Análisis Multivariado

para datos biológicos

una introducción

| Programa del taller

INTRODUCCIÓN TEÓRICA

Definición y alcance de los métodos uni bi y multi variados.

Conceptos centrales del análisis multivariado: Metodología. Métodos de ordenación, agrupamiento y combinados.

DESARROLLO PRÁCTICO

Implementación en python de método de agrupación (K-medias) y ordenamiento (PCA).

| definiciones preliminares

Unidad de Estudio (UE)

entidades que el investigador elige analizar

Variable (V)

característica o propiedad que difiere entre las Unidades de Estudio (UE) a analizar

- V Cualitativa: categorías o etiquetas, por ejemplo, los distintos tipos de tejido (hígado, riñón, pulmón)
- V Cuantitativa: mide una magnitud numérica, por ejemplo, concentración de proteína

ANÁLISIS

UNIVARIADO

BIVARIADO

MULTIVARIADO

ANÁLISIS

UNIVARIADO

| univariado: describir el comportamiento de una variable

Forma más simple de análisis estadístico

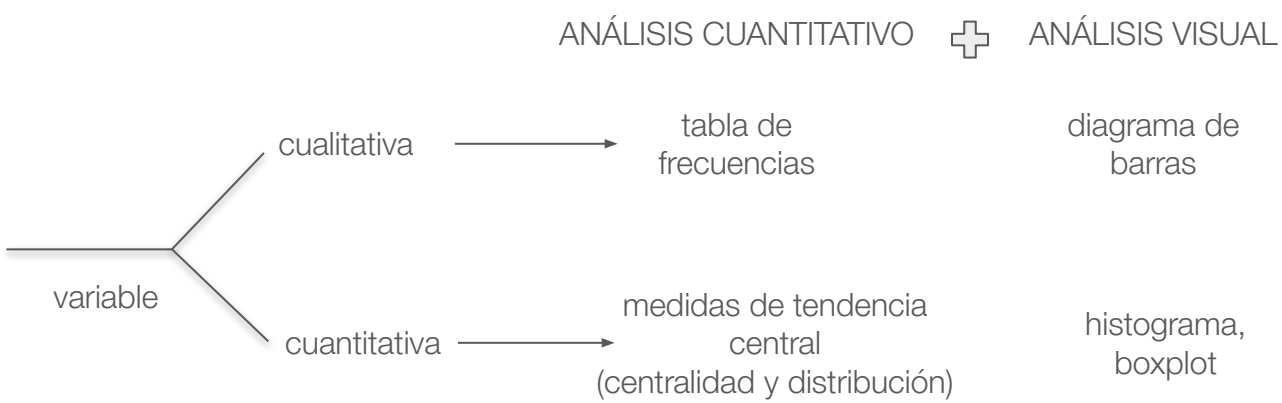
Análisis de una sola variable en un conjunto de datos sin tener en cuenta interacciones entre múltiples variables.

Comparaciones de subgrupos a través de una única variable.

Sea cual sea el número de medidas registradas en la investigación, se limita a **explorar cada una de las variables de manera independiente**.

METODOLOGÍA

1 Creación de un perfil descriptivo de la variable



Análisis visual y análisis cuantitativo ⇨ sustitutivos y complementarios

SUSTITUTIVOS: visualizar una distribución ya nos permite intuir cómo estarán situadas la media, la mediana, la moda o la desviación típica. Y si sabemos las medidas de centralidad y dispersión de una distribución podemos hacernos una idea de qué forma visual tendrá.

COMPLEMENTARIOS: si miramos una variable por los dos lados, la parte visual y la parte cuantitativa, podremos ver más matices

2 Pruebas de hipótesis
- t-student
- ANOVA

componente inferencial!

nos permite ir más allá de “veo que las medias difieren” y cuantificar la evidencia contra el azar

EJEMPLO

Variable: diámetro de roseta en *Arabidopsis thaliana*
medición, 21 días, el diámetro de la roseta como proxy de vigor.

1. Recolección de datos

Se toman 10 plantas al azar y, con un calibrador, se mide el diámetro (en mm) de la roseta al punto de —por ejemplo— inicio de floración temprana.
Datos obtenidos (mm):

```
55, 58, 60, 62, 57, 59, 61, 63, 56, 58
```

2. Tabla de frecuencias

Clase (mm)	f	fr	Fa	Fra
55–56	2	0,20	2	0,20
57–58	4	0,40	6	0,60
59–60	2	0,20	8	0,80
61–62	2	0,20	10	1,00
Total	10	1,00	—	—

EJEMPLO

3. Estadísticos descriptivos
Media = 59,9 mm
Mediana = 59,5 mm
Desviación estándar ≈ 2,5 mm
Rango = 63 mm – 55 mm = 8 mm

4. Visualización
Histograma: para ver si la distribución es aproximadamente normal.
Boxplot: detecta posibles outliers (nuestra tabla no sugiere extremos fuera del rango).

5. Prueba de normalidad
Shapiro–Wilk ($\alpha = 0,05$).
Supongamos $p = 0,45 \rightarrow$ no se rechaza normalidad.

Conclusiones univariadas
La variable “diámetro” está equilibradamente distribuida y cumple supuestos de normalidad.
No hay outliers a corregir ni transformaciones necesarias.
Con esto, ya podemos usar tests paramétricos (t-test, ANOVA) si quisiéramos comparar con otros tratamientos.

El análisis univariado
es crucial en las etapas iniciales
de cualquier análisis de datos

Explorar la naturaleza de la variable y su distribución:

- Para determinar tipo de análisis bivariado o multivariado a realizar
- Para identificar si se requiere una transformación de los datos, si hay valores atípicos, etc.

Antes de pasar a comparaciones o correlaciones, siempre debemos
entender cómo se comporta cada variable por separado

ANÁLISIS

BIVARIADO

| bivariado: cómo se relacionan 2 variables?

Explorar si el comportamiento de Y (respuesta)
está parcialmente determinado por X (predictora)

"¿Cómo cambia la expresión de mi gen de interés al aumentar la dosis del fármaco?"

Nos ayuda a generar hipótesis:
si dos variables se mueven juntas, podría
haber una relación causal o compartan
mecanismos biológicos.

METODOLOGÍA

BIVARIADO

1 Exploración inicial

Categórica – Categórica	Tabla de contingencia Diagrama de barras
Categórica – Numérica	Boxplot Violin plot
Numérica – Numérica	Scatter plot con línea de regresión

2 Prueba de hipótesis: testear si los datos apoyan estadísticamente la hipótesis de que las dos variables están relacionadas

Categórica – Categórica	coeficiente de asociación (Cramer's V) para fuerza χ^2/Fisher para significación
Categórica – Numérica	diferencia de medias con t-test/ANOVA + IC 95 %
Numérica – Numérica	coeficiente de correlación (Pearson o Spearman) + p-valor

3 Modelar la interacción (solo entre variables numéricas): crear una fórmula matemática que permita predecir el valor de la variable dependiente si se conoce el valor de la independiente

Numérica – Numérica	regresión lineal, R^2

EJEMPLO

Variables: área foliar vs. peso seco de hojas en Helianthus annuus (girasol) evaluar cómo el área de la hoja se traduce en producción de biomasa seca.
10 UE

Área (cm²)	Peso seco (mg)
10	50
15	75
20	95
25	120
30	145
35	165
40	190
45	210
50	240
55	260

EJEMPLO

Exploración inicial: Scatter plot (área en eje X, peso en Y) para ver tendencia general.

Medida de asociación
Coeficiente de Pearson:
r ≈ +0,99 (muy alta correlación positiva). p-valor < 0,001 (significativo a cualquier α razonable).
Modelo de regresión lineal
Parámetros (aprox.):
Pendiente
b≈4,5 mg/cm^2
Intercepto
a≈5 mg
Coeficiente de determinación
$R2$≈0,98
el 98 % de la variabilidad del peso se explica por el área.
Comprobación de supuestos
Residuos ~ N(0, σ^2) → boxplot y test de normalidad sobre residuos.
Homocedasticidad → gráfico de residuos vs. predicciones.
Interpretación biológica
La pendiente indica que, por cada cm^2 extra de área foliar, la hoja acumula en promedio ~4 mg de biomasa seca.
Un R^2 tan alto sugiere relación casi lineal en ese rango de áreas (típico en hojas jóvenes y plenamente desarrolladas).

FACTORES DE CONFUSIÓN

Variables ocultas que alteran la relación X → Y

Si, en el contexto de una investigación que tenga como objetivo poner a prueba una relación de causalidad, observamos una asociación entre una **variable independiente** –también llamada *variable predictora* o *explicativa*– y una **variable dependiente** –también conocida como *variable resultado* o *explicada*–, una tercera variable sería un factor de confusión si su incorporación al análisis comportara el incremento, el decrecimiento, la desaparición o, incluso, como hemos podido ver, la inversión de su relación.

factor de confusión

Se considera la relación entre cada una de las posibles parejas de variables pero en cada ocasión, de manera independiente. Por lo que **no es posible descartar que cualquier otra variable pueda interferir en estas relaciones actuando como un potencial factor de confusión** y, por lo tanto, alterando o incluso haciendo evidentes las relaciones entre dos variables que podrían no haber sido observadas inicialmente.

La determinación de una relación de causalidad implica la observación de una asociación entre dos variables, sin embargo **la mera evidencia de esta asociación desde el punto de vista estadístico no implica, necesariamente, la existencia de una relación causal.**

Helados vendidos ↑ y ahogamientos ↑ en verano → tercera variable! (temperatura).

En este sentido, como extensión del análisis bivariante, el análisis multivariante se presenta como el marco analítico general que se propone analizar e interpretar las relaciones entre diversas variables, pero lo hace, en este caso, mediante la construcción de modelos complejos que permiten determinar su existencia de manera simultánea. Así, más allá de la consideración de las variables dependientes e independientes, este tipo de análisis permite a los investigadores incorporar a sus estudios las **variables de control** que sean necesarias. Es decir, les permite tener en cuenta todas las variables extrañas que eventualmente podrían actuar como factores de confusión y que, por lo tanto, podrían interferir en las relaciones que son realmente objeto de interés.

la ventaja es que permite controlar confusores medidos, pero aún no puede controlar los NO no medidos

ANÁLISIS

MULTIVARIADO

ANÁLISIS MULTIVARIADO

conjunto de **métodos estadísticos** cuyo objetivo común es **estudiar simultáneamente más de una variable medida sobre un conjunto de unidades de estudio** (UE), distinguiendo la contribución de cada variable al conjunto de relaciones con el fin de **identificar patrones, agrupamientos o relaciones** en los datos que describan, expliquen o predigan los fenómenos

Modelar las múltiples relaciones existentes entre diversas variables de manera simultánea

Estimar el peso específico o la importancia relativa de cada una de ellas en sus modelos

