Nivelatorio: Estadística para la analítica

Presentación 3: Describir una población a partir de una muestra, mediante el uso de estimadores puntuales, intervalos de confianza y pruebas de hipótesis

Diego Antonio Bohórquez Ordóñez
dabohorquez@Icesi.edu.co

Maestría en Ciencia de Datos

Agenda

1. ¿Por qué necesitamos hacer muestreo?
2. Tipos de muestreo
3. Distribución muestral de la media y Teorema del Límite Central
4. Distribución Normal Estándar
5. Estimadores puntuales e intervalos de confianza
6. Elección del tamaño adecuado de muestra
7. Pruebas de hipótesis para la media poblacional
8. Pruebas de hipótesis para la proporción poblacional
9. Práctica en Python

¿Por qué necesitamos hacer muestreo?

¿Por qué necesitamos hacer muestreo?

- ¿De dónde vienen los datos que analizamos?
- La mejor opción para estudiar las características de una población es recoger la información de todos los elementos que la componen.
- Sin embargo, no siempre es posible tener todos los elementos de una población, por lo que se debe recurrir a muestras.
- ¿En qué situaciones debemos recurrir al uso de muestras?

¿Por qué necesitamos hacer muestreo?

1. Establecer contacto con toda la población requeriría mucho tiempo

Ejemplo: características sociodemográficas de los hogares en Colombia (GEIH)

2. Alto costo de estudiar todos los elementos de la población

Ejemplo: Intención de voto, GEIH

3. Puede que sea imposible verificar a todos los elementos de la población

Ejemplo: Nivel de contaminación de las aguas de un río

4. Naturaleza “destructiva” del ejercicio

Ejemplo: Catadores de vino, resistencia a la tensión de las placas metálicas.

¿Por qué necesitamos hacer muestreo?

- ¿Tomar una muestra permite obtener conclusiones sobre la población?
- Si el proceso de muestreo se hace correctamente, es posible contar con una muestra que representa, con la mayor precisión posible, a la población total.
- Es decir, tener información del 100% de la población no resulta ser primordial para estudiar sus características (de esto se encarga la estadística inferencial!)

Tipos de muestreo

Tipos de muestreo

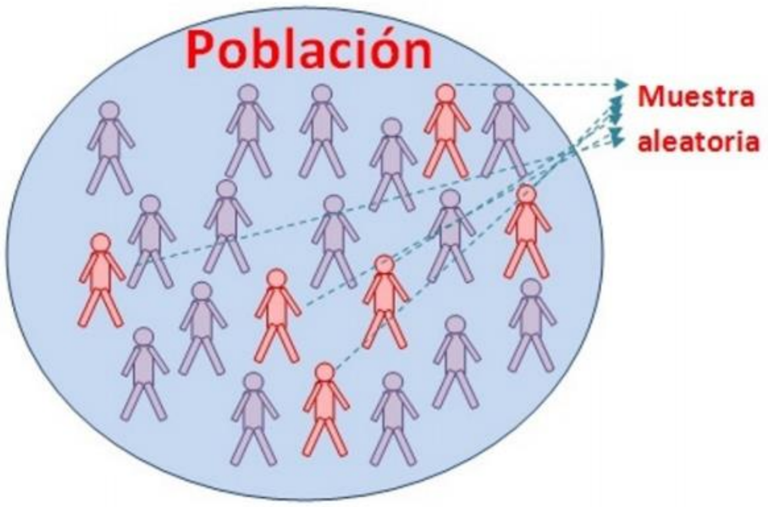
Tipos de muestreo

Muestreo no probabilístico: No se seleccionan los elementos aleatoriamente.

