# Biometría

Estimación de parámetros

# Estimación

- Las poblaciones son descriptas mediante sus parámetros
  - Para variables cuantitativas, las poblaciones son descriptas mediante $\mu$ y $\sigma$.
  - Para variables cualitativas, las poblaciones son descriptas mediante $p$.
- Si los valores de los parámetros son desconocidos, podemos estimarlos en base a muestras y esperamos que sean una buena aproximación al valor exacto

# Definiciones

- **estimación puntual:** se calcula un valor simple a partir de la muestra a fin de estimar el parámetro

- **estimación por intervalo de confianza:** se calculan dos números para crear un rango de valores que se espera contenga al parámetro con una cierta probabilidad o nivel de confianza

$$P(LI < \theta < LS) = 1 - \alpha$$

# ¿Qué tan buena es la estimación?

## Error muestral

- es la distancia entre el estimador puntual y el verdadero valor del parámetro
- Es el error que surge por estudiar a una parte de la población
- Posee las mismas unidades que la variable en estudio
- Su magnitud es desconocida y por lo tanto imposible de calcular con certeza
- Se sabe que disminuye cuando aumenta el tamaño de la muestra
- Si la muestra está diseñada de forma probabilística es posible controlar su magnitud y dar una estimación del mismo
- Pero para eso es necesario conocer la distribución de probabilidades (**distribución muestral**) del estimador

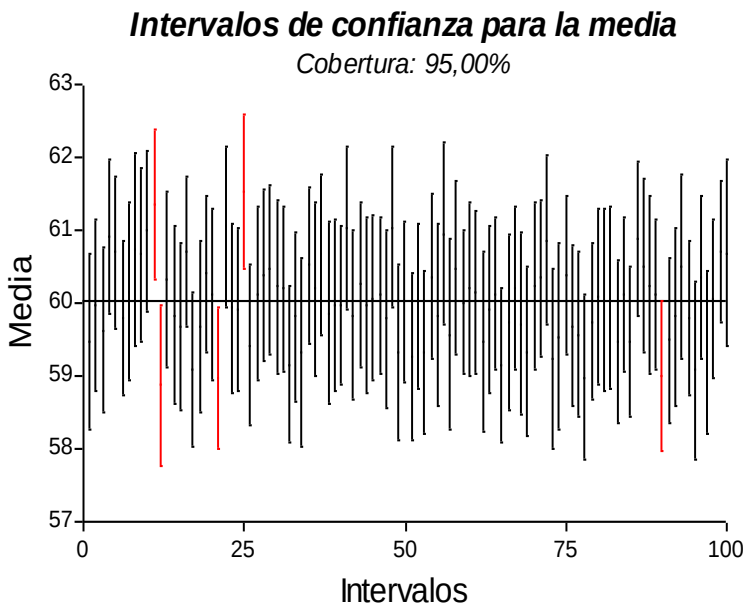
# ¿Qué son los errores no muestrales?

- Otros errores ajenos al muestreo: no respuesta, codificación, encuestador, encuestado, lógicos, de concepción, etc.
- No disminuyen cuando el tamaño de la muestra aumenta
- muy pero muy difíciles de medir!!!

