ESTIMACIÓN (de parámetros)

Estimación puntual y por intervalo. Intervalo de confianza de la media y de la proporción. Distribución t de Student. Cálculo del tamaño de muestra.

INTRODUCCIÓN

"I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them!" (de Gosset a Fisher)[1]

A. Definiciones (Diccionario de la Lengua Española – R.A.E.)

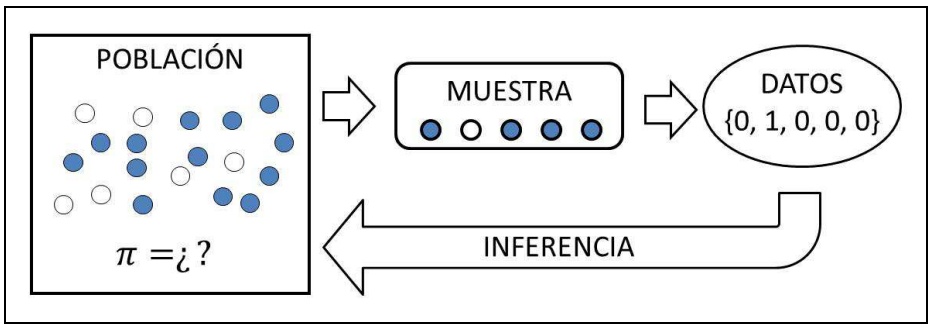
• Inferir. Del lat. inferre 'llevar a'.

1. tr. Deducir algo o sacarlo como conclusión de otra cosa. *Se infiere DE su rostro que está contento.*

• Estimar. Del lat. aestimāre.

1. tr. Calcular o determinar el valor de algo. *Estimaron los daños EN mucho dinero*.

B. "A study of the inferences made concerning a population by using samples drawn from it, together with indications of the accuracy of such inferences by using probability theory, is called *statistical inference*"[2]



C. De lo anterior ha de quedar claro que el interés está en la población. La muestra proporciona las observaciones para conseguir datos cuyo análisis dará una información aproximada de lo que sucede en la población.

D. Por tanto, la estadística inferencial contribuye a la investigación científica siempre que la cuestión a resolver tenga que ver con el mundo empírico y sea necesario la realización de un estudio para recoger datos.

[1] "William Sealy Gosset", en Wikipedia: The Free Encyclopedia https://en.wikipedia.org/wiki/William_Sealy_Gosset ; consultado: 6/10/2016

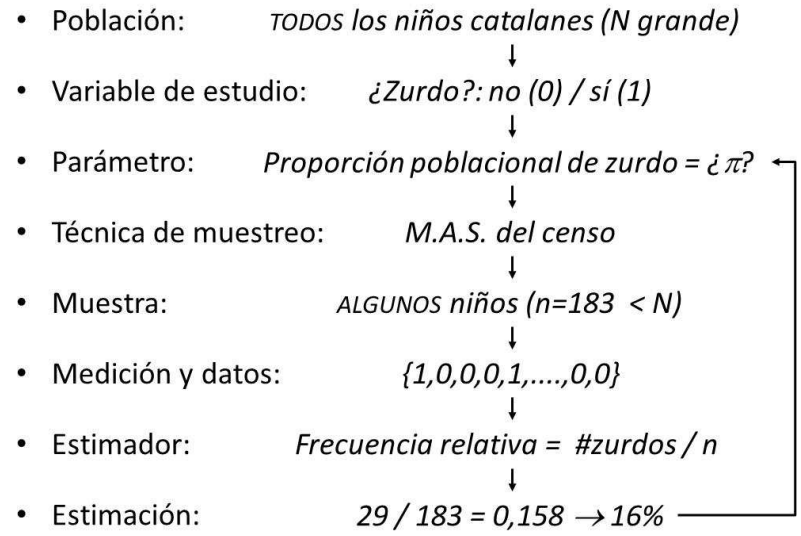
[2] En: Spiegel, MR (2011). Ver bibliografía.

E. Se distinguen dos tipos de preguntas:

- Las que pretenden valorar una característica desconocida: ¿Cuánto vale la media de colesterol en niños con diabetes tipo 1? ¿Cuánto vale tasa de curación de un nuevo fármaco en la población de pacientes de estudio?
- Las que buscan refutar o no una hipótesis: ¿Fumar causa cancer en los adolescentes?, ¿es el nuevo fármaco igual que el convencional en los pacientes atendidos en nuestro entorno?

F. La teoría de la estimación corresponde a la parte de la estadística inferencial cuyo objetivo es responder preguntas del primer tipo, es decir, determinar el valor de un parámetro poblacional.

G. Los conceptos básicos en estimación son:



H. POBLACIÓN es el conjunto de todos individuos que comparten un rasgo común, ya sea que viven en una misma zona o que tienen una misma enfermedad. Las poblaciones pueden ser pequeñas (finita) o muy grandes (infinita).

