

# THE BRIDGE

## Estadística Inferencial

# Introducción

## ESTADÍSTICA INFERENCIAL

1. Probabilidad
2. Variables aleatorias
3. Distribuciones de probabilidad
4. Distribución normal
5. Intervalos de confianza
6. Error absoluto y tamaño de la muestra
7. Contraste de hipótesis

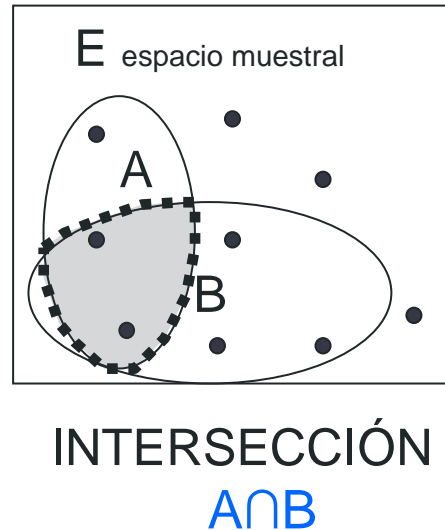
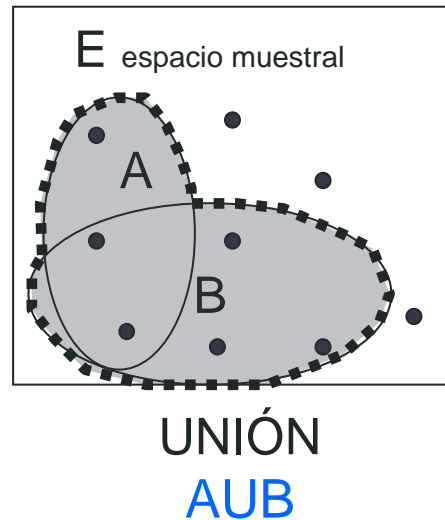
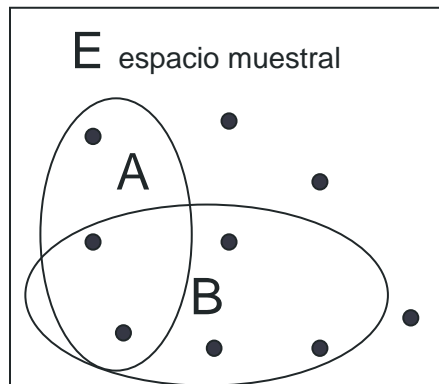
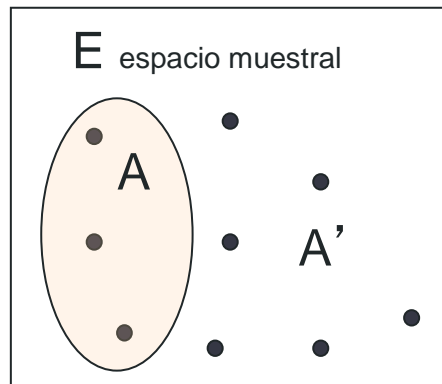
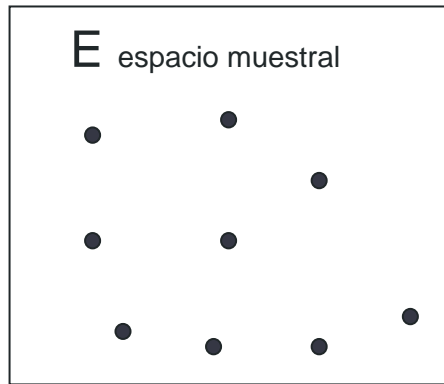
## 3. Estadística Inferencial

### 3.1. Probabilidad

---

# Conceptos básicos

- **Experimento aleatorio:** su resultado no puede predecirse con certeza
- **Espacio muestral:** El conjunto de todos los resultados posibles
- **Suceso:** Subconjunto del espacio muestral



# Conceptos básicos

Ejemplo: Se lanza una moneda al aire dos veces

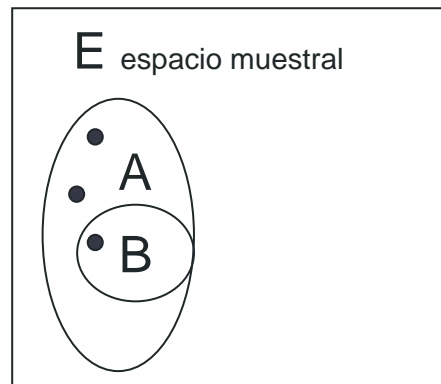
- Espacio muestral:  $\mathbf{E} = \{\mathbf{CC}, \mathbf{CX}, \mathbf{XC}, \mathbf{XX}\}$ , donde C (Cara) y X (Cruz)
- Suceso A: Al menos sale una cara  $\mathbf{A} = \{\mathbf{CC}, \mathbf{CX}, \mathbf{XC}\}$
- Suceso B: Las dos veces sale cara  $\mathbf{B} = \{\mathbf{CC}\}$

¿Cuál es el suceso complementario de A?

- $\mathbf{A}' = \{\mathbf{XX}\}$

¿Cuál es el suceso  $\mathbf{A} \cap \mathbf{B}$ ?

- $\mathbf{A} \cap \mathbf{B} = \{\mathbf{CC}\}$



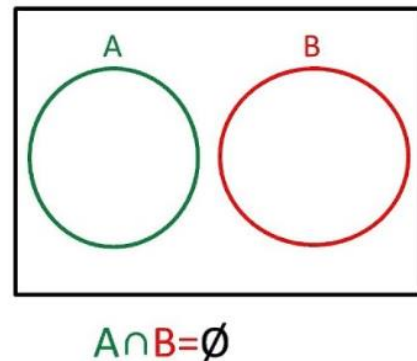
# Definición de probabilidad

La probabilidad de un suceso es **un número** que cuantifica en términos relativos las opciones de verificación de un suceso

$$P(A) = \frac{\text{Nº de casos favorables}}{\text{Nº de casos posibles}}$$

Propiedades:

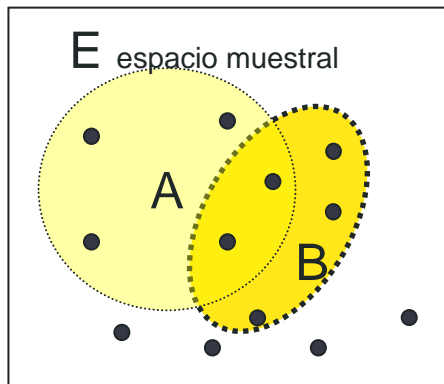
- $P(E) = 1$
- $0 \leq P(A) \leq 1$
- $P(A') = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A) \cdot P(B)$  si A y B son independientes



# Probabilidad condicional

- Probabilidad del suceso **A**, dada la verificación del suceso **B**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Ejemplo: El 50% de la población fuma y el 10% que fuma es hipertenso. ¿Cuál es la probabilidad de que un fumador sea hipertenso?

