

Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño
ccarrenovi@gmail.com

Abril, 2021

Modulo 2: Componentes Principales Y Análisis De Correspondencias

- Análisis estadístico preliminar de los datos
- Introducción al Análisis estadístico
- Introducción al análisis estadístico de datos multivariantes

Análisis Estadístico Preliminar de los Datos

- El Análisis estadístico preliminar de los datos consiste en realizar un análisis descriptivo de la muestra
- Permite controlar posibles errores en la selección de los datos como valores fuera de rango, outliers
- Proporciona una idea de la forma de los datos, su posible distribución de probabilidad y sus parámetros de centralización: como:
 - Media
 - Mediana
 - Moda
- Proporciona una idea de la dispersión de los datos
 - Varianza
 - Desviación típica

Medidas de Tendencia Central

- Media
- Mediana
- Moda

Media

- Formula para el calculo de la media

LA MEDIA \bar{x}

Si las observaciones son x_1, x_2, \dots, x_n .

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

$$\bar{x} = \frac{\sum x_i}{n}$$

Ejemplo: Medidas de Tendencia Central

- Mediana, la observación que divide en dos partes el conjunto de datos
- Moda, la observación que mas de repite o de mayor frecuencia

```
1 v1 = c(10,15,15,25,30,48,51)
2 median(v1)
3 mean(v1)
4 library(modes)
5 modes(v1)
6 |
```

6:1 (Top Level) ⬇

Console Terminal x Jobs x

~/ ↗

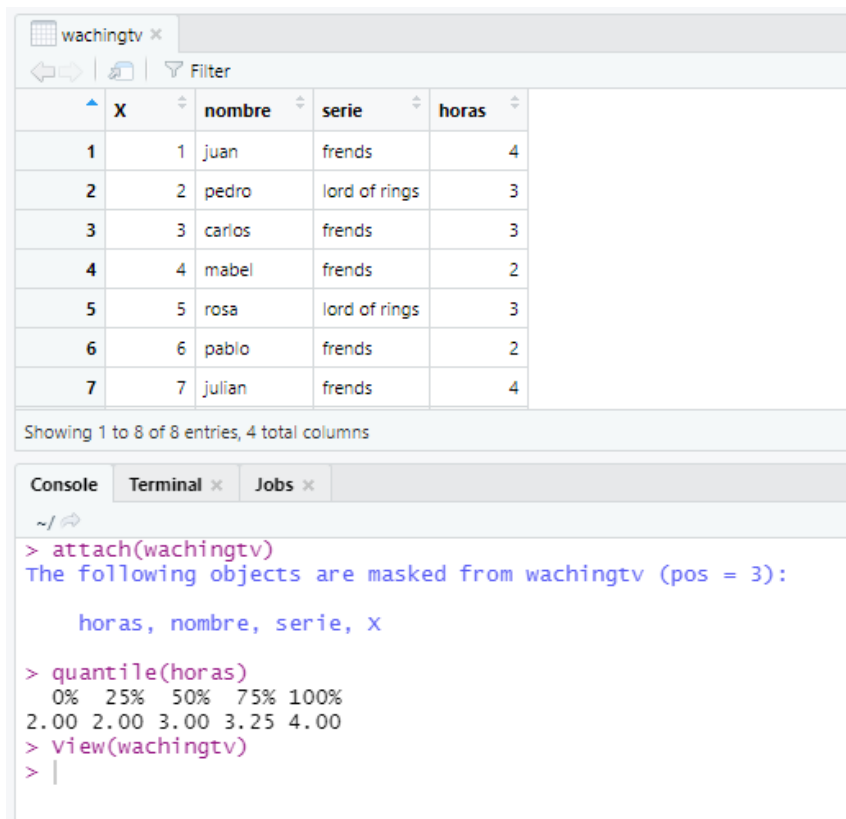
```
> v1 = c(10,15,15,25,30,48,51)
> median(v1)
[1] 25
> mean(v1)
[1] 27.71429
> library(modes)
> modes(v1)
      [,1]
value    15
Length    2
> |
```

Medidas de Dispersion

- Quantiles
- Varianza
- Desviación Estándar

Quantiles

- Ejemplo de mostrar los quantiles en R



The screenshot shows a RStudio window with a data table and a console. The table has 4 columns: X, nombre, serie, and horas. The console shows the following R commands and output:

```
> attach(wachingtv)
The following objects are masked from wachingtv (pos = 3):
    horas, nombre, serie, x

> quantile(horas)
 0%  25%  50%  75% 100% 
2.00 2.00 3.00 3.25 4.00 
> view(wachingtv)
> |
```

	X	nombre	serie	horas
1	1	juan	friends	4
2	2	pedro	lord of rings	3
3	3	carlos	friends	3
4	4	mabel	friends	2
5	5	rosa	lord of rings	3
6	6	pablo	friends	2
7	7	julian	friends	4

Showing 1 to 8 of 8 entries, 4 total columns

Varianza y Desviación Estándar

- Formula de la Varianza y Desviación Estándar

Varianza

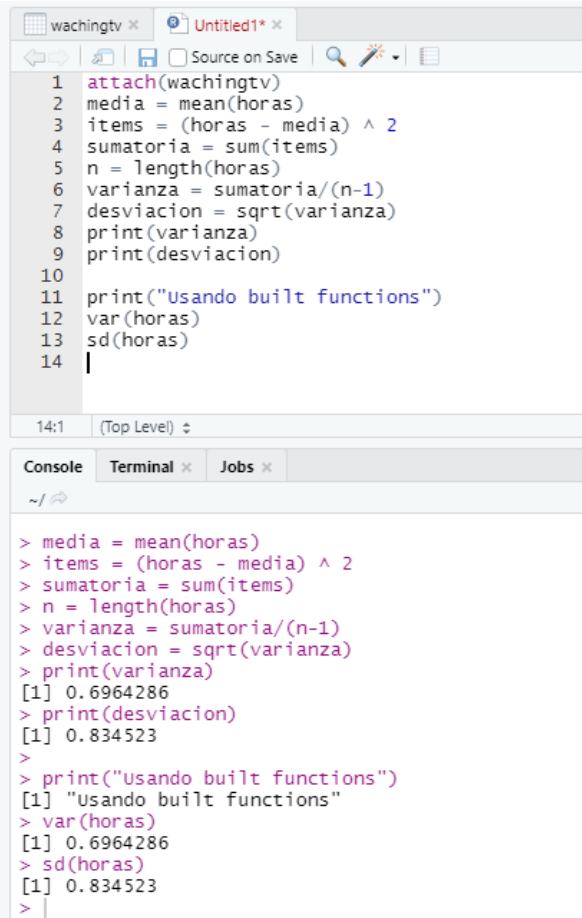
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Desviación Estándar

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Ejemplo Varianza y Desviación Estándar

- Implementando las funciones en R

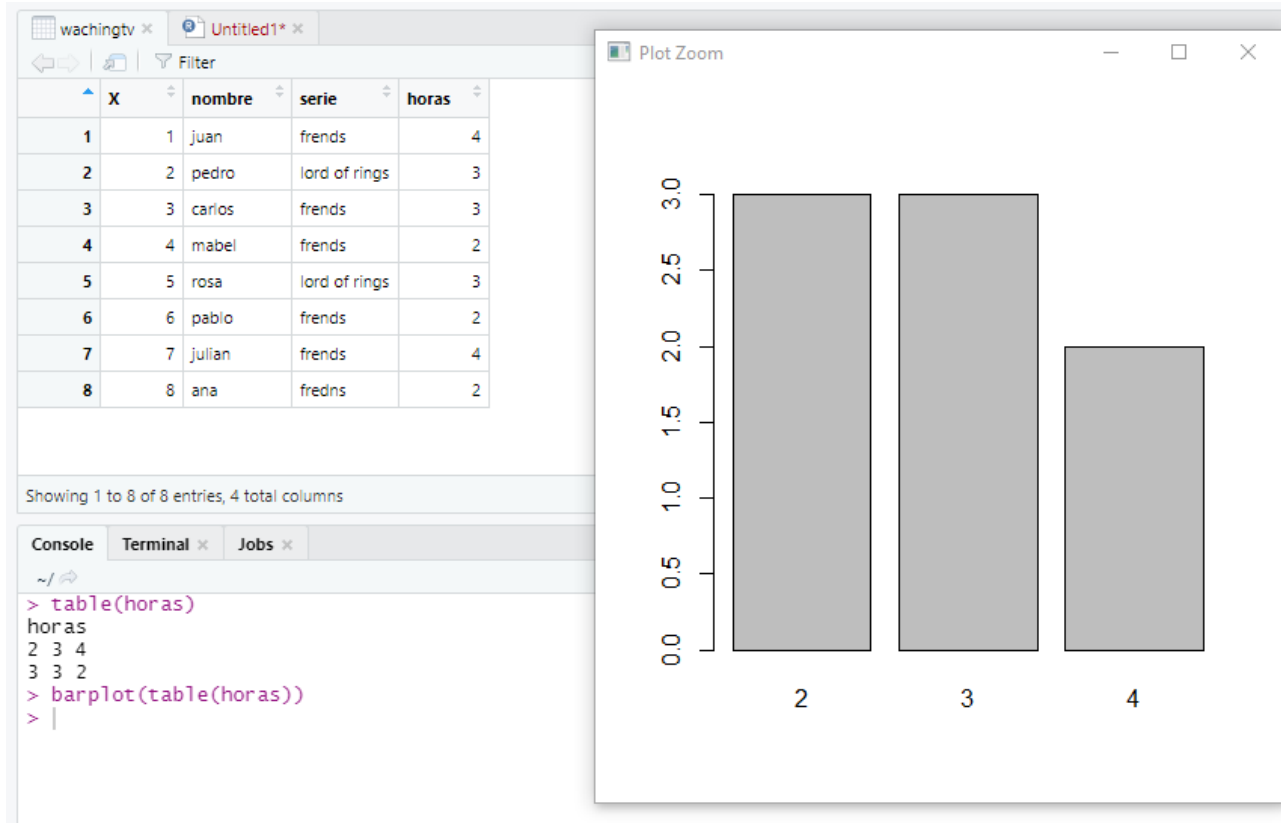


```
wachingt v x  Untitled1* x
Source on Save
1 attach(wachingt v)
2 media = mean(horas)
3 items = (horas - media) ^ 2
4 sumatoria = sum(items)
5 n = length(horas)
6 varianza = sumatoria/(n-1)
7 desviacion = sqrt(varianza)
8 print(varianza)
9 print(desviacion)
10
11 print("usando built functions")
12 var(horas)
13 sd(horas)
14 |

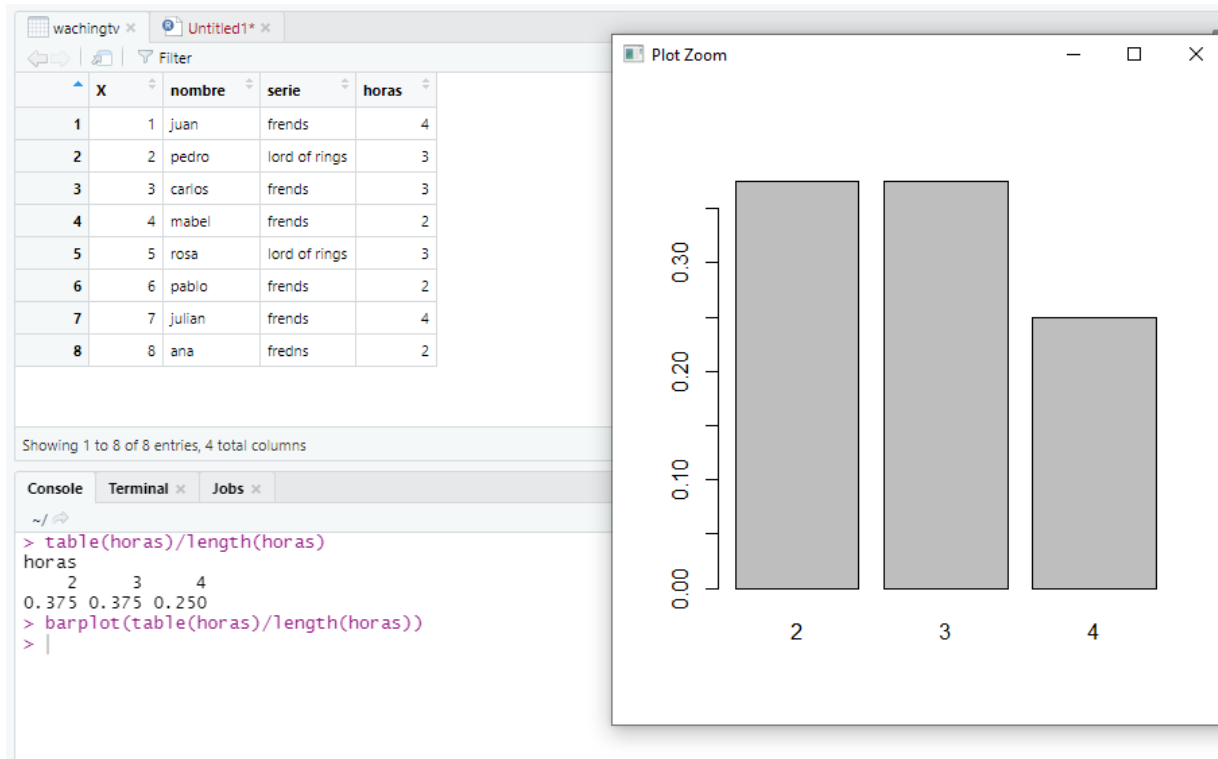
14:1 (Top Level)
Console Terminal x Jobs x
~/
> media = mean(horas)
> items = (horas - media) ^ 2
> sumatoria = sum(items)
> n = length(horas)
> varianza = sumatoria/(n-1)
> desviacion = sqrt(varianza)
> print(varianza)
[1] 0.6964286
> print(desviacion)
[1] 0.834523
>
> print("usando built functions")
[1] "usando built functions"
> var(horas)
[1] 0.6964286
> sd(horas)
[1] 0.834523
>
```