# ¿Qué tan buena es la estimación?

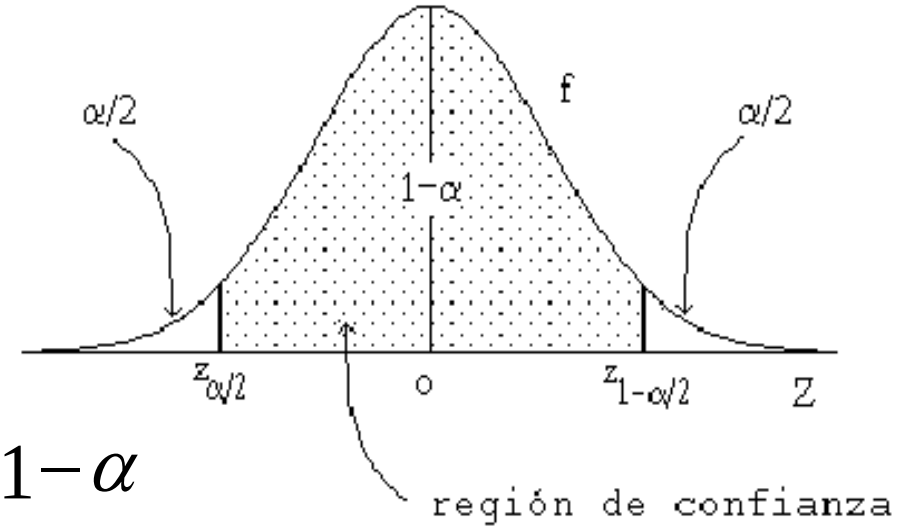
## Nivel de confianza

- es la probabilidad de que el intervalo contenga al parámetro
- Se lo simboliza como $1-\alpha$
- Lo fija el investigador. Valores típicos de $1-\alpha$ =0,90 ; **0,95** ; 0,99
- α es la probabilidad de error (no contener al parámetro) y se la denomina también riesgo
- Es el porcentaje de intervalos que se espera contengan al parámetro (para ese tamaño de muestra)



# ¿Cómo calcular el error muestral en la estimación de μ (siendo σ conocido)?

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




$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

$$P(z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < z_{1-\alpha/2}) = 1 - \alpha$$

$$P(z_{\alpha/2}\, \sigma / \sqrt{n} < \bar{x} - \mu < z_{1-\alpha/2}\, \sigma / \sqrt{n}) = 1 - \alpha$$

***EM***

# ¿Entre qué valores esperaría que se encuentre μ? Intervalo de confianza para μ

$$P(z_{\alpha/2}\,\sigma/\sqrt{n} < \bar{x} - \mu < z_{1-\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

$$P(\bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

***LI*** ***LS***

$$P(LI < \mu < LS) = 1-\alpha$$

$$\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$$

$$\bar{x} \pm EM$$

# ¿Cómo mejorar la estimación?

TABLA 4

Para disminuir el error muestral (mayor precisión):

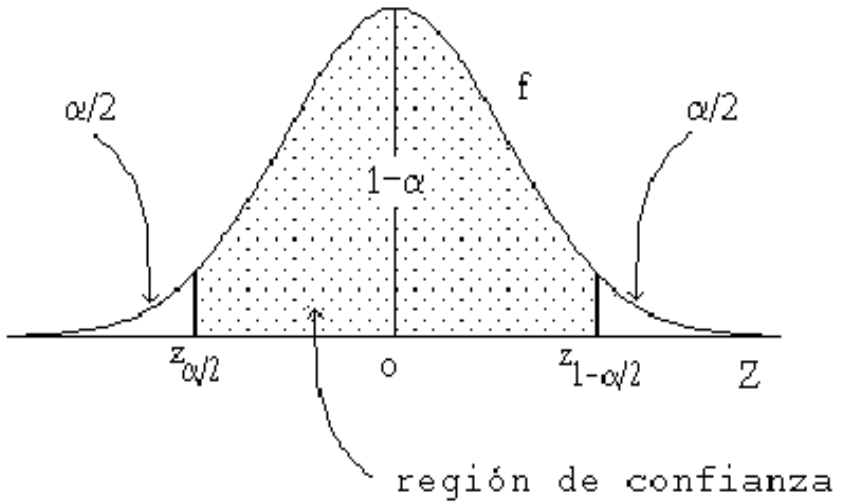
- Tamaño de la muestra
- Nivel de confianza
- Desvío estándar

| Nivel de confianza | $z_{1-\alpha/2}$ |
|---|---|
| 0.90 | 1.645 |
| 0.95 | 1.96 |
| 0.99 | 2.576 |

$$P(LI < \mu < LS) = 1 - \alpha$$

$$\bar{x} \pm z_{\alpha/2}\, \sigma / \sqrt{n}$$

$$\bar{x} \pm EM$$



# ¿De qué depende el tamaño de una muestra?

- De los recursos y del presupuesto: $$$$.
- Del tipo de población en estudio.
- De la variable a estudiar (cuali o cuantitativa).
- Del grado de  homogeneidad de ésta en la población.
- Del diseño muestral empleado.

