

Repaso de estadística

Junio, 2017

Rosangela Bando

Oficina de Planificación Estratégica y Efectividad en el Desarrollo



Objetivo

Revisión de conceptos.

I. Repaso de estadística: min, max, mediana, media, desviación estándar, histograma, muestreo aleatorio simple, teorema del límite central y ley de los grandes números.

II. Pruebas de diferencias en la media: Comparación de medias, Potencia, significancia y regresión para comparar medias.

III. Identificación: resultado, atributo, mecanismo de asignación, endógeno, identificación, observables, heterogeneidad.

Objetivo

variable	N	mean	sd	min	p50	max
puntaje	25	654	118	445	683	849

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	600	6.3	100	587.5	612.4
Escuela B	250	630	6.3	100	617.5	642.4
Diferencia		-30	8.9		-47.6	-12.4

Tratamiento	-.24 (0.09)***
Obs.	500

- * significativo al 90%,
- ** significativo at 95%,
- *** significativo at 99%

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

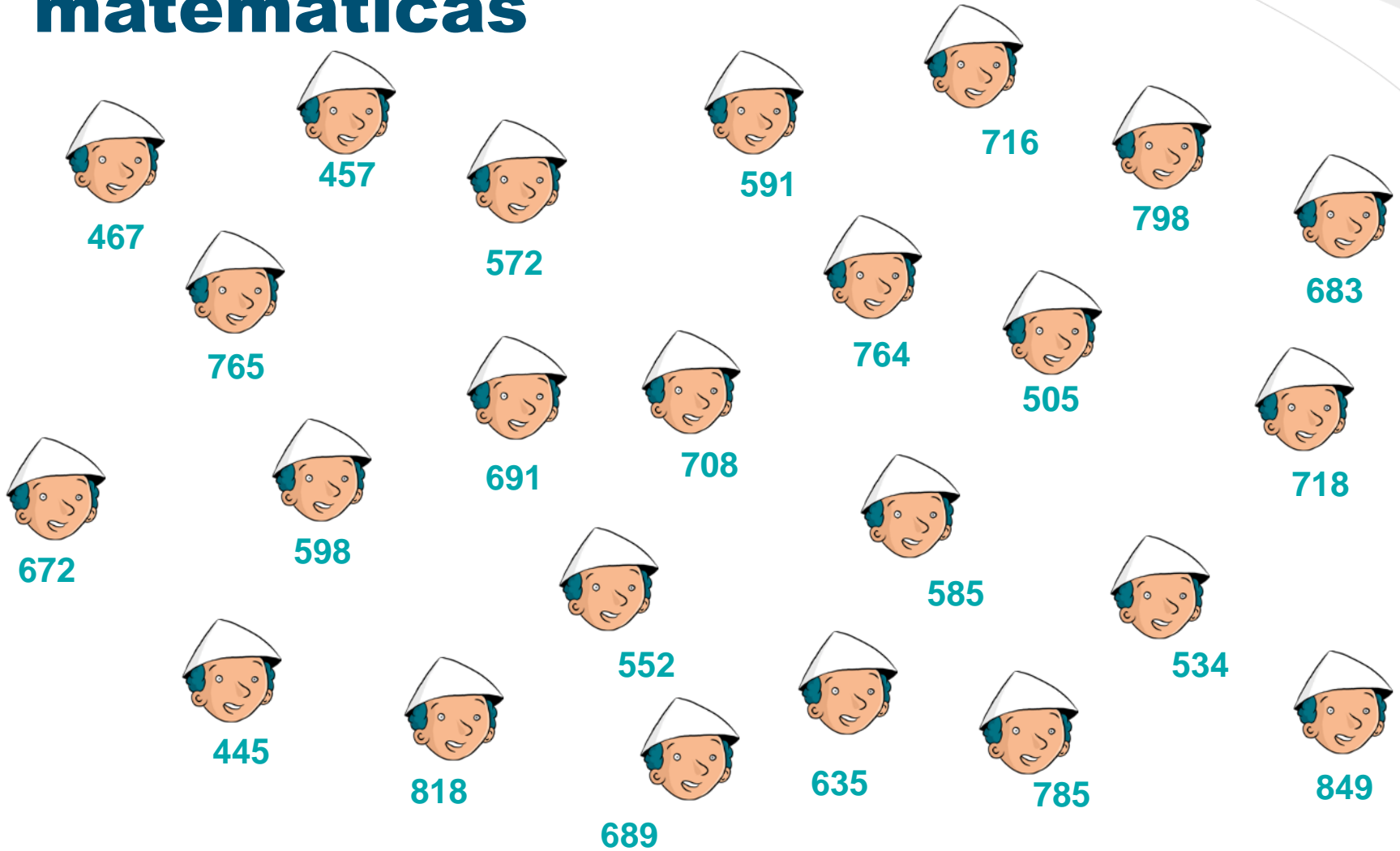
I. REPASO DE ESTADÍSTICA

Motivación

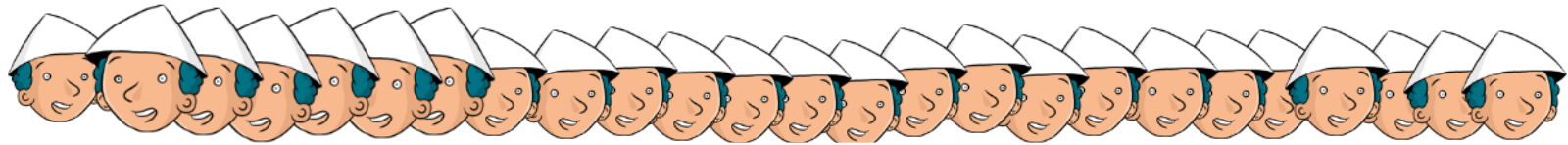
Tabla 1. Desempeño de estudiantes en matemáticas

variable	N	mean	sd	min	p50	max
-----+						
puntaje	25	654	118	445	683	849

Desempeño de estudiantes en matemáticas

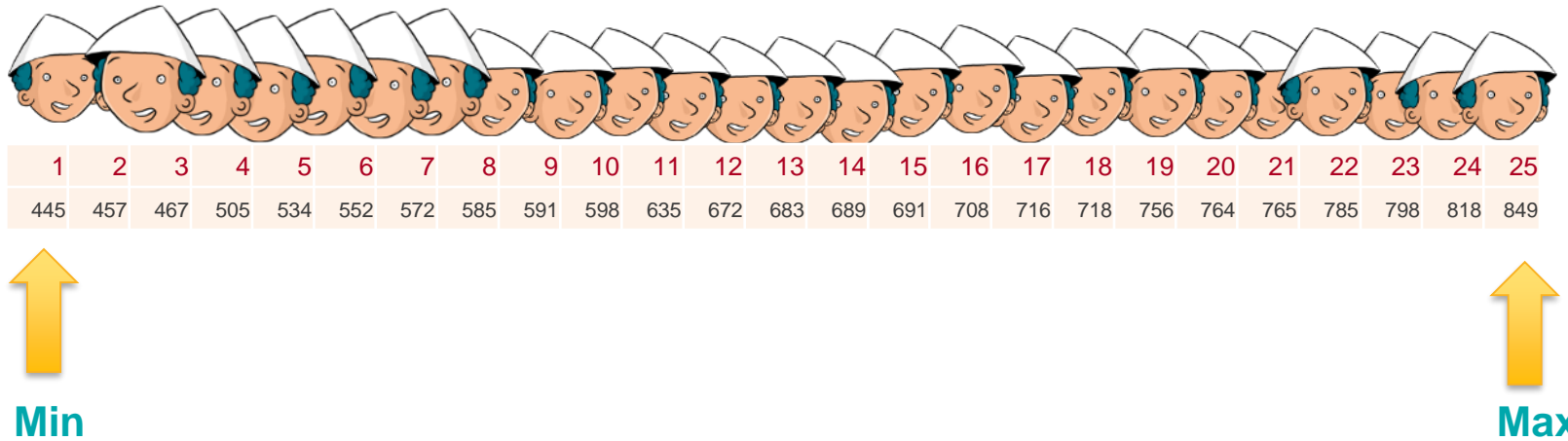


Estadística

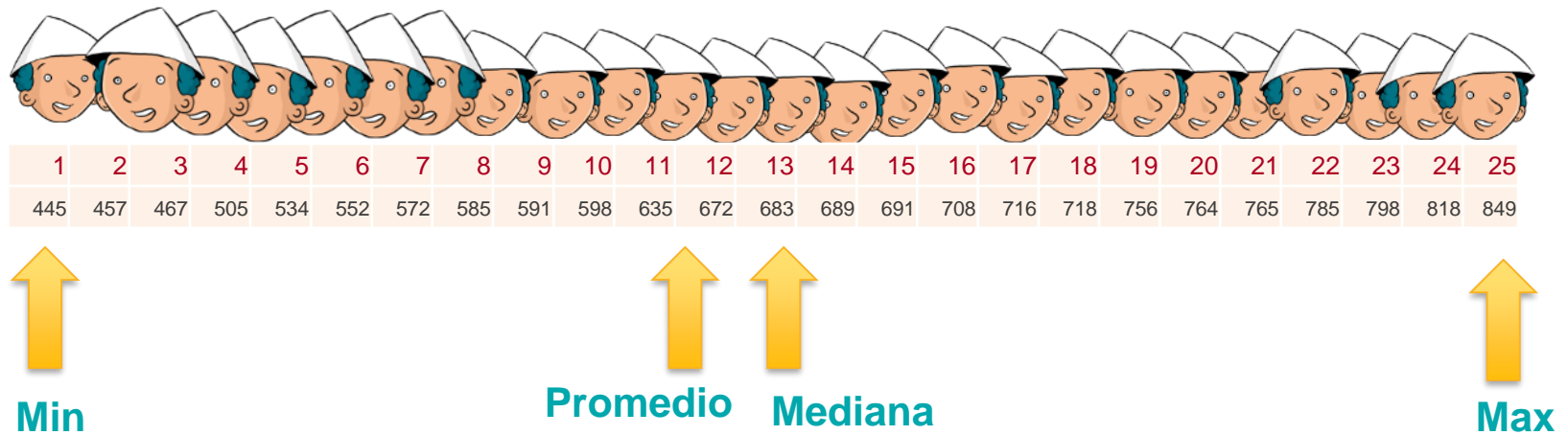


Orden:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Puntaje:	445	457	467	505	534	552	572	585	591	598	635	672	683	689	691	708	716	718	756	764	765	785	798	818	849

