Grado de Medicina – Universidad de Barcelona
Bioestadística básica, Epidemiologia y Introducción a la Investigación (2019/20)
Begoña Campos – Departamento de Fundamentos Clínicos

ESTIMACIÓN DE PARÁMETROS

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

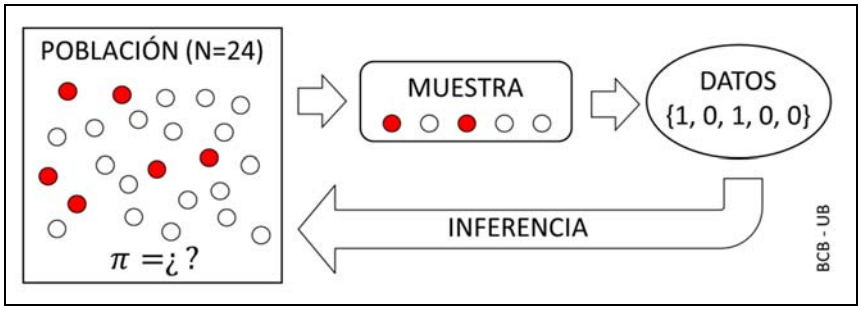
INTRODUCCIÓN

"After the visit he (Gosset) sent Fisher a copy of Student's tables, …
as you are the only man that's ever likely to use them"[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

[1] Box, JF. Gosset, Fisher and the t Distribution. The American Statistician, 1981, vol 35, n 2, 61-66.
[2] Spiegel, MR (2011). Ver bibliografía.

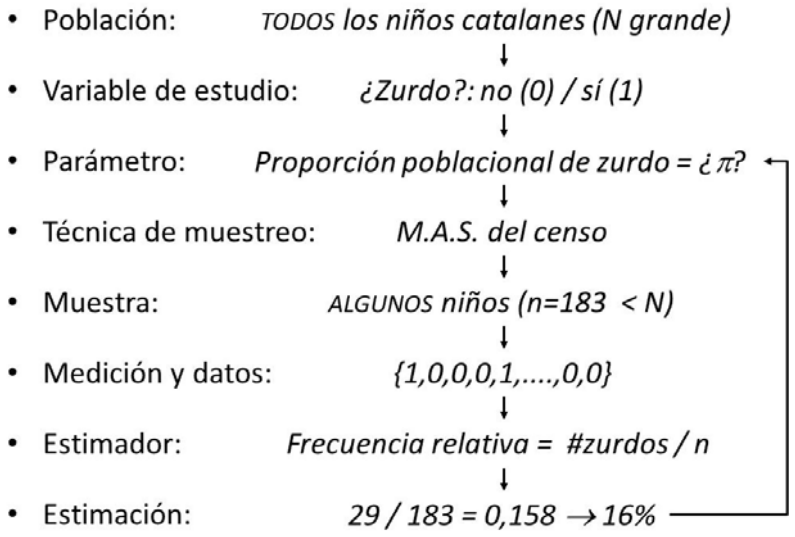
que la cuestión a resolver tenga que ver con el mundo empírico y sea necesario la realización de un estudio para recoger datos.

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 con cáncer?
- Las que buscan refutar o no una hipótesis: ¿Fumar causa cáncer 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:

- Cualitativa: X= ser zurdo (si/no)
- Cuantitativa: X= altura (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 no tienen por qué tener igual distribución, por ejemplo, albúmina sérica (normal) y urea sérica (no normal) en personas sanas[3].

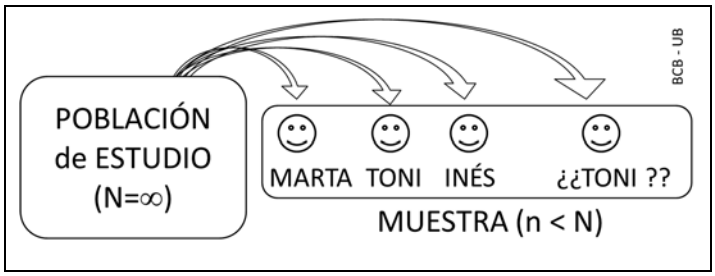
[3] Elveback, et al. Health, normality and the ghost of Gauss. JAMA, 1970, vol 211, n 1, 69-75.

K. PARÁMETRO es una característica numérica que resume el conjunto de todos los valores de una variable presentes en una población. Ejemplos:

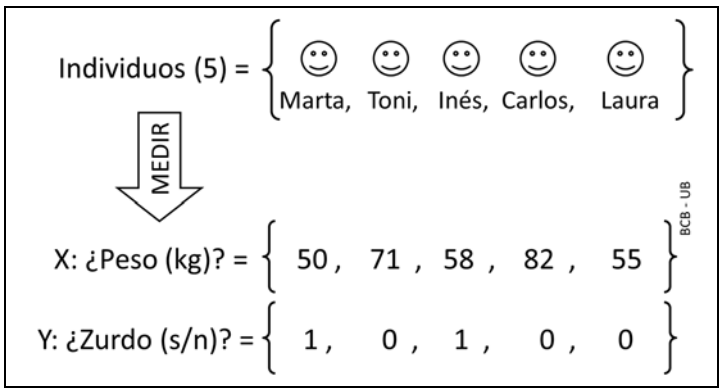
X= ser zurdo (si/no) → Proporción poblacional de zurdos (π)

X= altura (cm) → Media poblacional de altura (μ)

L. MUESTRA es un subconjunto de elementos de la población que han sido seleccionados con la finalidad de representarla. Una muestra aleatoria simple (M.A.S.) es la que se obtiene mediante un procedimiento que da a todos los elementos la misma probabilidad de pertenecer a la muestra.



