

# **FUNDAMENTOS DE LA BIOESTADÍSTICA INFERENCIAL EN R**

PROF. ASTRID LILIANA VARGAS SANCHEZ

M.S.C EN BIOINFORMÁTICA

UNIVERSIDAD INTERNACIONAL DE LA RIOJA

# CUÉNTAME SOBRE TI

¿DE DONDE ERES?

¿QUE NIVEL DE ESTUDIOS TIENES?

¿TIENES CONOCIMIENTOS DE  
BIOESTADÍSTICA?

¿TIENES CONOCIMIENTOS DE  
PROGRAMACIÓN?

# FUNDAMENTOS DE LA BIOESTADISTICA INFERENCIAL

MODULO 1



# BIOESTADISTICA

Rama de la estadística que recopila, organiza, procesa e interpreta datos para deducir características de un grupo. Se aplica al estudio de los seres vivos, especialmente en las áreas de la medicina, la biología, la salud pública y las ciencias de la vida.

Bioestadística  
descriptiva

Bioestadística  
inferencial

# BIOESTADÍSTICA DESCRIPTIVA

Se encarga de resumir y describir los datos mediante medidas numéricas (como promedios, medianas y desviaciones estándar) y representaciones gráficas (como tablas, gráficos de barras o histogramas). Su objetivo es organizar y presentar la información de forma comprensible, sin realizar inferencias o predicciones. Información acerca de la distribución, dispersión y forma de los datos.

# BIOESTADÍSTICA INFERENCIAL

Permite tomar decisiones o hacer generalizaciones sobre una población a partir de los datos obtenidos de una muestra representativa. No se limita solo a describir los datos (como la estadística descriptiva), sino que busca extraer conclusiones más amplias utilizando métodos matemáticos y probabilísticos.

Establece relaciones entre las características observadas.

Su misión es hacer inferencias científicas y contrastar hipótesis

## Ejemplo: Usos de facebook

Bioestadística descriptiva:  
Sabemos que en promedio 14  
millones de personas entran a  
Facebook en un dia

Bioestadística inferencial:  
Podemos conocer si las  
personas que duran mas  
tiempo en Facebook en un  
día pueden tener un riesgo  
mayor de tener adicción a  
las redes sociales

Ejemplo: Un investigador prueba dos medicamentos distintos para reducir la presión arterial en un grupo de pacientes.

Bioestadística descriptiva:  
Promedio de la presión arterial  
en cada grupo.

Bioestadística inferencial:  
Uso de una prueba t para  
comparar los promedios de  
presión arterial entre los  
dos grupos y ver si la  
diferencia observada se  
puede generalizar a toda la  
población.

# APLICACIONES DE LA BIOESTADÍSTICA INFERENCIAL

La bioestadística inferencial se utiliza en áreas como ensayos clínicos, epidemiología, genética y salud pública. Permite extraer conclusiones más allá de los datos observados, evaluando su grado de certeza. Esto ayuda a tomar decisiones informadas basadas en evidencia y a impulsar la investigación en salud y biomedicina.



# MUESTRA

Es un subconjunto representativo de una población que se selecciona para estudiar y obtener conclusiones sobre esa población sin tener que analizar a todos sus integrantes con el objetivo de estudiar sus características y hacer inferencias o generalizaciones sobre toda la población. La elección de una muestra adecuada es crucial en el análisis estadístico, ya que una muestra inadecuada puede llevar a conclusiones erróneas o sesgadas.



# VARIABLES EN BIOESTADISTICA

Tipo de Variable	¿Qué es?	Ejemplo	Usos en análisis
Variable independiente	Es la que se manipula o clasifica para observar su efecto sobre otra variable.	Tipo de tratamiento (placebo, dosis baja, dosis alta)	Determina los grupos o condiciones a comparar.
Variable dependiente	Es la que se mide o evalúa para observar el efecto de la variable independiente.	Presión arterial, nivel de glucosa, peso corporal	Resultado principal del estudio.
Variable emparejada	Se refiere a datos relacionados o repetidos de un mismo sujeto o unidad experimental.	Peso antes y después de un tratamiento en el mismo paciente	Pruebas pareadas como t de Student pareada o Wilcoxon
Variable de agrupación	Categórica que divide los datos en grupos o niveles para comparación.	Grupo sanguíneo, sexo, grupo experimental	Agrupa los datos en el análisis estadístico.

# MEDIDAS DE TENDENCIA CENTRAL

Valor que representa la posición central de una distribución de datos

## MEDIA

Es una medida estadística utilizada para obtener el valor promedio de un conjunto de datos. Es la suma de todos los valores dividida entre la cantidad de datos. Es sensible a datos atípicos.

$$m = \frac{\text{suma de los términos}}{\text{número de términos}}$$

# MEDIDAS DE TENDENCIA CENTRAL

## MEDIANA

Representa el valor central en un conjunto de datos ordenados. Reduce el impacto de los valores atípicos. Si hay un número par de datos, es el promedio de los dos valores centrales.

Ejemplo impar:

Datos: 3, 5, 7 → la mediana es 5

Ejemplo par:

Datos: 2, 4, 6, 8 → la mediana es  $(4+6)/2=5$

# MEDIDAS DE DISPERSIÓN

## VARIANZA

Mide cuánto se alejan, en promedio, los datos con respecto a la media, elevando al cuadrado esas diferencias. Se mide en unidades al cuadrado, por ejemplo, si tus datos están en centímetros, la varianza estará en  $\text{cm}^2$ .

$$Var(X) = \frac{\sum_1^n (x_i - \bar{X})^2}{n}$$

Donde:

- $x_i$ : Cada dato
- $\bar{x}$ : La media
- n: Número de datos

# MEDIDAS DE DISPERSIÓN

## DESVIACIÓN ESTANDAR

Es la raíz cuadrada de la varianza. Mide cuánto se desvían, en promedio, los datos con respecto a la media, pero en las mismas unidades del dato original.

$$\text{Desviación est\'andar} = \sqrt{\text{Varianza}}$$

# PROBABILIDAD

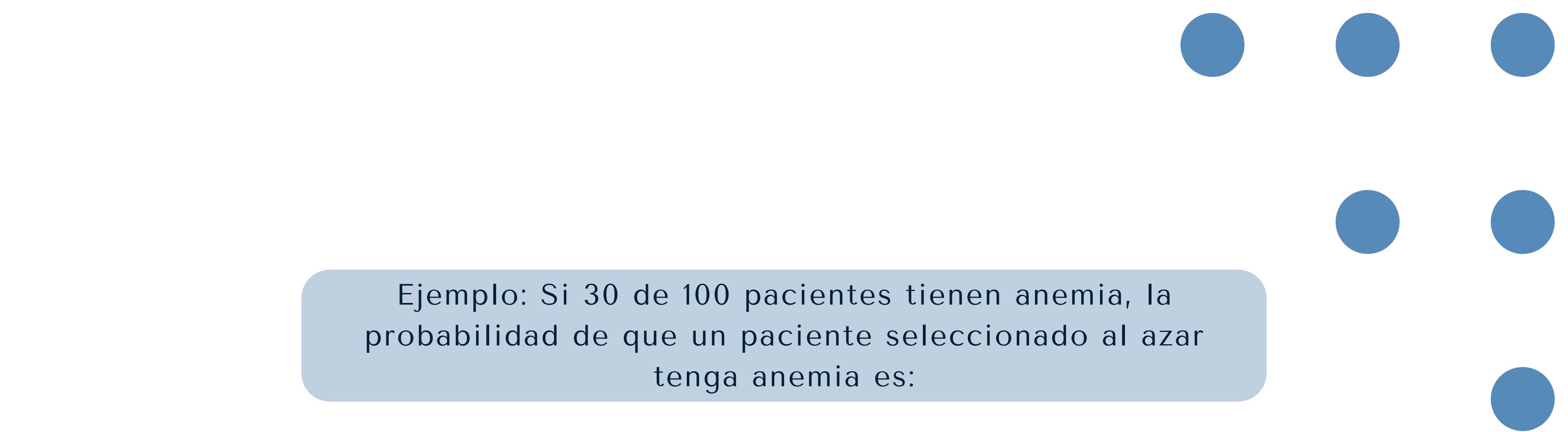
La probabilidad es una medida numérica que indica qué tan probable es que ocurra un evento. Se expresa como un número entre 0 y 1, donde:

0 significa que el evento nunca ocurrirá (imposible).

1 significa que el evento ocurrirá con certeza (seguro).

Un valor como 0.5 indica una probabilidad del 50%, es decir, hay la misma posibilidad de que ocurra o no.

$$\text{Probabilidad (P)} = \frac{\text{Número de casos favorables}}{\text{Número total de casos posibles}}$$



Ejemplo: Si 30 de 100 pacientes tienen anemia, la probabilidad de que un paciente seleccionado al azar tenga anemia es:

$$P(\text{anemia}) = \frac{30}{100} = 0.3$$

# DISTRIBUCIÓN DE PROBABILIDAD

Es una forma matemática de describir cómo se comporta una variable aleatoria. Según el tipo de datos (categóricos, discretos, continuos), usamos diferentes distribuciones.