$A = \{\text{ser hipertenso}\}$

$B = \{\text{ser fumador}\}$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.10}{0.50} = 0.2$$

## 3. Estadística Inferencial

### 3.2. Variables Aleatorias

---



# Variables aleatorias (VA)

- Supongamos el experimento de elegir al azar un universitario de España. El espacio muestral está formado por todos los distintos universitarios:  $E = \{\omega_i\}_{i=1}^N$
- Nuestro interés no está en el individuo  $\omega_i$  sino en un valor asociado a él que denotamos por  $X(\omega_i) = x$ , por ejemplo, su edad
- A la función que asocia a un resultado un valor numérico se le llama **variable aleatoria**

$$\begin{aligned} X: E &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = x \end{aligned}$$

# Variables aleatorias continuas

- Pueden tomar un número infinito de valores, por ejemplo, la concentración de azúcar de un refresco
- De estos valores aleatorios, nos interesa conocer la probabilidad que tenemos de que el valor que observemos esté entre dos números:

$$P(a < X \leq b) = \int_a^b f(x) d(x)$$

- La función  $f(x)$  recibe el nombre de **función densidad de probabilidad** de la variable aleatoria  $X$ , y nos informa sobre la cantidad de probabilidad que hay en un intervalo determinado. Cumple las siguientes propiedades:

$$1) f(x) \geq 0$$

$$2) \int f(x) = 1$$

## 3. Estadística Inferencial

### 3.3. Distribuciones de probabilidad

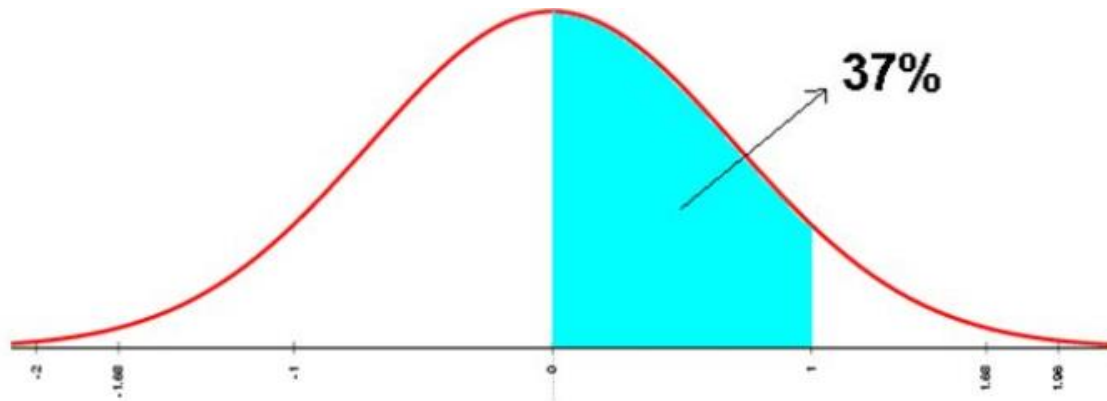
---

# Distribución de probabilidad

- La distribución de probabilidad describe las probabilidades de los posibles valores de una variable aleatoria
- Si la VA es discreta, le corresponderá una distribución discreta
- Si la VA es continua (puede tomar cualquier valor dentro de un intervalo), la distribución será continua.

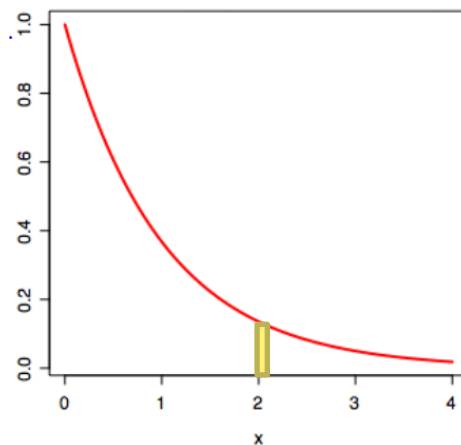
# Función densidad de probabilidad (fdp)

- La **función de densidad de probabilidad**, función de densidad, o, simplemente, densidad de una **variable aleatoria continua** describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor



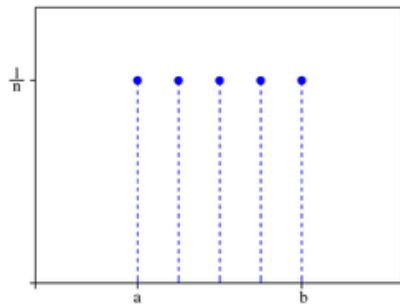
# Función densidad de probabilidad (fdp)

- Ejemplo: Una especie de bacteria típicamente vive entre 0 y 4 horas. ¿Cuál es la probabilidad de que una bacteria viva *exactamente* 2 horas?
- La respuesta es 0%. Muchas bacterias vivirán *aproximadamente* 2 horas, pero es improbable que dada una bacteria ésta viva *exactamente* 2.000000 horas
- En lugar de eso, la pregunta debería ser: ¿Cuál es la probabilidad de que la bacteria muera entre 2 y 2.01 horas?



