

Biometría



Distribuciones muestrales

¿Qué tienen en común estos ejemplos?

Banco de semillas de un pastizal uruguayo bajo diferentes condiciones de pastoreo

FEDERICO HARETCHE & CLAUDIA RODRÍGUEZ ✉

Ecología Austral 16:000-000. Diciembre 2006

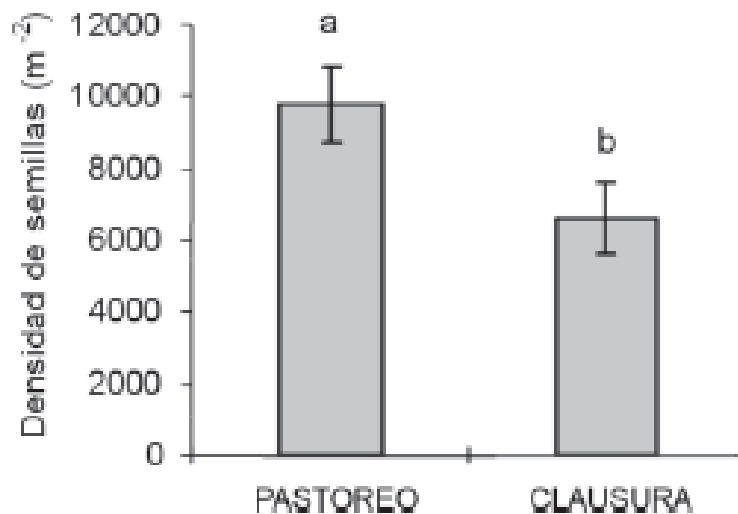


Figura 1. Densidad del banco de semillas (semillas/m² ± EE) de un pastizal uruguayo bajo condiciones de pastoreo y exclusión del ganado. Letras distintas encima de las barras indican diferencias significativas ($p < 0.05$).

Buenos Aires, domingo 30 de septiembre de 2007

Las elecciones presidenciales: encuesta e



FICHA TECNICA

- **Universo:** personas residentes en el territorio argentino, en hogares particulares en centros urbanos de más de 2000 habitantes, mayores de 18 años, en condiciones de votar
- **Tipo de encuesta:** domiciliaria por muestreo y telefónica con el sistema CATI for Windows
- **Tamaño de la muestra:** 1329 casos; 1069 domiciliarios y 260 telefónicos

■ **Error estadístico:** +/- 2,68% para un nivel de confianza del 95%

de septiembre de 2007

LA NACION

Inferencia estadística

- **Población o universo** es el conjunto de **todas** las unidades de interés. Normalmente es demasiado grande para poder abarcarla. El estudio de toda la población se denomina **censo**.

 \mathcal{N} 

- **Muestra** es un subconjunto suyo al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)

 n 

La inferencia estadística consiste en **generalizar** las conclusiones extraídas de una **muestra** sobre la **población**



Parámetros y estimadores

- **Parámetro:** Es una cantidad numérica calculada sobre la población
- **Estimador:** Es una cantidad numérica calculada sobre la muestra

¿Y en los ejemplos?



- **Población**
- **Parámetro**

- **Muestra**
- **Estimador**

Pero ¿y cómo generalizamos? ¿podemos equivocarnos?
Necesitamos manejar probabilidades

Una situación supuesta

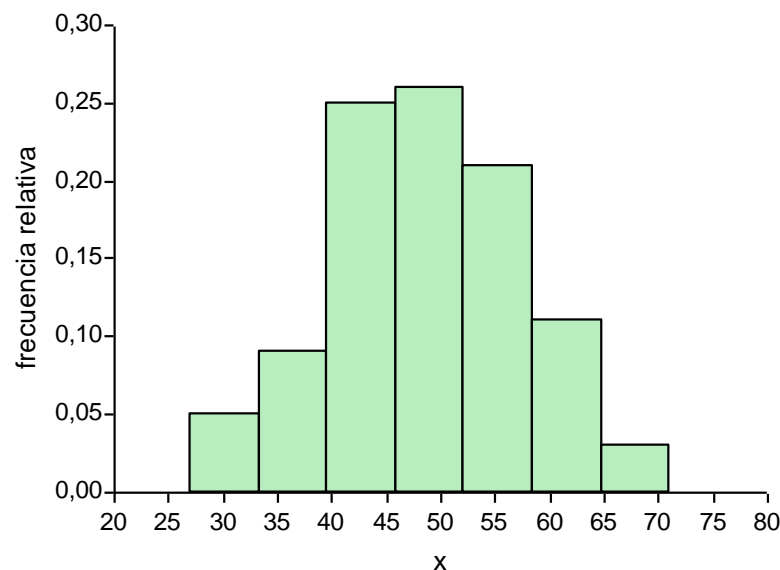
POBLACIÓN

41	44	34	42	53
54	41	61	44	44
52	59	36	57	61
60	53	54	43	48
57	43	51	51	52
32	45	55	36	47
49	42	38	71	46
54	27	55	45	42
46	45	58	53	43
42	54	44	39	49
62	54	36	61	59
57	43	63	47	49
32	56	44	44	53
49	45	52	58	59
42	67	58	32	39
48	37	49	47	64
52	33	35	51	41
45	47	46	55	42
43	50	61	47	67
57	49	57	52	47

- Contamos con una población integrada por 100 individuos; es decir $N=100$
- La media de la población es 50; es decir $\mu=50$
- La variabilidad de la población es de 10; es decir $\sigma = 10$

PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$

Histograma



¿Y si sacamos una muestra?