# ¿Qué se necesita para determinar el tamaño de una muestra para un promedio?

Tres elementos importantes:

1. Error muestral o margen de error deseado.
2. Nivel de Confianza o de Riesgo, y el valor del fractil de la distribución asociada a alguno de ellos.
3. Una magnitud de la dispersión o del grado de heterogeneidad de la variable a estudiar.

$$EM = z_{\alpha/2}\,\sigma/\sqrt{n} \quad \Rightarrow \quad n = \left(\frac{z_{\alpha/2}\,\sigma}{EM}\right)^2$$

2. 3. 1.

# Supuestos

Para que las estimaciones sean confiables se debe cumplir:

- Muestreo **aleatorio** probabilístico
- Muestreo con reposición o bien $n/N < 0.05$
- La variable *x* debe tener distribución **normal**; en caso contrario, el tamaño de la muestra debe ser lo **suficientemente grande** ($n \geq 30$)
- El desvío estándar poblacional $\sigma$ debe ser **conocido**

# Intervalos de confianza: un ejemplo



UNIVERSIDAD DE ANTIOQUIA

**Indicadores bioquímicos sanguíneos en ganado de lidia mantenido en pastoreo en la cordillera central colombiana**

Revista Colombiana de Ciencias Pecuarias

David Jordán[1], MVZ; Néstor A Villa[1], MVZ, MSc; Miguel Gutiérrez[2], MVZ, MSc; Ángela B Gallego[1], MVZ; Gustavo A Ochoa[1], MVZ; Alejandro Ceballos[1], MVZ, MSc
[1]Grupo: Salud Productiva en Bovinos, Porcinos y Equinos. Universidad de Caldas, A.A. 275. Manizales, Colombia.
[2]Ganadería "Ernesto Gutiérrez", Edificio El Castillo Of. 703, Manizales, Colombia.
aleceballos@ucaldas.edu.co

**Tabla 2.** Valor promedio, desviación estándar (DE), rango, intervalo de confianza (IC) y coeficiente de variación (CV) para las variables séricas analizadas en bovinos de lidia (Grupo 1, n=39) mantenidos en pastoreo en la cordillera central colombiana.

| Variable | ±DE | Rango | IC (95%) | CV% |
|---|---|---|---|---|
| Glucosa (mmol/L) | 6.4±2.1 | 3.8 – 14.7 | 5.5 – 7.0 | 33 |
| Colesterol (mmol/L) | 2.9±0.5 | 1.6 – 4.4 | 2.7 – 3.0 | 18 |
| ß-OHB (mmol/L) | 0.15±0.08 | 0.02 – 0.33 | 0.12 – 0.18 | 53 |
| TAG (mmol/L) | 0.31±0.11 | 0.06 – 0.68 | 0.27 – 0.35 | 38 |
| Proteínas totales (g/L) | 89±14 | 55 – 128 | 82 – 91 | 17 |
| Albúmina (g/L) | 31±4 | 26 – 44 | 30 – 33 | 14 |
| Globulinas (g/L) | 55±13 | 28 – 84 | 51 – 59 | 23 |
| Relación A/G | 0.6±0.2 | 0.4 – 1.2 | 0.5 – 0.7 | 29 |
| Urea (mmol/L) | 6.9±1.5 | 5.2 – 10.3 | 6.2 – 7.3 | 21 |

*Glucosa:* $IC_{95}$ : 5.5-7.0 *mmol/L*

1. El 95% de los ejemplares tiene entre 5.5-7.0 *mmol/L*.
2. La glucosa promedio de los 39 ejemplares se encuentra entre 5.5-7.0 *mmol/L*
3. El promedio de la especie se encuentra entre 5.5-7.0 *mmol/L*.

