Bioestadística

Sesión 1: Inferencia
Conceptos básicos Inferencia

José Aurelio Pina Romero
Ja.pina@ua.es
Bioestadística – Grado Enfermería
UA- Departamento de Enfermería

Universitat d'Alacant
Universidad de Alicante

¿Qué vamos a ver?

- Población us muestra
- Técnicas de muestreo
  - Muestreo aleatorio simple.
  - Muestreo sistemático.
  - Muestreo estratificado.
- Estimador puntual
- Teorema central del límite
- Intervalo de confianza para la media muestral

Introducción

Las poblaciones están formadas por individuos, pero sería mejor denominarlas **unidades de muestreo** o **unidades de estudio**:

- Personas, células, familias, hospitales, enfermedades,⋯

La población ideal que se pretende estudiar se denomina **población objetivo**.

- No es fácil estudiarla por completo. Aproximamos mediante muestras que den idealmente la misma probabilidad a cada individuo de ser elegido.
- Tampoco es fácil elegir **muestras** de la población objetivo:
  - Si llamamos por teléfono excluimos a los que no tienen.
  - Si elegimos indiv. en la calle, olvidamos los que están trabajando...

- El grupo que en realidad podemos estudiar se denomina **población de estudio**.

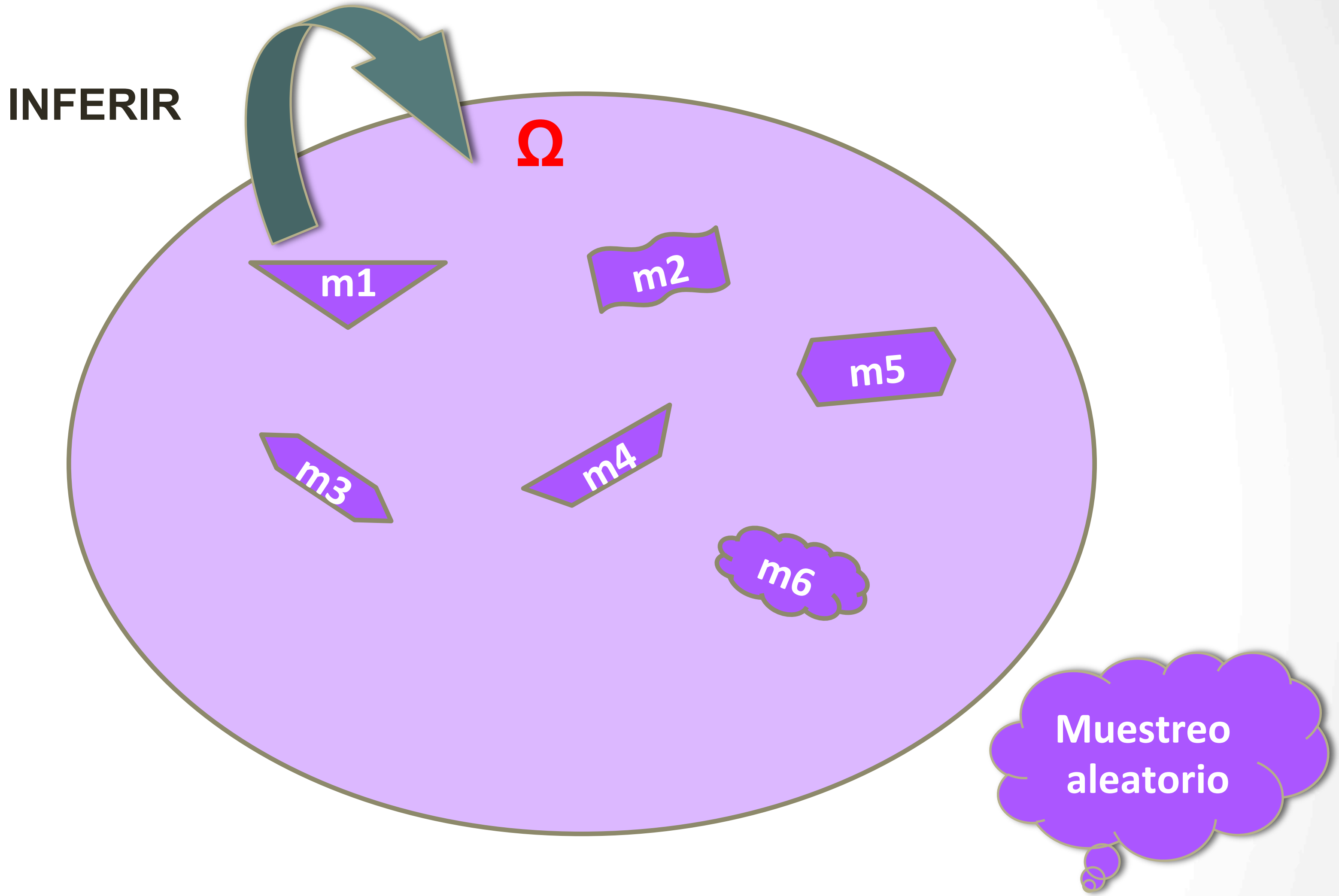
Técnicas de muestreo

Cuando elegimos individuo de una población de estudio para formar muestras podemos encontrarnos en las siguientes situaciones:

- **Muestreos probabilísticos o aleatorios**
  - Elementos de la población tienen una probabilidad conocida de ser incluidos en la muestra.
  - Interesantes para usar estadística matemática con ellos.
  - El **error muestral es cuantificable** → aplicar técnicas de inferencia estadística (determinación del tamaño muestral, estimación por intervalos, contrastes de hipóetes,⋯

- **Muestreos no probabilísticos**
  - No se conoce la probabilidad.
  - Son muestreos que seguramente esconden sesgos.