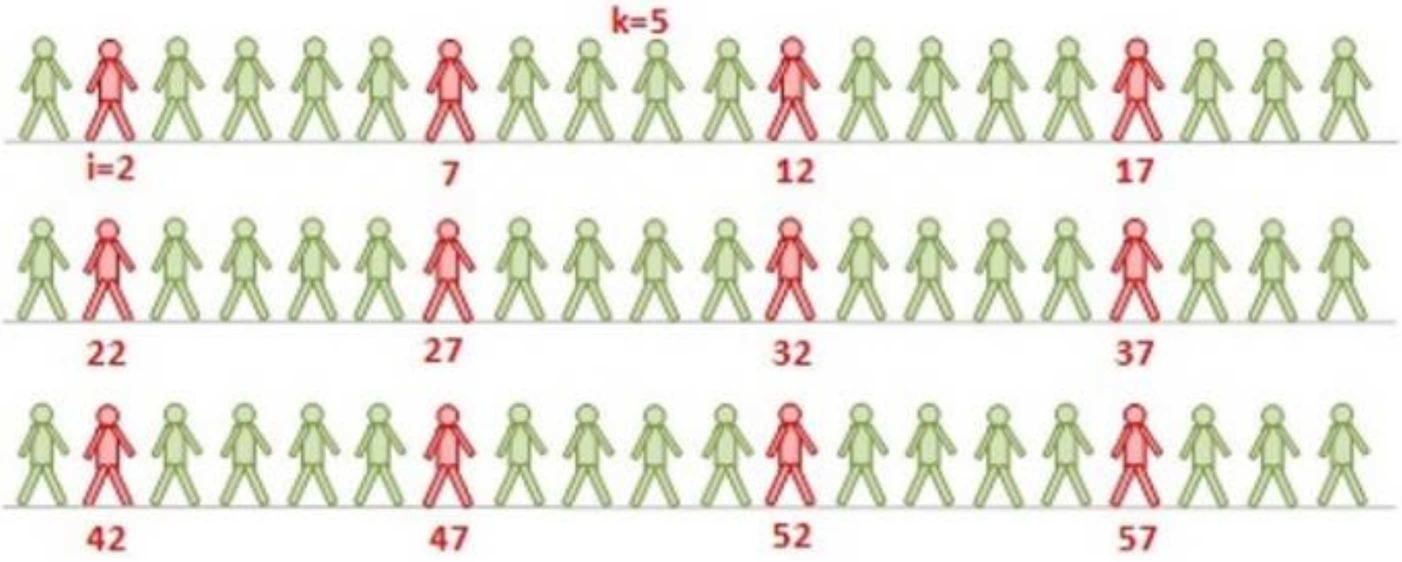
1. A juicio: El investigador, a su criterio, elige los elementos de la población.
2. Por conveniencia: El investigador elige los elementos más accesibles de la población.
3. Voluntario: Los elementos de la población son quienes deciden si suministran su información.
4. Bola de nieve: Los primeros elementos de la población localizados otorgan información para localizar otros elementos (y así sucesivamente).

Muestreo probabilístico: Se seleccionan los elementos aleatoriamente.

1. **Muestreo Aleatorio Simple**: Cada elemento tiene la mima probabilidad de ser seleccionado.



2. **Muestreo Aleatorio Sistemático**: Los elementos de la población están aleatoriamente ordenados en una secuencia. Se selecciona un punto aleatorio de inicio y se elige cada k-ésimo elemento de la población.



3. **<u>Muestreo Aleatorio Estratificado</u>**: Los elementos de la población están divididos en subgrupos, llamados *<u>estratos</u>*. Se selecciona aleatoriamente elementos de cada subgrupo, de tal forma que la cantidad seleccionada en cada subgrupo debe ser proporcional a su importancia dentro del grupo.

Los elementos dentro de los estratos son homogéneos en características, mientras que entre estratos son heterogéneos.

4. **Muestreo por Conglomerado**: Los elementos de la población están divididos en conglomerados mutuamente excluyentes. Se seleccionan aleatoriamente los conglomerados a tener en cuenta y luego se selecciona aleatoriamente elementos dentro de cada uno de esos conglomerados.

   Los elementos dentro de los conglomerados son heterogéneos en características, mientras que entre conglomerados son similares.

Distribución muestral de la media y Teorema del Límite Central

Definición clave:

- **Parámetro**: Es una medida que resume un atributo de una población.
- **Estadístico**: Es una medida que resume un atributo de una muestra.

Por tanto, el objetivo es conocer los parámetros que describen a una población, a partir de los estadísticos de la muestra. Es decir, los estadísticos se utilizan para estimar los parámetros.

<u>Distribución muestral de la media</u>: Dado que la media muestral varía de muestra en muestra, es posible construir una distribución de probabilidad de todas las posibles medias.

Ejemplo: Número de clientes atendidos diariamente en junio de 2019 por el asesor Pedro.

27	23	21
27	24	25
19	23	23
26	28	22
21	21	30
20	24	24
21	26	25
25	20	26
31	25	27
20	26	20

Media poblacional (μ): 24

M1	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*	*M9*	*M10*
27	20	20	27	25	28	23	31	20	22
27	21	26	20	25	20	20	22	26	23
27	24	30	25	23	23	25	21	20	27
21	20	21	19	26	24	27	25	24	25
23	20	26	24	27	31	19	31	22	27
23	24	20	30	27	20	21	27	23	28
22	25	27	26	23	20	24	26	26	24
30	25	30	20	24	30	22	20	26	27
28	21	24	22	26	23	19	27	31	26
21	28	25	23	25	21	20	21	25	26
24,9	22,8	24,9	23,6	25,1	24,0	22,0	25,1	24,3	25,5

Media muestral ($\bar{X}$)

Teorema del Límite Central (TLC):

- Es uno de los descubrimientos más poderosos en estadística!
- Establece que, si todas las muestras de un tamaño en particular se seleccionan de cualquier población, la distribución muestral de la media se aproxima a una distribución normal.
- A medida que el tamaño de muestra se hace más grande, dicha aproximación se hace más exacta.

