



# Introducción a la Estadística

RAQUEL SOCORRO LEÓN

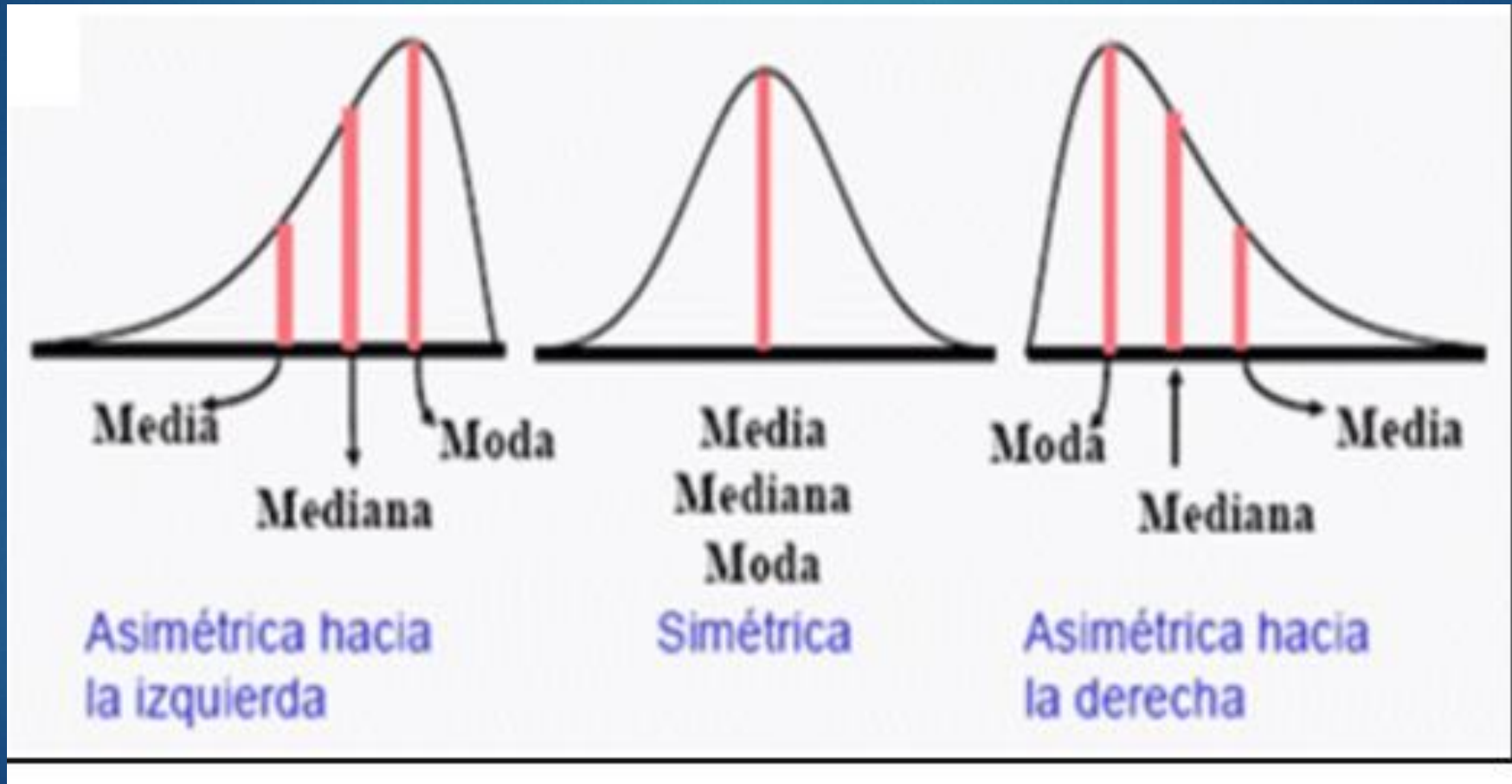
# Principales estimadores (I)

Distribución de una variable son las propiedades de los valores observados, por ejemplo: una variable puede ser las ventas de un producto y sus observaciones pueden ser las cifras de esta ventas.