Descripción básica de datos



Descripción básica de datos

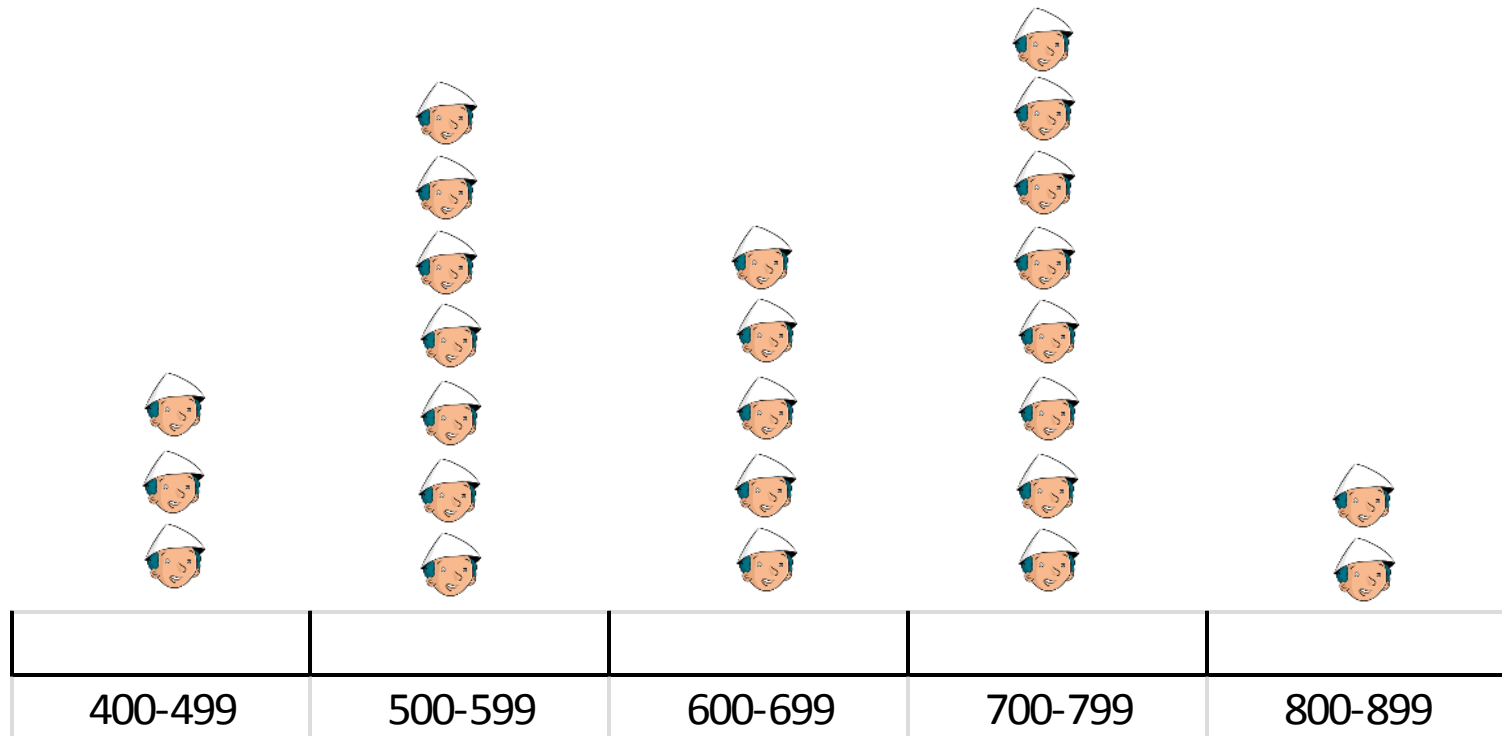


Suma de todos los puntajes: 16353
Total de estudiantes: 25



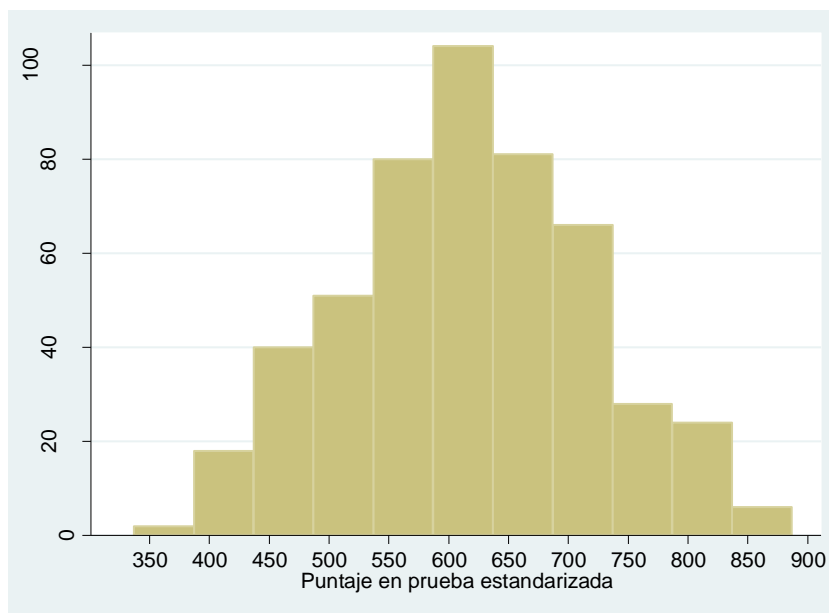
División de puntaje igual: 654

Histograma



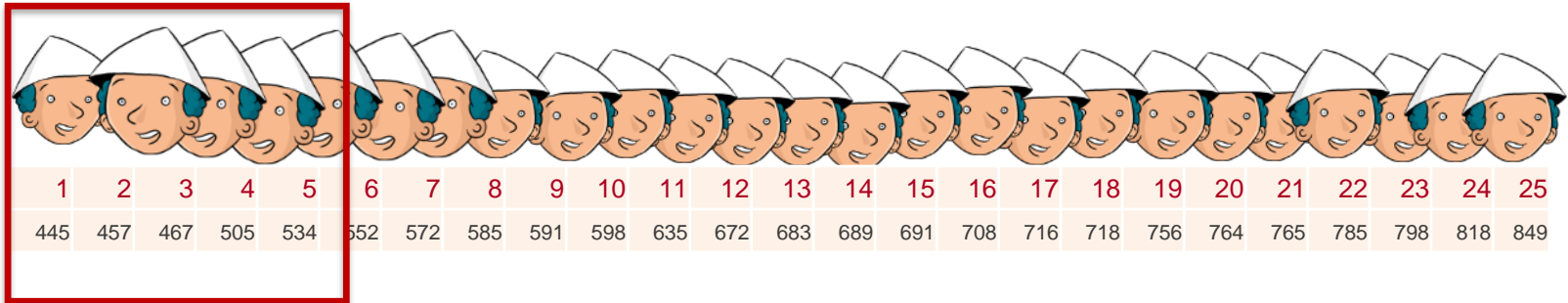
Descripción básica de datos

variable	N	μ mean	σ sd	min	p50	max
puntaje	500	615	101	362	615.5	879



Muestra

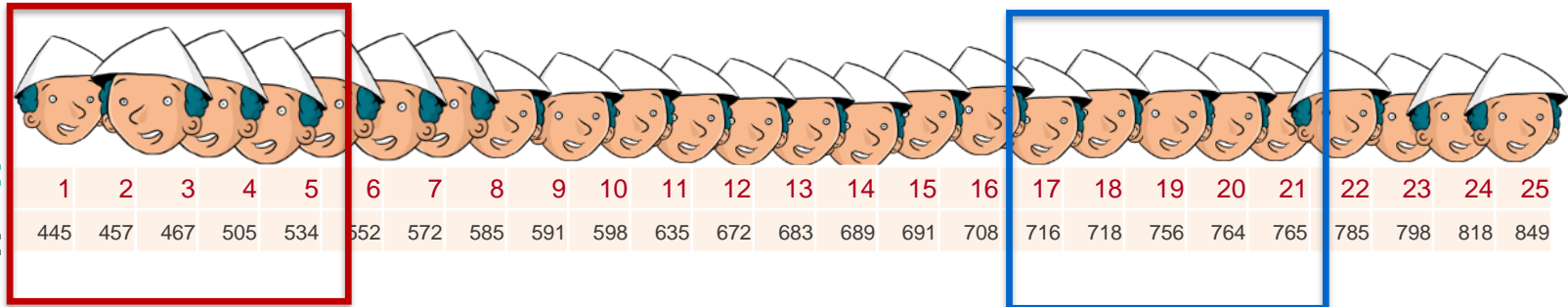
Orden:
Puntaje:



variable	N	mean	sd	min	p50	max
<hr/>						
puntaje	25	654	118	445	683	849
<hr/>						
Muestra 1	5	482	40	445	467	534

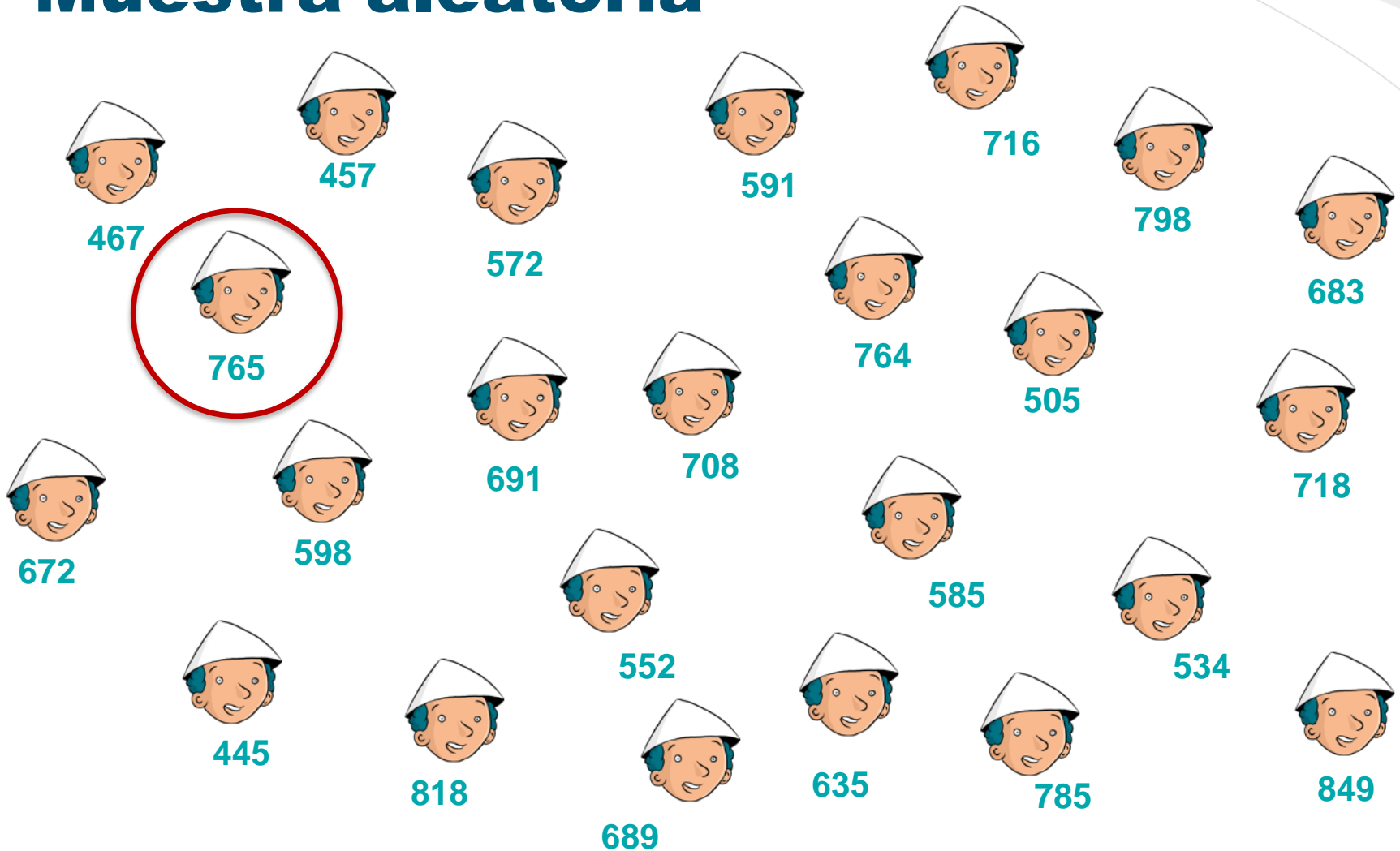
Muestra

Orden:
Puntaje:



variable	N	mean	sd	min	p50	max
Población	25	654	118	445	683	849
Muestra 1	5	482	40	445	467	534
Muestra 2	5	803	32	716	756	765