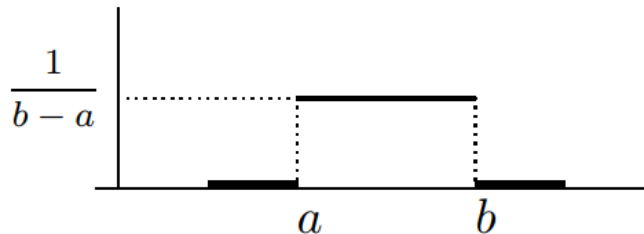
# Distribuciones discretas más comunes

- **Distribución binomial:** describe el número de aciertos en experimentos con posibles resultados binarios con probabilidad de acierto  $p$  y probabilidad de fallo  $q = 1 - p$ .
  - Para representar que una VA sigue una distribución binomial:  $X \sim B(n, p)$
  - Ejemplos: n° de caras al lanzar 20 veces una moneda  $X \sim B(20, 0.5)$
- **Distribución uniforme:** Asume un número finito de valores con la misma probabilidad
  - La probabilidad de cada resultado  $x_i$  es  $p(x_i) = \frac{1}{n}$
  - Ejemplo: En un dado, todos los resultados tienen la probabilidad  $1/6$

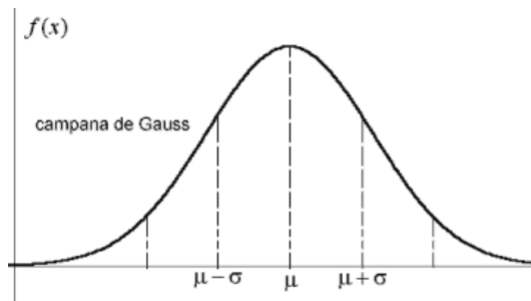


# Distribuciones continuas más comunes

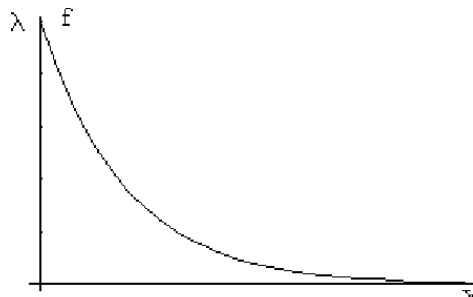
- **Distribución uniforme**



- **Distribución normal**



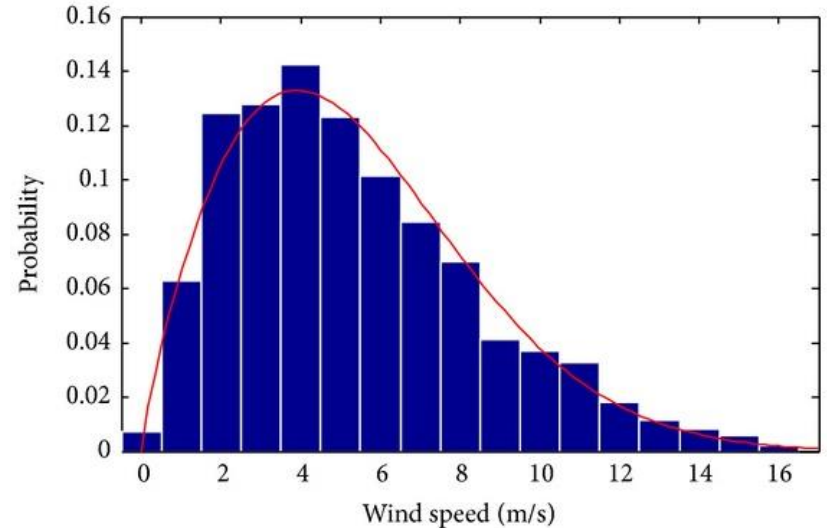
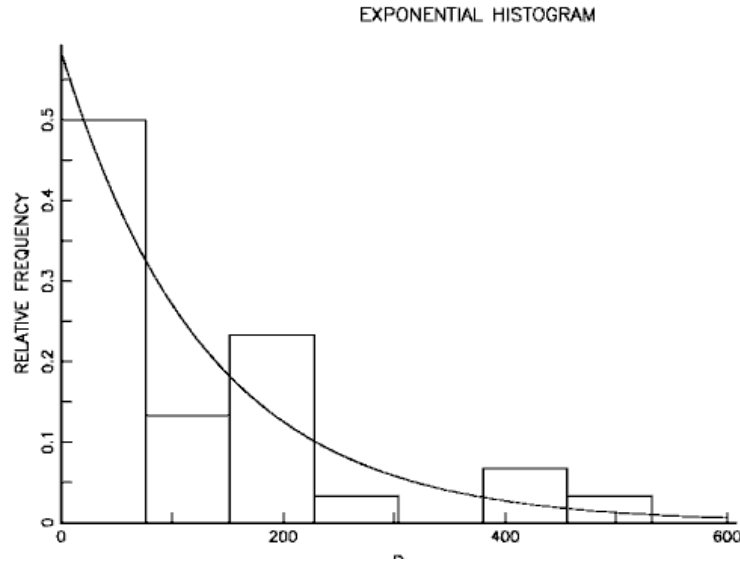
- **Distribución exponencial**





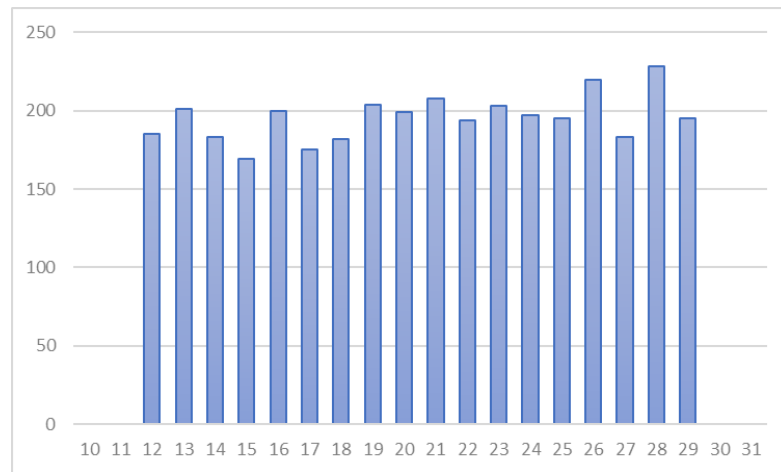
# Histogramas y fdps

- Los histogramas dan una idea aproximada de la distribución de los datos, y a menudo se utilizan para estimar la función densidad de probabilidad