- ▶ Media aritmética: conjunto de datos numéricos es su valor medio.
- ▶ Mediana: es el valor central de la distribución, es decir, es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor.
- ▶ Moda: es el valor con mayor frecuencia en una de las distribuciones de datos

Tanto la mediana como la media aritmética miden el centro de una distribución. Solo cuando la distribución es simétrica, ambos valores coinciden.

# Tipos de distribuciones



# Principales estimadores (II)

- ▶ Varianza: es la media aritmética de los cuadrados de las desviaciones respecto a la media aritmética, es decir, es el promedio de las desviaciones de la media elevadas al cuadrado.
- ▶ Desviación estándar: es la raíz cuadrada de la varianza, de modo que nos da una medida de la dispersión en torno a la media, pero expresada en la misma escala que la variable.



# Muestra VS Población

Para obtener información promedio de una muestra se llama Estadístico. Mientras que en población se llama Parámetro.

Estimador

Estadístico

Parámetro

Media

$\bar{X}$

$\mu$

Varianza

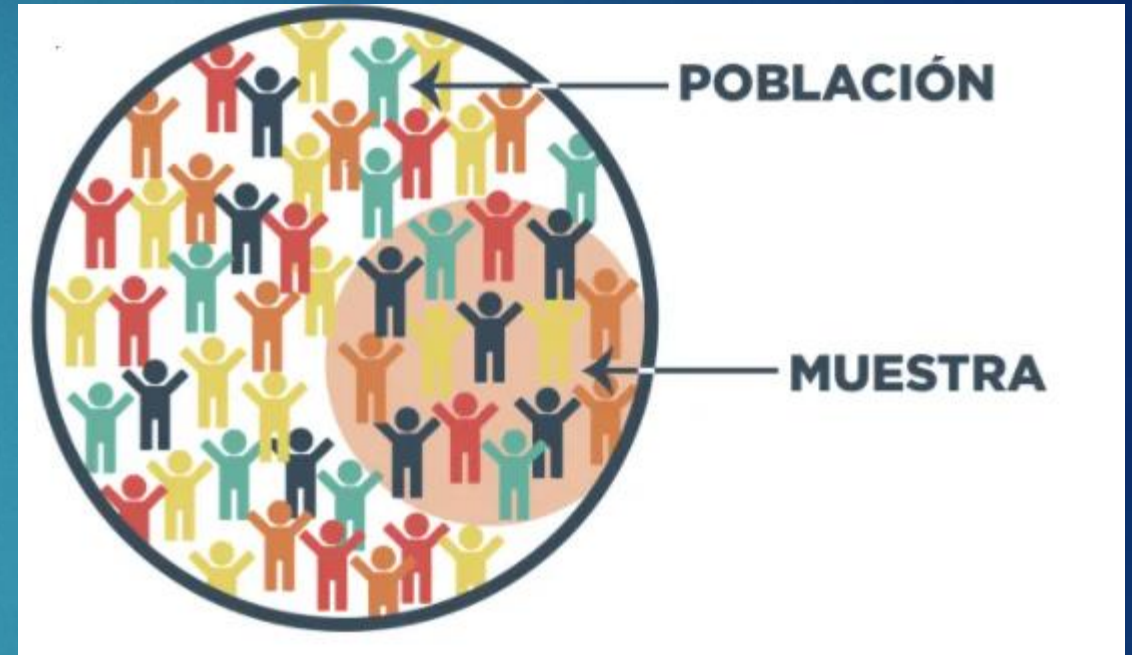
$S^2$

$\sigma^2$

Desviación típica

$S$

$\sigma$



Por ejemplo:

- Obtener la media de uso de coches en España → Parámetro
- Obtener la media de uso de coches de 100.000 habitantes en España → Muestra

# Distribución Normal (I)

- ▶ La curva de densidad es la representación gráfica de la distribución de una variable suponiendo que fuéramos capaces de obtener muchísimas observaciones.
- ▶ La distribución normal también es conocida como campana de Gauss por la forma que describe su función de densidad, que recuerda la forma de una campana.

