Análisis Cuantitativo I

Distribuciones Muestrales e Intervalos de Confianza

Carlos Cardona Andrade

Universidad del Rosario

30 de septiembre

Cuartiles

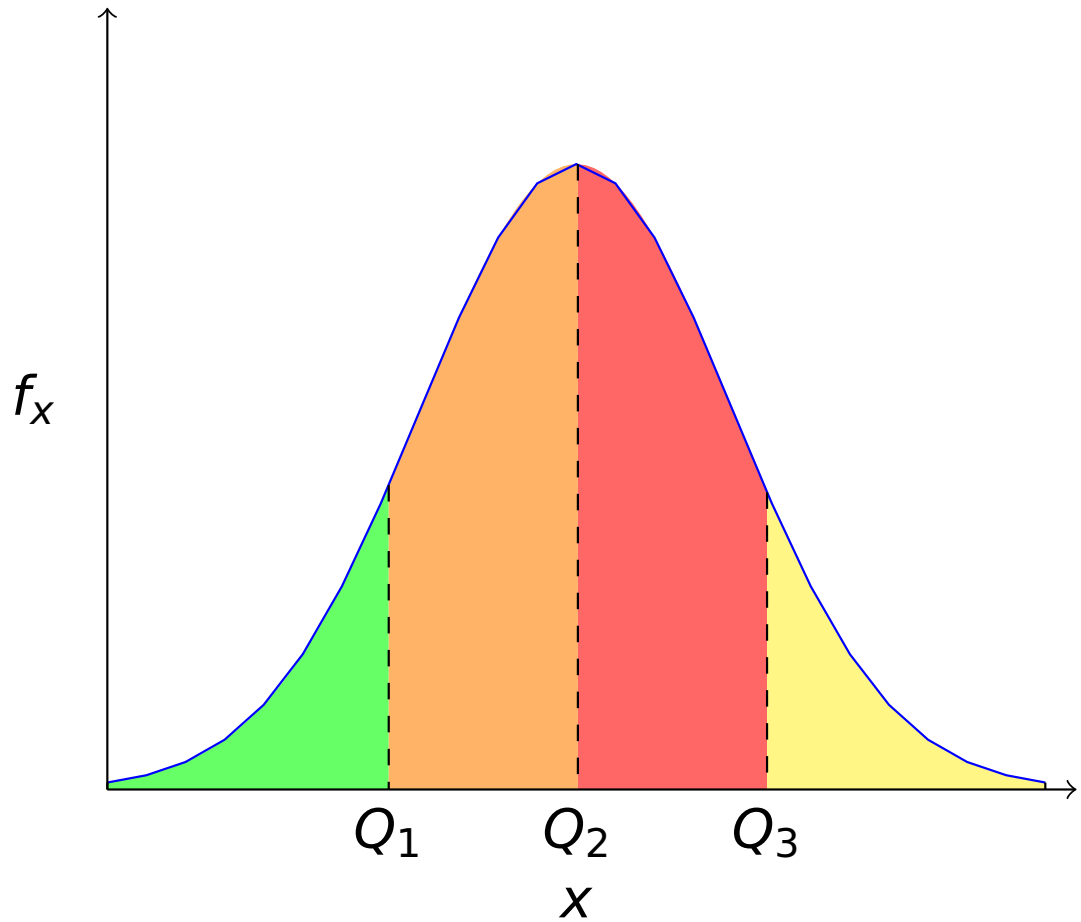
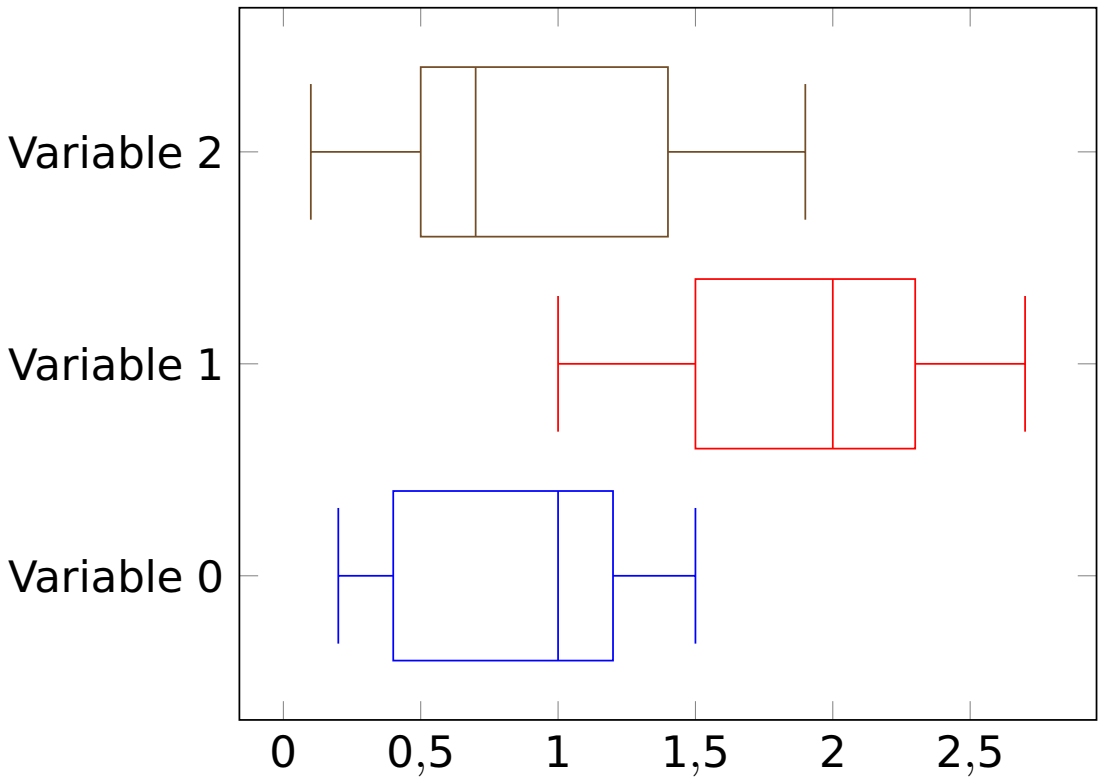