I. VARIABLE de estudio es una característica o atributo de los individuos que componen la población y que es objeto de interés. Ejemplos:

- Cualitativo: ¿zurdo: si/no?
- Cualitativo: ¿talla (cm)?

J. La distribución de los valores de una variable es específica para cada población. Por ejemplo, el reparto de tallas en la población masculina es diferente al de la población femenina. Por otro lado, dos variables distintas estudiadas en una misma población (colesterol y altura en mujeres) no tienen por qué tener igual distribución.

K. PARÁMETRO es una característica numérica que resume el conjunto de todos los valores poblacionales de una variable X. Ejemplos:

X: ¿zurdo: si/no? → Proporción poblacional de zurdos (π)
X: ¿talla (cm)? → Media poblacional de altura (μ)

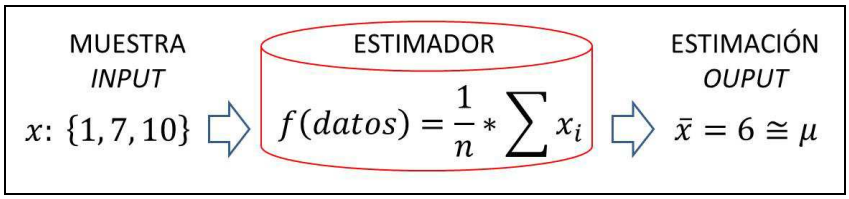
ESTIMACIÓN PUNTUAL y POR INTERVALO

A. En una población estable los parámetros que la caracterizan tienen valor único y constante, pero generalmente desconocido.

B. Un procedimiento de estimación es un intento para asignar valor a un parámetro a partir de los datos obtenidos mediante un muestreo aleatorio. El objetivo es tener una alta probabilidad de que el resultado sea próximo al verdadero valor del parámetro, aunque no coincida exactamente con él.

Población → Muestra de X → Parámetro (poblacional)

C. ESTIMADOR. Función matemática que se evalúa a partir del conjunto de datos de la muestra y sirve para inferir un parámetro. Por ejemplo, la media aritmética de una muestra es un estimador de la media poblacional. Todos los estimadores son estadísticos.



D. ESTIMACIÓN. Resultado numérico obtenido al aplicar el estimador a los datos de una muestra particular. No es necesariamente igual al valor del parámetro, pero se considerará una buena aproximación.

E. Existen muchos estimadores, incluso para un mismo parámetro. La media poblacional puede estimarse con la media aritmética y también con la mediana. Seleccionar el estimador apropiado para cada problema dependerá de sus propiedades, en especial la ausencia de sesgo y el error típico. Esto se conoce estudiando la distribución muestral del estimador.

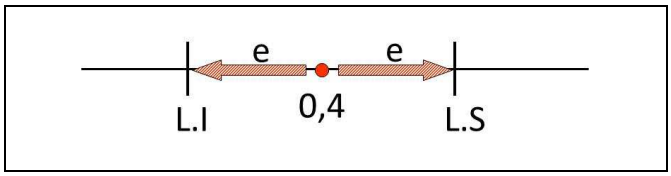
F. Estimadores importantes:

VARIABLE	PARÁMETRO	ESTIMADOR	DISTRIBUCIÓN del ESTIMADOR
Zurdo (si/no)	Proporción (π)	Frecuencia Relativa	Binomial → Normal
Altura (cm)	Media (μ)	Media Aritmética	Normal T de Student
	Variancia (σ^2)	Variancia Corregida	Ji-cuadrado

G. Tipos de Estimación:

- PUNTUAL. Asignar al parámetro el valor resultante de aplicar el estimador a la muestra. Se considera la mejor aproximación al verdadero valor del parámetro.
  Ejemplo: muestra = {0, 1, 0, 1, 0} → fr = 2/5= 0,4 → $0,4 \cong \pi$

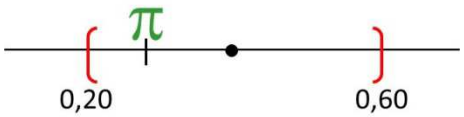
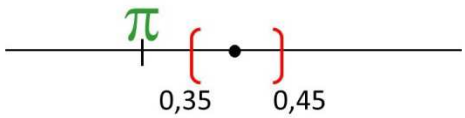
- POR INTERVALO. Rango de valores, definidos por un límite inferior y un límite superior derivados de la muestra, dentro del cual confiamos que esté el parámetro. Generalmente los intervalos son simétricos y se construyen sumando y restando a la estimación puntual una cantidad determinada de incertidumbre o margen de error (e).



La forma compacta de expresarlo es:

$$estimación\ puntual\ \pm\ incertidumbre$$

H. Si por azar la estimación puntual quedara lejos del valor del parámetro, entonces la estimación por intervalo puede ser errónea, en el sentido de que el parámetro no quede incluido (atrapado, encerrado, confinado) entre sus límites. Cuanto mayor sea la amplitud del intervalo menos riesgo habrá de que esto ocurra, pero menos útil será la estimación así obtenida. La estimación por intervalo tiene que ser un compromiso que de ciertas garantías de una buena estimación con la menor amplitud posible.

