variables es muy grande.

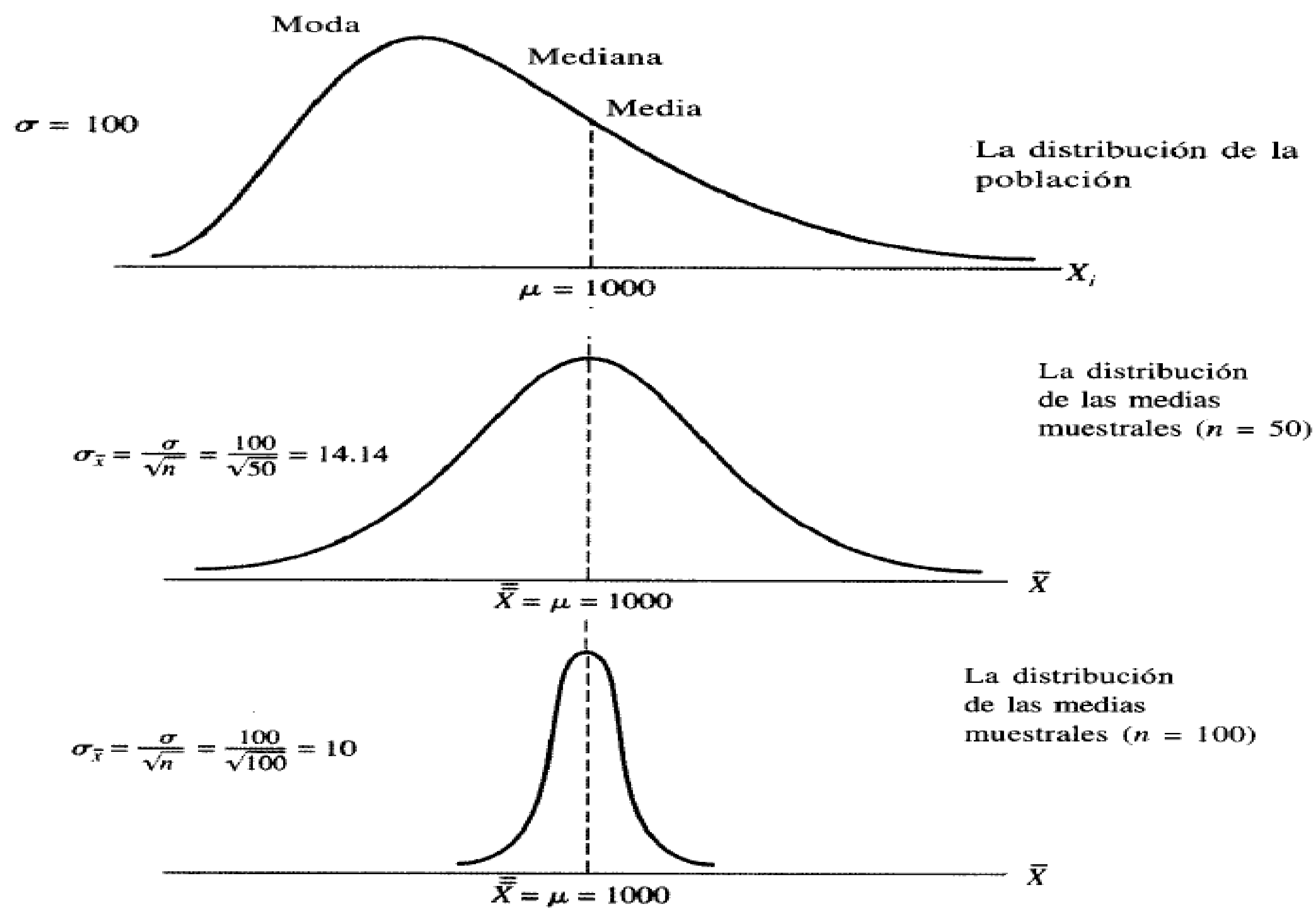
- Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con media μ y varianza σ^2. Entonces, si n grande, la v.a.:

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

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$

Teorema Central del Límite(2)

- Ej:



Teorema Central del Límite(3)

- El TCL permite el uso de la transformación Z para distribuciones muestrales:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Esta fórmula es válida cuando la desviación estándar **poblacional** es conocida.