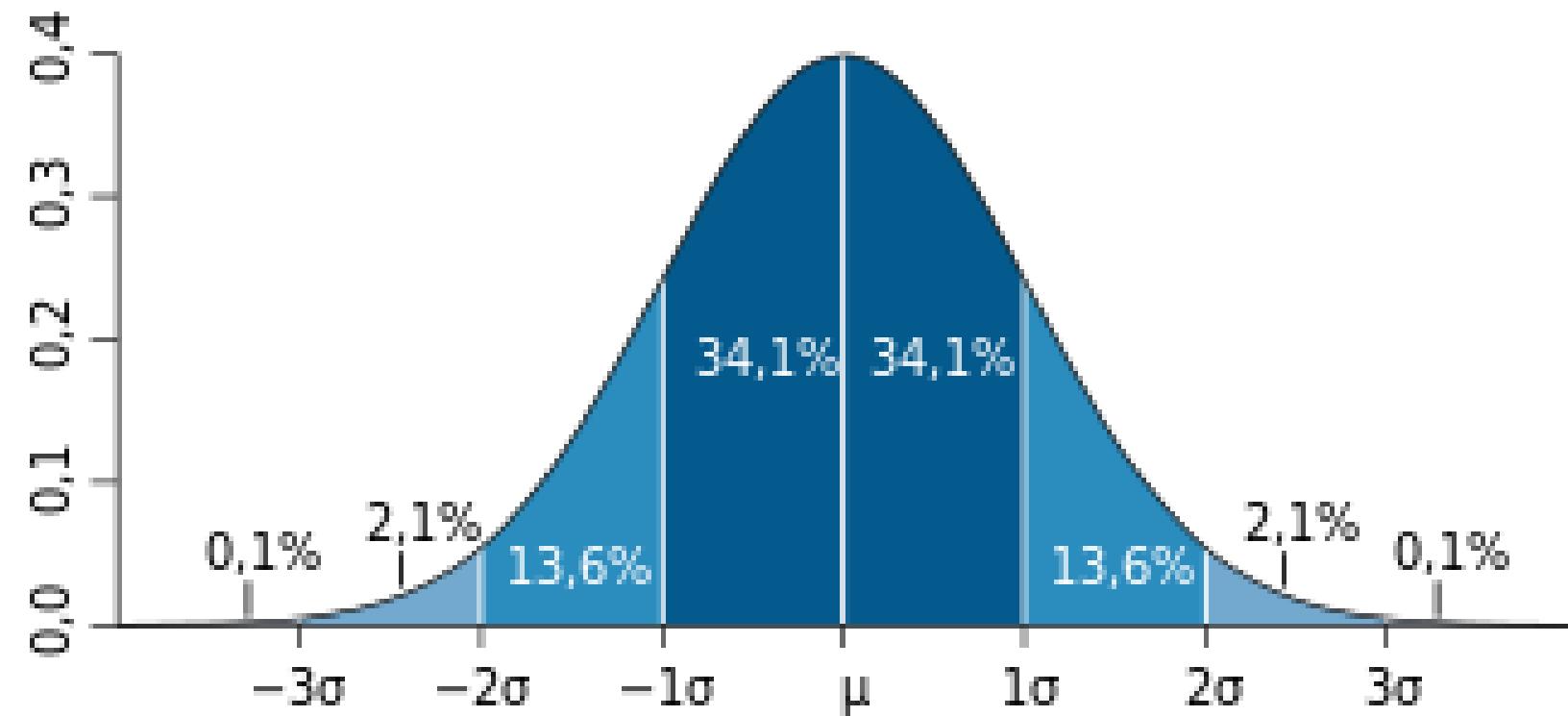
<https://www.wolfram.com/mathematica/new-in-8/parametric-probability-distributions/index.html>

Distribuciones discretas (valores contables)	¿Para qué sirve?
Binomial	Conteo de éxitos/fracaso (ej. pacientes que mejoran con un tratamiento: sí/no).
Poisson	Conteo de eventos raros por tiempo o espacio (ej. infecciones por día).

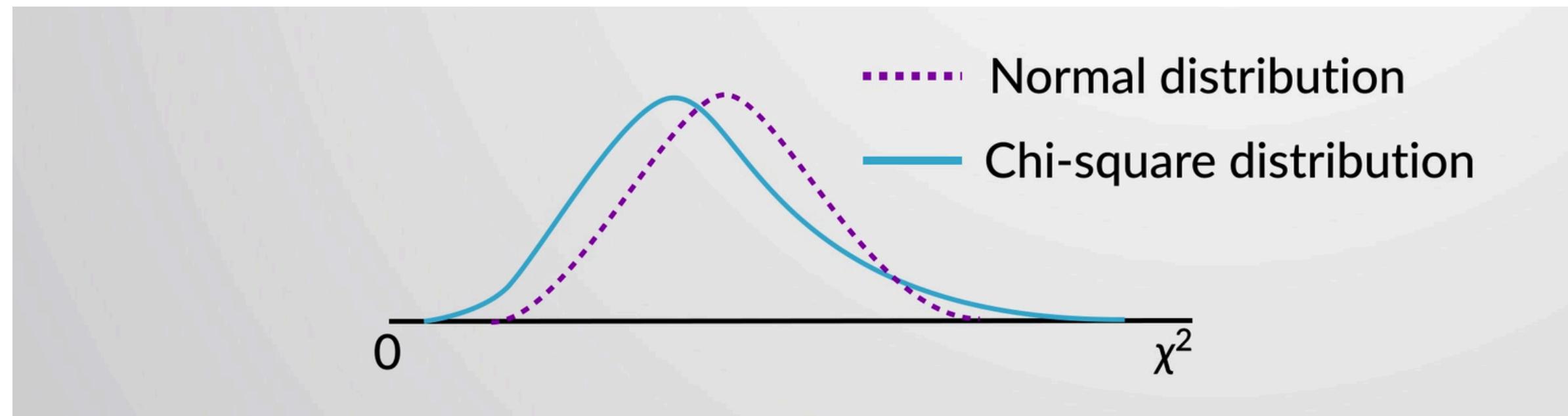
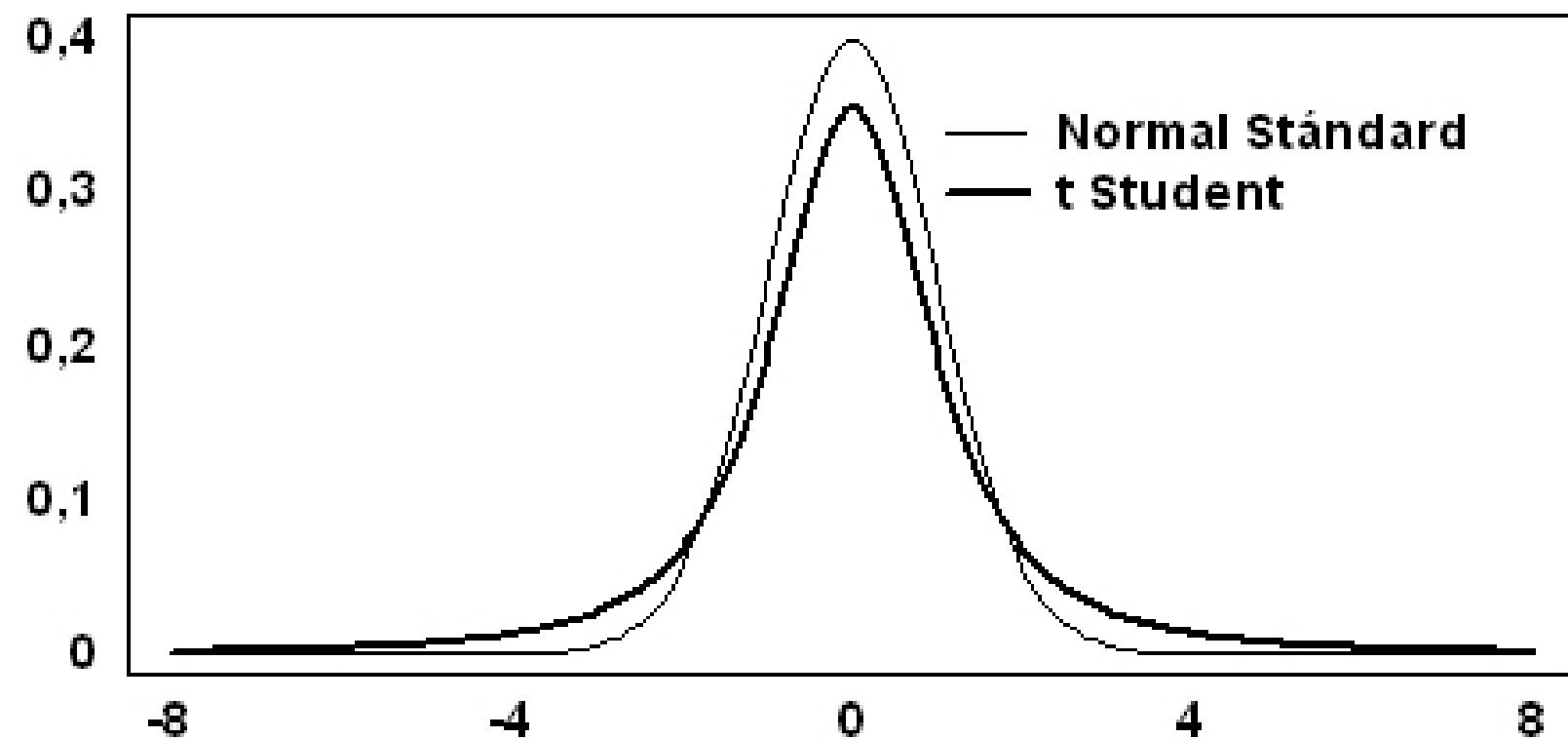
# DISTRIBUCIÓN DE PROBABILIDAD