Diagrama de Caja



$$\text{Rango Intercuartílico (RIC)} = Q_3 - Q_1$$
$$\text{Límite Inferior} = Q_1 - 1{,}5 * RIC$$
$$\text{Límite Superior} = Q_3 + 1{,}5 * RIC$$

Muestras y Poblaciones

- En clases anteriores se introdujo el concepto de valor z (*z-score*) y probabilidad.
- Para cualquier valor seleccionado de una población, es posible hallar el *z*-score. El cual describe la ubicación del valor dentro de la distribución.
- Si la distribución es normal, también es posible determinar la probabilidad (proporción) de obtener cualquier valor individual.
- Ambos conceptos están limitados (hasta ahora!) a situaciones en las cuales la muestra es un valor único $n = 1$.

- La mayoría de investigaciones involucran muestras mucho más grandes, como $n = 30$ estudiantes de esta clase o $n = 50000$ mujeres que contestan la ENDS.
- En estas situaciones, la media muestral, en lugar de ser un único valor, se utiliza para responder preguntas sobre una población.
- La idea de esta sesión es expandir los conceptos de z-score y probabilidad para cubrir escenarios con muestras más grandes.
- En otras palabras, la idea es comparar muestras y calcular qué tan probable es obtener cierta muestra.

Error de Muestreo

- Como ya hemos mencionado, las muestras no son una caracterización exacta de la población.
- La edad media de los alumnos de este salón seguramente no es igual a la edad media de todos los estudiantes de la universidad.
- Esta diferencia, o *error*, entre el estadístico muestral y el parámetro poblacional se llama **error de muestreo**.
- Las muestras son variables; si tomamos dos muestras de $n = 2$, con total seguridad la media de las edades de ambas muestras serán diferentes.

Distribuciones Muestrales

- Dos muestras separadas probablemente diferirán a pesar de ser tomadas de una misma población.
- Las muestras tienen diferentes individuos, diferentes valores, diferentes medias, etc.
- En muchos casos, es posible obtener infinitas muestras de una población.
- Por ejemplo, para Colombia existen más de 10000 muestras de 2 personas dados los 45 millones de habitantes.

- Aun cuando en muchas ocasiones es imposible obtener todas las muestras posibles de una población (e.g., más de 10000 muestras de 2 personas para Colombia) , existen ciertos patrones en el comportamiento de esas muestras.
- La habilidad de predecir características muestrales está basada en la distribución muestral de medias.

Definición

La distribución muestral de medias es la colección de las medias muestrales para todas las posibles muestras aleatorias de un tamaño particular (n) que pueden ser obtenidas de una población.

- Noten que la distribución muestral de medias contienen *todas las muestras posibles*.
- Es necesario tener todos los posibles valores para calcular probabilidades.
- Por ejemplo, si el conjunto entero contiene 100 muestras, entonces la probabilidad de obtener cualquier muestra específica es 1 de 100: $p = \frac{1}{100}$.
- Cabe destacar que la distribución muestral de medias es diferente de las distribuciones que hemos visto anteriormente.
- Antes hablábamos de distribuciones de puntajes; ahora los valores en la distribución no son puntajes sino estadísticos (medias muestrales).

- Como los estadísticos son obtenidos de muestras, la distribución de estadísticos es denominada como *distribución muestral*.