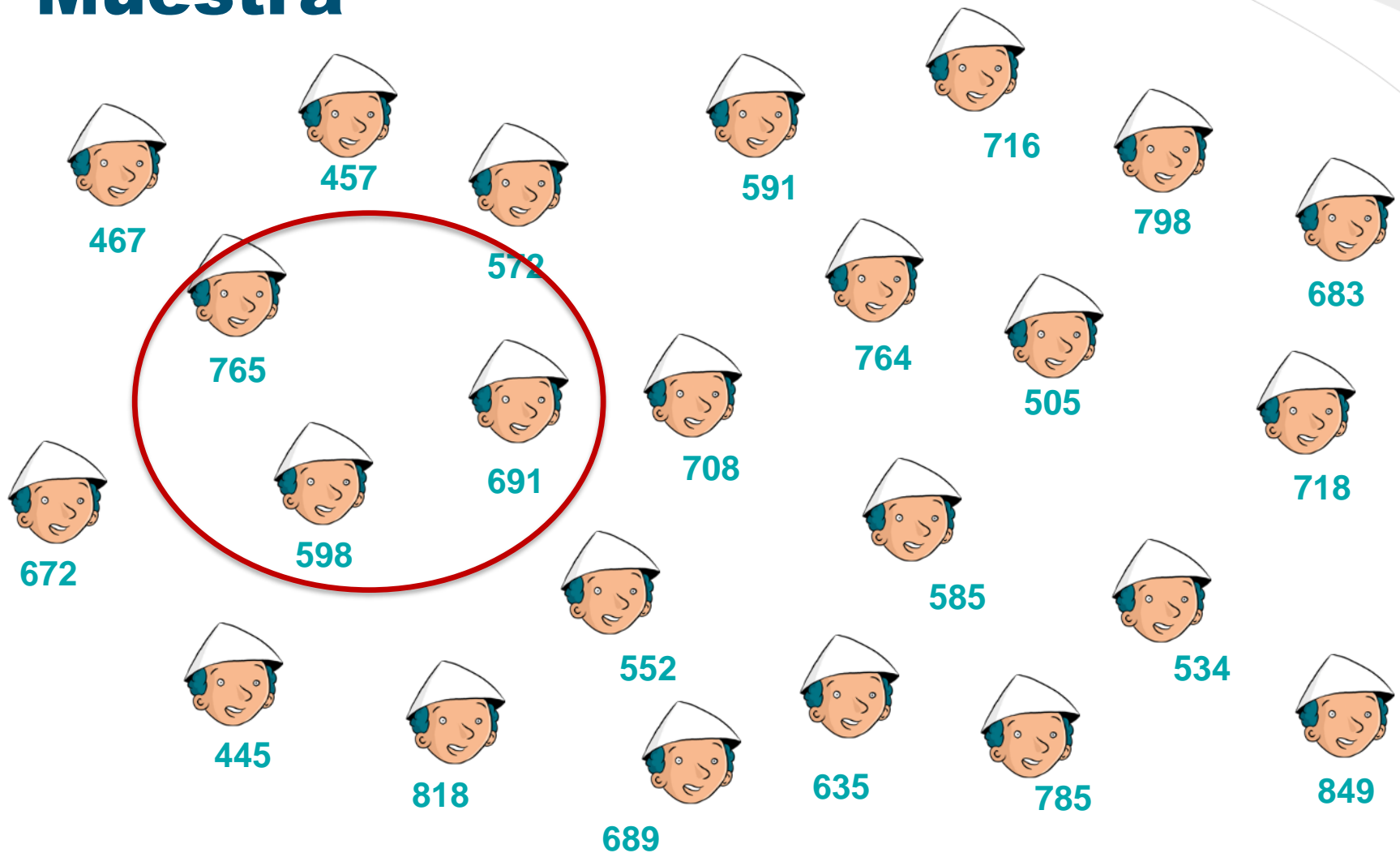
Muestra aleatoria



Media muestral: 765

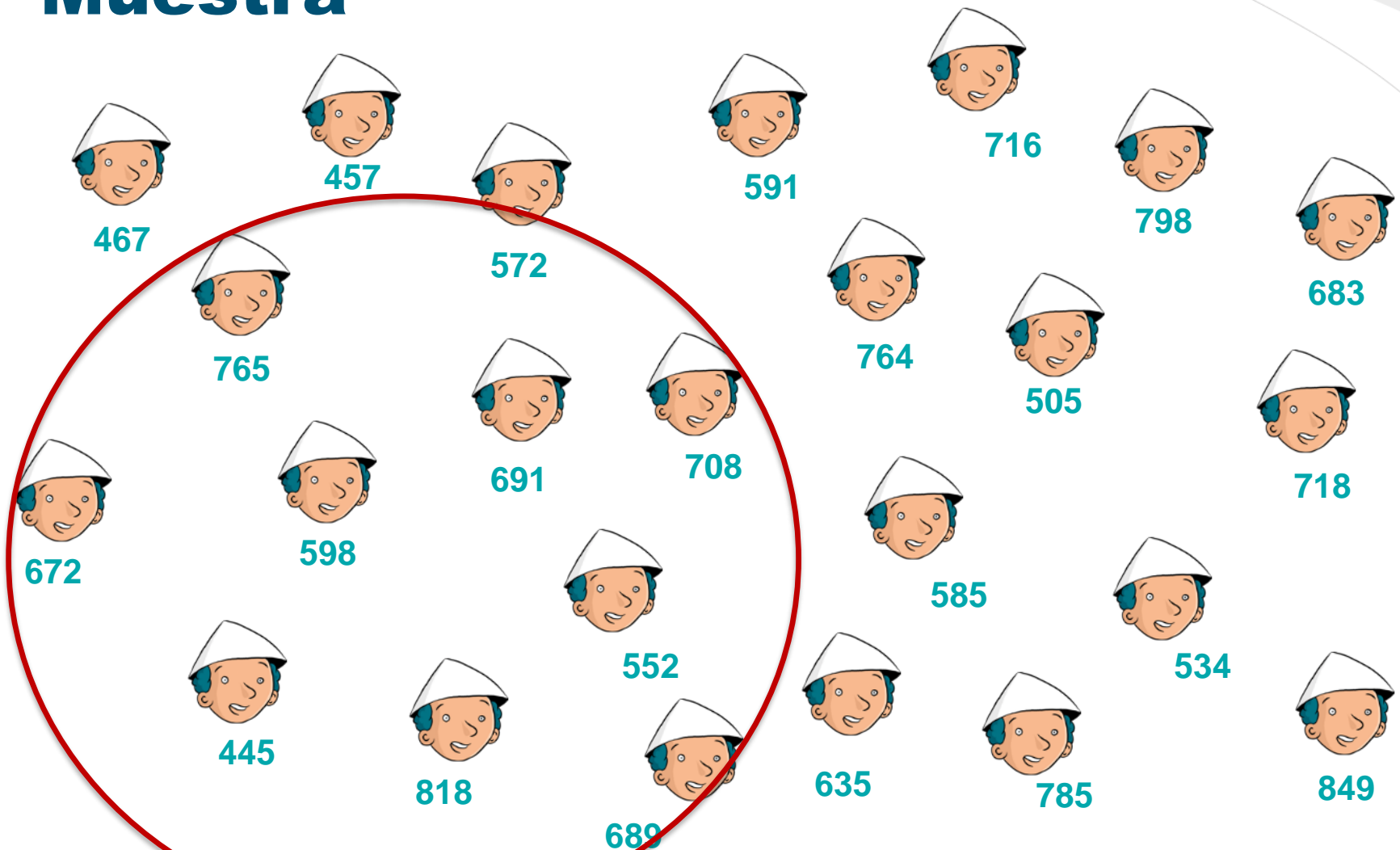
Media poblacional: 654

Muestra



Media muestral: 765
Media poblacional: 654

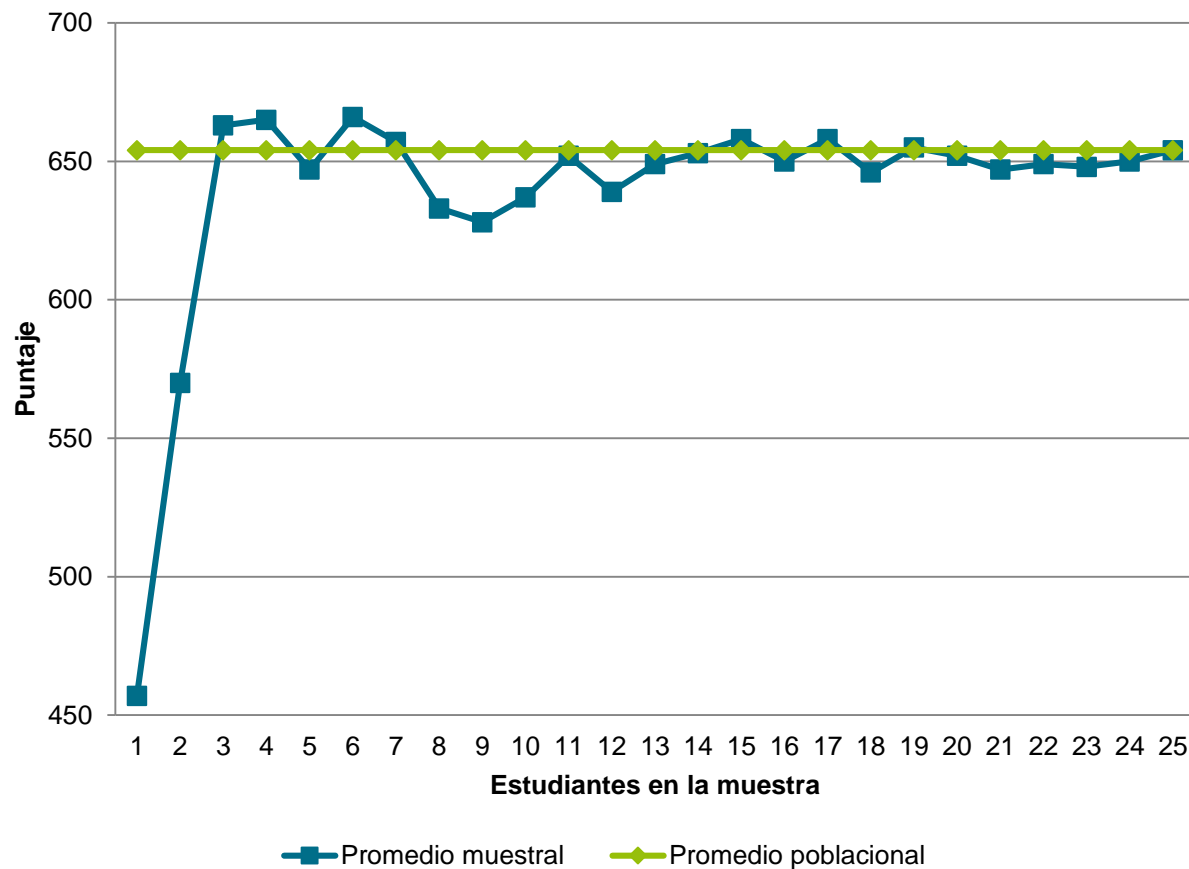
Muestra



Media muestral: 652

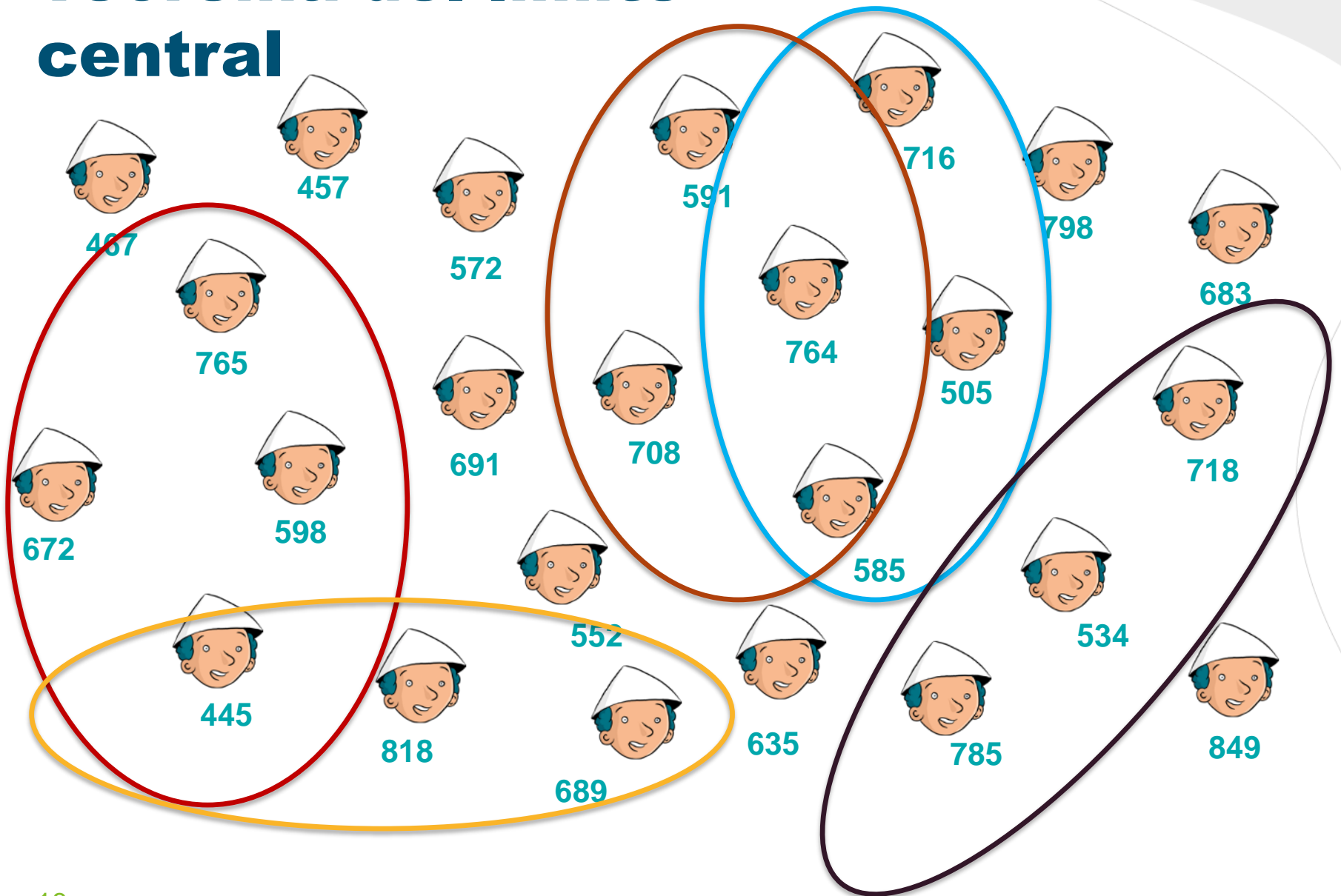
Media poblacional: 654

Ley de los GRANDES números

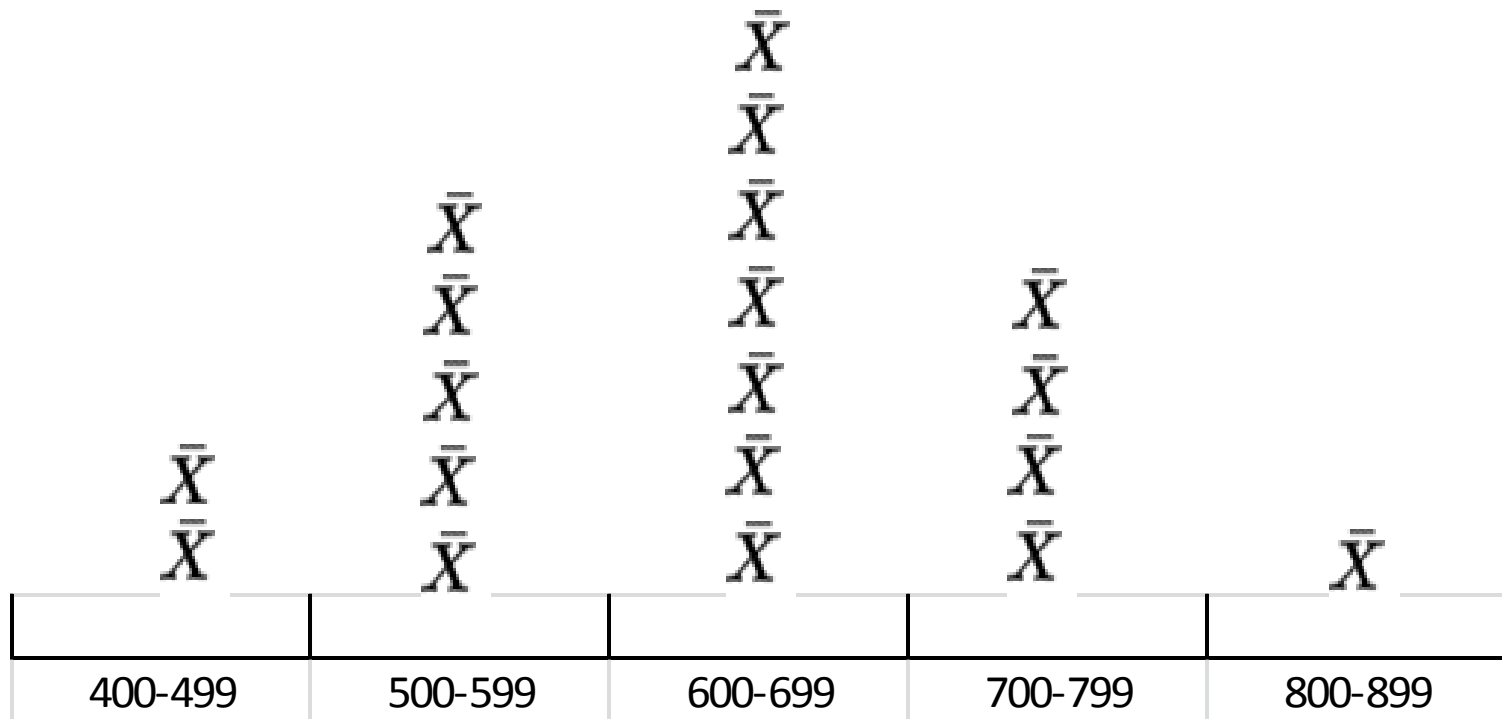


$$\Pr\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1.$$