M. La colección de DATOS se obtiene de medir la variable de interés en los individuos que componen la muestra



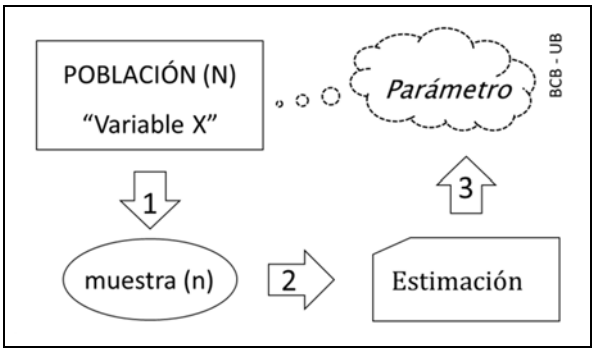
ESTIMACIÓN PUNTUAL y POR INTERVALO

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

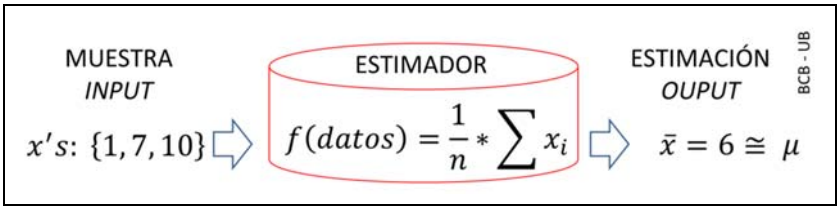
B. Un procedimiento de estimación es un intento para asignar valor a un parámetro a partir de los datos obtenidos mediante un muestreo aleatorio.

Población → Muestra de X → Parámetro (poblacional)

El objetivo es tener una alta probabilidad de que el resultado sea próximo al verdadero valor del parámetro, aunque no coincida exactamente con él.



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. Ya que los datos están sujetos a variabilidad, los estadísticos calculados a partir de ellos también están sujetos a variabilidad. Por ello se afirma que los estimadores se comportan como variables aleatorias y se caracterizan por una distribución muestral.



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 tipos de 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 particular se valorará que sea:

- Insesgado, es decir, con sesgo nulo. Esto significa que su distribución muestral está centrada en el valor del parámetro que se quiere estimar.
- Eficiente, es decir, con mínimo error cuadrático medio (MSE en inglés). Esto implica que su distribución muestral está muy concentrada en torno al centro.

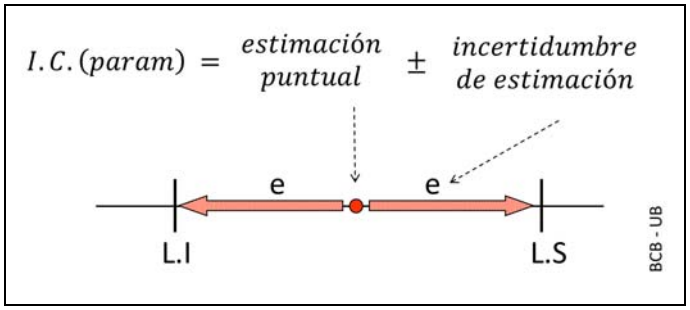
F. Estimadores importantes:

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

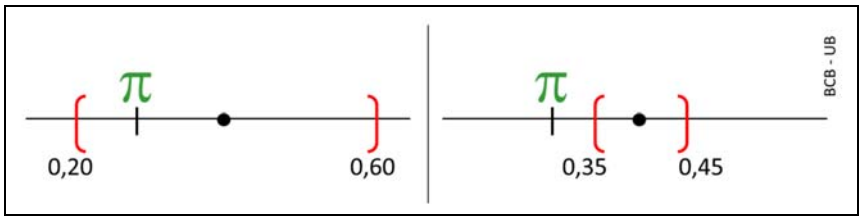
G. Tipos de Estimación:

- PUNTUAL. Asignar al parámetro el valor resultante de aplicar el estimador a la muestra. Se considera la mejor aproximación al verdadero valor del parámetro.
  muestra = {0, 1, 0, 1, 0, 1, 0, 0} → fr = 3/8= 0,375 $\cong \pi$
- 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. Ejemplo:
  muestra = {0, 1, 0, 1, 0, 1, 0, 0} → (LI a LS) = (0,0395 a 0,7105) $\ni \pi$

H. Aunque los intervalos pueden ser tanto bilaterales como unilaterales, lo más habitual es hacer la estimación con los bilaterales. Generalmente éstos son simétricos porque se construyen sumando y restando a la estimación puntual una cantidad determinada de incertidumbre (e), también denominada margen de error.