# Estimación de un promedio con desvío poblacional desconocido

- Es la situación más habitual

- El hecho de desconocer el valor paramétrico de $\sigma$ tiene un costo: se debe utilizar la distribución ***t de Student***, que posee mayor dispersión que la normal estándar

# Intervalo de confianza para $\mu$ cuando el desvío poblacional $\sigma$ es desconocido

- Con σ conocido

$$P(\bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

- Con σ desconocido

$$P(\bar{x} + t_{n-1,\alpha/2}\,s/\sqrt{n} < \mu < \bar{x} + t_{n-1,1-\alpha/2}\,s/\sqrt{n}) = 1-\alpha$$

***LI*** ***LS***

$$P(LI < \mu < LS) = 1-\alpha$$

$$\bar{x} \pm t_{n-1,\alpha/2}\,s/\sqrt{n}$$

$$\bar{x} \pm EM$$

TABLA 4

# ¿Cómo mejorar la estimación?

Para disminuir el error muestral (mayor precisión):

- Tamaño de la muestra
- Nivel de confianza
- Desvío estándar

$$EM = t_{n-1,\alpha/2}\, s/\sqrt{n} \quad \Rightarrow \quad n = \left[\frac{t_{n-1,\alpha/2}\, s}{EM}\right]^2$$

Como el *n* está a ambos lados de la ecuación, se debe utilizar un **método iterativo** para calcular el tamaño muestral

# Supuestos

Para que las estimaciones sean confiables se debe cumplir:

- Muestreo **aleatorio** probabilístico
- Muestreo con reposición o bien $n/N < 0.05$
- La variable *x* debe tener distribución **normal**; en caso contrario, el tamaño de la muestra debe ser lo **suficientemente grande** ($n \geq 30$)

# Prevalencia de la diabetes mellitus no dependiente de la insulina en Lejona (Vizcaya)

| | Diabéticos (grupo DM) | |
|---|---|---|
| | Varones | Mujeres |
| Número de individuos | 15 | 16 |
| Edad (años) | 61,0 ± 11,8 | 61,7 ± 13,1[a] |
| Talla (cm) | 164,5 ± 6,3 | 151,6 ± 7,1[b] |
| Peso (kg) | 75,8 ± 14,8 | 67,2 ± 9,9[a] |
| IMC (kg/m$^2$) | 27,9 ± 4,5 | 29,1 ± 3,7[a] |
| PAS (mmHg) | 148,6 ± 24 | 158,7 ± 29[a] |
| PAD (mmHg) | 82 ± 10,6 | 87,5 ± 14,4[a] |
| Antecedentes familiares (%) | 18,8 | 23,5[a] |

Estimar con una confianza del 95% la presión arterial sistólica de diabéticos

# Estimación de una proporción
## Un ejemplo

Las aves parásitas de cría como el tordo renegrido, *Molothrus bonaerensis*, depositan sus huevos en nidos de otras especies que proveen la totalidad del cuidado parental. Se llevó a cabo un estudio en la prov. de Mendoza durante el mes de octubre a fin de estimar la incidencia de parasitismo en el zorzal chalchalero, *Turdus amaurochalinus*. Se visitaron 108 nidos, observándose 72 parasitados.