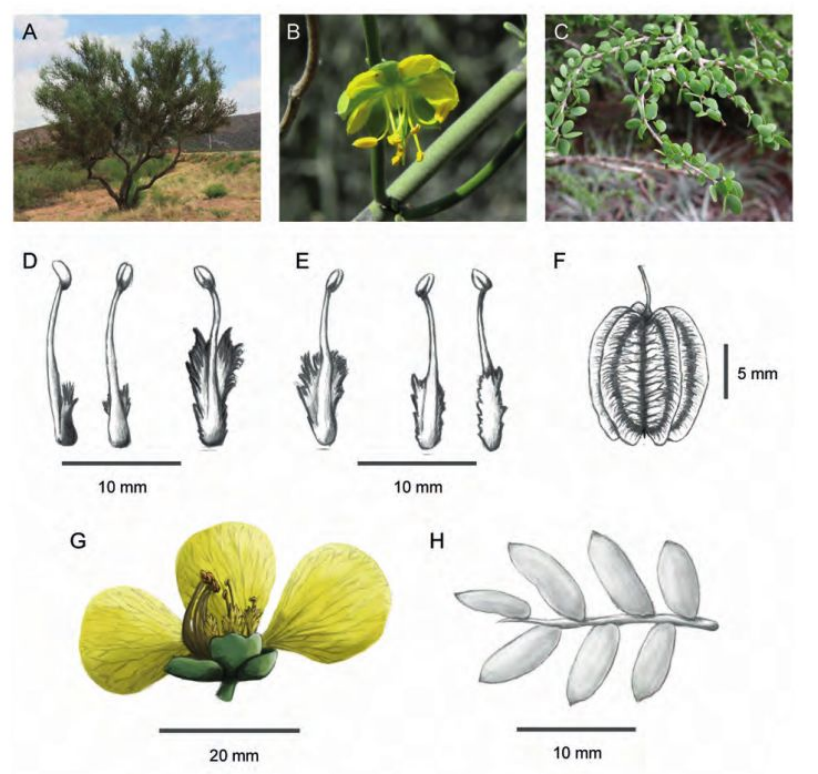
| serie de PASOS ELEMENTALES del análisis multivariado

1. Elección de Unidades de Estudio (UE)
2. Elección de Variables (V)
3. Construcción de Matriz Básica de Datos (MBD)
4. Construcción de Matriz Básica de Similitud (MS)
5. Aplicación de métodos cuantitativos
6. Detección de patrones e inferencias

1 Elección de Unidades de Estudio

Unidad de Estudio: elemento sobre el que se realiza el análisis

- individuos
- poblaciones
- especies
- localidades
- sitios paleontológicos
- secuencias génicas



8 especies del género *Bulnesia*

A) y (B) Bulnesia retama; (C) B. sarmientoi; (D) estambres de B. arborea; (E) estambres de B. carrapo; (F) fruto de B. schickendantzii; (G) flor de B. carrapo (se muestran sólo tres carpelos); (H) hojas de B. retama.

UE

1 Elección de Unidades de Estudio

cuales son las UE en nuestro ejemplo?

2 Elección de Variables

Variable: característica o propiedad que difiere entre las UE

Tipos de variables:

- Morfológicas
- Fisiológicas
- Químicas
- Ecológicas
- Geográficas
- Genéticas

VARIABLES

cualitativas no métricas

etiquetas o categorías sin sentido numérico

especies, tipos de muestras

cuantitativas métricas

- discretas
- continuas

representan un grado o cantidad mediante números

recuento de individuos, longitud de secuencias

Tipos de datos		Ejemplo	Estados
1. Cualitativos o categóricos	1.1. Nominales	Sexo cromosómico	XX, XY, XXY, X0, XYY
		Presencia-ausencia de una especie en un área	Presencia, ausencia
	1.2. Ordinales	Pubescencia de la hoja	Glabra, pelos poco abundantes, pelos muy abundantes
		Grado de disturbio	Bajo, medio, alto
2. Cuantitativos o numéricos	2.1. Continuos	Longitud del abdomen	10 mm, 10,2 mm, 20,1 mm, …
		Temperatura del cuerpo	36,2 °C, 37,1 °C, …
	2.2. Discretos	Número de inflorescencias	1, 2, 3, 10, …
		Número de tipos de aminoácidos en una proteína	10, 15, 20, …

Los datos pueden ser transformados

Los **cualitativos se transforman en valores numéricos**

Los **cuantitativos pueden dividirse en rangos y transformarse en variables binarias** si es necesario

VARIABLES

Variable	Estados	Codificación
1. Hábito	Arbustos	0
	Arbustos y árboles	1
	Árboles	2
2. Longitud del internodio (cm)	-	-
3. Diámetro del internodio (cm)	-	-
4. Longitud de la hoja (cm)	-	-
5. Ancho de la hoja (cm)	-	-
6. Longitud del pecíolo (cm)	-	-
7. Número de folíolos	-	-
8. Presencia de peciólulos	Folíolos no sésiles	0
	Folíolos sésiles o no sésiles	1
	Folíolos sésiles	2
9. Disposición de los folíolos en el raquis	Folíolos alternos	0
	Folíolos subopuestos	1
	Folíolos opuestos	2
10. Pubescencia de la hoja	Ausente	0
	Ausente y presente	1
	Presente	2
11. Longitud del folíolo (mm)	-	-
12. Ancho del folíolo (mm)	-	-
13. Número de nervaduras primarias del folíolo	-	-
14. Posición de los folíolos terminales	Paralelos	0
	Paralelos y divergentes	1
	Divergentes	2
15. Presencia de mucrón en folíolos	Folíolos no mucronados	0
	Folíolos mucronados	1
16. Tipo de inflorescencia	Flores solitarias	1
	Inflorescencia en dicasio	2
17. Longitud del pedúnculo (mm)	-	-
18. Longitud del sépalo (mm)	-	-
19. Ancho del sépalo (mm)	-	-

Variable	Estados	Codificación
20. Color de los pétalos	Blanco	1
	Amarillo	2
21. Longitud del pétalo (mm)	-	-
22. Ancho del pétalo (mm)	-	-
23. Número de nervaduras del pétalo	-	-
24. Tipo de estambres	Heterogéneos	0
	Heterogéneos y homogéneos	1
	Homogéneos	2
25. Modificación de los estambres	No modificados	0
	Modificados y no modificados	1
	Modificados	2
26. Presencia de gran escama junto al estambre	Ausente	0
	Ausente y presente	1
	Presente	2
27. Presencia de pelos en la base del filamento estaminal	Sin pelos	0
	Con o sin pelos	1
	Con pelos	2
28. Presencia de una escama suplementaria junto al estambre	Ausente	0
	Presente	1
29. Agrupación de los estambres	No agrupados	0
	Agrupados o no agrupados	1
	Agrupados	2
30. Longitud del filamento (mm)	-	-
31. Longitud de la antera (mm)	-	-
32. Longitud de la escama (mm)	-	-
33. Presencia de ápice laciniado en la escama estaminal	Sin ápice laciniado	0
	Con o sin ápice laciniado	1
	Con ápice laciniado	2
34. Número de carpelos	En número de 3	0
	En número de 3 y 5	1
	En número de 5	2
35. Curvatura del estilo	Estilo no curvado	0
	Estilo curvado o no curvado	1
	Estilo curvado	2

Variable	Estados	Codificación
36. Número de óvulos por carpelo	-	-
37. Pubescencia del fruto	Glabro	0
	Pubescente	1
38. Longitud del fruto (mm)	-	-
39. Ancho del fruto (mm)	-	-
40. Desarrollo del carpóforo	Reducido	1
	Bien desarrollado	2
41. Longitud del carpóforo (mm)	-	-
42. Forma de la semilla	Semicircular o semielíptica	1
	Oblongo-reniforme	2
43. Longitud de la semilla (mm)	-	-

2 Elección de Variables

qué tipos de variables usamos en nuestro ejemplo?

VARIABLES

43 variables morfológicas.
Caracteres:
- 19 cualitativos codificadas
- 19 cuantitativos continuos (media)
- 5 cuantitativos discretos (moda)
- *excepción: número de óvulos (media)

Variable	Estados	Codificación
36. Número de óvulos por carpelo	-	-
37. Pubescencia del fruto	Glabro	0
	Pubescente	1
38. Longitud del fruto (mm)	-	-
39. Ancho del fruto (mm)	-	-
40. Desarrollo del carpóforo	Reducido	1
	Bien desarrollado	2
41. Longitud del carpóforo (mm)	-	-
42. Forma de la semilla	Semicircular o semielíptica	1
	Oblongo-reniforme	2
43. Longitud de la semilla (mm)	-	-

VARIABLES

43 variables morfológicas.
Caracteres:
- 19 cualitativos codificadas
- 19 cuantitativos continuos (media)
- 5 cuantitativos discretos (moda)
- *excepción: número de óvulos (media)

Variable	Estados	Codificación
36. Número de óvulos por carpelo	-	-
37. Pubescencia del fruto	Glabro	0
	Pubescente	1
38. Longitud del fruto (mm)	-	-
39. Ancho del fruto (mm)	-	-
40. Desarrollo del carpóforo	Reducido	1
	Bien desarrollado	2
41. Longitud del carpóforo (mm)	-	-
42. Forma de la semilla	Semicircular o semielíptica	1
	Oblongo-reniforme	2
43. Longitud de la semilla (mm)	-	-

VARIABLES

43 variables morfológicas.
Caracteres:
19 cualitativos codificadas
19 cuantitativos continuos (media)
5 cuantitativos discretos (moda)
*excepción: número de óvulos (media)

Variación intra-UE
¿Cómo tratarla?

---

1. Considerar que la variación intra-UE puede reducirse a una medida de posición estadística (media, mediana o moda).
2. Método del ejemplar: elegir al azar un organismo de los que componen la UE, para considerar que los estados presentes en ese organismo son los estados representativos de la UE.

Variación intra-UE
¿Cómo tratarla?

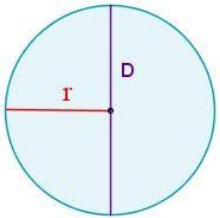
---

cómo se trata la variación intra-UE?

VARIABLES RELACIONADAS DETERMINISTICAMENTE
¿Cómo tratarlas?

¿Qué son? variables que pueden ser calculadas a partir de otra (una variable es función de otra)

Se debe excluir toda propiedad que sea consecuencia lógica de otra propiedad ya utilizada



D = 2.r

Longitud de la secuencia (bases) y número de codones **≈ longitud/3**

VARIABLES RELACIONADAS DETERMINISTICAMENTE
¿Cómo tratarlas?
---

hay en nuestros datos?

MALDICIÓN DE LA DIMENSIONALIDAD

La eficiencia y la precisión en la clasificación de las UE disminuye rápidamente a medida que aumenta el número de dimensiones o variables (Bellman 1957).

Gran problema: **alta dimensionalidad y bajo tamaño muestral**

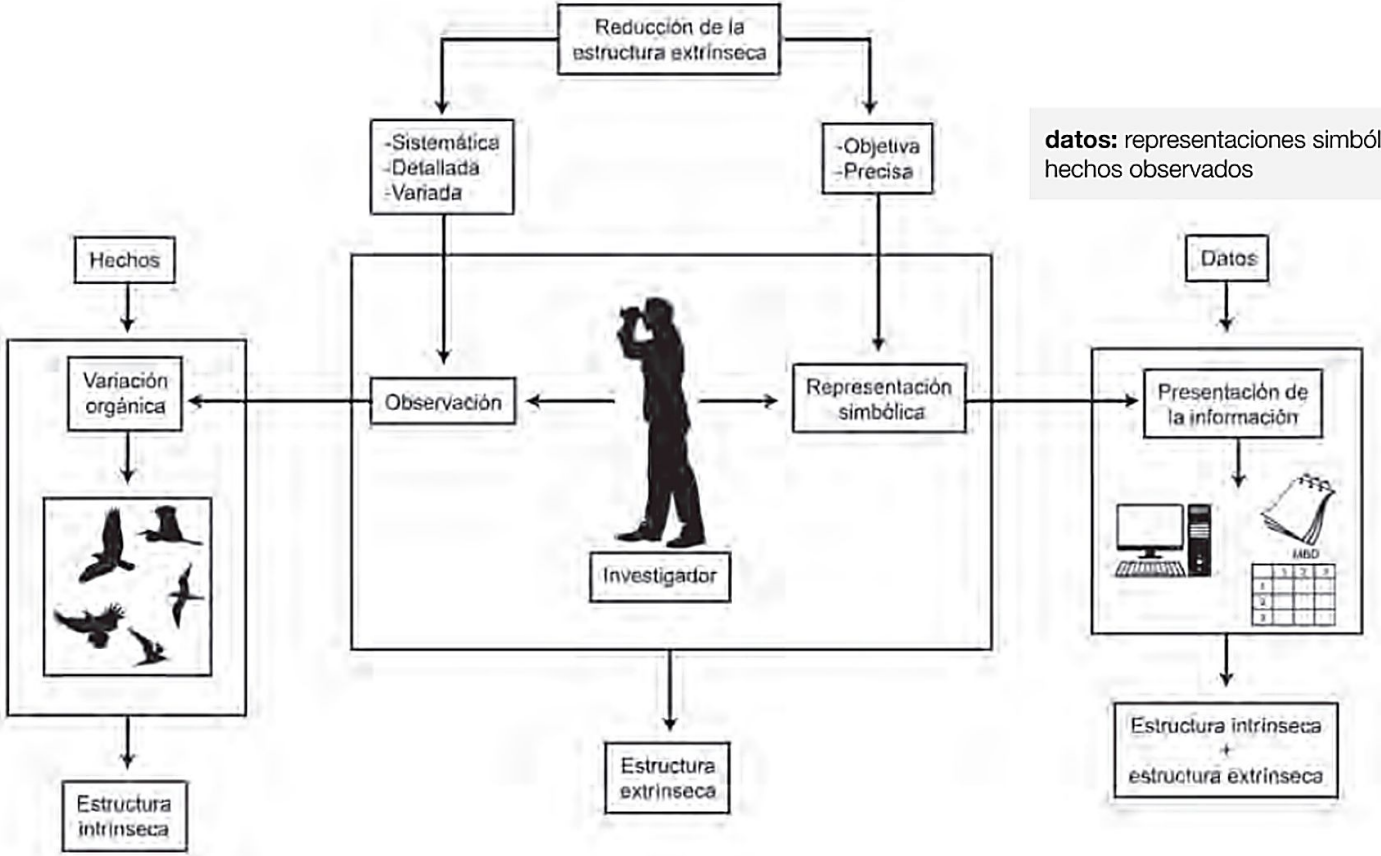
Bellman 1957:
al crecer las dimensiones, las distancias se vuelven menos informativas y el espacio se “vacía”

