UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Herramientas estadísticas para Big Data

Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

Máster **Big Data** Analytics

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

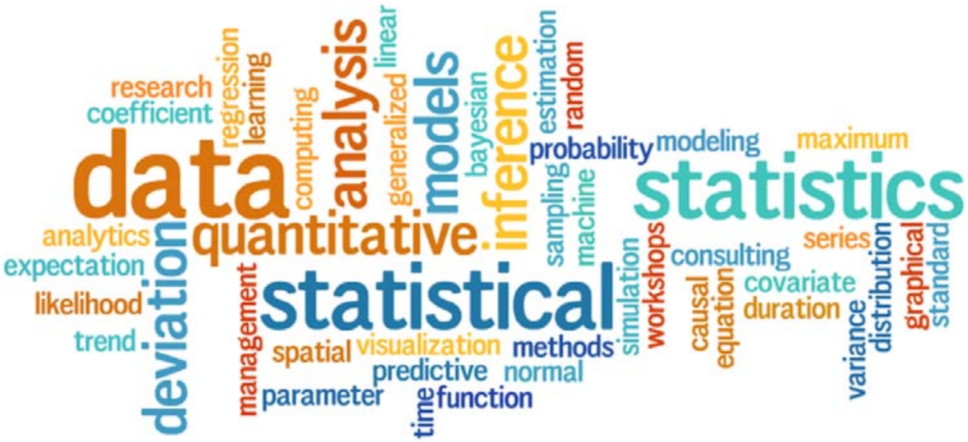
Valencia, Octubre 2017

Elena Vázquez

www.upv.es

bigdata.inf.upv.es

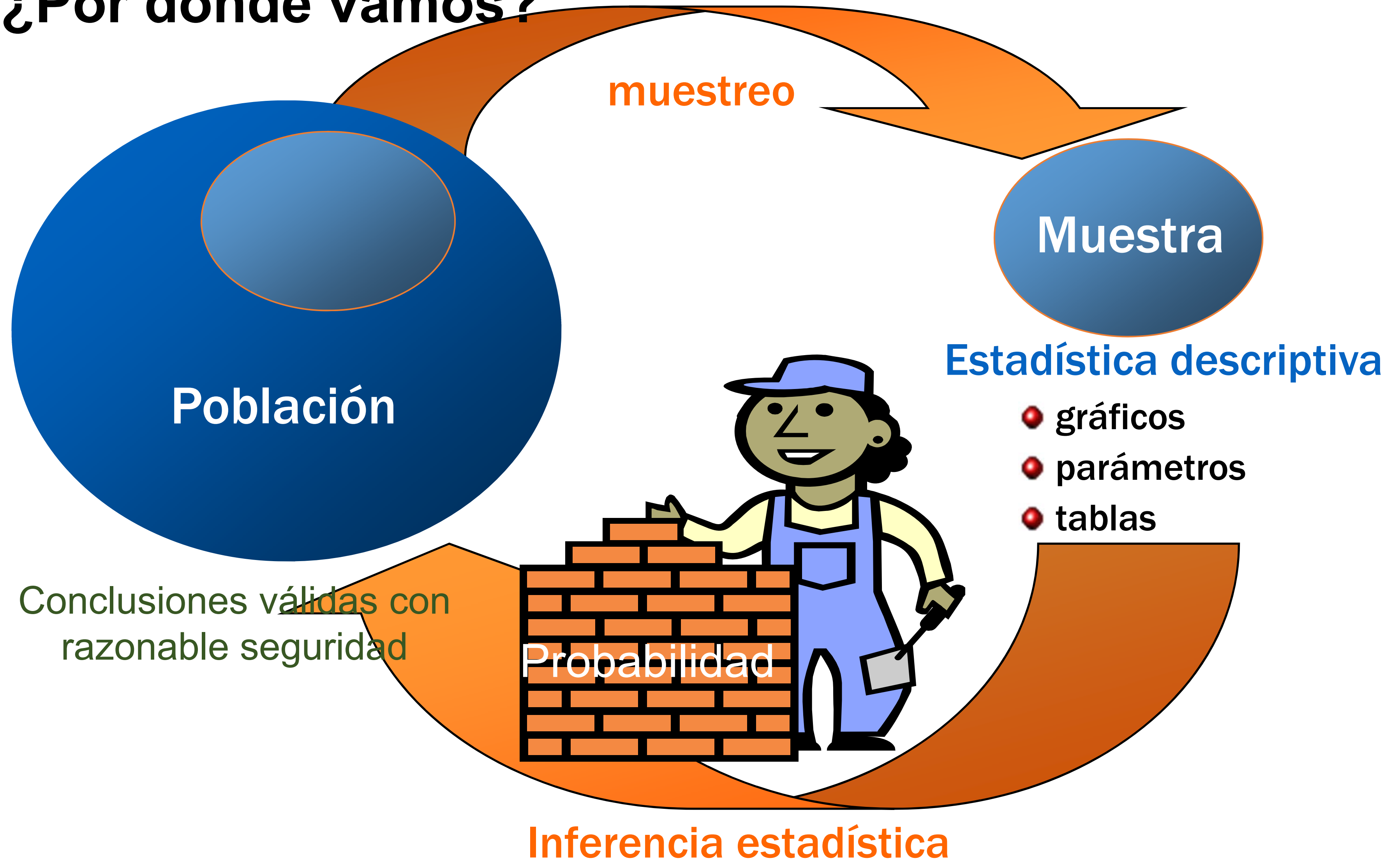
Contenidos



research
coefficient
regression
learning
computing
analysis
generalized
linear
models
bayesian
inference
estimation
random
probability
modeling
maximum
data
statistics
analytics
expectation
likelihood
trend
deviation
quantitative
sampling
machine
statistical
management
spatial
visualization
methods
predictive
normal
parameter
time
function
simulation
workshops
consulting
series
covariate
duration
causal
equation
variance
distribution