I. Si por azar la estimación puntual quedara lejos del valor del parámetro, entonces la estimación por intervalo será errónea, en el sentido de que el parámetro no estará 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 incluir el parámetro con la menor amplitud posible.



J. PRECISIÓN suele utilizarse como sinónimo de incertidumbre de estimación (e), y numéricamente correspondería a la semi-amplitud del intervalo, es decir la mitad de la distancia entre el límite inferior al límite superior.

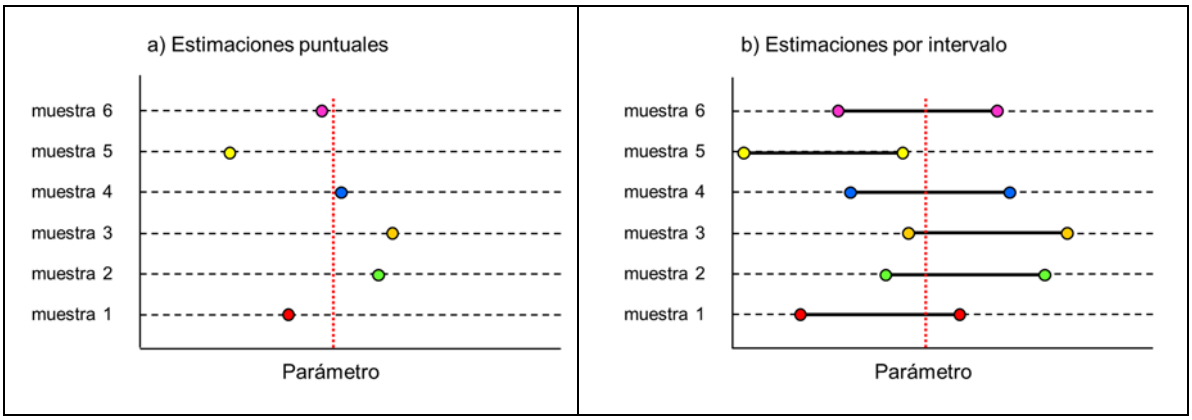
$$precisión = e = \frac{A}{2} = \frac{(LS - LI)}{2}$$

Sin embargo, cualitativamente se interpreta al revés, porque aumentar la precisión implica reducir la semi-amplitud:

- 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 acompañe a la estimación, menos precisa será.

K. CONFIANZA es un concepto de probabilidad que se aplica al conjunto de todas las estimaciones por intervalo que pueden resultar de infinitos muestreos de una misma población. Afirmar que la confianza es del 95% significa que el método de estimación es capaz de producir 95 intervalos que contienen el parámetro de un total de cien intervalos. En la siguientes dos figuras se representan las estimaciones obtenidas en seis muestras distintas: a) las estimaciones puntuales, b) las estimaciones por intervalo. La línea roja vertical marca en ambas gráficas la posición del parámetro. El conjunto de las estimaciones puntuales está centrado alrededor del parámetro, pero sólo dos de ellas, muestras 4 y 6, quedan muy próximas. Por el contrario, la mayoría de las estimaciones por intervalo consiguen incluir el parámetro dentro de sus límites, y solamente en un caso, muestra 5, 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 probabilidad de contener el parámetro.

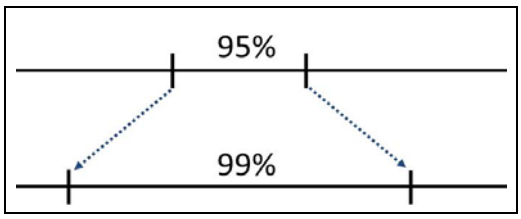


L. La incertidumbre (e) o precisión, ya sea interpretado en negativo o en positivo, es el resultado de combinar dos elementos que se obtienen de la distribución muestral del estimador. Generalmente:

$$\text{Incertidumbre} = k * \text{error típico}$$

- k es el valor de un percentil de la distribución del estimador, que dependerá de la confianza que se quiera atribuir al intervalo.
- Error típico de la distribución del estimador, que es inversamente proporcional al tamaño de la muestra

A mayor confianza demandada, mayor será el correspondiente percentil y por tanto mayor la amplitud. Por otro lado, a mayor tamaño de muestra, menor será el error típico, con lo cual la amplitud disminuirá.



M. Hacer una estimación por intervalo con una confianza total (100%) de incluir 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 el

riesgo de que el parámetro quede fuera de los límites. Este riesgo se denomina alfa y es complementario a la confianza:

Confianza = 100% *(1- Alfa)