3 Construcción de Matriz Básica de Datos

Matriz Básica de Datos: matriz de n × p. Las n filas representan las UE y las p columnas representan las variables

Materia prima de cualquier análisis multivariado!!

		Variables				
		1	2	3	…	p
UE	1	x_{11}	x_{12}	x_{13}	…	x_{1p}
	2	x_{21}	x_{22}	x_{23}	…	x_{2p}
	3	x_{31}	x_{32}	x_{33}	…	x_{3p}
	⋮	⋮	⋮	⋮	⋱	⋮
	n	x_{n1}	x_{n2}	x_{n3}	…	x_{np}



datos: representaciones simbólicas de hechos observados

3 Construcción de Matriz Básica de Datos

todo lo que esté en azul es código!

instalaciones

!apt-get -qq install -y openjdk-8-jdk-headless > /dev/null ### instalamos java
!pip -q install pandas scikit-learn matplotlib ### instalamos las librerías
!pip install adjustText ### instalamos ésta utilidad para usarla y que no se nos superpongan las etiquetas

importaciones

```
import numpy as np          # arrays n-dimensionales y operaciones matemáticas/vectoriales
import pandas as pd         # estructuras de datos (DF) para análisis y manipulación de datos tabulares

from sklearn.preprocessing import StandardScaler
# escalar variables — centra en la media (0) y escala a desviación estándar 1 — útil antes de clustering o PCA

from sklearn.cluster import KMeans
# KMeans: algoritmo de clustering no supervisado que agrupa los datos en K clusters minimizando la varianza intra-cluster

from sklearn.decomposition import PCA
# PCA

import matplotlib.pyplot as plt
# matplotlib: biblioteca de gráficos 2D — permite crear histogramas, scatterplots, líneas de tiempo, etc.
```

| carga de datos

```
import pandas as pd
```

```
data = {
    "Especie": [
        "B. arborea", "B. carrapo", "B. chilensis", "B. bonariensis",
        "B. retama", "B. foliosa", "B. schickendantzii", "B. sarmientoi"
    ],
    "C1": [2, 2, 0, 0, 1, 0, 0, 2],
    "C2": [35, 36, 24, 20, 40, 19, 10, 22],
    "C3": [2.1, 1.6, 2.6, 1.3, 2.0, 1.3, 1.9, 1.4],
    "C4": [85, 97, 14, 26, 13, 28, 20, 21],
    "C5": [57, 71, 8.9, 18, 11, 25, 12, 27],
    "C6": [7.7, 9.0, 1.8, 3.4, 3.1, 5.8, 2.7, 5.1],
    "C7": [13, 7, 8, 14, 5, 4, 10, 2],
    "C8": [2, 2, 2, 1, 1, 1, 0, 2],
    "C9": [0, 0, 1, 0, 1, 1, 0, 2],
    "C10": [2, 2, 0, 2, 2, 2, 2, 1],
    "C11": [30, 40, 5.2, 8.9, 6.6, 14, 5.7, 17],
    "C12": [8.6, 16, 2.4, 2, 2.6, 7.8, 1.9, 12],
    "C13": [6, 6, None, 1, 2, 3, 1, 5],
    "C14": [0, 1, 2, 2, 1, 2, 2, 2],
    "C15": [1, 1, 1, 1, 2, 1, 1, 1],
```

```
df = pd.DataFrame(data)
```

```
df
```

MBD

df ⟶ MBD de *Bulnesia*

objeto DataFRame como una hoja de cálculo de Excel

creado a partir de un diccionario, los keys pasan a ser los nombres de las columnas y las listas las filas

Especie	C1	C2	C3	C4
B. arborea	2	35	2.1	85
B. carrapo	2	36	1.6	97
B. chilensis	0	24	2.6	14
B. bonariensis	0	20	1.3	26
B. retama	1	40	2.0	13
B. foliosa	0	19	1.3	28
B. schickendantzii	0	10	1.9	20
B. sarmientoi	2	22	1.4	21

para mostrar en una celda la matriz simplemente llamamos a df

MBD

ésta primer columna no quiero que sea considerada una columna, no es una variable más como C_n sino que son las UE,quiero que sea la etiqueta de la fila,QUIERO QUE SEA EL ÍNDICE

```
df = df.set_index("Especie")
```

DATOS FALTANTES
¿Cómo tratarlos?

- Variables que no pueden ser medidas debido a la naturaleza de lo que se intenta medir: longitud de la hoja de una especie áfila en un estudio del género donde el resto de las especies tienen hojas.
- Variables que pueden ser medidas pero que debido a situaciones aleatorias la medición no pudo concretarse.

Estrategias:

Eliminación de registros:

- Listwise deletion: elimina la fila completa que contiene el dato faltante (reduce el tamaño de la muestra y puede introducir sesgos)
- Pairwise deletion: elimina solo las celdas faltantes (los NA no son compatibles con algunos programas)

Imputación de valores:

- Media/mediana
- KNN
- Basada en k-means

comparar distribución de los datos antes y después!
considerar si tratar los datos faltantes antes o después de normalizar (los datos escalados sean representativos del set original o respeten la escala de el nuevo espacio de referencia)

VARIABLES EN DIFERENTES ESCALAS Y/O UNIDADES DE MEDIDAS
¿Cómo tratarlas?

---

es imprescindible que todas las variables estén en la misma escala, porque el algoritmo utiliza distancia Euclídea y de lo contrario las variables con mayor varianza dominarían la asignación de clusters

CENTRADO

$$x'_{ij} = x_{ij} - \bar{x}_j$$

resta la media de una variable a cada UE
expresa los valores en términos de distancia con respecto a la media
la media de esta nueva variable es 0, eliminando los efectos de magnitud debido a la posición de la media en las diferentes variables
las UE siguen siendo expresadas en las mismas unidades que en la MBD. centrado es aconsejable en aquellas MBD con variables medidas en las mismas unidades

ESTANDARIZACIÓN
NORMALIZACIÓN
(centrar y escalar)

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

centrado/desvío estándar
expresa los valores de la MBD en unidades de desvío estándar
útil cuando la escala y las unidades de las variables difieren

Especie	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	...	C34	C35	C36	C37	C38	C39	C40	C41	C42	C43
B. arborea	2	35	2.1	85	57.0	7.7	13	2	0	2	...	2	1.0	2	1	46	41	2	7.7	1	13.0
B. carrapo	2	36	1.6	97	71.0	9.0	7	2	0	2	...	2	2.0	2	1	56	52	2	5.3	1	12.0
B. chilensis	0	24	2.6	14	8.9	1.8	8	2	1	0	...	1	NaN	7	1	13	12	2	0.6	2	2.7
B. bonariensis	0	20	1.3	26	18.0	3.4	14	1	0	2	...	2	2.0	1	1	36	33	2	4.8	1	11.0
B. retama	1	40	2.0	13	11.0	3.1	5	1	1	2	...	2	1.0	8	1	23	19	1	0.8	2	11.0
B. foliosa	0	19	1.3	28	25.0	5.8	4	1	1	2	...	2	1.0	4	2	16	13	1	0.7	2	4.9
B. schickendantzii	0	10	1.9	20	12.0	2.7	10	0	0	2	...	2	1.0	4	2	12	13	1	0.4	2	5.3
B. sarmientoi	2	22	1.4	21	27.0	5.1	2	2	2	1	...	0	0.0	2	1	52	48	2	5.2	1	14.0

8 rows × 43 columns

hay datos faltantes

MBD

ESTANDARIZACIÓN

¿Estandarizar antes de tratar los datos faltantes?

La estandarización se basa en las medias y desviaciones de todas las variables disponibles originalmente.
Si eliminas primero las columnas, la media/varianza de las restantes cambia ligeramente.

recordar media = 0 s = 1

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

no es trivial cuando estandarizar!
no es tan simple decidir, en éste caso también se podría haber decidio estandarizar después para que los datos que no van a ser usados no afecten la escala!

ESTANDARIZACIÓN

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler() ### instanciamos la clase (se crea el objeto vacío con los parámetros with_mean=True y with_std=True)

Z = scaler.fit_transform(df) ### fit calcula media y desviación estándar de cada columna,transform calcula el Z Score de cada valor. DEVUELVE ARRAY de np! sin nombres de fila ni de columna!