  - En principio no se pueden extrapolar los resultados a la población.

Entonces vamos a ver algunas técnicas básicas de muestreo probabilístico: aleatorio simple, sistemático y estratificado



El **sesgo de selección** se produce cuando la muestra no es representativa de la población. (**muestras no probabilísticas**)

Llegaremos a **malas estimaciones y conclusiones erróneas**

Muestreo aleatorio simple (m.a.s.)

- Se eligen individuos de la población de estudio (N), de manera que todos tienen la misma probabilidad de aparecer (1/N), hasta alcanzar el tamaño muestral deseado.

- Se puede realizar partiendo de **listas de individuos** de la población, y eligiendo individuos aleatoriamente con un ordenador.

- Su aplicación tiene un **coste alto**.

- En general, las técnicas de inferencia estadística suponen que la muestra ha sido elegida usando m.a.s., aunque en realidad se use alguna de las que veremos a continuación.

Muestreo sistemático

Se necesita una lista enumerada de todos los individuos de la población (población a estudio) ordenada según un variable cuantitativa, por ejemplo edad, altura, IMC, ingresos familiares..., pero que no tenga que ver con la variable que se pretende estudiar.

Si se cuenta con una población a estudio (N) y se quiere obtener una muestra de tamaño (n):

1. Calcular k como el número entero más próximo a N/n.
2. Seleccionar mediante m.a.s. n_0 entre 1 y k.
3. Seleccionar los individuos que ocupan el lugar n_0, n_0+k, n_0+2k, $n_0 + 3k$,...

Muestreo estratificado

Se aplica cuando sabemos que hay ciertos factores (variables, subpoblaciones o estratos) que pueden influir en el estudio y queremos asegurarnos de tener cierta cantidad mínima de individuos de cada tipo:

- Hombres y mujeres,
- Jóvenes, adultos y ancianos…
- Españoles y extranjeros

La idea fundamental es construir muestras de tamaños proporcionales a la población en cada uno de los estratos o categorías.

