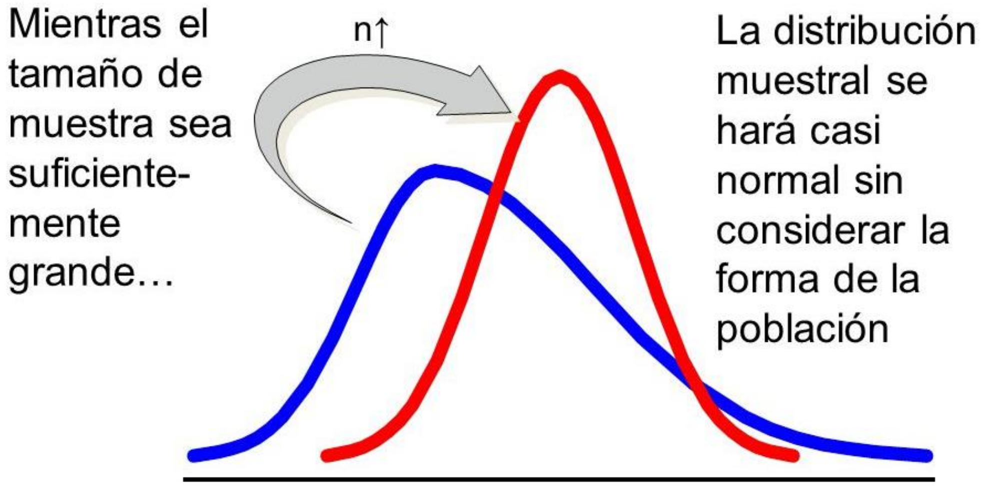
Distribuciones Muestrales



Teorema Central del Límite

Distribuciones Muestrales

Para hacer inferencias sobre una población, es necesario examinar un poco más los resultados muestrales.

Suponer que se toma una muestra aleatoria de n elementos y se calcula la media aritmética ($\bar{X}_n$) ¿Coincidirá con μ ?

Si se toma una segunda muestra ¿tendrá la media igual a la de la muestra anterior? ¿Coincidirá con μ ?

Es de esperar que los valores estén "próximos" al valor poblacional.

Pero ¿Qué se entiende por "próximos" ? ¿Cómo se determina o mide ésta proximidad o cercanía?

¿Cómo se distribuye un estadístico muestral cuando se ha muestreado repetidamente?

Un estadístico muestral es una variable aleatoria, porque el muestreo es aleatorio.

Medidas Estadísticas Descriptivas

Una *distribución muestral* es aquella distribución teórica de probabilidades de los valores posibles de un **estadístico** muestral que ocurriría si obtuviéramos **todas las muestras posibles** de un **tamaño fijo**, de una variable correspondiente a una población dada.

Estadísticos muestrales importantes:

Sea $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_n$ una muestra aleatoria de una variable cuantitativa **X**, entonces las variables:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \boldsymbol{T}_n = \sum_{i=1}^{n} X_i$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Si **Y** es una variable aleatoria con distribución binomial

$$Y \sim Bin(n, p)$$

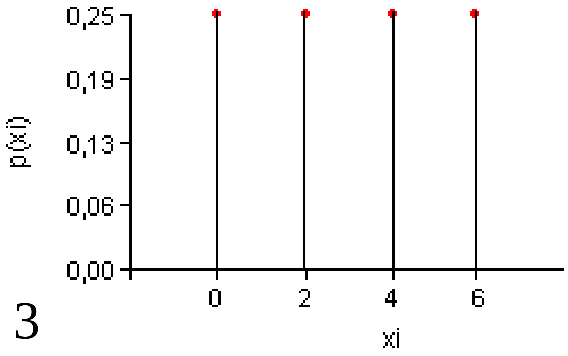
$\frac{Y}{n}$ = “la proporción muestral de éxitos”.

Distribuciones Muestrales

Una *distribución muestral* es aquella distribución teórica de probabilidades de los valores posibles de un estadístico muestral que ocurriría si obtuviéramos todas las muestras posibles de un tamaño fijo, de una variable correspondiente a una población dada.

Ejemplo: X={0; 2; 4; 6} ← es la Distribución poblacional de X

x_i	$p(x_i)$	$x_i\, p(x_i)$	$x_i^2\, p(x_i)$
0	0,25	0	0
2	0,25	0,5	1
4	0,25	1	4
6	0,25	1,5	9
tot	1	3	14



$E(X) = 3$

$V(X) = E(X^2) - [E(X)]^2$

$V(X) = 14 - (3^2) = 14 - 9 = 5$

Construyan todas las muestras posibles de tamaño 2

Calcular esperanza y varianza.

Distribuciones Muestrales

Ejemplo: X={0; 2; 4; 6} ← es la Distribución poblacional de X

Ayuda para encontrar todos los pares posibles: colocamos los números en la primer fila y la primer columna

	0	**2**	**4**	**6**
0	(0,0)	(0,2)	(0,4)	(0,6)
2	(2,0)	(2,2)	(2,4)	(2,6)
4	(4,0)	(4,2)	(4,4)	(4,6)
6	(6,0)	(6,2)	(6,4)	(6,6)

Variable: Distribución de la media

y_i	$p(y_i)$
0	1/16=0,0625
1	2/16=0,125
2	3/16=0,1875
3	4/16=0,25
4	3/16=0,1875
5	2/16=0,125
6	1/16=0,0625
Total	1

Tabla con el promedio de cada par:

	0	**2**	**4**	**6**
0	0	1	2	3
2	1	2	3	4
4	2	3	4	5
6	3	4	5	6

Distribuciones Muestrales

Ejemplo: X={0; 2; 4; 6} ← es la Distribución poblacional de X

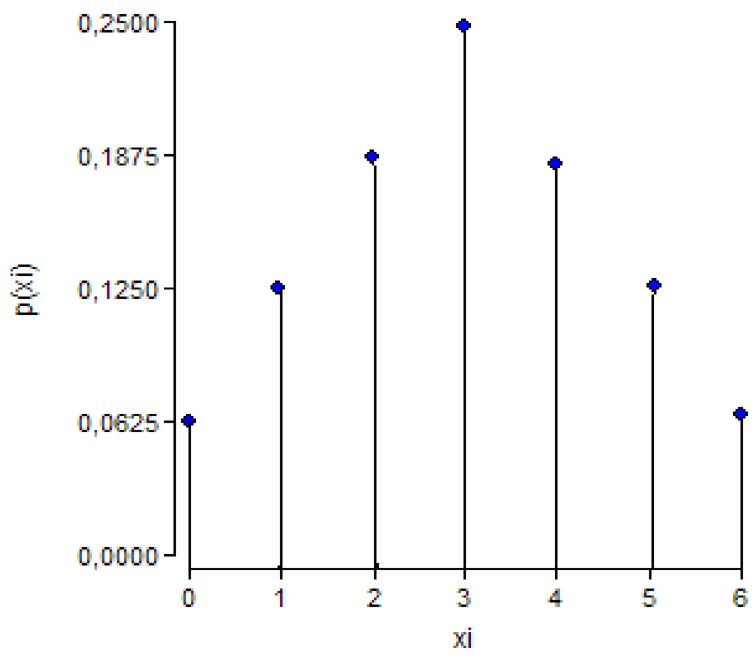
y_i	$p(y_i)$	$y_i\, p(y_i)$	$y_i^2 p(y_i)$
0	0,0625	0	0
1	0,125	0,125	0,125
2	0,1875	0,375	0,75
3	0,25	0,75	2,25
4	0,1875	0,75	3
5	0,125	0,625	3,125
6	0,0625	0,375	2,25
tot	1	3	11,5

Si n=2 y llamamos Y a la media

E(Y)= 3 **(igual que antes)**

$V(Y)= E(Y^2)-[E(Y)]^2$

V(Y)= 11,5 – 9 =2,5 = 5/2 **(la mitad que antes)**



Distribuciones Muestrales

Ejemplo: X={0; 2; 4; 6} ← es la Distribución poblacional de X

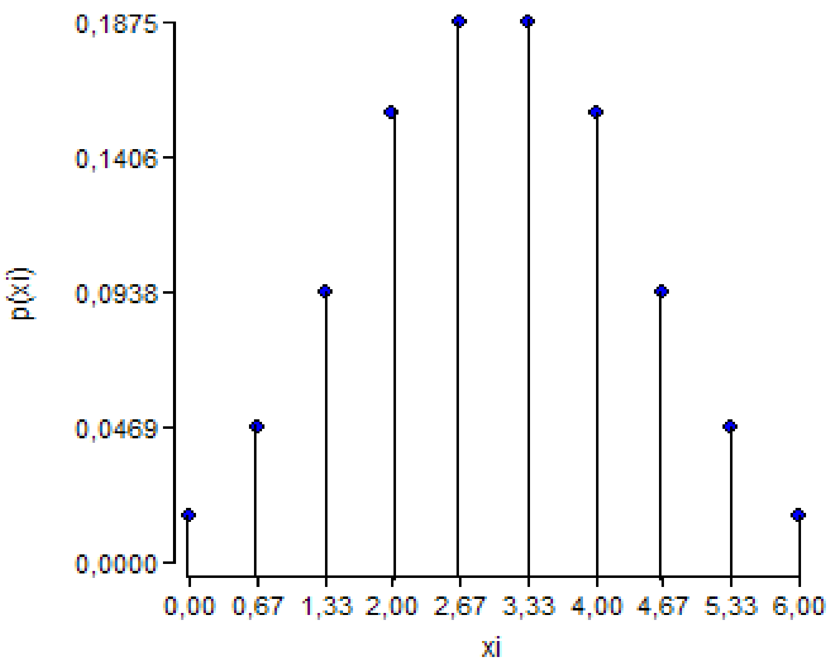
Si n=3 y llamamos Y a la media

E(Y)= 3 **(igual que antes)**

$V(Y)= E(Y^2)-[E(Y)]^2$

V(Y)= 10,66 – 9 =1,66 = 5/3 **(1/3 de la poblacional)**

y_i	$p(y_i)$	$y_i\, p(y_i)$	$y_i^2\, p(y_i)$
0	0,016	0	0
0,666667	0,047	0,03125	0,020833
1,333333	0,094	0,125	0,166667
2	0,15625	0,3125	0,625
2,666667	0,1875	0,5	1,333333
3,333333	0,1875	0,625	2,083333
4	0,15625	0,625	2,5
4,666667	0,09375	0,4375	2,041667
5,333333	0,047	0,25	1,333333
6	0,016	0,09375	0,5625
tot	1	3	10,66667