Z_df = pd.DataFrame(Z, index=df.index, columns=df.columns) ### reformatar como dataframe

Z_df
```

ESTANDARIZACIÓN

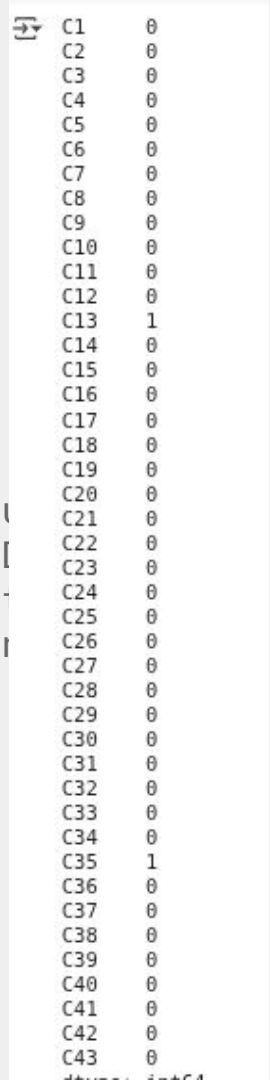
Z_df ⟶ MBD de *Bulnesia estandarizada*

Especie	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	...	C34	C35	C36	C37	C38	C39	C40	C41	C42	C43
B. arborea	1.21356	0.963398	0.756889	1.510247	1.313211	1.206921	1.286918	0.898027	-0.898027	0.538816	...	0.538816	-0.223607	-0.733799	-0.577350	0.843081	0.776447	0.774597	1.678380	-1.0	0.941586
B. carrapo	1.21356	1.067549	-0.407556	1.895842	1.963718	1.752659	-0.219718	0.898027	-0.898027	0.538816	...	0.538816	1.341641	-0.733799	-0.577350	1.434717	1.480853	0.774597	0.785724	-1.0	0.691331
B. chilensis	-0.94388	-0.182264	1.921334	-0.771190	-0.921745	-1.269891	0.031388	0.898027	0.538816	-2.334869	...	-0.898027	NaN	1.362770	-0.577350	-1.109317	-1.080623	0.774597	-0.962395	1.0	-1.636045
B. bonariensis	-0.94388	-0.598869	-1.106223	-0.385595	-0.498916	-0.598213	1.538024	-0.538816	-0.898027	0.538816	...	0.538816	1.341641	-1.153113	-0.577350	0.251445	0.264152	0.774597	0.599754	-1.0	0.441075
B. retama	0.13484	1.484154	0.524000	-0.803323	-0.824169	-0.724153	-0.721930	-0.538816	0.538816	0.538816	...	0.538816	-0.223607	1.782084	-0.577350	-0.517681	-0.632364	-1.290994	-0.888007	1.0	0.441075
B. foliosa	-0.94388	-0.703020	-1.106223	-0.321329	-0.173662	0.409304	-0.973035	-0.538816	0.538816	0.538816	...	0.538816	-0.223607	0.104828	1.732051	-0.931826	-1.016586	-1.290994	-0.925201	1.0	-1.085483
B. schickendantzii	-0.94388	-1.640380	0.291111	-0.578393	-0.777704	-0.892072	0.533600	-1.975658	-0.898027	0.538816	...	0.538816	-0.223607	0.104828	1.732051	-1.168481	-1.016586	-1.290994	-1.036783	1.0	-0.985381
B. sarmientoi	1.21356	-0.390567	-0.873334	-0.546260	-0.080733	0.115445	-1.475247	0.898027	1.975658	-0.898027	...	-2.334869	-1.788854	-0.733799	-0.577350	1.198062	1.224706	0.774597	0.748530	-1.0	1.191842

8 rows × 43 columns

¿Cuántos valores faltantes hay y dónde están?

```
print(Z_df.isnull().sum()) ### crea un DF del mismo tamaño que Z_df, pero con valores booleanos (True donde había u valor nulo -NaN- False donde había un dato válido). sobre ese D de booleanos, suma por columnas todos los True (Python los trat como 1), de modo que obtienes un conteo de valores faltantes en cada columna. muestra en pantalla el conteo
```



```
C1     0
C2     0
C3     0
C4     0
C5     0
C6     0
C7     0
C8     0
C9     0
C10    0
C11    0
C12    0
C13    1
C14    0
C15    0
C16    0
C17    0
C18    0
C19    0
C20    0
C21    0
C22    0
C23    0
C24    0
C25    0
C26    0
C27    0
C28    0
C29    0
C30    0
C31    0
C32    0
C33    0
C34    0
C35    1
C36    0
C37    0
C38    0
C39    0
C40    0
C41    0
C42    0
C43    0
```

MBD sin NaN

optamos por eliminar columnas con datos faltantes

```
df2 = Z_df.drop(columns=["C13", "C35"]) ### crea un nuevo DF a partir de Z_df eliminando las columnas llamadas "C13" y "C35"
```

```
print(df2.isnull().sum())
```

```
df2
```

MBD sin NaN

MBD lista!!

- Sin datos faltantes
- Estandarizada

MBD sin NaN

Especie	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	...	C33	C34	C36	C37	C38	C39	C40	C41	C42	C43
B. arborea	1.21356	0.963398	0.756889	1.510247	1.313211	1.206921	1.286918	0.898027	-0.898027	0.538816	...	0.898027	0.538816	-0.733799	-0.577350	0.843081	0.776447	0.774597	1.678380	-1.0	0.941586
B. carrapo	1.21356	1.067549	-0.407556	1.895842	1.963718	1.752659	-0.219718	0.898027	-0.898027	0.538816	...	0.898027	0.538816	-0.733799	-0.577350	1.434717	1.480853	0.774597	0.785724	-1.0	0.691331
B. chilensis	-0.94388	-0.182264	1.921334	-0.771190	-0.921745	-1.269891	0.031388	0.898027	0.538816	-2.334869	...	-0.538816	-0.898027	1.362770	-0.577350	-1.109317	-1.080623	0.774597	-0.962395	1.0	-1.636045
B. bonariensis	-0.94388	-0.598869	-1.106223	-0.385595	-0.498916	-0.598213	1.538024	-0.538816	-0.898027	0.538816	...	-1.975658	0.538816	-1.153113	-0.577350	0.251445	0.264152	0.774597	0.599754	-1.0	0.441075
B. retama	0.13484	1.484154	0.524000	-0.803323	-0.824169	-0.724153	-0.721930	-0.538816	0.538816	0.538816	...	-0.538816	0.538816	1.782084	-0.577350	-0.517681	-0.632364	-1.290994	-0.888007	1.0	0.441075
B. foliosa	-0.94388	-0.703020	-1.106223	-0.321329	-0.173662	0.409304	-0.973035	-0.538816	0.538816	0.538816	...	-0.538816	0.538816	0.104828	1.732051	-0.931826	-1.016586	-1.290994	-0.925201	1.0	-1.085483
B. schickendantzii	-0.94388	-1.640380	0.291111	-0.578393	-0.777704	-0.892072	0.533600	-1.975658	-0.898027	0.538816	...	0.898027	0.538816	0.104828	1.732051	-1.168481	-1.016586	-1.290994	-1.036783	1.0	-0.985381
B. sarmientoi	1.21356	-0.390567	-0.873334	-0.546260	-0.080733	0.115445	-1.475247	0.898027	1.975658	-0.898027	...	0.898027	-2.334869	-0.733799	-0.577350	1.198062	1.224706	0.774597	0.748530	-1.0	1.191842

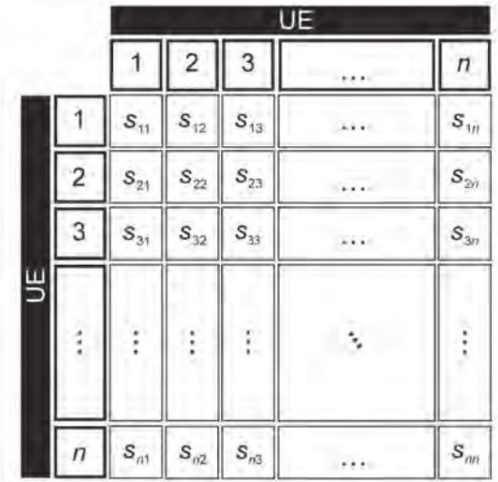
8 rows × 41 columns

4 Construcción de Matriz de Similitud

Matriz de Similitud: matriz cuadrada y simétrica que representa el grado de similitud o disimilitud entre pares de Unidades Experimentales (UE).

Se construye comparando cada par de UE mediante un coeficiente adecuado, que puede ser una medida de:

- distancia (p. ej. Euclídea, Manhattan),
- asociación (p. ej. Jaccard, Bray-Curtis),
- o correlación (p. ej. Pearson, si las variables están estandarizadas).



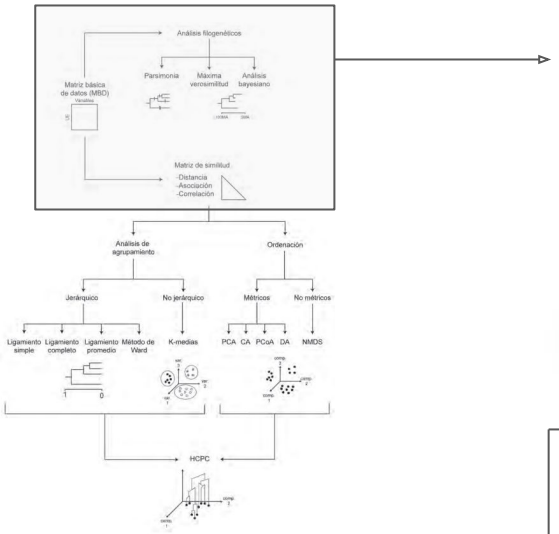
4 Construcción de Matriz de Similitud

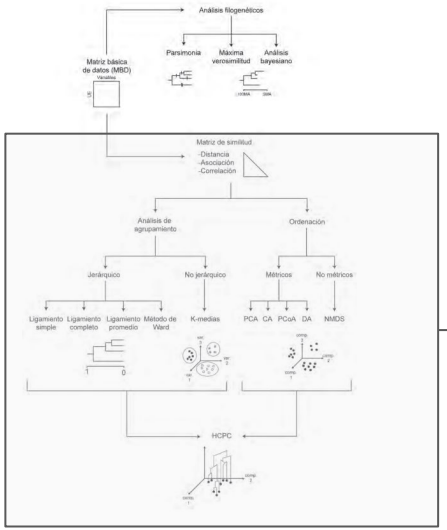
en nuestro ejemplo trabajamos directamente sobre la MBD porque el algoritmo busca minimizar la suma de distancias cuadradas de cada punto a su centroide, **si en vez de las coordenadas de las muestras se le da una matriz de semejanzas o distancias, no sabría "dónde" poner los centroides**, porque éstos son combinaciones lineales de las variables originales.

La MS es insuficiente para expresar relaciones entre la totalidad de las unidades UE porque sólo expone similitudes entre pares de unidades.

para reconocer las relaciones entre todas las UE analizadas → técnicas de análisis de la MS o de la MBD

Aplicación de Análisis Cuantitativo





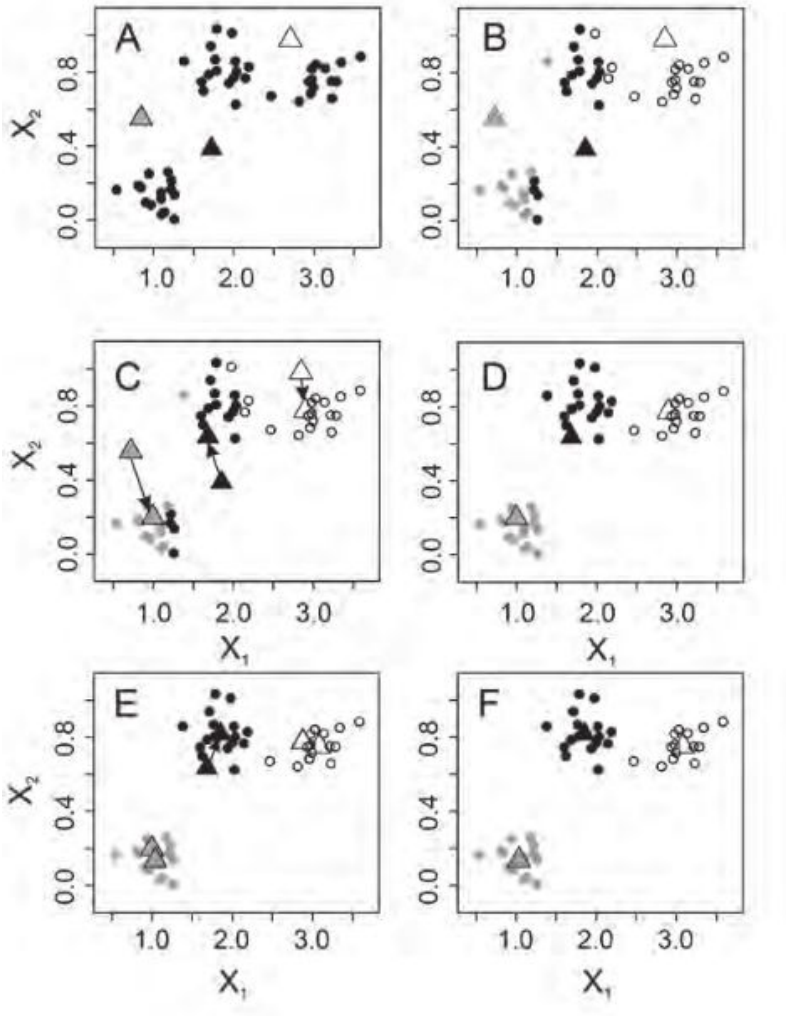
MULTIVARIADO

ANÁLISIS DE AGRUPAMIENTOS

k-medias

ANÁLISIS DE AGRUPAMIENTOS

		k-medias
exclusivas	**cada UE es exclusiva de un grupo**	**X**
no exclusivas	cada UE puede pertenecer a más de un grupo	
jerárquicas	grupos con rangos	
no jerárquicas	**grupos sin rangos**	**X**
secuenciales	grupos formado de a uno por vez	
simultáneas	**grupos formados todos simultáneamente**	**X**
directas	cuando una UE es asignada a un grupo este agrupamiento no se modifica	
iterativas	**la pertenencia de las UE a un determinado grupo puede ir cambiando durante el análisis**	**X**
no supervisadas	número de grupo establecido a posteriori	
supervisadas	**número de grupos establecido a priori**	**X**



dada una MBD, determinar e**l agrupamiento de las UE en K grupos,** de forma tal que las UE dentro de cada grupo sean más parecidas entre sí que a las UE de otros grupos

Procedimiento

1. Considerar a priori k grupos
2. Elegir k puntos aleatorios, funcionaran como centroides iniciales de cada grupo
3. Asignar cada UE al centroide más cercano (distancia euclidiana)
4. Calcular nuevo centroide para cada grupo usando la media de las coordenadas de las UE que se asignaron a ese grupo
5. Repetir 3 y 4 hasta convergencia, en algún momento las asignaciones de las UE dejan de cambiar!

Enunciados

1. El centro de un grupo es la media (**centroide**) de todas las UE pertenecientes al grupo.

2. Cada UE está más cerca de su propio centroide que de los centroides de otros grupos.

Objetivo: minimizar la suma de errores (distancia euclídea entre centroide y punto) al cuadrado

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

x_i vector de características
μ_k es el centroide (promedio) de los puntos en C_k
k grupos

$$\text{cluster}(x_i) = \arg\min_k \|x_i - \mu_k\|^2.$$

Desventajas:

- la solución depende de la posición inicial de los centroides de cada grupo (probar con varias configuraciones iniciales)
- sensible a la presencia de UE atípicas (con valores que se alejan mucho del resto de las UE) !!!!
- asume clusters de forma convexa y similar tamaño (esferas en el espacio de características)

Suponen que las especies se agrupan en tres grupos

k = 3

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=3,n_init=1000,random_state=42)

kmeans.fit(df2) ### se aplica k-means a la MBD
```