Muestreo estratificado

Se quiere obtener una muestra de tamaño (n), de un población de tamaño (N) con tres estratos con tamaños N_1, N_2, N_3.

1. Se seleccionan con m.a.s. en cada estrato, una muestra de tamaño

$$n_1 = \left(\frac{N_1}{N}\right) \cdot n, \; n_2 = \left(\frac{N_2}{N}\right) \cdot n, \; n_3 = \left(\frac{N_3}{N}\right) \cdot n$$

Muestreo por grupos o conglomerados

Se aplica cuando es difícil tener una lista de todos los individuos que forman parte de la población de estudio, pero sin embargo sabemos que se encuentran agrupados naturalmente en grupos.

Se realiza eligiendo varios de esos grupos al azar, y ya elegidos algunos podemos estudiar a todos los individuos de los grupos elegidos o bien seguir aplicando dentro de ellos más muestreos por grupos, por estratos, aleatorios simples,…

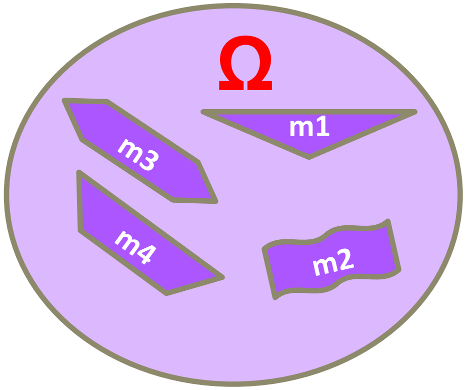
- Para conocer la opinión de los médicos del sistema nacional de salud, podemos elegir a varias regiones de España, dentro de ellas varias comarcas, y dentro de ellas varios centros de salud, y…

Distribuciones en el muestreo

Si utilizamos una técnica de muestreo para recoger la muestra aleatoria, entonces puede construirse un **estimador puntual.**

Puesto que la muestra es aleatoria:

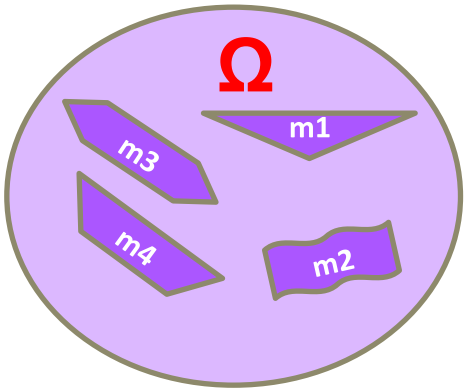
1. Se pueden obtener $\binom{N}{n}$ muestras.
2. Y para cada una de las muestras se dispondría de un estimador puntual $\bar{x}$ del parámetro poblacional (μ).



Estimador Parámetro

MUESTRA	MEDIA MUESTRAL	MEDIA POBLACIONAL
m1	$\overline{X}_1$	μ
m2	$\overline{X}_2$	μ
m3	$\overline{X}_3$	μ
m4	$\overline{X}_4$	μ
m5	$\overline{X}_5$	μ
m6	$\overline{X}_6$	μ

Distribuciones en el muestreo



Estimador **Parámetro**

MUESTRA	MEDIA MUESTRAL	MEDIA POBLACIONAL
m1	$\overline{X}_1$	μ
m2	$\overline{X}_2$	μ
m3	$\overline{X}_3$	μ
m4	$\overline{X}_4$	μ
m5	$\overline{X}_5$	μ
m6	$\overline{X}_6$	μ

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

Puesto que el estimador puntual varía en cada muestra, entonces este **estimador es una variable aleatoria.** → **se puede hallar la estimación de los parámetros poblaciones.**

Recordamos

El parámetro es una característica o medida obtenida mediante el uso de todos los valores de datos de una población específica.

Estimación, es el proceso de estimar el valor de un parámetro a partir de la información obtenida de una muestra.