A) intervalo ancho	B) intervalo estrecho
	

I. PRECISIÓN: semi-amplitud del intervalo.

$$precisión = \frac{A}{2} = \frac{(LS - LI)}{2}$$

Es sinónimo de incertidumbre y margen de error salvo que se interpreta al revés:

- si “e” es un número pequeño, entonces diremos que hay una gran precisión
- si “e” es un número grande, entonces diremos que hay baja precisión.

Por tanto cuanta más incertidumbre, menos precisión.

J. La incertidumbre (e) o precisión, ya sea visto en negativo o en positivo, es el resultado de combinar dos elementos que pertenecen a la distribución muestral del estimador:

- Percentil de la distribución, el cual depende de la confianza que se quiera atribuir al intervalo.
- Error típico de la distribución, que es inversamente proporcional al tamaño de la muestra

Generalmente:

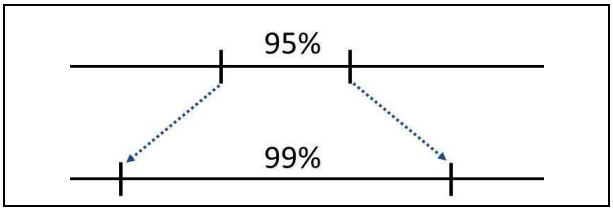
$$\text{Precisión} = \text{percentil} * \text{error típico}$$

A mayor confianza, mayor el valor del percentil y por tanto aumentará la amplitud. Sin embargo a mayor tamaño de muestra, menor error típico y por tanto disminuirá la amplitud.

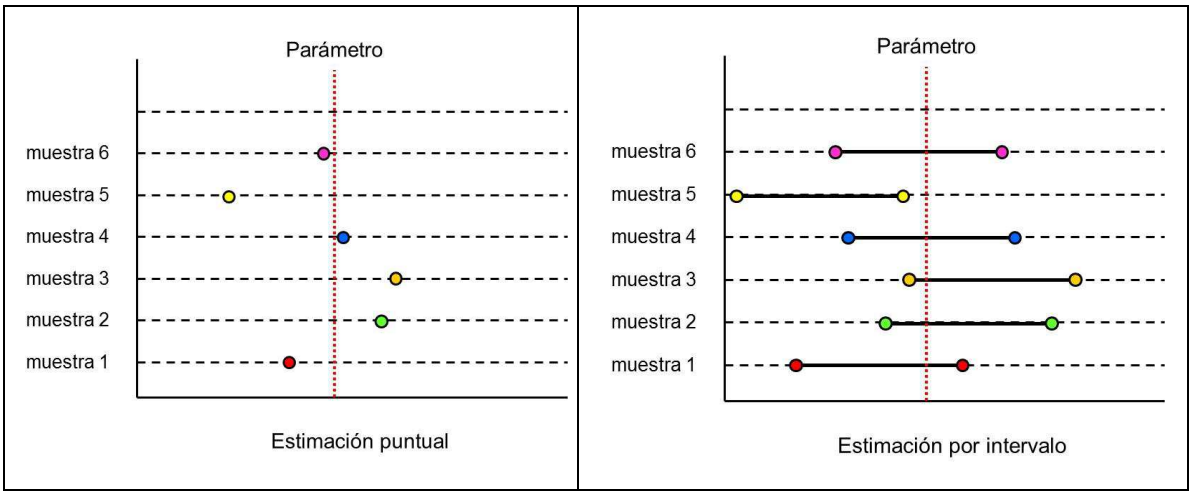
K. Hacer una estimación por intervalo con una total confianza de que incluya el parámetro es inviable pues supone construirlo con una amplitud excesiva y por tanto precisión mínima. Disminuir la confianza por debajo del 100% significa aceptar la posibilidad de que el parámetro quede fuera de los límites. Ese riesgo se denomina alfa y por tanto:

Confianza = (1-alfa).

En la práctica se trabaja con valores de confianza entre 90 y 99%.



L. CONFIANZA es un concepto de probabilidad cuando se aplica al conjunto de todas las estimaciones por intervalo que pueden resultar de infinitos muestreos de una misma población. Se afirma que un intervalo tiene una confianza del 95% cuando el método que se aplica es capaz de producir 95 intervalos que contienen el parámetro de un total de cien intentos. En la siguientes dos figuras se representan las estimaciones obtenidas en seis muestras: puntual y por intervalo. La línea roja vertical marca en ambas la posición verdadera del parámetro. El conjunto de las estimaciones puntuales está centrado alrededor del parámetro, pero sólo dos de ellas quedan justo al lado. Por el contrario, la mayoría de las estimaciones por intervalo consiguen incluir el parámetro dentro de sus límites y en solo un caso el parámetro ha quedado fuera. De un intervalo individual se puede decir que el parámetro está dentro o fuera, pero es incorrecto afirmar que tiene una probabilidad (1-alfa) de contener el parámetro.