Medidas de Tendencia Central - Media

- Si se cuentan con *n* observaciones (muestra) de una variable *X*, la *media* (aritmética) de los valores observados es:

$$\bar{X} = \frac{\sum_{i-1}^{n} X_i}{n}$$

- El parámetro que define la media poblacional (promedio real de las N observaciones de una población) es:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Medidas de Dispersión - Varianza

- La varianza es el promedio de las desviaciones de las observaciones con respecto a su media al cuadrado.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

- La desviación estándar es la raíz de la varianza. Es una medida muy útil de dispersión ya que tiene las mismas unidades que la variable estudiada.

$$\sigma = \sqrt{\sigma^2}$$

Medidas de Dispersión – Varianza(2)

- La *varianza muestral* sigue la misma lógica:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n-1} \qquad s = \sqrt{s^2}$$

- Llama la atención que se divida por *n-1*, lo que se debe a que este estadístico tiene *n-1 grados de libertad*.

- Los grados de libertad equivalen al número de observaciones menos el número de *restricciones* impuesta en tales observaciones.

Intervalos de Confianza

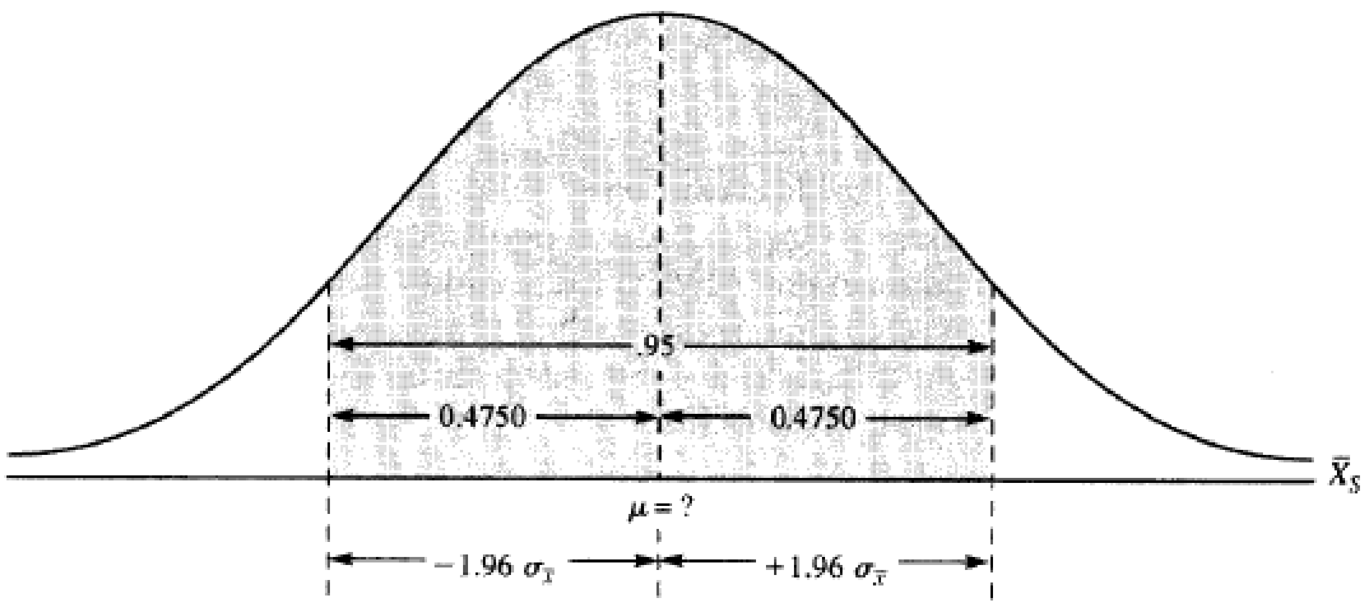
- Un intervalo de confianza denota un rango dentro del cual puede encontrarse un parámetro, y el nivel de confianza que el intervalo considera del parámetro.

- Los niveles de confianza más usados son 90%, 95%, 99%. Se denomina α a la probabilidad de que el intervalo NO contenga el parámetro desconocido.