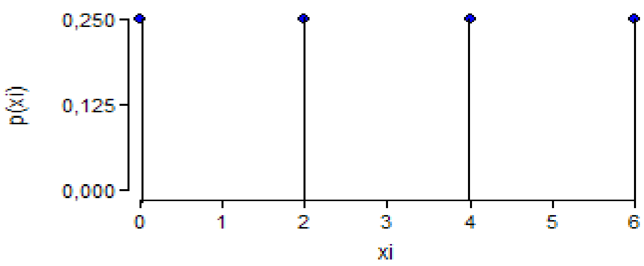


Distribuciones Muestrales

X={0; 2; 4; 6}

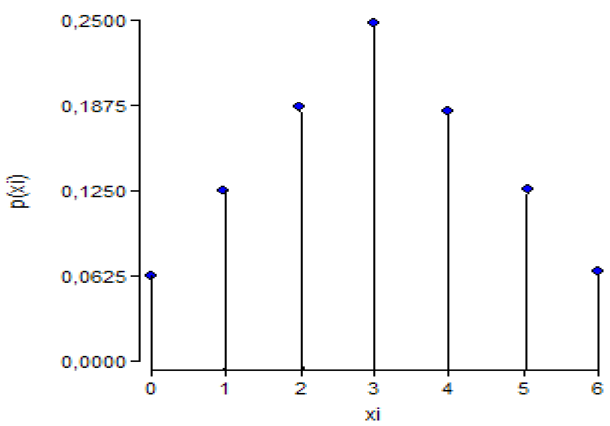
Cuando n=1

$$E(\bar{X}_1) = 3$$
$$V(\bar{X}_1) = E(\bar{X}_1^2) - [E(\bar{X}_1)]^2 =$$
$$= 14 - 9 = 5/1$$



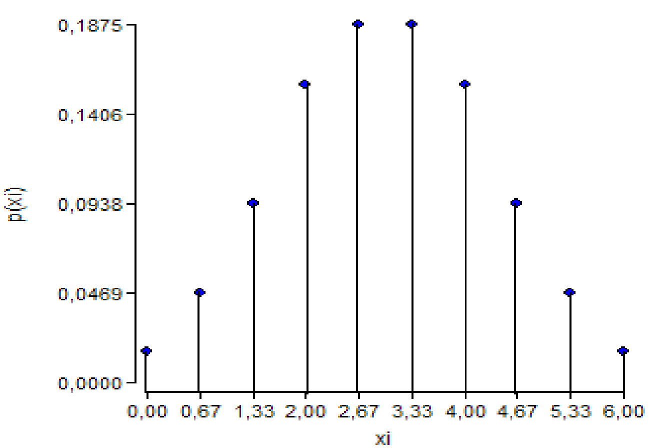
Cuando n=2

$$E(\bar{X}_2) = 3$$
$$V(\bar{X}_2) = E(\bar{X}_2^2) - [E(\bar{X}_2)]^2 =$$
$$11{,}5 - 9 = 5/2$$



Cuando n=3

$$E(\bar{X}_3) = 3$$
$$V(\bar{X}_3) = E(\bar{X}_3^2) - [E(\bar{X}_3)]^2 =$$
$$10{,}66 - 9 = 5/3$$



Distribución de la media Muestral

Conclusión:

- La media de la distribución muestral no varía al cambiar el tamaño de la muestra. (Siempre es μ; E($\bar{X}$)= μ para todo n)

-La desviación estándar de la media muestral (llamada *error estándar* de la media) indica que la dispersión de la media muestral decrece a media que crece el tamaño de muestra

$$\sigma_{\bar{X}}^{2} = \frac{\sigma_X^2}{n} \Longrightarrow$$

La varianza de la media aritmética de las muestras disminuye al aumentar n

$$\sqrt{\frac{\sigma_X^2}{n}} = \frac{\sigma_X}{\sqrt{n}} \Longrightarrow$$

El desvío estándar de la media aritmética de las muestras disminuye al aumentar n

Estadísticos Muestrales

<u>Distribución muestral de</u> $\bar{X}_n$

Sea X una v.a. tal que $E(X) = \mu$ y $V(X) = \sigma^2$

y una muestra aleatoria de tamaño n $X_1, X_2, X_3, \ldots, X_n$

Entonces:

a) $E(\bar{X}) = \mu$

b) $V(\bar{X}) = \dfrac{\sigma^2}{n}$

c) Si $X \sim N(\mu; \sigma^2) \rightarrow \bar{X} \sim N\left(\mu; \dfrac{\sigma^2}{n}\right)$

<u>Nota</u>:

1) no importa cual es el tamaño de la muestra; si X se distribuye *normalmente* entonces *la media aritmética se distribuye normalmente*.

2) Importante: la varianza de la media muestral es siempre menor que la varianza original.

Teorema Central del Límite

El *Teorema Central del Límite*, es la herramienta más importante de la teoría de distribuciones pues *garantiza las propiedades distribucionales de la media muestral*, la que tiene un rol destacado en la estadística inferencial (estimación y prueba de hipótesis).

Dada una población cualquiera con media μ y varianza σ^2 la distribución muestral de $\bar{x}$, calculada a partir de las muestras aleatorias de tamaño n con reemplazo (reposición) de esta población, estará distribuida en forma aproximadamente normal con $\mu_{\bar{x}} = \mu$ y varianza $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ y la aproximación a la normal será mejor cuanto mayor sea n.

¿Qué significa “n” suficientemente grande?