graphical
standard

2 Probabilidad y variables aleatorias

¿Por dónde vamos?



Sitios web para repasar… jugando

Proyecto Descartes

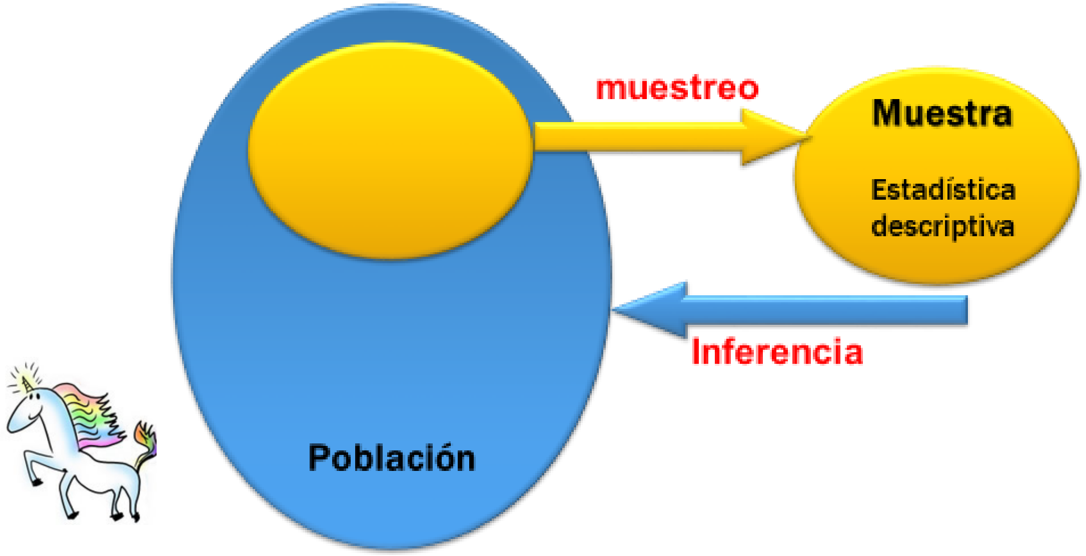
http://recursostic.educacion.es/descartes/web/materiales_didacticos/Azar_y_probabilidad/index.htm#intro

Proyecto ATICA

http://www.educa.jcyl.es/educacyl/cm/gallery/recursos_atica/IES%20Campo%20Charro/portada.swf

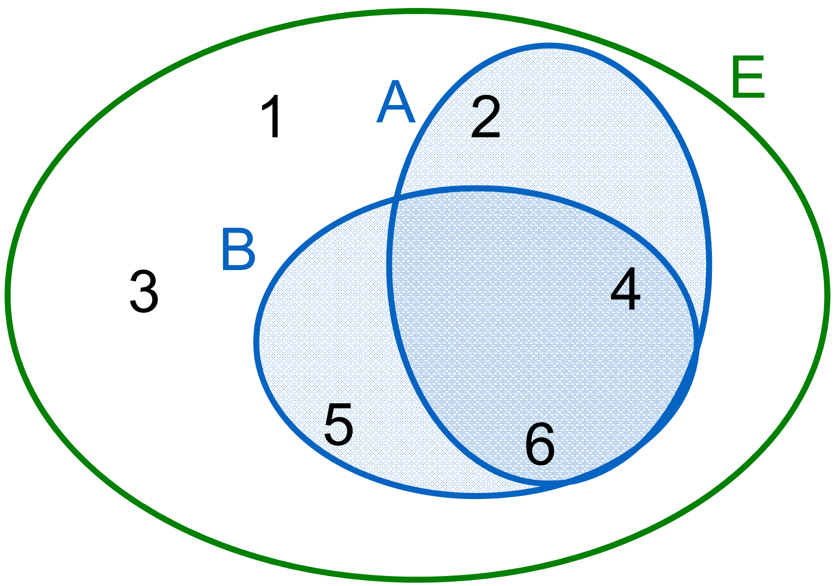
Sucesos y operaciones con sucesos

- **Población**: conjunto de individuos acerca de los cuales se quiere obtener información o inferir conclusiones.
- **Variable aleatoria**: valores que toma la característica a estudiar de los individuos de la población.
- **E** (Espacio muestral): conjunto de valores que puede tomar una determinada variable aleatoria
- **A** (**Suceso**): cualquier subconjunto **A** de **E**