- Si se desea construir un intervalo para la media muestral de 95%, se debe dividir el área en 2: 47.5% y luego hallar el valor de Z para esta área: Z=1.96.

Intervalos de Confianza(2)

- Este valor representa el número de desviaciones estándar que se permite para la variación de la media muestral, tanto por encima como por debajo.



Intervalos de Confianza – Muestras Grandes*

- Si σ^2 es conocido, el intervalo de confianza para la media muestral μ es $\bar{X} \pm Z\sigma_{\bar{x}}$
- Z=1.96 para un 95% de confianza
- Si σ^2 es desconocido, se utiliza la varianza muestral s^2

$$\text{I.C. para estimar } \mu = \bar{X} \pm Z s_{\bar{X}}$$

*>30 obs.

Intervalos de Confianza – Muestras pequeñas

- El TCL asegura normalidad sólo en muestras grandes

- Cuando las muestras son pequeñas se utiliza la distribución t de Student. Se deben cumplir 3 condiciones:
  - Muestra pequeña
  - σ^2 desconocido (si es conocido se usa Z)
  - Población normal o casi normal (si no se cumple se deben considerar tests no parámetricos)

Intervalos de Confianza – Muestras pequeñas(3)

- El estadístico t se calcula de manera similar a Z:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

- El intervalo de la media muestral de confianza en muestras pequeñas:

$$\text{I.C. para estimar } \mu = \bar{X} \pm (t)(s_{\bar{x}}) = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

Intervalos de Confianza – Proporción poblacional

- Las decisiones dependen con frecuencia de parámetros que son binarios (ej proporción de clientes que paga sus créditos)

- La distribución de las prop. muestrales se distribuye normal en muestras grandes (*np, n(1-p)*>5) con media igual p y desv. estandar:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

- El I.C. para estimar $\pi = p \pm Z s_p$

Población finita

- Cuando la población es finita, se corrige la desviación estándar por un factor de corrección por poblaciones finitas (fpc) .

- La nueva desviación estándar muestral queda entonces:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad s_p = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

- El efecto del FPC es que el error se va a cero cuando la muestra de tamaño *n* es igual a la población N.

Determinación tamaño de muestra

- El error del muestreo se asocia al intervalo de confianza:

$$e = Z\frac{\sigma}{\sqrt{n}} \qquad e = Z\sqrt{\frac{p(1-p)}{n}}$$

- Asumiendo un error de muestreo máximo permitido e, podemos despejar *n* en función de este error:

$$n_0 = \frac{Z^2\sigma^2}{e^2} \qquad n_0 = \frac{Z^2 p(1-p)}{e^2}$$

Determinación tamaño de muestra(2)

- Para poblaciones finitas el error muestral es:

$$e = Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \qquad e = Z \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

- Se puede determinar el tamaño muestral en función de N y de n_o (tamaño de muestra para población infinita):

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

Varianza Muestral - Población finita

- El estimador insesgado de la varianza muestral cuando se tiene una muestra finita es:

$$\hat{s}^2 \frac{N-1}{N}$$

- Por lo tanto: $\hat{\sigma}_{\bar{x}}^2 = \frac{\hat{s}^2}{n} \frac{N-n}{N}$

- Cuando *n* es suficientemente grande se considerará la normal, en caso contrario la t -Student

Varianza Muestral - Población finita

- La misma corrección puede hacerse para el caso de las proporciones muestrales:

- $\hat{p} = \dfrac{x}{n}$, con x= obs. que presentan la característica en *n*

- Intervalos de confianza: $p \in (\hat{p} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}})$

$$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n}\frac{N-n}{N}$$

Estimación del total

- A veces se desea estimar el total poblacional Nμ en vez de la media.