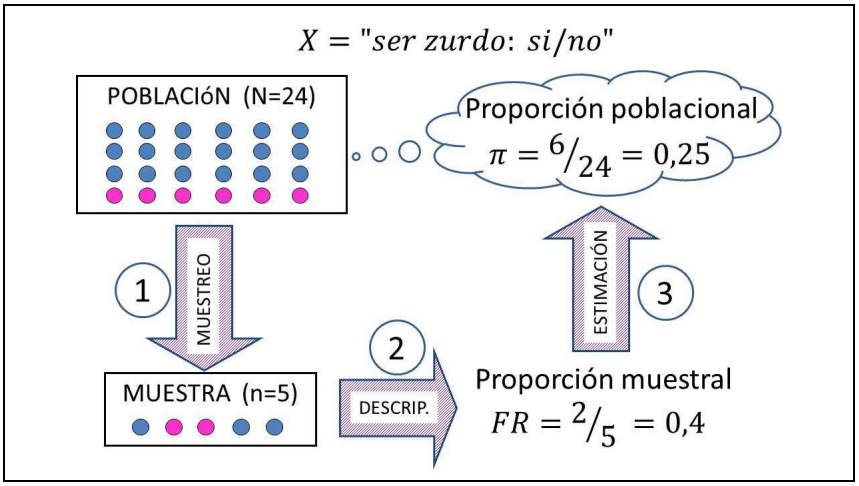
INTERVALO DE CONFIANZA DE LA PROPORCIÓN

A. La estimación de una proporción poblacional (π) tiene sentido cuando la variable que se estudia es de tipo binario, es decir, en un individuo se mide si tiene (éxito) o no tiene (fracaso) un atributo. Ejemplos: “ser zurdo”, “ser diabético”, “tener el grupo sanguíneo A”.

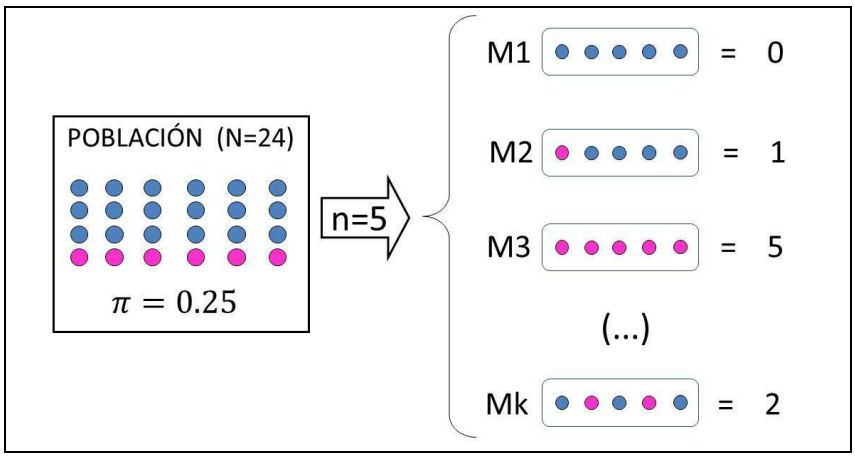
B. Se denomina PROPORCIÓN MUESTRAL a la frecuencia relativa (FR) de éxitos observados en una muestra de tamaño n. En caso de población infinita y selección de individuos mediante muestreo aleatorio simple, se afirma que la proporción muestral (FR) es un buen estimador de la proporción poblacional.

$$FR = \frac{\#exitos}{n} \quad \rightarrow \quad P(éxito) = \pi$$

C. La inferencia más simple de π es utilizar la estimación puntual como valor del parámetro:



D. El numerador del estimador FR es el recuento de éxitos, o frecuencia absoluta, que se comporta como variable aleatoria porque el resultado cambia de muestra en muestra.



E. La distribución de la variable "recuento de éxitos", en las condiciones indicadas más arriba, sigue un modelo Binomial de parámetros n=tamaño de la muestra y p=π:

$$FA = \#exitos \ \sim Binomial(n, \pi)$$

A medida que aumenta el tamaño de la muestra (n), y siempre que la proporción poblacional (π) no sea muy extrema, la distribución Binomial converge a un modelo Normal de parámetros media igual a n∗π y desviación típica igual a √[n∗π∗(1-π)]:

$$FA = \#exitos \ \xrightarrow{n\ grande} \ \sim \ Normal\left(n * \pi, \sqrt{n\pi(1-\pi)}\right)$$

Para usar la Normal la regla práctica es n ≥ 30.

F. La distribución del estimador FR en muestras de gran tamaño también es Normal, pues dividir por n es un re-escalamiento que no modifica el modelo, aunque sí el valor de los parámetros:

$$FR = \frac{FA}{n} \ \xrightarrow{n\ grande} \ \sim \ Normal\left(\pi, \sqrt{[\pi(1-\pi)/n]}\right)$$

G. El centro de la distribución de FR es π, por ello se afirma que FR es un estimador insesgado.

H. El error típico es √[π∗(1-π)/n] que depende de n. La dispersión, o incertidumbre, en torno al parámetro se puede reducir aumentando el tamaño de la muestra.