Teorema del Límite Central (TLC):

- La media de la población es la media de las medias muestrales:

$$\mu_{\bar{X}} \approx \mu \text{ (Si } \underline{\text{no}} \text{ tomamos en cuenta todas las muestras posibles)}$$
$$\mu_{\bar{X}} = \mu \text{ (Si tomamos en cuenta todas las muestras posibles)}$$

- El error estándar de la media es:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- ¿Qué creen que sucede si n es muy grande?

Distribución muestral de la media y Teorema del Límite Central

Volviendo al ejemplo: Número de clientes atendidos diariamente en junio de 2019 por el asesor Pedro.

27	23	21
27	24	25
19	23	23
26	28	22
21	21	30
20	24	24
21	26	25
25	20	26
31	25	27
20	26	20

M1	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*	*M9*	*M10*
27	20	20	27	25	28	23	31	20	22
27	21	26	20	25	20	20	22	26	23
27	24	30	25	23	23	25	21	20	27
21	20	21	19	26	24	27	25	24	25
23	20	26	24	27	31	19	31	22	27
23	24	20	30	27	20	21	27	23	28
22	25	27	26	23	20	24	26	26	24
30	25	30	20	24	30	22	20	26	27
28	21	24	22	26	23	19	27	31	26
21	28	25	23	25	21	20	21	25	26
24,9	22,8	24,9	23,6	25,1	24,0	22,0	25,1	24,3	25,5

Media muestral ($\bar{X}$)

Media poblacional (μ): 24

Media de las medias muestrales ($\mu_{\bar{X}}$): 24,2

<u>Distribución muestral de proporciones</u>: La proporción muestral también varía de muestra en muestra, por lo que es posible construir una distribución de probabilidad de todas las posibles proporciones.

Ejemplo: Número de mujeres en un curso de 50 personas.

Hombre	Mujer	Mujer	Hombre	Mujer
Mujer	Mujer	Hombre	Mujer	Hombre
Hombre	Hombre	Mujer	Mujer	Hombre
Mujer	Mujer	Mujer	Hombre	Mujer
Mujer	Hombre	Hombre	Hombre	Mujer
Hombre	Mujer	Mujer	Mujer	Mujer
Hombre	Mujer	Hombre	Mujer	Hombre
Mujer	Hombre	Hombre	Mujer	Mujer
Hombre	Mujer	Mujer	Hombre	Hombre
Mujer	Hombre	Mujer	Mujer	Mujer

Proporción poblacional (P): 60%

M1	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*	*M9*	*M10*
Hombre	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Hombre	Mujer
Mujer	Hombre	Hombre	Mujer	Mujer	Hombre	Hombre	Hombre	Mujer	Mujer
Mujer	Hombre	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer
Hombre	Mujer	Mujer	Hombre	Hombre	Hombre	Mujer	Hombre	Hombre	Mujer
Hombre	Mujer	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Mujer
Mujer	Hombre	Hombre	Mujer	Hombre	Mujer	Mujer	Hombre	Hombre	Hombre
Mujer	Mujer	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Hombre	Mujer
Mujer	Mujer	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Mujer
Mujer	Mujer	Mujer	Mujer	Hombre	Hombre	Mujer	Mujer	Mujer	Hombre
Hombre	Mujer	Mujer	Hombre	Hombre	Mujer	Mujer	Mujer	Hombre	Hombre
60,00%	60,00%	60,00%	80,00%	60,00%	50,00%	70,00%	50,00%	30,00%	70,00%

Proporción muestral (p)

- El conteo de éxitos se puede convertir en proporción, si se divide por el número total de experimentos.

- Si n es lo suficientemente grande tal que se cumple que np y n(1-p) son ambos mayores que 5, entonces la binomial puede aproximarse usando la distribución normal.

- En el cálculo de probabilidades, es equivalente trabajar con una distribución normal con media np y varianza np(1-p) (es decir, conteo de éxitos), que una normal con media p y varianza p(1-p) (es decir, proporciones).

- La proporción de la población es la media de las proporciones muestrales:

$p \approx P$ (Si no tomamos en cuenta todas las muestras posibles)

$p = P$ (Si tomamos en cuenta todas las muestras posibles)

- El error estándar de la proporción es:

$$\sigma_p = \frac{\sqrt{P(1-P)}}{\sqrt{n}}$$

- ¿Qué creen que sucede si n es muy grande?

Ejemplo: Número de mujeres en un curso de 50 personas.