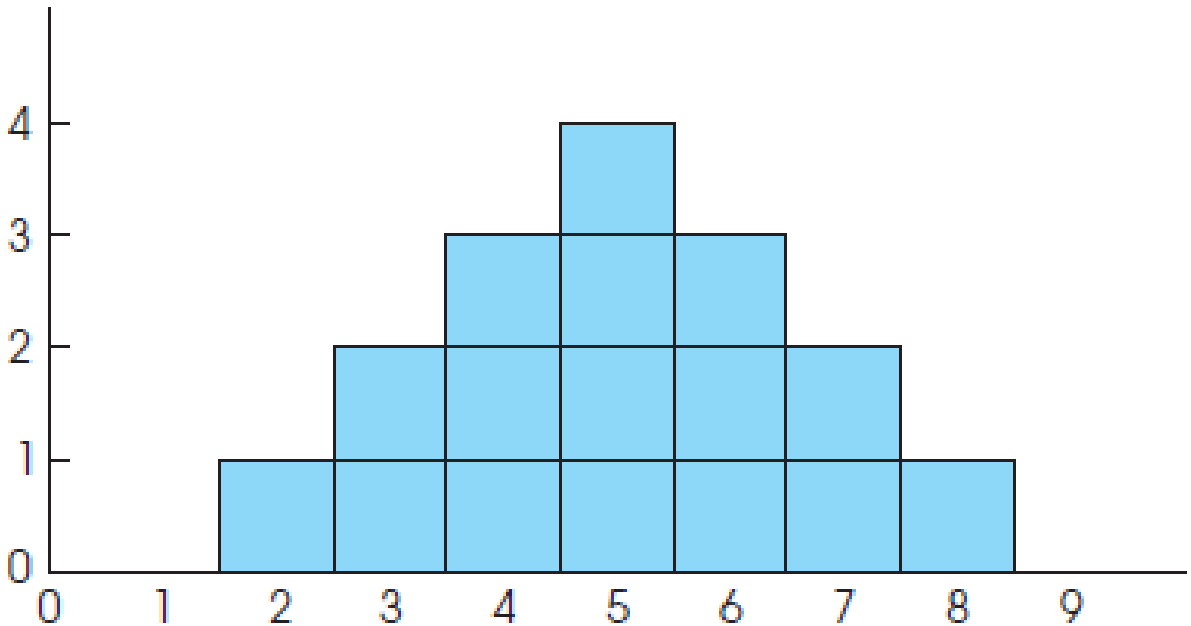
Definición

Una distribución muestral es una distribución de estadísticos obtenidos al seleccionar todas las muestras posibles de un tamaño (n) específico de una población.

- De esta manera, la distribución muestral de medias es un ejemplo de una distribución muestral.
- Consideren una población que consiste de sólo 4 valores: 2, 4, 6, 8.
- A partir de esta distribución, vamos a construir la distribución muestral de medias para $n = 2$.

Muestra	Valores		Media Muestral
	Primer	Segundo	$(\bar{X})$
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

Histograma de la Distribución Muestral



- Dos caracterízticas se destacan del histograma de la distribución muestral de medias:
  1. Las medias muestrales se mueven alrededor de la media.
  2. La distribución muestral de medias se aproxima a una curva normal.
- Finalmente, se puede usar la distribución muestral para responder preguntas de probabilidad relacionadas a las medias muestrales.
- Por ejemplo, si tomamos una muestra $n = 2$ de la población original, ¿cuál es la probabilidad de obtener una media muestral mayor a 7?

Teorema del Límite Central

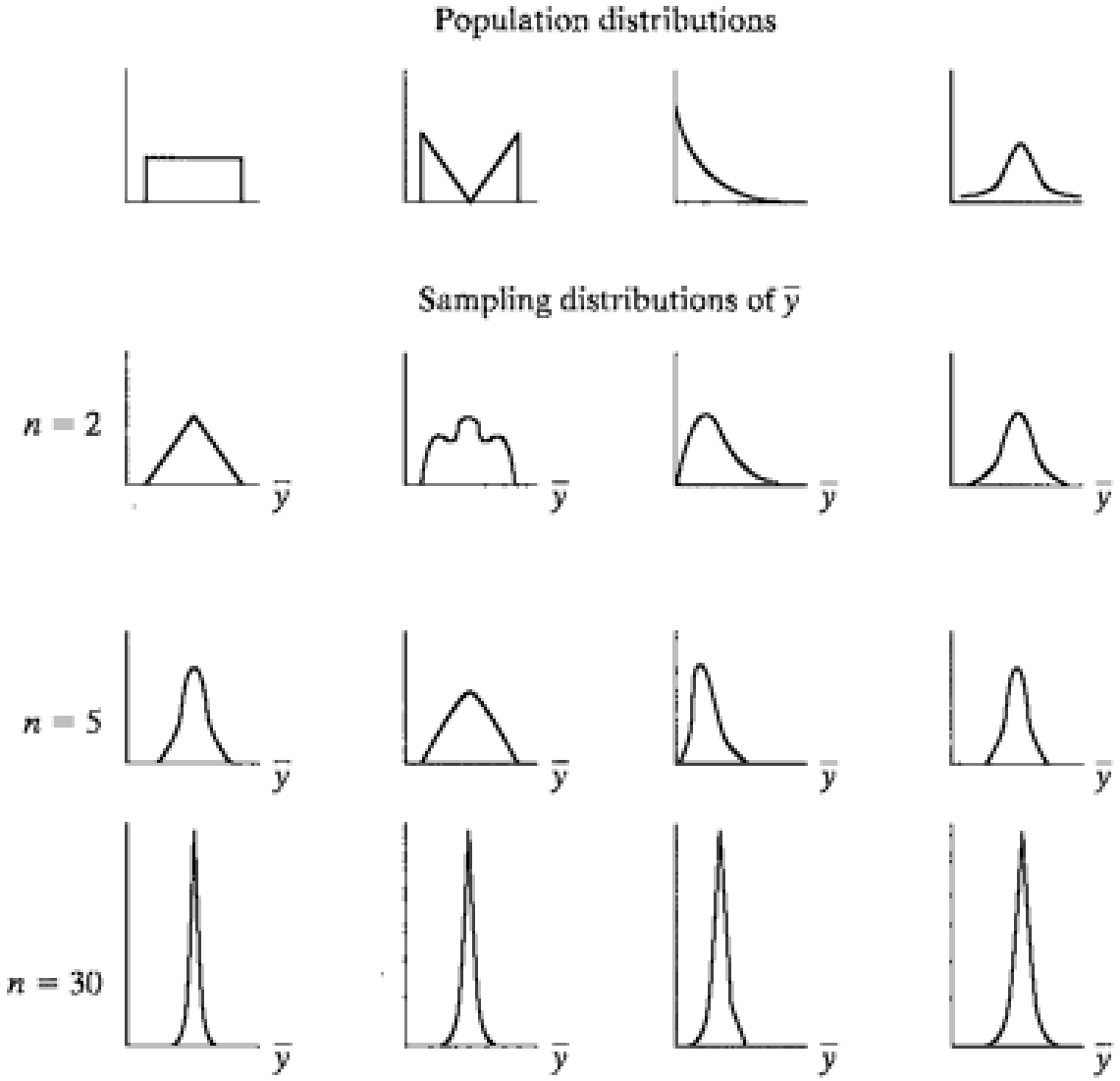
- En situaciones más reales, con poblaciones y muestras mucho más grandes, el número de muestras posibles aumenta drásticamente.
- Por lo tanto, es imposible tener cada muestra posible.
- A pesar de esto, el teorema del límite central provee una descripción precisa de la distribución resultante si se seleccionan todas las muestras posibles.