Tabla de Frecuencias

- Usa la función `table()` para ver las frecuencias de las observaciones

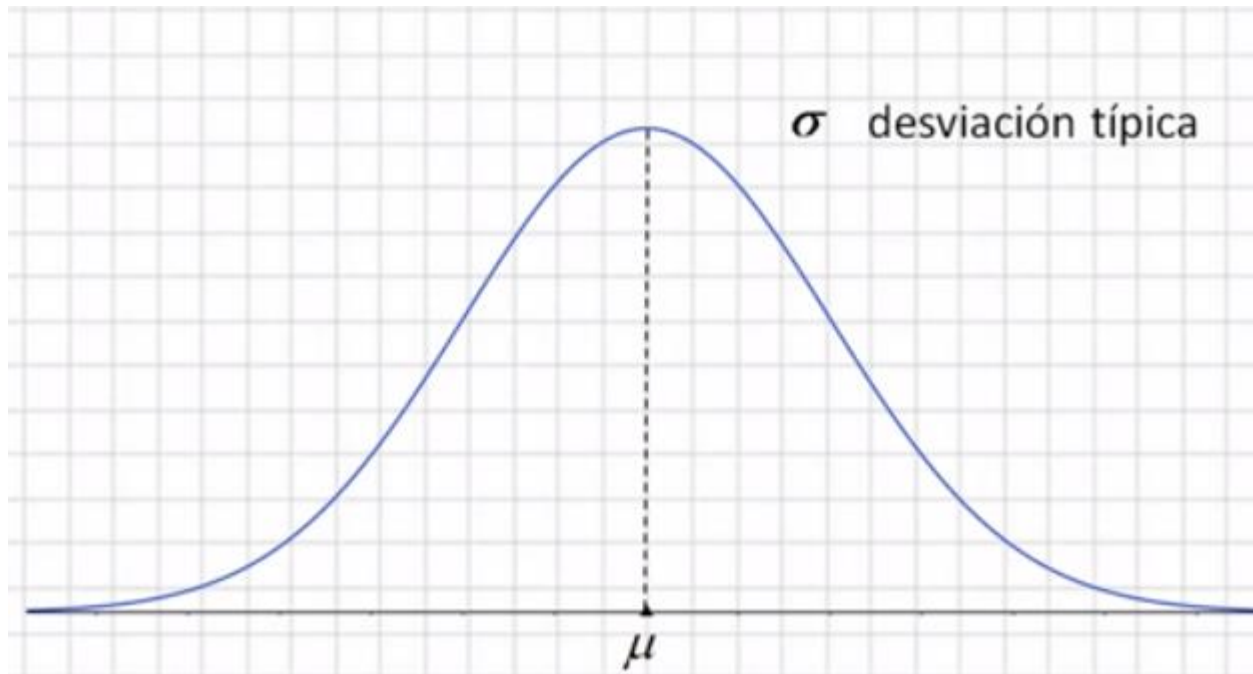


- Tabla de frecuencias relativas
- ¿Que ocurre si multiplicas por 100 a los valores relativos?



Distribución Normal

- Que es la distribución normal

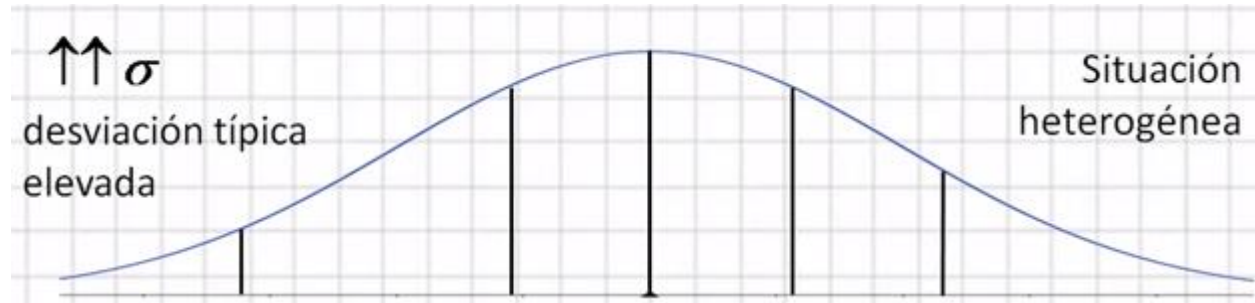


Distribución Normal

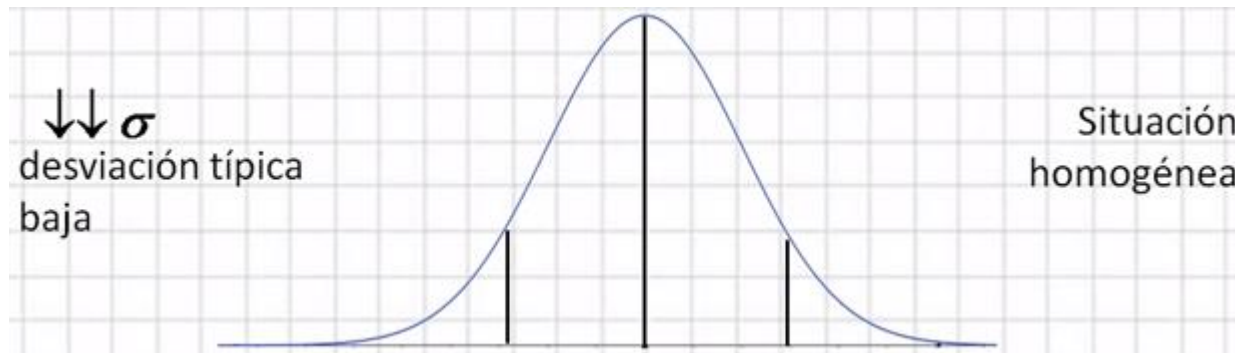
$$N(\mu, \sigma)$$

Desviación Típica

- Desviación típica elevada

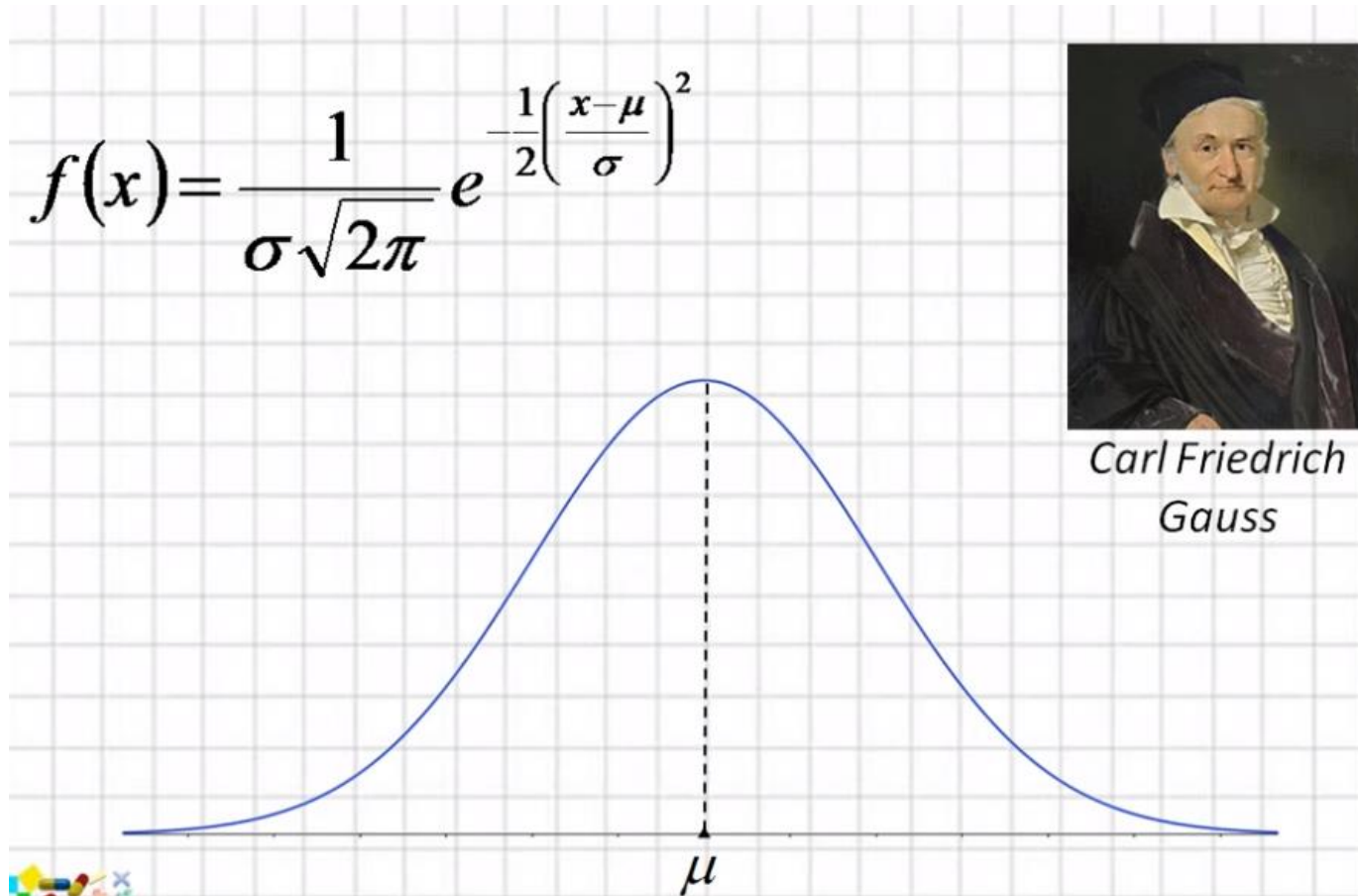


- Desviación típica baja



Función de Distribución Normal

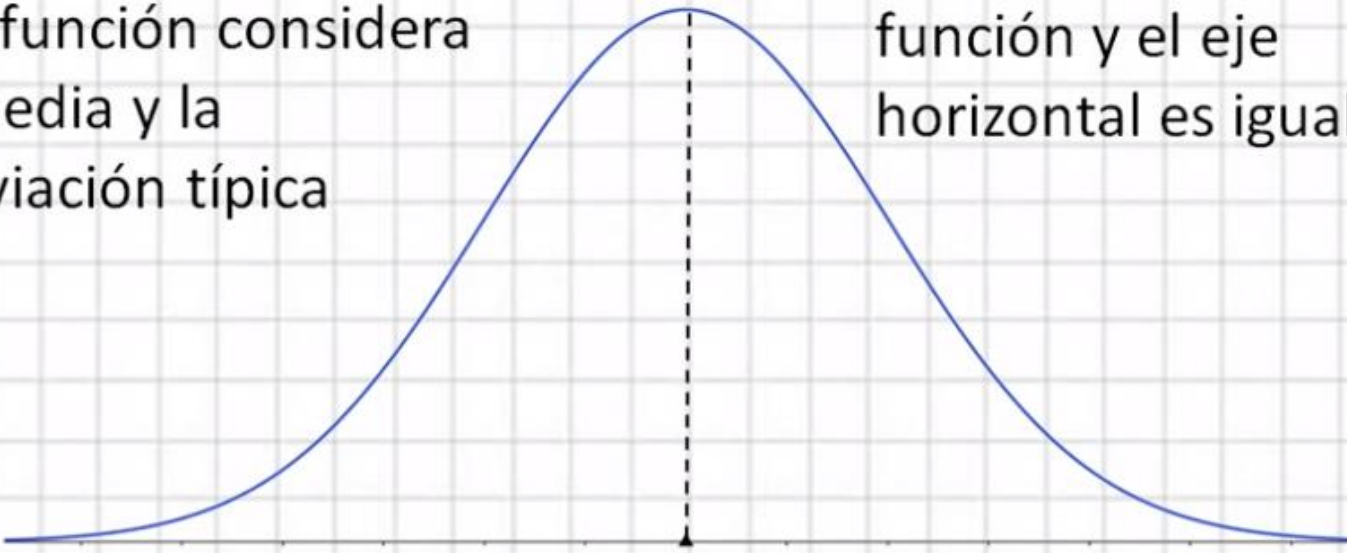
- Función Matemática que se ajusta a la Distribución Normal



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

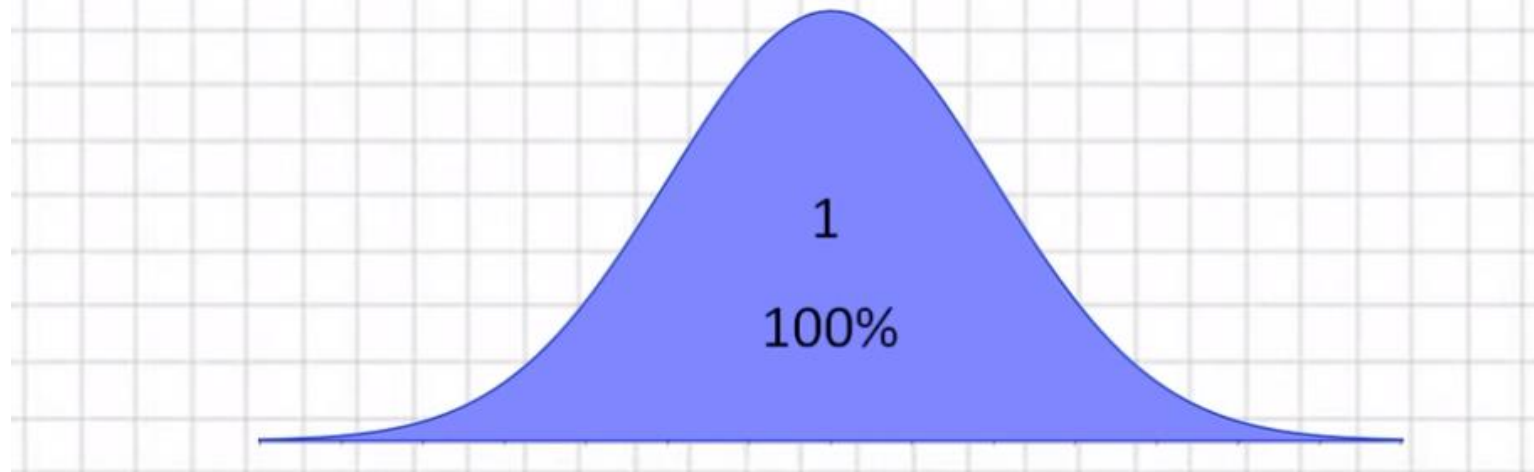
- La función considera la media y la desviación típica