Estimador (o punto estimado) es una estimación de un valor numérico específico de un parámetro utilizando los valores de datos de una muestra.

Un **estimador** es una cantidad numérica calculada sobre una muestra y que esperamos que sea una buena aproximación de cierta cantidad con el mismo significado en la población (parámetro).

Recordamos

Los **estimadores** son características cuantitativas calculadas a partir de los datos de la muestra observada que, por su conducción, intentan acercarse al verdadero valor del parámetro desconocido de la población.

Estimadores puntuales (media)

Propiedades de E[X]

X,Y Variables aleatorias y β una constante

1. **E[X+Y] = E[X]+E[Y]**
2. **E[X-Y] = E[X]-E[Y]**
3. **E[β X]= βE[X]**
4. **E[β]= β**

Suponemos X_1 ,X_2 ,…,X_n n variables aleatorias con media μ.
(E[X_i]= μ unknown)

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

$$E\left[\overline{X}\right] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = E\left[\frac{X_1 + \ldots + X_n}{n}\right] = E\left[\frac{1}{n}\left(X_1 + \ldots + X_n\right)\right] =$$

❸ ❶

$$= \frac{1}{n} E\left[\left(X_1 + \ldots + X_n\right)\right] = \frac{1}{n}\left(E\left[X_1\right] + \ldots + E\left[X_n\right]\right) = \frac{1}{n}\left(\mu + \ldots + \mu\right) = \frac{1}{n} n\mu = \mu$$

MUESTRA	MEDIA MUESTRAL	MEDIA POBLACIONAL
m1	$\overline{X}_1$	μ
m2	$\overline{X}_2$	μ
m3	$\overline{X}_3$	μ
m4	$\overline{X}_4$	μ
m5	$\overline{X}_5$	μ
m6	$\overline{X}_6$	μ

Estimadores puntuales (varianza)

Propiedades VAR[X]
X,Y variables aleatorias independientes y β es una constante

1. VAR[X+Y]= VAR[X]+VAR[Y]
2. VAR[X-Y]= VAR[X]+VAR[Y]
3. VAR[β X]= β²VAR[X]
4. VAR[β]=0

Suponemos $X_1, X_2, \ldots, X_n$ n variables aleatorias con varianza σ².

(VAR[X_i]= σ² unknown)

$$VAR\left[\overline{X}\right] = VAR\left[\sum_{i=1}^{n}\frac{X_i}{n}\right] = VAR\left[\frac{X_1+\ldots+X_n}{n}\right] = VAR\left[\frac{1}{n}\left(X_1+\ldots+X_n\right)\right] =$$

❸ ❶

$$= \frac{1}{n^2}VAR\left[\left(X_1+\ldots+X_n\right)\right] = \frac{1}{n^2}\left(VAR\left[X_1\right]+\ldots+VAR\left[X_n\right]\right) = \frac{1}{n^2}\left(\sigma^2+\ldots+\sigma^2\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Conclusión:

- ✓ **Media población= Media Muestral** $E\left[\overline{X}\right] = \mu$
- ✓ **Varianza población = varianza muestral / n** $VAR\left[\overline{X}\right] = \frac{\sigma^2}{n}$

Estimadores puntuales máximo verosímiles

Parámetro (población)	Estimador puntual (muestra)	
Media μ	Media muestral	$\overline{X} = \hat{\mu} = \sum_{i=1}^{n} \frac{X_i}{n}$
Proporción ρ	Proporción muestral	$p = \hat{\rho} = r/n$
Varianza σ^2	Varianza muestral corregida	$s^2 = \hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}$
Diferencia de medias $\mu_1 - \mu_2$	Diferencia de medias muestrales	$\overline{X}_1 - \overline{X}_2 = \hat{\mu}_2 - \hat{\mu}_2$
Diferencia de proporciones $\rho_1 - \rho_2$	Diferencia de proporciones muestrales	$p_1 - p_2 = \hat{\rho}_1 - \hat{\rho}_2$

Teorema central del límite

Dado que un estimador es una v. aleatoria, entonces tiene sentido preguntarse,

¿Podría ser conocida la distribución de la probabilidad asociada a cada uno de los estimadores puntuales descritos?