Ejemplo: Lanzamientos de un dado

- **Población** = {todos los lanzamientos que se puedan efectuar con el dado}
- Variable aleatoria: nº de puntos al lanzar el dado
- **E** (Espacio muestral) = {1, 2, 3, 4, 5, 6}
  - **Suceso A**: sacar un número **par**
  - **Suceso B** : sacar un número **>3**



Tipos de sucesos

- **Suceso seguro :** asociado a **E**, que es un subconjunto de sí mismo. Para todos los individuos de la población se verifica dicho suceso.
- **Suceso imposible**: es el asociado al subconjunto vacío Φ de E.
- Suma o **unión de sucesos** A y B : $\mathbf{C = A \cup B}$
- **Producto o intersección de sucesos A y B : $\mathbf{A \cap B}$**
- **Sucesos excluyentes:** son aquellos cuya intersección es el suceso imposible Φ.
- **Suceso contrario o complementario** ($\bar{A}$) a uno dado **A**, es aquél que se verifica si, y sólo si, no se verifica **A**

Probabilidad: concepto, cálculo y propiedades

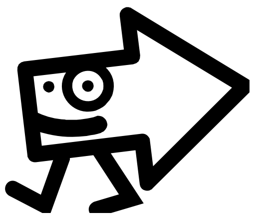
- A todo suceso **A** se le puede asociar un número comprendido entre **0** y **1** al que se denomina **probabilidad** de dicho suceso, y se le representa por **P(A)**

- **Interpretación intuitiva**: la probabilidad de un suceso es la proporción de individuos de la población considerada en los que se verifica dicho suceso.

Muestra

Frecuencia relativa (%)

% de veces que sale un 5 al lanzar el dado



Población

Probabilidad
(tanto por uno)

P(A) =
P(Obtener un nº=5)

Probabilidad: Propiedades

1. **0** ≤ P(**A**) ≤ **1**
2. P(**E**) = 1 → Probabilidad de sacar un 1, 2, 3, 4, 5 o 6
3. P(No-A) = 1 - P(A) → Prob. de NO sacar un 6 = 1 – Prob de sacar un 6
4. P(Φ) = 0 → Probabilidad de sacar un 9

$P(A \cup B)$

El caso general unión de 2 sucesos es:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

Si los sucesos son excluyentes :

$$\boxed{P(A \cup B) = P(A) + P(B) - \cancel{P(A \cap B)}} \qquad \boxed{P(A \cap B) = 0}$$

El caso general probabilidad de la unión de 3 sucesos

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)}$$

Esta expresión se puede generalizar para la suma de k sucesos

Independencia de sucesos

Intuitivamente:

- A el hecho de que se verifique **B** no afecta, en absoluto, a que se verifique **A**.
- Que se cumpla **B** no modifica la probabilidad de **A**

Cuando 2 sucesos son independientes:

$$\boxed{P(A \cap B) = P(A).P(B)}$$

Ejemplo sucesos NO independientes

Lanzamos un dado simétrico y sean los sucesos:

A = obtener un número **par**

B = obtener un número **mayor que 3**

A y B ¿Son independientes?

Si se trata de adivinar,

SÍ

¿el saber que en una tirada ha salido un número mayor que 3, aporta información sobre la probabilidad de que el número sea par?

$P(A) \times P(B) = 1/2 \times 1/2 = 1/4 \neq P(A \cap B) = 1/3$

$\Downarrow$

NO son Independientes

Ejemplo sucesos independientes

Sacamos una carta al azar de una baraja española (40 cartas, 10 de cada palo) y sean los sucesos:

A = sacar un **as** **B** = sacar un **oro**

A y B ¿Son independientes?

Si se trata de adivinar,

¿el saber que ha salido un as,

aporta información sobre la probabilidad de que la carta sea de oro?