# Distribución Normal (II)

## ► PROPIEDADES:

Es una distribución simétrica. El valor de la media, mediana y moda coinciden.

Parámetros →

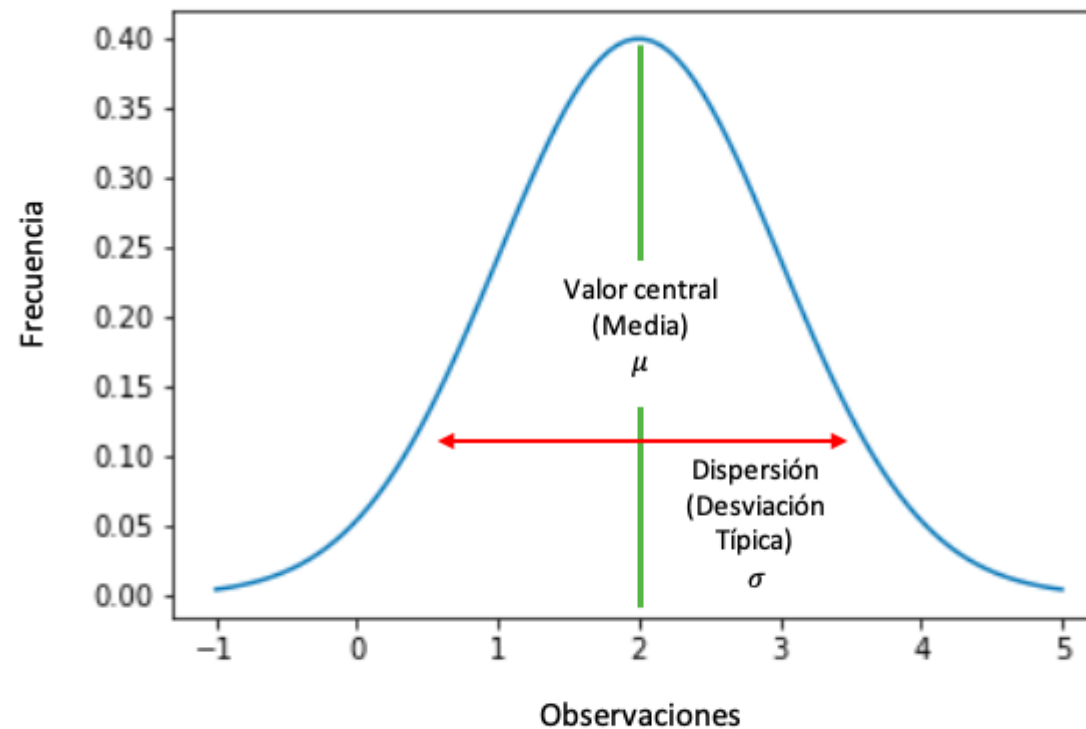
*media o valor central* =  $\mu$

*desviación típica* =  $\sigma$

# Distribución Normal (III)

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

Variable  
aleatoria  $X$   
aproximada  
a una  
distribución  
normal.



*Función de densidad de una distribución normal.*



# Distribución Normal (IV)

- Estandarización o Normalización de variables: es un procedimiento de transformación de la variable original, restándole a cada valor la media y dividiendo esto entre la desviación típica.

Z-scores



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



Distribución Normal Estándar

cuya media es 0 y desviación típica es 1

# Distribución t de Student (I)

## ► PROPIEDADES:

Estimar la media de una poblacional normalmente distribuida a partir de una muestra pequeña.

Es una distribución simétrica. El valor de la media, mediana y moda coinciden.