Teorema central del límite

Teorema central del límite

Sean $\{X_1, ..., X_n\}$ variables aleatorias independientes con media μ y varianza σ^2 finita y conocida, entonces

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

EE = des. típica

Cuando el tamaño de la muestra fuera lo suficiente grande (n>30)

Normal estandarizada

Si X~N(μ,σ), E[X]=μ VAR[X]= σ^2 entonces Z= (X-μ)/ σ ~N(0,1)

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$$

Intervalo de confianza para la media (μ)

Definición:

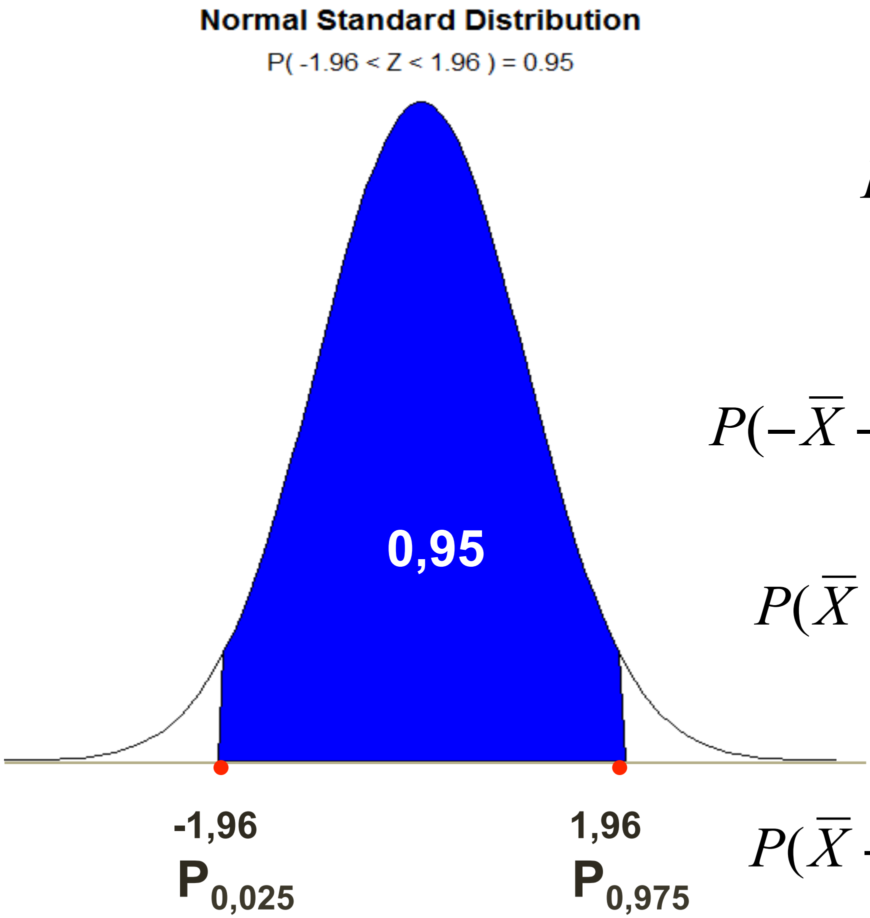
Intervalo [a,b] que contiene al verdadero valor de la media (poblacional) con una confianza $1-\alpha$.

Nótese que α es la probabilidad de error (no contener al parámetro). En el tema de contrastes de hipótesis se llamará probabilidad de error tipo I o nivel de significación.

En general el tamaño del intervalo disminuye con el tamaño muestral y aumenta con $1-\alpha$.