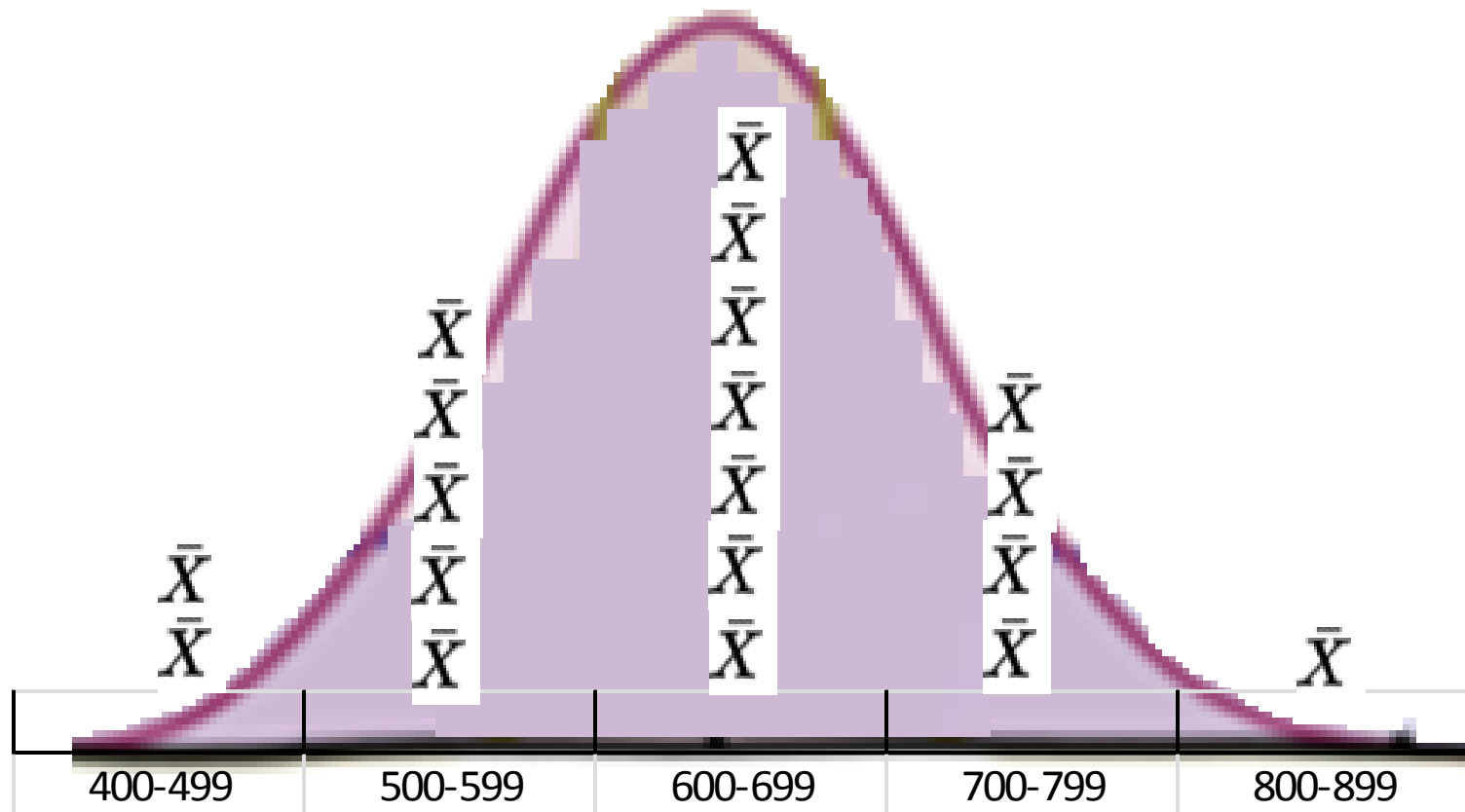
Teorema del límite central

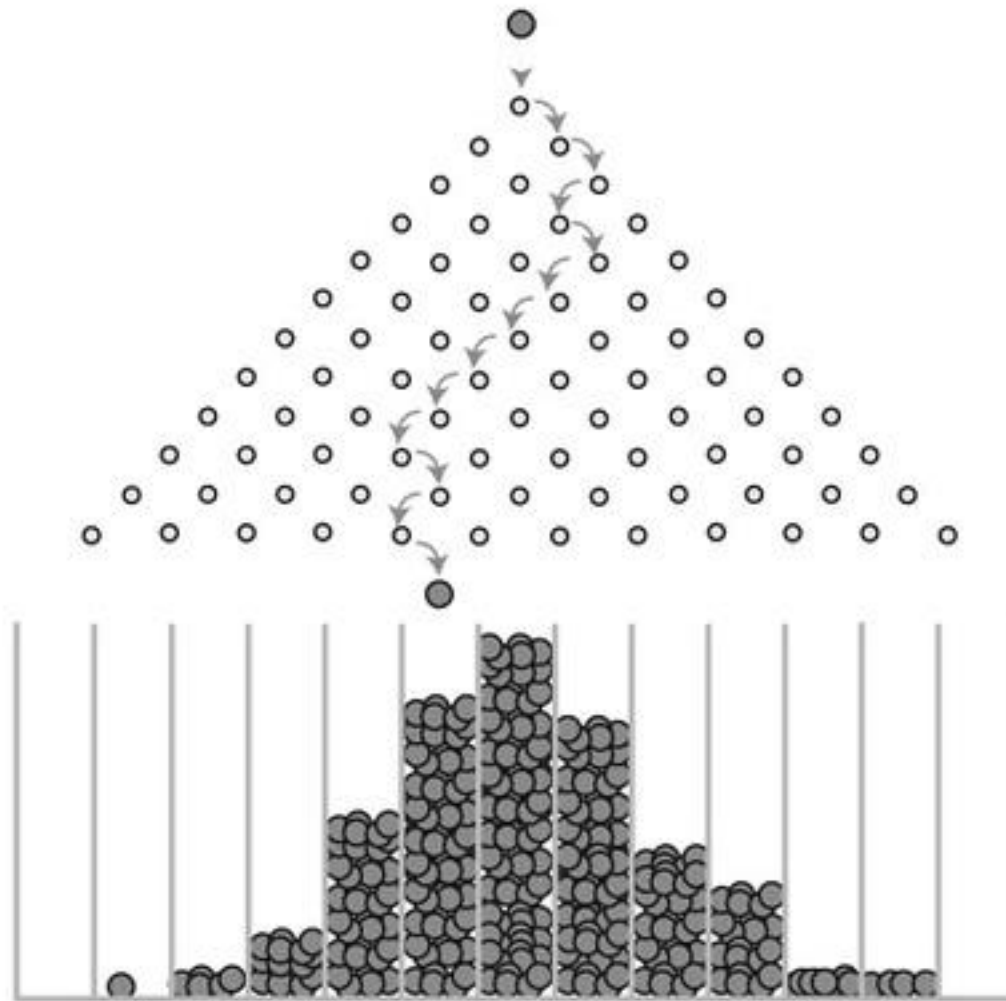


Histograma

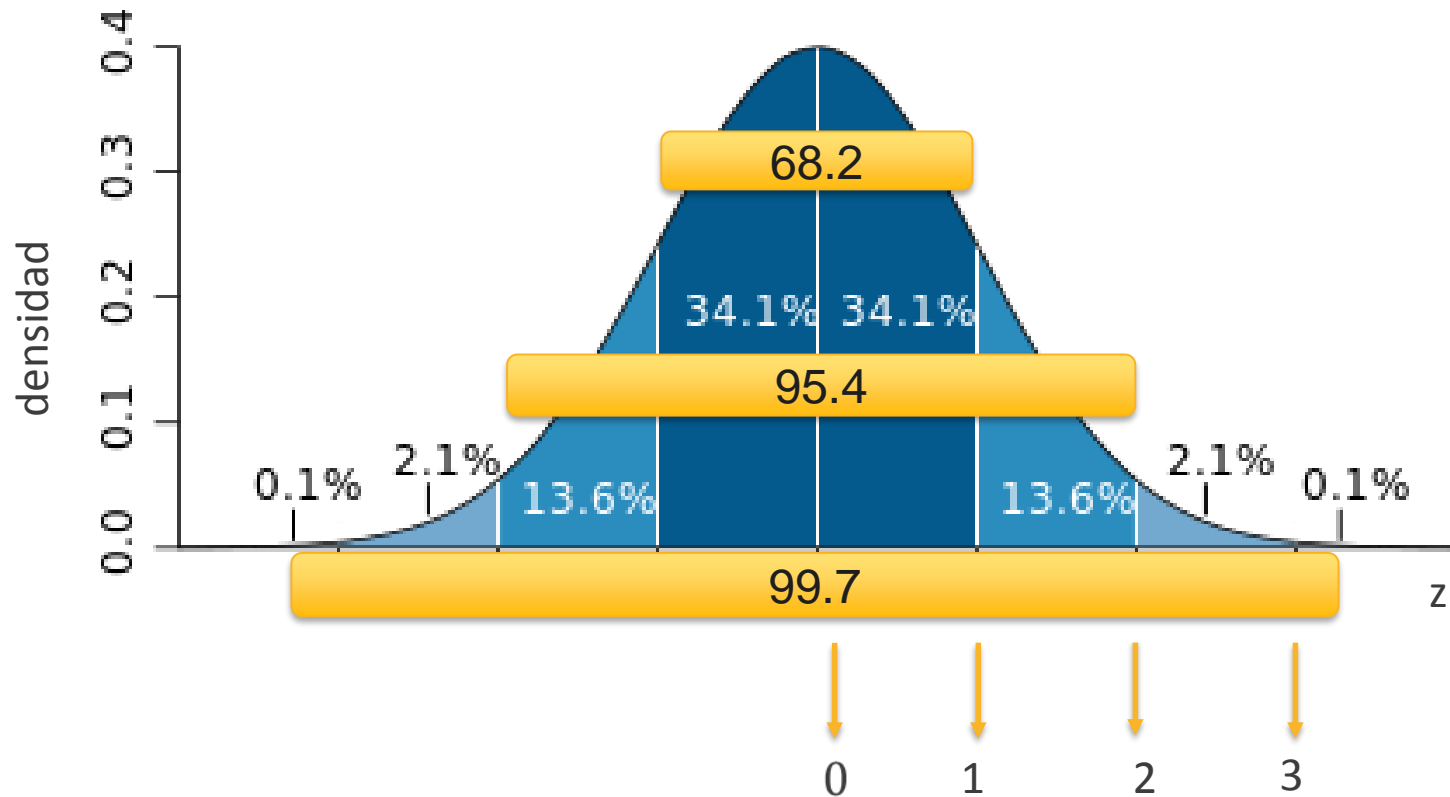


Teorema del límite central





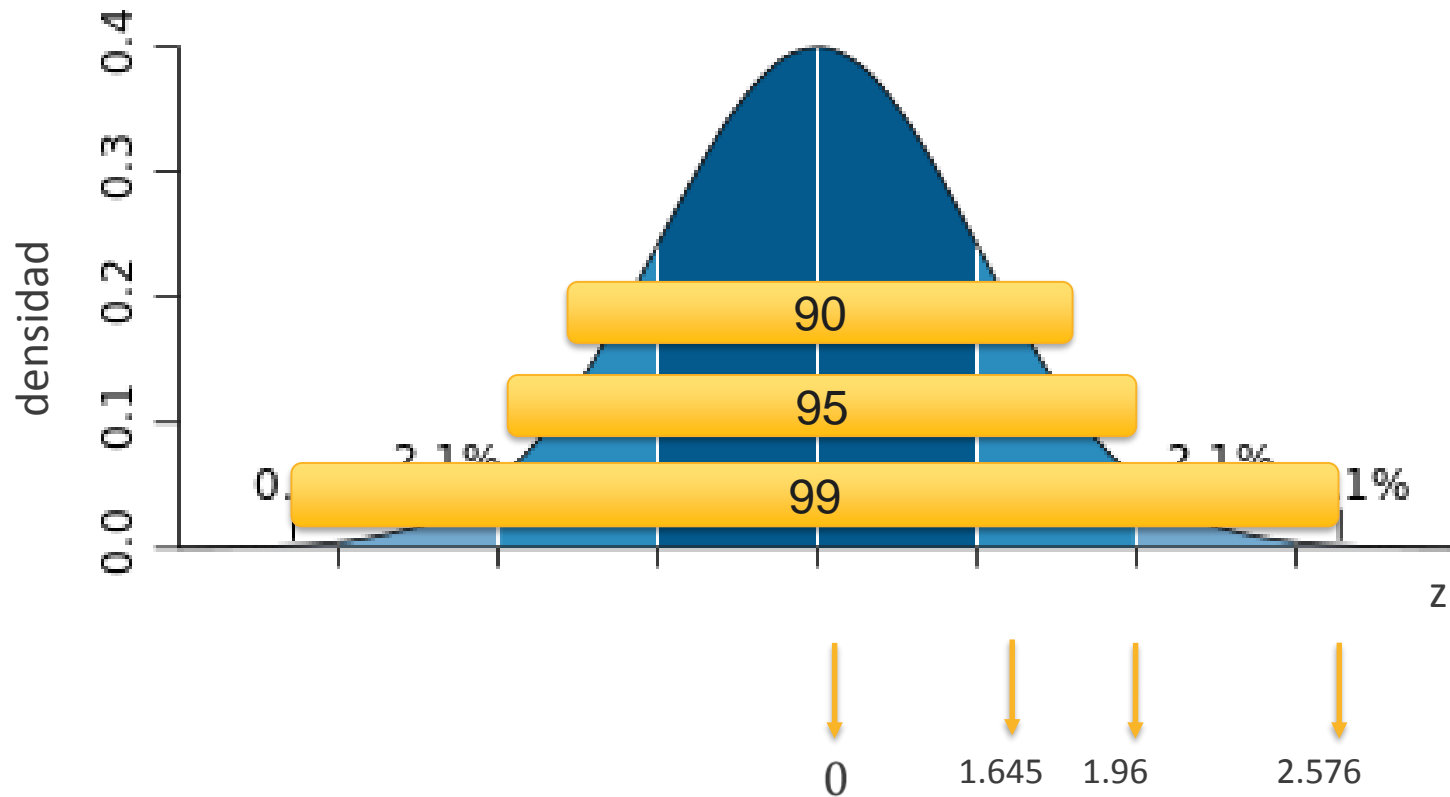
La regla tres-sigma o la regla empírica



Esta es una distribución normal con media 0 desviación estándar 1

$$z \sim N(0,1)$$

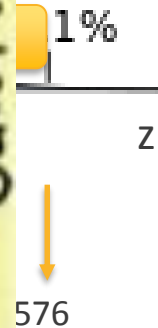
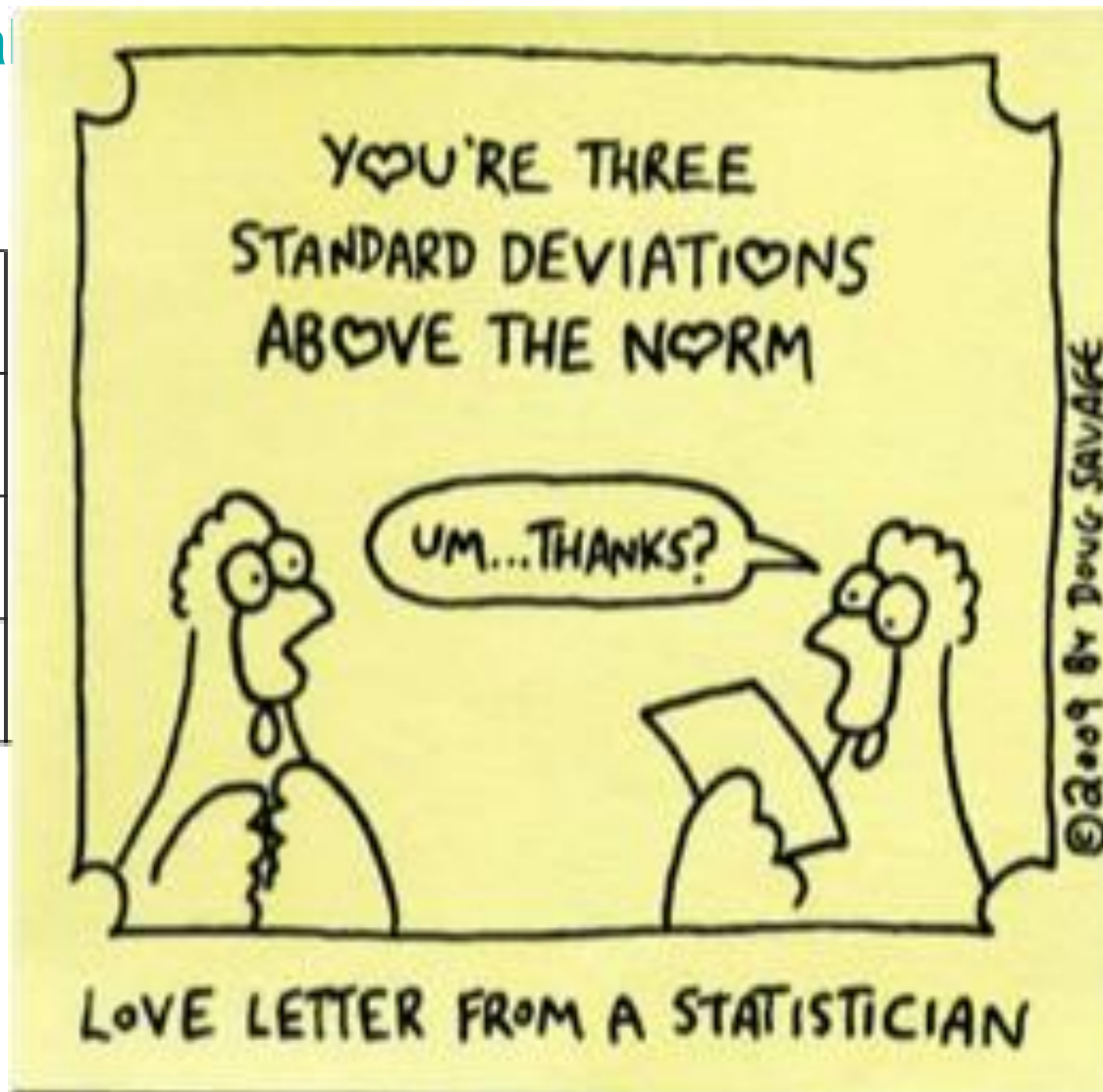
Intervalo de confianza