# Distribuciones para la asignación de valores aleatorios

- A menudo, se realizan experimentos escogiendo números aleatorios, donde se especifica la distribución que deben seguir estos números
- Ejemplo: generar 100 números que sigan una distribución uniforme entre 12 y 29



## 3. Estadística Inferencial

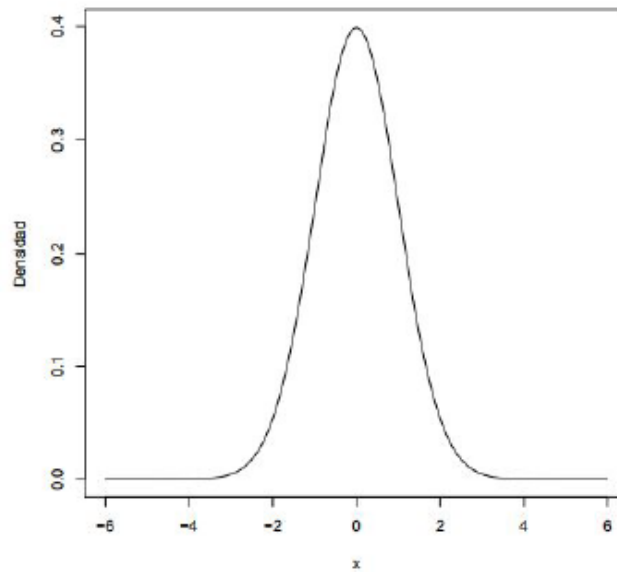
### 3.4. Distribución normal

---

# Distribución normal

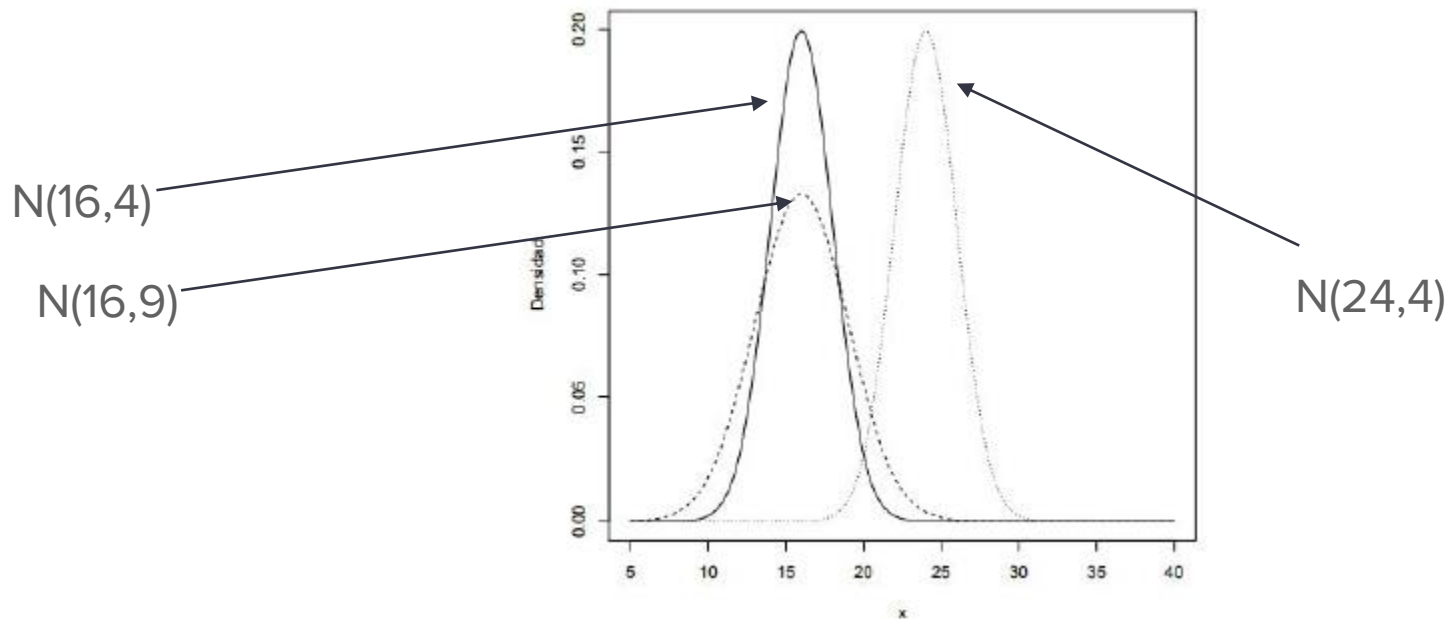
- Una variable aleatoria  $X$  se dice que sigue una distribución normal con media  $\mu$  y varianza  $\sigma^2$  (o, simplemente, que es una variable aleatoria normal) y se denota con  $X \sim N(\mu, \sigma)$  si su función de densidad viene dada por

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



# Distribución normal

- Ejemplo: ¿A cuál de las siguientes curvas corresponden las distribuciones normales  $N(16,4)$ ,  $N(24,4)$  y  $N(16,9)$ ?