NO

$P(A) \times P(B) = 4/40 \times 10/40 = 1/40 \quad = \quad P(A \cap B) = 1/40$

$\Downarrow$

SÍ son Independientes

Probabilidad Condicional

Dados dos sucesos **A** y **B**, decimos que la

Probabilidad de A condicionado a B

$$P(A/B)$$

es la probabilidad de que se haya presentado el suceso **A** sabiendo que se ha presentado el suceso **B**

Se calcula:

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

Ejemplo probabilidad condicional

- **Población** = {todos los lanzamientos de un dado}
- **E** (Espacio muestral) = {1, 2, 3, 4, 5, 6}
  - **Suceso A**: sacar un número **par**
  - **Suceso B** : sacar un número **>3**

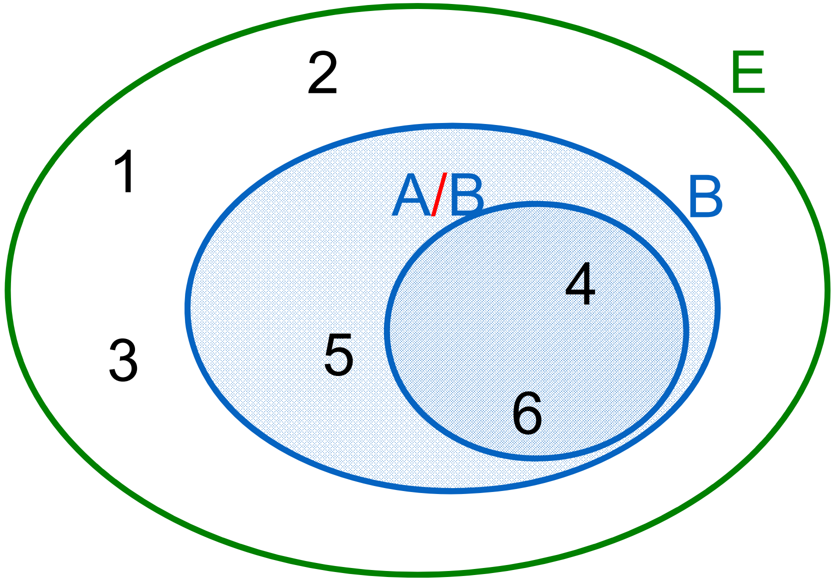
¿ P(A/B) ?

Si es >3 → 3 valores posibles.

De esos 3 valores,

hay 2 que son par

P(A/B) = 2/3



Prob Condicional e Intersección de sucesos

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

¿ $P(A \cap B)$?

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

¿ $P(A \cap B)$?

Caso general para obtener la probabilidad del producto de 2 sucesos:

$$P(A \cap B) = P(A).P(B/A)$$

$$P(A \cap B) = P(B).P(A/B)$$

Caso particular, cuando los sucesos son independientes:

$$P(A \cap B) = P(A).P(B)$$

Otras equivalencias sucesos independientes

Dos sucesos **A** y **B** son **independientes** si se verifica:

$$P(A/B) = P(A)$$

$$P(B/A) = P(B)$$

$$P(A \cap B) = P(A) \times P(B)$$

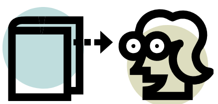
$$P(A/B) = P(A/\bar{B})$$

$$P(B/A) = P(B/\bar{A})$$

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B})$$

$$P(\bar{A} \cap B) = P(\bar{A}) \times P(B)$$

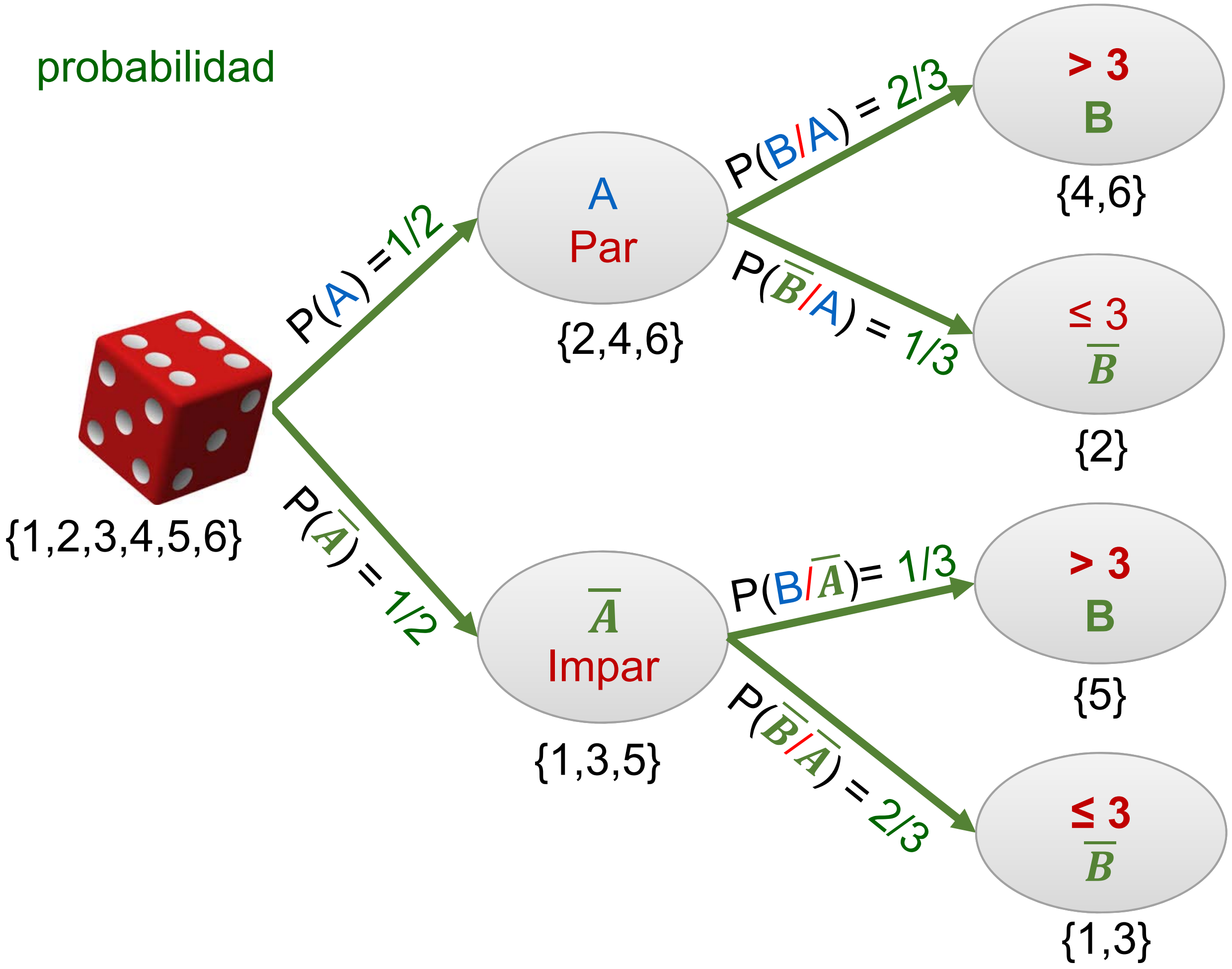
$$P(A \cap \bar{B}) = P(A) \times P(\bar{B})$$