$$z \sim N(0,1)$$

Interval

densidad
0.0 0.1 0.2 0.3 0.4



$$Z \sim N(0,1)$$

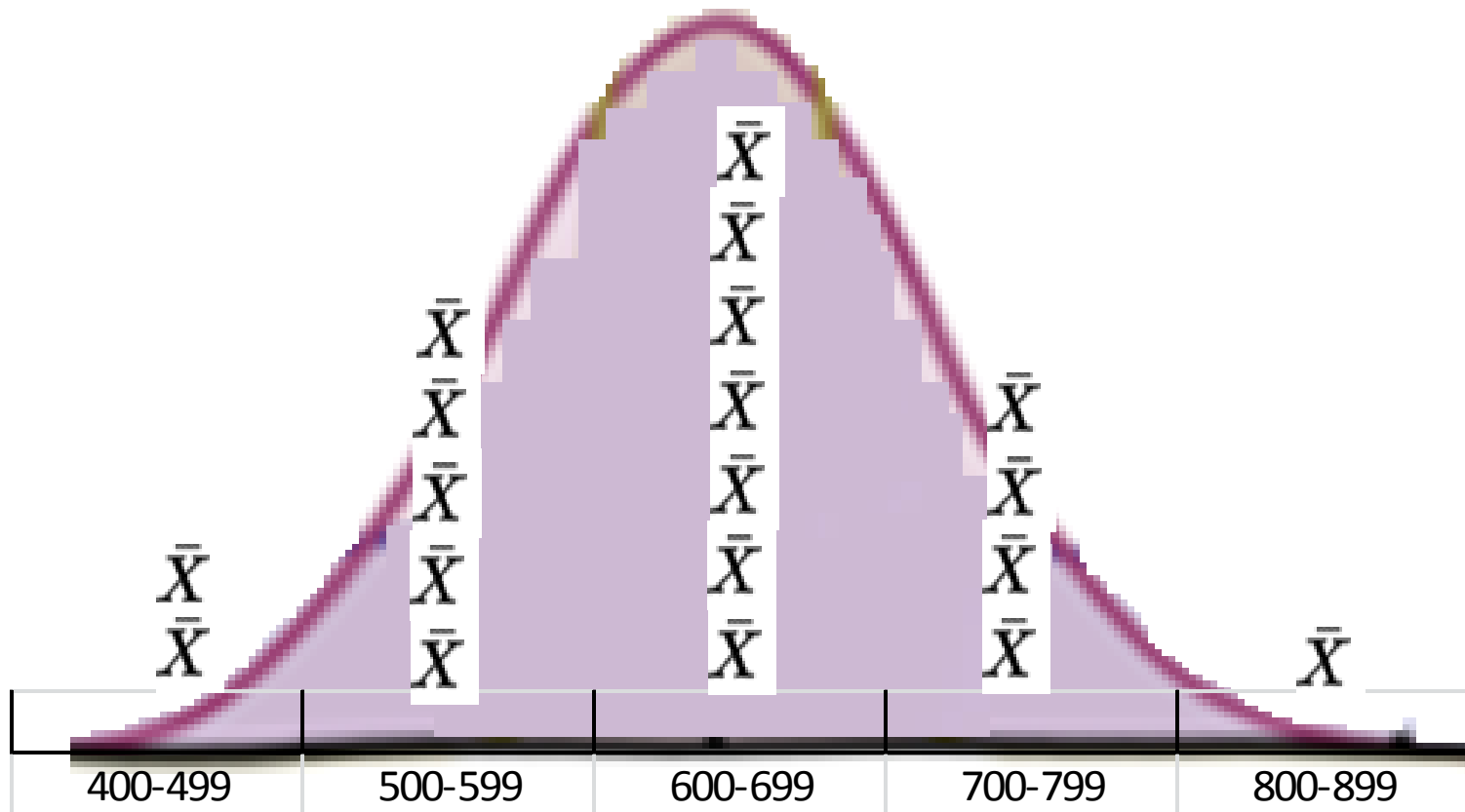
Teorema (del límite central): Sea X_1, X_2, \dots, X_n un conjunto de variables aleatorias, independientes e idénticamente distribuidas de una distribución con media μ y varianza $\sigma^2 \neq 0$. Entonces, si n es suficientemente grande, la variable aleatoria

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

tiene aproximadamente una distribución normal con

$$\mu_{\bar{X}} = \mu \text{ y } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Teorema del límite central

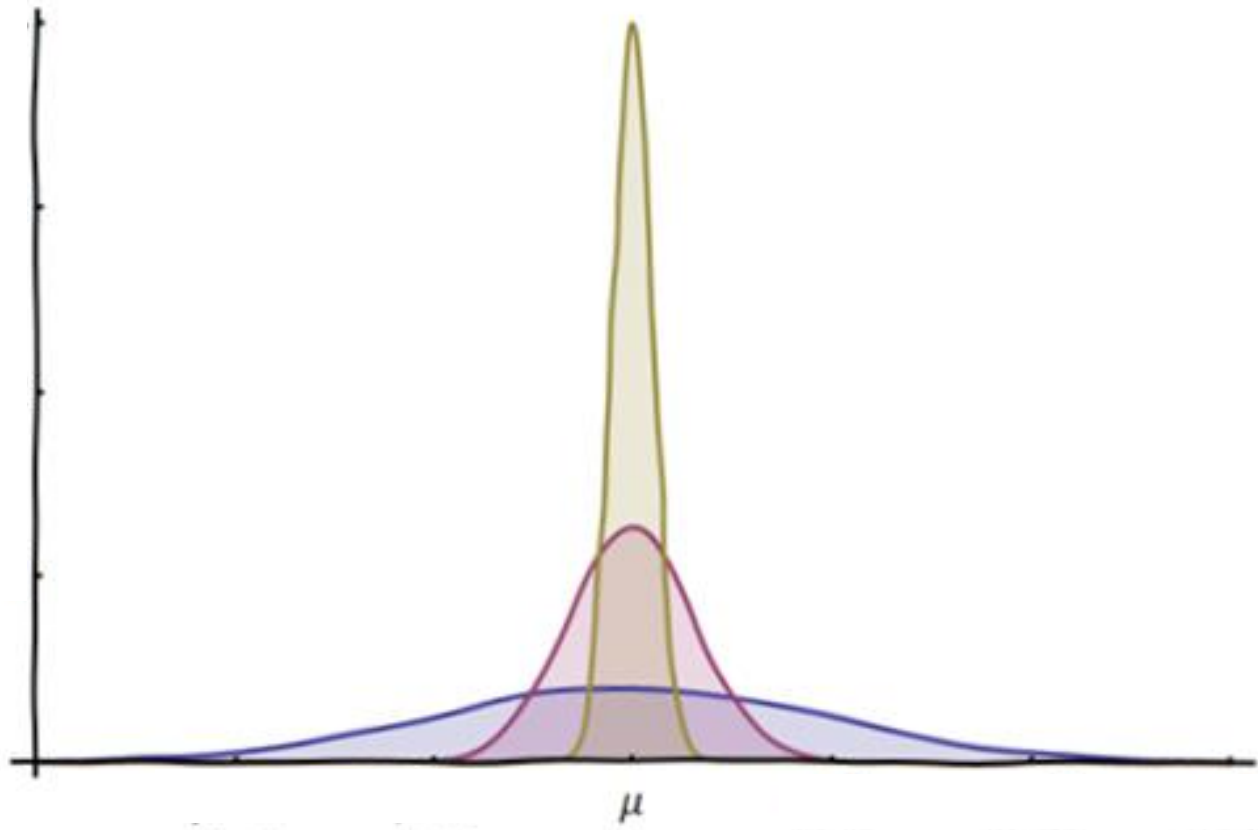


La desviación estándar del promedio se llama **error estándar**.

$$SE = SD / \sqrt{N}$$

Lección clave:

↑ **tamaño de la muestra** → ↓ **SE**



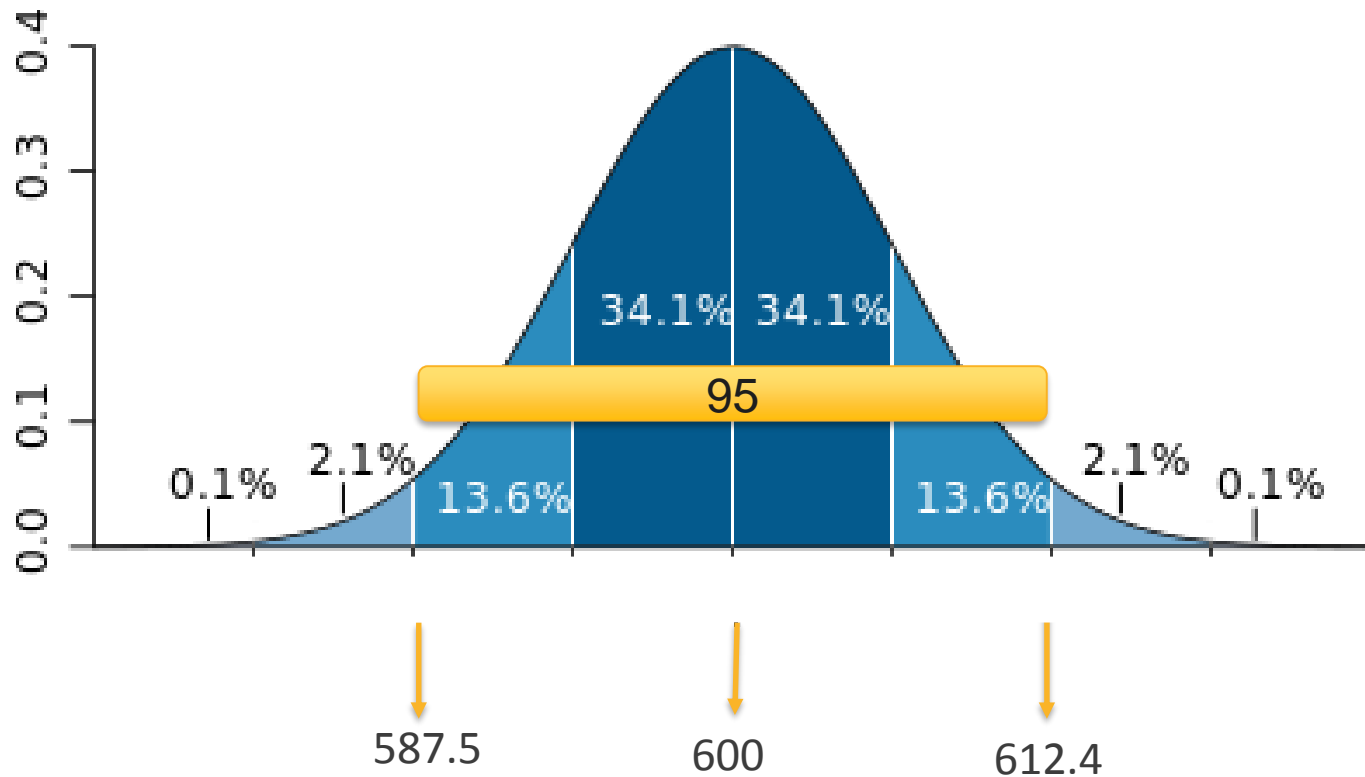
Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	600	6.3	100	587.5	612.4

Intervalo de confianza

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	600	6.3	100	587.5	612.4
	n	\bar{X}	σ/\sqrt{n}	σ		