- Es simétrica
- Tiene una asíntota horizontal
- El área entre la función y el eje horizontal es igual a 1



- Toda la población el 100% de la probabilidad la representa el área bajo la curva

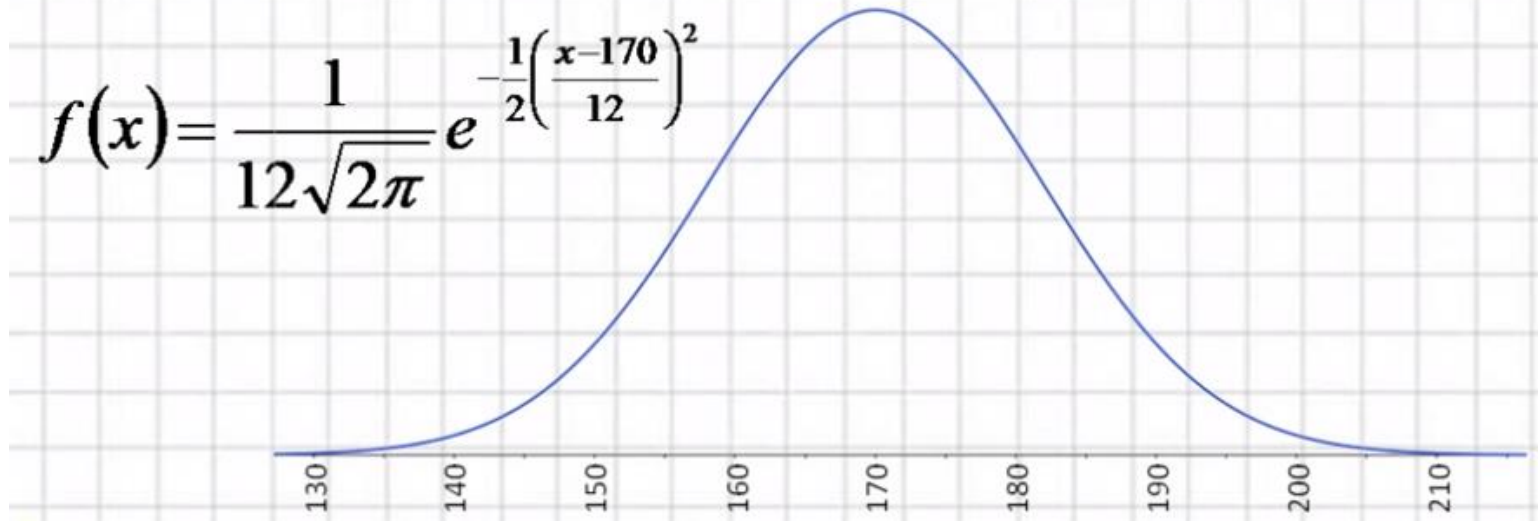
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \int_{-\infty}^{\infty} f(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$$



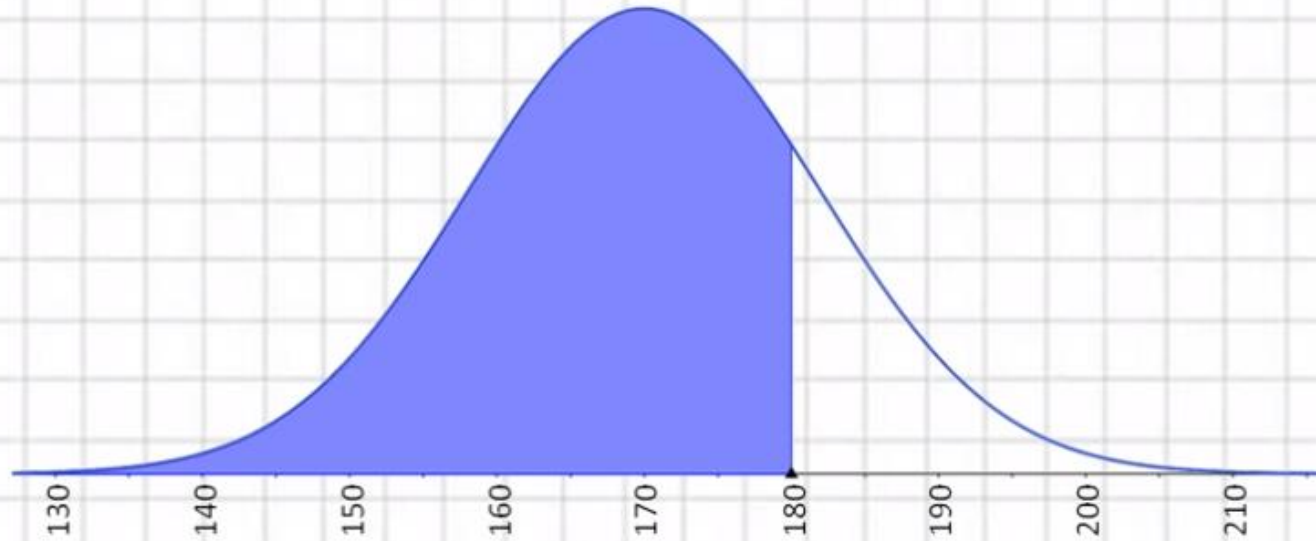
Ejemplo: Función de Distribución Normal

Supongamos que en un determinado país la estatura de la población adulta sigue una distribución normal de media 170 cm y desviación típica igual a 12 cm

$$\mu = 170 \quad \sigma = 12 \quad N(170, 12)$$

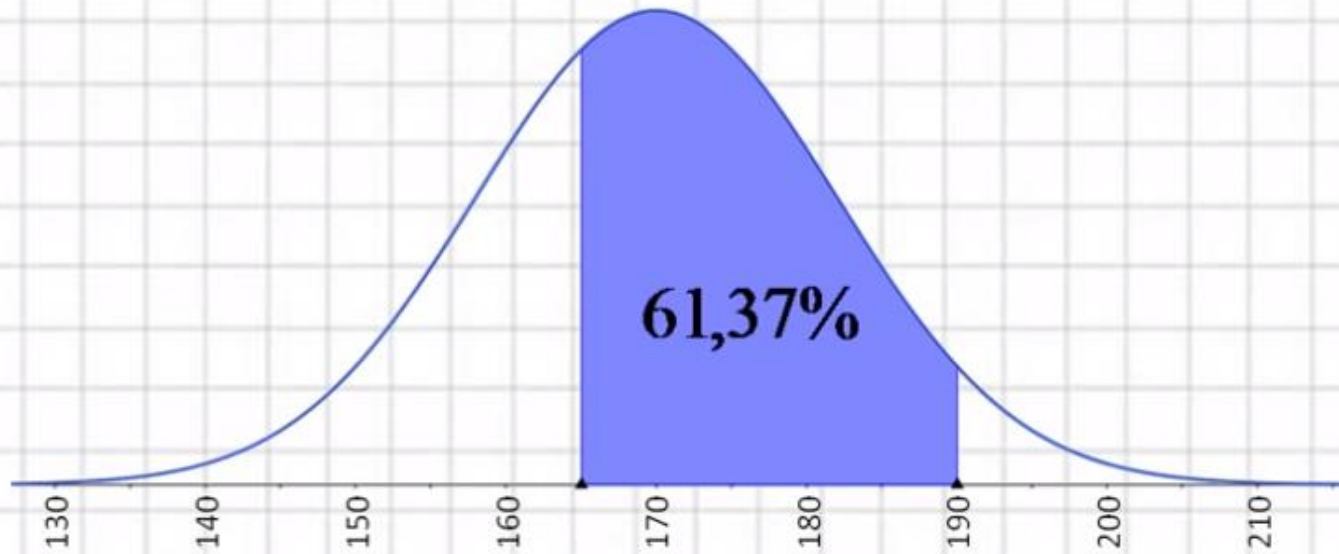


¿Qué porcentaje de esa población mide menos de 180 cm?



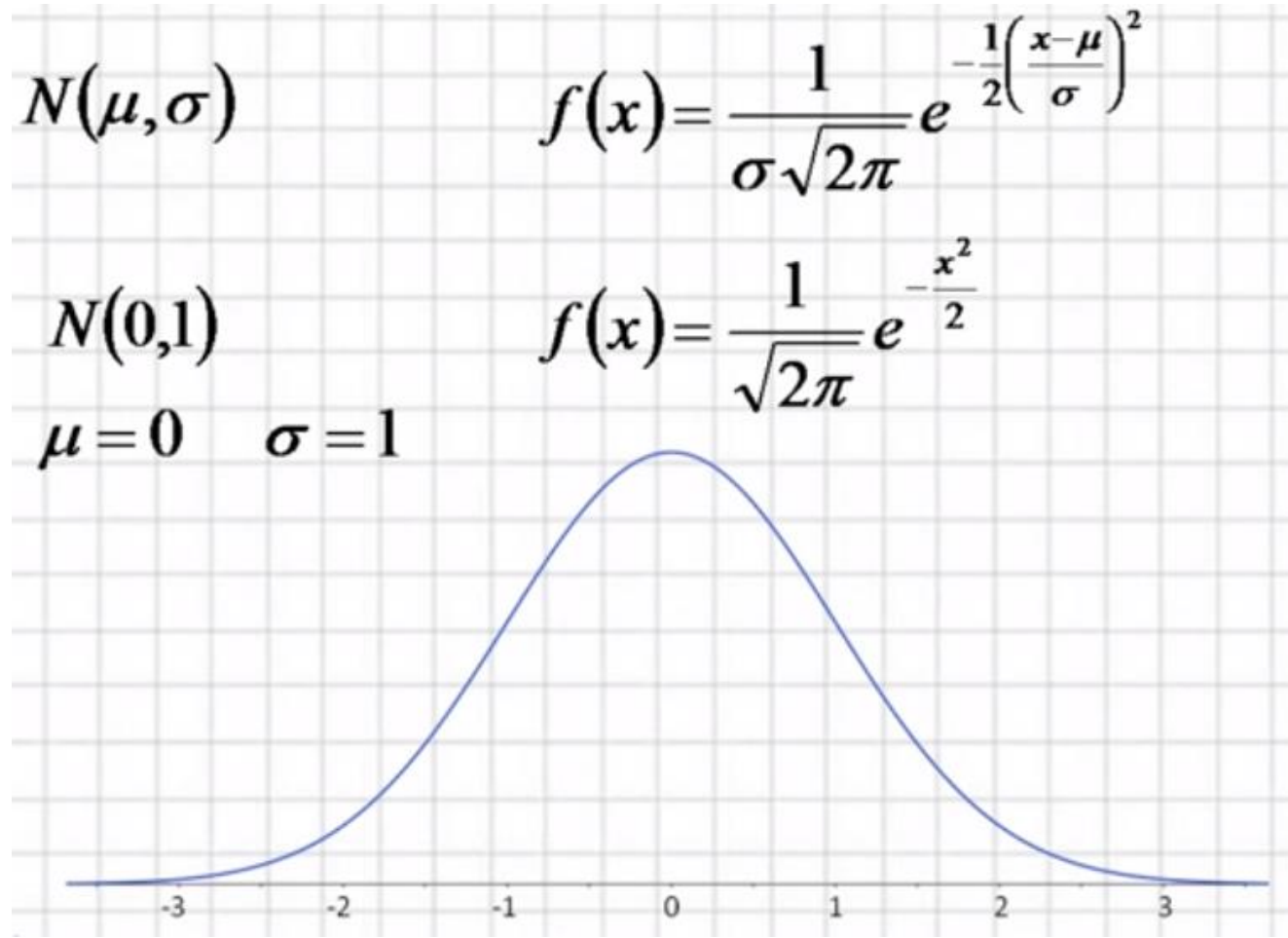
$$\int_{-\infty}^{180} \frac{1}{12\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-170}{12}\right)^2} dx = 0,7977$$

¿Qué porcentaje de esa población mide entre 165 y 190 cm?