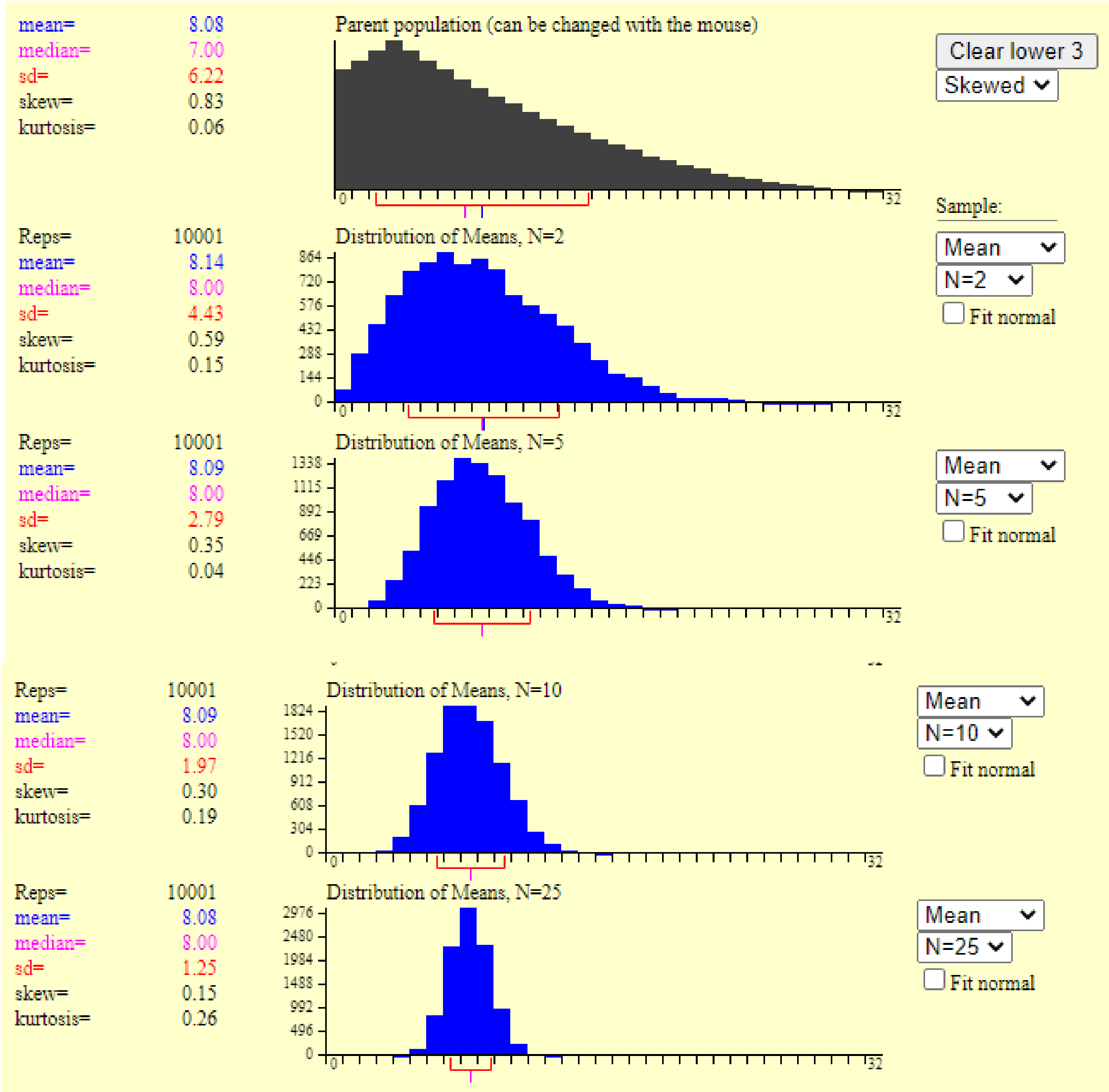
¿Cómo sabemos si la aproximación es buena?

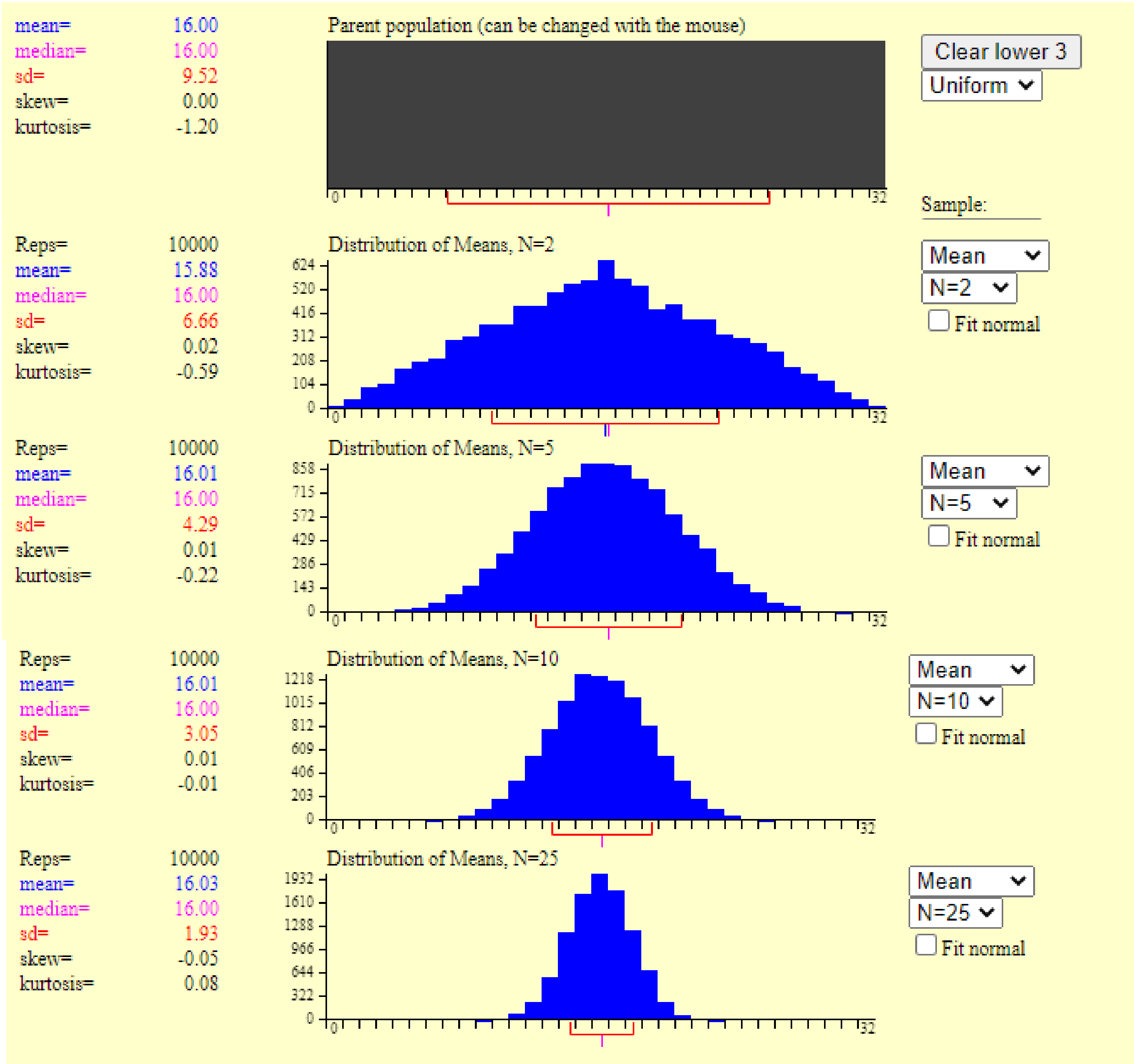
El tamaño *n* de la muestra necesario para que la aproximación sea razonable depende la distribución de la variable original. Mientras más acampanada y simétrica es la distribución de las observaciones, más rápidamente se logra una buena aproximación; pero en términos generales basta con $n \geq 30$ para obtener buenas aproximaciones.