- Población
- Muestra
- Tipo de muestreo
- Individuo
- Parámetro
- Estimador

# Distribución muestral de $\hat{p}$

Si de una población con cierta proporción de éxitos $p$ se extraen infinitas muestras aleatorias de tamaño $n$ y a cada una de ellas se le calcula la **proporción muestral** $\hat{p}$ , se demuestra que esta se comporta según una distribución **normal** siempre y cuando se cumplan las condiciones de aproximación de la distribución binomial a la normal, es decir:

$$n > 30, pn > 5 \; y \; qn > 5$$

# Distribución muestral de $\hat{p}$

1. La media de $\hat{p}$ es: $p$
2. El desvío estándar (ES) de $\hat{p}$ es: $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$
3. Si el tamaño de la muestra es lo suficientemente grande, $pn > 5$ $y$ $qn > 5$, la distribución de $\hat{p}$ es **normal**
   Por lo tanto es posible calcular probabilidades utilizando:

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

# Intervalo de confianza para ***p***

- Para μ con σ conocido

$$P(\bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

- Para $p$

$$P(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}) = 1-\alpha$$

***LI*** ***LS***

$$P(LI < p < LS) = 1-\alpha$$

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad \hat{p} \pm EM$$

# Prevalencia de hipertensión arterial y factores asociados en la población rural marginada

Jesús Fernando Guerrero-Romero, M.C.,[1] Martha Rodríguez-Morán M.C., M. en C.[1]

*salud pública de méxico / vol.40, no.4, julio-agosto de 1998*

**Objetivo.** Determinar la prevalencia y los factores asociados a la hipertensión arterial sistémica (HAS) en la población rural marginada de Durango, México. **Material y métodos.** Se realizó un estudio transversal comparativo en 627 comunidades rurales, de las que aproximadamente 90% tiene 250 o menos habitantes. Se determinaron las cifras de presión arterial y las variables sociodemográficas. **Resultados.** Se estudiaron 5 802 sujetos, es decir, 4 452 mujeres (76.7%) y 1 350 hombres (23.3%). Se identificó HAS en 1 271 individuos (21.9%; IC95% 20.8-23.0), de los cuales 1 011 eran mujeres (22.71%; IC95% 21.5-23.9), y 260, hombres (19.26%; IC95% 17.2-21.4). Del total de la población blanco, 3 018

# ¿Cómo mejorar la estimación?

Para disminuir el error muestral (mayor precisión):

- Tamaño de la muestra
- Nivel de confianza

$$EM = z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \Rightarrow \quad n = \frac{z_{1-\alpha/2}^{\ 2}\,\hat{p}\hat{q}}{EM^{2}}$$

Si no existe muestreo previo, se asume ***p* = 0.5**

| p | q | pxq |
|---|---|---|
| 0,9 | 0,1 | 0,09 |
| 0,8 | 0,2 | 0,16 |
| 0,7 | 0,3 | 0,21 |
| 0,6 | 0,4 | 0,24 |
| **0,5** | **0,5** | **0,25** |
| 0,4 | 0,6 | 0,24 |
| 0,3 | 0,7 | 0,21 |
| 0,2 | 0,8 | 0,16 |
| 0,1 | 0,9 | 0,09 |

# Supuestos

Para que las estimaciones sean confiables se debe cumplir:

- Muestreo **aleatorio** probabilístico
- Muestreo con reposición o bien n/N < 0.05
- Para que sea válida la aproximación a la **normal** el tamaño de la muestra debe ser lo suficientemente grande (n≥30), $pn > 5$ $y$ $qn > 5$

# En resumen:

$$IC_{1-\alpha} : \hat{\theta} \pm EM$$

$$IC_{1-\alpha} : \hat{\theta} \pm P_{1-\alpha/2} ES_{\hat{\theta}}$$

$$\bar{x} \pm z\, ES_{\bar{x}} \qquad \bar{x} \pm t\, ES_{\bar{x}} \qquad \hat{p} \pm z\, ES_{\hat{p}}$$

$$\bar{x} \pm z\, \sigma / \sqrt{n} \qquad \bar{x} \pm t\ s / \sqrt{n} \qquad \hat{p} \pm z \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Todos los EM son proporcionales a √n ⇨ para reducir un IC a la mitad, se debe cuadriplicar el tamaño de la muestra