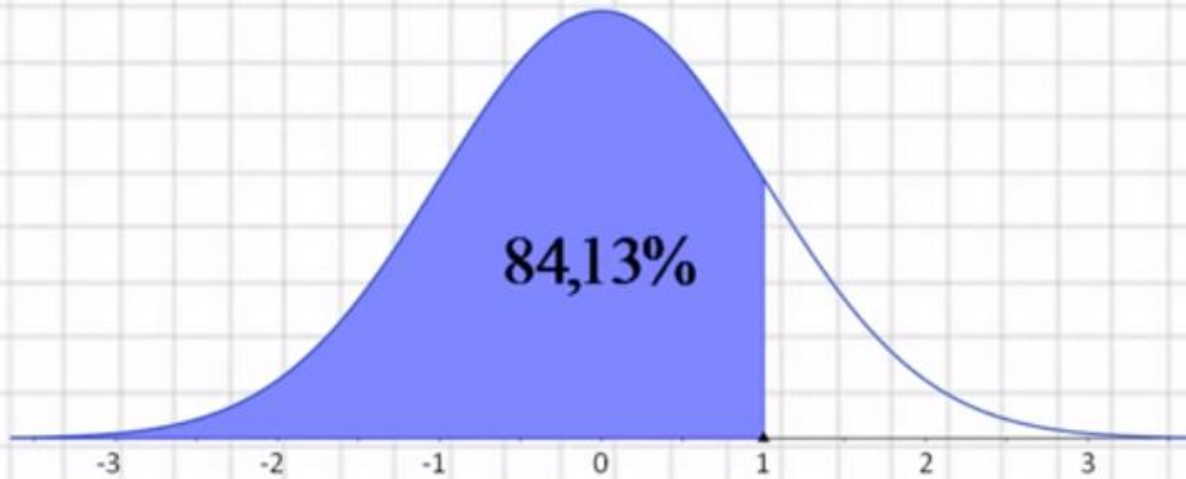
$$\int_{165}^{190} \frac{1}{12\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-170}{12}\right)^2} dx = 0,6137$$

Distribución Normal Estándar



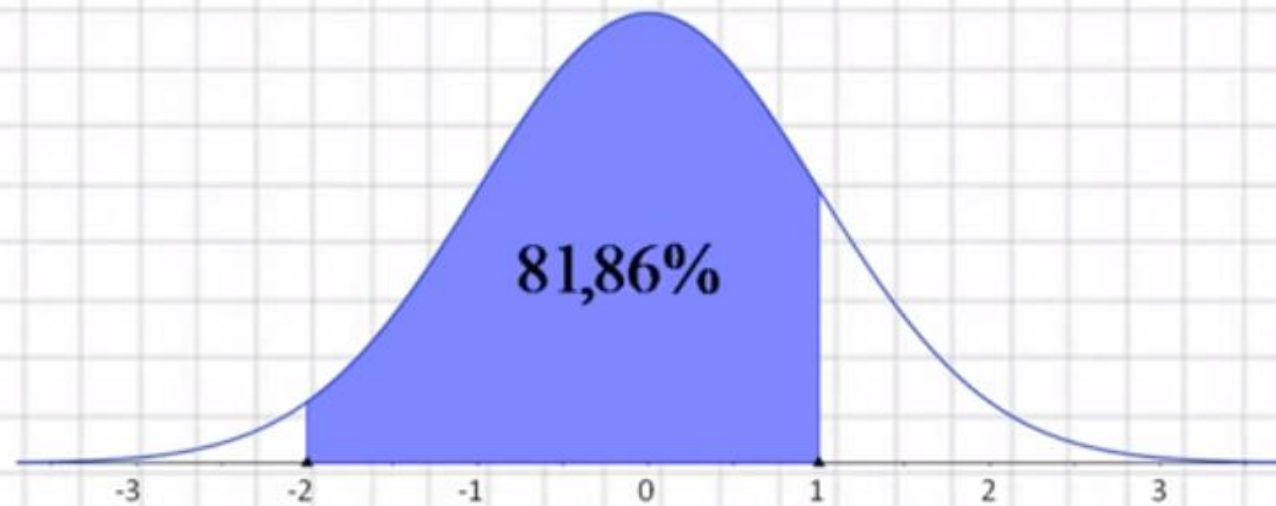
Ejemplo: Distribución Normal Estándar

¿Qué porcentaje queda por debajo del valor 1?



$$\int_{-\infty}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0,8413$$

¿Qué porcentaje queda entre -2 y 1?



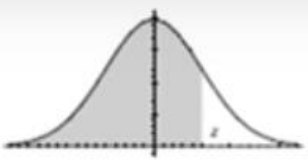
$$\int_{-2}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0,8186$$

Tabla de Áreas Bajo La Distribución Normal Estándar

- Datos de las áreas calculadas bajo la curva para $Z \leq z$

ÁREAS BAJO LA DISTRIBUCIÓN DE PROBABILIDAD NORMAL ESTÁNDAR

Los valores en la tabla representan el área bajo la curva normal hasta un valor positivo de z .

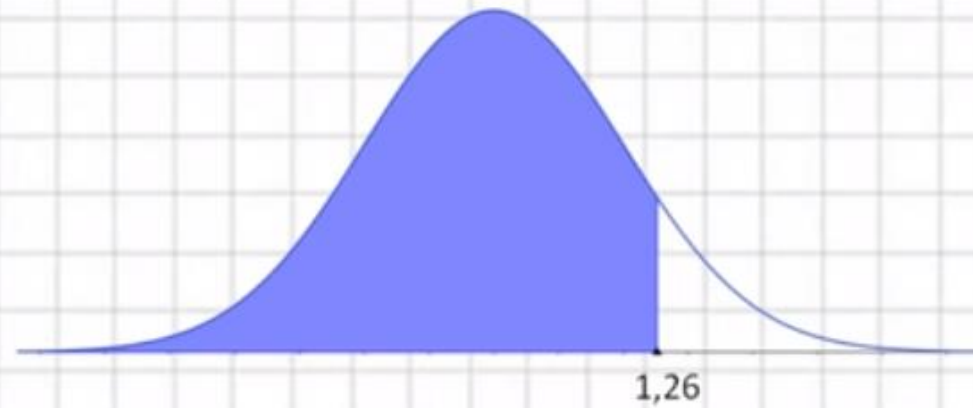


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936

Ejemplo: Usando Tabla de Distribución Normal Estándar

¿Qué área queda por debajo del valor 1,26?

$$P(Z \leq 1,26)$$



<i>z</i>	,00	,01	,02	,03	,04	,05	,06	,07
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5
0,2	0,5793	0,5833	0,5873	0,5910	0,5948	0,5987	0,6026	0,6
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9



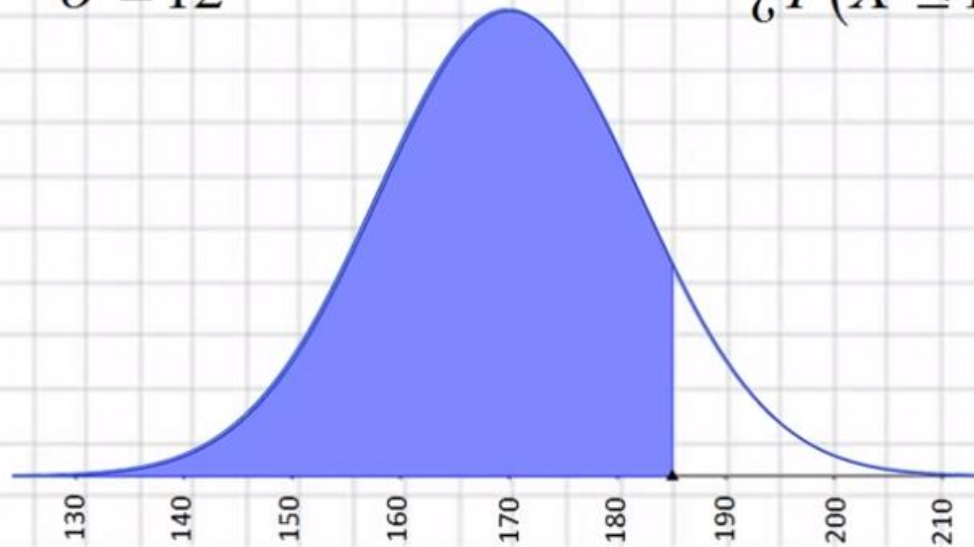
Tipificar Una Variable

- Como se resolvería el siguiente caso?

Supongamos que en un determinado país la estatura de la población adulta sigue una distribución $N(170,12)$
¿Qué porcentaje de esa población mide menos de 185 cm?

$$\mu = 170 \quad \sigma = 12$$

$$¿P(X \leq 185)?$$



...

- Tipificar consiste en transformar la variable de nuestro ejercicio en su equivalente en una distribución $N(0,1)$, para poder usar la tabla o una función en R

$$z = \frac{x - \mu}{\sigma}$$

Ejemplo: Tipificar la variable

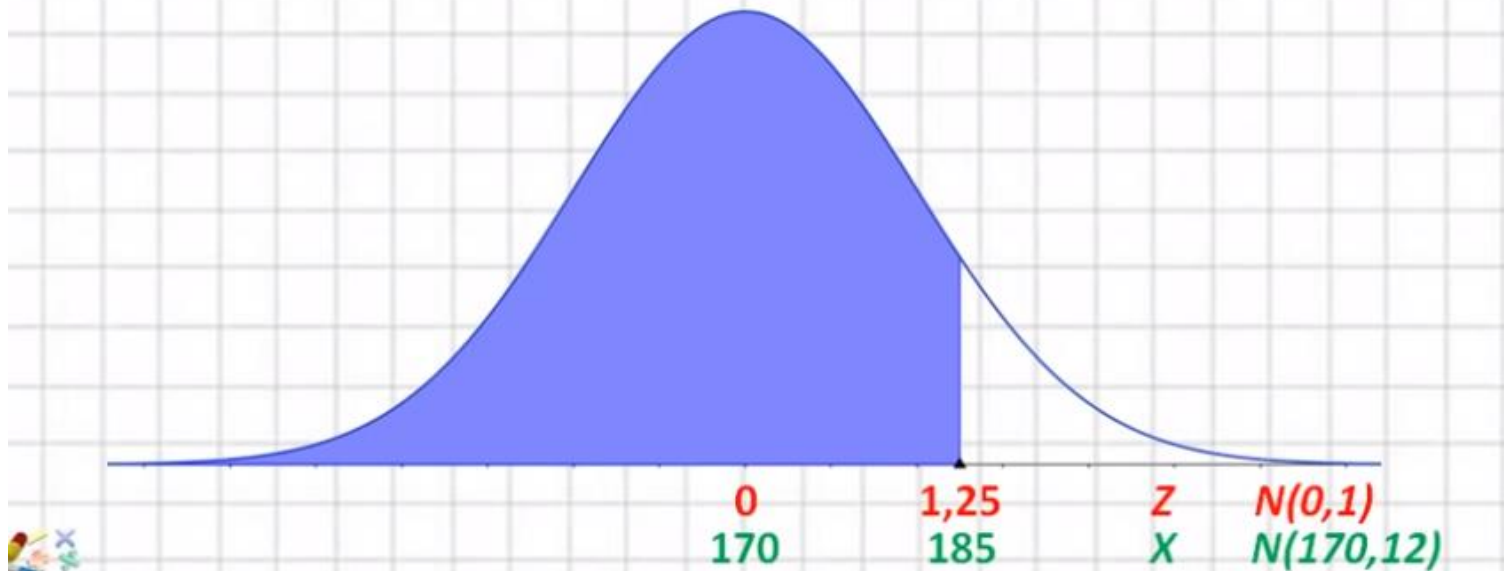
$N(170,12)$

¿ $P(X \leq 185)$?

$$z = \frac{x - \mu}{\sigma} \quad z = \frac{185 - 170}{12} \quad z = 1,25$$

$P(X \leq 185) = P(Z \leq 1,25) = 0,8944$

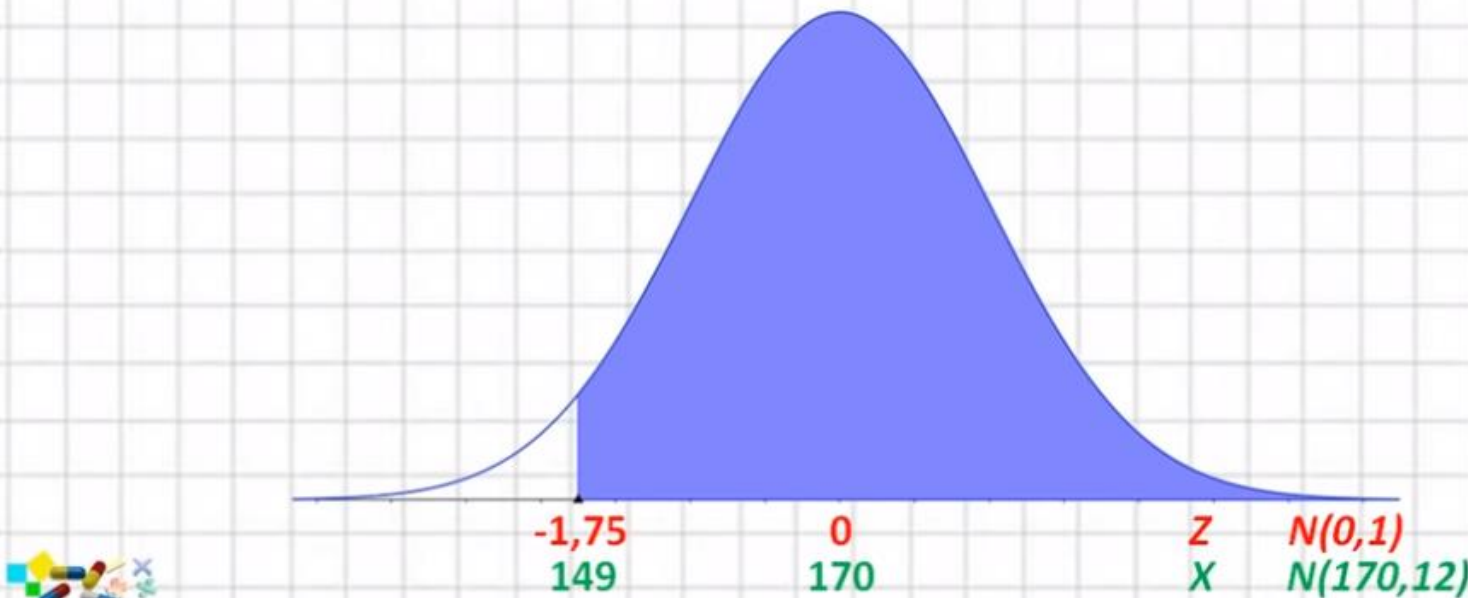
TABLA



¿Y qué porcentaje de esa población mide más de 149 cm?