Hombre	Mujer	Mujer	Hombre	Mujer
Mujer	Mujer	Hombre	Mujer	Hombre
Hombre	Hombre	Mujer	Mujer	Hombre
Mujer	Mujer	Mujer	Hombre	Mujer
Mujer	Hombre	Hombre	Hombre	Mujer
Hombre	Mujer	Mujer	Mujer	Mujer
Hombre	Mujer	Hombre	Mujer	Hombre
Mujer	Hombre	Hombre	Mujer	Mujer
Hombre	Mujer	Mujer	Hombre	Hombre
Mujer	Hombre	Mujer	Mujer	Mujer

M1	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*	*M9*	*M10*
Hombre	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Hombre	Mujer
Mujer	Hombre	Hombre	Mujer	Mujer	Hombre	Hombre	Hombre	Mujer	Mujer
Mujer	Hombre	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer	Mujer
Hombre	Mujer	Mujer	Hombre	Hombre	Hombre	Mujer	Hombre	Hombre	Mujer
Hombre	Mujer	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Mujer
Mujer	Hombre	Hombre	Mujer	Hombre	Mujer	Mujer	Hombre	Hombre	Hombre
Mujer	Mujer	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Hombre	Mujer
Mujer	Mujer	Hombre	Mujer	Mujer	Mujer	Hombre	Mujer	Hombre	Mujer
Mujer	Mujer	Mujer	Mujer	Hombre	Hombre	Mujer	Mujer	Mujer	Hombre
Hombre	Mujer	Mujer	Hombre	Hombre	Mujer	Mujer	Mujer	Hombre	Hombre
60,00%	60,00%	60,00%	80,00%	60,00%	50,00%	70,00%	50,00%	30,00%	70,00%

Proporción poblacional (P): 60%

Proporción muestral (p)

Media de la proporción muestral = 59%

¿En términos prácticos que significa todo esto que acabamos de ver del TLC?

- Claramente no vamos a recoger más de una muestra para hallar la media de las muestras (ya de por sí recolectar datos de una sola es costoso…)
- La distribución muestral de la media se aproxima a una distribución normal, siendo el centro de la distribución la media poblacional.
- La media de nuestra muestra será "cercana" a la media de la población, siempre y cuando el muestreo haya sido adecuado (es muy poco probable que sean iguales, pero serán cercanas).
- Adicionalmente, entre mayor sea la muestra, menor será el error estándar de la media. Es decir, más precisa será la estimación de la media poblacional.

Distribución Normal Estándar

Distribución Normal Estándar

- Es un caso particular de la distribución normal, en el que la media es 0 y la desviación estándar 1.

- Se utiliza para calcular las probabilidades de cualquier caso de la distribución normal, ya que cualquier variable que siga dicha distribución puede ser reescalada para convertirse en una distribución normal estándar (tablas de probabilidad normal).

- ¿Cómo?

$$z = \frac{x - \mu}{\sigma}$$

- Ejemplo: Cuál es el valor z de x=8, media=5 y desviación estándar=3?