POBLACIÓN

41	44	34	42	53
54	41	61	44	44
52	59	36	57	61
60	53	54	43	48
57	43	51	51	52
32	45	55	36	47
49	42	38	71	46
54	27	55	45	42
46	45	58	53	43
42	54	44	39	49
62	54	36	61	59
57	43	63	47	49
32	56	44	44	53
49	45	52	58	59
42	67	58	32	39
48	37	49	47	64
52	33	35	51	41
45	47	46	55	42
43	50	61	47	67
57	49	57	52	47

MUESTRA $n = 5$

44	52	47	33	42
----	----	----	----	----

PROMEDIO $\bar{x} = 43.6$

- ☐ El promedio de la muestra no coincide con el de la población...
- ☐ La diferencia entre el valor muestral y el poblacional se denomina **error muestral**.
En este caso, $EM = 43.6 - 50 = -6.4$
- ☐ Es el costo que pagamos por no haber efectuado un censo

PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$

¿Y si sacamos otra muestra?

POBLACIÓN

41	44	34	42	53
54	41	61	44	44
52	59	36	57	61
60	53	54	43	48
57	43	51	51	52
32	45	55	36	47
49	42	38	71	46
54	27	55	45	42
46	45	58	53	43
42	54	44	39	49
62	54	36	61	59
57	43	63	47	49
32	56	44	44	53
49	45	52	58	59
42	67	58	32	39
48	37	49	47	64
52	33	35	51	41
45	47	46	55	42
43	50	61	47	67
57	49	57	52	47

MUESTRA $n = 5$

61	45	38	67	51
----	----	----	----	----

PROMEDIO $\bar{x} = 52.4$

$$EM = 52.4 - 50 = 2.4$$

Los **parámetros** se calculan sobre los N valores de la población, por lo tanto no cambian a menos que cambie la población, son **constantes**.

Los **estimadores** se calculan sobre n valores muestrales, por lo tanto varían de muestra en muestra y por lo tanto son **variables aleatorias**.

PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$

POBLACIÓN

41	44	34	42	53
54	41	61	44	44
52	59	36	57	61
60	53	54	43	48
57	43	51	51	52
32	45	55	36	47
49	42	38	71	46
54	27	55	45	42
46	45	58	53	43
42	54	44	39	49
62	54	36	61	59
57	43	63	47	49
32	56	44	44	53
49	45	52	58	59
42	67	58	32	39
48	37	49	47	64
52	33	35	51	41
45	47	46	55	42
43	50	61	47	67
57	49	57	52	47

Si repitiésemos este proceso muchas veces,
¿Qué comportamiento esperaríamos para los
75.287.520 promedios muestrales posibles?

MUESTRAS $n = 5$

44	52	47	33	42
----	----	----	----	----

61	45	38	67	51
----	----	----	----	----

51	54	50	33	71
----	----	----	----	----

.....

41	58	49	34	49
----	----	----	----	----

\bar{x}_1

\bar{x}_2

\bar{x}_3

.....

$\bar{x}_{75287520}$

?

PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$

Distribuciones muestrales

Definición: La distribución muestral de un estimador es la distribución de probabilidades de todos los posibles valores de un estimador que se pueden obtener extrayendo infinitas muestras aleatorias de tamaño n de la población.

La distribución de un estimador, como la de cualquier variable aleatoria, se pueden caracterizar por:

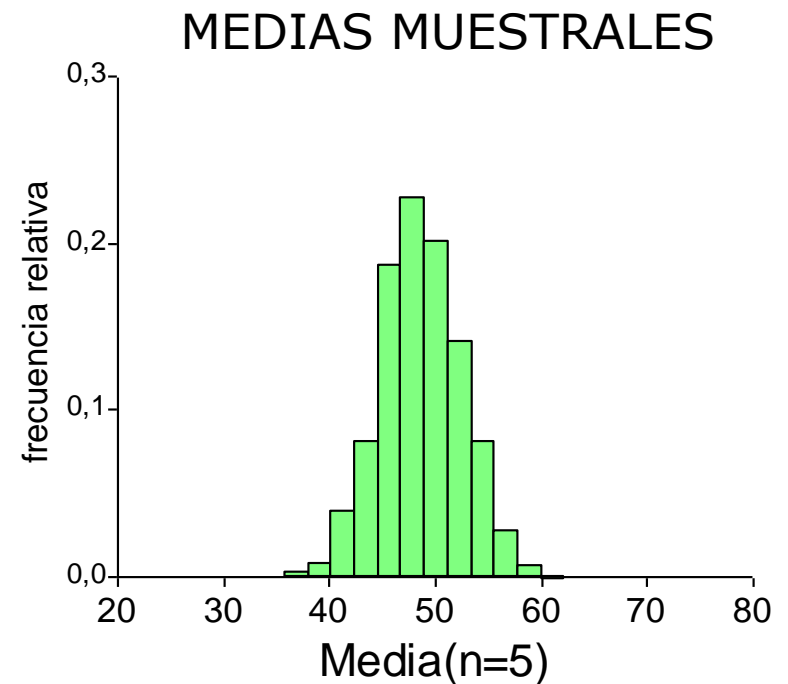
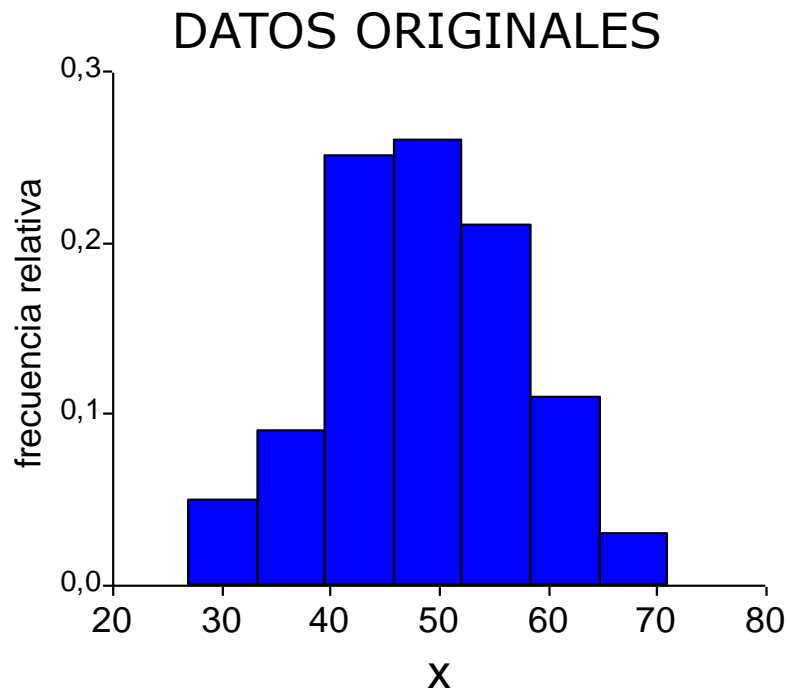
- ☐ tendencia central
- ☐ variabilidad
- ☐ función de probabilidad

Las distribuciones muestrales de los estimadores pueden ser:

- ☐ aproximadas mediante técnicas de **simulación**
- ☐ derivadas matemáticamente

Volviendo al ejemplo

Distribución muestral de \bar{x}



PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$

Distribución muestral de \bar{x}

¿Y si promediamos todas las medias muestrales?

$$\mu_{\bar{x}} = \mu$$

ESTIMADOR
INSESADO

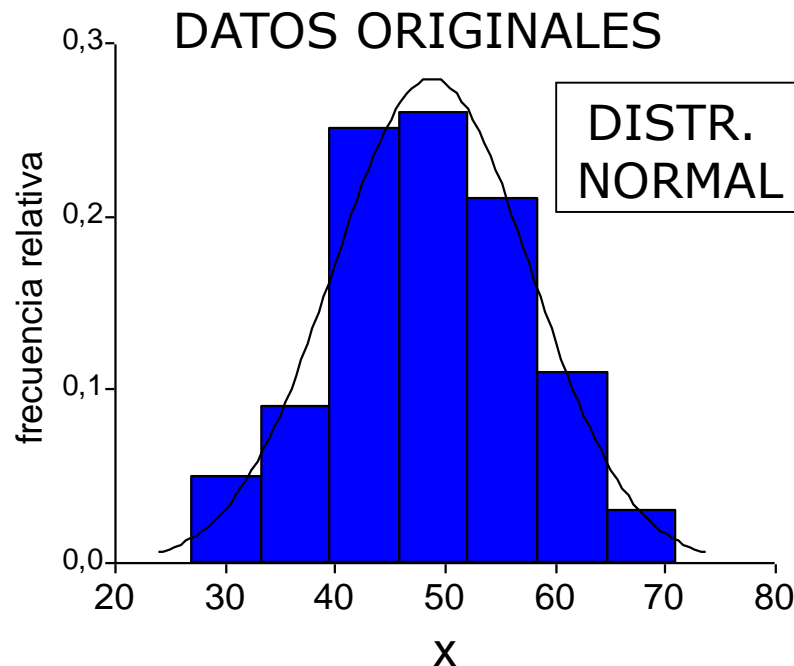
¿Cuál será la variabilidad de las medias muestrales?