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

INTERVALO DE CONFIANZA DE LA PROPORCIÓN

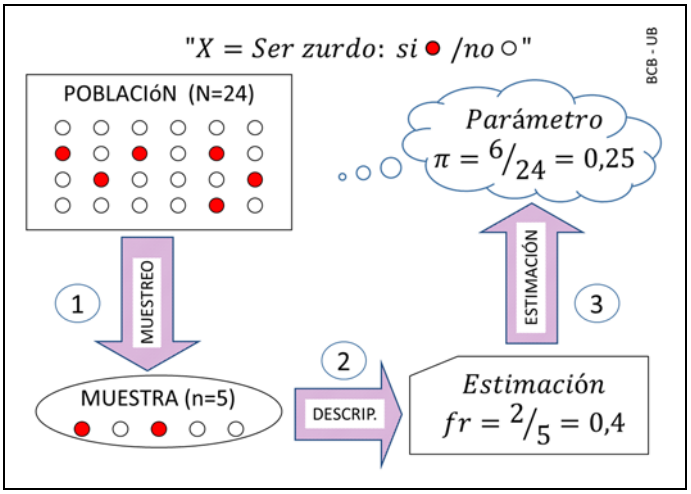
A. Plantear la estimación de una proporción poblacional (π) tiene sentido cuando la variable estudiada mide en un individuo la presencia de un atributo (éxito) o su ausencia (fracaso). Ejemplos: “ser zurdo”, “ser diabético”, “tener grupo sanguíneo A”.

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.

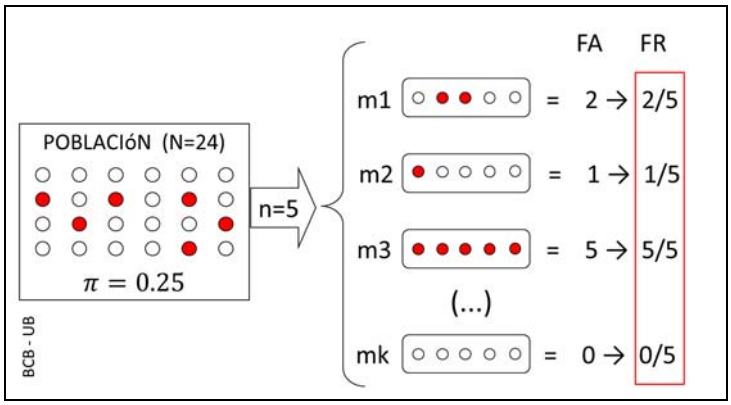
$$FR = \frac{\#exitos}{n} = \frac{1}{n} * FA \quad \rightarrow \quad P(éxito) = \pi$$

C. La estimación puntual de π consiste en asignarle el valor resultante de aplicar el estimador a la muestra, lo cual se etiquetará como fr minúscula.

muestra = {1, 1, 0, 0, 0) → FR → fr = 2/5= 0,4 → 0,4 ≅ π



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. La distribución de probabilidad de esta variable es una distribución muestral.



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)]:

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

La regla práctica para considerar válida esta aproximación es n ≥ 30 si π=0.5. Para otros valores de π habrá que compensar con mayores valores de n, requiriéndose entonces que tanto n*π como n*(1-π) sean mayores que 5.

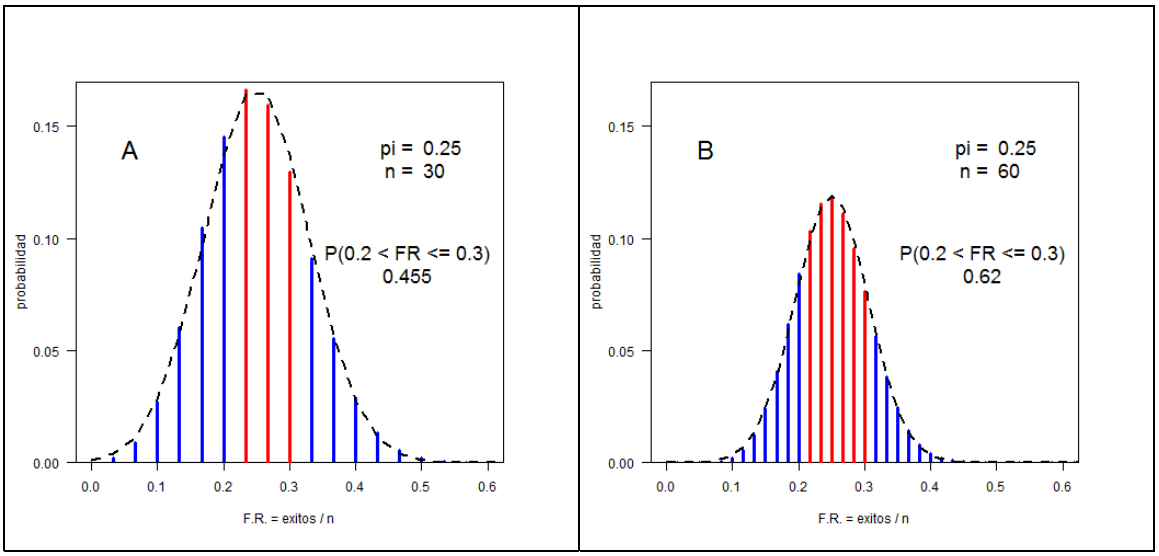
F. En muestras de gran tamaño, la distribución del estimador FR también es Normal, pues dividir por n es un re-escalamiento que no modifica el modelo, aunque sí el valor de los parámetros:

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

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

H. El error típico de FR es √[π∗(1-π)/n] que depende de n. Por tanto aumentar n reduce la dispersión de la distribución, condensándola en torno al centro que es π.