I. Derivar la fórmula para hacer la estimación por intervalo de una proporción tiene algunas complicaciones que generalmente se simplifican. La fórmula más clásica es una aproximación que se obtiene con el método de Wald y tiene la forma general de un intervalo simétrico. Otros métodos como el "score" de Wilson da lugar a intervalos asimétricos[3] . La precisión de la fórmula de Wald se compone de un percentil zeta y del error típico de la Normal para muestras grandes. Además, dado que el valor de π es desconocido se sustituye éste por su estimación puntual (fr):

$$fr \pm precisión \quad \rightarrow \quad fr \ \pm \left[\ z_{(1-\alpha/2)} \ * \ \sqrt{\frac{fr * (1-fr)}{n}} \ \right]$$

J. El percentil de la fórmula corresponde al valor de una Normal (0,1) que acumula por debajo una probabilidad de (1-alfa/2).
Ejemplo: si se impone una confianza del 95%, entonces alfa valdrá 5% y la zeta es el valor que acumula una probabilidad de (1-0,025) = 0,975, es decir, z=1,96.

K. Es muy importante que se cumplan las condiciones de aplicación para usar esta fórmula basada en la Normal. Si la muestra es n=30 y la probabilidad del éxito muy pequeña, tal que sólo se observa un éxito, entonces ocurrirá que:

- estimación puntual: FR=1/30 = 0,033 OK
- estimación por intervalo: (-0,031 a 0,098) ERROR, ¡LI es negativo!

[3] En: Pruim, R (2011). Ver bibliografía.

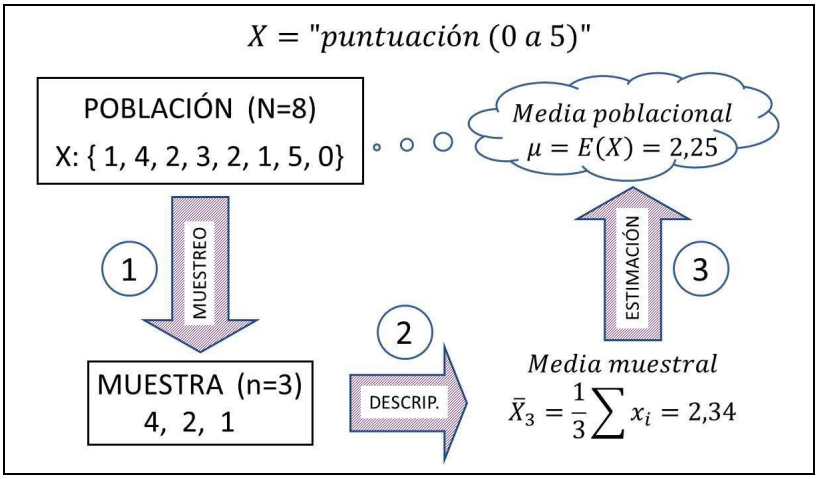
A. La estimación de una media poblacional (μ) tiene sentido cuando la variable que se estudia es cuantitativa. Ejemplos: “altura (cm)”, “concentración de colesterol (mg/dL)”.

B. Se denomina MEDIA MUESTRAL a la media aritmética ($\bar{X}_n$) de los datos recogidos en una muestra de tamaño n. En caso de población infinita, selección de individuos mediante muestreo aleatorio simple y variable de estudio razonablemente simétrica, la media muestral es uno de los mejores estimadores de la media poblacional.

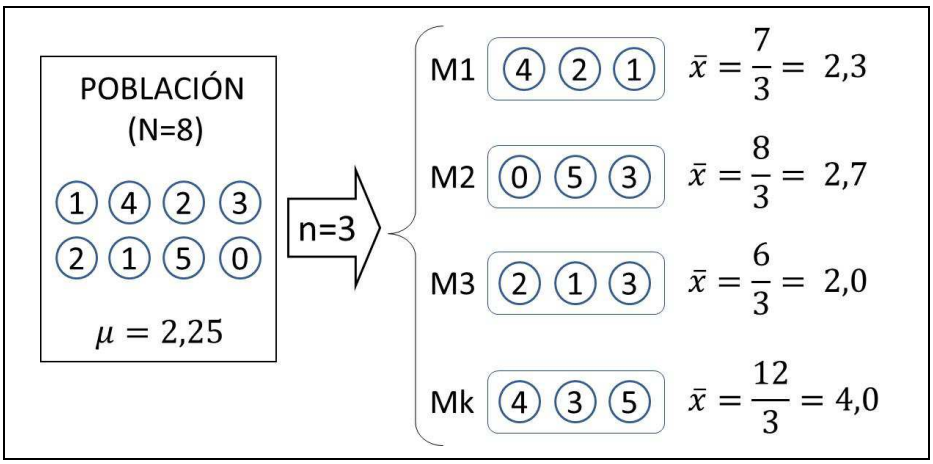
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \rightarrow \quad E(X) = \mu$$

La mediana es un estimador alternativo, pero es menos eficiente (mayor error típico) y es más costoso de cálculo en muestras de gran tamaño.