Distribuciones continuas (valores medibles)	¿Para qué sirve?
Normal (gaussiana)	Muy común. Curva en forma de campana. La media, mediana y moda son iguales. Controla la dispersión mediante la desviación estándar. Modela datos como estatura, glucosa, colesterol (continuos).
t de Student	Parecida a la normal, pero se usa cuando el tamaño de muestra es pequeño ( $n < 30$ ). Es menos ancha que la distribución normal. Se usa cuando tenemos desconocimiento de la desviación estándar de la población.
Chi-cuadrado ( $\chi^2$ )	Modela frecuencias. Se usa en pruebas de asociación entre variables categóricas. Asimétrica. Se usa para realizar pruebas de hipótesis con datos categóricos.

# CAMPANA DE GAUSS



Es una representación gráfica de la distribución normal. Los datos están simétricamente distribuidos alrededor de la media. Cuanto más te alejas de la media, menos frecuencia tienen los valores.



# DISTRIBUCIÓN DE PROBABILIDAD

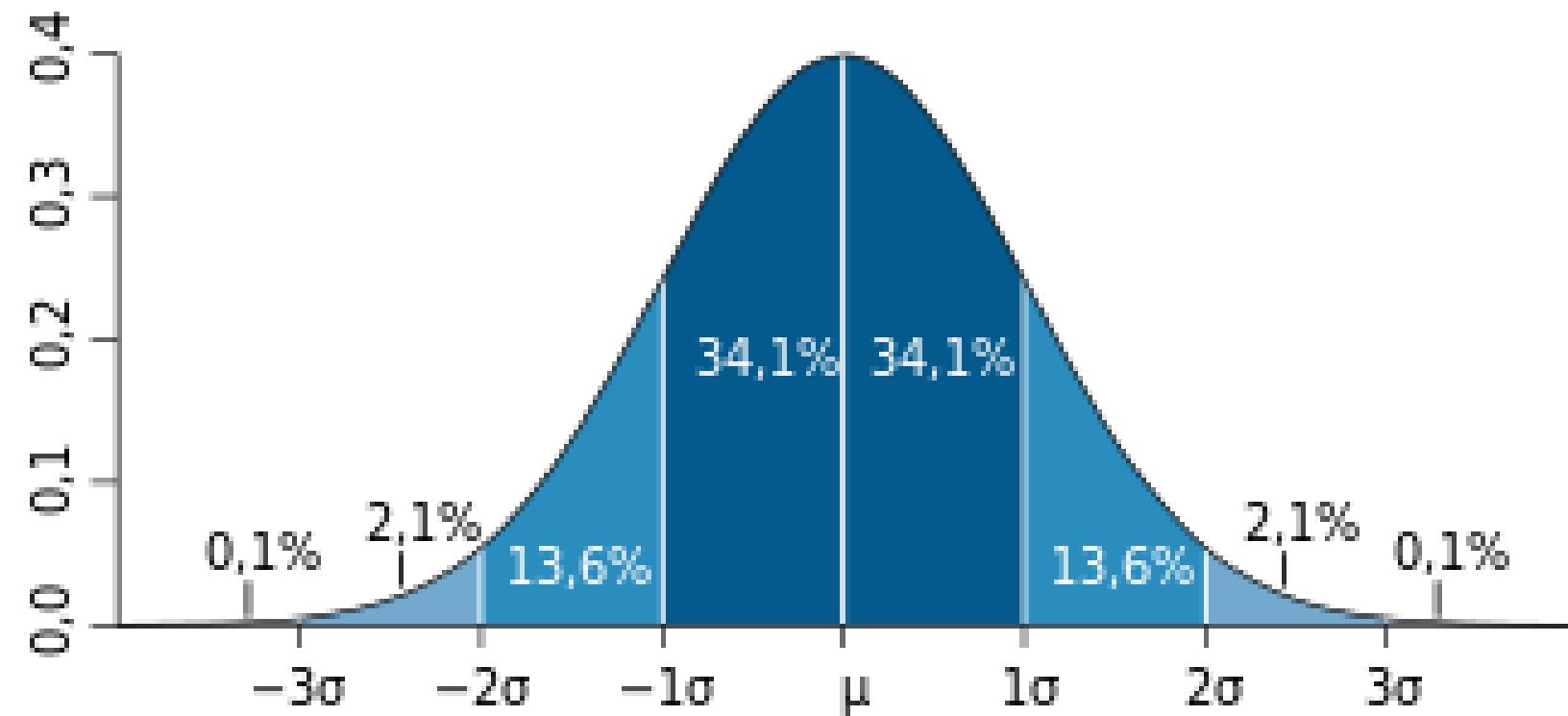
Análisis	Distribución usada	Por qué
Comparar glucosa (numérica) entre sexos (2 grupos)	t de Student	Compara medias
Ver si diabetes está relacionada con tipo de sangre (ambos categóricos)	Chi-cuadrado	Ver asociación

# TEOREMA DEL LÍMITE CENTRAL

Cuando se toman muchas muestras aleatorias de una población (del mismo tamaño), la distribución de las medias muestrales tiende a ser una distribución normal, sin importar cómo sea la distribución original de la población, siempre que el tamaño de la muestra sea suficientemente grande (usualmente  $n \geq 30$ ).

<https://www.geogebra.org/m/KFKpqe4c>

# CAMPANA DE GAUSS



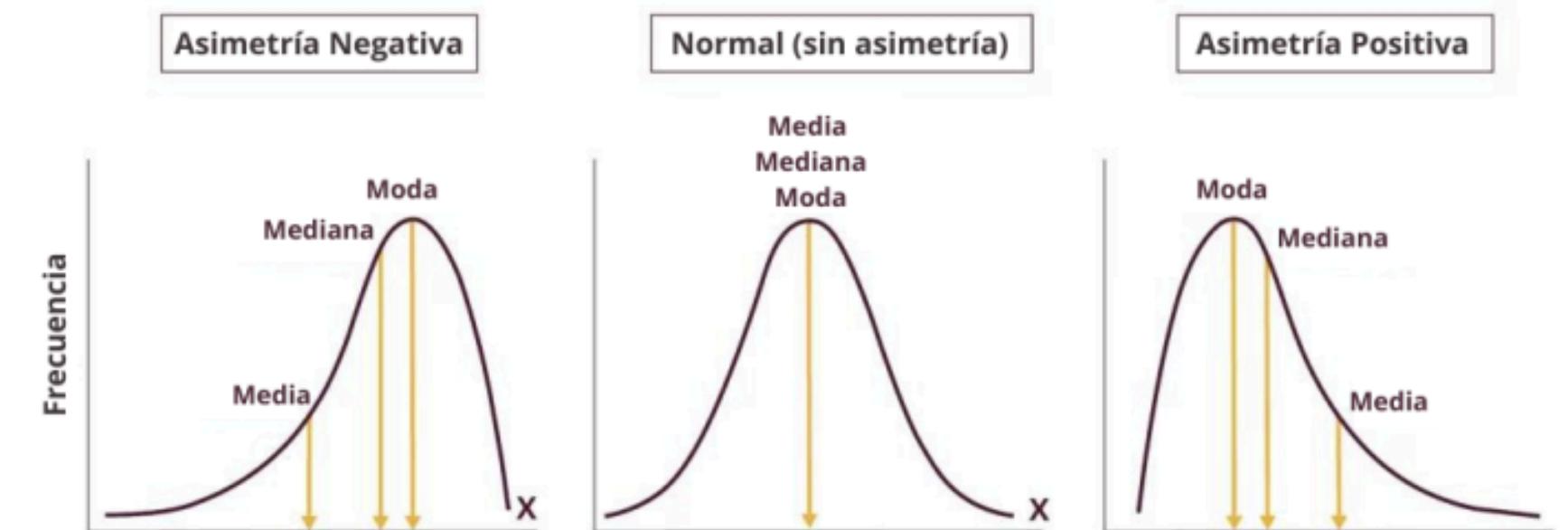
Es una representación gráfica de la distribución normal. Los datos están simétricamente distribuidos alrededor de la media. Cuanto más te alejas de la media, menos frecuencia tienen los valores.