Nivel de confianza es frecuente que se exprese en %

Intervalo de confianza para la media (μ)



$$P(-1{,}96 < \overline{X} < 1{,}96) = 0{,}95$$

$$P\left(-1{,}96 < \frac{\overline{X}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} < 1{,}96\right) = 0{,}95$$

$$P\left(-\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

$$P\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

μ desconocida
σ conocida

$$Z = \frac{\overline{X}-\mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0{,}1)$$

Intervalo de confianza para la media (μ)

μ desconocida
σ conocida

$$Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$$

$$P(\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

$$IC_{0,95}(\mu) = \left(\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right)$$

$$IC_{0,95}(\mu) = \left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Entonces el 95% de los intervalos que podrían construirse a partir de las diferentes muestras que podrían haberse obtenido por muestreo aleatorio, contendrán al verdadero valor del parámetro poblacional.

Intervalo de confianza para la media (μ)

μ desconocida
σ desconocida

$$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \sim t_{n-1}$$

N-1 grados de libertad

$$P(\bar{X} - t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{1-\alpha/2}^{n-1} \frac{s}{\sqrt{n}}) = 0{,}95$$

$$IC_{1-\alpha}\left(\mu\right) = \bar{x} \pm t_{1-\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

$$IC_{0{,}95}\left(\mu\right) = \bar{x} \pm t_{1-\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

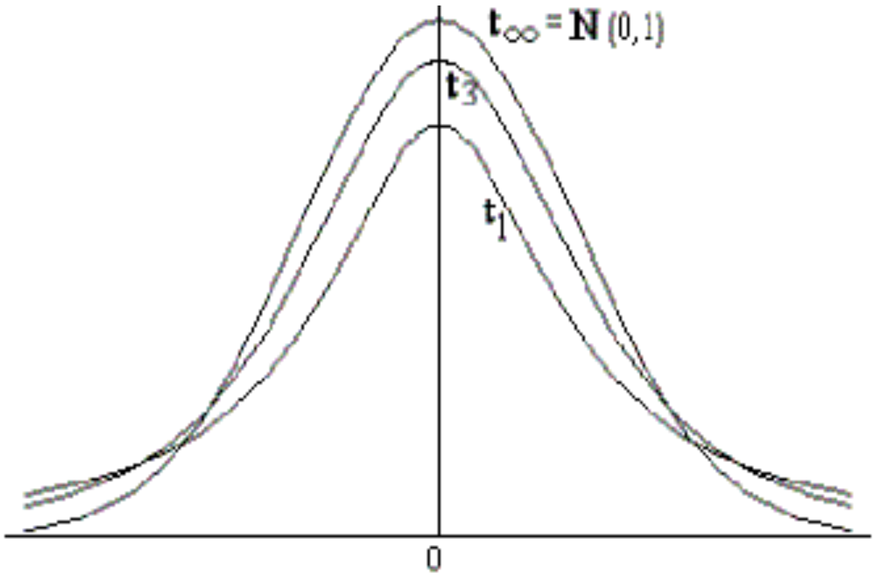
IC para la media (μ) - Pasos

- 1) Parámetro poblacional
- 2) Estimador puntual
- 3) Distribución muestral:
- 4) IC = [estimador ± coeficiente x EE]
- 5) Calculamos el Coeficiente
- 6) Construimos el Intervalo de confianza

$$IC_{1-\alpha}\left(\mu\right) = \overline{x} \pm t_{1-\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

T de student

- Tiene un parámetro denominado grados de libertad.
- Cuando aumentan los grados de libertad, más se acerca a N(0,1).
- Es simétrica con respecto al cero.
- Se consideran valores anómalos los que se alejan de cero (positivos o negativos).