Puedes usar el árbol de probabilidades

Árbol de probabilidad

probabilidad



P(A∩B) = P(B/A) P(A) =
2/3 x 1/2 = 2/6

P(A∩$\bar{B}$) = P($\bar{B}$/A) P(A) =
1/3 x 1/2 = 1/6

P($\bar{A}$∩ B) = P(B/$\bar{A}$)P($\bar{A}$) =
1/3 x 1/2 = 1/6

P($\bar{A}$∩$\bar{B}$) = P($\bar{A}$/$\bar{B}$)P($\bar{A}$) =
2/3 x 1/2 = 2/6

Teorema de Bayes y de la Probabilidad Total

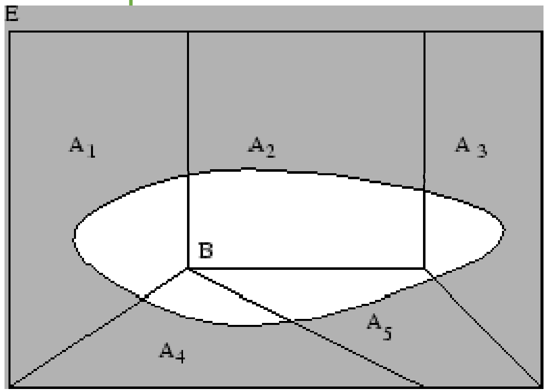
- **Determinación de la probabilidad de las causas a partir de los efectos que han podido ser observados**.

- En el año 1763, dos años después de la muerte de **Thomas Bayes** (1702-1761), se publicó una memoria en la que aparece, por vez primera, cómo obtener estas probabilidades.

- **El cálculo de dichas probabilidades recibe el nombre de teorema de Bayes**.