C. La inferencia más simple de μ es utilizar la estimación puntual como valor del parámetro:



D. El estimador $\bar{X}_n$ se comporta como variable aleatoria porque el resultado cambia de muestra en muestra.



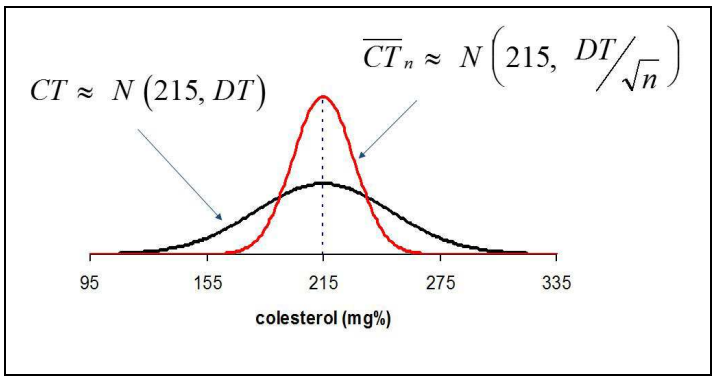
E. Su distribución muestral depende de la distribución de la variable original de estudio (X) en la población. Existen dos teoremas que pueden aplicarse:

- Teorema de la Adición. si la variable X medida sigue N(μ, sigma)

$$\bar{X}_n \sim Normal\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$$

- Teorema del Límite Central. si la variable X NO sigue una normal, pero n>30

$$\bar{X}_n \ \ aprox \ \ Normal\left(\mu, {}^{\sigma}/_{\sqrt{n}}\right)$$



F. El centro de la distribución del estimador $\bar{X}_n$ es μ , por ello se afirma que es un estimador insesgado.

G. El error típico es √[σ/n] que depende de n. La dispersión, o incertidumbre, en torno al parámetro se puede reducir aumentando el tamaño de la muestra.

H. La fórmula para hacer la estimación por intervalo tiene la forma general de un intervalo simétrico, sacando los componentes de la precisión de la distribución muestral correspondiente. Dado que el valor de μ es desconocido se sustituye por su estimación puntual ($\bar{x}$):

$$\bar{x} \pm precisión \quad \rightarrow \quad \bar{x} \pm \left[\ z_{(1-{}^{\alpha}/_{2})} \ * \frac{\sigma}{\sqrt{n}} \ \right]$$

I. El uso de la fórmula anterior requiere conocer el valor del parámetro desviación típica poblacional (σ), algo que en la práctica no sucede. La solución pasa por sustituir el parámetro por su mejor estimación, que es la estimación puntual obtenida con la desviación típica muestral corregida. En términos cuadráticos:

$$S_{n-1}{}^2 = \left[\frac{n}{n-1} * S_n{}^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \rightarrow \quad V(X) = \sigma^2$$

La corrección consiste en multiplicar la desviación sin corregir por [n/(n-1)]. Si n es muy grande, el cociente vale prácticamente uno y la corrección es mínima. Sin embargo, para “enes” pequeñas puede tener importancia. Por ejemplo, si n=6 entonces hay que multiplicar 1,2 lo que significa inflar la desviación un 20%. La variancia corregida será siempre numéricamente mayor que la sin corregir.

J. La variancia poblacional, σ^2, es una constante, sin embargo el estimador ${S_{n-1}}^2$ es una variable aleatoria. Si la variable original de medida X sigue una distribución Normal, entonces se afirma que la "variancia muestral corregida" multiplicada por el factor $(n-1)/\sigma^2$ sigue un modelo de distribución Ji-cuadrado con su parámetro "grados de libertad" (gl) igual a (n-1):

$$\left[\frac{(n-1)}{\sigma^2}\right] * {S_{n-1}}^2 \sim \chi^2(gl = n-1)$$

El estimador sin corregir es sesgado, mientras que el corregido es insesgado. El factor de corrección consigue que la distribución muestral esté centrada allí donde le toca.

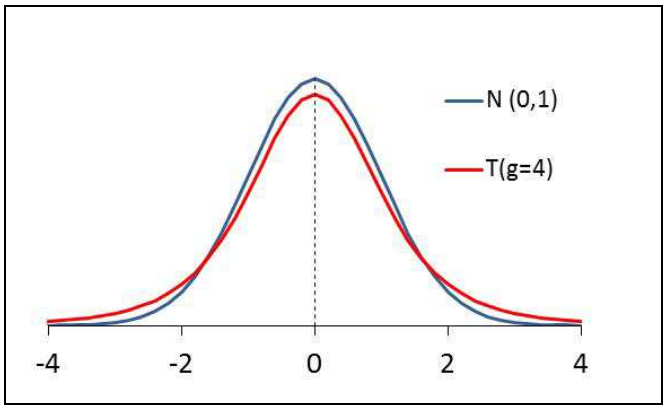
K. Sustituir el parámetro de dispersión por un estimador aportará un extra de incertidumbre a la estimación por intervalo de la media poblacional. Como consecuencia el percentil de la fórmula de la estimación por intervalo hay que buscarlo en una distribución t de Student en lugar de una Normal (0,1).