- En este caso: $Var(N\overline{x}) = N^2\sigma^2_{\overline{x}} = N^2 \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

- El intervalo es $N\mu \in (N\overline{x} \pm z_{\alpha/2} N\hat{\sigma}_{\overline{x}})$

- Cuando la varianza es desconocida:

$$N^2\hat{\sigma}^2_{\overline{x}} = N^2 \frac{\hat{s}^2}{n}\frac{N-n}{N} = \frac{\hat{s}^2}{n} N(N-n)$$

Muestreo Estratificado

- Se considera cada estrato i como un MAS independiente:

Población dividida en *k* estratos:	$N_1 + N_2 + \dots + N_k = N$
Tamaños muestrales de los estratos:	$n_1 + n_2 + \dots + n_k = n$
Medias poblacionales en los estratos:	$\mu_1 \quad \mu_2 \ \dots\dots \ \mu_k$
Medias muestrales en los estratos:	$\bar{x}_1 \quad \bar{x}_2 \ \dots\dots \ \bar{x}_k$

Muestreo No Aleatorio

- Muestreo por Cuotas: En este tipo de muestreo se fijan unas "cuotas" que consisten en un número de individuos que reúnen unas determinadas condiciones (variables de control)
- por ejemplo: 20 individuos de 25 a 40 años, de sexo femenino y residentes en Santiago. Siempre que se ajuste a las cuotas fijadas, el entrevistador tiene libertad para elegir a los entrevistados.
- Este método se utiliza mucho en las encuestas de opinión.

Muestreo No Aleatorio(2)

- Muestreo de juicio o de opinión: Este tipo de muestreo se caracteriza por un esfuerzo deliberado de obtener muestras "representativas“ mediante la inclusión en la muestra de grupos supuestamente típicos.

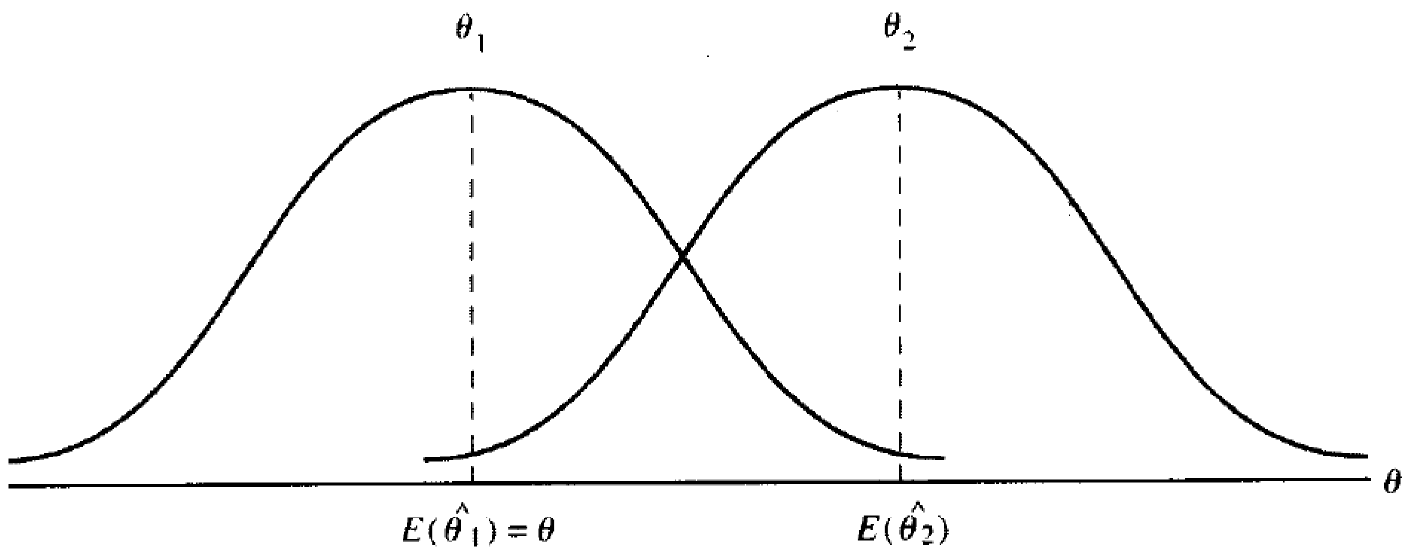
- Por ejemplo, cuando el interés del estudio se centra en comparar las características diferenciadoras de los clientes que han presentado cierta insatisfacción en el servicio respecto a aquellos otros que no, se determina a juicio del investigador un número predeterminado de encuestas a los clientes satisfechos y otro número de encuestas a clientes insatisfechos.