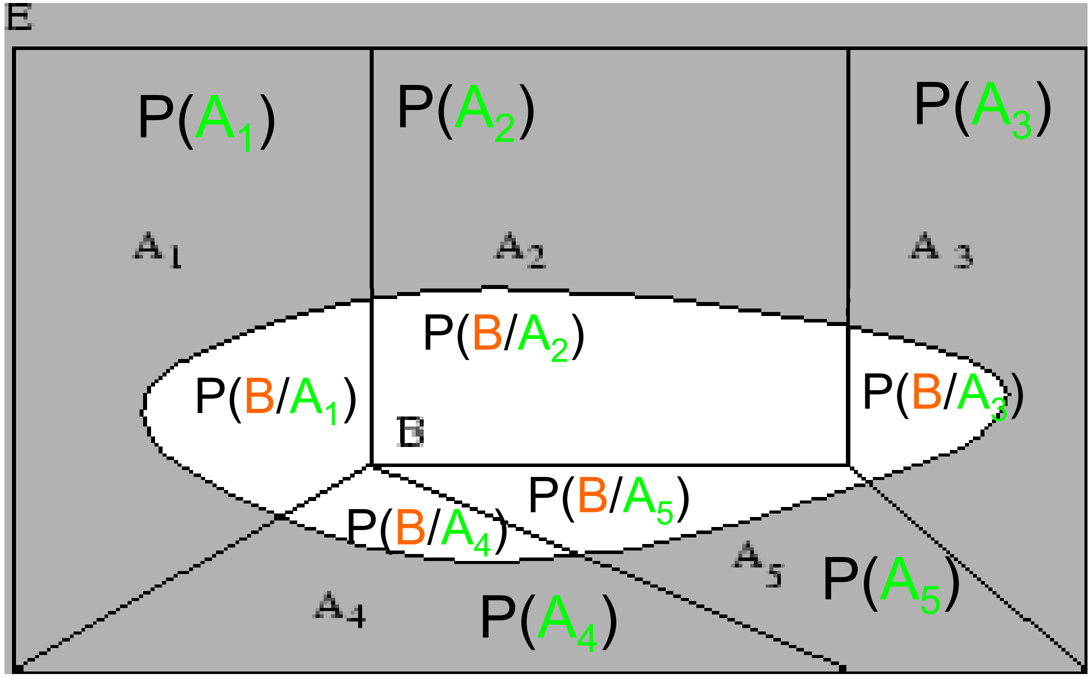
Teorema de la Probabilidad Total

A_i mutuamente



$P(A_i)$ $P(B/A_i)$

¿$P(B)$?



$$P(B) = P(A_1)\,P(B/A_1) + \ldots + P(A_n)\,P(B/A_n)$$

Teorema de la Probabilidad Total

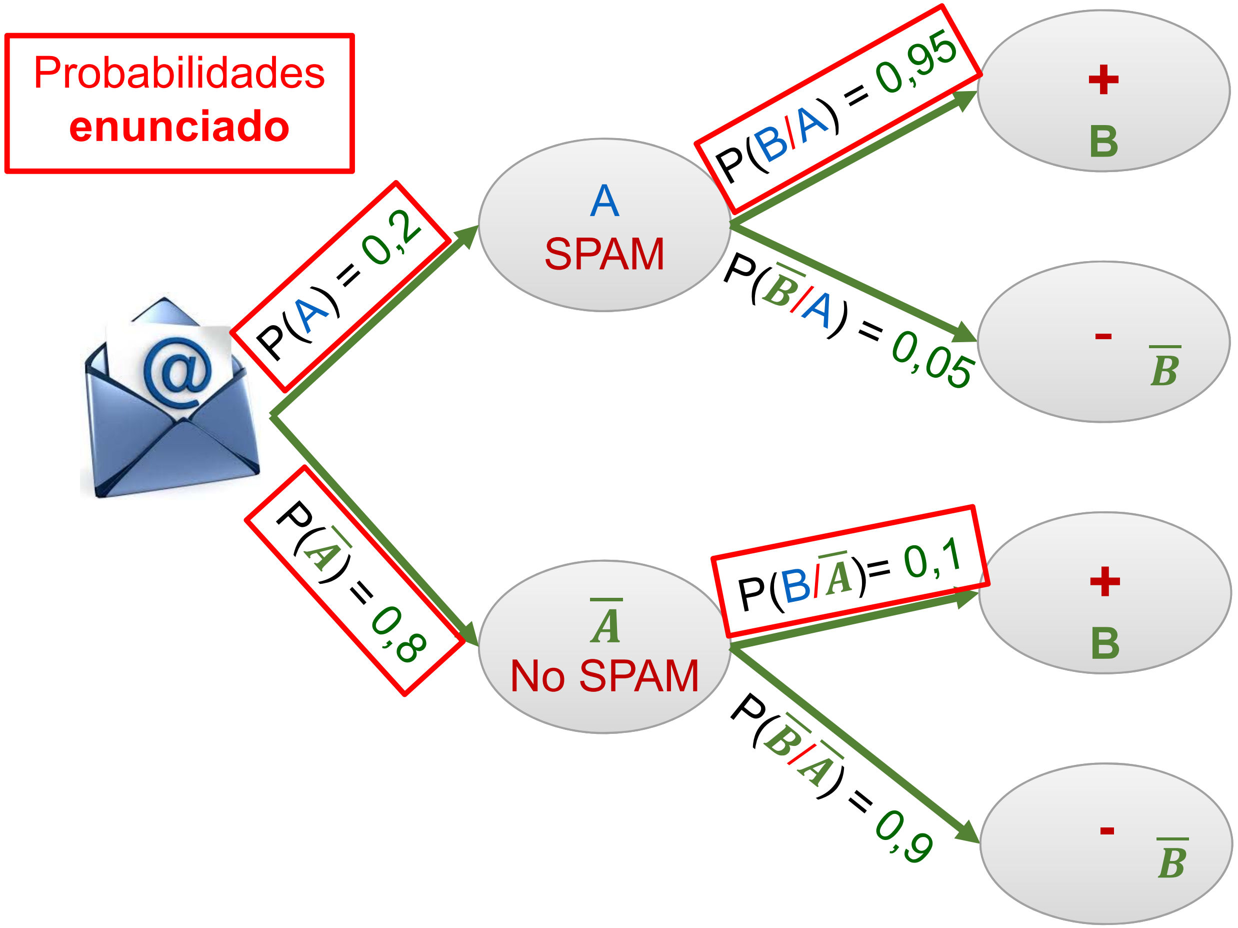
En un servidor de correo electrónico, el 20% de los correos resultan ser SPAM, mientras que el otro 80% no lo es.