el objeto kmeans almacena:

kmeans.labels_: array con la etiqueta (0, 1 o 2) de cluster para cada muestra.

kmeans.cluster_centers_: las coordenadas de los 3 centroides en el espacio de características.

kmeans.inertia_: la suma de distancias al cuadrado desde cada punto hasta su centroide, la métrica que el algoritmo minimiza! cuanto más baja la inercia, más compactos son los grupos!

selecciona la partición con la menor inertia y descarta las otras 999!

$$\text{totSS} = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

variabilidad global de todos los datos respecto a la media global

$$\text{withinSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

suma de la variabilidad dentro de cada clúster, respecto al centroide de ese clúster

$$\text{betweenSS} = \sum_{k=1}^{K} |C_k| \, (\mu_k - \bar{x})^2$$

variabilidad entre los diferentes clústeres, cuánto difieren los centroides de la media global.

$$\text{totSS} = \text{withinSS} + \text{betweenSS}$$

Sumando esos dos aportes cuadráticos, para todas las observaciones y todos los grupos, recuperamos exactamente la desviación de cada punto a la media global.

$(x_i - \mu_k)$: cuánto se aleja cada observación de su propio centroide

$(\mu_k - \bar{x})$: cuánto se aleja ese centroide de la media global

$$\text{totSS} = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

variabilidad global de todos los datos respecto a la media global

```
totss = np.sum((df2 - df2.mean(axis=0))**2)
totss_scalar = totss.sum()
```

$$\text{withinSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

suma de la variabilidad dentro de cada clúster, respecto al centroide de ese clúster

```
betweenss = totss_scalar - inertia
```

$$\text{betweenSS} = \sum_{k=1}^{K} |C_k| \, (\mu_k - \bar{x})^2$$

variabilidad entre los diferentes clústeres, cuánto difieren los centroides de la media global.

```
inertia = kmeans.inertia_
```

$$\text{totSS} = \text{withinSS} + \text{betweenSS}$$

$$\% \text{ Varianza explicada} = \frac{\text{betweenSS}}{\text{totSS}} \times 100$$

qué proporción de toda la dispersión original (TotSS) está capturada o explicada por la estructura de clusters (BetweenSS)

el x % de la variabilidad de tus datos se debe a la separación entre grupos; el x % restante ocurre dentro de los clusters (WithinSS)

```
pct_explained = betweenss / totss_scalar * 100
```

K-medias

```
cluster_por_especie = pd.Series(labels, index=df2.index,
name="Cluster")
print(cluster_por_especie)
print("Tamaño de cada clúster:", np.bincount(labels))
print("Varianza total (totss):", totss_scaler)
print("Suma cuadrados entre (betweenss):", betweenss)
print("% Varianza explicada:", pct_explained)
```

```
Especie
B. arborea          0
B. carrapo          0
B. chilensis        1
B. bonariensis      1
B. retama           1
B. foliosa          1
B. schickendantzii  1
B. sarmientoi       2
Name: Cluster, dtype: int32
[0 0 1 1 1 1 1 2]
Tamaño de cada clúster: [2 5 1]
Varianza total (totss): 328.0
Suma cuadrados entre (betweenss): 198.63683187122263
% Varianza explicada: 60.56000971683617
```

El clustering capta buena parte de la estructura morfológica, pero todavía hay variación interna que podría deberse a micro-diferencias o ruido

Hay un singleton

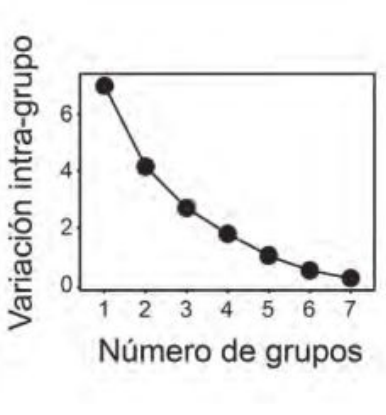
EXplorar otros K?

Aunque el K es determinado a priori (supervisado) se puede evaluar cuál es el número óptimo de grupos a posteriori

---

Existe una buena partición cuando los grupos son homogéneos.
Baja variación intra-grupo y una alta variación entre grupos (cociente mímimo INTRA/ENTRE)

Elbow method (Within-Cluster Sum of Squares): analiza la variación intra-grupo como función de la cantidad de grupos, el número óptimo de grupos es aquel que, al ir subdividiendo los grupos, los subgrupos resultantes no disminuyen de manera significativa la variación intra-grupo.



Aunque el K es determinado a priori (supervisado) se puede evaluar cuál es el número óptimo de grupos a posteriori

---

- Para cada muestra, el *silhouette* combina dos cosas:
  1. **Cohesión** (a_i): lo bien que una muestra está dentro de su propio clúster (distancia media a los demás del mismo clúster).
  2. **Separación** (b_i): lo lejos que está esa muestra del clúster más cercano distinto (la distancia media al siguiente clúster más próximo).
- El coeficiente de cada muestra es

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

  y varía entre –1 y +1:
  - Valores cerca de +1 → bien agrupada y lejos de otros clústeres.
  - Valores cerca de 0 → en el límite entre clústeres.
  - Valores negativos → mal asignada (está más cerca de otro clúster que del suyo).
- El *silhouette score* global que calculas (`silhouette_score`) es el promedio de todos los s_i.

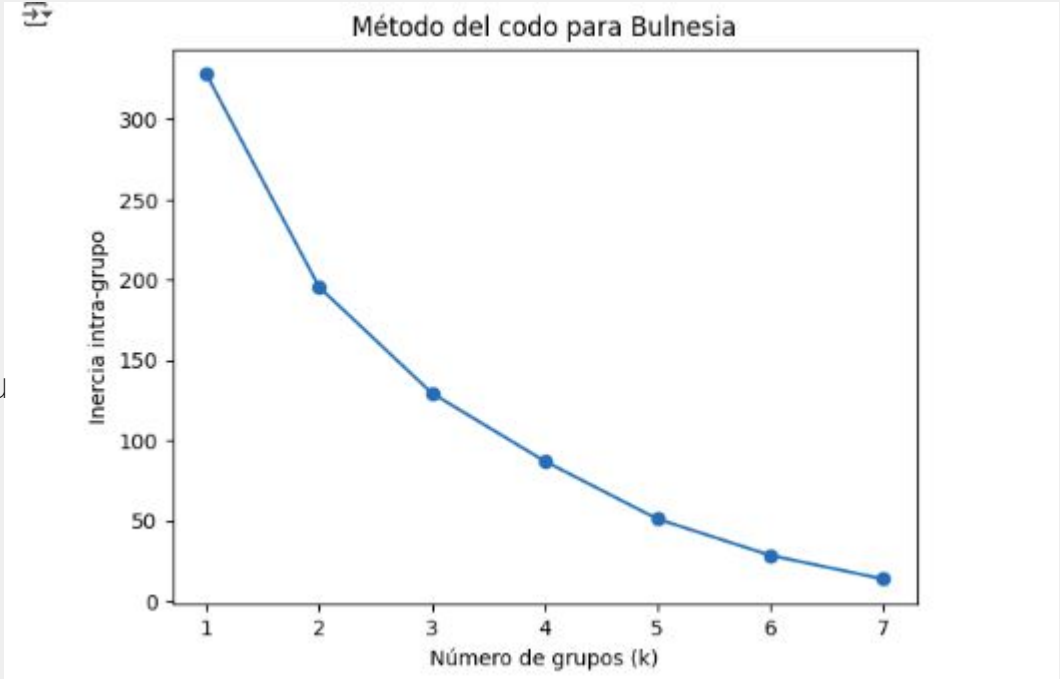
método del codo

K óptimos

```
import matplotlib.pyplot as plt

inertias = []
K_range = range(1, 8)
for k in K_range:
    km = KMeans(n_clusters=k, n_init=50, random_state=42)
    km.fit(df2)
    inertias.append(km.inertia_)

plt.figure()
plt.plot(list(K_range), inertias, marker='o')
plt.xlabel("Número de grupos (k)")
plt.ylabel("Inercia intra-grupo")
plt.title("Método del codo para Bulnesia")
plt.show()
```

K óptimos

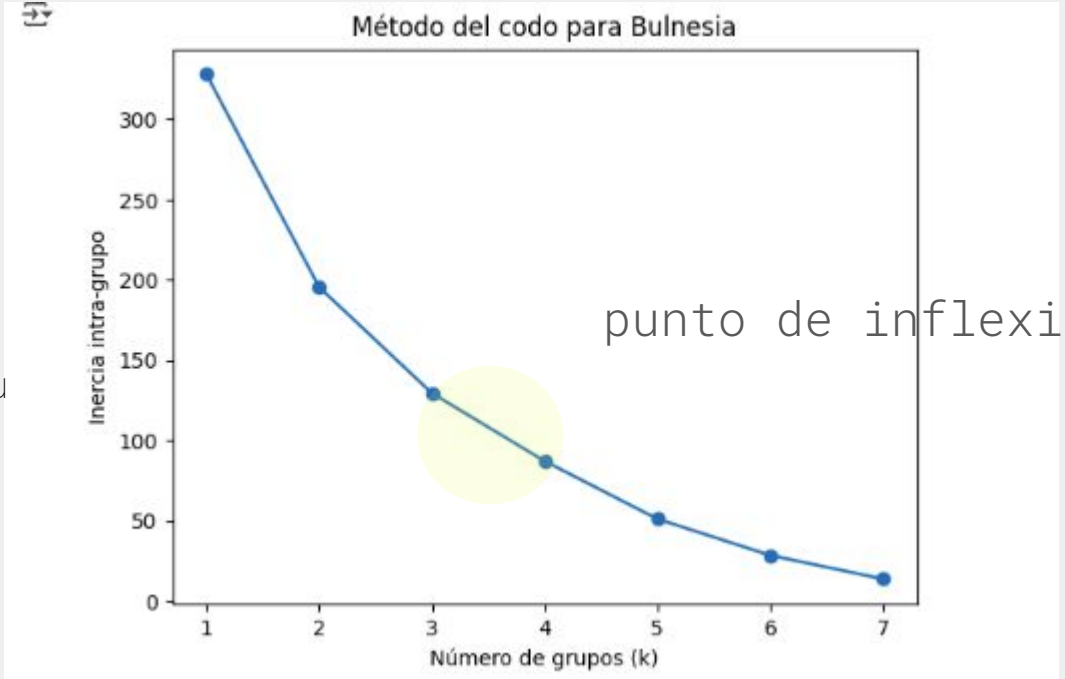
distancias cuadráticas de cada punto a su centroide



en qué K la reducción de inercia deja de ser significativa?