# Estimación de la variabilidad
## Un ejemplo

Se desea estimar la variabilidad en la concentración de hemoglobina en jugadores de fútbol profesionales. Una muestra aleatoria de 9 jugadores arrojó los siguientes valores (en g/dl):

15.3 16.0 14.4 16.2 16.2 14.9 15.7 15.3 14.6

- Población
- Muestra
- Tipo de muestreo
- Individuo
- Parámetro
- Estimador

# Distribución muestral

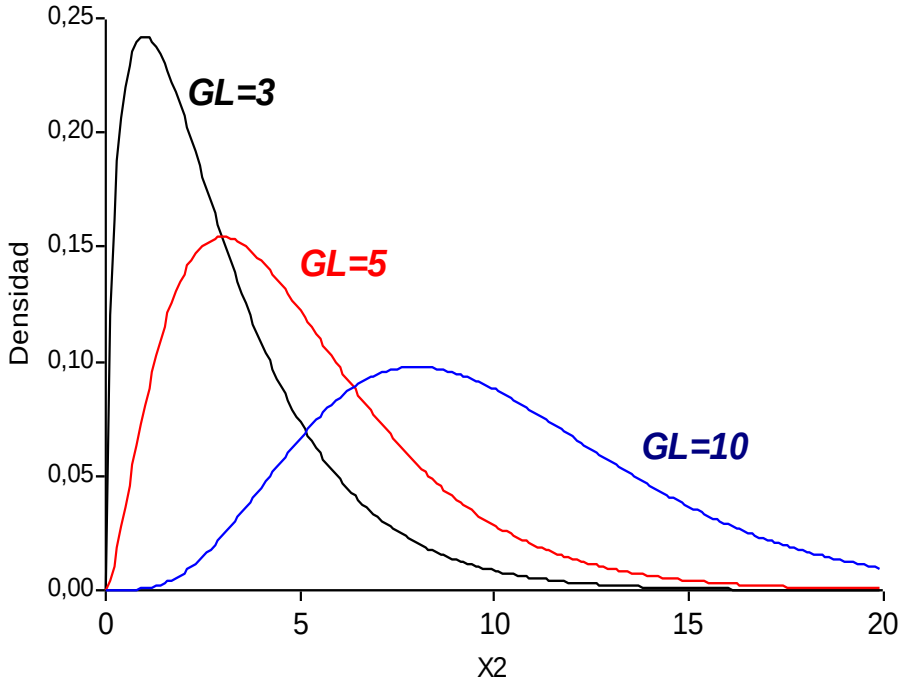
Si de una población **con distribución normal** se extraen infinitas muestras aleatorias de tamaño $n$ y a cada una de ellas se le calcula la varianza muestral $s^2$ , se demuestra que el estadístico

$$\frac{(n-1)s^2}{\sigma^2}$$

se comporta según una distribución **chi-cuadrado ($\chi^2$)** con $n-1$ grados de libertad

# Distribución chi-cuadrado ($\chi^2$)

- Es una distribución asimétrica positiva
- Solo toma valores positivos, es decir que $\chi^2 \geq 0$
- No se trata de una única curva, sino de infinitas curvas, cada una caracterizada por un parámetro denominado **grados de libertad (GL)**
- Los GL dependen del tamaño de la muestra
- A medida que aumentan los GL la distribución tiende a hacerse simétrica



# Intervalo de confianza para la varianza $\sigma^2$

$$\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}$$

***LI*** ***LS***

TABLA 5

$$P(LI < \sigma^2 < LS) = 1-\alpha$$

- Para el desvío estándar se debe aplicar raíz cuadrada
- Observar que los límites del intervalo no son simétricos con respecto al estimador

# Supuestos

Para que las estimaciones sean confiables se debe cumplir:

- Muestreo **aleatorio** probabilístico
- Muestreo con reposición o bien $n/N < 0.05$
- La variable debe seguir una distribución **normal**