Re-escalar la distribución normal estándar



$$587.5 = 600 - 1.96 * 6.3$$

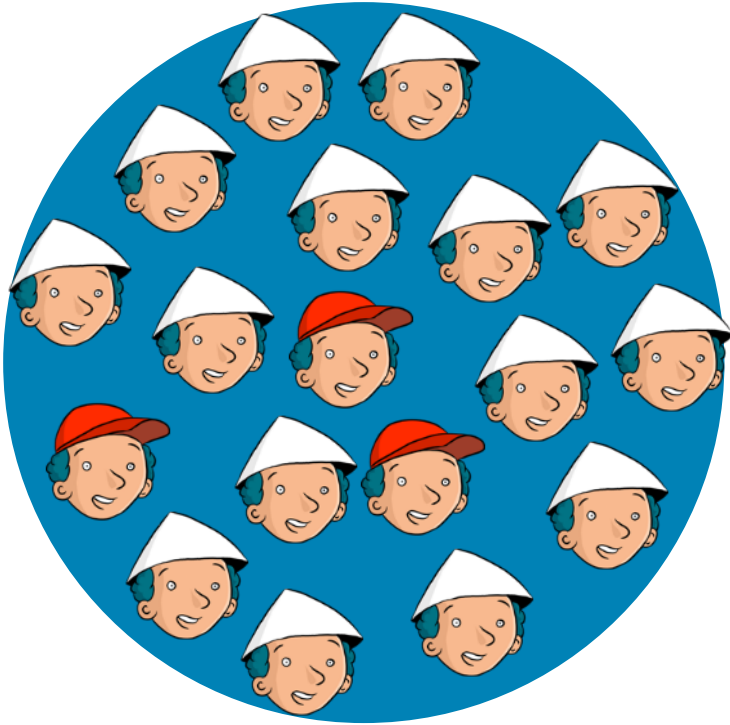
$$612.4 = 600 + 1.96 * 6.3$$

Intervalo de confianza

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

II. DIFERENCIA DE MEDIAS

Ahora tenemos 2 poblaciones



Escuela A



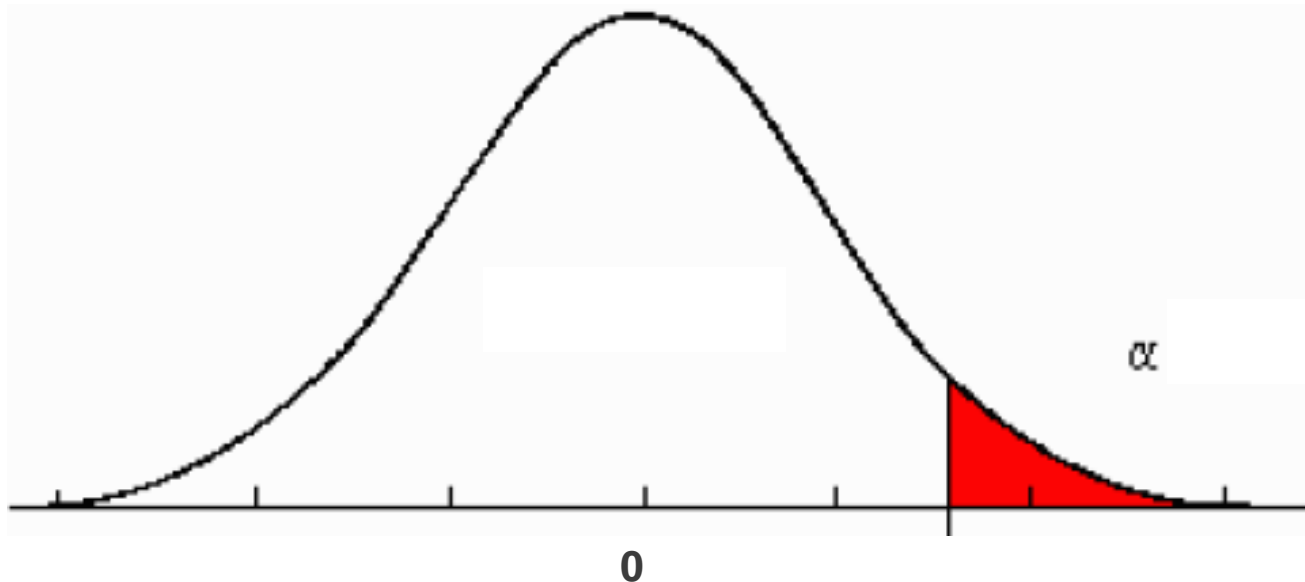
Escuela B

¿Son diferentes las escuelas A y B?

	N	mean	sd	min	p50	max
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849

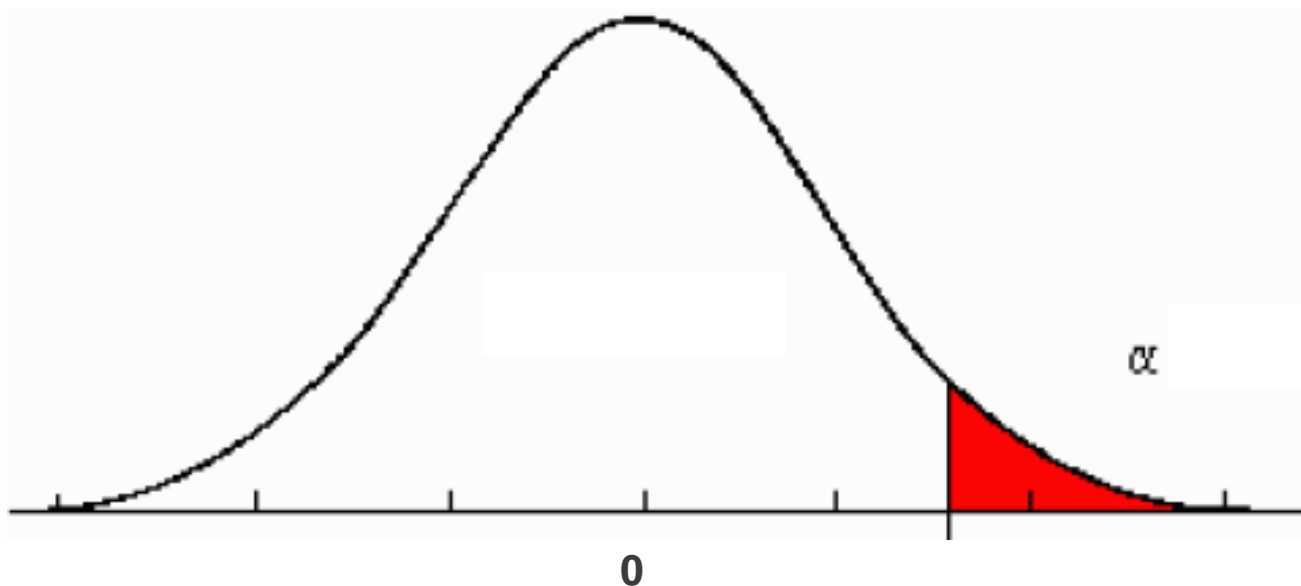
Son diferentes? Supongamos que no

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849
Diferencia		30	9		12	48



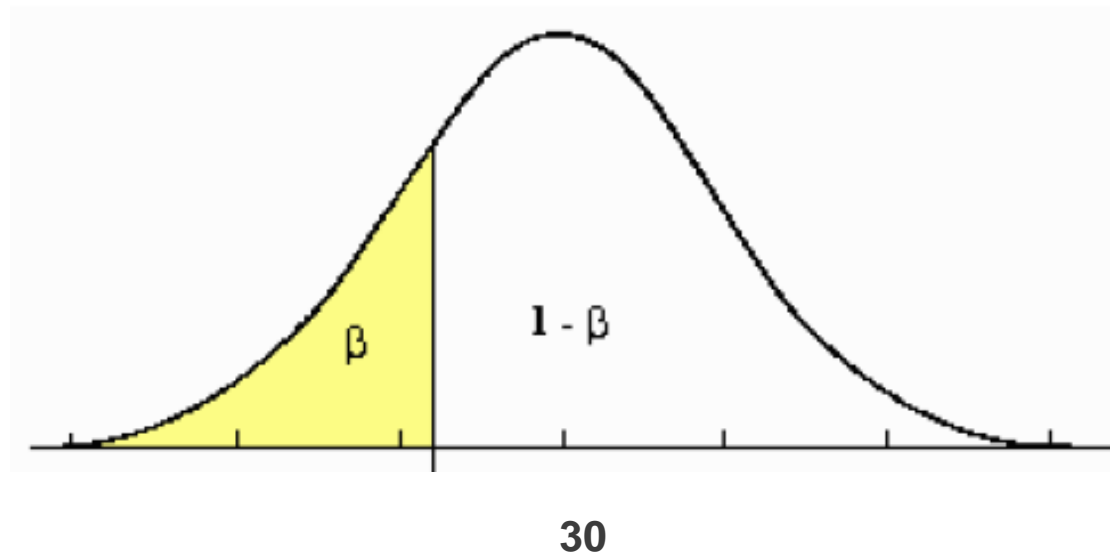
Son diferentes? Supongamos que no

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849
Diferencia		30	9		12	48



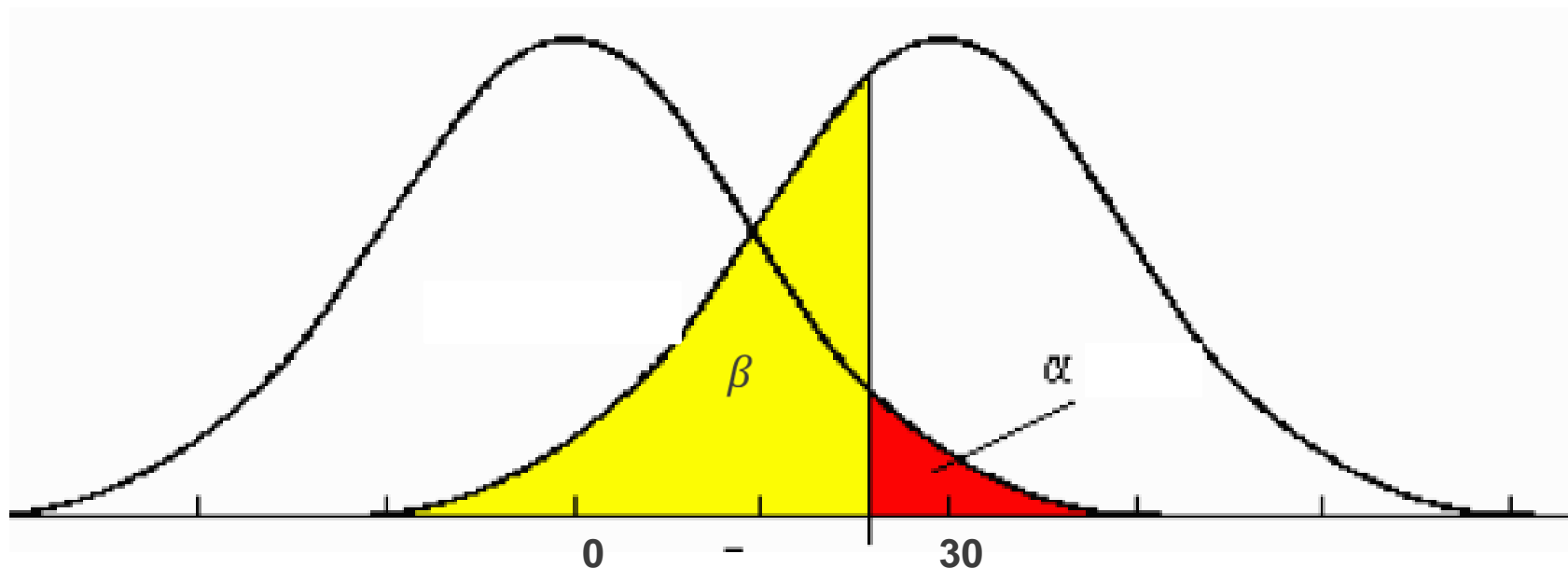
Ahora supongamos que si

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849
Diferencia		30	9		12	48



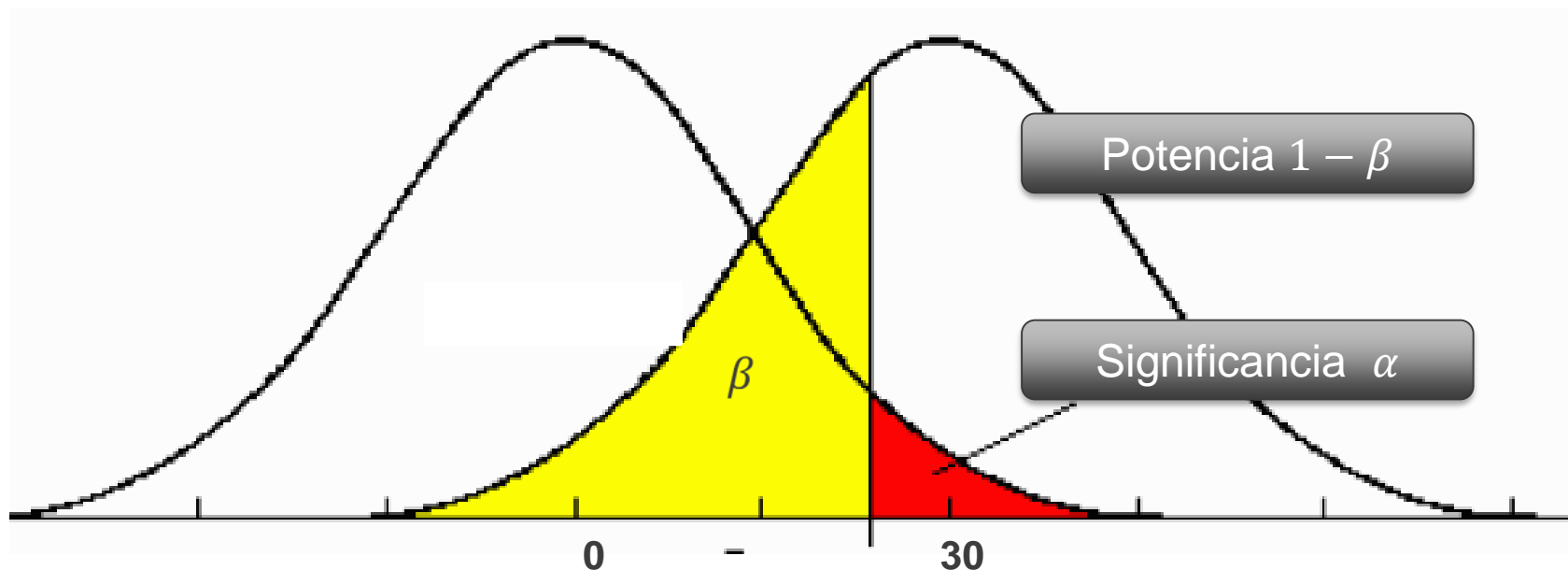
Errores tipo I y tipo II

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849
Diferencia		30	9		12	48



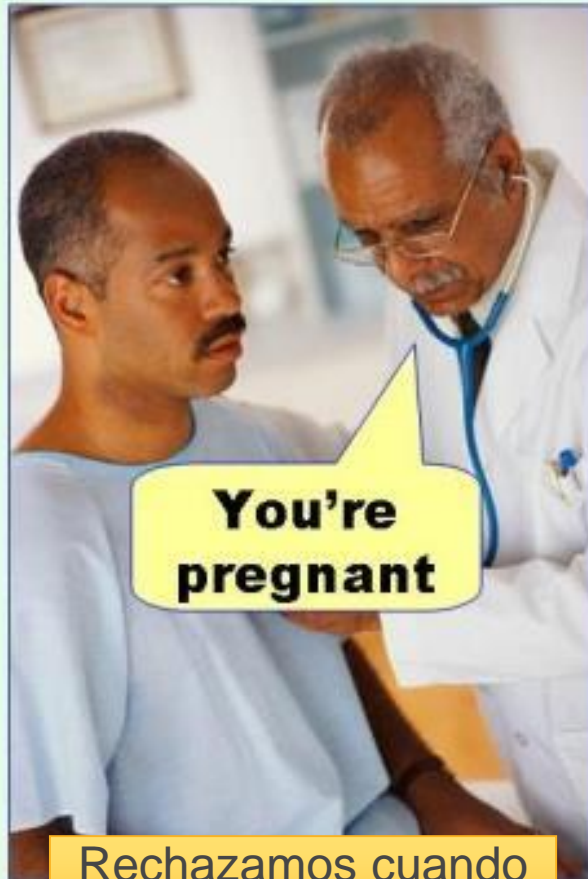
Errores tipo I y tipo II

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	630	100	392	628	879
Escuela B	250	600	100	362	598	849
Diferencia		30	9		12	48



Hipótesis nula H_0
Son iguales

Type I error
(false positive)



Rechazamos cuando
NO hay diferencia

Type II error
(false negative)



No rechazamos cuando
SI hay diferencia

¿Hubo impacto?