$$¿P(X > 149)? \quad z = \frac{x - \mu}{\sigma} \quad z = \frac{149 - 170}{12} \quad z = -1,75$$

$$P(X > 149) = P(Z > -1,75) = P(Z \leq 1,75) \overset{\text{TABLA}}{=} 0,9599$$



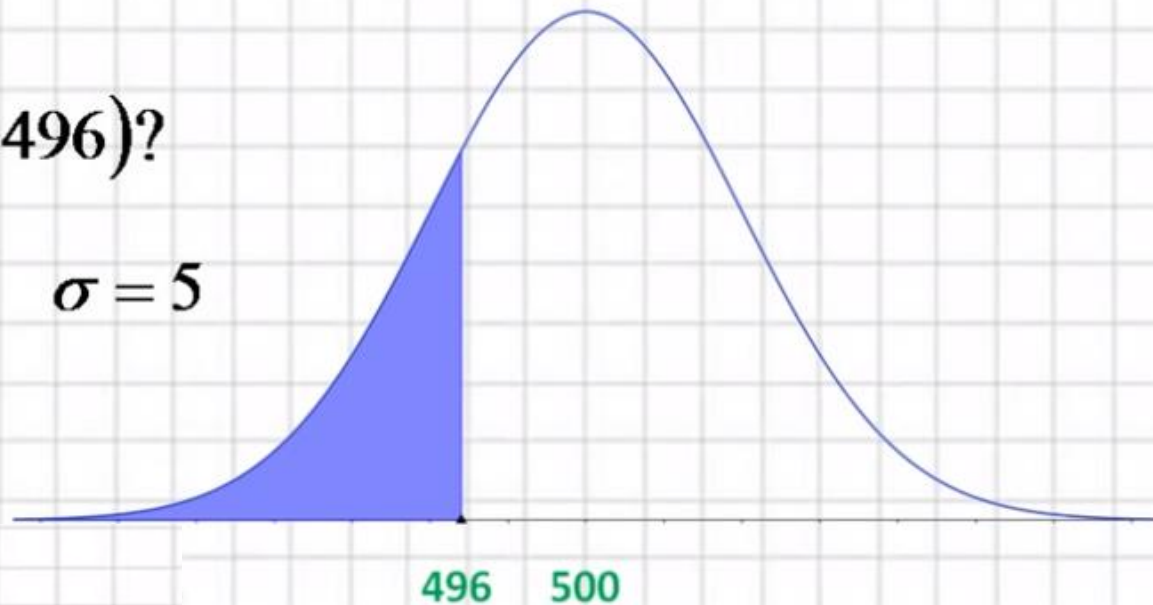
Ejercicio:

El peso en gramos de las cajas de cereales de cierta marca sigue una distribución $N(500,5)$

Calcula la probabilidad de encontrar una caja que pese menos de 496 g

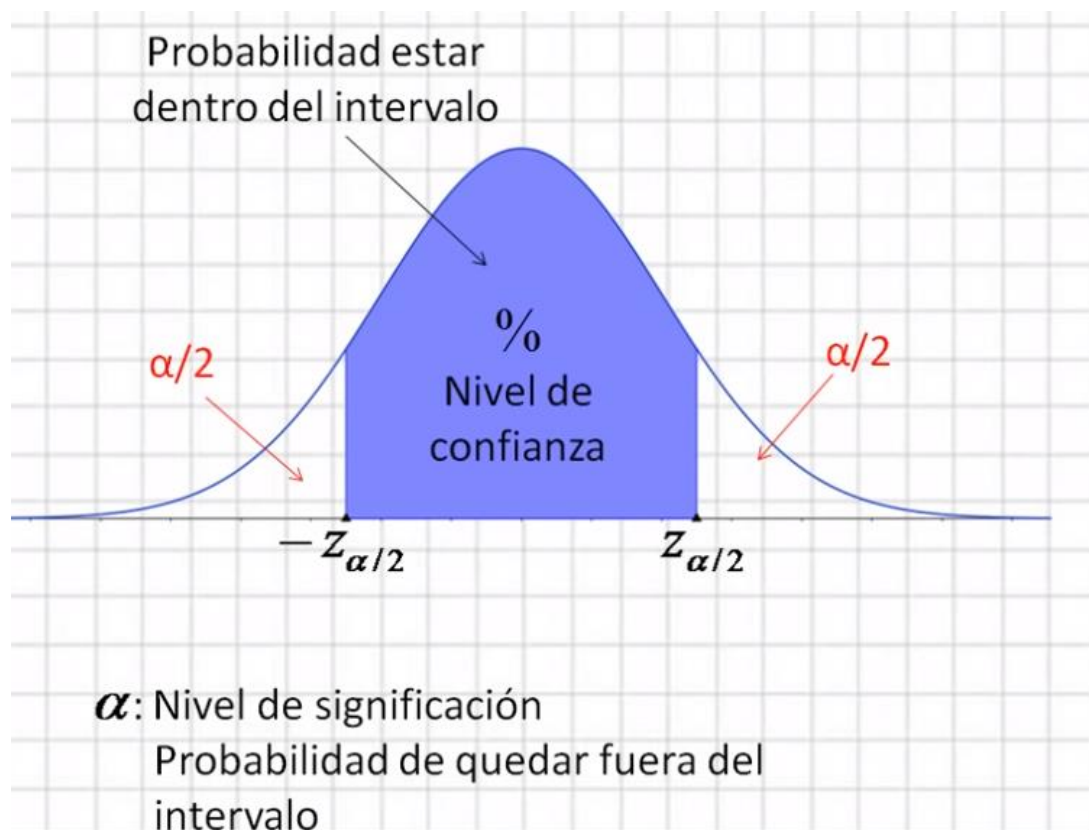
$$¿P(X \leq 496)?$$

$$\mu = 500 \quad \sigma = 5$$



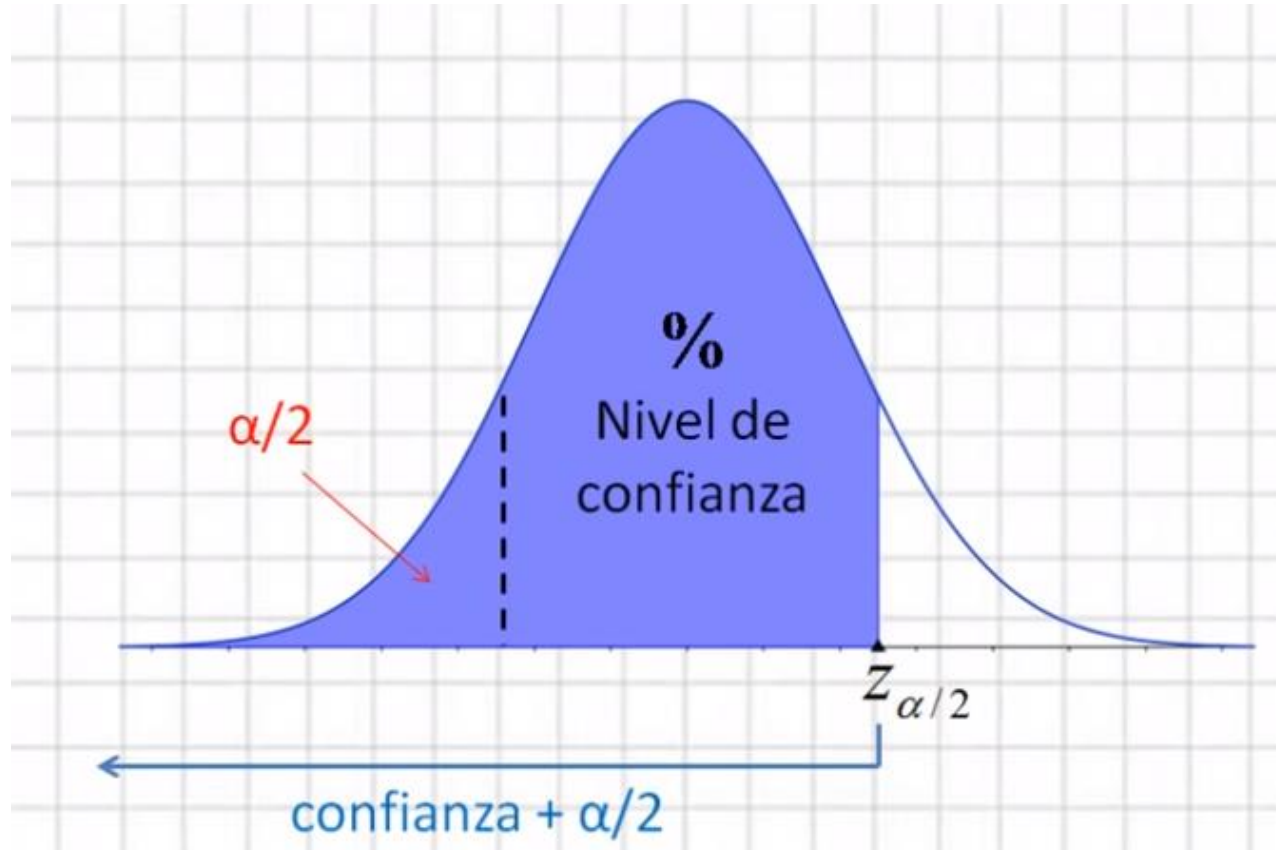
Intervalo de Confianza

- Es el intervalo definido por dos valores simétricos que contienen el porcentaje de observaciones que se requiere conocer



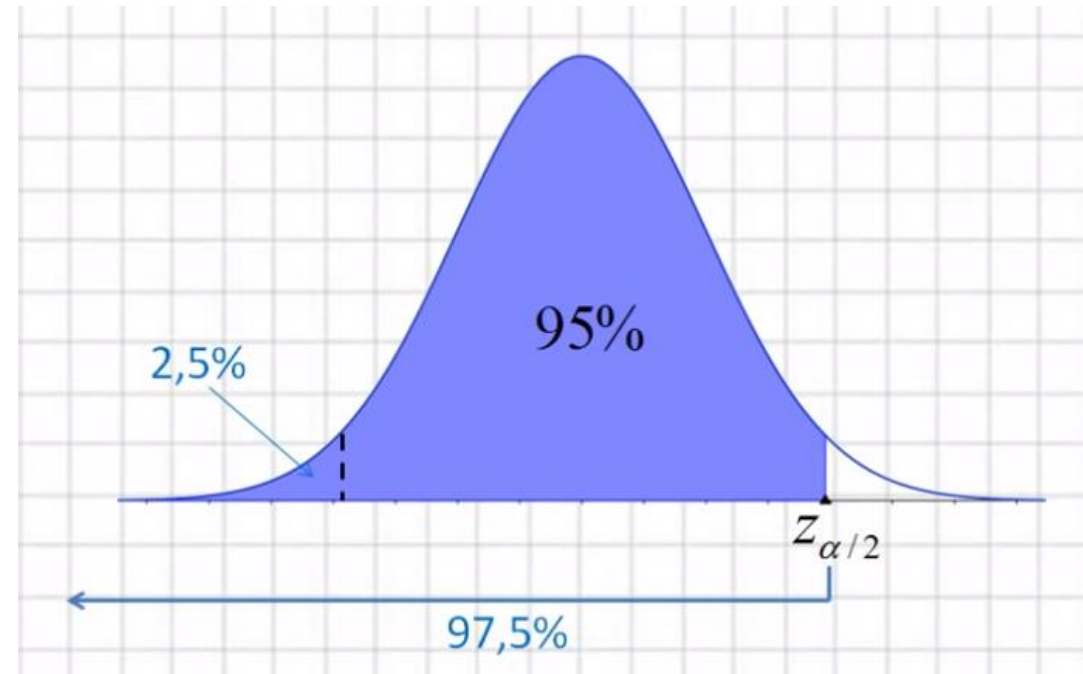
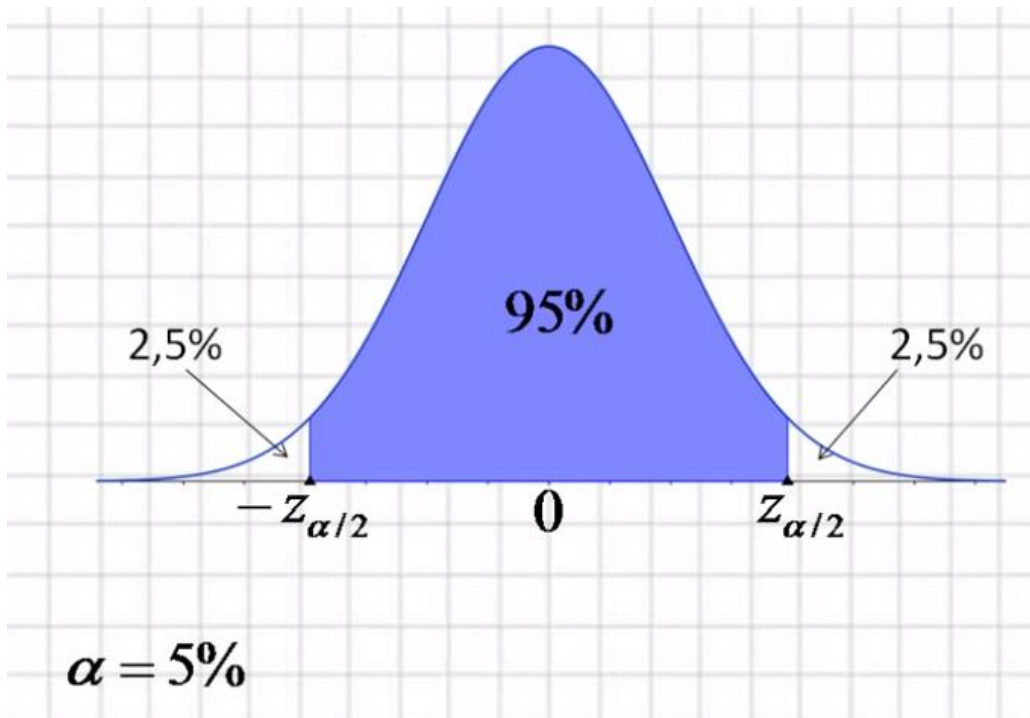
Intervalo de Confianza y Tabla de Distribución Normal Estándar