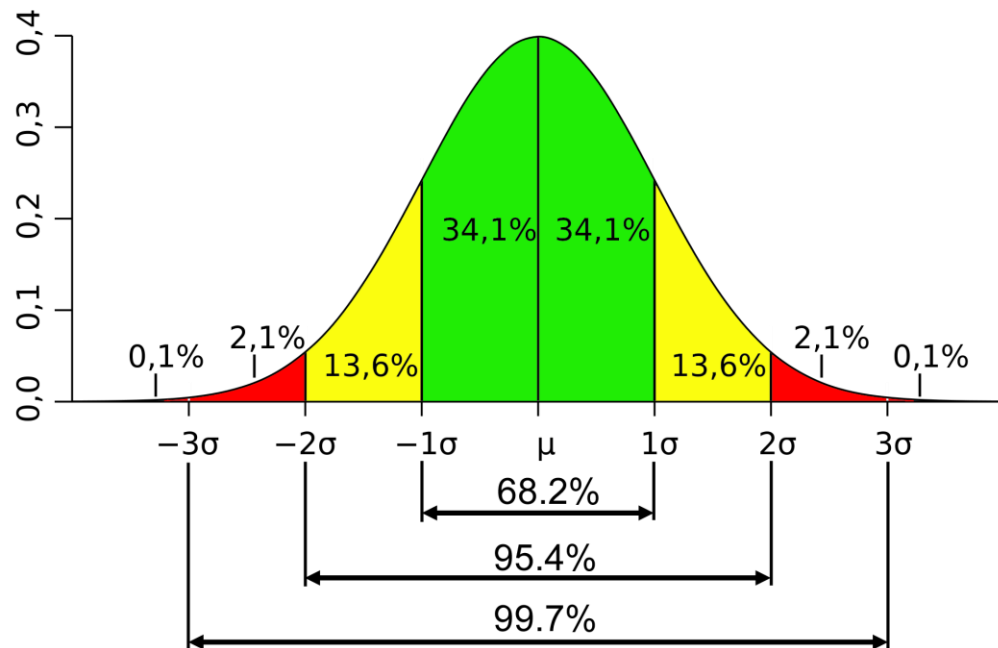
# Distribución normal

Muchos fenómenos físicos se pueden modelar de manera adecuada a través de esta distribución.



# Distribución normal

Se pueden conocer las proporciones de datos/probabilidades en función de la desviación estándar:



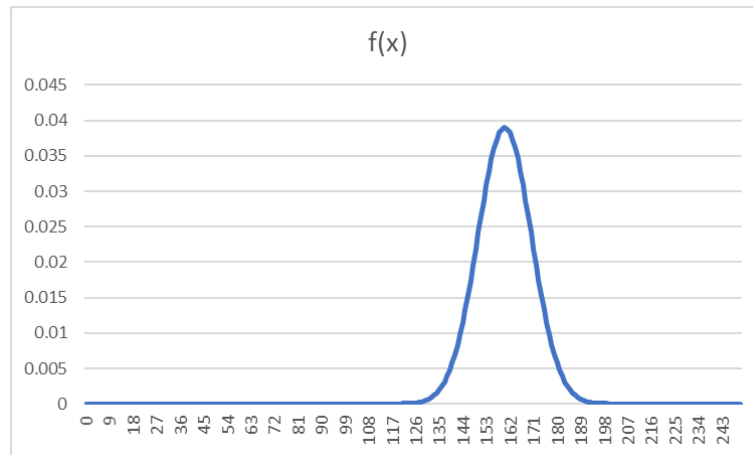
# Distribución normal

Ejemplo: Supongamos la VA de la altura de los universitarios españoles, la cual se distribuye de forma normal con media 160cm y desviación estándar de 10.23

Suponiendo que hay 1500 personas en el estudio, vamos a generar aleatoriamente estos valores

¿Cuál es la probabilidad de que la altura esté entre 159 y 162cm?

Solución: 15.49%





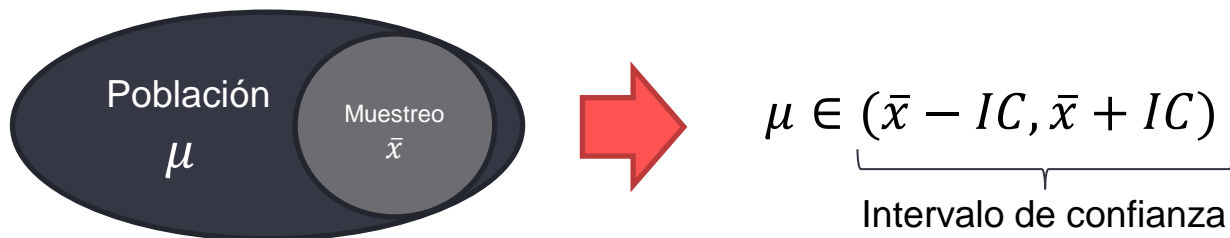
## 3. Estadística Inferencial

### 3.5. Intervalos de confianza

---

# Intervalos de confianza

- El intervalo de confianza nos da una idea del “margen de error” al realizar un muestreo
- El intervalo de confianza nos da un rango en el que podemos estar seguros con cierta probabilidad (normalmente del 95%) de que la media **real** de la población estará en ese rango.



# Intervalos de confianza

- Ejemplo: Disponemos de 100.000 clavos y queremos conocer la longitud media de cada uno de ellos. Para ello, se realiza un muestreo de 100 clavos y se calcula la media de ellos.



10.000 clavos con media  
de longitud  $\mu$



100 clavos con media de  
longitud  $\bar{x}$

$\mu$  estará  
comprendido entre  
 $\bar{x} \pm IC$   
con un 95% de  
probabilidad

# Error estándar de la muestra

- Es la desviación estándar de todas las posibles muestras escogidas en una población