La noción de potencia

Lo que pasó en el mundo		
Lo que decidimos	Hipótesis nula H_0 Son iguales	Hipótesis alternativa H_A Son diferentes
No rechazar	$P(\text{aceptar correctamente } H_0) = 1 - \alpha$ OK	$P(\text{Error tipo II}) = \beta$ Error tipo II
Rechazar	$P(\text{Error tipo I}) = \alpha$ Error tipo I	$P(\text{rechazar correctamente } H_0) = 1 - \beta$ OK
Tamaño del efecto	Nivel de significancia	Potencia

Tratamiento	-0.24 (0.09)***
Obs.	500

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

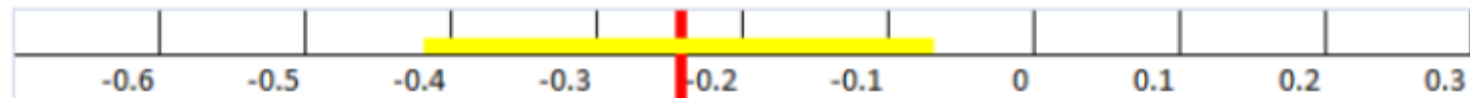


Figure: 500 observaciones

Tratamiento	-0.24 (0.1)**
Obs.	400

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

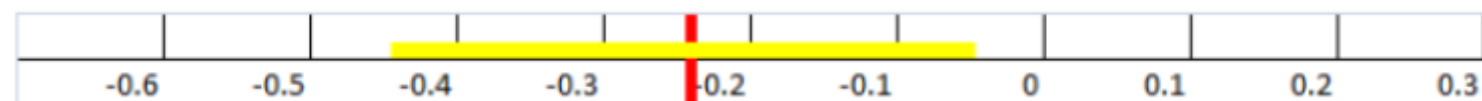


Figure: 400 observaciones

Tratamiento	-.25 (0.12)**
Obs.	300

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

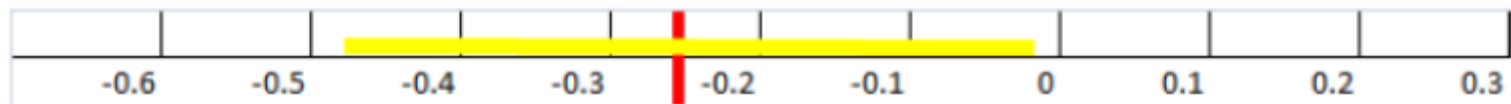


Figure: 300 observaciones

Tratamiento	-.26 (0.14)*
Obs.	200

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

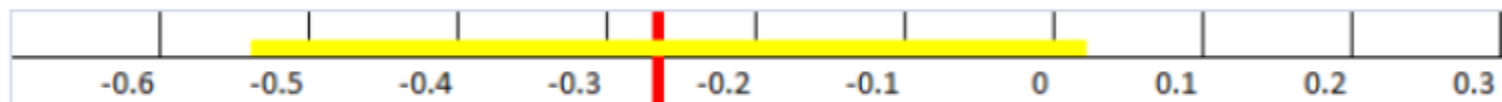


Figure: 200 observaciones

Tratamiento	-.19 (0.22)
Obs.	100

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

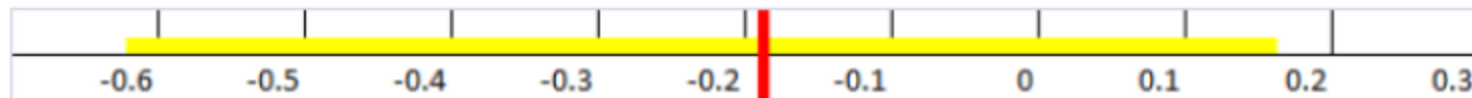


Figure: 100 observaciones

Ejercicio

Considera los siguientes datos:

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	600	6	100	588	612

Ejercicio 1: Dibuja la distribución de las calificaciones de la escuela A. Indica la media, la desviación estándar e indica valores para la regla 3 sigma.

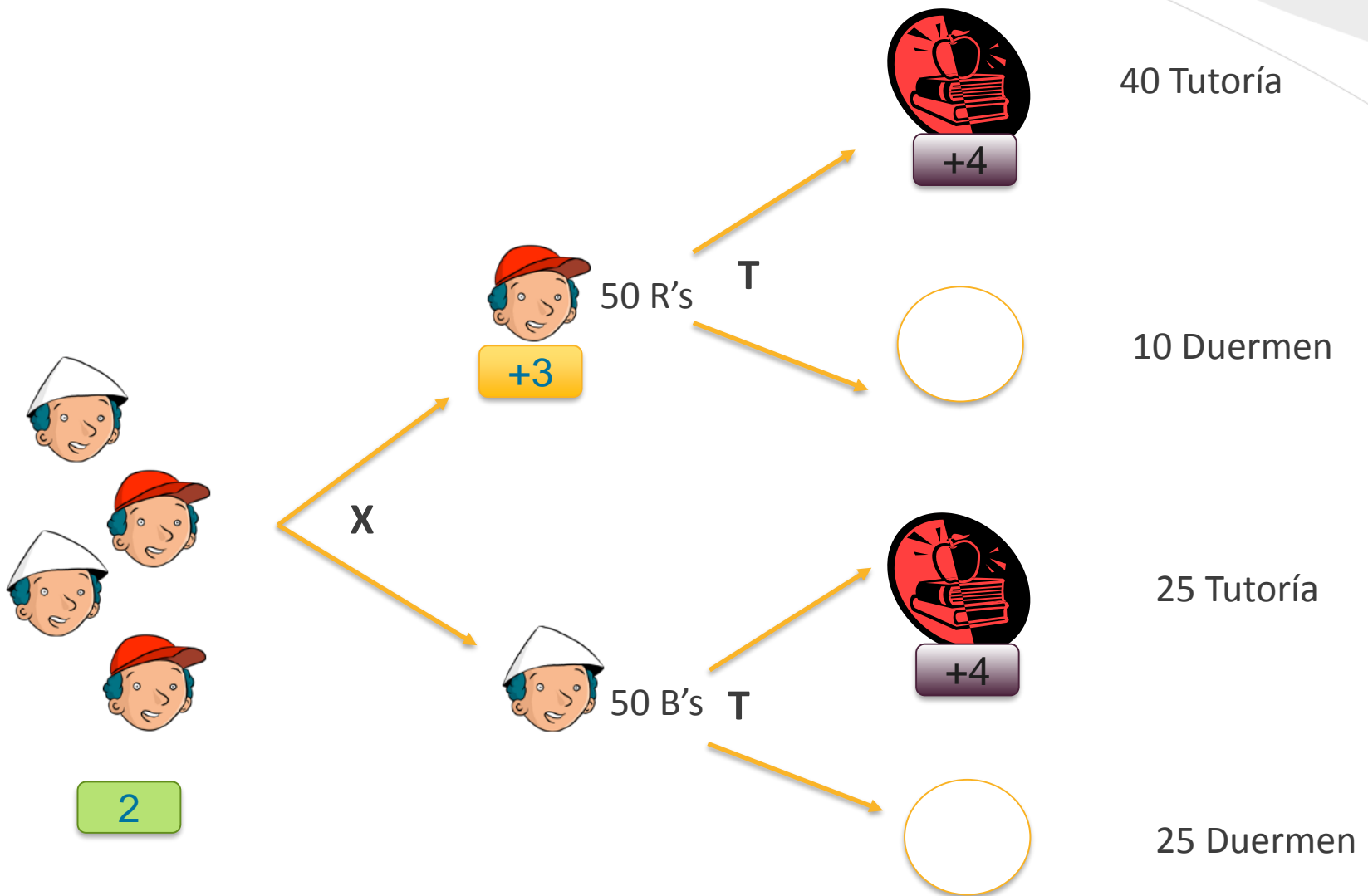
Ejercicio 2: Dibuja la distribución de **la media** de las calificaciones de la escuela A. Indica la media, el error estándar y el intervalo de confianza.

Ejercicio 3: El Subsecretario de educación dice que no se alcanzó la meta de lograr un puntaje promedio de 607 puntos. ¿Tiene razón?

Ejercicio 4: ¿Qué ventaja tendría tener una muestra de estudiantes más grande?

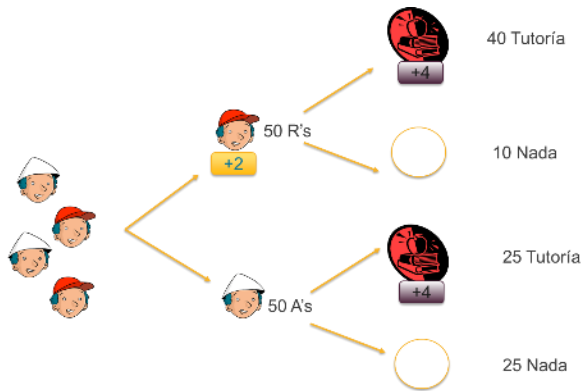
RETO EXTRA: ¿De qué tamaño tendría que ser la muestra para poder rechazar que la media es 607 al 5% de significancia?

III. IDENTIFICACIÓN



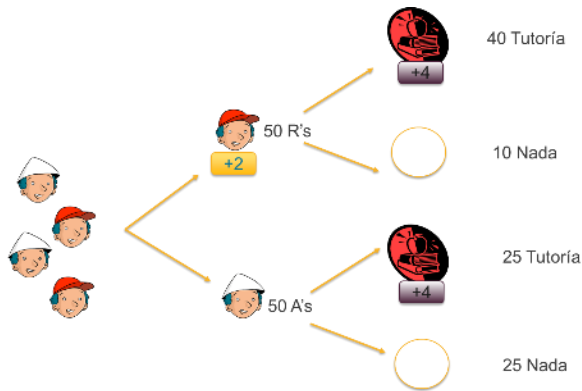
Intuición

1. Supongamos que hay una prueba de conocimiento con puntajes 0 a 10 denotado por Y (indicador de resultado).
2. Supongamos que hay 100 personas. Atributos:
 1. 50 con gorro rojo – “R’s”
 2. 50 con gorro blanco – “B’s”
3. Supongamos que todos conocen la respuesta de 2 preguntas.
4. Supongamos los R conocen la respuesta de 3 preguntas mas, pero los B’s no (es decir, el color del gorro X se asocia a Y. En otras palabras, el atributo esta asociados a resultados)
5. Supongamos que hay un programa de tutoría T en donde se les proporciona la respuesta a 4 preguntas. Éstas no coinciden con las que ya tienen los R’s.
6. Supongamos que la participación al programa es voluntaria (mecanismo de asignación) y resulta en 40 R’s y 25 B’s (Tratamiento asociado a atributo)



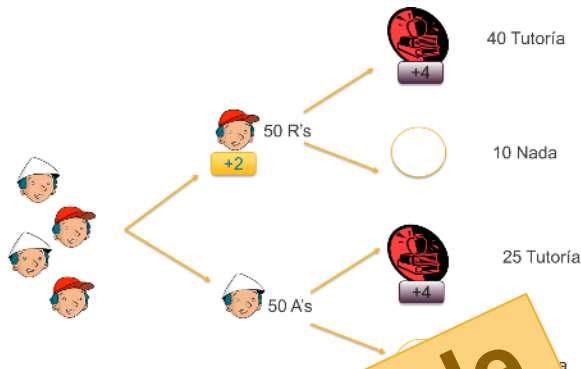
$$y_i = 2 + 4T_i + 3X_i$$

	T=1 (tutoría)	T=0 (duermen)	Promedio
X=1 (Rs)	40 personas ____ puntos	10 personas ____ puntos	____ puntos
X=0 (Bs)	25 personas ____ puntos	25 personas ____ puntos	____ puntos
Promedio	____ puntos	____ puntos	



$$y_i = 2 + 4T_i + 3X_i$$

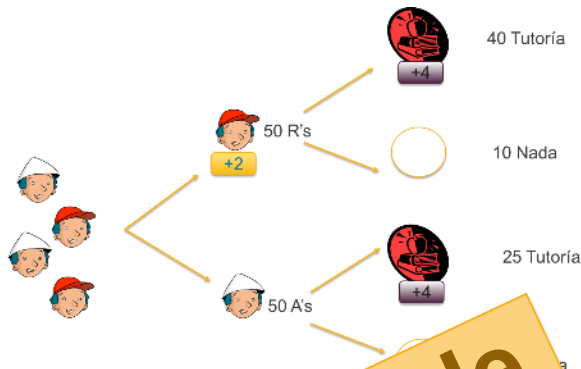
	T=1 (tutoria)	T=0 (duermen)	Promedio
X=1 (Rs)	40 personas $Y=2+3+4$ $Y=9$ puntos	10 personas $Y=2+3+0$ $Y=5$ puntos	$E(Y X=1)=(40/50)(9)+(10/50)(5)=$ 8.2 puntos
X=0 (Bs)	25 personas $Y=2+0+4$ $Y=6$ puntos	25 personas $Y=2+0+0$ $Y=2$ puntos	$E(Y X=0)=(25/50)(6)+(25/50)(2)=$ 4 puntos
Promedio	$E(Y T=1)=(40/65)(9)+(25/65)(6)=$ 7.8 puntos	$E(Y T=0)=(10/35)(5)+(25/35)(2)=$ 2.9 puntos	



$$y_i = 2 + 4T_i + 3X_i$$

Diferencia de promedios entre T y C es 7.8-2.9=4.9!