Es mas aplanada del centro y tiene colas mas anchas

Tamaño de la muestra es inferior a 30, es decir  $n < 30$

Grados de libertad: es igual al número de observaciones menos el número de relaciones requeridas entre las observaciones (por ejemplo, estimar la media). Por tanto,  $n - 1$

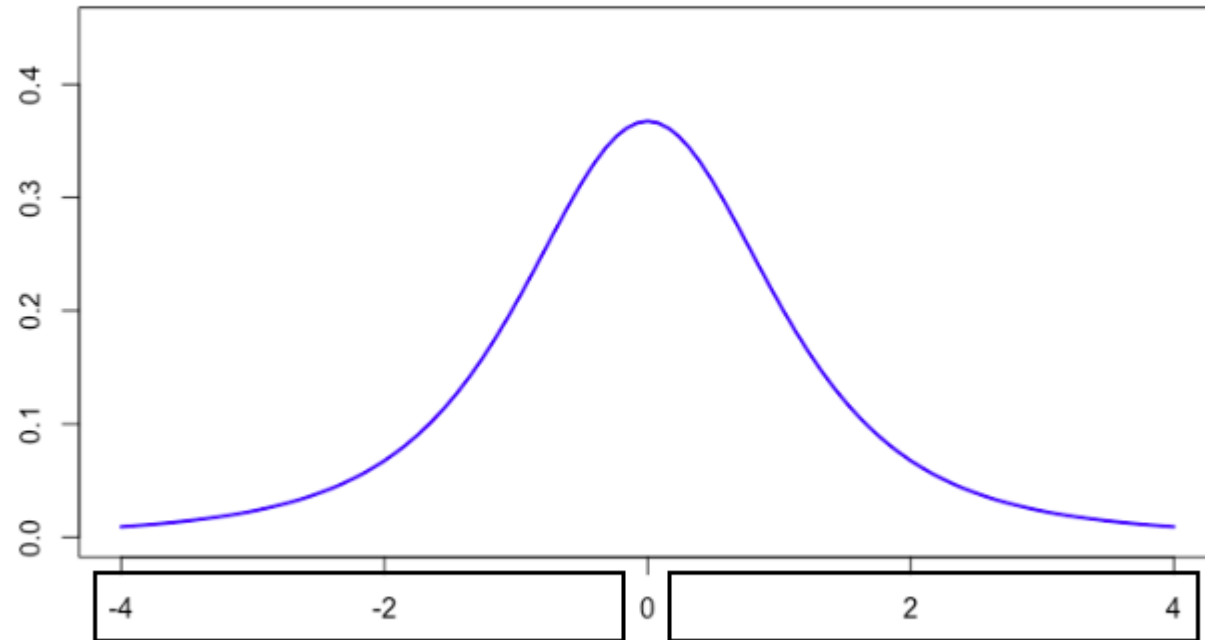
# Distribución t de Student (I)

$$L \sim t_g$$

La variable aleatoria L  
sigue una distribución t  
con g grados de libertad



Función de densidad de la distribución t de Student con df = 3



*Función de densidad de la distribución t con 3 grados de libertad.*

# Distribución Normal VS Distribución t

- ▶ Ambas distribuciones son similares, aunque en la Distribución t dispone de lo siguiente:
  - ▶ Grados de libertad  $\rightarrow n - 1$
  - ▶ Las colas son más gruesas
  - ▶  $\downarrow$  tamaño muestral  $\uparrow$  tamaño de las colas
  - ▶  $\uparrow$  tamaño muestral se asemeja a la distribución normal

Distribución t de Student (azul) y distribución Normal estándar  $N(0,1)$  (naranja)



*Función de densidad de una distribución t (línea azul) y de una distribución Normal estándar (línea naranja).*