$$\bar{x} \pm precisión \quad \rightarrow \quad \bar{x} \pm \left[t_{(1-\alpha/2;gl=n-1)} * \frac{S_{n-1}}{\sqrt{n}} \right]$$

L. La función de densidad de una distribución T de STUDENT es muy similar a una campana de Gauss. Es siempre simétrica, pero leptocúrtica (coeficiente de curtosis positivo) ya que el pico es más agudo y las colas más gruesas, cuando se compara con una Normal de igual dispersión. Sólo tiene un parámetro que se llama "grados de libertad" (gl) y determina los valores esperados:

$$E(T) = 0 \quad y \quad V(T) = gl/(gl-2) > 1$$

El parámetro grados de libertad vale (n-1) cuando se utiliza para la estimación de una media. A medida que aumenta el valor de este parámetro, es decir el tamaño de muestra, la curva de la T de Student se va modificando hasta adaptarse a una zeta en n igual a infinito.



M. Nota histórica. La distribución T de Student fue concebida por William S. Gosset mientras trabajaba como científico para la empresa cervera Guiness[4]. Su preocupación era adaptar el uso de la Normal a los experimentos de mejora que se hacían con muestras de tamaño pequeño. En 1908 publicó, con el pseudónimo

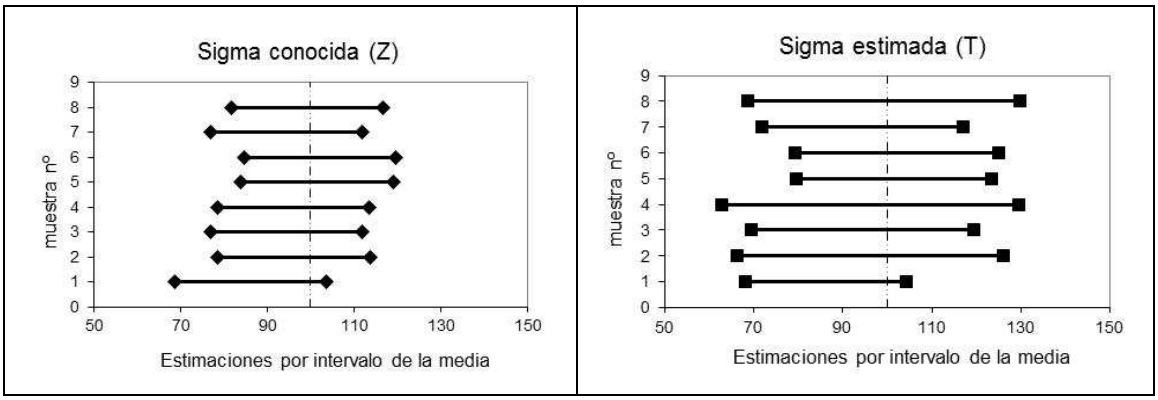
[4] Boland, PJ. A Biographical Glimpse of William Sealy Gosset. The American Statistician, Vol. 38, No. 3 (Aug., 1984), pp. 179-183
-Zabell, SL. On Student's 1908 Article "The Probable Error of a Mean". Journal of the American Statistical Association, Vol. 103, No. 481 (Mar., 2008), pp.1-20

Student, su método estadístico en la revista Biometrika que se convertiría en su artículo más importante[5]:
"A curve has been found representing the frequency distribution of values of the means of such samples, when these values are measured from the mean of the population in terms of the standard deviation of the sample."

Luego en 1925 Ronald A. Fisher, otro estadístico con quien Gosset se carteaba desde 1912, publicó la demostración matemática y denominó la curva t de Student.[6]

N. Utilizar T de Student en lugar de zeta para calcular el intervalo de confianza afecta a las estimaciones por intervalo de dos maneras: a) la amplitud de los intervalos no es constante ya que la estimación de la dispersión varía con cada muestra, b) los intervalos con T son más anchos, porque el percentil t es mayor que z para una misma confianza:

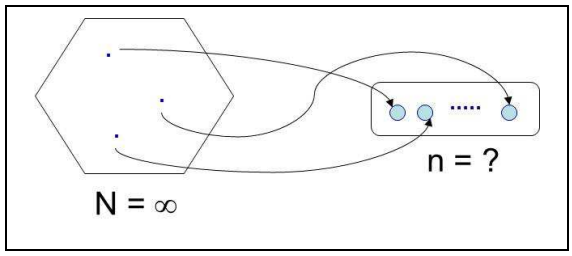
percentil 80% (Z) = 0.842 < percentil 80% (T) = 0.941



[5] Student The Probable Error of a Mean. Biometrika, Vol. 6, No. 1 (Mar., 1908), pp. 1-25
[6] Box, JF. Gosset, Fisher, and the t Distribution. The American Statistician, Vol. 35, No. 2 (May, 1981), pp. 61-66

CÁLCULO del TAMAÑO de MUESTRA

A. Una de las tareas a realizar durante la planificación de un estudio es determinar el tamaño de muestra (n) necesario para que el resultado final de estimación tenga la precisión debida. La pregunta a resolver es: ¿Cuántos individuos de la población de estudio hay que seleccionar para incluirlos en la muestra?.