I. Esta última característica es muy importante porque determina la incertidumbre asociada al proceso de estimación. La figura siguiente ilustra la distribución de FR centrada en π=0,25 y con dos tamaños de muestra distintos: a) n=30, b) n=60. En el primer caso, n=30, la probabilidad de que la proporción muestral (FR) esté entre 0,2 y 0,3 vale 0,455. Doblando el tamaño de muestra, n=60, la probabilidad para el mismo intervalo es mayor, 0,62, es decir, se la probabilidad se ha concentrado en torno al valor del parámetro. Esto último significa que más estimaciones puntuales quedarán más cerca del parámetro.



J. Hay varias fórmulas para hacer la estimación por intervalo bilateral de una proporción, que se derivan de aplicar distintos métodos[4]. El método asintótico simple, también conocido como método de Wald, se basa en la aproximación normal dando lugar a la fórmula clásica del intervalo simétrico. El cálculo es fácil, pero el resultado es sólo una aproximación. Por el contrario, el método exacto de Clopper-Pearson se basa en la Binomial y da lugar a un intervalo asimétrico cuyo cálculo es más elaborado. Otro método es el "score" de Wilson que también aplica la aproximación normal, pero que genera intervalos asimétricos[5].

K. La fórmula de Wald, método asintótico simple, se deduce de la siguiente manera. Primero se fija un valor de confianza, por ejemplo 95%, para luego definir un intervalo (a, b) que bajo la distribución muestral del estimador FR contenga esa probabilidad[6]:

$$P(a < FR \leq b) = (1 - \alpha) = 0{,}95$$

El siguiente paso es asumir que la distribución de FR se puede aproximar por una normal. De esta manera se puede tipificar FR con la media (π) y el error típico correspondiente, ($\sqrt{[\pi*(1-\pi)/n]}$):

$$P\left(za < \frac{FR - \pi}{\sqrt{\pi(1-\pi)/n}} \leq zb\right) = 0{,}95$$

Dado que se busca un intervalo centrado, el valor de alfa se reparte equitativamente en los dos extremos de la distribución. En consecuencia, los extremos za y zb son los percentiles 0,025 y 0,975 de una normal zeta que valen -1,96 y +1,96 respectivamente.

A continuación hay que despejar el centro de la doble desigualdad para conseguir que el parámetro a estimar (π) quede aislado y encerrado entre dos extremos. Hacer bien esto implica resolver una ecuación cuadrática, pues el parámetro está dentro de la raíz cuadrada del denominador, además de estar en el numerador. La simplificación de este método consiste en manejar en bloque el denominador ($\sqrt{[\pi*(1-\pi)/n]}$):

$$P\left(FR - 1{,}96 * \sqrt{\pi(1-\pi)/n} < \pi \leq FR + 1{,}96 * \sqrt{\pi(1-\pi)/n}\right) = 0{,}95$$

[4] Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine, 1998, 17, 857-872.
[5] Pruim R (2011). Ver bibiografía.
[6] Rosner B (2011). Ver bibliografía.

para después sustituir el parámetro (π) por la estimación puntual (fr):

$$\sqrt{\pi(1-\pi)/n} \cong \sqrt{fr(1-fr)/n} = c \leq \sqrt{0{,}5/n}$$

El intervalo final que se obtiene se ajusta a la forma siguiente:

$$P(FR - 1{,}96c < \pi \leq FR + 1{,}96c) = 0{,}95$$

L. Es importante destacar que los extremos del intervalo anterior son aleatorios, mientras que π es una constante. Por ello, sólo antes de elegir la muestra tiene sentido afirmar que la probabilidad del intervalo de incluir el parámetro es del 95%. Después de conseguir la muestra ya no hay probabilidad involucrada[7].

M. El INTERVALO de CONFIANZA se obtiene después de seleccionar la muestra y calcular el extremo inferior y superior. Este es un intervalo particular de los muchos posibles, y contiene o no el parámetro:

$$fr \pm incertidumbre \;\rightarrow\; fr \pm [z * et] \;\rightarrow\; fr \;\pm \left[z_{(1-\alpha/2)} * \sqrt{\frac{fr*(1-fr)}{n}} \right]$$

donde fr es la estimación puntual y z es el percentil de una Normal (0,1) que acumula por debajo una probabilidad de (1-alfa/2).

N. 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!

INTERVALO DE CONFIANZA DE LA MEDIA y T-STUDENT

A. Plantear la estimación de una media poblacional (μ) tiene sentido cuando la variable que se estudia es de tipo cuantitativa. Ejemplos: “altura (cm)”, “concentración de colesterol (mg/dL)”, “producción diaria de orina (mL/dia)” o “consumo semanal de azúcar (gr/sem)”.