- En la tabla tendremos que buscar el área bajo la curva para un $Z = \text{confianza} + \alpha/2$



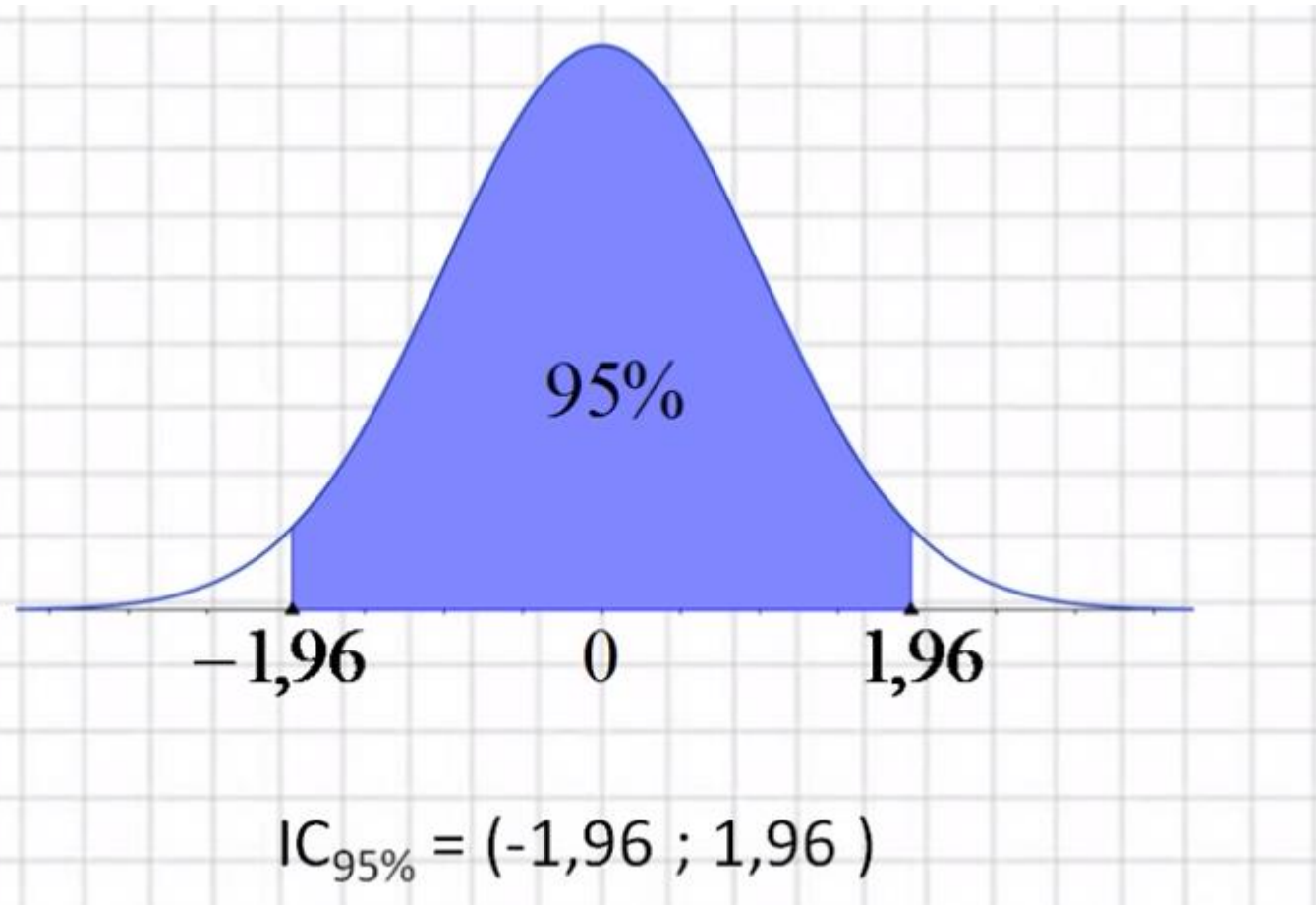
Ejemplo: Intervalo de Confianza

- Cual es el intervalo de confianza del 95%?



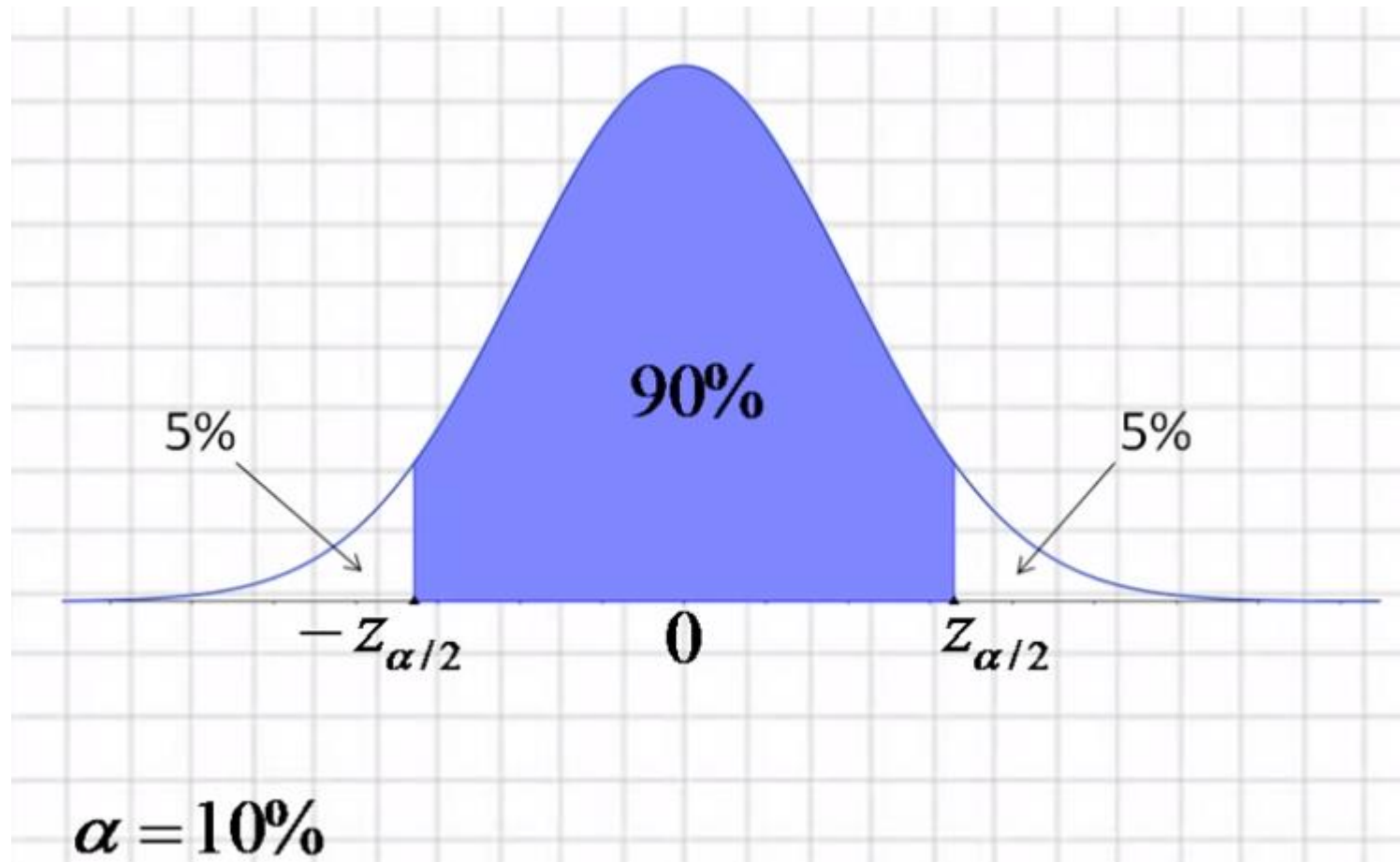
Calcula el intervalo de confianza del 95%