# Contrastes de hipótesis e Intervalos de Confianza (I)

- ▶ **Inferencia Estadística:** El proceso de analizar una muestra para llegar a una conclusión sobre una propiedad de la población.
- ▶ **Contraste de hipótesis:** Se trata de comparar las predicciones con la realidad que observamos ocurrida en una muestra aleatoria y significativa, que permita aceptar o rechazar una hipótesis, sobre el valor de un parámetro desconocido de una poblacional. Se trata de un proceso estadístico que permite elegir una hipótesis de trabajo de entre dos posibles y antagónicas.
- ▶ **Tipos de hipótesis:**
  - ▶ **Hipótesis nula:** partimos del supuesto de que las diferencias entre el valor verdadero del parámetro y su valor hipotético, en realidad no son tales sino debidas al azar, es decir no hay diferencia o dicho de otra forma la diferencia es nula. Se representa como  $H_0$ .
  - ▶ **Hipótesis alternativa:** Es la negación de la hipótesis nula y generalmente representa la afirmación que se pretende probar. Se representa como  $H_1$

Normalmente cuando queremos plantear las hipótesis de una determinada situación debemos tener en cuenta que aquello que queramos demostrar irá siempre a la hipótesis alternativa ya que el error que cometemos cuando rechazamos  $H_0$  lo podemos medir (está fijado de antemano por el nivel de significación).

# Contrastes de hipótesis e Intervalos de Confianza (II)

## ► Tipos de contrastes:

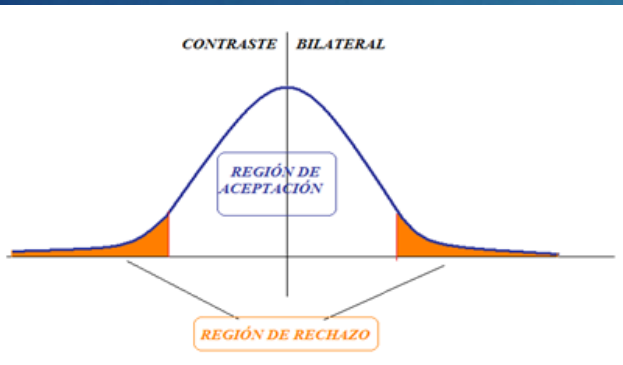
### Bilateral

Hipótesis nula

$$H_0 \quad \mu = \mu_0$$

Hipótesis alternativa

$$H_1 \quad \mu \neq \mu_0$$



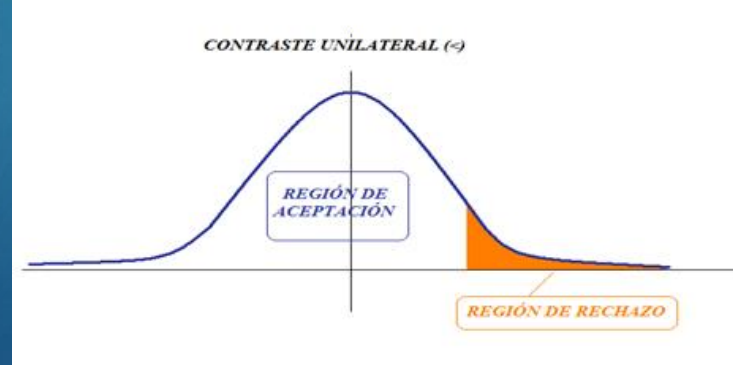
### Unilateral Derecho

Hipótesis nula

$$H_0 \quad \mu \leq \mu_0$$

Hipótesis alternativa

$$H_1 \quad \mu > \mu_0$$



### Unilateral Izquierdo

Hipótesis nula

$$H_0 \quad \mu \geq \mu_0$$

Hipótesis alternativa

$$H_1 \quad \mu < \mu_0$$



# Contrastes de hipótesis e Intervalos de Confianza (III)

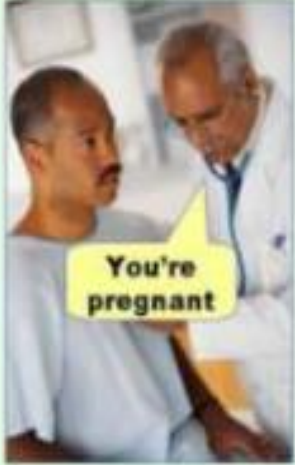

## Tipos de Errores:

Cometeremos un **error de tipo I** cuando rechazamos la hipótesis nula cuando en realidad es cierta. Llamaremos  $\alpha$  a la probabilidad de cometer un error de tipo I, esto es:  $P(\text{elegir } H_1 \mid H_0 \text{ es cierta}) = \alpha$ .