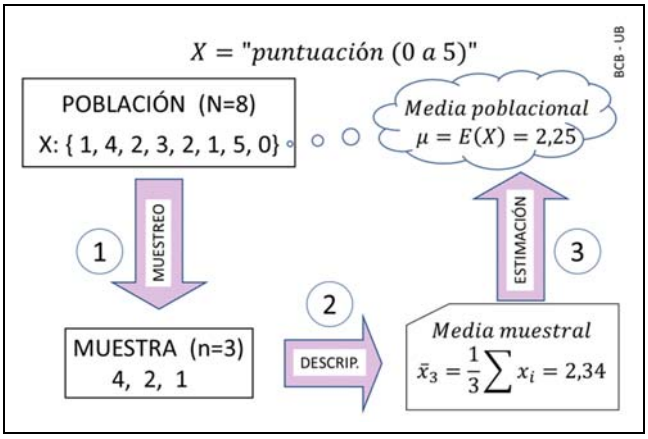
B. Se denomina MEDIA MUESTRAL a la media aritmética ($\bar{X}_n$) de los datos recogidos en una muestra de tamaño n. En caso de población infinita, selección de individuos mediante muestreo aleatorio simple y variable de estudio razonablemente simétrica, la media muestral es uno de los mejores estimadores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

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

[7] Pagano M, Gauvreau K (2011). Ver bibliografía.

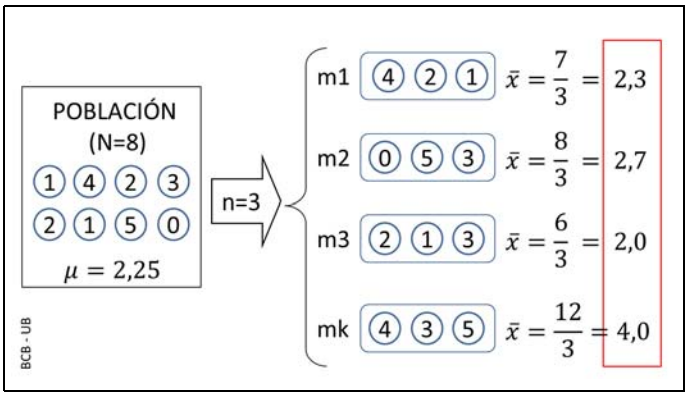
C. La estimación puntual de μ consiste en asignarle el valor resultante de aplicar el estimador a la muestra, lo cual se etiquetará como x-barra minúscula.

muestra = {4, 2, 1) → $\bar{X}_n$ → $\bar{x}$ = 7/3= 2,34 → 2,34 ≅ μ



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

E. El estimador $\bar{X}_n$ se comporta como variable aleatoria porque el resultado cambia de muestra en muestra. La distribución de probabilidad de $\bar{X}_n$ es una distribución muestral.



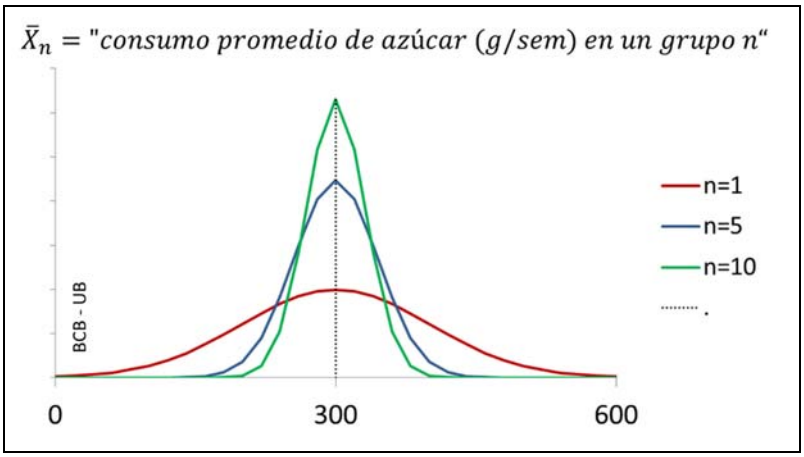
F. La distribución de la variable "media muestral" depende de la distribución de la variable original X que se mide en los individuos. Existen dos teoremas que pueden aplicarse:

- 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

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



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

H. El error típico de la media muestral es [$\sigma/\sqrt{n}$] que depende de n. En consecuencia, la incertidumbre respecto al valor del parámetro se puede reducir aumentando el tamaño de la muestra.

I. La fórmula clásica para hacer la estimación por intervalo bilateral parte de fijar un valor de confianza, por ejemplo 95%, para luego definir un intervalo (a, b) que contenga esa probabilidad:

$$P(a < \bar{X}_n \leq b) = (1 - \alpha) = 0{,}95$$

Si por aplicación del teorema de la adición o del teorema del límite central se puede asumir que la media muestral sigue un modelo normal, entonces por tipificación:

$$P\left(za < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq zb\right) = 0{,}95$$

Y dejando el parámetro aislado en el centro se llega a:

$$P\left(\bar{X}_n - 1{,}96 * {}^{\sigma}/_{\sqrt{n}} < \mu \leq \bar{X}_n + 1{,}96 * {}^{\sigma}/_{\sqrt{n}}\right) = 0{,}95$$

J. Es importante destacar que los extremos del intervalo anterior son aleatorios, mientras que μ es una constante. Por tanto sólo antes de elegir la muestra tiene sentido afirmar que la probabilidad del intervalo de incluir el parámetro es del 95%. Después de conseguir la muestra ya no hay probabilidad involucrada.

K. El INTERVALO de CONFIANZA se obtiene después de seleccionar la muestra y calcular el extremo inferior y superior. Este es un intervalo particular de los muchos posibles, y contiene o no el parámetro:

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

Donde $\bar{x}$ es la estimación puntual y z es el percentil de una Normal (0,1) que acumula por debajo una probabilidad de (1-alfa/2).