Para discernir entre ambas situaciones se instala un filtro que puede dar positivo en SPAM o negativo. Se sabe que la probabilidad de que el filtro resulte positivo es 0,95 cuando los correos son SPAM y 0,10 cuando los correos no lo son.

a) Elegido, al azar, un correo recibido, ¿cuál es la probabilidad de que al pasar el filtro de positivo?

b) Sabiendo que en un correo el filtro lo ha catalogado como SPAM, ¿cuál es la probabilidad de que sea realmente un correo SPAM?

Teorema de la Probabilidad Total



P(A∩B) = P(B/A) P(A) =
0,95 x 0,2 = 0,19

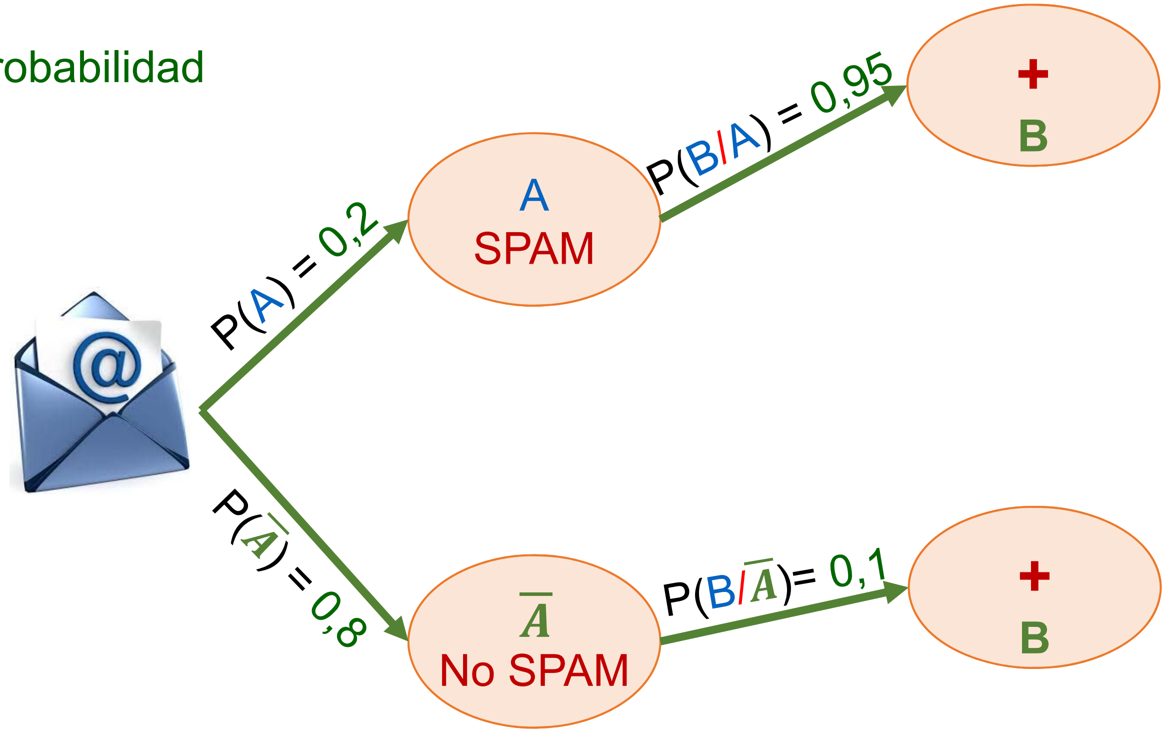
P(A∩$\overline{B}$) = P($\overline{B}$/A) P(A) =
0,05 x 0,2 = 0,01

P($\overline{A}$∩ B) = P(B/$\overline{A}$)P($\overline{A}$) =
0,1 x 0,8 = 0,08

P($\overline{A}$∩$\overline{B}$) = P($\overline{A}$/$\overline{B}$)P($\overline{A}$) =
0,9 x 0,8 = 0,72

Teorema de la Probabilidad Total

probabilidad



$P(A \cap B) = P(B/A)\,P(A) =$
$0{,}95 \times 0{,}2 = 0{,}19$

$P(\bar{A} \cap B) = P(B/\bar{A})P(\bar{A}) =$
$0{,}1 \times 0{,}8 = 0{,}08$

$\underline{P(B)} = P((A \cap B) \cup P(\bar{A} \cap B)) = P(A \cap B) + P(\bar{A} \cap B) =$
$P(B/A)P(A) + P(B/\bar{A})P(\bar{A}) = 0{,}95 \times 0{,}2 + 0{,}1 \times 0{,}8 = \underline{0{,}27}$