Teorema del Límite Central

Para cualquier población con media μ y desviación estándar σ, la distribución muestral de medias para un tamaño de muestra n tendrá una media igual a μ y una desviación estándar de $\frac{\sigma}{\sqrt{n}}$.

Además, se aproximará a una normal a medida que n tiende a infinito.

- El valor de este teorema recae en dos hechos:
  1. Describe la distribución muestral de medias para *cualquier población*, sin importar su forma, media o desviación estándar.
  2. La distribución muestral de medias se aproxima a una normal de manera rápida. Cuando la muestra alcanza un $n = 30$, la distribución es muy cercana a una normal.
- En resumen, el teorema del límite central identifica las tres características básicas de una distribución:
  - I Forma $\rightarrow$ Normal (Si la distribución poblacional es normal o si $n > 30$)
  - II Tendencia central $\rightarrow \mu$
  - III Dispersión $\rightarrow \dfrac{\sigma}{\sqrt{n}}$


Population distributions
Sampling distributions of $\bar{y}$
$n = 2$
$n = 5$
$n = 30$
$\bar{y}$

Error Estándar

- La desviación estándar de una distribución muestral de medias se denomina *error estándar* y se identifica con el símbolo $\sigma_{\bar{X}}$.
- El error estándar cumple los dos propósitos de una desviación estándar:
  1. Describe la distribución al decir si las medias muestrales están agrupadas o se dispersan a lo largo de un intervalo amplio.
  2. Mide qué tan bien a una media muestral representa a toda la distribución de medias.
- Por tanto, un error estándar muy grande significa que existen grandes diferencias entre las muestras.
- Dado que la media de la distribución es μ, el error estándar provee un estimado de la distancia entre una media muestral $\bar{X}$ y la media poblacional μ.

Tamaño de la Muestra

- El tamaño de la muestra influencia la precisión con la cual la muestra representa la población.
- Específicamente, una muestra grande es más precisa que una muestra pequeña.
- En general, a medida que el tamaño de muestra aumenta, el error entre la media muestral y la media poblacional decrece.

Ley de los Grandes Números

A medida que el tamaño de muestra n aumenta, es más probable que la media muestral sea cercana a la media poblacional.

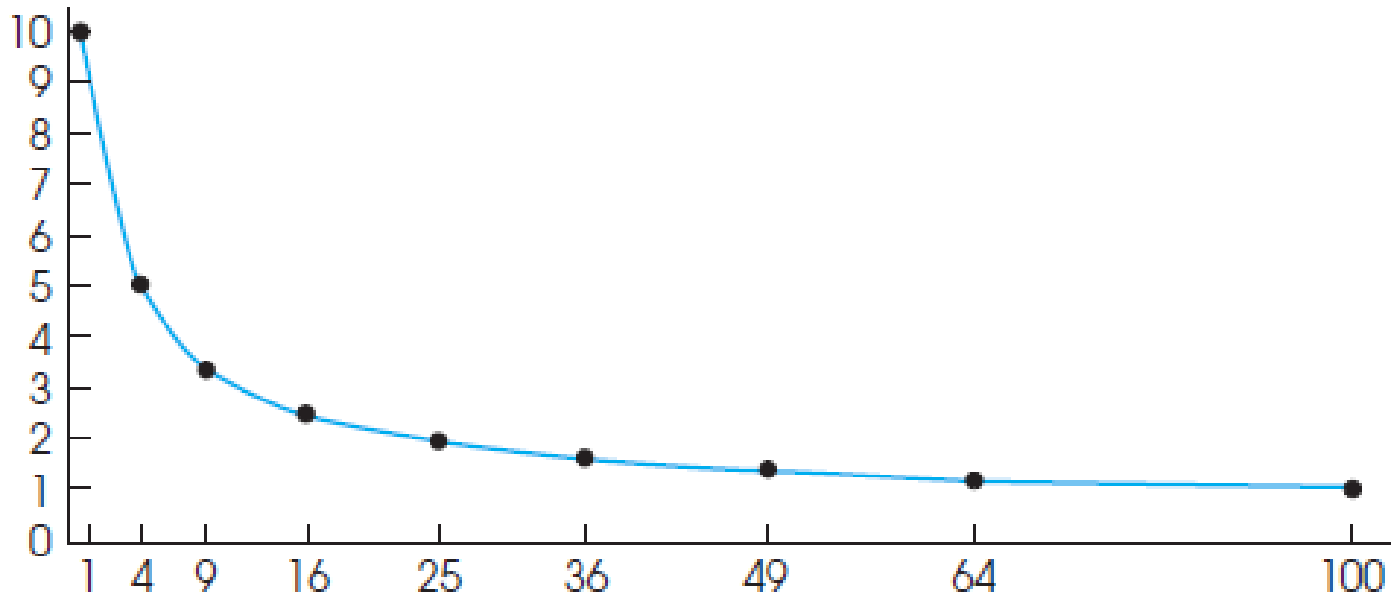
$$\bar{X} \to \mu \quad sii \quad n \to \infty$$