$$z = \frac{x - \mu}{\sigma} = \frac{8 - 5}{3} = 1$$

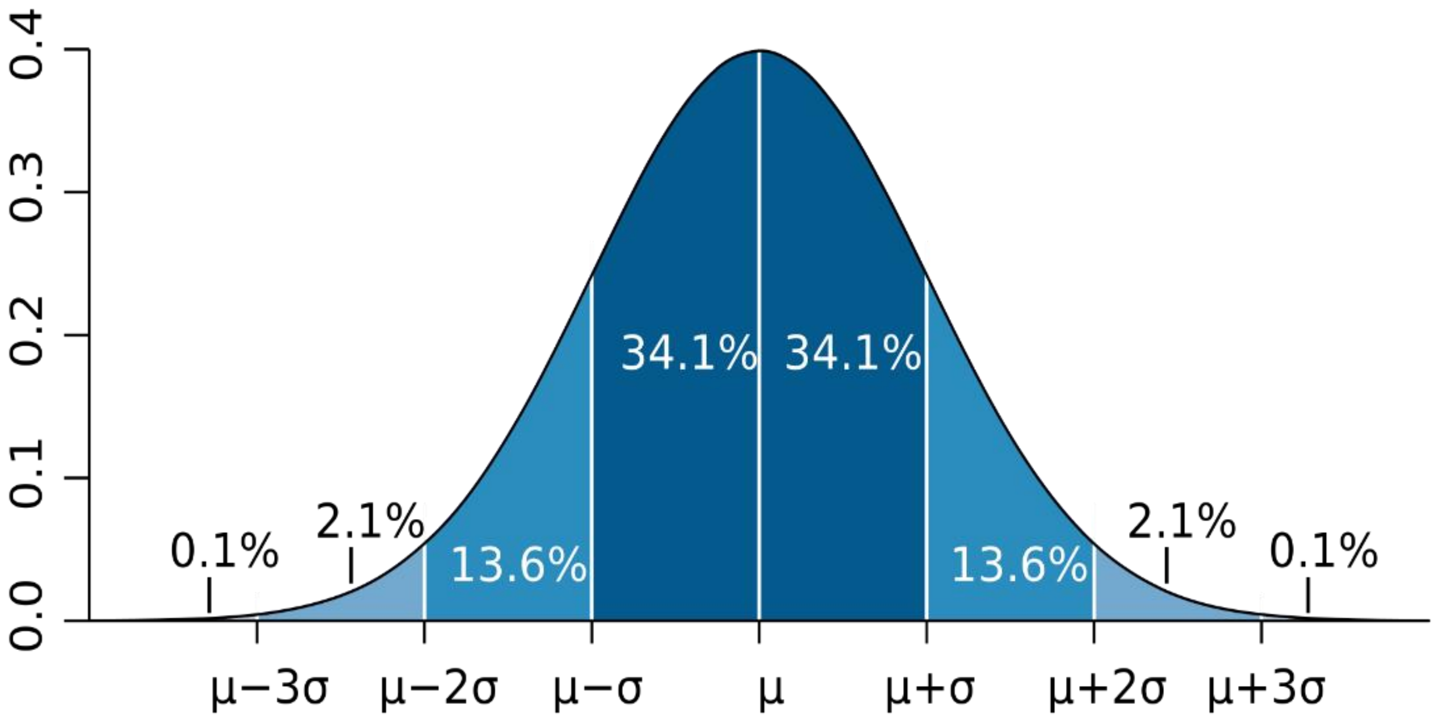
8 está a 1 desviación estándar por encima de la media.

- Ejemplo 2: Cuál es el valor z de x=4, media=5 y desviación estándar=3?

$$z = \frac{x - \mu}{\sigma} = \frac{4 - 5}{3} = -0{,}33$$

4 está a 0,33 desviaciones estándar por debajo de la media.

- Esto está relacionado con la “regla empírica” de la que hemos hablado:



Distribución Normal Estándar

- ¿Cómo se relaciona esto con lo que vimos hasta aquí?

- La distribución de la media muestral puede reescalarse para tener una distribución normal estándar con la siguiente fórmula:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- La distribución de la proporción muestral puede reescalarse para tener una distribución normal estándar con la siguiente fórmula:

$$z = \frac{p - P}{\sqrt{P(1 - P)} / \sqrt{n}}$$

Estimadores puntuales e intervalos de confianza

- Estimador puntual: Estadístico calculado a partir de la información de la muestra, para estimar el parámetro poblacional.

  - La media muestral, $\bar{X}$, es un estimador puntual de la media poblacional, μ.

  -La proporción muestral, p, es un estimador puntual de la proporción poblacional, P.

  - La desviación estándar muestral, s, es un estimador puntual de la desviación estándar poblacional, σ.

- Aunque el estimador puntual es “cercano” al parámetro poblacional, no corresponde exactamente al mismo valor. Por tal motivo, surge la necesidad de calcular el rango donde probablemente se encuentra el parámetro.

- Intervalo de confianza: Es un conjunto de valores en el que se espera, con una determinada probabilidad, que se encuentre el parámetro poblacional.

- Dicho nivel de probabilidad recibe el nombre de nivel de confianza $(1-\alpha)$.

- El intervalo de confianza se estima usando el estimador puntual y un margen de error. El margen de error depende del error estándar del estimador puntual y de la amplitud del intervalo que se quiera seleccionar.

- Ejemplo de interpretación: Con un 95% de confianza, se espera que la media de visitas a la oficina más concurrida esté entre 25 y 35.

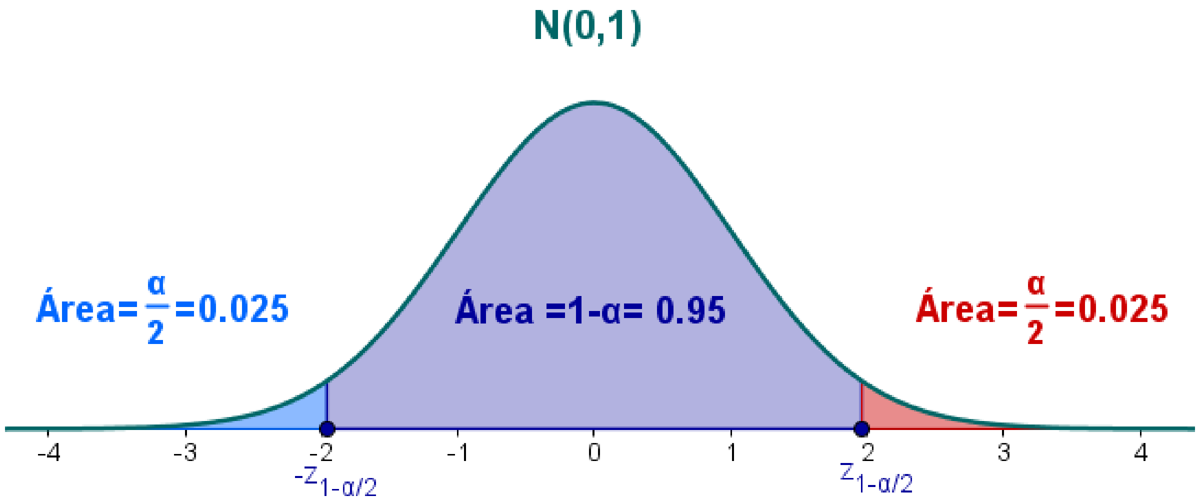
- <u>Intervalo de confianza para la media:</u> Se consideran dos casos.