# MEDIDAS DE FORMA

## ASIMETRÍA

Grado de distribución de una distribución de datos.

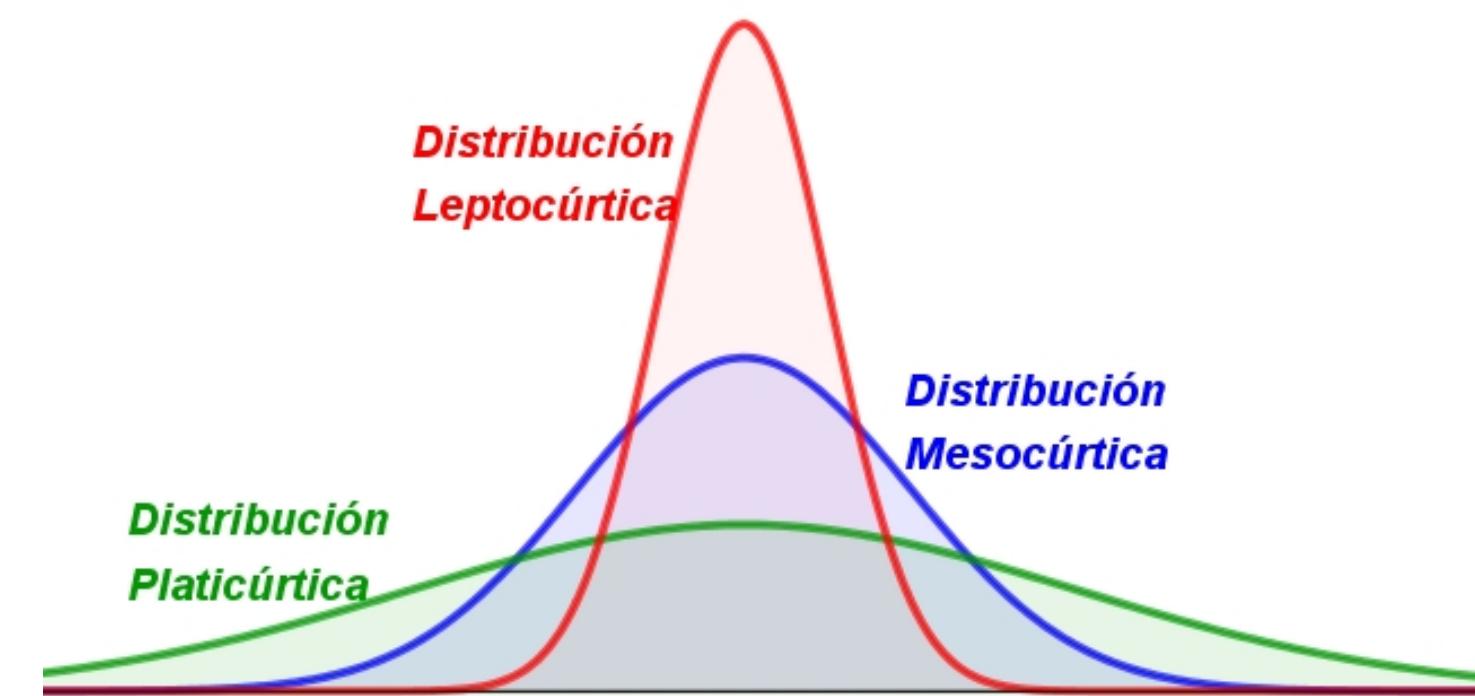
- Asimetría = 0 → Distribución simétrica (como la campana de Gauss). También se conoce como: simetría.
- Asimetría > 0 → Sesgo a la derecha (cola más larga hacia la derecha). También se conoce como: asimetría positiva.
- Asimetría < 0 → Sesgo a la izquierda (cola más larga hacia la izquierda). También se conoce como: asimetría negativa.



# MEDIDAS DE FORMA

## CURTOSIS

Permite evaluar el grado de concentración de una distribución de datos en torno a su media. Determina si una distribución es más puntiaguda o aplanada.



- Curtosis = 3 → Distribución normal (mesocúrtica).
- Curtosis > 3 → Más alta y estrecha (leptocúrtica).
- Curtosis < 3 → Más plana (plasticúrtica).

# PRUEBAS DE NORMALIDAD

Son pruebas que ayudan a determinar si los datos de una muestra se ajustan bien a una distribución normal:

- $p > 0.05 \rightarrow$  No se rechaza la normalidad (los datos podrían ser normales).
- $p \leq 0.05 \rightarrow$  Se rechaza la normalidad (los datos no parecen normales).

Prueba	¿Para qué sirve?	Características	Función en R
Shapiro-Wilk	Prueba para datos ordenados; muy sensible y potente. Muy usada para muestras pequeñas o medianas ( $n < 50$ , pero también hasta 5000)	Alta potencia. Muy recomendable.	shapiro.test(x)
Kolmogorov-Smirnov (K-S)	Prueba no paramétrica (No asume que los datos ya siguen una distribución específica), basada en la función de distribución empírica (EDF). Compara la distribución empírica con una teórica (como la normal). Muy usada para muestras grandes.	Menos sensible con muestras pequeñas.	ks.test(x, "pnorm", mean=media, sd=desviación)
Anderson-Darling	Una versión mejorada del K-S	Pone más peso en las colas de la distribución.	library(nortest) ad.test(x)

# ESTANDARIZACIÓN DE DATOS

La estandarización de datos es una técnica usada en estadística para transformar los datos y que puedan compararse entre sí o ajustarse a ciertos supuestos, como en el caso de muchas pruebas estadísticas que requieren datos con distribución normal estándar.

Estandarizar un dato significa convertirlo a una escala común, donde:

La media de los datos se convierte en 0

La desviación estándar se convierte en 1

Para cada dato  $x$ , el valor estandarizado  $z$  se calcula así:

$$z = \frac{x - \bar{x}}{s}$$

Donde:

- $x$  = valor original
- $\bar{x}$  = media de todos los valores
- $s$  = desviación estándar

Un valor  $z = 0$  está justo en la media.

Un valor  $z = 1$  está una desviación estándar por encima de la media.  
Un valor  $z = -2$  está dos desviaciones estándar por debajo de la media.

# HOMOGENEIDAD DE VARIANZAS (HOMOCEDASTICIDAD)

Es un supuesto estadístico que significa que los distintos grupos que estamos comparando tienen varianzas (dispersión de los datos) similares o iguales.

Muchas pruebas estadísticas —como el ANOVA o la prueba t de Student— asumen que las varianzas de los grupos son iguales. Si esta condición no se cumple, los resultados pueden ser inexactos  $p>0.05 \rightarrow$  Homogeneidad

Nombre de la prueba	Función en R	Para qué sirve
Prueba de Levene	leveneTest() (del paquete car)	Evalúa si las varianzas entre grupos son iguales
Prueba de Bartlett	bartlett.test()	También evalúa igualdad de varianzas (más sensible a la normalidad)
Prueba de Fligner-Killeen	fligner.test()	Alternativa robusta incluso si los datos no son normales

Prueba	¿Requiere normalidad?	Robustez	Cuándo usar
Bartlett	Sí	Baja	Si los datos son claramente normales
Levene	No estrictamente	Media	Cuando sospechas que hay no normalidad
Fligner-Killeen	No	Alta	Cuando los datos no son normales

Característica	Pruebas Paramétricas	Pruebas No Paramétricas
Definición	Pruebas estadísticas que suponen que los datos siguen una distribución específica (normal) y que cumplen ciertos supuestos.	Pruebas estadísticas que no requieren supuestos estrictos sobre la distribución de los datos.
Supuestos principales	- Normalidad- Homogeneidad de varianzas	- No requieren normalidad- Pueden usarse con datos ordinales o rangos
Tipo de datos	Cuantitativos continuos (normalmente distribuidos)	Cuantitativos no normales u ordinales
Ejemplos	- t de Student- ANOVA- Regresión lineal	- U de Mann-Whitney- Kruskal-Wallis- Test de Wilcoxon
Ventajas	- Mayor potencia estadística si se cumplen los supuestos	- Mayor robustez ante datos atípicos o distribuciones no normales
Desventajas	- Resultados pueden ser inválidos si los supuestos no se cumplen	- Menor potencia en comparación con pruebas paramétricas cuando se cumplen los supuestos
Uso recomendado	Cuando se cumplen los supuestos de normalidad y homogeneidad de varianzas	Cuando los datos son ordinales, no normales o hay pequeñas muestras

# TRANSFORMACIÓN DE LOS DATOS

La transformación de datos es un proceso matemático que consiste en aplicar una función (como logaritmo, raíz cuadrada, inverso, etc.) a cada valor de un conjunto de datos, con el fin de modificar su distribución y cumplir con los supuestos estadísticos, como la normalidad o la homogeneidad de varianzas.