B. El tamaño de muestra no puede ser exageradamente grande porque sería costoso, requeriría más tiempo y además sería poco ético. Por otro lado, el tamaño no puede ser pequeño porque el esfuerzo invertido sería inútil y también poco ético.

C. Las formulas de la estadística inferencial sirven para determinar un valor mínimo. En la práctica *"ordinarily the sample size calculation should be based on the statistics used in the analysis of the data"*[7] . Sin embargo,*"begin with a basic formula for sample size".* En resumen, no hay una única fórmula para calcular n.

D. En general los pasos a seguir son:
1. Identificar el objetivo del estudio y el diseño para seleccionar la fórmula apropiada
2. Decidir los valores de entrada de la fórmula
3. Realizar los cálculos
4. Aumentar la n calculada para prever las posibles pérdidas o retiradas del estudio

E. En un estudio cuyo objetivo sea estimar una proporción poblacional seleccionando un grupo de personas por muestreo aleatorio simple, la fórmula para la determinación de n se deriva de la precisión en una estimación por intervalo simétrica:

$$precisión\ (e) = \ z_{(1-\alpha/2)} \ * \sqrt{\pi * (1-\pi)/n}$$

y despejando n:

$$n = \ z^2_{(1-\alpha/2)} * \frac{\pi * (1-\pi)}{e^2}$$

F. Antes de hacer el estudio no se dispone de una estimación puntual para cambiar por π, ¿qué hacer entonces?

- Situación teórica más exagerada. Puesto que en el numerador aparece π multiplicado por su complementario (1-π), el valor que maximiza el producto, y por tanto el valor de n, es π = 0.5:

  π*(1-π) = 0.5*0.5 = 0.25
- Situación basada en información obtenida en un estudio previo:

[7] En: Van Belle, G. (2002). Ver bibliografía

- usar la estimación puntual fr
- usar el límite más desfavorable de la estimación intervalo, es decir, LI ó LS más próximo a 0.5.

G. En un estudio cuyo objetivo sea estimar una media poblacional seleccionando un grupo de personas por muestreo aleatorio simple, la fórmula para la determinación de n se deriva de la precisión en una estimación por intervalo:

$$precisión\ (e) = \ z_{(1-\alpha/2)} \ * \ \sigma/\sqrt{n}$$

y despejando n:

$$n = \ z^2_{(1-\alpha/2)} \ * \frac{\sigma^2}{e^2}$$

F. Para resolver el cálculo de n con esta fórmula es necesario conocer el valor del parámetro σ. A falta de una estimación propia habrá que utilizar resultados de estudios previos o la mejor suposición que se tenga. Por otro lado, el percentil que se aplica es de zeta y no de la t de Student, como una primera aproximación al problema (recordar que t se acerca a z con n infinita). Si tras un primer cálculo el valor de n es pequeño, por ejemplo 8, entonces se repetiría el cálculo usando un percentil de t con grados de libertad igual a 8-1.

BIBLIOGRAFÍA RECOMENDADA

- Daniel, WW. Bioestadística: base para el análisis de las ciencias de la salud. 4ªed. México D.F.: Limusa; 2002.
- Johnson RA, Bhattacharyya GK. Statistics: principles and methods. Hoboken, N.J: Wiley; cop. 2010, 6th ed., International student ed.
- Larson HJ. Introduction to probability theory and statistical inference. New York [etc.] : Wiley, cop. 1982, 3rd ed.
- Pruim, R Foundations and Applications of Statistics: an introduction using R" (2011)
- Rosner B. Fundamentals of biostatistics. Pacific Grove, Calif. : Brooks/Cole, Cengage Learning, 2011. 7th ed., International ed
- Spiegel, MR. Statistics. Shaum's easy outlines. McGrawHill. 2011. 2ªed
- Van Belle, G. Statistical Rules of thumb. Wiley Series in probability and statistics. 2002

GLOSARIO

Amplitud de un IC
Confianza
Distribución de muestreo
Distribución t de Student
Error típico (estándar)
Estadístico t de Student
Estimación por intervalo
Estimación puntual
Estimador
Estimador "desviación típica corregida"
Estimador "media muestral"
Estimador "proporción muestral"
Grados de libertad (gl)
Intervalo de confianza (1-alfa)%
Límites de confianza
Muestreo aleatorio simple
Parámetro
Parámetro desviación típica (sigma)
Parámetro media (mu)
Parámetro proporción (pi)
Precisión de un IC
Tamaño de la muestra