1. EJEMPLOS DE DATOS MULTIVARIADOS

INTRODUCCIÓN

El **análisis multivariado** considera **varias variables aleatorias**, relacionadas entre si, **simultáneamente**. Cada variable se considera **igualmente importante** al iniciar el análisis.

Ej.1: Gorriones sobrevivientes a una tormenta

- Después de una tormenta severa el 1.Feb.1898, un conjunto de gorriones moribundos fueron llevados al Laboratorio de biología, Brown University, Rhode Island. Aproximadamente 50% de los pájaros murieron.
- Hermon Bumpus: Una oportunidad para estudiar el efecto de la selección natural en los pájaros. Contexto histórico: Teoría de Darwin sobre la selección natural.
- Experimento: Determinación de medidas morfológicas y peso de cada pájaro.

Ej.1: Gorriones sobrevivientes a una tormenta

- Medidas (en mm) del cuerpo de 49 pájaros hembra.
- Los pájaros 1 a 21 sobrevivieron, mientras que los restantes perecieron.

  X1 = longitud total, X2= extensión de las alas,
  X3 = longitud del pico y la cabeza, X4 = longitud del húmero, X5 = longitud de la quilla del esternón.

Pájaro	X1	X2	X3	X4	X5
1	156	245	31,6	18,5	20,5
2	154	240	30,4	17,9	19,6
3	153	240	31,0	18,4	20,6
4	153	236	30,9	17,7	20,2
5	155	243	31,5	18,6	20,3
6	163	247	32,0	19,0	20,9
7	157	238	30,9	18,4	20,2
8	155	239	32,8	18,6	21,2
9	164	248	32,7	19,1	21,1
10	158	238	31,0	18,8	22,0
11	158	240	31,3	18,6	22,0

Ej.1: Gorriones sobrevivientes a una tormenta

- Bumpus concluyó que "…los pájaros que perecieron, no lo hicieron por accidente, perecieron porque estaban fisicamente descalificados; y los pájaros que sobrevivieron, lo hicieron porque poseían ciertas características físicas."
- Específicamente: Los sobrevivientes "…eran más cortos y pesaban menos… sus huesos de las alas eran más largos, tenían piernas más largas y esternón más largo y mayor capacidad cerebral" que los no sobrevivientes.

Ej.1: Gorriones sobrevivientes a una tormenta

- Bumpus también concluyó que "el proceso de eliminación selectiva es más severo con los individuos extremadamente variables, sin importar en cual dirección ocurre la variación. Es tan peligroso estar claramente por encima de un cierto estándar de excelencia orgánica, como lo es estar visiblemente por debajo del estándar."
- **Interpretación:** Los individuos con medidas próximas a la media sobrevivieron mejor que los individuos con medidas que se apartan de la media.

Ej.1: Gorriones sobrevivientes a una tormenta

- **Contexto histórico:** El desarrollo de la **estadística multivariada** apenas comenzaba en 1898. El coeficiente de correlación entre dos variables, fue introducido por Francis Galton en 1877.
- Hotelling introduce el **análisis de componentes principales** en 1933, una de las técnicas multivariadas más simples aplicable a los datos de Bumpus.
- Bumpus ni siquiera calculó las desviaciones estándar. Sin embargo, sus métodos de análisis eran sensatos. Muchos autores han re-analizado sus datos y, en general, han confirmado sus conclusiones.

Ej.1: Gorriones sobrevivientes a una tormenta

Ilustración de las técnicas multivariadas:

1. ¿Cómo se relacionan las diferentes medidas? Ej.: Si una variable muestra valores altos ¿tienden éstos a ocurrir cuando las otras variables también muestran valores altos?
2. Sobrevivientes y no sobrevivientes: ¿Muestran las medias de sus variables diferencias significativas?
3. ¿Muestran los s y no-s similar variación en las variables medidas?
4. Si los s y no-s difieren con relación a la distribución de sus variables, ¿es posible construir alguna función de esas variables f(X1, X2, X3, X4, X5) que separe los dos grupos?