La Desviación Estándar Poblacional

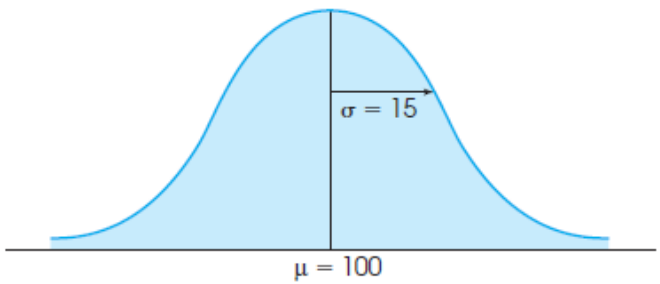
- Ya se mencionó que existe una relación inversa entre el tamaño de muestra y el error.
- El caso más extremo es la muestra que consiste de un sólo valor $n = 1$.
- En este caso, cada muestra es un valor único y la distribución muestral es idéntica a la distribución poblacional.
- La desviación estándar de la distribución muestral, la cual es el error estándar, es la misma desviación estándar de la distribución poblacional.

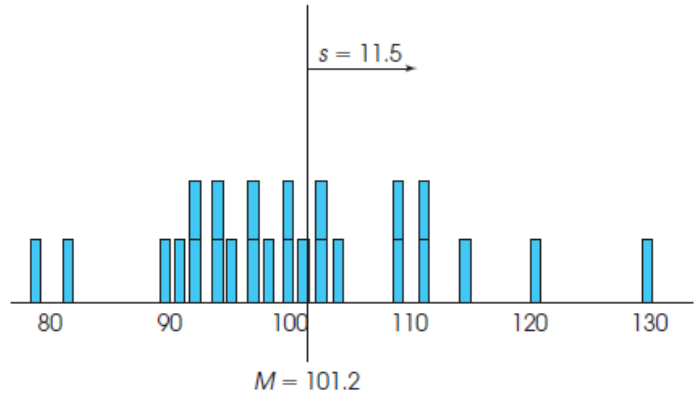
$$Si \quad n = 1 \rightarrow \sigma_{\bar{X}} = \frac{\sigma}{1} = \sigma$$

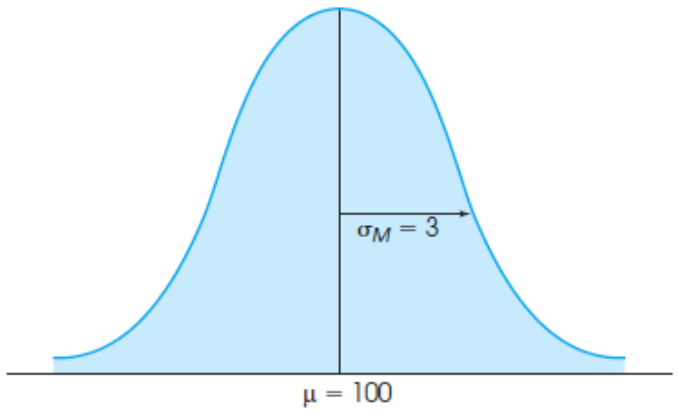
- Para una distribución poblacional con $\sigma = 10$, el error estándar varía de la siguiente manera según el tamaño de muestra:



Tres Distribuciones Distintas







Distribuciones muestrales para variables nominales

- Hasta el momento hemos considero variables continuas debido a que las variables nominales no se distribuyen normal.
- Por ejemplo, la proporción de solteros en una población o la proporción de personas que votan por un partido en particular.
- Asumamos que para el último censo, 32 % de la población colombiana es soltera. Es decir, 100-32=68 % de los colombianos no son solteros.
- Bajo esta situación, si tomamos 10000 muestras de 300 colombianos, la proporción de solteros debe estar alrededor de 0.32

- Denotemos la proporción poblacional de éxito como P_U. En este caso, $P_U = 0.32$ es la proporción de solteros.
- La letra Q_U la usamos para nombrar a la proporción poblacional de fracasar. Para el ejemplo, $Q_U = 1 - P_U = 1-0.32=0.68$ es la proporción de no ser soltero.
- El error estándar se calcula de la siguiente manera:

$$\sigma_P = \sqrt{\frac{P_U * Q_U}{n}} = \sqrt{\frac{0{,}32 * 0{,}68}{300}} = 0{,}026$$

- Por lo tanto, la distribución muestral de la proporción de solteros en Colombia, tiene una media de 0.32 y una desviación estándar de 0.026

Error de muestreo vs Error Estándar

- El concepto general de error de muestreo es que una muestra usualmente no proporciona una representación perfecta de la población.
- De hecho, 50 % de las muestras tienen medias que son más pequeñas que μ (todo el lado izquierdo de la distribución). Similarmente, 50 % de las muestras producen medias que sobreestiman la verdadera media de la población.
- Asumiendo que la distribución muestral es normal, para cada muestra se puede medir el error (o distancia) entre la media muestral y la media de la población.
- El *error estándar* proporciona una manera de medir el "promedio" de la distancia entre una media muestral y la media poblacional. Es decir, el error estándar es una manera de medir el error de muestreo.