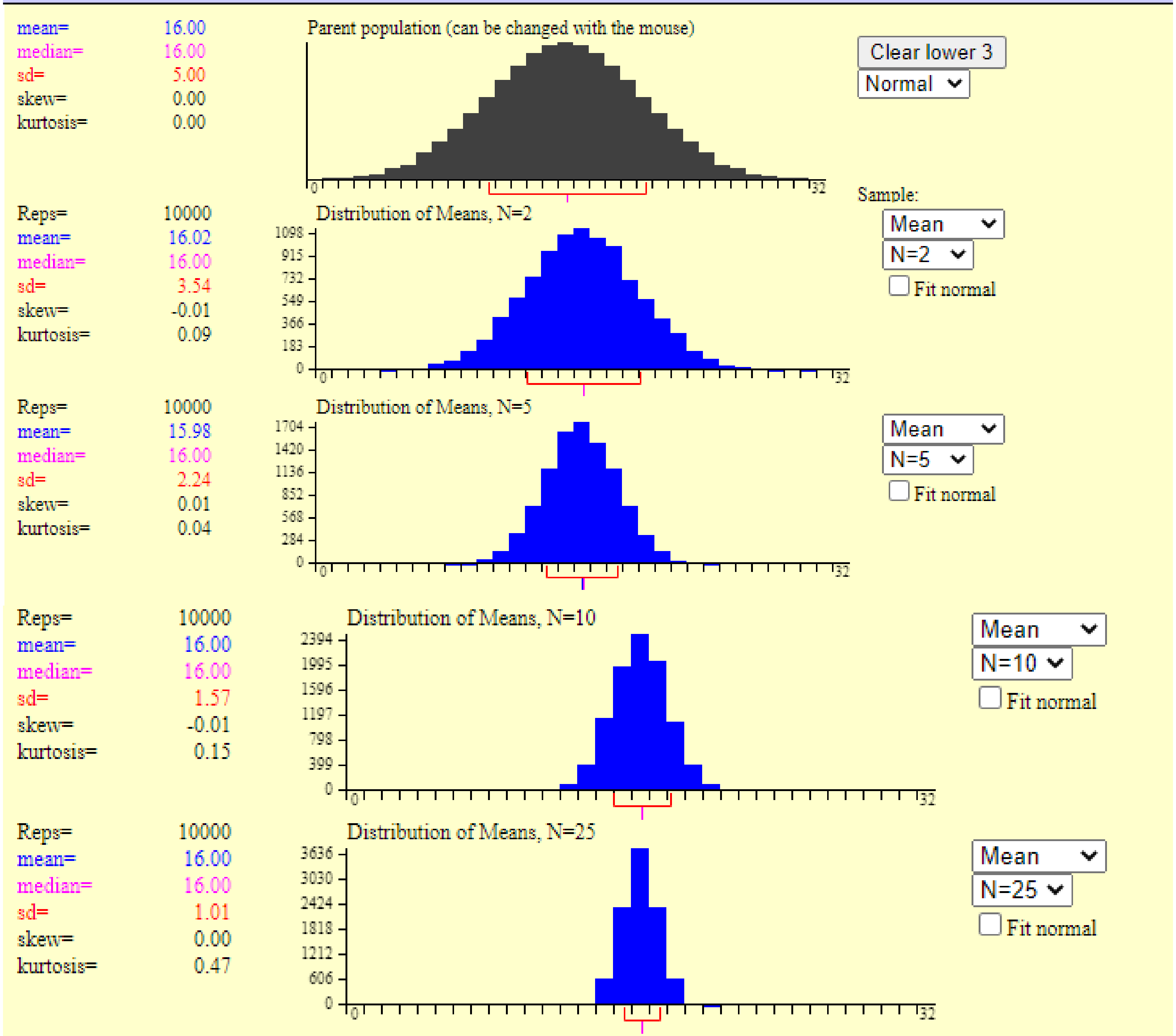
Si de ***cualquier población*** independiente de la forma de la distribución se extraen muestras aleatorias de tamaño n fijo, a medida que ***n crece*** *la distribución de las medias muestrales* tiende a convertirse en una distribución *normal* con media μ y varianza σ^2/n