1) Se conoce la desviación estándar de la población:

$$\bar{X} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} \pm 1{,}96 * \frac{\sigma}{\sqrt{n}} \text{ (IC 95\%)}$$

$$\bar{X} \pm 2{,}58 * \frac{\sigma}{\sqrt{n}} \text{ (IC 99\%)}$$



- Ejemplo: Se cuenta con una muestra aleatoria de 100 gerentes de cuenta, a quienes se les preguntó por el salario. La media muestral resultó ser $6’500.000, mientras que la desviación estándar del salario de la población es de $300.000.

- ¿Cuál es la media de la población de los gerentes?

No lo sabemos exactamente, pero podemos usar la media muestral como la mejor estimación de dicho valor.

- ¿Entre que valores se puede esperar que se encuentre la media poblacional, con un nivel de confianza del 95%?

$$\bar{X} \pm z_{0,05/2} * \frac{\sigma}{\sqrt{n}}$$

IC del 95%:

$$6'500.000 - 1,96 * \frac{300.000}{\sqrt{100}} = 6'441.200$$

$$6'500.000 + 1,96 * \frac{300.000}{\sqrt{100}} = 6'558.800$$

- Intervalo de confianza para la media: Se consideran dos casos.

2) NO se conoce la desviación estándar de la población:

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}$$

- Se usa la desviación estándar de la muestra, ya que es un estimador de la desviación estándar poblacional.

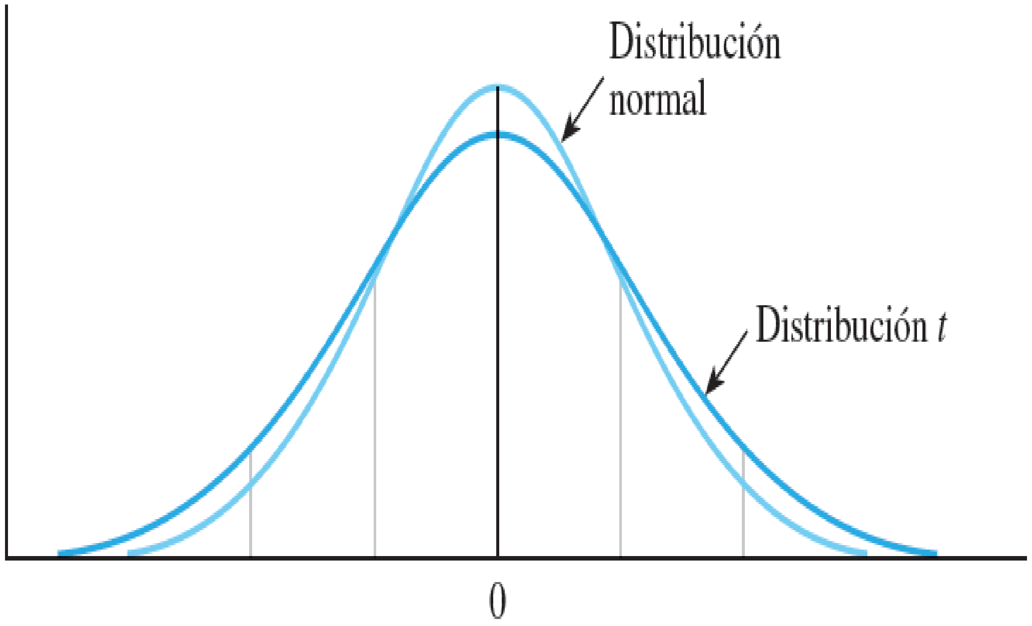
- Dado este cambio, no puede utilizarse la distribución normal, sino la distribución t de Student.

- Distribución t de Student:

- Tiene forma de campana y es simétrica.