Intervalo de confianza para la media (μ)

0	E	Ac1	Pr	Ac2	T	0	E	Ac1	Pr	Ac2	T
1	28	1	1	1	6.5	32	31	1	2	1	2.0
2	30	2	1	2	3.5	33	32	1	2	1	3.5
3	22	1	1	1	8.5	34	17	1	2	1	7.5
4	25	1	1	1	2.5	35	28	2	2	1	4.0
5	32	2	1	2	0.0	36	26	1	2	1	3.5
6	21	1	1	1	6.0	37	26	1	2	1	3.5
7	18	1	1	2	0.0	38	23	1	2	1	6.0
8	25	1	1	2	0.0	39	28	2	2	2	7.0
9	36	2	1	2	1.5	40	22	1	2	1	6.0
10	30	1	1	2	0.0	41	22	1	2	1	7.0
11	23	1	1	1	6.0	42	33	2	2	2	0.0
12	24	1	1	1	3.0	43	34	2	2	1	0.5
13	28	2	1	2	1.5	44	25	1	2	2	0.0
14	39	2	1	2	0.0	45	24	1	2	1	8.0
15	25	1	1	1	4.5	46	26	2	2	1	5.0
16	25	1	1	2	0.0	47	27	2	2	1	4.0
17	22	1	1	1	6.5	48	29	2	2	1	3.0
18	27	2	1	2	2.0	49	31	2	2	1	0.0
19	29	2	1	2	0.0	50	34	1	2	1	2.5
20	33	2	1	2	1.0	51	25	1	2	1	3.0
21	34	1	1	1	4.5	52	38	2	2	1	2.0
22	30	1	1	2	0.5	53	32	2	2	1	2.0
23	21	1	1	1	4.5	54	21	1	2	1	5.0
24	20	1	1	1	6.0	55	20	1	2	1	7.0
25	19	1	1	1	8.5	56	27	2	2	1	3.0
26	23	1	1	1	9.0	57	26	1	2	1	3.5
27	23	2	1	1	9.0	58	25	1	2	1	1.0
28	27	2	1	2	0.5	59	24	1	2	1	7.0
29	29	2	1	1	3.0	60	22	1	2	1	6.5
30	22	1	1	2	0.0	61	32	2	2	1	2.5
31	40	2	1	2	0.0	62	31	2	2	1	2.5

Estudio sobre lactancia (n=62), la media muestral (Edad) toma un valor 26,95 años, y una desviación típica de 5,20.

a) Calcula un intervalo de confianza con un nivel de confianza del 95% para la media de edad.

$$IC_{0,95}(\mu) = \overline{x} \pm t_{1-\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

Intervalo de confianza para la media (μ)