- Cometeremos un **error de tipo II** cuando rechazamos la hipótesis alternativa cuando en realidad es cierta. Llamaremos  $\beta$  a la probabilidad de cometer un error de tipo II, esto es:  $P(\text{elegir } H_0 \mid H_1 \text{ es cierta}) = \beta$ .

	Situación cierta	
	$H_0$	$H_1$
Elegimos $H_0$	Correcta	Error de tipo II
Elegimos $H_1$	Error de tipo I	Correcta

$H_0$  : You are not pregnant  
 $H_A$  : You are pregnant

Type I error (false positive)	Type II error (false negative)
	

# Contrastes de hipótesis e Intervalos de Confianza (IV)

## ► Intervalos de Confianza

Nos permite calcular los valores alrededor de una media muestral (uno superior y otro inferior) con un nivel de confianza  $1 - \alpha$

Los valores más habituales del nivel de confianza  $1 - \alpha$  son 0.9, 0.95 o 0.99 (la confianza de 90%, 95% 99%). En ocasiones también se emplea la terminología **nivel de significación** para el valor  $\alpha$



# Contrastes de hipótesis e Intervalos de Confianza (V)

Tipos de PRUEBAS:

- z.test → Contrasta media
- t.test → Contrasta media
- ANOVA → Contrasta varianza

# Prueba z test (I)

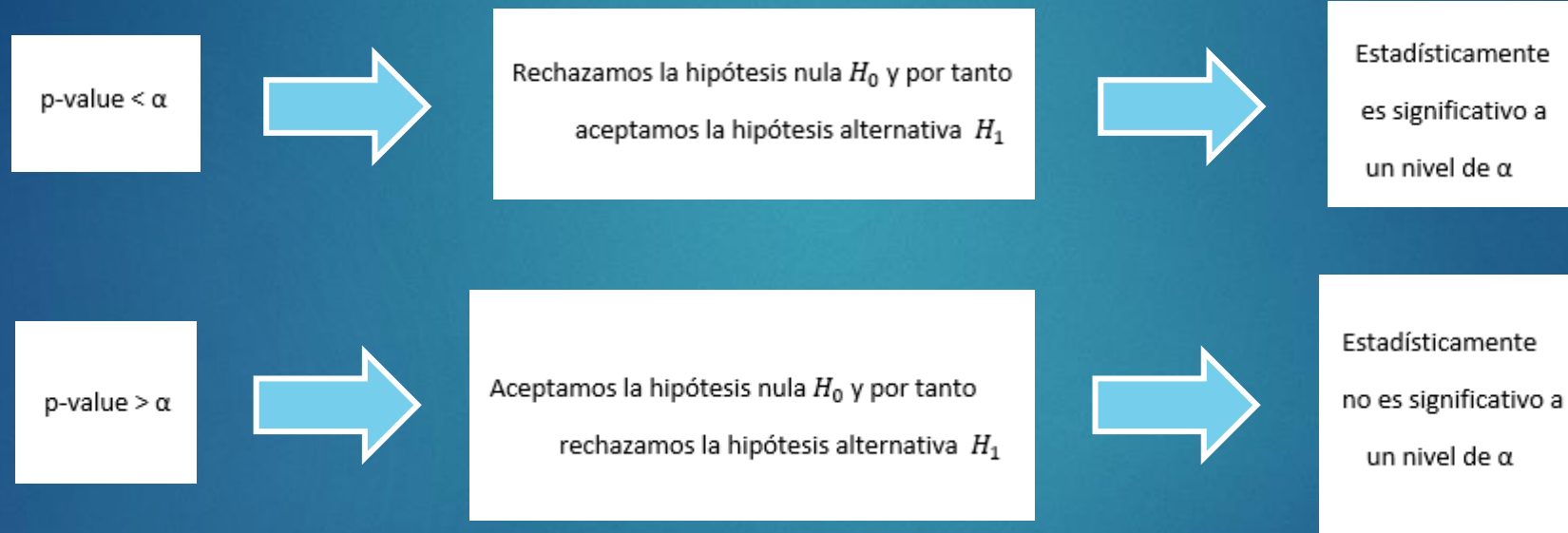
- ▶ Cuando se tiene una población con distribución normal y se conoce la varianza es posible emplear el estadístico Z para la prueba de hipótesis.
- ▶ Se utiliza para tamaño muestrales  $> 30$
- ▶ Su estadístico de prueba: Es una variable aleatoria que se calcula a partir de datos de muestra y se utiliza en una **prueba** de hipótesis. El estadístico de prueba se utiliza para calcular el valor  $p$ . En este caso es el siguiente:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- ▶  $p$  value es una medida de la fuerza de la evidencia en sus datos en contra de  $H_0$ . Se utiliza para compararse con el nivel de significación. De esta manera determinamos aceptar o rechazar la hipótesis nula.