- Corregir asimetría (skewness).
- Reducir el efecto de valores extremos (outliers).
- Mejorar la homogeneidad de varianzas entre grupos.
- Permitir el uso de pruebas estadísticas paramétricas, que requieren normalidad.

Transformación	Se usa cuando...	Ejemplo en R
Logarítmica (log)	Datos con distribución sesgada a la derecha. Solo funciona con valores positivos (mayores que cero).	$\log(x)$
Raíz cuadrada	Datos con varianzas crecientes o conteos. Solo se aplica a valores $\geq 0$ (no sirve con negativos).	$\text{sqrt}(x)$
Inversa	Distribuciones fuertemente sesgadas. No se puede usar con valores iguales a 0.	$1/x$

# TRANSFORMACIÓN DE LOS DATOS

$$y = \frac{1}{x}$$

Transformación Inversa:  
Invierte los valores: los grandes se vuelven pequeños y viceversa.

Aumenta la simetría de los datos con valores muy altos.

Ejemplos:  $1/1=1$   
 $1/2=0.5$   
 $1/10=0.1$

$$y = \frac{1}{x}$$

# INTERVALOS DE CONFIANZA (IC)

Es un método estadístico que permite estimar un parámetro desconocido de una población utilizando una muestra. En lugar de ofrecer un solo valor como estimación, los intervalos de confianza presentan un rango de posibles valores en el que, con cierto nivel de certeza, se espera que se encuentre el valor real del parámetro.

Ejemplo:

Si la media de glucosa en una muestra es 100 mg/dL, un IC del 95% podría ser:  
(95, 105)

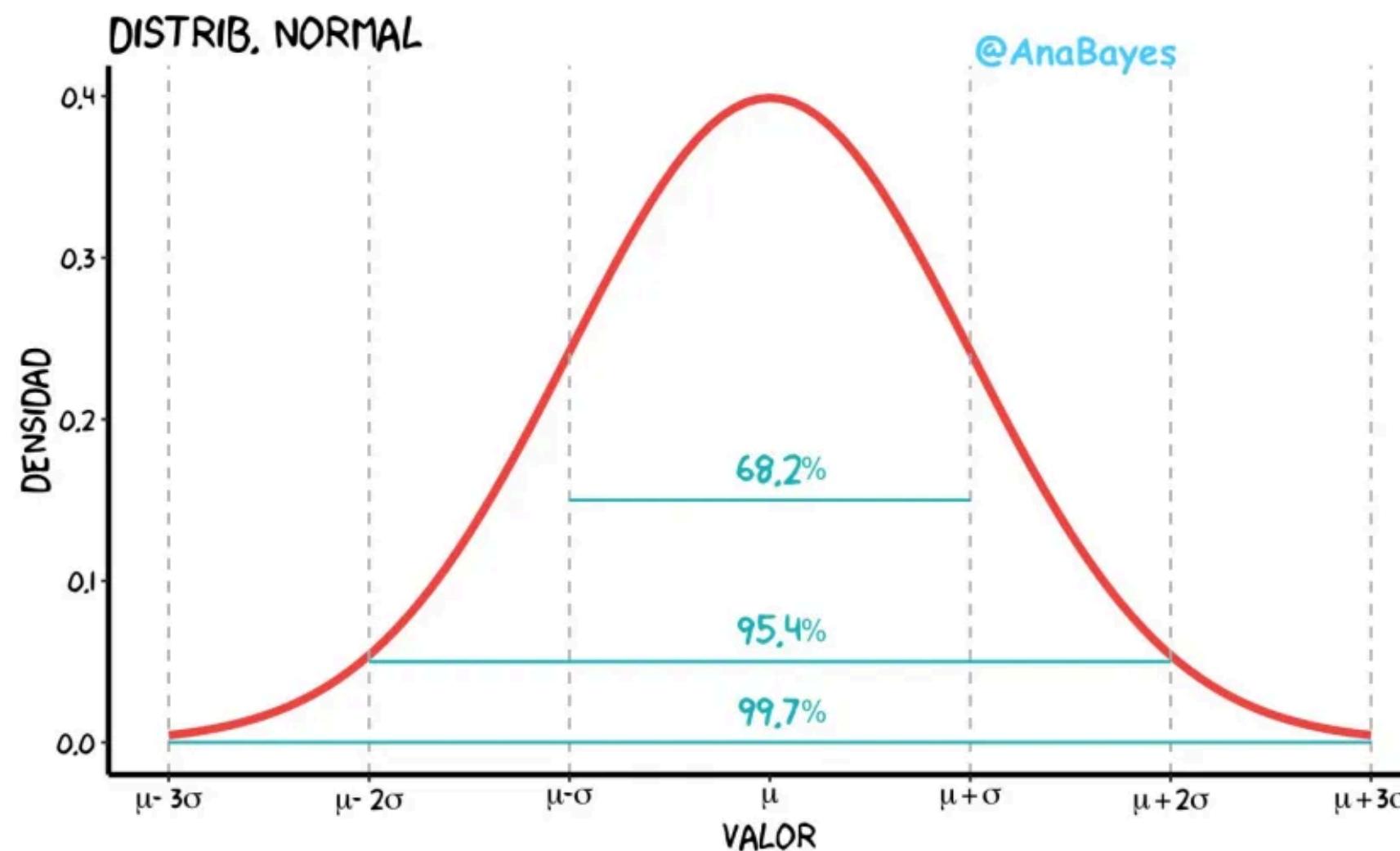
Ejemplo: <https://anabelforte.com/2020/05/16/confianza-para-recuperar-una-flor/>

# PARA QUE SIRVEN LOS IC?

Los intervalos de confianza resultan fundamentales en la interpretación de resultados y en el proceso de toma de decisiones. Proporcionan una medida de cuán precisa es una estimación, permitiendo comparar condiciones o grupos dentro de una investigación. Asimismo, ofrecen una idea clara del grado de incertidumbre presente y facilitan una comunicación más detallada y confiable de los hallazgos obtenidos.

# CALCULO DE IC

Nivel de confianza: Se refiere a la probabilidad de que el intervalo de confianza capture el verdadero valor del parámetro. Es habitual emplear un nivel de confianza del 95 %, lo que significa que, si repitiéramos el estudio muchas veces, aproximadamente el 95 % de los intervalos obtenidos incluirían ese valor verdadero.



68%: El intervalo es más estrecho, corresponde a un menor riesgo de error.

95% Buen equilibrio: te da alta confianza sin hacer el intervalo demasiado amplio.

99%: Tendrías más confianza, pero el intervalo sería mucho más amplio (menos preciso).

# CALCULO DE IC

Estimación puntual: Primero calculamos un valor específico, como la media o la proporción de una muestra, para obtener una aproximación del valor del parámetro en la población. A partir de esta estimación, y utilizando la distribución de probabilidad adecuada (por ejemplo, la t de Student), se determina la amplitud del intervalo de confianza.

$$CI = \bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

Mean value      Lower/Upper limit      z-value for the confidence level      Standard deviation      Sample size

Calculo del Intervalo de confianza  
<https://datatab.es/tutorial/confidence-interval>

Intervalos de confianza  
en R: <https://fhernanb.github.io/Manual-de-R/ic.html>

# HIPÓTESIS

Una hipótesis es una suposición o afirmación que se hace sobre un parámetro desconocido de una población (como la media) o de un fenómeno. Se pone a prueba mediante técnicas estadísticas como pruebas de hipótesis utilizando datos muestrales.

Hipótesis nula ( $H_0$ ): Es una afirmación que se asume como cierta inicialmente y que representa la ausencia de efecto, cambio o diferencia.

Hipótesis alternativa ( $H_1$  o  $H_a$ ): También llamada hipótesis afirmativa, es la contraparte de la hipótesis nula. Representa lo que queremos probar o detectar

# HIPÓTESIS

Ejemplo: Queremos saber si una nueva medicina reduce la presión arterial.

Hipótesis nula ( $H_0$ ): "La nueva medicina no tiene efecto diferente".

Hipótesis alternativa ( $H_1$ ): "La nueva medicina sí reduce más la presión arterial".

Posteriormente, recolectamos datos y realizamos una prueba estadística para decidir si rechazamos  $H_0$  o no.

# SIGNIFICANCIA ESTADÍSTICA

Probabilidad de que el resultado observado haya ocurrido por azar, bajo la suposición de que la hipótesis nula es verdadera.

Cuando un resultado es estadísticamente significativo, significa que hay evidencia suficiente para rechazar la hipótesis nula, y por tanto, es probable que haya un efecto real o una diferencia verdadera.



# VALOR P (P-VALUE)

Es una medida estadística que nos permite evaluar la evidencia contra la hipótesis nula ( $H_0$ ) en una prueba de hipótesis. Es la probabilidad de que ocurra un evento de forma al azar, tomando una premisa como cierta. Se calcula con una prueba estadística.