K óptimos

distancias cuadráticas de cada punto a su centroide



punto de inflexión

No cuantifica la calidad de separación entre clusters, solo mide compacidad interna.

El “codo” puede no ser claro o no existir de forma nítida en algunos datasets.

silhouette score

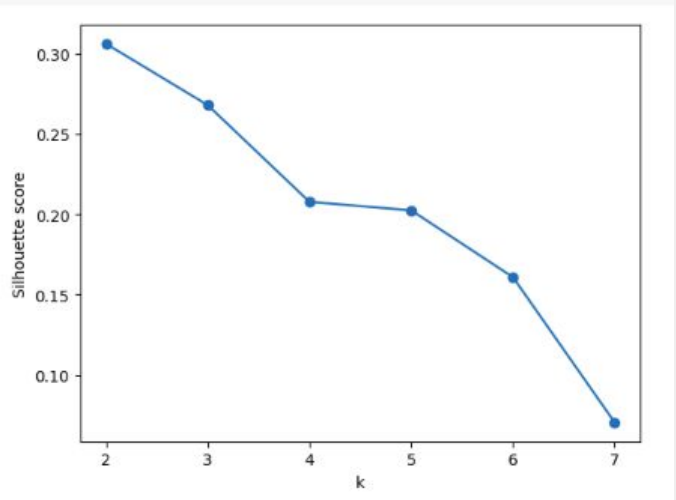
```
from sklearn.metrics import silhouette_score

sils = []
K_range = range(1, 8)
for k in K_range[1:]:  # no tiene sentido k=1
    km = KMeans(n_clusters=k, random_state=42,
n_init=50)
    labels = km.fit_predict(df2)
    sils.append(silhouette_score(df2, labels))

plt.plot(K_range[1:], sils, '-o')
plt.xlabel('k')
plt.ylabel('Silhouette score')
plt.show()
```

K óptimos

silhouette score

K óptimos

mide cohesión interna y separación externa

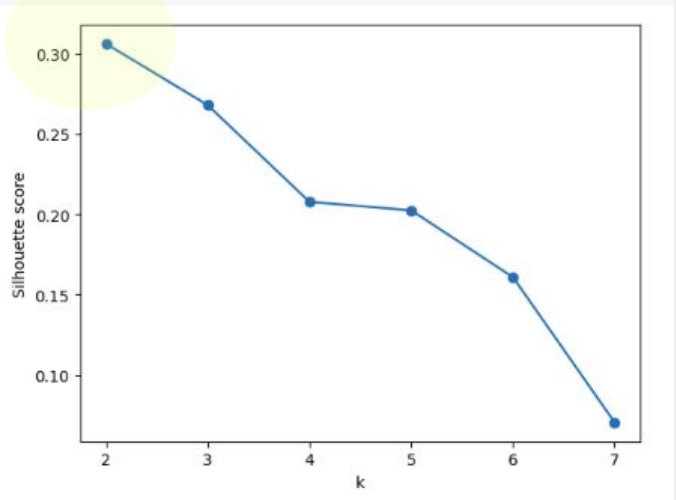


silhouette promedio alto indica clusters compactos y bien separados
silhouette negativo puntos mal asignados
SÍ ES COMPARABLE ENTRE DATASETS!!

silhouette score

valor máximo

mide cuán bien se separan los clusters



K óptimos

¿Y ahora? ¿Cuál K elijo?

K óptimos

K=2

Ventaja: mejor calidad de clusterización (silhouette más alto), indica dos grupos muy bien diferenciados.

Desventaja: explica menos varianza total (inercia todavía relativamente alta).

K=3

Ventaja: “codo” en la inercia, con un gran salto de reducción entre 2→3; explica más varianza (~60 %), capturando un tercer grupo (tu especie - outlier).

Desventaja: silhouette ya baja un poco, clusters menos compactos.

K óptimos

Si lo que buscas es la partición más clara posible, k=2 es óptimo según el silhouette.

Si quieres capturar más heterogeneidad (p.ej. ese singleton muy distinto que vimos), k=3 es un buen compromiso: baja un poco la cohesión global, pero te da un tercer grupo que puede ser biológicamente relevante.

hagan k-mean con k=2

Visualizar en PCA coloreando por K=2K=2K=2 y por K=3K=3K=3 para ver cuál separa mejor las especies.

Calcular el silhouette “por cluster” para ver si alguno tiene silhouette negativo o muy bajo.

MÉTODOS DE ORDENACIÓN

Análisis de Componentes Principales (PCA)

MÉTODOS DE ORDENACIÓN

¿Por qué se llaman métodos de ordenación?

Las UE pueden ubicarse en un espacio de menores dimensiones que el espacio original de la matriz, mientras se preservan de la mejor forma posible las relaciones entre dichas unidades.

Se reducen dimensiones!

Para visualizar (porque no puedo visualizar más de 3D)

Para disminuir el efecto de la maldición de la multidimensionalidad

Análisis de Componentes Principales

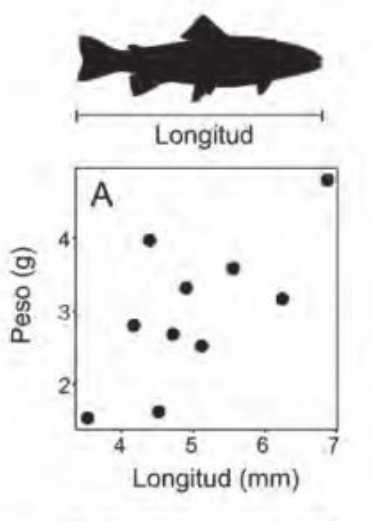
Objetivo: representar un conjunto de variables mediante un número reducido de combinaciones lineales de las mismas, denominadas **componentes principales (PC)**.

Útil cuando hay correlación o redundancia entre las variables, cuando la información de una variable se solapa parcialmente con otra para comprimir la información en menos dimensiones sin perder mucho. La reducción de dimensiones puede ser usada para visualizar o como paso previo a un agrupamiento.

Componentes principales:

- no están correlacionados entre sí permitiendo interpretación independiente
- cada uno contiene una parte de la variabilidad total la MBD original
- el PC1 contiene la mayor variabilidad, el PC2 captura la mayor variabilidad restante, y así sucesivamente hasta distribuir toda la variabilidad en los PCs.
- cada PC contiene información de todas las variables pero en diferentes proporciones.

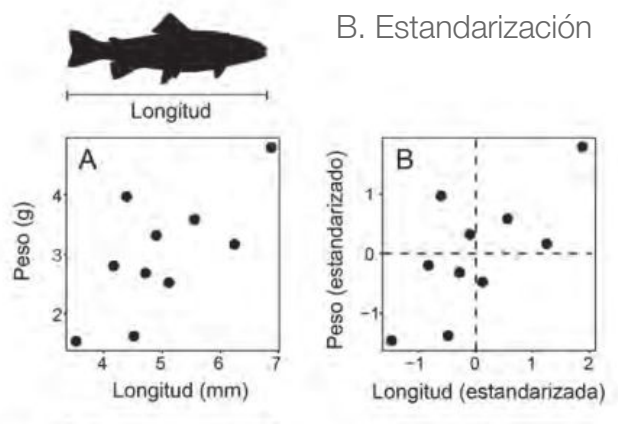
A. Gráfico de dispersión



10 Unidades Experimentales: individuos de la misma especie de pez

2 Variables: longitud y peso (correlación hipotética 0,7)

B. Estandarización



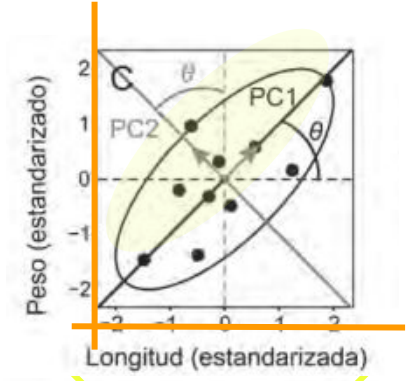
variables en diferentes escalas → estandarización
media = 0
varianza = 1

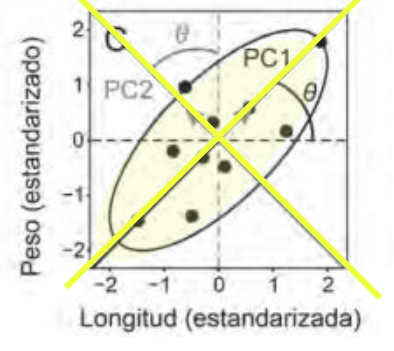
$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

La **dispersión de las UE** en relación con cualquiera de los dos ejes **depende de la escala utilizada**. Ambas variables deben ser expresadas en la misma unidad de medida. Estandarizar.

Las UE están referidas ahora a un **nuevo par de ejes ortogonales** (perpendiculares entre sí) que se cortan en un punto que corresponde al promedio de cada variable.
Cada UE tendrá un **nuevo par de coordenadas definidas en función de unidades de desvío estándar**.

C. Elipse que engloba las UE y PC 1 y 2, estos ejes corresponden a la rotación de la ordenada y la abscisa un ángulo θ





Si se mide cuánto varían los puntos en dirección horizontal (X, longitud) o vertical (Y, peso), no se está considerando la dirección en la que los puntos están más dispersos.
La dirección de máxima dispersión no coincide con X ni con Y: está inclinada, en diagonal. **PCA rota el sistema de coordenadas para alinear los nuevos ejes con las direcciones donde los puntos están más dispersos**:

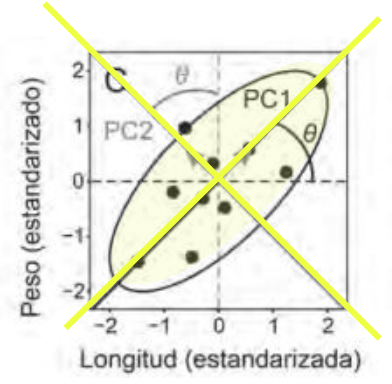
El PC1 apunta en la dirección donde hay más variación → eje largo de la elipse.

El PC2 es perpendicular a PC1 → el eje corto.

Cada punto se puede proyectar sobre estos nuevos ejes, que ahora sí reflejan mejor las diferencias entre UE.

Geométricamente, la d**isposición espacial de las UE para dos variables correlacionadas es la de una nube elíptica**. El eje de mayor variación coincide con el eje mayor de esa elipse y corresponde al PC1.