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

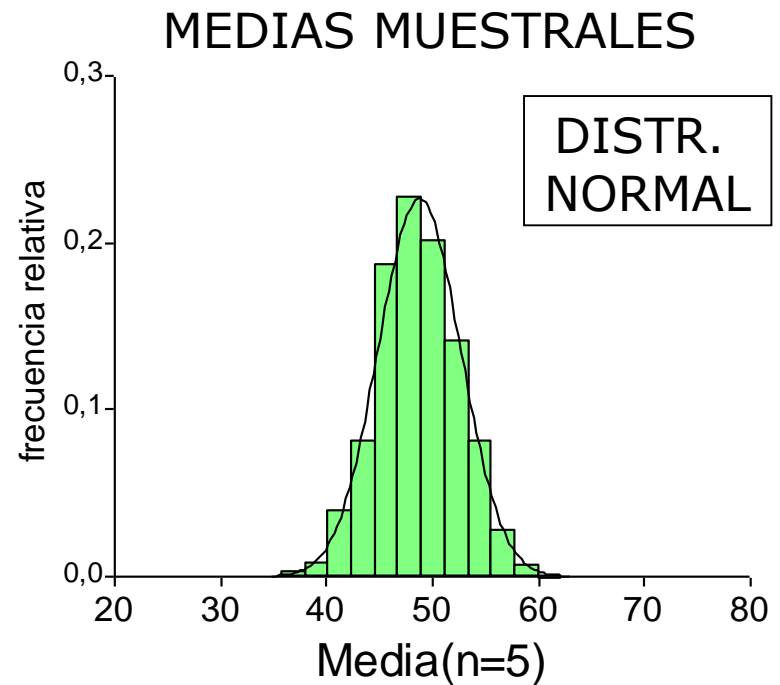
El desvío estándar de un estimador se conoce como **error estándar** y da idea de la precisión en la estimación

Distribución muestral de \bar{x}

¿Cuál será la distribución de probabilidades de \bar{x} ?

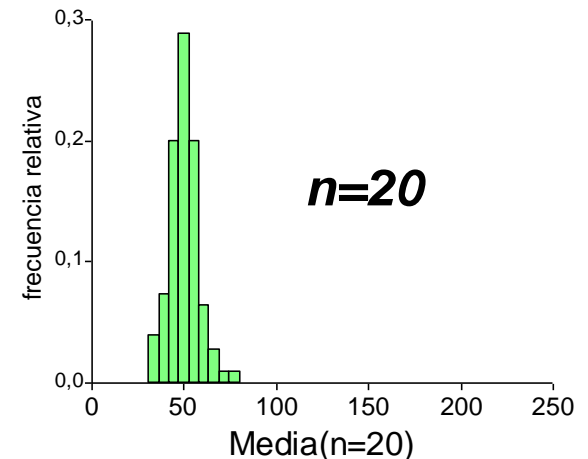
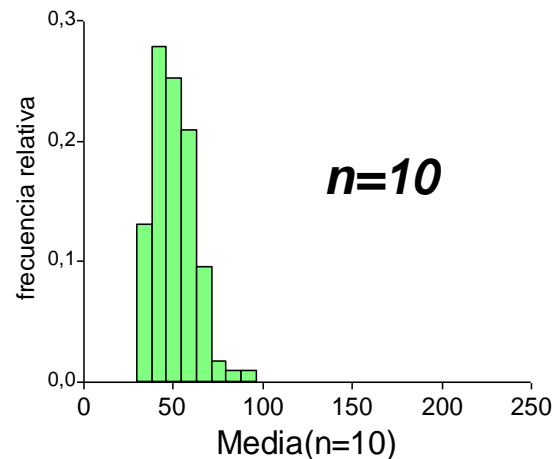
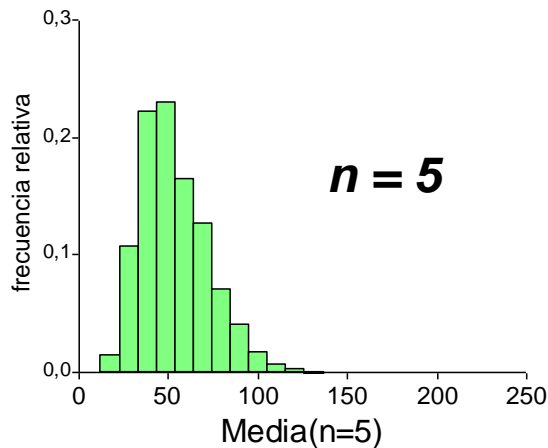
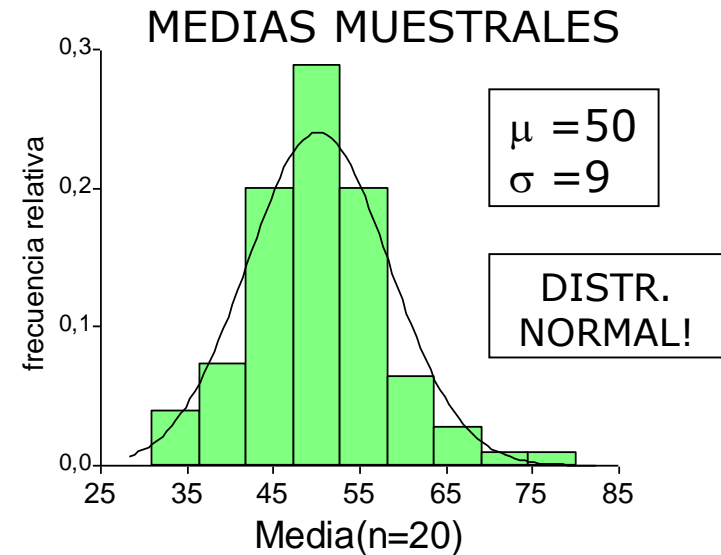
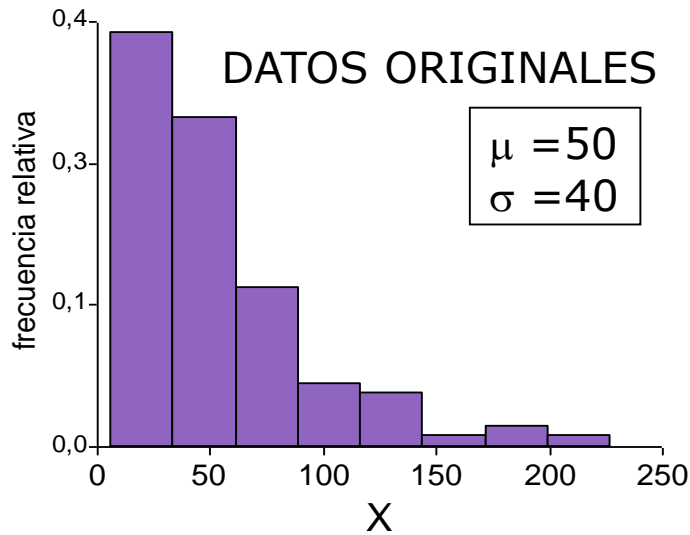