z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817



Ejercicio: Intervalo de Confianza

- Calcula el intervalo de confianza del 90%

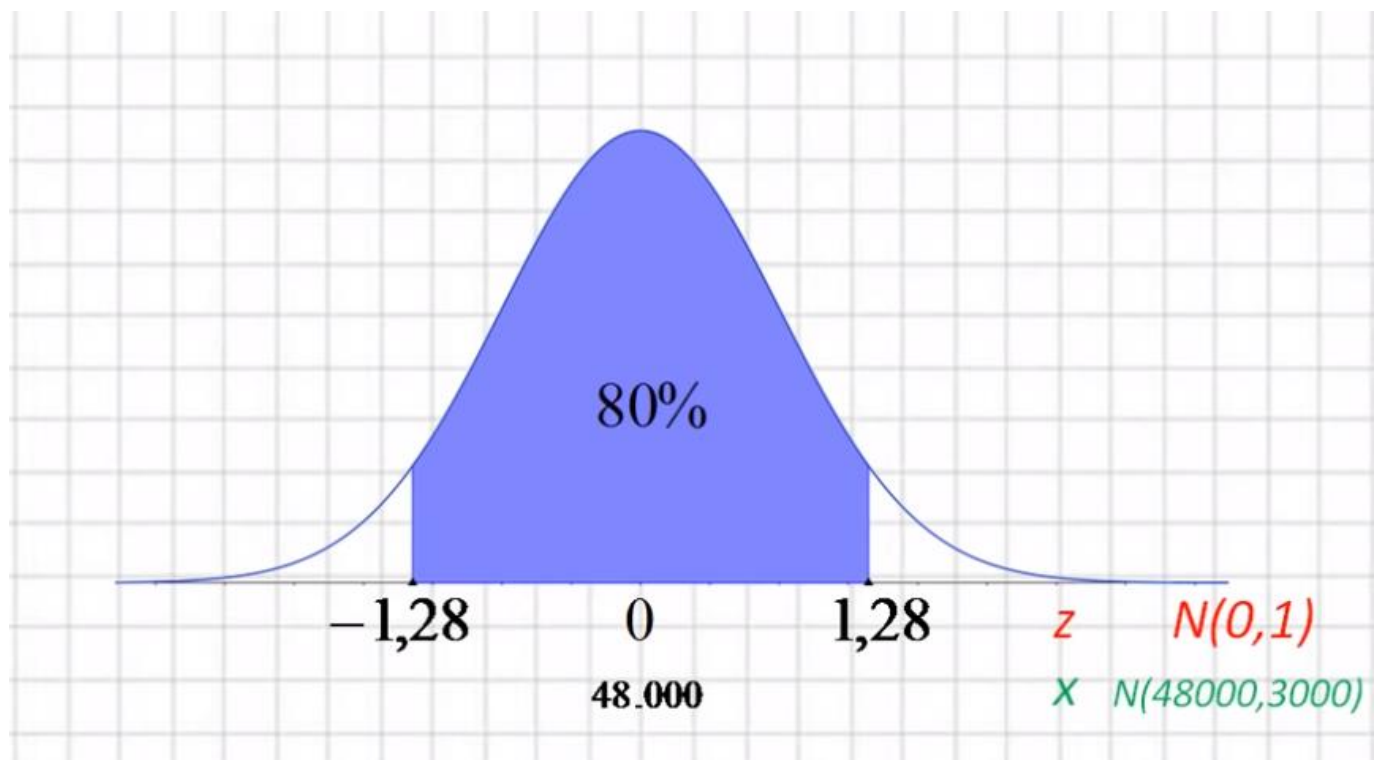


Ejercicio: Intervalo de Confianza

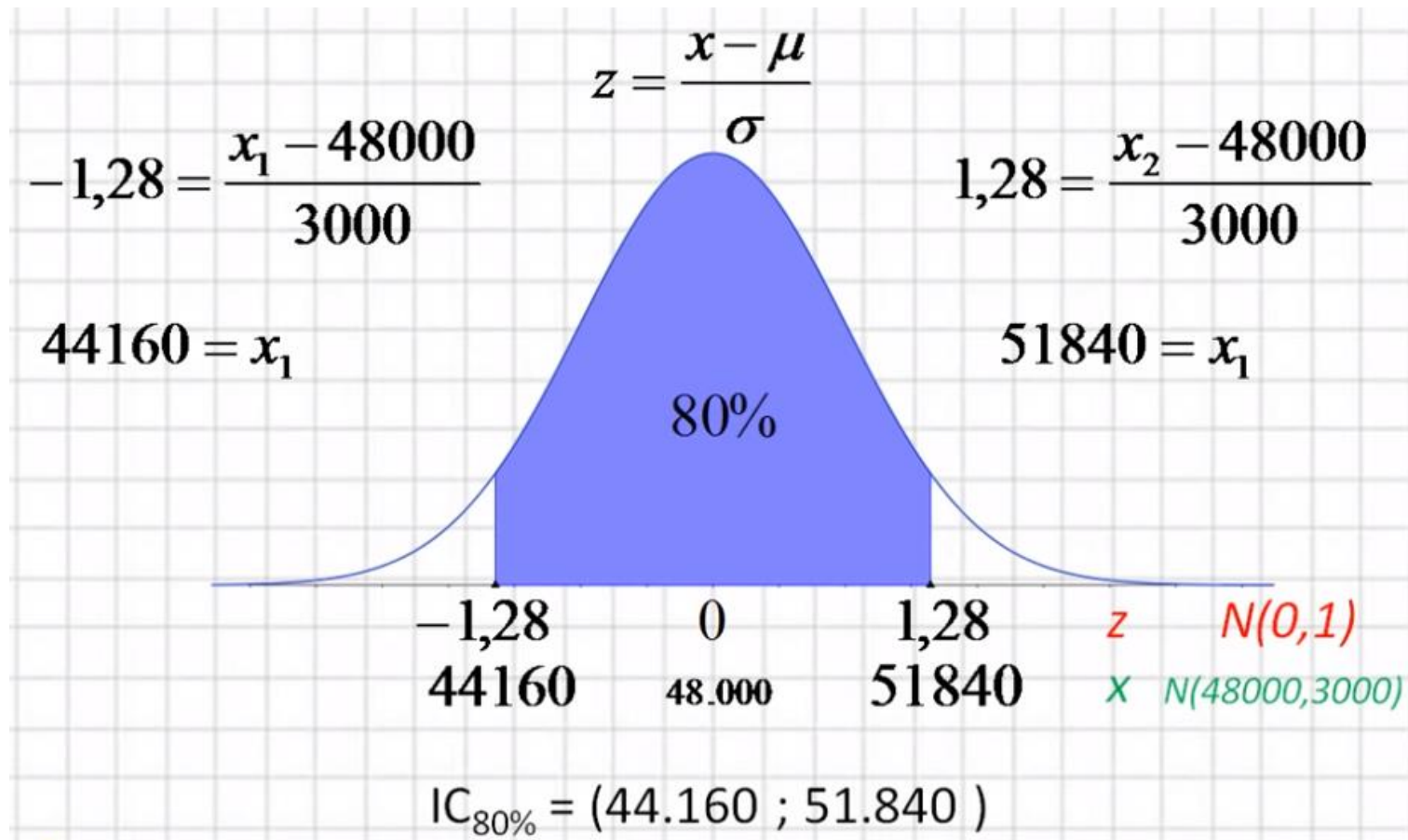
- La duración en km de los neumáticos de una marca se ajusta a una distribución $N(48000, 3000)$, ¿Calcula el intervalo de confianza del 80%?

Solución:

- Usando la tabla encuentro el Z

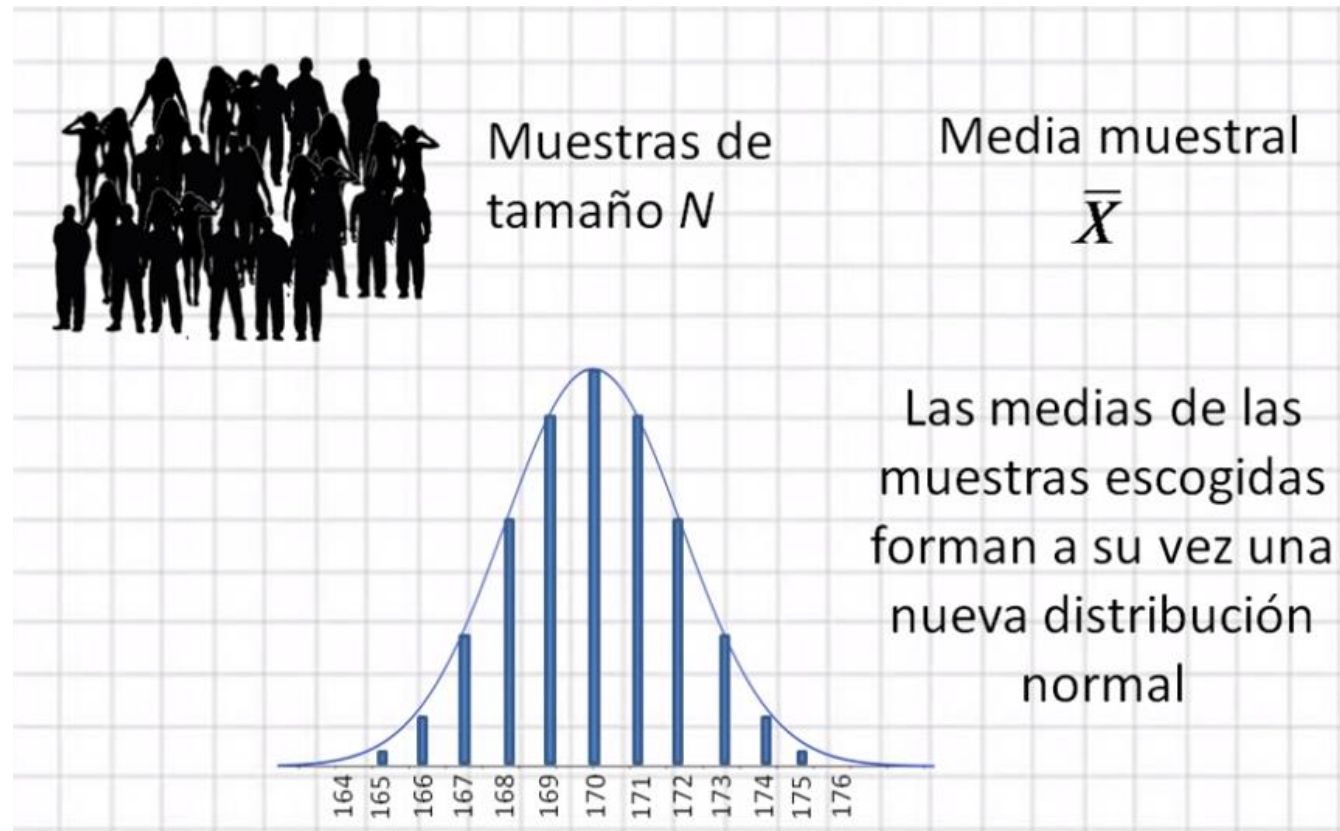


- Usando la formula de tipificar, encuentro el intervalo.

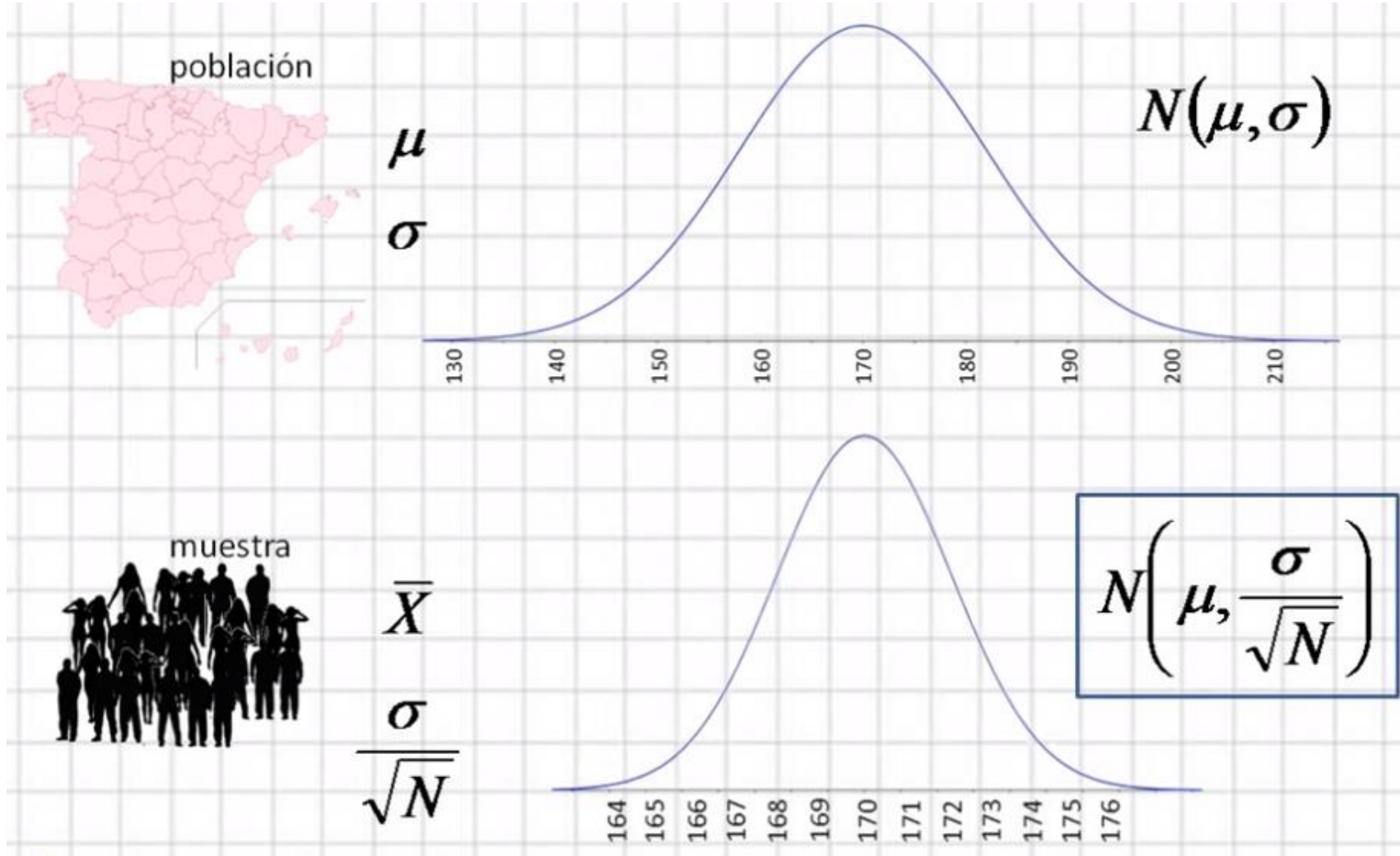


Inferencia Estadística

- Métodos Estadísticos para determinar parámetros estadísticos de una población basados en muestras



Desviación Poblacional y Muestral



Ejemplo: Inferencia Estadística

- Supongamos que sabemos que la duración media de las bombillas de una determinada marca sigue una $N(1500, 160)$
- Si tomamos una bombilla al azar, Cual es la probabilidad de que funcione mas de 1524 horas?
- Si escogemos una muestra de 100 bombillas y calculamos su duración media, Cual es la probabilidad de que sea superior a 1524 horas?

Solución: Para una bombilla

Una bombilla ¿ $P(X > 1524)$?

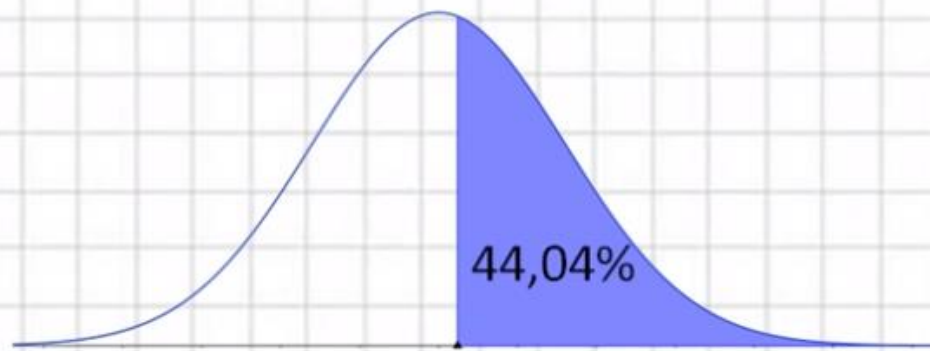


$N(\mu, \sigma)$

$N(1500, 160)$

$$z = \frac{X - \mu}{\sigma} = \frac{1524 - 1500}{160} = 0,15$$

$$P(X > 1524) = P(Z > 0,15) = 1 - P(Z \leq 0,15) \stackrel{\text{TABLA}}{=} 0,4404$$



Solución: Muestra de 100 bombillas

Una muestra de
100 bombillas

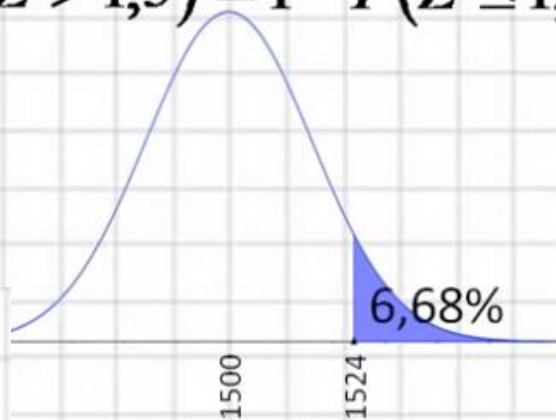


¿ $P(\bar{X} > 1524)$?