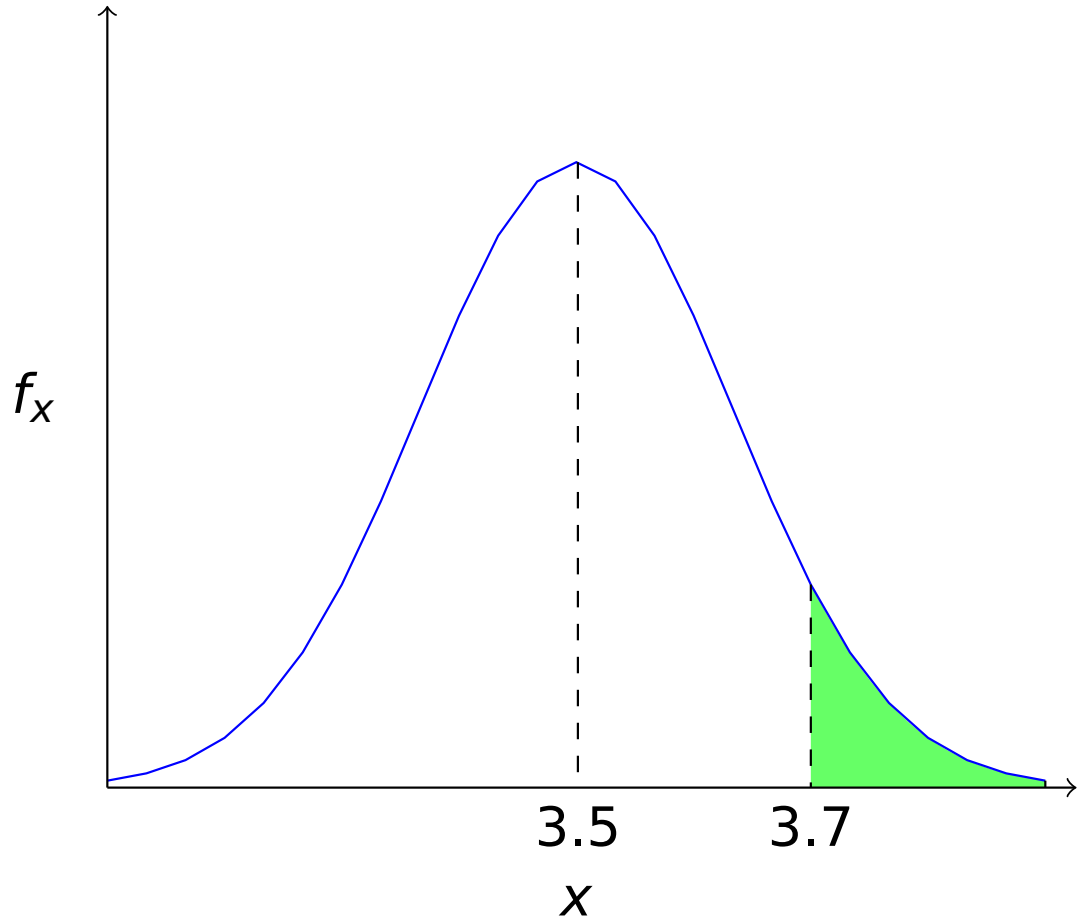
Probabilidad y Distribución Muestral

- El principal uso de la distribución muestral de medias es encontrar la probabilidad asociada a cualquier muestra específica.
- Recuerden que la probabilidad es equivalente a una proporción!
- Dado a que la distribución muestral de medias presenta el conjunto de todas las posibles medias muestrales, podemos utilizar proporciones de esta distribución para determinar las probabilidades.
- Además, gracias al teorema del límite central podemos utilizar la tabla de la distribución normal.

- Asumamos que se realiza un examen a todos los estudiantes de la universidad. La media de la distribución es $\mu = 3.5$ y la desviación estándar es $\sigma = 0.5$. La distribución de la nota del examen es normal.
- Si se toma una muestra de $n = 25$ estudiantes, ¿cuál es la probabilidad que la media muestral sea mayor a $\bar{X} = 3.7$?
- ¿Qué sabemos?
  1. La distribución muestral es normal porque la distribución del examen es normal.
  2. La distribución muestral tiene una media de 3.5 dado que la media poblacional es $\mu = 3.5$
  3. La distribución muestral tiene una error estándar $\sigma_{\bar{X}} = 0.1$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0{,}5}{\sqrt{25}} = \frac{0{,}5}{5} = 0{,}1$$

La distribución muestral de medias



$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{3{,}7 - 3{,}5}{0{,}1} = \frac{0{,}2}{0{,}1} = 2$$

- El z-score para una media muestral $\bar{X} = 3.5$ será $z = +2$.
- Por lo tanto, la probabilidad o proporción de encontrar una muestra con una media mayor a 3.7 es $p(z > 2) = 0.0228$=2.8 %.
- Por lo tanto, es posible utilizar el z-score para describir la ubicación exacta de cualquier muestra específica dentro de la distribución muestral de medias.
- Considerando la misma distribución del ejemplo anterior, ahora encontremos el rango de valores que son esperados para la media muestral el 80 % de las veces.
- Ya sabemos que la distribución es normal con una media esperada de $\mu = 3.5$ y desviación estándar $\sigma_{\bar{X}} = 0.1$.