PROMEDIO $\mu = 50$
DESVÍO STD $\sigma = 10$



PROMEDIO $\mu_{\bar{x}} = 50$
ERROR STD $\sigma_{\bar{x}} = 4.5$

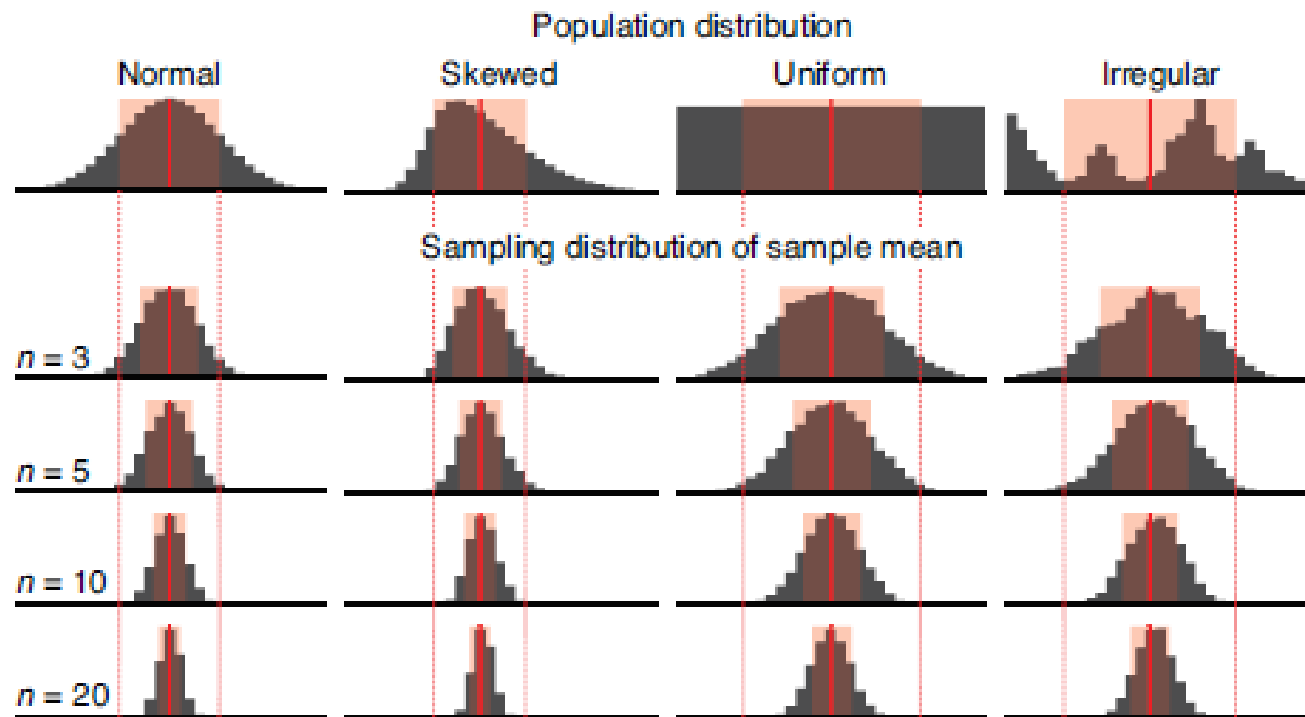
¿Y si los datos originales no siguen una distribución normal?



ampliando...