En esto radica la importancia del Teorema Central del Límite, ya que sabemos que ***sin importar la forma de la distribución de la población en estudio*** (ya que es muy difícil conocer la tendencia de una población muy grande), si se toma un número suficiente de muestras repetidas veces y se les calcula la media, ***la distribución de las medias muestrales va a tener un comportamiento normal*** (entonces podremos usar Z o t para estimar ciertos parámetros poblacionales)

https://onlinestatbook.com/stat_sim/sampling_dist/index.html


mean= 8.08
median= 7.00
sd= 6.22
skew= 0.83
kurtosis= 0.06
Parent population (can be changed with the mouse)
Clear lower 3
Skewed
Sample:
Reps= 10001
mean= 8.14
median= 8.00
sd= 4.43
skew= 0.59
kurtosis= 0.15
Distribution of Means, N=2
Mean
N=2
Fit normal
Reps= 10001
mean= 8.09
median= 8.00
sd= 2.79
skew= 0.35
kurtosis= 0.04
Distribution of Means, N=5
Mean
N=5
Fit normal
Reps= 10001
mean= 8.09
median= 8.00
sd= 1.97
skew= 0.30
kurtosis= 0.19
Distribution of Means, N=10
Mean
N=10
Fit normal
Reps= 10001
mean= 8.08
median= 8.00
sd= 1.25
skew= 0.15
kurtosis= 0.26
Distribution of Means, N=25
Mean
N=25
Fit normal


mean= 16.00
median= 16.00
sd= 9.52
skew= 0.00
kurtosis= -1.20
Parent population (can be changed with the mouse)
Clear lower 3
Uniform
Sample:
Reps= 10000
mean= 15.88
median= 16.00
sd= 6.66
skew= 0.02
kurtosis= -0.59
Distribution of Means, N=2
Mean
N=2
Fit normal
Reps= 10000
mean= 16.01
median= 16.00
sd= 4.29
skew= 0.01
kurtosis= -0.22
Distribution of Means, N=5
Mean
N=5
Fit normal
Reps= 10000
mean= 16.01
median= 16.00
sd= 3.05
skew= 0.01
kurtosis= -0.01
Distribution of Means, N=10
Mean
N=10
Fit normal
Reps= 10000
mean= 16.03
median= 16.00
sd= 1.93
skew= -0.05
kurtosis= 0.08
Distribution of Means, N=25
Mean
N=25
Fit normal


mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent population (can be changed with the mouse)
0
32
Clear lower 3
Normal
Sample:
Reps= 10000
mean= 16.02
median= 16.00
sd= 3.54
skew= -0.01
kurtosis= 0.09
Distribution of Means, N=2
1098
915
732
549
366
183
0
Mean
N=2
Fit normal
Reps= 10000
mean= 15.98
median= 16.00
sd= 2.24
skew= 0.01
kurtosis= 0.04
Distribution of Means, N=5
1704
1420
1136
852
568
284
0
Mean
N=5
Fit normal
Reps= 10000
mean= 16.00
median= 16.00
sd= 1.57
skew= -0.01
kurtosis= 0.15
Distribution of Means, N=10
2394
1995
1596
1197
798
399
0
Mean
N=10
Fit normal
Reps= 10000
mean= 16.00
median= 16.00
sd= 1.01
skew= 0.00
kurtosis= 0.47
Distribution of Means, N=25
3636
3030
2424
1818
1212
606
0
Mean
N=25
Fit normal

En el caso de muestras sin reposición

Volvemos al ejemplo original; X={0; 2; 4; 6}

Pares	$\bar{X}$		
0,2	1		
0,4	2		
0,6	3		
2,4	3		
2,6	4		
4,6	5		
tot			

E(X)= 3

V(Xmed)= $E(X^2)-[E(X)]^2$

V(Xmed)= 64/6 – 9 = 5/3

No hace falta hacer estos cálculos. Si tomamos el valor de la varianza para muestras de tamaño 2 con reemplazo y lo multiplicamos por el factor de corrección por población finita y llegamos al valor que calculamos arriba.

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{5}{2}\left(\frac{4-2}{4-1}\right) = \frac{5}{3}$$

Estadísticos Muestrales

Distribución muestral de $\bar{X}_n$

Sea X una v.a. tal que $E(X) = \mu$ y $V(X) = \sigma^2$

y una muestra aleatoria de tamaño n $X_1, X_2, X_3, \ldots, X_n$

Entonces:

a) $E(\bar{X}) = \mu$

b) $\mathrm{V}(\bar{X}) = \dfrac{\sigma^2}{n}$

c) Si $X \sim N(\mu; \sigma^2) \rightarrow \bar{X} \sim N\left(\mu; \dfrac{\sigma^2}{n}\right)$

Nota:

1) no importa cual es el tamaño de la muestra; si X se distribuye *normalmente* entonces *la media aritmética se distribuye normalmente*.

2) Importante: la varianza de la media muestral es siempre menor que la varianza original.