- Nuestro objetivo es encontrar el rango de valores límites del 80 % medio de la distribución.
- Dado que la distribución es normal, podemos usar la tabla para encontrar los valores que dividen el 10 % a lado y lado de la distribución.
- En la tabla encontramos que para una proporción o probabilidad de 0.9, el z-score es 1.28
- De esta manera, los límites del 80 % medio de la distribución corresponden a $z = -1.28$ y $z = +1.28$

- Por definición, un z-score de 1.28 representa una ubicación que está a 1.28 desviaciones estándar de la media.
- Con un error estándar de 0.1, la distancia a la media es de 1.28*0.1 = 0.128
- La media es $\mu = 3.5$, por lo cual, una distancia de 0.128 en ambas direcciones produce un rango de valores entre 3.372 y 3.628
- De esta manera, 80 % de todas las posibles medias muestrales se encuentran dentro de un intervalo entre 3.372 y 3.628
- Otra interpretación es que si seleccionamos una muestra $n = 25$, estamos 80 % seguros que la media de la muestra va a encontrarse en ese intervalo.

Intervalos de Confianza

- Es claro que sólo en contadas ocasiones se tendrán los valores para la media y la desviación estándar poblacional.
- Al trabajar con una de las posibles muestras, es necesario acercarnos a los parámetros poblacionales a partir de los estadísticos muestrales que se tienen disponibles.
- Un **intervalo de confianza** es un rango de valores posibles de un parámetro expresado en un grado o nivel específico de confianza.
- Con los intervalos de confianza tomamos una estimación puntual de la muestra y la acoplamos con el conocimiento que tenemos sobre las distribuciones muestrales.
- Con los intervalos de confianza proyectamos un rango conocido y calculable de error respecto a la estimación puntual.

- Supongamos que tomamos una muestra de $n = 300$ estudiantes de la universidad con un promedio de edad de $\bar{X} = 20.5$ años y $s_X = 1.5$.
- Sabemos que la distribución muestral de medias toma forma de curva normal cuando $n > 30$.
- La edad promedio de la muestra debe estar cerca del parámetro poblacional real (la edad media de todos los estudiantes de la universidad).
- Dado lo anterior, podemos decir con seguridad que el 95 % de las muestras caen dentro de casi 2 errores estándar del parámetro real (Regla 68-95-99).
- Al calcular un intervalo de confianza, no trabajamos con muchas muestras ni calculamos áreas bajo la curva de distribución muestral.

Nivel de Confianza

- Por el contrario, trazamos una única muestra y calculamos una estimación puntual como la media.
- El **nivel de confianza** nos dice nuestra tasa de éxito, es decir, con qué frecuencia el parámetro poblacional se encuentra en el rango del intervalo de confianza.
- Usualmente, los grados de confianza más utilizados en las ciencias sociales son el 95 y 99 %.
- Al confiar en una muestra, sabemos que podemos fallar en la predicción debido a la existencia del error de muestreo.

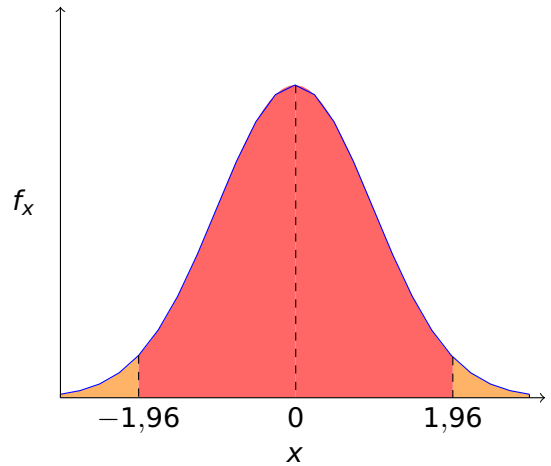
Nivel de Significancia

- La única manera de tener total certeza sobre nuestras conclusiones es reunir datos de la población.
- Cabe destacar que la cantidad de error es conocida. El nivel de error esperado es la diferencia entre el nivel de confianza y la "confianza perfecta" del 100 %.
- En otras palabras, si estamos 95 % seguros acerca de nuestro resultado, estamos 5 % inseguros acerca de este.

$$Nivel\,de\,confianza = 95\,\%$$

$$Nivel\,de\,significancia = \alpha = 100\,\% - 95\,\% = 5\,\%$$

- ¿A qué distancia del parámetro poblacional se encuentra nuestra media muestral para un nivel de confianza del 95 %?
- A partir dela tabla de la normal podemos encontrar los valores Z críticos para el nivel de significancia α (Z_α).



- El área roja representa el 95 % de las medias muestrales, mientras que, las dos áreas naranjas representan el 5 % de las medias que están fuera de los dos valores críticos.
- 95 % de las observaciones de una normal están dentro de 1.96 SD.