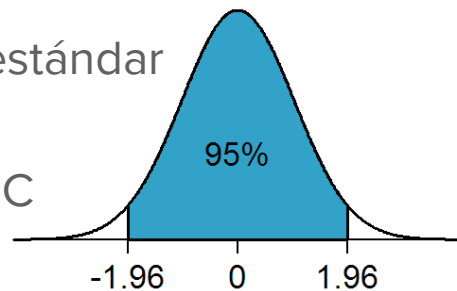
$$SE = \frac{s}{\sqrt{n}}$$

donde  $s$  es la desviación estándar de la muestra

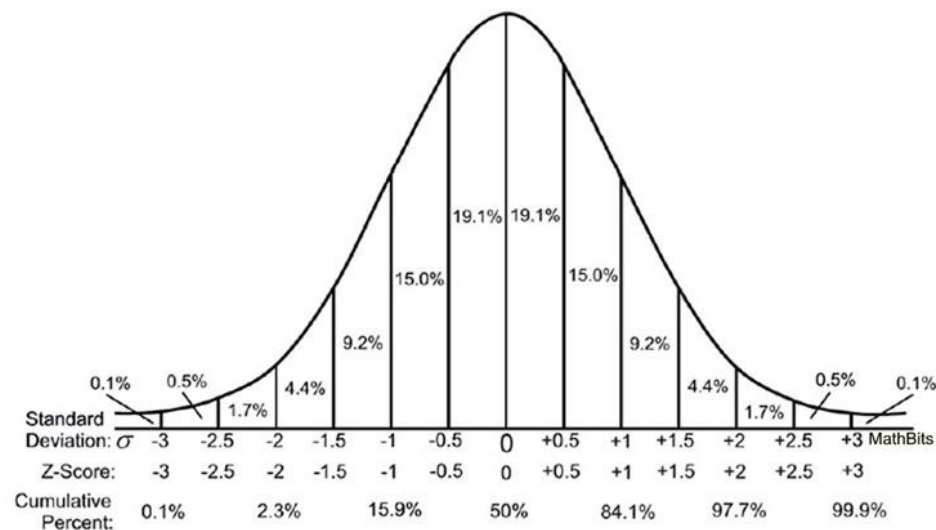
- El intervalo de confianza se calcula como  $IC = \pm 1.96 \cdot SE$

Factor multiplicador

- El valor  $z = 1.96$  proviene del 95% de la distribución normal estándar
- Cuanto más se quiera aumentar la confianza, mayor será el IC



# Error estándar de la muestra



## Confidence Level

## $z^*$ - value

80%

1.28

85%

1.44

90%

1.64

95%

1.96

98%

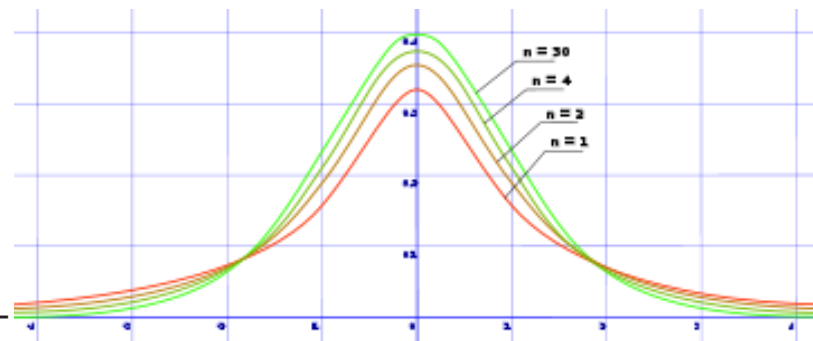
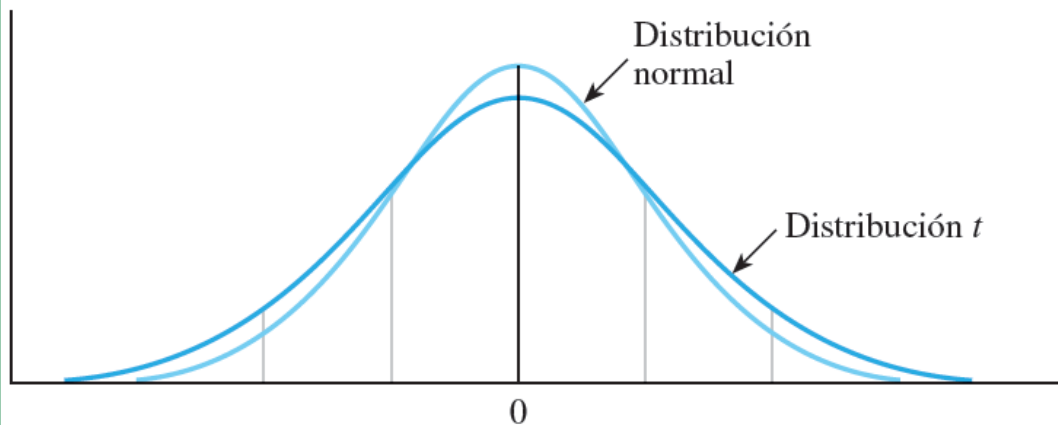
2.33

99%

2.58

# Distribución de t-student

- La distribución t - student es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeña ( $n < 30$ )



- En función de los **grados de libertad** ( $n^\circ$  de muestras -1) se tienen diferentes fdp
- Los valores suelen consultarse en tablas

# Ejemplo 1

- Al salir de una película en el cine, se entrevistan a 11 personas para saber qué puntuación entre 0 y 10 le darían a la película que acaban de ver. Se quiere conocer la media muestral el intervalo de confianza. Las puntuaciones fueron:

2,6,5,5,6,3,7,4,1,1,7

1. Calculamos la media muestral  $\bar{x} = \frac{\sum x_i}{n} = 4.27$
2. Calculamos la desviación estándar de la muestra  $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 2.24$
3. Calculamos el error estándar de la media  $SE = \frac{s}{\sqrt{n}} = \frac{2.24}{\sqrt{11}} = 0.675$
4. Busco en la tabla de t-student el factor multiplicador con  $n = 11$  (10 grados de libertad) y  $p = 0.95$ , obteniendo  $t = 1.8125$
5. Intervalo de confianza:  $(\bar{x} - t \cdot SE, \bar{x} + t \cdot SE) = (4.27 - 0.675 \cdot 1.8125, 4.27 + 0.675 \cdot 1.8125)$   
 $= (3.05, 5.49)$  (La nota media real de la película está entre esos valores con un 95% de confianza)

## Ejemplo 2

- Se quiere conocer la altura de los estudiantes de cursos de análisis de datos con una confianza del 99%, teniendo las siguientes muestras:

180,165,176,165,169,179,168,176,191,178,173,157,175,179,169,185,168,  
170,166,178,177,180,168,179,173,162,175,175,180,167

1.  $\bar{x} = 173.43$

2.  $s = 7.24$

3.  $SE = \frac{s}{\sqrt{n}} = \frac{7.24}{\sqrt{30}} = 1.32$

4.  $z \cdot SE = 2.58 \cdot 1.32 = 3.40$

5. *Intervalo de confianza:*  $(\bar{x} - z \cdot SE, \bar{x} + z \cdot SE) = (170, 176)$