Recordando la transformación de nuestras distribuciones a una Distribución Z, cuyas probabilidades se hayan tabuladas:

$$Z = \frac{X - \mu}{\sigma}$$

Como en este caso trabajamos con “la media de las medias”, vamos a utilizar esta fórmula, en donde el desvío estándar está corregido, ya que fue obtenido a partir de las múltiples muestras a las que luego se les calculó la media

$$Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$$

Ejemplo 1:

Asuma que el aumento de peso en un cobayo, alimentado con un determinado balanceado, se distribuye normal con media igual a 105 g y desvío estándar igual a 10,5 g. Si se extrae de esta población de cobayos una muestra aleatoria de tamaño 16:

a) ¿Con qué probabilidad un cobayo tomado al azar tendrá un aumento de peso de al menos 104 g?

b) ¿Cuál es la probabilidad de que los cobayos de una muestra de tamaño 16 tengan un aumento de peso promedio inferior a 106 g?

a) ¿Con qué probabilidad un cobayo tomado al azar tendrá un aumento de peso de al menos 104 g?

a) En este caso estamos calculando la probabilidad de que un cobayo pese 104 g o más (al menos 104 g), nos interesa la distribución de X. Entonces se procede:

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

$X \sim N(105;\ 110{,}25)$

$$P(X \geq 104) = P\left(Z \geq \frac{104 - 105}{10{,}5}\right) = P(Z \geq -0{,}095) = 1 - P(Z < -0{,}095)$$

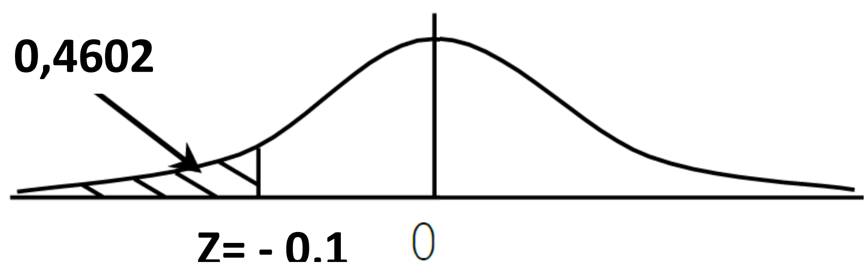
Buscamos en la tabla de la distribución normal estándar el valor -0,095 pero como la tabla tiene valores de z con dos decimales tomamos como valor aproximado -0,1 y obtenemos:

$P(X \geq 104) = P(Z \geq -0{,}1) =$

$P(X \geq 104) = 1 - P(Z < -0{,}1)$

$P(X \geq 104) =$ 1- 0,4602 (de tabla)

$P(X \geq 104) =$ 0,5398 ≈ 0,54



Un cobayo de la muestra tendrá un aumento de peso de al menos 104 g con una probabilidad de 0,54.

b) ¿Cuál es la probabilidad de que los cobayos de una muestra de tamaño 16 tengan un aumento de peso promedio inferior a 106 g?

b) En este caso nos preguntan sobre el aumento medio (o promedio) de los cobayos de una muestra de tamaño 16, entonces se debe usar la distribución de $\bar{X}_{16}$.

$$\boxed{Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}}$$

$$P(\bar{X}_{16} < 106) \equiv P\left(Z < \frac{106 - 105}{\left(\frac{10{,}5}{\sqrt{16}}\right)}\right) = P(Z < +0{,}38) = 0{,}6480$$

La probabilidad de que el promedio de los individuos de una muestra de tamaño 16 sea menor que 106 g es igual a 0,65.

Ejemplo 2:

Una máquina dosificadora está programada para que la cantidad de alimento balanceado que arroja en cada bolsa sea una variable aleatoria con una media de 2000g y una desviación estándar de 150g. ¿Cuál es la probabilidad de que la cantidad media de alimento servido en una muestra aleatoria de 36 bolsas sea mayor a 2040g ?

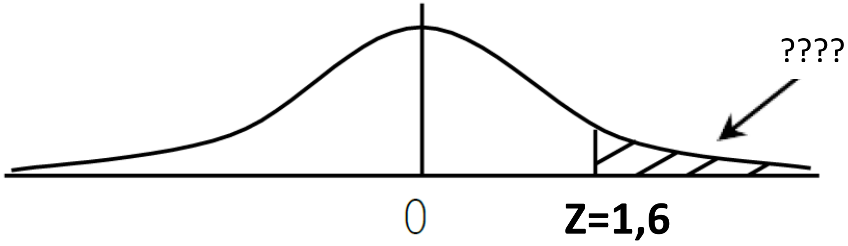
Realizar los cálculos usando el TCL.

$$\boxed{Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}}$$

$$P(\bar{X} > 2040) = P\left(Z > \frac{2040 - 2000}{\frac{150}{\sqrt{36}}}\right) = P(Z > 1{,}6) =$$

1- $P(Z = 1{,}6) =$

$1 - 0{,}9452 = 0{,}0548$



Corolario del Teorema Central del Límite de De-Moivre- Laplace

Este corolario permite aproximar la distribución binomial por la normal:

Sea $X \sim Bin(n,p)$ y sea $\frac{X}{n}$ = "la proporción muestral de éxitos".

	Variable	Media	Desvío
Binomial	X	$E(x)= \mu_x=np$	$\sigma_x = \sqrt{npq}$
Proporción	x/n	$E(x/n) = \frac{np}{n} = \pi$	Desv (x/n) = $\sqrt{\frac{pq}{n}}$

por el Teorema Central del límite:

$$X \approx N(np; np(1-p)) \rightleftharpoons \text{y} \rightleftharpoons \frac{X}{n} \approx N\left(p; \frac{p(1-p)}{n}\right)$$

$$z \approx \frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}}$$

E(Bin)=n.p
Var(Bin)= n p q= n. p. (1-p)

Se considera que la aproximación es buena cuando: $np \geq 5$ y además $np(1\text{-}p) \geq 5$

***Ejemplo 3*:**

Se ha determinado que 60% de los estudiantes de una universidad grande trabaja más de 6 hs diarias.