C. Los vectores (flechas grises) representan los eigenvectores del PCA cuyas longitudes son 1.



Los PCs corresponden a una rotación de la ordenada y la abscisa un cierto ángulo θ. Los vectores que definen la ubicación y la dirección de los ejes mayor y menor se denominan **eigenvectores, autovectores o vectores propios,** y tienen longitud igual a 1.

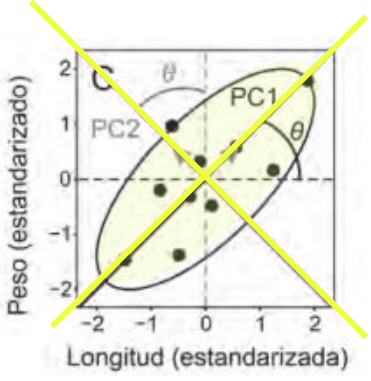
Los eigenvalores, autovalores o valores propios λ_i, reflejan la variación de cada PC (mayor en el primero, menor en el segundo, y así sucesivamente). La suma de todos los eigenvalores constituye la varianza total de la MBD original, y en el caso de una MBD estandarizada corresponde al número total de variables. **$\lambda_i > 1$** indica que un componente representa más de una variable.

Si analizamos tres variables, la elipse se transforma en un elipsoide y el tercer componente principal estaría representado por el tercer eje del elipsoide, perpendicular a los dos primeros. Si el estudio incluyera más de tres variables, se necesitarían dimensiones adicionales cuya representación geométrica no puede ser visualizada, pero de igual forma puede aplicarse el tratamiento matemático.

En nuestro ejemplo, los eigenvalores son 1,7 y 0,3 (PC1 y PC2, respectivamente), los eigenvectores para el PC1 son 0,707 (longitud) y 0,707 (peso).

Eigenvalor: varianza explicada por un PC

Eigenvector: vector que define la dirección de un PC



Cada componente es una nueva variable hipotética que se construye utilizando todas las variables de la MBD, dado que representa una rotación de los ejes originales. Cada PC está compuesto de la **suma del producto de los eigenvectores por las variables estandarizadas.**

$$PC1 = a_{11} \times \text{longitud} + a_{12} \times \text{peso}$$
$$PC2 = a_{21} \times \text{longitud} + a_{22} \times \text{peso}$$

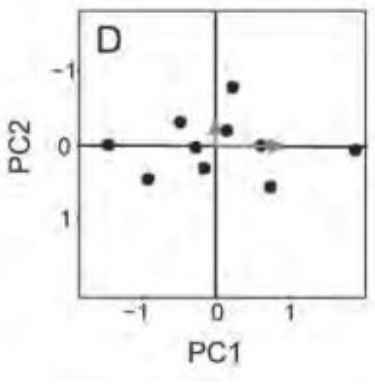
eigenvector eigenvector

Los coeficientes a_{ij} (i = número de componente, j = número de variable) corresponden a los eigenvectores y están sujetos a la condición $a_{11} + a_{12} = 1$ y $a_{21} + a_{22} = 1$ (longitud unitaria).

Al calcular la suma del producto de los eigenvectores por las variables estandarizadas obtenemos las coordenadas de las UE en el nuevo espacio de ordenación, denominadas **scores**.

Score: coordenada de la UE en el espacio de los PCs

D. Resultado del PCA en dos dimensiones, se muestran los eigenvectores multiplicados por la raíz cuadrada de sus eigenvalores (loadings).



Un concepto sumamente importante en el PCA es el concepto de **loading**, definido como el producto de un eigenvector por la raíz cuadrada de su eigenvalor. El loading corresponde al **coeficiente de correlación de Pearson (r) entre una variable y un componente principal**, son como la **contribución relativa de cada variable a cada componente**. Así, todas las variables contribuyen a todos los componentes pero de manera diferencial; es decir, la variable 1 puede ser un importante aporte para el PC1, pero pobre para el PC2.

$$\ell_{ij} = e_{ij}\sqrt{\lambda_j}$$

En nuestro ejemplo, los loadings para el PC1 son las correlaciones entre la longitud y el PC1 ($r_{\text{longitud, PC1}}$), y entre el peso y el PC1 ($r_{\text{peso, PC1}}$):

$$r_{\text{longitud, PC1}} = 0,707 \times \sqrt{1,7} = 0,92$$

$$r_{\text{peso, PC1}} = 0,707 \times \sqrt{1,7} = 0,92$$

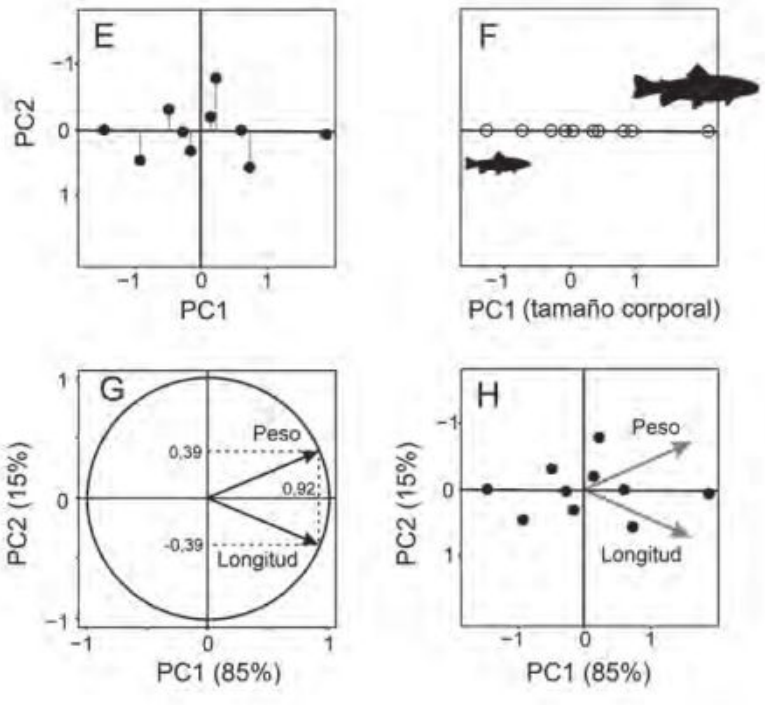
Loading: contribución de cada variable a cada PC

Cuantos PCs retener?

La suma de eigenvalores constituye la varianza total de la MBD original, puede calcularse el porcentaje de variación contenido en cada componente principal según su aporte a esa suma, como el cociente entre el eigenvalor de un componente y la suma total de eigenvalores.

$$\text{Varianza}\%_i = \frac{100\,\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

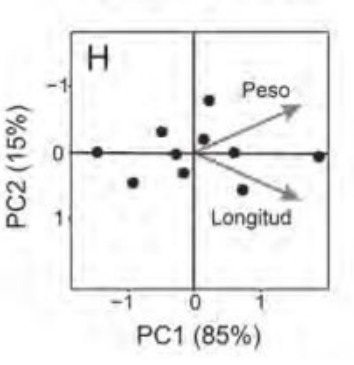
En el ejemplo los eigenvalores son 1,7 y 0,3 (PC1 y PC2, respectivamente), por lo que su contribución relativa es 100%× 1,7 / (1,7 + 0,3) = 85%. Esto significa que el PC1 representa el 85% de la variación total de la MBD, y se considera suficiente retener e interpretar sólo este componente. Esto se debe a que ambas variables están muy correlacionadas y contienen información redundante. Sería trivial mantener
ambos componentes, ya que equivaldría a interpretar las dos variables de la MBD original (por definición, **utilizar todos los componentes equivale a utilizar todas las variables de la MBD y por lo tanto, acumulan el 100% de la variación total**).



E. Las UE se proyectan únicamente sobre el PC1 (líneas perpendiculares)
F. Las nuevas coordenadas de las UE proyectadas sobre el PC1 (círculos blancos) se denominan scores.
G. Círculo de correlación con radio igual a 1, donde se muestran las variables (vectores) y sus coordenadas (loadings) que definen la calidad de la representación
H. Biplot de UE vs. variables.

La posición de cada UE en el espacio de componentes principales está dada por sus coordenadas, denominadas scores. Las coordenadas de las UE y de las variables no están construidas sobre el mismo espacio, por lo tanto debemos enfocarnos en la dirección y en el sentido de las variables y no en sus posiciones absolutas sobre el gráfico.

H. Biplot de UE vs. variables



Reglas de interpretación de un biplot:

(1) las UE cercanas en el espacio tienen características similares en cuanto a sus variables
(2) una UE que está cercana a una variable tiene un alto valor para esa variable
(3) una UE que se encuentra opuesta a una variable tiene un bajo valor para esa variable.

Por ejemplo en la Figura 6.1H se observa que las UE del lado derecho tienen valores altos de longitud y peso corporal.

La posición de uno u otro componente en la abscisa o
en la ordenada es indistinta.
Las representaciones gráficas deben ir acompañadas de una tabla que contenga la siguiente información acerca del PCA: eigenvalores, porcentaje de variación explicada por cada componente, acumulación de dicho porcentaje y loadings.
La suma de los porcentajes contenidos en los ejes seleccionados da una idea de la cantidad de variación expresada por la ordenación. Por ejemplo, si el primer componente contiene el 50% de la variación total y el segundo componente 20%, el gráfico bidimensional de estos componentes expresará el 70% de la variación total de la MBD.
Una alta variación en los primeros componentes indica que las variables están muy correlacionadas o son redundantes, porque comparten información similar.

PCA para visualización

Reducir el número de dimensiones con alguna técnica de ordenación, como el análisis de componentes principales, y luego utilizar los primeros componentes para graficar.

```
pca = PCA(n_components=2) ### cuantas PCs conservar cuando se calculen
coords = pca.fit_transform(df2) ### caplica el PCA para calcular los autovectores y autovalores, y proyecta cada punto en el nuevo sistema de referencia (fit aprende, transform aplica)
centroids_2d = pca.transform(centroids) ### es necesario proyectar los centroides en el nuevo espacio de referencia también
print("Varianza explicada por PC1 y PC2:", pca.explained_variance_ratio_)
print("Componentes principales:\n", pca.components_)
```

```
Varianza explicada por PC1 y PC2: [0.45608031 0.20564175]
Componentes principales:
 [[ 0.17109653  0.13024702 -0.03137832  0.21796688  0.21742585  0.19725688
    0.07566073  0.1362651  -0.10473369  0.07199614  0.21425187  0.16207819
   -0.16137667 -0.06102761  0.21445727  0.16278322  0.03326517  0.10828488
    0.01047979  0.22412356  0.22781795  0.20226769 -0.12155416  0.2115917
    0.22293856 -0.09601601 -0.08696299  0.14344199  0.14375337 -0.11568996
    0.15891411  0.07947645  0.04346792 -0.14485514 -0.11827942  0.19675542
    0.19552324  0.14748212  0.20150728 -0.18944028  0.15475148]
 [-0.15056862  0.07791052  0.17738159  0.03457507 -0.03137766 -0.07460972
    0.21131412 -0.09816558 -0.23950338  0.06503582 -0.05302349 -0.18645067
   -0.10089075  0.09377487  0.05320942  0.18898928  0.33326881  0.19952209
    0.32758133  0.02778461  0.03563898  0.10596024 -0.23665847  0.07006399
    0.077985    0.13518884  0.04402691  0.18581121  0.22532205  0.2624724
    0.0260338  -0.13145028  0.26601344  0.13712016  0.00461333 -0.15809577
   -0.16098029 -0.06818671 -0.09647861  0.12909583 -0.13834235]]
```

PC1 captura el 45.6 % de la variación total entre los tres centroides.
PC2 añade otro 20.6 %.
En conjunto ~66.2 % de la variación morfológica se resume en estas dos dimensiones.

con dos ejes ya estamos explicando dos tercios de la heterogeneidad global, el tercio restante queda en PC3–PC43, donde probablemente están las diferencias más finas o el ruido.