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} = \frac{1524 - 1500}{16} = 1,5$$

$$N\left(\mu, \frac{\sigma}{\sqrt{N}}\right) \quad N(1500, 16)$$

$$P(\bar{X} > 1524) = P(Z > 1,5) = 1 - P(Z \leq 1,5) \stackrel{\text{TABLA}}{=} 0,0668$$



Intervalo de Confianza para la Media Poblacional

Supongamos que tomamos una muestra de 100 personas y obtenemos una media muestral de 169 cm

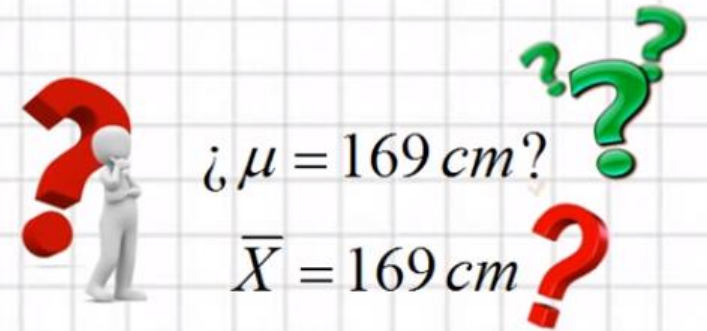
$$N = 100 \quad \bar{X} = 169$$

Podríamos afirmar con bastante certeza que la media de la población, μ , está entre:

120 cm

218 cm

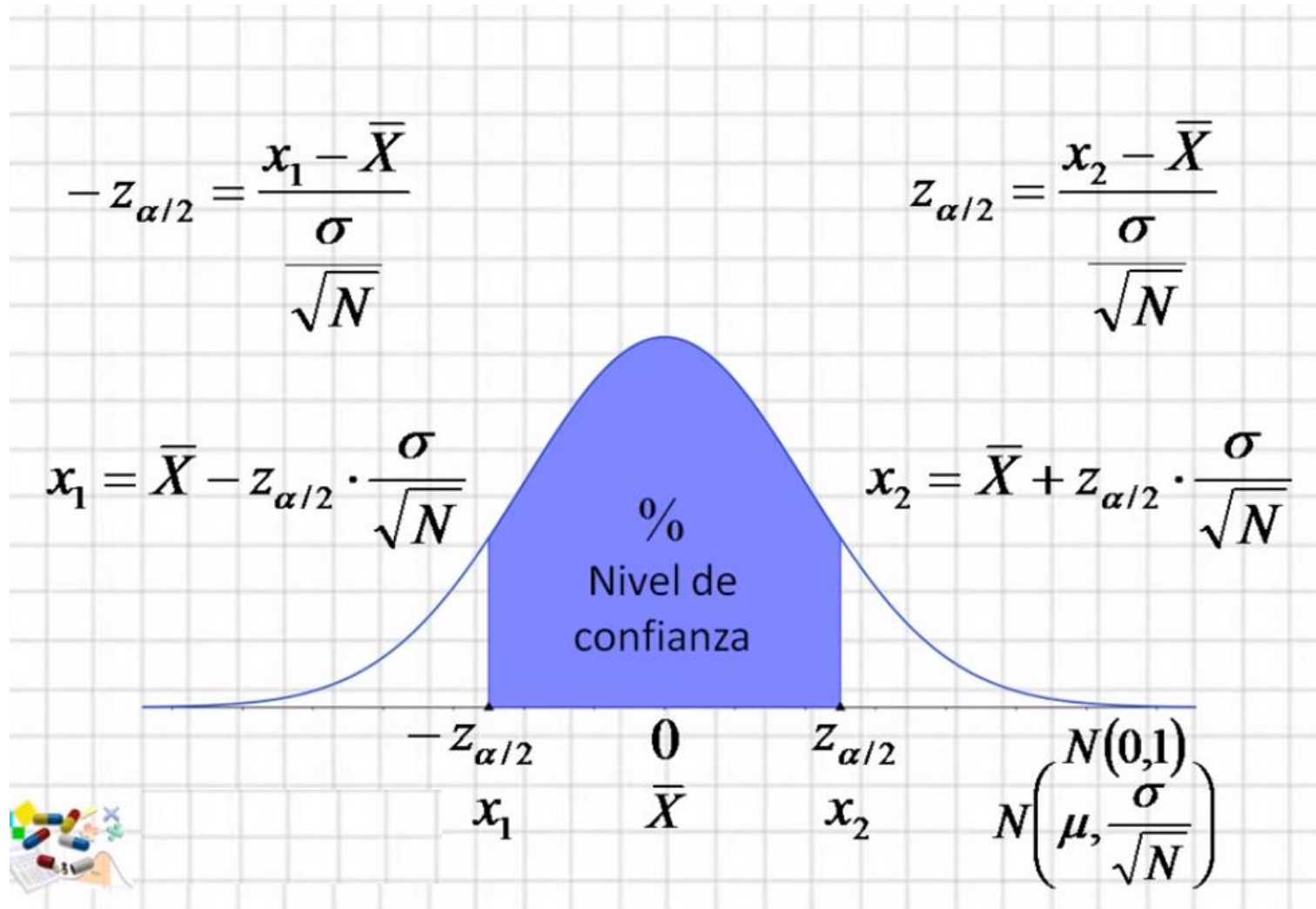
$$\bar{X} = 169 \text{ cm}$$



Intervalo de Confianza u



A partir de la media muestral \bar{X} puedo generar un Intervalo de Confianza dentro del cual tenemos cierta seguridad de que estará la media poblacional μ



$$IC = (x_1, x_2)$$

$$IC = (\bar{X} - error, \bar{X} + error)$$

$$IC = \left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}} \right)$$

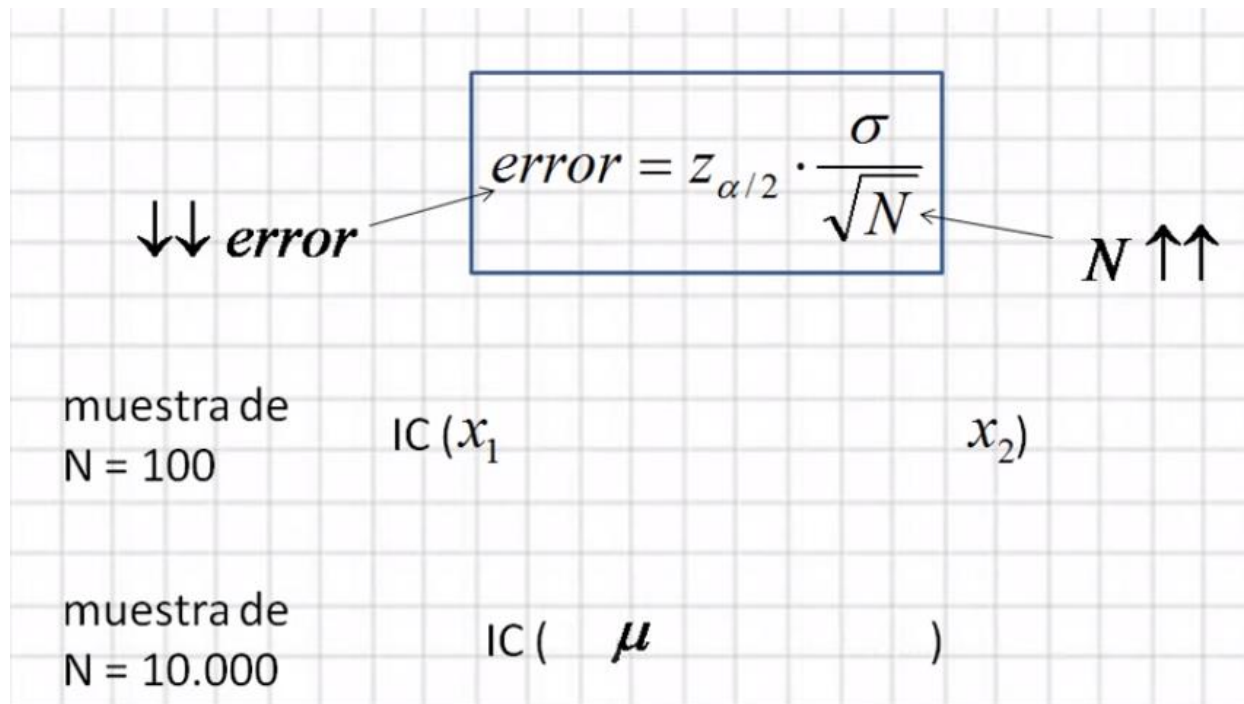
...

- Error

$$error = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

Ejemplo: Intervalo de confianza u

- El tamaño de la muestra reduce el error



A partir de la media muestral \bar{X} genero un Intervalo de Confianza dentro del cual tengo cierta seguridad de que se encuentra la media poblacional μ

$$IC = (\bar{X} - error, \bar{X} + error)$$

$$error = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

Cuanto más grande sea la muestra, más fiable será la media muestral \bar{X} , más pequeño será el error y más específico será el Intervalo de Confianza

Ejercicio: Intervalo de confianza u

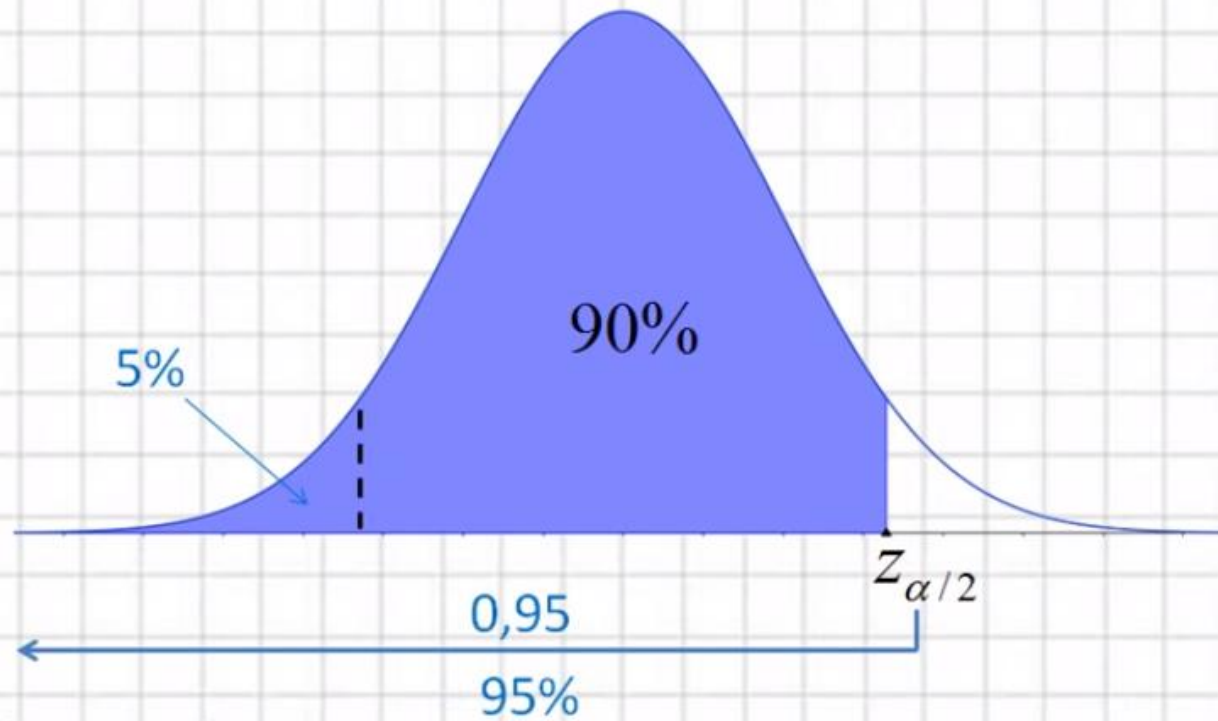
El tiempo diario que los adultos de una determinada ciudad dedican a actividades deportivas, expresado en minutos, se puede aproximar por una variable aleatoria con distribución normal de desviación típica $\sigma = 20$ minutos.

a) Para una muestra aleatoria simple de 250 habitantes de esa ciudad se ha obtenido un tiempo medio de dedicación a actividades deportivas de 90 minutos diarios. Calcúlese un intervalo de confianza al 90% para μ

Solución: Intervalo de confianza u

Población $N(\mu, 20)$

Muestra $N = 250$ $\bar{X} = 90$ ¿ $IC_{90\%}$ para μ ?



$$error = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}} = 1,645 \cdot \frac{20}{\sqrt{250}} = 2,08$$

$$IC = (\bar{X} - error, \bar{X} + error)$$

$$IC_{90\%} = (90 - 2,08; 90 + 2,08)$$

$$IC_{90\%} = (87,92 ; 92,08)$$

b) ¿Qué tamaño mínimo debe de tener una muestra aleatoria simple para que el error máximo cometido en la estimación de μ por la media muestral sea menor que 1 minuto con el mismo nivel de confianza del 90%?

...

- Para tener un error menor a 1, debo entrevistar a mínimo 1083 personas

$$error \leq 1 \quad z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}} \leq 1 \quad 1,645 \cdot \frac{20}{\sqrt{N}} \leq 1$$

$$1,645 \cdot \frac{20}{1} \leq \sqrt{N} \quad 32,9 \leq \sqrt{N} \quad 32,9^2 \leq N$$

$$1082,41 \leq N$$

Preguntas

- Alguna pregunta?



- Resolver con R :
 - Se conoce que el peso de las truchas en una picigranja se puede aproximar por una distribución $N(600,100)$. Considerando una muestra aleatoria simple de 20 truchas,
 - a) calcúlese la probabilidad de que su peso medio sea inferior a 550 gramos.
 - La altura en centímetros de los individuos de una población se puede aproximar por una distribución normal de desviación típica igual a 20 cm. En una muestra aleatoria simple de 500 individuos se ha obtenido una altura media de 174 cm.
 - a) Obténgase un intervalo de confianza al 95% para μ (media poblacional)
 - b) Cual debe ser el tamaño mínimo de la muestra para que el correspondiente intervalo de confianza para μ al 90%, tenga una amplitud de a lo sumo 5 cm?

- La cantidad de fruta, medida en gramos, que contiene los botes de mermelada de una cooperativa se puede aproximar mediante una variable aleatoria con distribución normal de desviación típica de 10 gramos
- a) Se selecciono una muestra aleatoria simple de 100 botes de mermelada y la cantidad total de fruta que contenían fue de 16000 gramos. Determinése un intervalo de confianza 95% para la media μ
- b) A partir de una muestra aleatoria simple de 64 botes de mermelada se ha obtenido un intervalo de confianza para la media μ con error de estimación de 2.35 gramos. Determinése el nivel de confianza utilizado para construir el intervalo