¿Y si los datos originales no siguen una distribución normal?

De cada distribución se extrajeron 10000 muestras del tamaño indicado, se calculó la media y se construyó la distribución muestral



Teorema central del límite

Si de una población con distribución **no normal o desconocida** con media μ y desvío estándar σ se extraen infinitas muestras aleatorias de tamaño n y a cada una de ellas se le calcula el promedio \bar{x} , se demuestra que éste se comporta según una **distribución normal** si n es **lo suficientemente grande**

¿A qué consideramos un n “lo suficientemente grande”?

- ❑ Si la variable original es **normal**, entonces \bar{x} será normal, para cualquier n
- ❑ Si la variable original es aproximadamente **simétrica y unimodal**, entonces \bar{x} tenderá a una distribución aproximadamente normal para n relativamente bajos
- ❑ Si la variable original es marcadamente **asimétrica**, entonces n deberá ser de mayor para que la distribución de \bar{x} sea normal

Algo para pensar...

- Una media muestral puede ser considerada como la suma de n variables aleatorias:

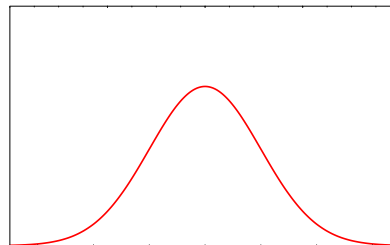
$$\bar{x} = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

- Por lo que el **Teorema central del límite** podría enunciarse: “Si se suman n variables aleatorias e independientes, la suma de dichas variables seguirá una distribución normal siempre y cuando n sea lo suficientemente grande”
- Cuando una variable (como la altura) es el resultado de la acción de muchos factores (variables), genéticos o ambientales, dicha variable seguirá una distribución **normal**
- En cambio, cuando es el resultado del efecto de un único o principal factor, la distribución tenderá a ser **asimétrica o bimodal**

En resumen:

Distribución muestral de \bar{x} cuando σ es conocido

1. La media de \bar{x} es: $\mu_{\bar{x}} = \mu$
2. El desvío estándar de \bar{x} (ES) es: $\sigma_{\bar{x}} = \sigma / \sqrt{n}$
3. Si el tamaño de la muestra es lo suficientemente grande o x es normal, la distribución de \bar{x} es **normal**



Por lo tanto es posible calcular probabilidades utilizando:

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

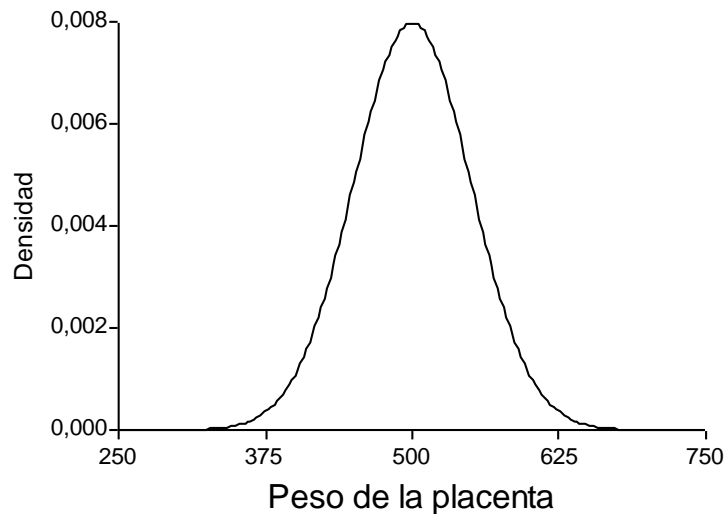
¿Es útil conocer la distribución de un estimador?

Nos permite calcular probabilidades \Rightarrow **es la clave para hacer inferencia!**

Por ejemplo:

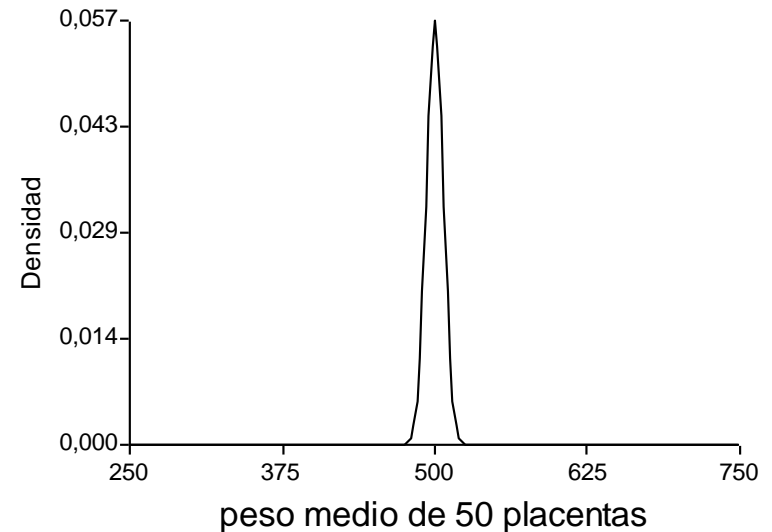
- Se sabe que el peso de la placenta de embarazos normales a término sigue una distribución normal con un promedio de **500g** y un desvío estándar de **50g**.
- Se determinó el peso de la placenta en **50** partos a término de madres fumadoras elegidas al azar y se obtuvo un promedio de **480g**.
- ¿Cuál es la probabilidad de que la media muestral sea de 480g o menor?