	(Tutoría)	T=0 (duermen)	Promedio
X=1	40 personas Y=2+3+0 Y=5 puntos	10 personas Y=2+3+0 Y=5 puntos	$E(Y X=1)=(40/50)(9)+(10/50)(5)=$ 8.2 puntos
X=0	25 personas Y=2+0+4 Y=6 puntos	25 personas Y=2+0+0 Y=2 puntos	$E(Y X=0)=(25/50)(6)+(25/50)(2)=$ 4 puntos
Promedio	$E(Y T=1)=(40/65)(9)+(25/65)(6)=$ 7.8 puntos	$E(Y T=0)=(10/35)(5)+(25/35)(2)=$ 2.9 puntos	

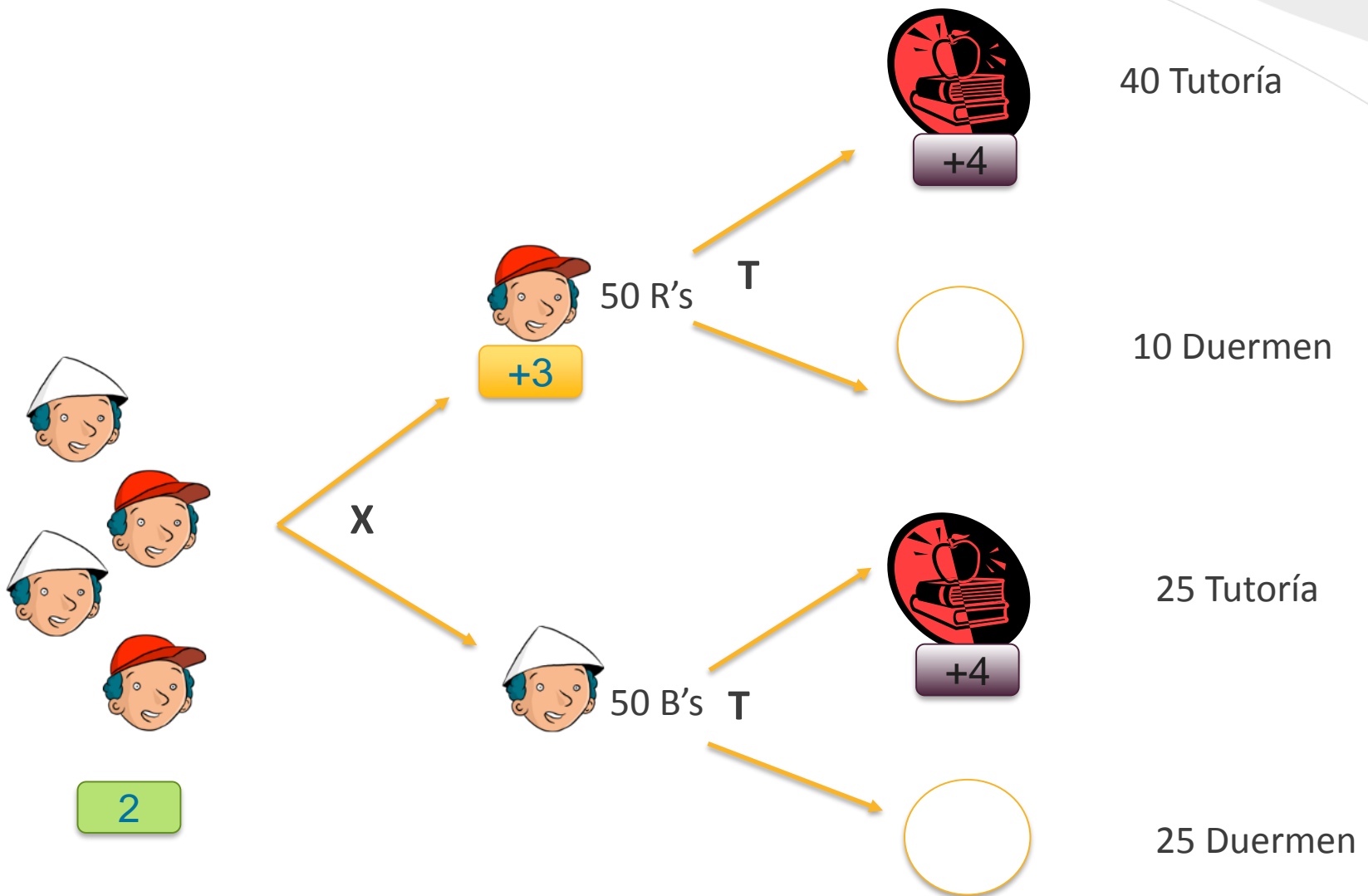


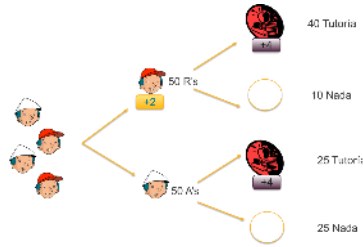
$$y_i = 2 + 4T_i + 3X_i$$

Diferencia de promedios entre T y C es $7.8 - 2.9 = 4.9!$

Decimos que una comparación de tratamientos no identifica el efecto del programa (es decir, que se puede separar del efecto del color de la gorra)

	Tutoría	Control	Promedio
X	1	0	$E(X=1) = (40/50) = 0.8$
Y	$Y = 2 + 0 + 4$	$Y = 2 + 0 + 0$	
	$Y = 6$ puntos	$Y = 2$ puntos	
Promedio	$E(Y T=1) = (40/65)(9) + (25/65)(6) = 7.8$ puntos	$E(Y T=0) = (10/35)(5) + (25/35)(2) = 2.9$ puntos	





No te puede ignorar el color de la gorra para evaluar

¿Por qué no?

Porque el color de la roja se relaciona con la tutoría.

Decimos que el tratamiento es **endógeno** (que se origina en virtud de causas internas.)

2 ejemplos de variables endógenas a calificaciones para niños

- La **riqueza del hogar** facilita el acceso a los programas y a la compra de libros, tutorías y actividades culturales. (Relativamente fácil de observar)
- La **motivación** facilita el acceso a los programas, y el número de horas invertidas en estudiar. (Difícil de observar)
- Algo de "jargón" de evaluadores:
 - Los evaluadores decimos "**observar**" pero pensamos "medir".
 - Los evaluadores decimos "difícil" pero pensamos "costoso"

Una dosis de realidad...

$$y_i = 2 + 4T_i + 3X_i$$



No podemos observar todos los atributos

Modelo

$$y_i = 2 + 4T_i + 3X_i$$

$$y_i = 2 + 4T_i + \varepsilon_i$$

Regresión

$$y_i = \alpha + \varepsilon_i$$



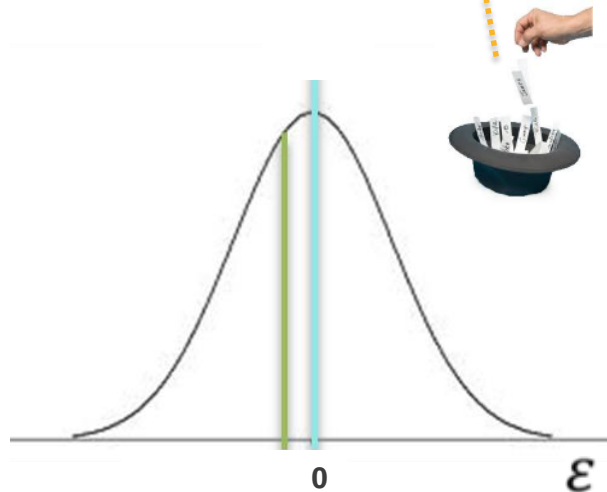
Carlos: $y = 600$

$$+ 20 = 620$$



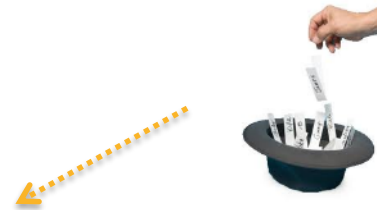
Tomás: $y = 600$

$$+ 1 = 601$$



Efecto del programa

$$Y_i = \alpha + \varepsilon_i$$



$$Y_{Tomás} = 600 + 20 = 620$$



$$Y_{Carlos} = 600 + (-1) = 599$$


$$E[Y] = \alpha$$

Efecto del programa

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

Promedio sin programa

Efecto promedio del programa



The diagram illustrates the components of the regression equation $Y_i = \alpha + \beta T_i + \varepsilon_i$. Three orange arrows point from the equation to descriptive text: one from α to 'Promedio sin programa', one from βT_i to 'Efecto promedio del programa', and one from ε_i to an illustration of a hand pulling a card from a top hat. The hat contains several cards with names like 'Katie', 'Mike', 'Garry', and 'Andrew' written on them, representing the random assignment of individuals to treatment or control groups.

$T_i=0$ Sin programa

$$Y_i = \alpha + \varepsilon_i$$

$T_i=1$ Con programa


$$Y_i = \alpha + \beta + \varepsilon_i$$

Efecto del programa

$$Y_i = \alpha + \beta T_i + \varepsilon_i$$

Promedio sin programa

Efecto promedio del programa

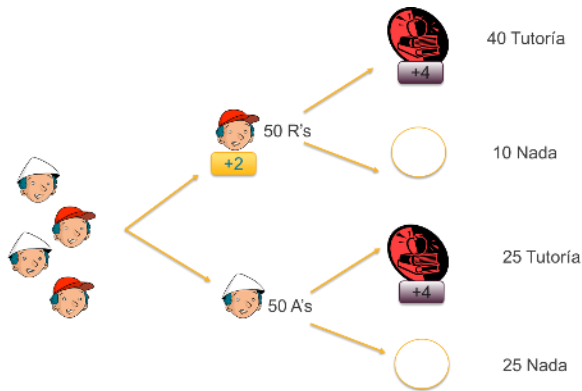


The diagram illustrates the components of the regression equation $Y_i = \alpha + \beta T_i + \varepsilon_i$. An arrow points from the term α to the text 'Promedio sin programa'. Another arrow points from the term βT_i to the text 'Efecto promedio del programa'. A third arrow points from the error term ε_i to a top hat filled with cards, with a hand pulling one out, representing the random assignment of the treatment variable T_i .

T=0 Sin programa $E[Y|T=1] = \alpha$

T=1 Con programa $E[Y|T=0] = \alpha + \beta$

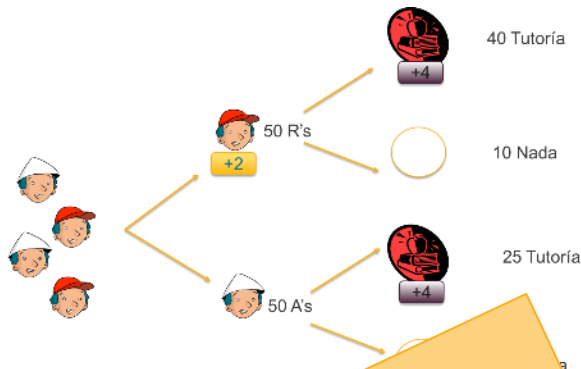
$$\beta = E[Y|T=1] - E[Y|T=0]$$



$$y_i = \alpha + \beta T_i + \varepsilon_i$$

4.9

	T=1 (tutoria)	T=0 (duermen)	Promedio
X=1 (Rs)	40 personas Y=2+3+4 Y=9 puntos	10 personas Y=2+3+0 Y=5 puntos	$E(Y X=1)=(40/50)(9)+(10/50)(5)=$ 8.2 puntos
X=0 (Bs)	25 personas Y=2+0+4 Y=6 puntos	25 personas Y=2+0+0 Y=2 puntos	$E(Y X=1)=(25/50)(6)+(25/50)(2)=$ 4 puntos
Promedio	$E(Y T=1)=(40/65)(9)+(25/65)(6)=$ 7.8 puntos	$E(Y T=1)=(10/35)(5)+(25/35)(2)=$ 2.9 puntos	



$$y_i = \alpha + \beta T_i + \varepsilon_i$$

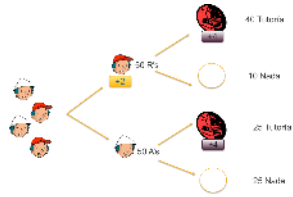
4.9

Estimación
sesgada en 0.9
porque el error
es endógeno

Tutoría)	T=0 (duermen)	Promedio
----------	---------------	----------

Condición para
estimador insesgado.
Exogeneidad: $\varepsilon \perp T$

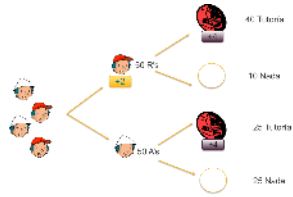
Promedio	$E(Y T=1) = (40/65)(9) + (25/65)(6) =$ 7.8 puntos	$E(Y T=0) = (10/35)(5) + (25/35)(2) =$ 2.9 puntos
----------	---	---



$$y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$$

4

	T=1 (tutoria)	T=0 (duermen)
X=1 (Rs)	40 personas Y=9-3=6 puntos	10 personas Y=5-3=2 puntos
X=0 (Bs)	25 personas Y=6 puntos	25 personas Y=2 puntos
Promedio	$E(Y T=1)=(40/65)(9-3)+(25/65)(6)=$ 6 puntos	$E(Y T=1)=(10/35)(5-3)+(25/35)(2)=$ 2 puntos



$$y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$$

4

Los controles
pueden ayudar
a disminuir
sesgo

T=1 (tutoria)

T=0 (duermen)

Condición para
estimador insesgado.
Exogeneidad: $\varepsilon \perp T, X$

6 puntos

2 puntos

Resumen

variable	N	mean	sd	min	p50	max
puntaje	25	654	118	445	683	849

Grupo	Obs	Media	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Escuela A	250	600	6.3	100	587.5	612.4
Escuela B	250	630	6.3	100	617.5	642.4
Diferencia		-30	8.9		-47.6	-12.4

Tratamiento	-.24 (0.09)***
Obs.	500

* significativo al 90%,
 ** significativo at 95%,
 *** significativo at 99%

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

Resumen

Estadística básica descriptiva: min, max, mediana, media y desviación estándar, histogramas

Muestreo: Muestreo aleatorio simple, teorema del límite central y ley de los grandes números.

Comparación de poblaciones: Comparación de medias, potencia, significancia, atributo, mecanismo de asignación, identificación, exogeneidad, sesgo, identificación.

Referencias

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2016 **Impact Evaluation in Practice**. 2nd edition. Washington, D.C.: World Bank



 Rosangelab@iadb.org

 [@the_IDB](https://twitter.com/the_IDB)