## 3. Estadística Inferencial

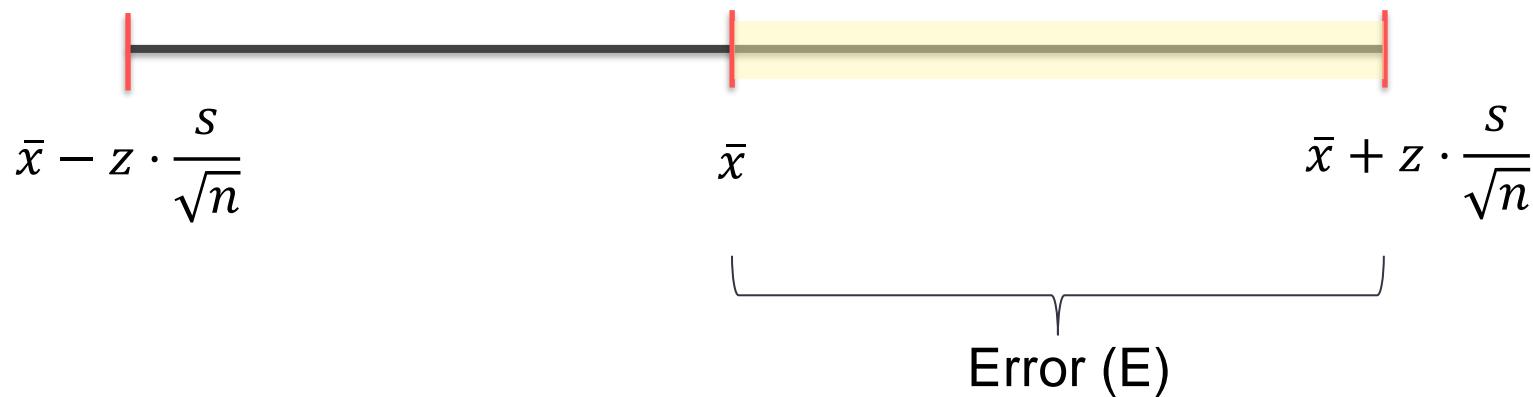
### 3.6. Error absoluto y tamaño de la muestra

---

# Error absoluto y tamaño de la muestra

- ¿Cuántos datos hemos de tener para que nuestro estudio tenga validez? Es una pregunta muy genérica y sin respuesta
- ¿Cuántos datos necesitamos que al estimar una media poblacional el error máximo que cometemos sea menor que una cantidad que previamente especificamos?
- En Estadística nunca podemos afirmar con seguridad nada. Siempre hacemos afirmaciones basadas en la probabilidad
- El error absoluto es la mitad de la longitud del intervalo de confianza

# Estimar una media



$$E = z \cdot \frac{s}{\sqrt{n}} \Rightarrow n = \left( \frac{z \cdot s}{E} \right)^2$$

“Para asumir un cierto error acerca de la media, con un grado de confianza determinado, necesito  $n$  muestras”

# Estimar una media

- Ejemplo: deseamos conocer la media del nivel de azúcar en un refresco, con una seguridad del 95% y una precisión de  $\pm 3$  mg/dl y tenemos información bibliográfica de que la varianza es de 250 mg/dl

$$n = \left( \frac{z \cdot s}{E} \right)^2 = \left( \frac{1.96 \cdot \sqrt{250}}{3} \right)^2 = 106.7$$

- Necesitaría tomar al menos 107 muestras para mantener esa precisión

# Estimar una proporción (población total desconocida)

$$n = \left(\frac{z}{E}\right)^2 \cdot p \cdot (1 - p)$$

- Si deseamos estimar una proporción, debemos tener una idea aproximada del parámetro que queremos medir (en este caso una proporción). En caso de no tener dicha información utilizaremos el valor  $p=0.5$  (50%), que maximiza el tamaño muestral.
- Ejemplo: Sabiendo que un 5% de la población tiene diabetes, ¿a cuántas personas habría que examinar para conocer la proporción de diabetes con una precisión del 3% y una confianza del 95%?
  - $n = \left(\frac{1.96}{0.03}\right)^2 \cdot 0.05 \cdot 0.95 = 203$

## Estimar una proporción (población total conocida)

$$n = \frac{N \cdot z^2 \cdot p \cdot (1 - p)}{E^2 \cdot (N - 1) + z^2 \cdot p \cdot (1 - p)}$$

donde N es el tamaño de la población

- Ejemplo: ¿A cuántas personas tendría que estudiar de una población de 15.000 habitantes para conocer la prevalencia de diabetes?

$$\text{➤ } n = \frac{15000 \cdot 1.96^2 \cdot 0.05 \cdot 0.95}{0.03^2 (15000 - 1) + 1.96^2 \cdot 0.05 \cdot 0.95} = 200$$

# Estimar una proporción (población total conocida)

Tamaño de la población	Tamaño de la muestra por margen de error		
	$\pm 3\%$	$\pm 5\%$	$\pm 10\%$
500	345	220	80
1000	525	285	90
3000	810	350	100
5000	910	370	100
10 000	1000	385	100
100 000	1100	400	100

# Calculadora online

- <https://www.netquest.com/es/gracias-calculadora-muestra>

5000

## TAMAÑO DEL UNIVERSO

Número de personas que componen la población a estudiar.

50

## HETEROGENEIDAD %

Es la diversidad del universo. Lo habitual suele ser 50%.

5

## MARGEN DE ERROR

Menor margen de error requiere mayores muestras.

95

## NIVEL DE CONFIANZA

Cuanto mayor sea el nivel de confianza, mayor tendrá que ser la muestra (95% - 99%).

# El tamaño de muestra que necesitas es...

# 357

### El resultado anterior debe interpretarse así:

Si encuestas a 357 personas, el 95% de las veces el dato que quieres medir estará en el intervalo  $\pm 5\%$  respecto al dato que observes en la encuesta.