Margen/Término de Error

- Ya sabemos que el $Z_\alpha = 1.96$ y que $s_X = 1{,}5$. Por lo tanto:

$$\sigma_{\bar{X}} = \frac{1{,}5}{\sqrt{300}} = \frac{1{,}5}{17{,}3} = 0{,}08$$

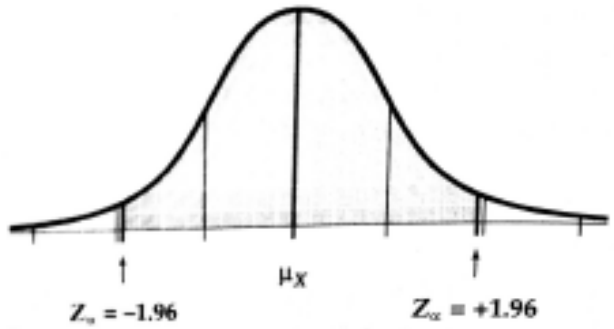
- El margen de error será igual a $Z_\alpha * \sigma_{\bar{X}} = 0{,}08 * 1{,}96 = 0{,}169$
- El intervalo de confianza estará entre:

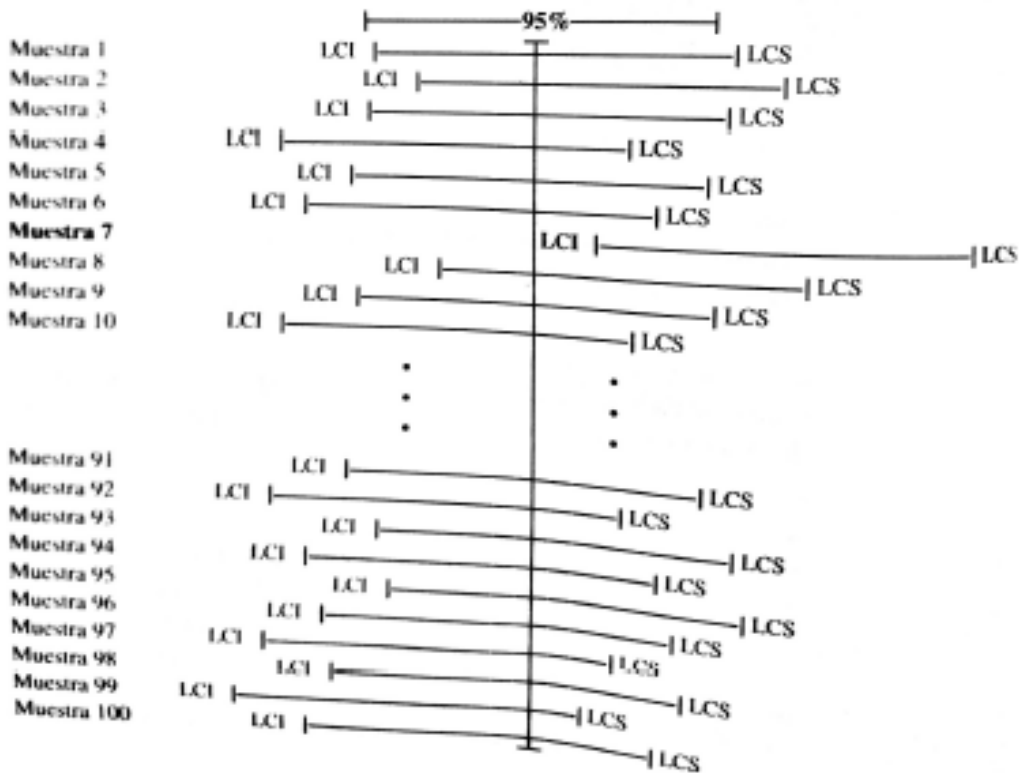
$$IC\,de\,95\,\%\,de\,\mu = \bar{X} \pm (Z_\alpha)(\sigma_{\bar{X}})$$

$$\text{Límite Inferior} = 20.5 - 0.169 = 20.3$$
$$\text{Límite Superior} = 20.5 + 0.169 = 20.6$$

¿Cuál es la interpretación?

- Estoy 95 % seguro de que la edad promedio de los estudiantes de la universidad se ubica entre 20.3 y 20.6 años.
- En otras palabras, si se realizan los mismos procedimientos muestrales 100 veces, el parámetro poblacional μ estará entre los intervalos calculados el 95 de esas veces.


μ_X
$Z_\alpha = -1.96$
$Z_\alpha = +1.96$


95%
Muestra 1
Muestra 2
Muestra 3
Muestra 4
Muestra 5
Muestra 6
Muestra 7
Muestra 8
Muestra 9
Muestra 10
Muestra 91
Muestra 92
Muestra 93
Muestra 94
Muestra 95
Muestra 96
Muestra 97
Muestra 98
Muestra 99
Muestra 100
LCI
LCS

Grado de Precisión

- Entre mayor sea el nivel de confianza estipulado, mayor será el margen de error y por lo tanto será menos preciso el intervalo de confianza.

$$Z_{0,05} = 1{,}96 \quad vs \quad Z_{0,01} = 2{,}58$$

- Entre mayor sea el tamaño de la muestra, más preciso será el intervalo de confianza.

$$\sigma_{\bar{X}} = \frac{s_X}{\sqrt{n}} \; entonces \; si \uparrow n \rightarrow \downarrow \sigma_{\bar{X}}$$