# Prueba z test (II)

Por tanto, concluimos con p-value:



# Prueba t-Student

- ▶ Es una prueba de hipótesis sobre la media de una población normal con varianza desconocida.
- ▶ Se utiliza para tamaño muestrales  $< 30$
- ▶ Su estadístico de prueba: Es una variable aleatoria que se calcula a partir de datos de muestra y se utiliza en una **prueba** de hipótesis. El estadístico de prueba se utiliza para calcular el *valor p*. En este caso es el siguiente:

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

- ▶ p value es una medida de la fuerza de la evidencia en sus datos en contra de  $H_0$ . Se utiliza para compararse con el nivel de significación. De esta manera determinamos aceptar o rechazar la hipótesis nula.



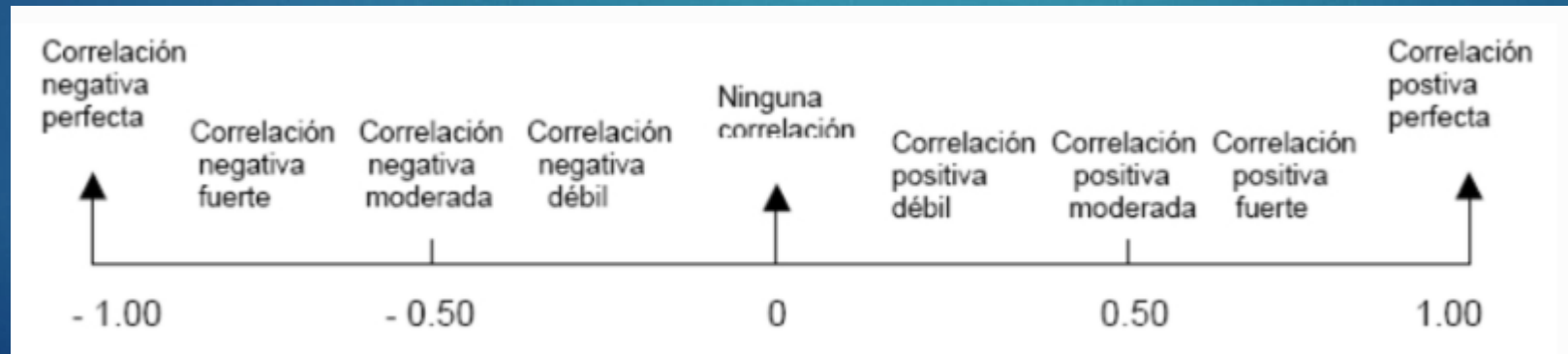
# ANOVA

- ▶ Análisis de varianza (ANOVA) permite contrastar si dos muestras presentan igualdad de varianzas (homocedasticidad).
- ▶ Utiliza el estadístico F o también conocido como Fisher.
- ▶ Las varianzas son una medida de dispersión, es decir, qué tan dispersos están los datos con respecto a la media. Los valores más altos representan mayor dispersión. La varianza es el cuadrado de la desviación estándar