Si se toma una muestra aleatoria de 800 de dichos estudiantes, ¿cuál es la probabilidad de que la proporción de los que trabaja más de 6hs en la muestra sea menor que 0,55?

Solución:

Aplicando el TCL sabemos que

$$p \xrightarrow{d} N\left(p, \frac{p(1-p)}{n}\right)$$

$$z \approx \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}$$

$$P(p < 0{,}55) = P\left(Z < \frac{0{,}55 - 0{,}60}{\sqrt{\frac{0{,}60 0{,}40}{800}}}\right) = P(Z < -2{,}89) = \mathit{0{,}00193}$$

Recordemos

Parámetro θ	**Estimador $\widehat{\theta}$**
μ	$\bar{X}$
σ^2	S^2
π	p

Un estimador es un valor que puede calcularse a partir de los datos muestrales y que proporciona información sobre el valor del parámetro. Un estimador es una variable aleatoria.

$\hat{\theta}$ estimador de θ

PROPIEDADES DE LOS ESTIMADORES

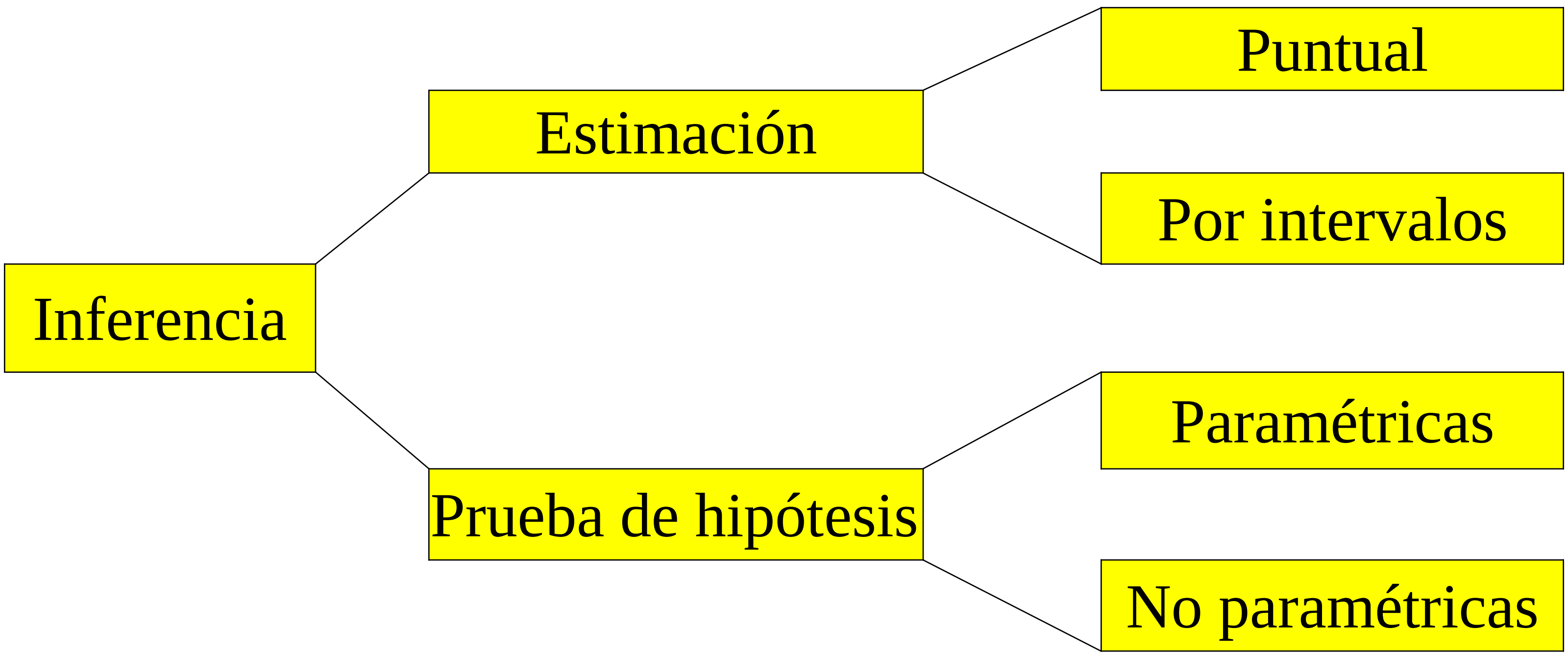
1.-Se dice que un estimador es insesgado si $E\left[\hat{\theta}\right] = \theta$

2.-El estimador $\hat{\theta}_1$ **es más eficiente que** $\hat{\theta}_2$ **si** $Var(\hat{\theta}_1) < Var\left(\hat{\theta}_2\right)$

3.-Un estimador es consistente si $lim_{n\to\infty}\,\hat{\theta} = \theta$

4.-Un estimador suficiente para un parámetro es aquel capaz de recoger o resumir toda la información que la muestra de una Variable aleatoria X contiene.

INFERENCIA


Inferencia
Estimación
Puntual
Por intervalos
Prueba de hipótesis
Paramétricas
No paramétricas

En la práctica, interesa no sólo dar una estimación de un parámetro, sino que además, un intervalo que permita precisar la incertidumbre existente en la estimación (lo veremos en el próximo tema:

INTERVALOS DE CONFIANZA

¡Muchas Gracias por su atención!