Tabla 5: Percentiles de la distribución t de Student $P(t_n \leq t_p)=p$

grados de libertad	$t_{0.005}$	$t_{0.01}$	$t_{0.025}$	$t_{0.05}$	$t_{0.10}$	$t_{0.25}$	$t_{0.50}$	$t_{0.75}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	-63,657	-31,821	-12,706	-6,314	-3,078	-1,000	0,000	1,000	3,078	6,314	12,706	31,821	63,657
2	-9,925	-6,965	-4,303	-2,920	-1,886	-0,816	0,000	0,816	1,886	2,920	4,303	6,965	9,925
3	-5,84									2,353	3,182	4,541	5,841
4	-4,60									2,132	2,776	3,747	4,604
5	-4,03									2,015	2,571	3,365	4,032
6	-3,70									1,943	2,447	3,143	3,707
7	-3,49									1,895	2,365	2,998	3,499
8	-3,35									1,860	2,306	2,896	3,355
9	-3,25									1,833	2,262	2,821	3,250
10	-3,16									1,812	2,228	2,764	3,169
11	-3,10									1,796	2,201	2,718	3,106
12	-3,055	-2,681	-2,179	-1,782	-1,356	-0,695	0,000	0,695	1,356	1,782	2,179	2,681	3,055
13	-3,012	-2,650	-2,160	-1,771	-1,350	-0,694	0,000	0,694	1,350	1,771	2,160	2,650	3,012
14	-2,977	-2,624	-2,145	-1,761	-1,345	-0,692	0,000	0,692	1,345	1,761	2,145	2,624	2,977
15	-2,947	-2,602	-2,131	-1,753	-1,341	-0,691	0,000	0,691	1,341	1,753	2,131	2,602	2,947
16	-2,921	-2,583	-2,120	-1,746	-1,337	-0,690	0,000	0,690	1,337	1,746	2,120	2,583	2,921
17	-2,898	-2,567	-2,110	-1,740	-1,333	-0,689	0,000	0,689	1,333	1,740	2,110	2,567	2,898
18	-2,878	-2,552	-2,101	-1,734	-1,330	-0,688	0,000	0,688	1,330	1,734	2,101	2,552	2,878
19	-2,861	-2,539	-2,093	-1,729	-1,328	-0,688	0,000	0,688	1,328	1,729	2,093	2,539	2,861
20	-2,845	-2,528	-2,086	-1,725	-1,325	-0,687	0,000	0,687	1,325	1,725	2,086	2,528	2,845
21	-2,831	-2,518	-2,080	-1,721	-1,323	-0,686	0,000	0,686	1,323	1,721	2,080	2,518	2,831
22	-2,819	-2,508	-2,074	-1,717	-1,321	-0,686	0,000	0,686	1,321	1,717	2,074	2,508	2,819
23	-2,807	-2,500	-2,069	-1,714	-1,319	-0,685	0,000	0,685	1,319	1,714	2,069	2,500	2,807
24	-2,797	-2,492	-2,064	-1,711	-1,318	-0,685	0,000	0,685	1,318	1,711	2,064	2,492	2,797
25	-2,787	-2,485	-2,060	-1,708	-1,316	-0,684	0,000	0,684	1,316	1,708	2,060	2,485	2,787
26	-2,779	-2,479	-2,056	-1,706	-1,315	-0,684	0,000	0,684	1,315	1,706	2,056	2,479	2,779
27	-2,771	-2,473	-2,052	-1,703	-1,314	-0,684	0,000	0,684	1,314	1,703	2,052	2,473	2,771
28	-2,763	-2,467	-2,048	-1,701	-1,313	-0,683	0,000	0,683	1,313	1,701	2,048	2,467	2,763
29	-2,756	-2,462	-2,045	-1,699	-1,311	-0,683	0,000	0,683	1,311	1,699	2,045	2,462	2,756
30	-2,750	-2,457	-2,042	-1,697	-1,310	-0,683	0,000	0,683	1,310	1,697	2,042	2,457	2,750
35	-2,724	-2,438	-2,030	-1,690	-1,306	-0,682	0,000	0,682	1,306	1,690	2,030	2,438	2,724
40	-2,704	-2,423	-2,021	-1,684	-1,303	-0,681	0,000	0,681	1,303	1,684	2,021	2,423	2,704
45	-2,690	-2,412	-2,014	-1,679	-1,301	-0,680	0,000	0,680	1,301	1,679	2,014	2,412	2,690
50	-2,678	-2,403	-2,009	-1,676	-1,299	-0,679	0,000	0,679	1,299	1,676	2,009	2,403	2,678
55	-2,668	-2,396	-2,004	-1,673	-1,297	-0,679	0,000	0,679	1,297	1,673	2,004	2,396	2,668
60	-2,660	-2,390	-2,000	-1,671	-1,296	-0,679	0,000	0,679	1,296	1,671	2,000	2,390	2,660

$$IC_{0,95}(\mu) = \overline{x} \pm t^{n-1}_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

Intervalo de confianza para la media (μ)

b) Calcula un intervalo de confianza con un nivel de confianza del 90% para la media de edad.

c) Calcula un intervalo de confianza con un nivel de confianza del 99% para la media de edad.

USO GEOGEBRA PARA CALCULAR INTERVALOS DE CONFIANZA

En general el tamaño del intervalo disminuye con el tamaño muestral y aumenta con $1-\alpha$.