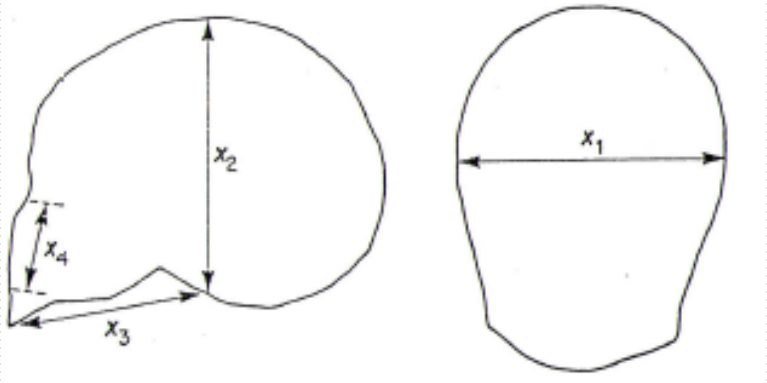
Ej. 2: Cráneos egipcios

- Mediciones realizadas en cráneos egipcios masculinos del área de Tebas, Egipto.

- 5 muestras de 30 cráneos cada una:
  - período predinastía temprano (4000 AC)
  - periodo predinastía tardío (3300 AC)
  - 12 y 13 dinastías (1850 AC)
  - período Ptolemaico (200 AC)
  - período romano (150 DC)

Ej. 2: Cráneos egipcios

Para cada cráneo se disponen 4 medidas:

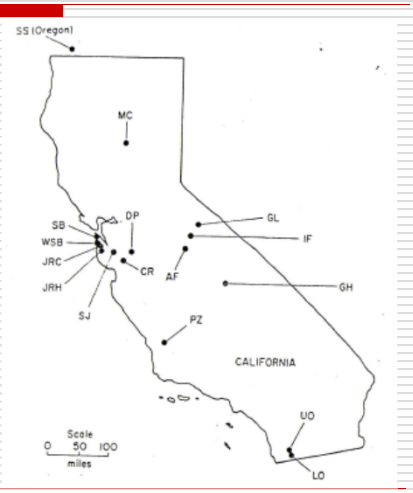


Ej. 2: Cráneos egipcios

1. ¿Cómo se relacionan las 4 medidas?
2. ¿Existen diferencias estadísticamente significativas en las **medias** de las muestras de las variables? ¿Reflejan esas diferencias cambios graduales con el tiempo, en el tamaño y la forma de los cráneos?
3. ¿Existen diferencias significativas en las **desviaciones estándares** de las muestras de las variables? ¿Reflejan estas diferencias cambios graduales con el tiempo?
4. ¿Es posible construir una función de las 4 variables, f(X1, X2, X3, X4), que describa los cambios en el tiempo?

Ej. 3: Distribución de una mariposa

Estudio de 16 colonias de la mariposa *Euphydryas editha* en California y Oregon.



Ej. 3: Distribución de una mariposa

Estudio de 16 colonias de la mariposa *Euphydryas editha* en California y Oregon.

Colony	Altitude (feet)	Annual Precipitation (inches)	Annual Max Temperature (ºF)	Annual Min Temperature (ºF)	Pgi movility gene frequencies (%)					
					0,40	0,60	0,80	1,00	1,16	1,30
SS	500	43	98	17	0	3	22	57	17	1
SB	800	20	92	32	0	16	20	38	13	13
WSB	570	28	98	26	0	6	28	46	17	3
JRC	550	28	98	26	0	4	19	47	27	3
JRH	550	28	98	26	0	1	8	50	35	6
SJ	380	15	99	28	0	2	19	44	32	3
CR	930	21	99	28	0	0	15	50	27	8
UO	650	10	101	27	10	21	40	25	4	0

- 4 variables ambientales (altitud, precipitación, temperatura mínima y máxima)
- 6 variables genéticas (porcentaje de frecuencias para diferentes genes Pgi determinados por electroforesis)

Ej. 3: Distribución de una mariposa

Las frecuencias describen, en cierto grado, la distribución genética de las mariposas.

- ¿Son similares las frecuencias Pgi para colonias próximas en el espacio?
- ¿Existe una relación entre las frecuencias Pgi y las variables ambientales? En caso afirmativo, ¿en que medida?

Estas interrogantes son importantes cuando se trata de decidir que variables influyen en las frecuencias Pgi.

Ej. 3: Distribución de una mariposa

- Si la composición genética de las colonias estuviera determinada por la migración presente y pasada, las frecuencias genéticas tenderán a ser similares para colonias próximas en el espacio, pero pueden no mostrar relación alguna con las variables ambientales.