- Las distribuciones t tienen media 0 y sus desviaciones estándares difieren de acuerdo al tamaño de la muestra.

- En términos generales, es más plana y tiene colas más pesadas que la normal estándar. Sin embargo, a medida que aumenta el tamaño de muestra, tiende hacia ella.



- Distribución t de Student:
- Se trabaja con n-1 grados de libertad para intervalos de confianza de la media.
- Grados de libertad, de manera coloquial, son la cantidad de información suministrada por la muestra que usted puede "gastar" para estimar los valores de parámetros de población desconocidos.
- Entre mayor tamaño de muestra, más información se tiene para estimar a partir de la muestra.

Valores t para diferentes grados de libertad:

GL= 5, IC 95%: $\bar{X} \pm 2{,}57 * \frac{s}{\sqrt{n}}$

GL= 10, IC 95%: $\bar{X} \pm 2{,}23 * \frac{s}{\sqrt{n}}$

GL= 20, IC 95%: $\bar{X} \pm 2{,}08 * \frac{s}{\sqrt{n}}$

GL= 100, IC 95%: $\bar{X} \pm 1{,}98 * \frac{s}{\sqrt{n}}$

GL= 500, IC 95%: $\bar{X} \pm 1{,}965 * \frac{s}{\sqrt{n}}$

Factor de corrección de una población finita:

- Cuando se cuenta con una población finita, los errores muestrales deben ser ajustados. Si la muestra es un porcentaje importante de la población, el estimador será más preciso.

$$\bar{X} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

- Intervalo de confianza para la proporción:

$$p \pm z_{\alpha/2} * \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

- Intervalo de confianza para la proporción con corrección:

$$p \pm z_{\alpha/2} * \frac{\sqrt{p(1-p)}}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

Elección del tamaño adecuado de muestra

Tamaño de muestra para estimar la media de la población:

$$\bar{X} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$e = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$n = \frac{{z_{\alpha/2}}^2 N \sigma^2}{\sigma^2 {z_{\alpha/2}}^2 + (N-1)e^2}$$

UNIVERSIDAD ICESI | A OTRO NIVEL

<u>Tamaño de muestra para estimar la proporción de la población:</u>

$$p \pm z_{\alpha/2} * \frac{\sqrt{p(1-p)}}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$e = z_{\alpha/2} * \frac{\sqrt{p(1-p)}}{\sqrt{n}} * \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

$$n = \frac{z_{\frac{\alpha}{2}}{}^{2} N p(1-p)}{p(1-p) z_{\alpha/2}{}^{2} + (N-1)e^{2}}$$

Pruebas de hipótesis para la media poblacional

Pruebas de hipótesis para la media poblacional

- ¿Qué es una hipótesis?

Afirmación relativa a un parámetro de la población sujeta a verificación.

- ¿Qué es una prueba de hipótesis?

Procedimiento basado en evidencia de muestra para determinar si una hipótesis es razonable.

1. Se establecen las hipótesis nula y alternativa
2. Se selecciona un nivel de significancia
3. Se identifica el estadístico de la prueba
4. Se formula una regla para tomar una decisión (valor crítico)
5. Se toma una muestra y se llega a una decisión

- ¿Qué es el nivel de significancia?

Probabilidad de rechazar la hipótesis nula cuando es verdadera.

- ¿Qué es el estadístico de prueba?

Es el valor calculado a partir de la muestra, para determinar si se rechaza la hipótesis nula.

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \text{o} \quad t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

- <u>¿Qué es un valor crítico?</u>
- Punto de división entre la región de rechazo y no rechazo de la hipótesis nula.
- Al definir el nivel de significancia, es posible obtener el valor crítico.
- <u>¿Cómo se toma la decisión?</u>
- Si el estadístico es más extremo que el valor crítico, entonces se rechaza la hipótesis nula.
- Si el estadístico es menos extremo que el valor crítico, entonces no hay suficiente información para rechazar la hipótesis nula.

- ¿Cómo se toma la decisión? (otra opción: el valor-p)

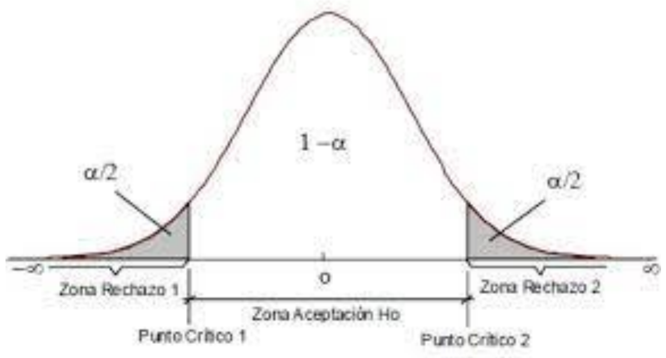
- Valor-p es la probabilidad de observar un valor muestral tan extremo o más que el observado, si la hipótesis nula es verdadera.