# Conociendo la distribución muestral de un estimador se puede construir un IC para el parámetro

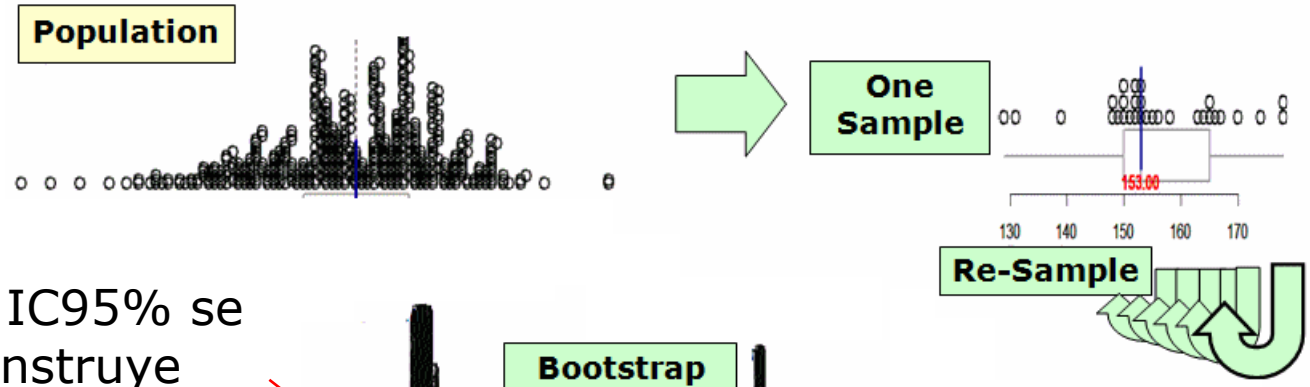
| | Estimador | Esperanza | Desvío estándar (error estándar) | Estadístico |
|---|---|---|---|---|
| 12 | $\bar{x}$ | $\mu$ | $\sigma/\sqrt{n}$ | $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ |
| 16 | CV | $CV$ | $\frac{CV}{\sqrt{2n}}\sqrt{1+2\left[\frac{CV}{100}\right]^2} \approx \sqrt{2n}$ | $t_{n-1} = \frac{\hat{C}V - CV}{\sqrt{2n}}$ |
| 17 | $g_1$ | $\gamma_1$ | $\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}$ | $t_{n-1} = \frac{g_1 - \gamma_1}{\sqrt{\frac{6}{n}}}$ si n > 100 |
| 18 | $g_2$ | $\gamma_2$ | $\sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} \approx \sqrt{\frac{24}{n}}$ | $t_{n-1} = \frac{g_2 - \gamma_2}{\sqrt{\frac{24}{n}}}$ si n > 100 |
| 19 | $\bar{x}_1 - \bar{x}_2$ | $\mu_1$-$\mu_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ |

# Estimación por Bootstrap

- Cuando no se conoce la distribución teórica del estimador
- Es el caso de la mediana, de muchos índices en biología que surgen de funciones complejas (diversidad, árboles filogenéticos, etc)
- Solo se cuenta con datos muestrales (n). Consideramos que su distribución constituye una buena aproximación a la distribución real de la variable
- Entonces aproximamos la distribución muestral mediante la simulación de experimentos repetidos sobre nuestros datos muestrales
- Mediante la simulación podemos obtener EE, predecir sesgo e incluso comparar varias formas de estimar el mismo parámetro
- El único requisito es que los datos hayan sido independientemente muestreados de una única distribución

# Bootstrap: Procedimiento

- Se extraen muchas muestras **con reposición** (i.e. 1000) de tamaño n de la muestra  original (se "re-muestrea")
- En cada muestra se calcula el estimador de interés
- Con esos valores se construye la distribución muestral
- El estimador por boostrap del parámetro será la media de dicha distribución y su error estándar sería el desvío estándar de la distribución




El IC95% se construye con los percentiles 2,5 y 97,5