- Si el ambiente es lo más importante, luego las frecuencias Pgi deberían estar relacionadas con las variables ambientales. Sin embargo, colonias próximas en el espacio tendrán diferentes frecuencias si el ambiente es diferente. Obvio que, las colonias próximas en el espacio tienden a tener ambientes similares. Por lo tanto, puede resultar dificultoso arribar a una conclusión clara.

Ej. 4: Perros prehistóricos de Tailandia

	X1	X2	X3	X4	X5	X6
Modern dog	9,7	21,0	19,4	7,7	32,0	36,5
Golden Jackal	8,1	16,7	18,3	7,0	30,3	32,9
Chinese wolf	13,5	27,3	26,8	10,6	41,9	48,1
Indian wolf	11,5	24,3	24,5	9,3	40,0	44,6
Cuon	10,7	23,5	21,4	8,5	28,8	37,6
Dingo	9,6	22,6	21,1	8,3	34,4	43,1
Prehistoric dog	10,3	22,1	19,1	8,1	32,3	35,0

Excavaciones en sitios prehistóricos, al noreste de Tailandia, permitieron obtener series de huesos de caninos, que cubren el período desde 3500 AC hasta el presente.

Para clarificar los ancestros del perro prehistórico, se realizaron medidas de la mandíbula inferior de las especies disponibles.

1. ¿Qué sugieren las mediciones sobre las relaciones entre los grupos?
2. ¿Cómo parece relacionarse el perro prehistórico con los otros grupos?

Ej. 5: Empleo en los países europeos

Table 1.5 Percentages of the Workforce Employed in Nine Different Industry Groups in 30 Countries in Europe

Country	Group	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7.0
France	EU	5.1	0.3	20.2	0.9	7.1	16.7	10.2	33.1	6.4
Germany	EU	3.2	0.7	24.8	1.0	9.4	17.2	9.6	28.4	5.6
Greece	EU	22.2	0.5	19.2	1.0	6.8	18.2	5.3	19.8	6.9
Ireland	EU	13.8	0.6	19.8	1.2	7.1	17.8	8.4	25.5	5.8
Italy	EU	8.4	1.1	21.9	0.0	9.1	21.6	4.6	28.0	5.3
Luxembourg	EU	3.3	0.1	19.6	0.7	9.9	21.2	8.7	29.6	6.8
Netherlands	EU	4.2	0.1	19.2	0.7	0.6	18.5	11.5	38.3	6.8

Porcentajes de fuerza laboral en 9 diferentes tipos de industrias, en 30 países europeos.

Objetivo del análisis multivariado: Aislar grupos de países con patrones de empleo similares. Ayudar a la comprensión de las relaciones entre los países. Ej.: agrupamiento político (UE, EFTA Zona de Libre Comercio Europea, países del Este) pueden ser de particular interés.

MÉTODOS MULTIVARIADOS

- Varias variables de interés

Ejemplos

- Obviamente, no independientes

- Análisis de Componentes Principales (ACP)
- Análisis de Factores (AF)
- Análisis de la Función Discriminante (AFD)
- Análisis de Conglomerados (clusters) (AC)
- Análisis de Correlaciones Canónicas (ACC)
- Escalado MultiDimensional (EMD)
- Ordenación:
  - Análisis de Coordenadas Principales (ACoP)
  - Análisis de correspondencias (Acorr)

Análisis de Componentes Principales (ACP)

- Reducir el número de variables que es necesario considerar a un número menor de **índices (componentes principales), que son combinaciones lineales de las variables originales**
- **Ejemplo 1.** La variación en la medida de los cuerpos de los gorriones estará relacionada con el tamaño general de los pájaros, y el total I_1 debería medirlo bastante bien:

$$I_1 = X_1 + X_2 + X_3 + X_4 + X_5$$

Otro índice: contraste entre las tres primeras medidas y las dos últimas:

$$I_2 = X_1 + X_2 + X_3 - X_4 - X_5$$

- El ACP proporciona una manera objetiva de encontrar índices de este tipo, de modo que la varianza de los datos pueda ser explicada lo más concisamente posible.

Análisis de Componentes Principales (ACP)

- La consideración de los valores de los componentes principales en lugar de los valores de las variables originales puede facilitar la comprensión de la información contenida en los datos. En resumen, **el ACP es un método que permite simplificar los datos, reduciendo el número de variables.**