- Un valor p pequeño indica que sería poco probable obtener los datos observados si la hipótesis nula fuera cierta → lo que nos lleva a cuestionar o rechazar la hipótesis nula. Significativo.
- Un valor p grande sugiere que los datos son compatibles con la hipótesis nula, por lo tanto, no hay evidencia suficiente para rechazarla. No significativo.

# NIVEL DE SIGNIFICANCIA

También llamado nivel de significación o alfa, representado como  $\alpha$  es un valor que se utiliza en estadística inferencial para decidir si se rechaza o no la hipótesis nula ( $H_0$ ) en una prueba de hipótesis. Es la probabilidad máxima de cometer un error de tipo I, es decir, rechazar la hipótesis nula cuando en realidad es verdadera.

Valores comunes de $\alpha$ :	¿Cómo se usa?
<ul style="list-style-type: none"><li>• 0.05 (5%): el más común. Aceptamos hasta un 5% de probabilidad de equivocarnos al rechazar <math>H_0</math>.</li><li>• 0.01 (1%): más estricto.</li><li>• 0.10 (10%): más flexible, menos exigente.</li></ul>	<p>Se compara con el valor <math>p</math> de la prueba:</p> <ul style="list-style-type: none"><li>• Si <math>p \leq \alpha \rightarrow</math> se rechaza <math>H_0</math>. Significativo</li><li>• Si <math>p &gt; \alpha \rightarrow</math> no se rechaza <math>H_0</math>, No significativo</li></ul>

# NIVEL DE SIGNIFICANCIA

Ejemplo:

Una empresa farmacéutica quiere saber si un nuevo fármaco reduce la presión arterial más que el medicamento actual.

Hipótesis:

- $H_0$  (hipótesis nula): el nuevo medicamento NO es mejor que el actual.
- $H_1$  (hipótesis alternativa): el nuevo medicamento SÍ es mejor.

Se toma una muestra de 30 pacientes.

Se hace una prueba estadística (por ejemplo, una prueba t).

El resultado da un valor  $p = 0.03$ .

El investigador decidió antes usar un nivel de significancia  $\alpha = 0.05$

Comparamos el valor  $p$  con  $\alpha$ :

- $p = 0.03$
- $\alpha = 0.05$

Como  $p < \alpha$ , rechazamos la hipótesis nula  $H_0$ .

Con un nivel de significancia del 5%, hay evidencia suficiente para decir que el nuevo medicamento sí es mejor que el actual.

Nota: Al rechazar  $H_0$ , estás aceptando un 5% de probabilidad de estar cometiendo un error tipo I, es decir, de estar equivocado y que en realidad no haya diferencia, pero igual dijiste que sí la hay.

# ERROR TIPO I Y II

$H_0$	Aceptamos	Rechazamos
Verdadera	✓ <i>Decisión correcta</i>	<i>Error tipo I</i> $\alpha$
Falsa	<i>Error tipo II</i> $\beta$	✓ <i>Decisión correcta</i>

Error de Tipo I ( $\alpha$ ):

Es el error de rechazar la hipótesis nula ( $H_0$ ) cuando en realidad es verdadera. También se conoce como falso positivo.

Este error se controla con el nivel de significancia.

# ERROR TIPO I Y II

$H_0$	Aceptamos	Rechazamos
Verdadera	✓ <i>Decisión correcta</i>	<i>Error tipo I</i> $\alpha$
Falsa	<i>Error tipo II</i> $\beta$	✓ <i>Decisión correcta</i>

Ejemplo Error de Tipo I ( $\alpha$ ):

Un médico concluye que un tratamiento nuevo es efectivo (rechaza  $H_0$ ), cuando en realidad no lo es.

$H_0$ : El medicamento nuevo NO es efectivo.

# ERROR TIPO I Y II

$H_0$	Aceptamos	Rechazamos
Verdadera	✓ <i>Decisión correcta</i>	<i>Error tipo I</i> $\alpha$
Falsa	<i>Error tipo II</i> $\beta$	✓ <i>Decisión correcta</i>

Error de Tipo II ( $\beta$ ):

Es el error de no rechazar la hipótesis nula ( $H_0$ ) cuando en realidad es falsa. También se conoce como falso negativo.

# ERROR TIPO I Y II

$H_0$	Aceptamos	Rechazamos
Verdadera	✓ <i>Decisión correcta</i>	<i>Error tipo I</i> $\alpha$
Falsa	<i>Error tipo II</i> $\beta$	✓ <i>Decisión correcta</i>

Ejemplo. Error de Tipo II ( $\beta$ ):

Un médico concluye que el tratamiento no es mas efectivo (no rechaza  $H_0$ ), cuando en realidad sí es más efectivo.

$H_0$ : El tratamiento NO es mas efectivo

# PRUEBAS ESTADISTICAS PARA CALCULAR EL VALOR P

Son métodos que nos permiten evaluar si los resultados observados en un conjunto de datos son estadísticamente significativos, es decir, si es probable que hayan ocurrido por azar o no.

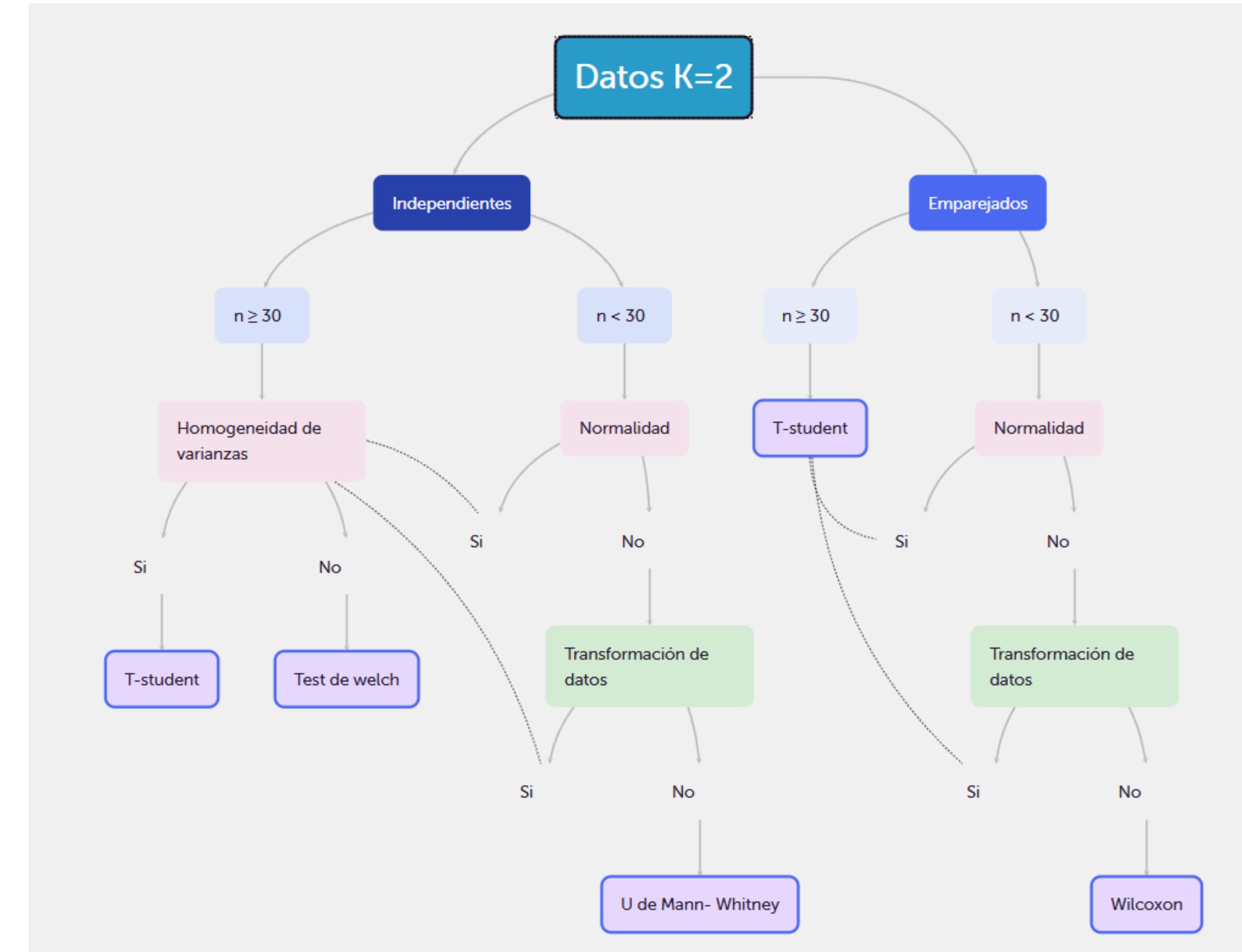
Estas pruebas comparan los datos observados con lo que esperaríamos bajo la hipótesis nula, y nos devuelven un valor p que nos ayuda a tomar una decisión.