DATOS ORIGINALES



PROMEDIO $\mu = 500$
DESvíO STD $\sigma = 50$

MEDIAS MUESTRALES



PROMEDIO $\mu_{\bar{x}} = 500$
ERROR STD $\sigma_{\bar{x}} = 50/\sqrt{50} = 7$

$$P(\bar{x} < 480) = F(2.86) = 0.002$$

P-valor

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{480 - 500}{7} = 2.86$$

¿Qué necesitamos para hacer inferencia?

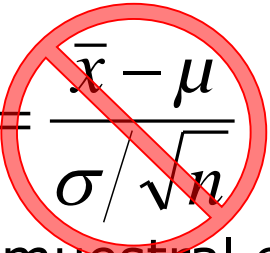
- ❑ **una** muestra aleatoria
- ❑ observaciones independientes
- ❑ un tamaño de muestra lo suficientemente grande

Algunas dudas que surgen...

- ❑ ¿es necesario sacar muchas (infinitas) muestras para poder aplicar el TCL?
- ❑ ¿A mayor n más cerca del parámetro estará mi estimador?
- ❑ ¿A mayor n menor variabilidad de los datos?

Distribución muestral de \bar{x} cuando el desvío estándar poblacional es desconocido

- ❑ En la práctica es habitual que TODOS los parámetros poblacionales son **desconocidos**, es decir que ni el promedio μ ni el desvío estándar poblacional σ son conocidos!
- ❑ Como se desconoce σ se utiliza su estimador $s \rightarrow$ mayor incertidumbre
- ❑ No es correcto utilizar la distribución **normal** para \bar{x}

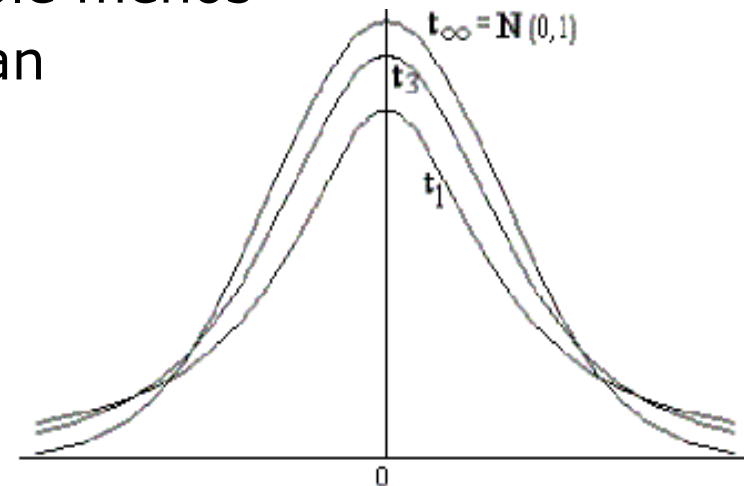

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- ❑ Se demuestra que la media muestral en estos casos ajusta a una distribución conocida como **t de Student**

$$t_{GL} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Distribución t de Student

- ❑ Tiene forma acampanada como la normal estándar, pero su dispersión es mayor (es más aplanada). Esto se debe a que al desconocer σ hay mayor incertidumbre
- ❑ Es simétrica con respecto al cero, es decir que $\mu=0$
- ❑ No se trata de una única curva, sino de infinitas curvas, cada una caracterizada por un parámetro denominado **grados de libertad** (GL)
- ❑ Los GL indican la cantidad de datos **independientes**, es decir el número de observaciones de la variable menos el número de restricciones que verifican
- ❑ Los GL dependen del tamaño de la muestra y en este caso valen $n-1$
- ❑ A medida que aumentan los GL más se asemeja a la normal estándar (porque s converge a σ)



Distribución muestral de \bar{x} cuando no se conoce σ

1. La media de \bar{x} es: $\mu_{\bar{x}} = \mu$
2. El desvío estándar (ES) de \bar{x} es: $\sigma_{\bar{x}} = s/\sqrt{n}$
3. Si el tamaño de la muestra es lo suficientemente grande o x es normal, la distribución de \bar{x} es ***t de Student***, con $n-1$ grados de libertad

Por lo tanto es posible calcular probabilidades utilizando:

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Propiedades de un buen estimador

- **Insesgado:** Un estimador es insesgado cuando la esperanza del estimador es igual al valor del parámetro que se desea estimar. O sea:

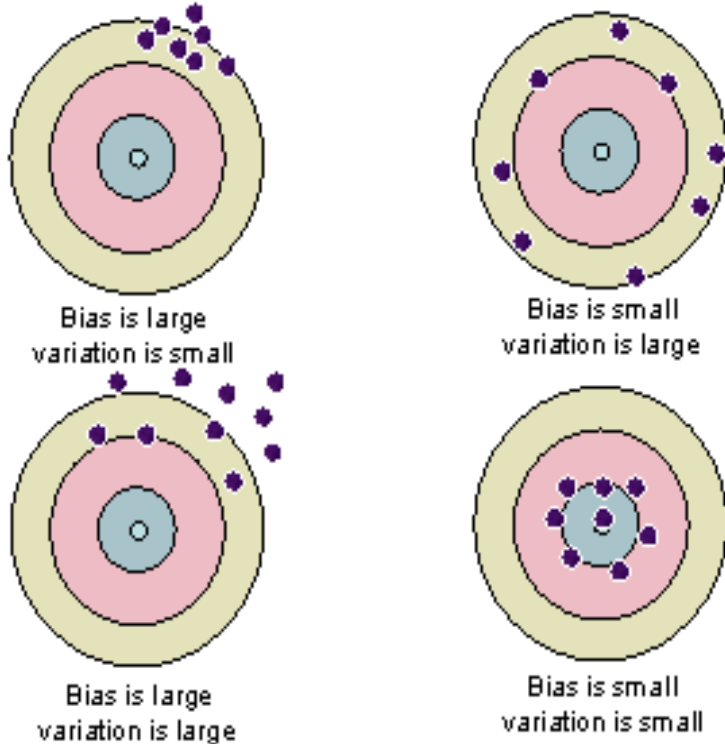
$$\mu_{(\hat{\theta})} = \theta$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \text{ es un estimador insesgado de } \sigma^2$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} \text{ no lo es}$$

- **Consistente:** A medida que el tamaño de la muestra aumenta el estimador debe tender al valor del parámetro y su variancia debe tender a cero

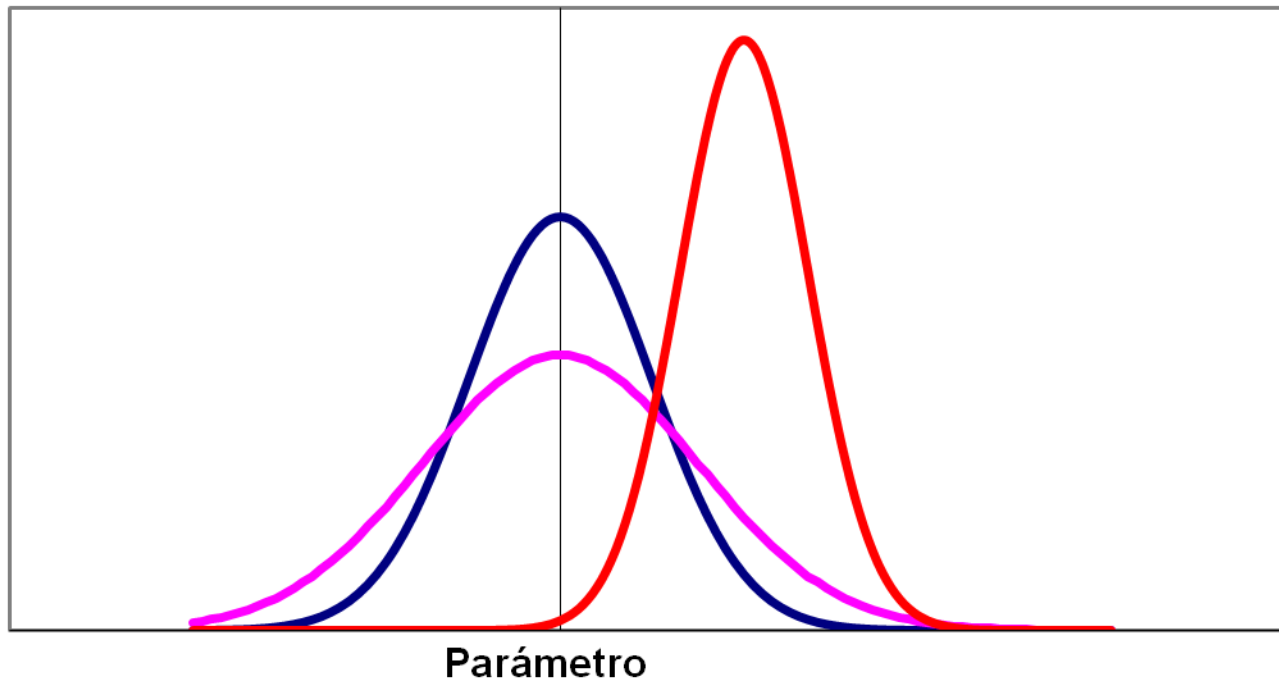
Propiedades de un buen estimador



▪ **Insesgado:** significa que el promedio del estimador es igual al parámetro (no sobre ni subestima sistemáticamente al parámetro)

▪ De los estimadores insesgados, se prefieren aquellos con **menor variabilidad** (más consistentes)

Distribución de 3 estimadores



— Estimador 1 — Estimador 2 — Estimador 3

¿Cuál es el mejor estimador?