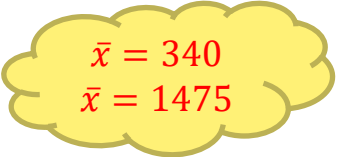
## 3. Estadística Inferencial

### 3.7. Contraste de hipótesis

---

# Ejemplo

- Un fabricante de bombillas afirma que sus bombillas tienen una duración media de 1500 horas.
  - Nuestro problema es tomar una entre dos decisiones: admitir que lo que afirma es correcto o bien que no lo es y la duración media real no es igual a 1500
  - Lo primero que necesitamos para tomar la decisión son datos. Se toma una muestra de bombillas y se repite el experimento consistente en tenerlas en funcionamiento ininterrumpido hasta que la bombilla deja de funcionar.
  - La afirmación del fabricante la consideramos como una hipótesis que hemos de evaluar. En concreto nos planteamos dicha hipótesis y su negación
- 
- $H_0: \mu = 1500$  (hipótesis nula)
  - $H_1: \mu \neq 1500$  (hipótesis alternativa)


$$\bar{x} = 340$$

$$\bar{x} = 1475$$

# Contraste de hipótesis

- Es un método estadístico que nos va a permitir aceptar o rechazar una determinada afirmación que realizamos (hipótesis nula) en función de los valores obtenidos en una muestra
- El objetivo del contraste de hipótesis no es decidir si la hipótesis válida es  $H_0$  o  $H_1$ , solo podemos rechazar  $H_0$  basándonos en que la probabilidad de que sea errónea es elevada
- Haciendo un símil con un juicio, donde  $H_0$  representaría la inocencia de un acusado y  $H_1$  la culpabilidad:
  - Para poder asegurar que es culpable (rechazar  $H_0$ ) tengo que tener muchas pruebas
  - Si no existen pruebas suficientes, no podemos asegurar que el acusado sea inocente.

# Pasos a seguir

1. Identificar el parámetro que vamos a estudiar (media, varianza, desviación...)
2. Especificar la hipótesis nula  $H_0$  y la hipótesis alternativa  $H_1$
3. Fijar un valor para el nivel de confianza
4. Obtener el valor del estadístico para la muestra elegida
5. Determinar la región de aceptación y la región de rechazo
6. Decidir si rechazamos o no rechazamos la hipótesis nula
7. Interpretar los resultados obtenidos

# Contraste de hipótesis bilateral: media

$H_0: \mu = k$  (hipótesis nula)

$H_1: \mu \neq k$  (hipótesis alternativa)

Estadísticos:

$Z = \frac{\bar{x} - k}{\sigma/\sqrt{n}}$  si se conoce  $\sigma$ ; sigue una distribución  $N(0,1)$

$T = \frac{\bar{x} - k}{s/\sqrt{n}}$  si no se conoce  $\sigma$ ; sigue una distribución t-student con  $n-1$  grados de libertad



# Contraste de hipótesis bilateral: media

Ejemplo: Un fabricante afirma que la duración media de sus bombillas es de 1500 horas. Se toma una muestra de 100 bombillas y se utilizan hasta fundirse, obteniendo una media de 1405 horas y una desviación típica de 323 horas

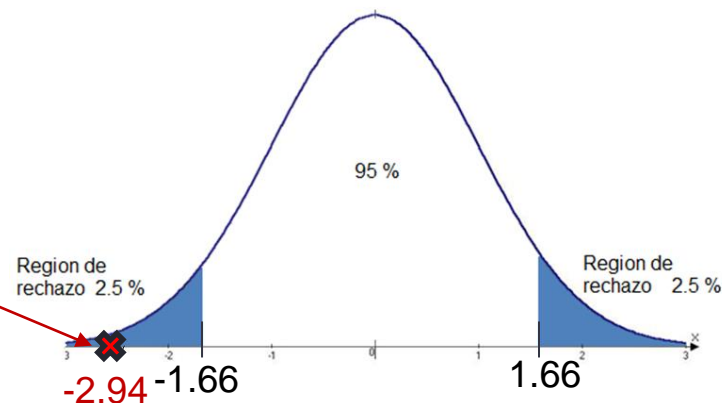
$H_0: \mu = 1500$  (hipótesis nula)

$H_1: \mu \neq 1500$  (hipótesis alternativa)

Estadístico  $T = \frac{1405-1500}{323/\sqrt{100}} = -2.94$

(sigue una distribución t-student)

El valor de la distribución t-student con 99 grados de libertad y  $p=95\%$  es de 1.66



⇒ **Se rechaza la hipótesis nula**  
(el fabricante miente)

# Contraste de hipótesis unilateral: media

$H_0: \mu \geq k$  (hipótesis nula)

$H_1: \mu < k$  (hipótesis alternativa)

Estadísticos:

$Z = \frac{\bar{x} - k}{\sigma/\sqrt{n}}$  si se conoce  $\sigma$ ; sigue una distribución  $N(0,1)$

$T = \frac{\bar{x} - k}{s/\sqrt{n}}$  si no se conoce  $\sigma$ ; sigue una distribución t-student con  $n-1$  grados de libertad



# Contraste de hipótesis unilateral: media

Ejemplo: Otro fabricante afirma que la duración media de sus bombillas es de al menos 1500 horas. Se toma una muestra de 100 bombillas y se utilizan hasta fundirse, obteniendo una media de 1595 horas y una desviación típica de 323 horas

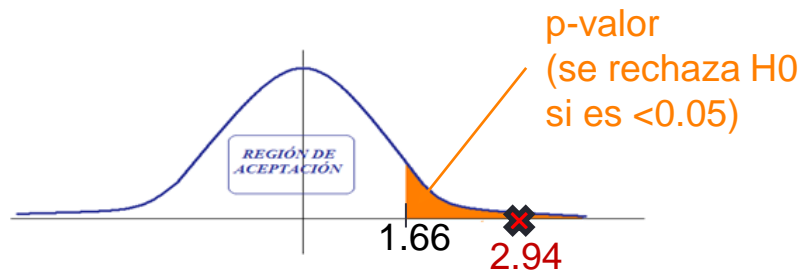
$H_0: \mu \leq 1500$  (hipótesis nula)

$H_1: \mu > 1500$  (hipótesis alternativa)

$$\text{Estadístico } T = \frac{1595 - 1500}{323 / \sqrt{100}} = 2.94$$

(sigue una distribución t-student)

El valor de la distribución t-student con 99 grados de libertad y  $p=95\%$  es de 1.66



⇒ **Se rechaza la hipótesis nula**  
(el fabricante dice la verdad)





# ¡Gracias!

Contacto: Rafael Zambrano  
rafael@thebridgeschool.es