- Si el valor-p es menor que el nivel de significancia, entonces se rechaza la hipótesis nula.

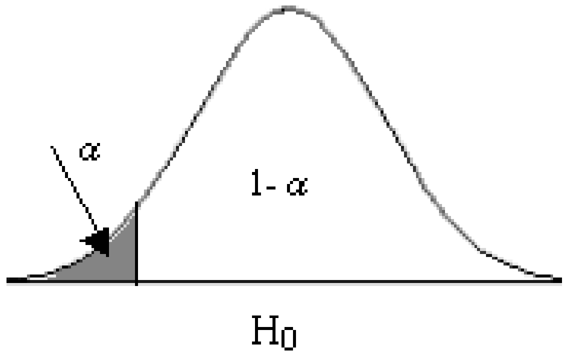
- Si el valor-p es mayor que el nivel de significancia, entonces no hay suficiente información para rechazar la hipótesis nula.

- Existen dos tipos de pruebas:

1. Prueba de dos colas: La región de rechazo está en ambas colas. Está asociada con hipótesis de igualdad (= o ≠).



2. Prueba de una cola: La región de rechazo está en una de las dos colas. Está asociado con hipótesis de desigualdad (> o <).



La clave para detectar el tipo de prueba que se requiere es observar cómo está escrita la hipótesis **alterna**.

- Ejemplo: Se recolectó una muestra de 60 llamadas, para estudiar el tiempo promedio de la llamada. La muestra arroja una media muestral de 5 minutos, con una desviación estándar muestral de 0,5 minutos. La jefa del área asegura que la media del tiempo es mayor que 4,8 minutos.

- Ho: La media del tiempo de una llamada es menor o igual a 4,8.
- Ha: la media del tiempo de una llamada es mayor que 4,8.

Estadístico de prueba:

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{5-4{,}8}{0{,}5/\sqrt{60}} = 3{,}09$$

- <u>Ejemplo:</u> Se recolectó una muestra de 60 llamadas, para estudiar el tiempo promedio de la llamada. La muestra arroja una media muestral de 5 minutos, con una desviación estándar muestral de 0,5 minutos. La jefa del área asegura que la media del tiempo es mayor que 4,8 minutos.

Nivel de significancia: 5%

$$t_{crítico} = t_{0,05;59} = 1{,}67$$

<u>Decisión</u>: Dado de que el estadístico es mayor que el valor crítico, entonces es posible rechazar la hipótesis nula.

El valor p es 0,1%, por lo que es menor que el nivel de significancia.

Pruebas de hipótesis para la proporción poblacional

Pruebas de hipótesis para la proporción poblacional

- <u>¿Cuál es el estadístico de prueba?</u>

Es el valor calculado a partir de la muestra, para determinar si se rechaza la hipótesis nula.

$$z = \frac{p - P}{\sqrt{P(1-P)} \big/ \sqrt{n}}$$

Las demás instrucciones se mantienen igual:

1. Se establecen las hipótesis nula y alternativa
2. Se selecciona un nivel de significancia
3. Se identifica el estadístico de la prueba
4. Se formula una regla para tomar una decisión (valor crítico)
5. Se toma una muestra y se llega a una decisión

En resumen…

Práctica en Python

Ejercicio práctico

Ejercicio Práctico

Se tiene una muestra de estudiantes que presentaron el Saber TYT:

1. Conociendo que el tamaño total de la población es 52712 ¿es óptimo el tamaño de muestra escogido, teniendo en cuenta que se quería tener un margen de error de 5% y un nivel de confianza del 95%? Justifique y demuestre.

2. Muestre las estadísticas descriptivas del puntaje de razonamiento cuantitativo. Grafique la distribución. Describa la distribución.

3. ¿Cuál es la estimación de la media del puntaje de razonamiento cuantitativo poblacional? Justifique.

4. Construya intervalos de confianza para la media de dicha prueba, con niveles de confianza del 90%, 95% y 99%. ¿Qué concluye sobre la media de dicha prueba?

5. Un asesor del Icfes asegura que la media del puntaje de razonamiento cuantitativo es 90. ¿Qué puede decir sobre tal afirmación?

6. Calcule la proporción de hombres y mujeres de la muestra. Grafique dichas proporciones.

7. ¿Cuál es la estimación de la proporción de mujeres poblacional?

8. Construya intervalos de confianza para la proporción de mujeres, con niveles de confianza del 90%, 95% y 99%. ¿Qué concluye sobre la proporción de mujeres?

9. Un asesor del Icfes considera que 2 de cada 3 estudiantes que presentan el Saber TYT son mujeres ¿Qué puede decir sobre tal afirmación?