Propiedades de un buen estimador

- **Insesgadez**: un estimador es insesgado si la media de su distribución muestral es igual al parámetro en estudio.

- Sea $\hat{\theta}$ un estimador de θ . $\hat{\theta}$ es insesgado ssi:

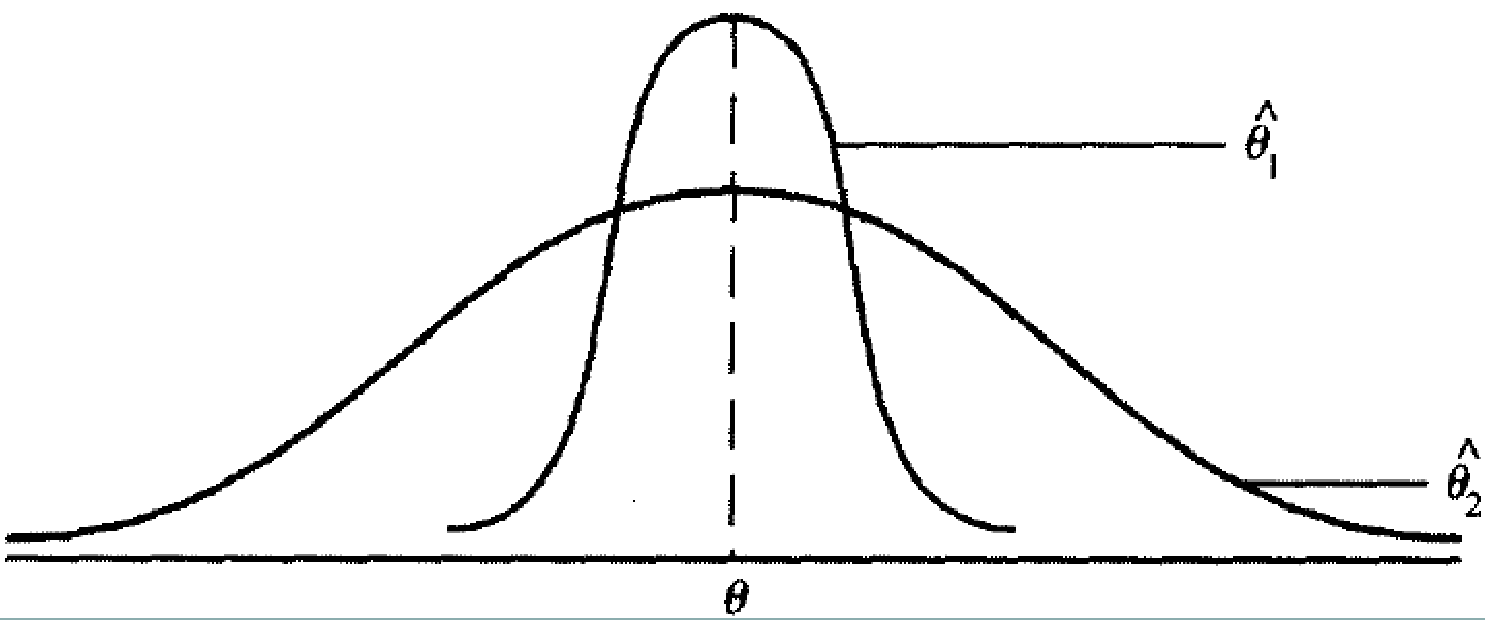
$$E(\hat{\theta}) = \theta$$



Propiedades de un buen estimador(2)

- **Eficiencia**: de todo estimador insesgado, el estimador más eficiente es aquel que tenga menor varianza:

- Sean 2 estimadores insesgados ($E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$)



Propiedades de un buen estimador(2)

- **Consistencia**: un estimador es consistente si, a medida que *n* aumenta, el valor del estadístico se aproxima al parámetro

- Para que un estimador sea consistente debe ser insesgado y su varianza debe aproximarse a cero a medida que la muestra aumente.

- Por ejemplo, la varianza de la media muestral $\sigma_{\bar{X}}^2 = \sigma^2/n$ se aproxima a cero cuando *n* aumenta, y como $E(\bar{X}) = \mu$ entonces $\bar{X}$ es un estimador consistente.