# Correlaciones (I)

- El coeficiente de correlación de Pearson es una medida del nivel de asociación lineal entre dos variables. Puede tomar valores entre -1 y +1 porque es una medida normalizada; por este motivo, es independiente de la escala en la que esté representada cada variable. Su fórmula:







$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$



# Correlaciones (II)

Para visualizar la relación de dos variables se utiliza el diagrama de dispersión o nube de puntos. A continuación, se presenta diferentes tipos de relaciones:

- Sin relación
- Alta correlación positiva
- Baja correlación positiva
- Fuerte correlación negativa
- Débil correlación negativa
- Relación compleja

Diagrama	Tipo de Relación
	Sin relación. No se aprecia ninguna correlación entre las dos variables.
	Alta correlación positiva. El valor de Y se incrementa nitidamente a medida que el valor de X aumenta.
	Baja correlación positiva. El valor de X aumenta ligeramente a medida que aumenta el valor de Y.
	Fuerte correlación negativa. El valor de X claramente disminuye a medida que aumenta el valor de Y.
	Débil correlación negativa. El valor de X disminuye ligeramente a medida que aumenta el valor de Y.
	Relación compleja. El valor de X parece estar relacionado con el valor de Y, pero esa relación no es fácil de establecer.

# Regresión Lineal (I)

- ▶ Utilizamos este tipo de modelo cuando queremos predecir – o explicar – la relación que existe entre una variable dependiente/explicada/predicha/respuesta **Y** a partir de una o más variables independientes/explicativas/predictora **x**
- ▶ La regresión lineal puede ser: Simple o Múltiple
- ▶ Se trata del tipo de estructura más simple, en el que los valores de y dibujan una línea aproximadamente recta. Definiremos esta **recta** mediante la siguiente ecuación.

The diagram shows the equation  $y = mx + b$  on a light blue background. Above the equation, the word "Pendiente" is positioned above  $m$  and "Intercepto" is positioned above  $b$ , both with red arrows pointing down to their respective variables. Below the equation, the word "Variable" is positioned above "Dependiente" (under  $y$ ) and "independiente" (under  $x$ ), with red arrows pointing up to the variables  $y$  and  $x$ .

Y = variable dependiente  
m = pendiente de la recta, nos indica como un cambio en x afecta a y.  
X = variable independiente  
b = intercepto, ordenada al origen, es decir,  $x = 0$



# Regresión Lineal (II)

## Coeficiente de determinación o R cuadrado

- ▶ Es el porcentaje de variación de la variable de respuesta/dependiente que explica su relación con una o más variables predictoras/independientes.
- ▶ Por lo general, mientras mayor sea el  $R^2$ , mejor será el ajuste del modelo a sus datos, es decir, cuanto mayor sea la varianza que explica el modelo de regresión, más cerca estarán los puntos de los datos de la línea de regresión ajustada.
- ▶ El  $R^2$  siempre se encuentra entre 0 y 1.
- ▶ Su fórmula es la siguiente:

$$R^2 = \frac{\sum_{t=1}^T (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^T (Y_t - \bar{Y})^2}$$

# Regresión Lineal (III)

## Error Residual Estándar

La varianza del error  $\epsilon$  es en general desconocida, pero se puede estimar a partir de los datos. Esta estimación es conocida como el error estándar residual (RSE), que no es más que la raíz cuadrada de la media de la suma de los residuos al cuadrado.

RSE nos dará una estimación sobre la desviación promedio de cualquier punto respecto a la verdadera recta de regresión.

Su fórmula:

$$s_{\epsilon} = \sqrt{\frac{SSE}{n-2}}$$