PCA para visualización

PCA para visualización

```
import pandas as pd

# Construimos un DataFrame de loadings
loadings = pd.DataFrame(
    pca.components_.T,           # trasponemos: variables como filas
    index=df2.columns,           # nombre de cada variable
    columns=['PC1', 'PC2']       # etiqueta de cada componente
)

# Mostramos los 5 loadings (positivos o negativos) de mayor magnitud
top_PC1 = loadings['PC1'].abs().sort_values(ascending=False).head(5)
top_PC2 = loadings['PC2'].abs().sort_values(ascending=False).head(5)

print("Variables que más pesan en PC1:")
print(loadings.loc[top_PC1.index, 'PC1'])
print("\nVariables que más pesan en PC2:")
print(loadings.loc[top_PC2.index, 'PC2'])
```

PCA para visualización

```
Variables que más pesan en PC1:
C22    0.227818
C21    0.224124
C26    0.222939
C4     0.217967
C5     0.217426
Name: PC1, dtype: float64

Variables que más pesan en PC2:
C18    0.333269
C20    0.327581
C34    0.266013
C31    0.262472
C9    -0.239503
Name: PC2, dtype: float64
```

chequear qué variables son!!

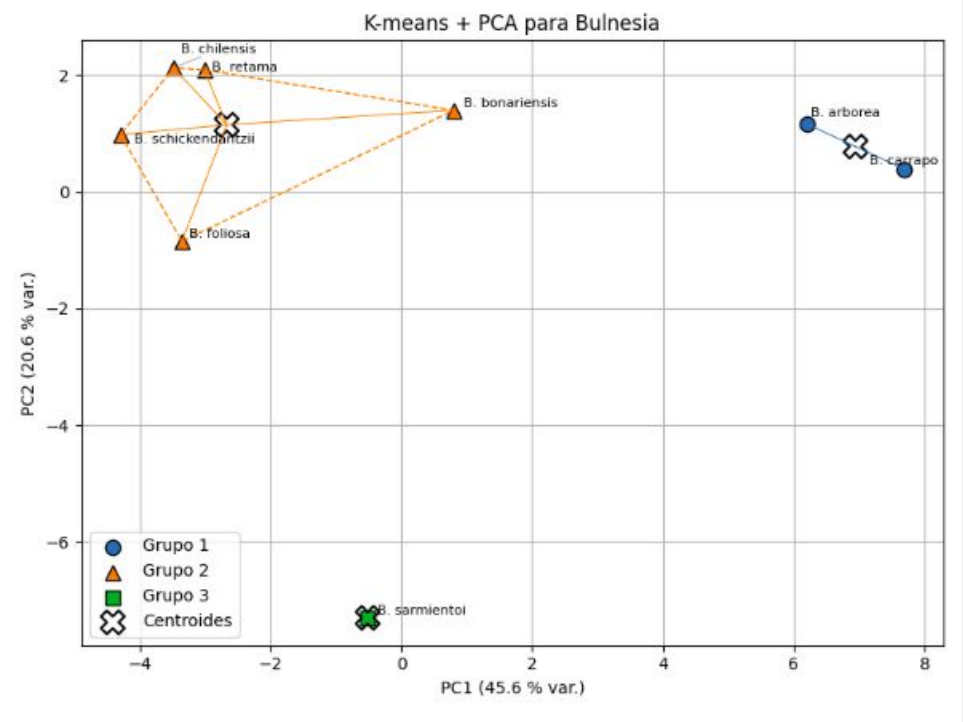
PCA para visualización

código para graficar !

Al combinar líneas estrella (densidad y outliers) con convex hull (contorno y solapamientos), el gráfico te permite en un solo vistazo:

- Cuán compactos son cada uno de los clusters.
- Dónde se ubican respecto a otros clusters.
- Qué puntos pueden requerir revisión (líneas muy largas o puntos fuera del hull).

PCA para visualización



B. sarmientoi se diferencia del resto especialmente en PC2

Comparar clústeres con filogenia o distribuciones geográficas: ¿coinciden estos grupos morfológicos con linajes evolutivos o áreas de endemismo?

Analizar rasgos discriminantes: usar ANOVA o tests no paramétricos para ver qué caracteres (de los 43) difieren más entre clusters.

Recordar que K-MEANS es sensible a la presencia de UE atípicas (con valores que se alejan mucho del resto de las UE) !!!!

scientific reports

Article | Open access | Published: 29 August 2022

Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

Eran Elhaik ✉

Scientific Reports **12**, Article number: 14683 (2022) | Cite this article

https://www.nature.com/articles/s41598-022-14395-4

Medidas de tendencia central

media (suma de todos los valores dividido por el número de frecuencias)
mediana (valor en el centro de la distribución)
moda (valor que se repite más veces (categoría con más frecuencias en una variable categórica) no tiene mucho sentido en variables numéricas continuas, si en discretas)

Medidas de dispersión

rango
varianza (poblacional o muestral)
desviación estándar
coeficiente de variación

$$\text{Rango} = x_{(\max)} - x_{(\min)}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma = \sqrt{\sigma^2} \quad , \quad s = \sqrt{s^2}$$

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Prueba de hipótesis

Test de Student (t-test): comparar medias cuando la población se asume normal y la varianza es desconocida.

Se plantea
H_0: “las medias son iguales” vs. H_A: “difieren”.

Se calcula t y su p-valor asociado según la distribución t con los grados de libertad.

Un p-valor bajo ($p < 0.05$) sugiere rechazar H_0

Prueba de hipótesis

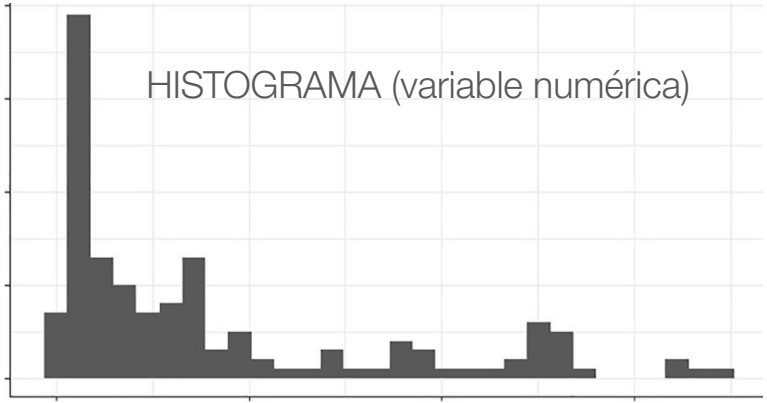
ANOVA (análisis de la varianza): extensión del t-test a más de dos grupos. Evalúa si al menos uno de los grupos difiere en su media.

Se plantea
H_0: “todas las medias de los grupos son iguales” vs. H_A: “al menos una difiere”.

Se calcula F y su p-valor en la tabla ANOVA.

Un p-valor bajo ($p < 0.05$) sugiere rechazar H_0 . Evidencia de diferencias de medias pero no sabes cuáles. Hacer comparaciones múltiples (Tukey, Bonferroni,...)

Representación gráfica



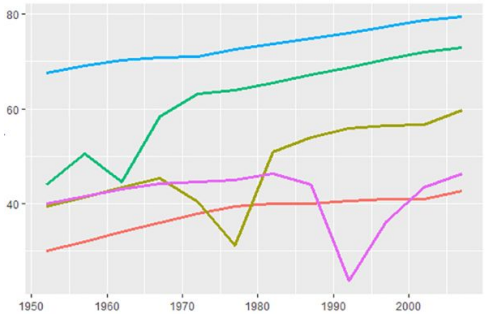


DIAGRAMA DE LÍNEAS
(variable numérica a través del tiempo)

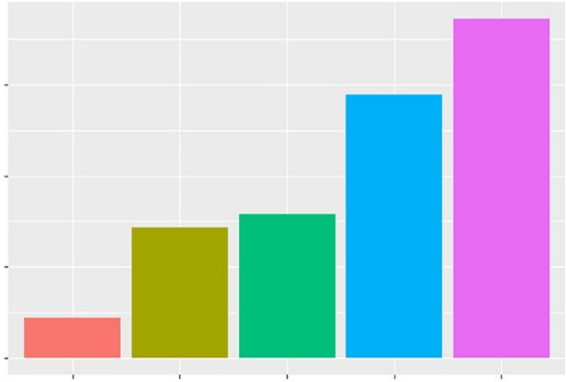


DIAGRAMA DE BARRAS
()

Representación gráfica

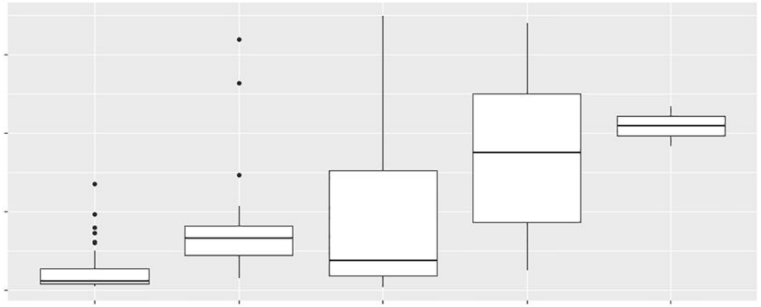


DIAGRAMA DE CAJAS
()

DIAGRAMA DE DISPERSIÓN
()

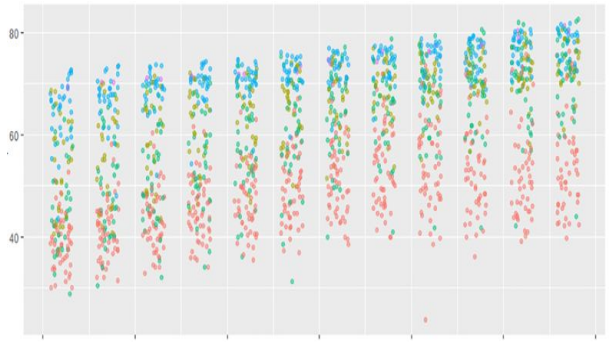


TABLA DE FRECUENCIAS

Estado	Frecuencia absoluta (f)	Frecuencia relativa (fr)	Frecuencia acumulada (Fa)	Frecuencia relativa acumulada (Fra)
Germinadas	16	0,80	16	0,80
No germinadas	4	0,20	20	1,00
Total	**20**	**1,00**	—	—

ANOVA

El **ANOVA de un factor** (one-way ANOVA) compara la **media** de una variable dependiente en tres o más grupos.

- **Variable dependiente**: la que mides (p. ej. altura).
- **Factor**: la categoría que define los grupos (p. ej. tipo de fertilizante A, B y C).

Lo que hace internamente es descomponer la **variabilidad total** de esa variable en:

1. **Variabilidad entre grupos** (debida al efecto del factor)
2. **Variabilidad dentro de los grupos** (ruido o error)

Y luego calcula un estadístico F = (Varianza entre grupos) / (Varianza dentro de grupos).

TABLA DE CONTINGENCIA

Resistencia ↓ \ Tratamiento →	Sin antibiótico	Con antibiótico	Total fila
Resistente (R)	40	10	50
Sensible (S)	10	40	50
Total columna	50	50	100

Para dos variables categóricas

distancia euclidiana

distancia “recta” o “en línea recta” entre dos puntos en un espacio euclidiano (espacio en el que se aplican las reglas de la geometría clásica) hipotenusa de un triángulo rectángulo formado entre las coordenadas de esos dos puntos.

Si tenés dos puntos:

- $A = (x_1, y_1)$
- $B = (x_2, y_2)$

La distancia euclidiana entre ellos es:

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Este es simplemente el **teorema de Pitágoras** aplicado a dos coordenadas.

En n dimensiones (espacio $\mathbb{R}^n$)

Si tenés dos vectores o puntos:

- $A = (a_1, a_2, ..., a_n)$
- $B = (b_1, b_2, ..., b_n)$

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$