L. 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 usar un estimador que de una buena estimación puntual de σ .

M. La DESVIACIÓN TÍPICA MUESTRAL CORREGIDA (S_{n-1}) es una medida de dispersión de las observaciones de una muestra de tamaño n. Se denomina corregida porque resulta de multiplicar la desviación típica sin corregir (S_n) por un factor. En términos cuadráticos (variancias) la expresión es la siguiente:

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

Si n es muy grande, el cociente (n/n-1) vale prácticamente uno y la corrección es mínima. Sin embargo, para "enes" pequeñas el efecto de la corrección puede ser importante. Por ejemplo, si n=6 entonces hay que multiplicar por 1,2 lo que significa inflar la medida de dispersión en un 20%. La variancia corregida será siempre numéricamente mayor que la sin corregir.

N. 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" dividida por la variancia poblacional (σ^2) y multiplicada por (n-1) sigue un modelo de distribución Ji-cuadrado con su parámetro "grados de libertad" (gl) igual a (n-1):

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

La justificación del factor de corrección es desplazar la distribución muestral para conseguir que el estimador S_{n-1} sea insesgado. Por el contrario, el estimador sin corregir es sesgado y por tanto no es un buen estimador.

Ñ. La transformación de la media muestral, $\bar{X}_n$, usando la desviación típica muestral corregida, S_{n-1} , en lugar de la desviación poblacional, σ, da lugar a una nueva variable que no tiene distribución normal, sino t de Student. Esto es debido a que S_{n-1} es una variable aleatoria y no una constante.

$$\frac{\bar{X}_n - \mu}{S_{n-1}/\sqrt{n}} \sim t - Student$$

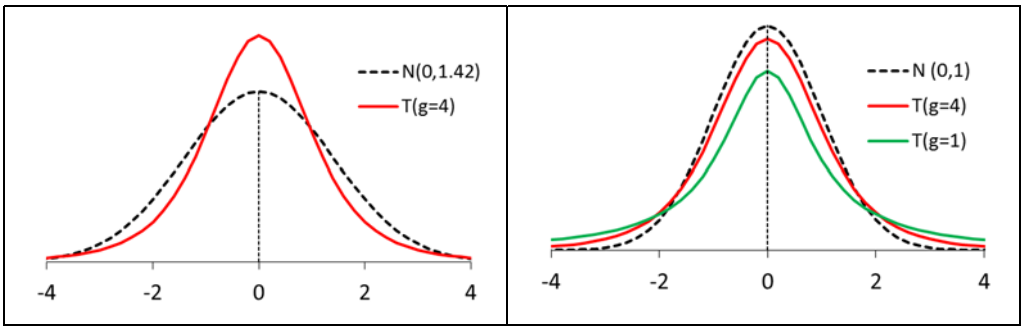
O. Ejemplo numérico:

				mu=2,25	sigma = 1,56			
	X1	X2	X3	media	dt(n)	dt(n-1)	z	t
m1	4	2	1	2,3	1,2	1,5	0,053	0,055
m2	0	5	3	2,7	2,1	2,5	0,267	0,166
m3	2	1	3	2,0	0,8	1,0	-0,160	-0,250
mk	4	3	5	4,0	0,8	1,0	1,122	1,750

P. La función de densidad de una distribución T de STUDENT es muy similar a una campana de Gauss. Es simétrica con un pico en el centro que está siempre en cero. Sólo tiene un parámetro que se llama "grados de libertad" (gl) y determina la variancia:

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

Esta distribución es leptocúrtica (coeficiente de curtosis positivo) cuando se compara con una distribución Normal de media cero e igual dispersión (ver gráfica).

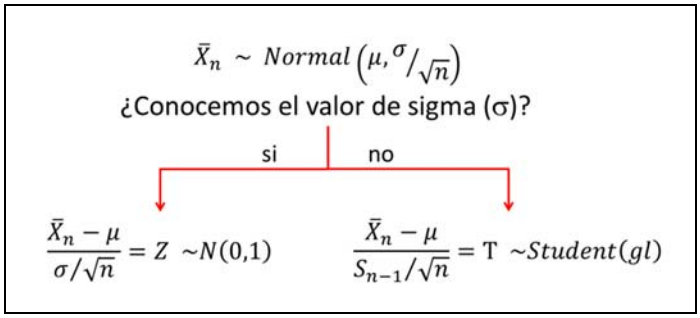


El parámetro gl vale (n-1) cuando se utiliza para la estimación de una media. Por tanto, si el tamaño de muestra es 3, los grados de libertad de la t serán 2. En la siguiente tabla se presentan los valores de t correspondientes al percentil 97,5% para distintos valores grados de libertad:

	2	3	5	11	19	29	49	79	99
t(0,975) =	4,303	3,182	2,571	2,201	2,093	2,045	2,010	1,990	1,984

A medida que aumentan los grados de libertad, es decir el tamaño de muestra, la variancia va disminuyendo y la curva de la t de Student se acerca a la curva de una Zeta, es decir, Normal (0,1) (ver gráfica). Con 200 grados de libertad el percentil t(0,975) vale 1,97190 y con 500 vale 1,96472.