Prueba	¿Qué te dice?
Pearson	Si hay una relación lineal significativa entre dos variables.
Spearman	Si hay una relación creciente o decreciente, no necesariamente lineal.
Regresión lineal	Si una variable predice a la otra, y cuánto la afecta (coeficientes).

Prueba	¿Qué es?	¿Para qué se usa?	Código en R
T de Student	Prueba paramétrica que compara medias entre dos grupos independientes, asumiendo igualdad de varianzas y distribución normal.	Determinar si hay una diferencia significativa entre las medias de dos grupos independientes.	<code>t.test(grupo1, grupo2, var.equal = TRUE)</code>
Test de Welch	Variante de la prueba t de Student que no asume igualdad de varianzas. También requiere distribución normal.	Comparar medias de dos grupos independientes cuando las varianzas son distintas.	<code>t.test(grupo1, grupo2, var.equal = FALSE)</code> (por defecto)
U de Mann-Whitney	Prueba no paramétrica para comparar dos grupos independientes. No requiere normalidad ni igualdad de varianzas.	Comparar si las distribuciones de dos grupos independientes son diferentes (útil con datos ordinales o no normales).	<code>wilcox.test(grupo1, grupo2, exact = FALSE)</code> (para independientes)
Wilcoxon (signed-rank)	Prueba no paramétrica para comparar dos grupos emparejados o relacionados.	Determinar si hay diferencias entre dos condiciones medidas en los mismos sujetos (pre-post, por ejemplo).	<code>wilcox.test(pre, post, paired = TRUE)</code>



Prueba	¿Qué es?	¿Para qué se usa?	Código en R
ANOVA-1 vía	Prueba paramétrica que compara medias entre más de dos grupos independientes.	Determinar si hay diferencias significativas entre las medias de tres o más grupos independientes.	<code>aov(variable ~ grupo, data = datos) summary(modelo)</code>
Kruskal-Wallis	Prueba no paramétrica alternativa al ANOVA-1. Compara rangos entre más de dos grupos.	Comparar distribuciones de tres o más grupos independientes cuando no se cumple normalidad o varianzas.	<code>kruskal.test(variable ~ grupo, data = datos)</code>
ANOVA de medidas repetidas	Prueba paramétrica para comparar más de dos condiciones medidas en los mismos sujetos.	Evaluar si hay diferencias entre tres o más mediciones repetidas sobre el mismo grupo.	<code>aov(variable ~ tiempo + Error(sujeto/tiempo), data = datos)</code>
Test de Friedman	Prueba no paramétrica alternativa al ANOVA de medidas repetidas.	Comparar rangos de tres o más mediciones emparejadas (no normales).	<code>`friedman.test(variable ~ tiempo</code>
ANOVA factorial	Prueba paramétrica que evalúa el efecto de más de una variable independiente y su interacción.	Determinar el efecto individual y combinado de varios factores sobre una variable continua.	<code>aov(variable ~ factor1 * factor2, data = datos)</code>
Otros modelos	Modelos más complejos (lineales mixtos, MANOVA, etc.) para datos con múltiples factores o estructuras.	Analizar datos con medidas repetidas y múltiples factores simultáneamente.	<code>`lme(variable ~ factor, random = ~1</code>

## Pruebas de hipótesis

Continuos +  
Categoricos

Datos  $K > 2$

1 variable de agrupación

> 1 variable de agrupación

Independientes

Emparejados

Independientes

Medidas repetidas

P

NP

P

NP

P y NP

P y NP

ANOVA-1

Kruskal-Wallis

ANOVA medidas  
repetidas

Test de Friedman

ANOVA Factorial

Otros modelos



Prueba	¿Qué es?	¿Para qué sirve?	Código en R
Chi-cuadrado ( $\chi^2$ )	Prueba estadística que compara las frecuencias observadas con las esperadas bajo la hipótesis nula.	Determinar si hay asociación entre dos variables categóricas en tablas de contingencia.	<code>chisq.test(tabla)</code>
Prueba exacta de Fisher	Alternativa a la prueba chi-cuadrado cuando los tamaños de muestra son pequeños.	Evaluar la asociación entre dos variables categóricas (ideal para tablas 2x2).	<code>fisher.test(tabla)</code>
Prueba de McNemar	Prueba no paramétrica para datos pareados categóricos (mismas personas antes/después).	Evaluar cambios en respuestas dicotómicas en dos momentos (pre/post o emparejados).	<code>mcnemar.test(tabla)</code>

# INTRODUCCIÓN A R PARA BIOESTADÍSTICA

MODULO 2



# LENGUAJE DE PROGRAMACIÓN R



- Lenguaje de programación para análisis estadístico y visualización.
- Fue creado por Ross Ihaka y Robert Gentleman en 1996.
- Popular en bioestadística, bioinformática y ciencia de datos.
- Es software libre y gratuito
- Línea de comandos
- Se puede utilizar sin RStudio

# INSTALACIÓN DE R

Descarga desde <https://cran.r-project.org/>



[CRAN  
Mirrors](#)  
[What's new?](#)  
[Search](#)  
[CRAN Team](#)

[About R](#)  
[R Homepage](#)  
[The R Journal](#)

[Software](#)  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Task Views](#)  
[Other](#)

[Documentation](#)  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

[Donations](#)  
[Donate](#)

The Comprehensive R Archive Network

**Download and Install R**

Precompiled binary distributions of the base system and contributed packages. Windows and Mac users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

**Source Code for all Platforms**

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2025-04-11, How About a Twenty-Six) [R-4.5.0.tar.gz](#), read [what's new](#) in the latest version.
- The CRAN directory [src/base-prerelease](#) contains R alpha, beta, and rc releases as daily snapshots in time periods before a planned release.
- Between releases, the same directory [src/base-prerelease](#) contains snapshots of current patched and development versions. Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Alternatively, daily snapshots are [available here](#).
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#).

**Questions About R**

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

**Supporting CRAN**

- CRAN operations, most importantly hosting, checking, distributing, and archiving of R add-on packages for various platforms, crucially rely on technical, emotional, and financial support by the R community.

Please consider making [financial contributions](#) to the R Foundation for Statistical Computing.

[What are R and CRAN?](#)

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series

# R STUDIO (POSIT RSTUDIO)



- Entorno de desarrollo integrado (IDE).
- Interfaz amigable que facilita la escritura de código, la visualización de gráficos, la administración de proyectos y la depuración de scripts.
- Mejora la experiencia de programación en R, pero no es un lenguaje en sí mismo.
- Incluye una consola, un editor de scripts, un visor de gráficos y pestañas para explorar variables, archivos y paquetes.

# INSTALACIÓN DE R STUDIO

Descarga desde: <https://posit.co/download/rstudio-desktop/>



DOWNLOAD

## RStudio Desktop

Used by millions of people weekly, the RStudio integrated development environment (IDE) is a set of tools built to help you be more productive with R and Python.

Don't want to download or install anything? Get started with RStudio on [Posit Cloud for free](#). If you're a professional data scientist looking to download RStudio and also need common enterprise features, don't hesitate to [book a call with us](#).

Want to learn about core or advanced workflows in RStudio? Explore the [RStudio User Guide](#) or the [Getting Started](#) section.

### 1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

*R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website.*

### 2: Install RStudio

[DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS](#)

Size: 281.27 MB | [SHA-256: 9E6F68CA](#) | Version: 2025.05.0+496 |  
Released: 2025-05-05



[Watch video on YouTube](#)

Error 153

Video player configuration error



<https://www.youtube.com/watch?v=hbgzW3Cvda4>

# VARIABLES

Una variable permite asociar un valor u objeto a un identificador. En la memoria del ordenador habrá una zona donde se almacena el objeto asociado a la variable

Integer	Números enteros (1, 5, -5, 8, -23). Se debe asignar con la letra L.	Date	Fecha (día, mes y año)
Numeric	Números con decimales (3,2; -6,32; 4,12)	POSIX	Fecha, hora y huso horario
Logical	Datos booleanos: True o false.	NA	Valor no disponible
Character	Texto	NULL	Elemento vacío
Factor	Información categórica	Inf	Infinito

# TIPOS DE VARIABLES

Variables cualitativas (categóricas) : Estas no se expresan con números. Representan cualidades, atributos o categorías.

Nominales  
No tienen un orden lógico.  
Solo identifican o clasifican.

Ejemplo: Género:  
femenino, masculino,  
no binario.

Ordinales  
Sí tienen un orden lógico, pero sin una distancia numérica clara entre categorías.

Ejemplo: Puesto en una competencia: 1.<sup>o</sup>, 2.<sup>o</sup>, 3.<sup>o</sup>.

Variables cuantitativas (numéricas): Son aquellas que se expresan con números y permiten hacer operaciones matemáticas.

Discretas  
Toman valores enteros y contables.  
No tienen decimales.

Ejemplo: Número de hijos.

Continuas  
Toman valores infinitos dentro de un intervalo, incluyen decimales.

Se obtienen mediante medición.

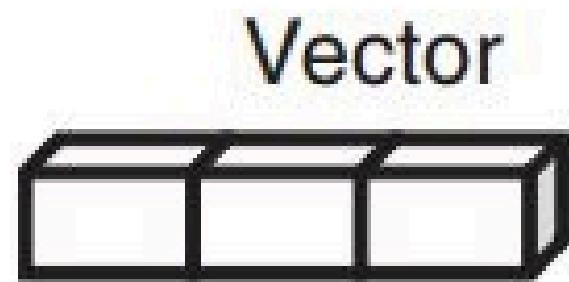
Ejemplo: Peso (62.3 kg).

# OPERADORES

Un operador es un símbolo o conjunto de símbolos que indica a R que debe realizar una operación específica sobre uno o más valores (llamados operandos)

Tipo de Operador	Operador(es)	Descripción	Ejemplo
Aritméticos	+, -, *, /	Suma, resta, multiplicación, división	$2 + 3 \rightarrow 5$
	$\wedge$	Potencia	$2^3 \rightarrow 8$
	$\%$	Resto de dividir un número por otro	$10 \% 3 \rightarrow 1$
	$\%/%$	División entera	$10 \%/% 3 \rightarrow 3$
Comparativos	$==, !=$	Igualdad, diferente	$5 == 5 \rightarrow \text{TRUE}$
	$<, >, <=, >=$	Comparaciones	$4 <= 3 \rightarrow \text{FALSE}$
Lógicos	$\&$	Sentencia AND. Ambos elementos deben ser TRUE para que se obtenga TRUE	$\text{TRUE} \& \text{FALSE} \rightarrow \text{FALSE}$
	$ $	Sentencia OR. Al menos un elemento debe ser TRUE para que se obtenga TRUE	$\text{TRUE}   \text{FALSE} \rightarrow \text{TRUE}$
	!	Se invierte el dato booleano	$!\text{TRUE} \rightarrow \text{FALSE}$
Asignación	$<-$ , $->$ , $=$	Asignación de valores	$x <- 10$
Secuencia	:	Secuencia de números	$1:5 \rightarrow 1 2 3 4 5$

# ESTRUCTURAS DE DATOS EN R

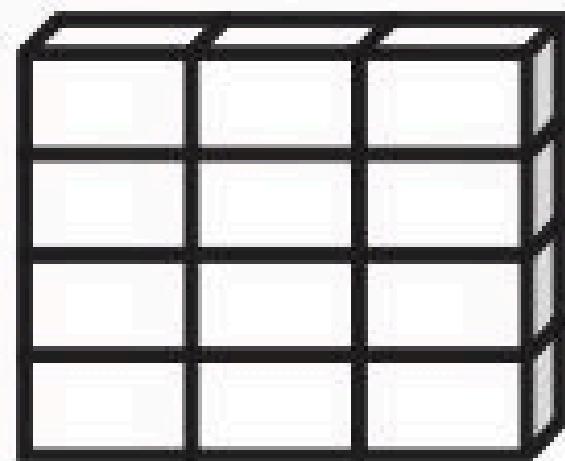


## VECTOR

Array de una única dimensión que almacena datos del mismo tipo.

# ESTRUCTURAS DE DATOS EN R

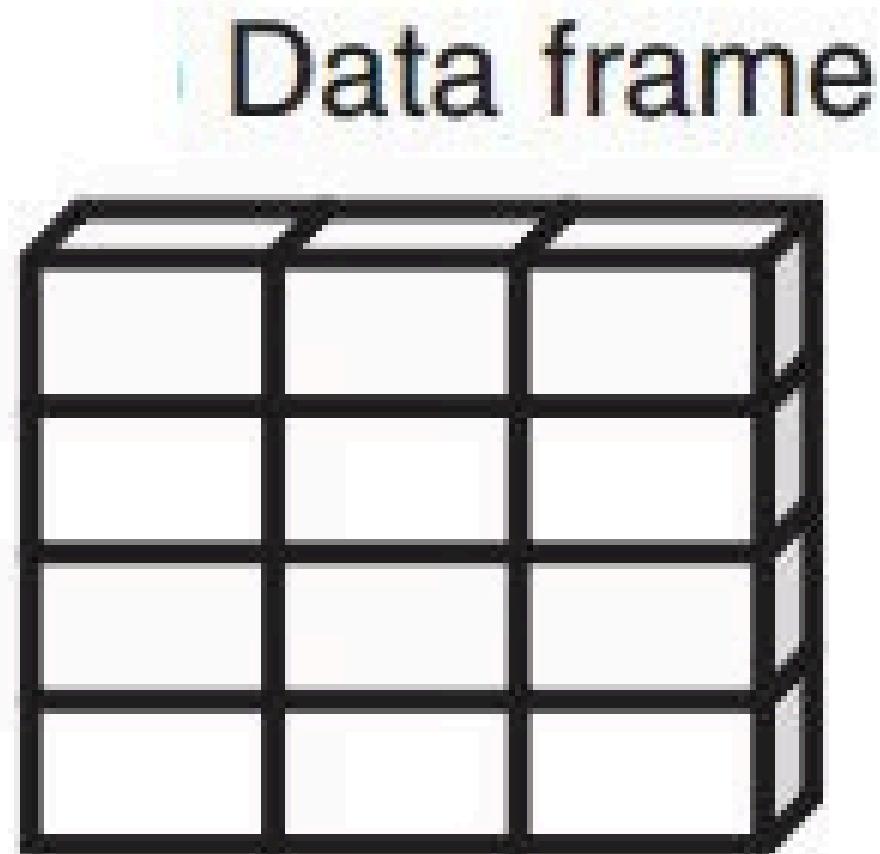
Matrix



## MATRIZ

Objeto de dos dimensiones con formato tabla (Con filas y columnas) con elementos del mismo tipo.

# ESTRUCTURAS DE DATOS EN R



## DATA FRAME

Objeto de dos dimensiones con formato tabla (Con filas y columnas) con elementos diferentes. Las filas en un data frame representan casos, individuos u observaciones, mientras que las columnas representan atributos, rasgos o variables. Las columnas deben tener la misma longitud

Material del curso:

Link: <https://github.com/Liliana223/Fundamentos-de-la-bioestadistica-inferencial/tree/main>

# RETO DE PROGRAMACIÓN!!

Material del curso:

Link: <https://github.com/Liliana223/Fundamentos-de-la-bioestadistica-inferencial/tree/main>

O en Kaggle: <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data>

# RECURSOS ADICIONALES

dplyr: <https://nyu-cdsc.github.io/learningr/assets/data-transformation.pdf>

ggplot2: Elegant Graphics for Data Analysis (3e): <https://ggplot2-book.org/>

Tidyverse packages: <https://www.tidyverse.org/packages/>

Documentación de R: <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/>

Fundamentos de Programación - BUCLES (Ciclos) EN PROGRAMACION: <https://www.youtube.com/watch?v=EbVT3XIf2TU>

Introducción a R. Bucles: <https://www.youtube.com/watch?v=w55TI908-Tw>

R for Data Science: <https://r4ds.had.co.nz/index.html>

R for Data Science Cookbook — Mark P. J. van der Loo & Yihui Xie (O'Reilly)

Iniciación a la probabilidad y la estadística. De Rosario Delgado de la Torre · 2004:

[https://www.google.com.co/books/edition/Iniciaci%C3%B3n\\_a\\_la\\_probabilidad\\_y\\_la\\_estad/qxdz9wGa5ZAC?hl=es-419&gbpv=1&dq=Shapiro%20Wilk&pg=PA143&printsec=frontcover](https://www.google.com.co/books/edition/Iniciaci%C3%B3n_a_la_probabilidad_y_la_estad/qxdz9wGa5ZAC?hl=es-419&gbpv=1&dq=Shapiro%20Wilk&pg=PA143&printsec=frontcover)

PROBABILITY STATISTICS for Engineering and the Sciences. JAY L. DEVORE:

[https://drhuang.com/science/mathematics/book/probability\\_and\\_statistics\\_for\\_engineering\\_and\\_the\\_sciences.pdf](https://drhuang.com/science/mathematics/book/probability_and_statistics_for_engineering_and_the_sciences.pdf)

Zar, Jerrold H. (2010). Biostatistical Analysis (5th ed.): <https://lib.zu.edu.pk/ebookdata/Biostatistics/Biostatistical%20Analysis-by%20Jerrold%20H%20Zar.pdf>

Field, Andy. Discovering Statistics Using R (2012):

[https://batrachos.com/sites/default/files/pictures/Books/Field\\_ea\\_2012\\_Discovering%20Statistics%20using%20R.pdf](https://batrachos.com/sites/default/files/pictures/Books/Field_ea_2012_Discovering%20Statistics%20using%20R.pdf)

Dalgaard, Peter. Introductory Statistics with R (2nd ed., 2008): <https://www.cin.ufpe.br/~maod/ESAP/R/Introductory-Statistics-With-R-2nd-Edition.pdf>

**¡MUCHAS  
GRACIAS!**