Q. En resumen



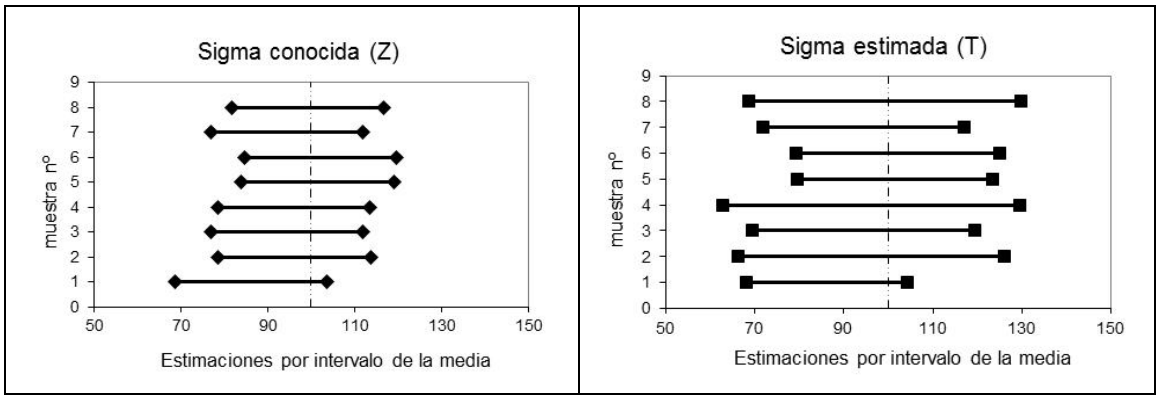
R. En consecuencia, para calcular una estimación por intervalo del parámetro media poblacional hay que usar los percentiles de la distribución t de Student en lugar de una Normal (0,1) siempre que se use una estimación de sigma. La fórmula correcta es:

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

S. Utilizar la distribución 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 (ver ejemplo numérico anterior), b) los intervalos con T suelen ser más anchos, porque el percentil t es mayor que z para una misma confianza:

percentil 80% (Z) = 0.842 < percentil 80% (T-4gl) = 0.941



T. Nota histórica. El primer artículo que abordó el problema de sustituir el parámetro del cual se derivó la distribución t de Student fue:

Student. The probable error of a mean. Biometrika, 1908, vol 6, n.1, pp.1-25[8]

Su autor fue William Sealy Gosset, un científico inglés que desde 1899 trabajaba en Dublín para la cervecera Guinness. Usó un pseudónimo, Student, porque la empresa no permitía a sus empleados publicar con sus propios nombres. Sin embargo, fue un permiso de la empresa lo que permitió a Gosset visitar al laboratorio de biometría de Karl Pearson en la University College London y trabajar con él en el tema que le preocupaba[9]. El objetivo de su estudio era adaptar la teoría estadística al análisis de situaciones experimentales en las que había que usar un estimador del parámetro desviación típica y las muestras eran de tamaño pequeño. La segunda conclusión del artículo decía lo siguiente:

"*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.*"

El impacto de este artículo en la estadística moderna se debe principalmente al reconocimiento que le dio Sir Ronald A. Fisher. La buena interacción entre ambos hizo que Fisher formalizara parte de las demostraciones, sustituyera el estimador sesgado por el insesgado y expandiera su uso al caso de dos muestras. El nombre "t de Student" para la distribución que hoy usamos fue acordado por ambos. Estos resultados quedaron recogidos en "Statistical Methods for Resarch Workers" que Fisher publicó en 1925 y que fue un libro de texto rompedor[10].

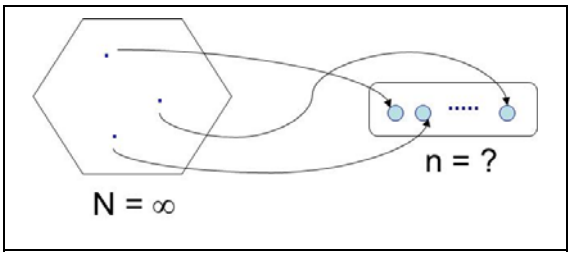
[8] Zabell, SL. On Student's 1908 Article "The Probable Error of a Mean". Journal of the American Statistical Association, 2008, vol 103, n 481, 1-20

[9] Boland, PJ. A Biographical Glimpse of William Sealy Gosset. The American Statistician, 1984, vol 38, n 3, 179-183

[10] Stigler SM. The seven pillars of statistical wisdom. Harvard University Press, 2016.

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"*[11] . 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:
- Identificar el objetivo del estudio y el diseño para seleccionar la fórmula apropiada
- Decidir los valores de entrada de la fórmula
- Realizar los cálculos
- 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. Un problema con la fórmula anterior es que antes de hacer el estudio no se dispone de una estimación puntual para cambiar por π, ¿qué hacer entonces?

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

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

[11] 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:

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

H. 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.
- Pagano M, Gauvreau K. Bioestadística. Thomson Learning, 2011. 2ª ed.
- Pruim R. Foundations and Applications of Statistics: an introduction using R. American Mathematical Society, 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