Teorema de la probabilidad total

Ejemplo Teorema de Bayes

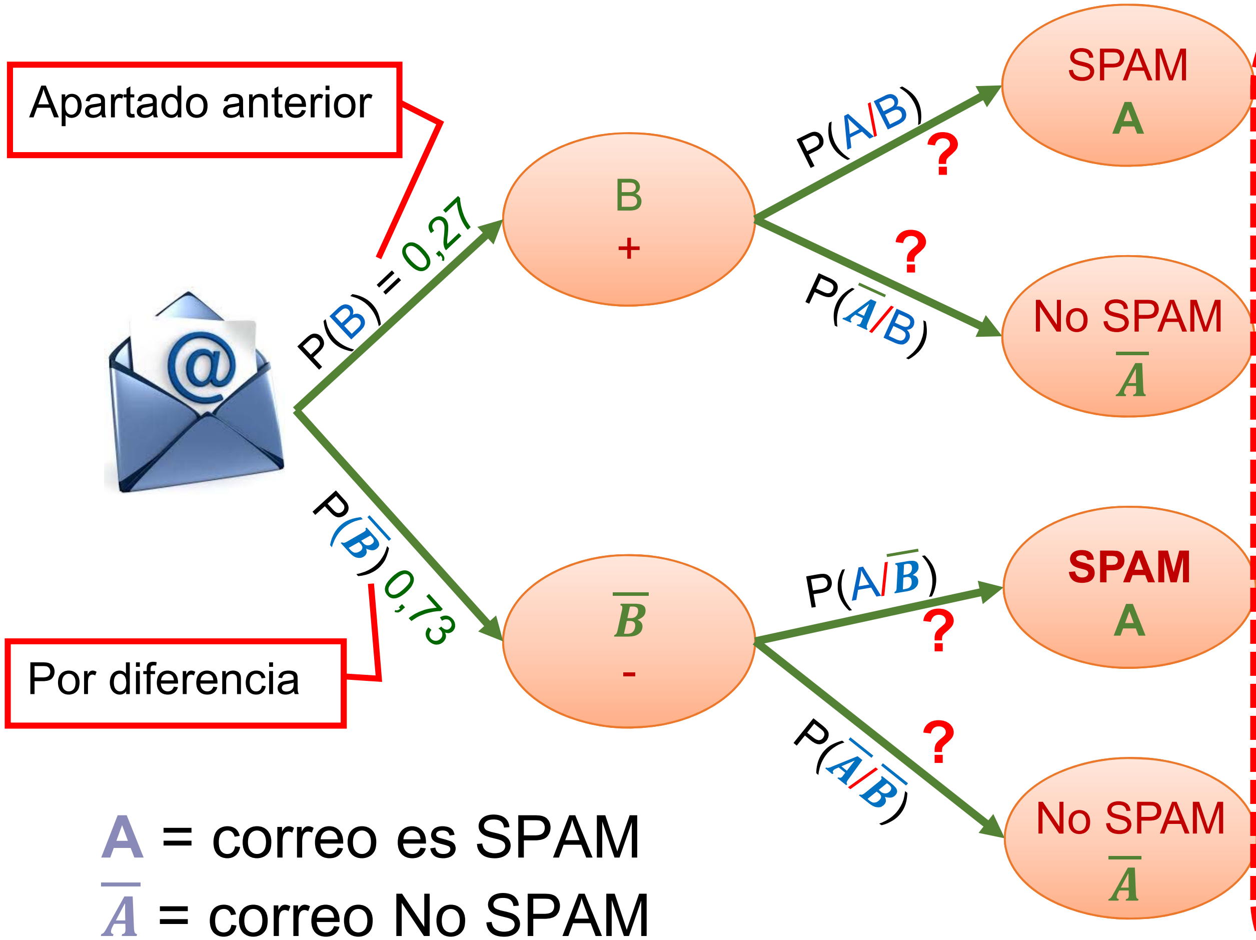
En un servidor de correo electrónico, el 20% de los correos resultan ser SPAM, mientras que el otro 80% no lo es.

Para discernir entre ambas situaciones se instala un filtro que puede dar positivo en SPAM o negativo. Se sabe que la probabilidad de que el filtro resulte positivo es 0,95 cuando los correos son SPAM y 0,10 cuando los correos no lo son.

a) Elegido, al azar, un correo recibido, ¿cuál es la probabilidad de que al pasar el filtro de positivo?

b) Sabiendo que en un correo el filtro ha dado positivo, ¿cuál es la probabilidad de que sea realmente un correo SPAM?

Ejemplo Teorema de Bayes



$P(A \cap B) = P(A/B)\, P(B) =$ 0,19

$P(\overline{A} \cap B) = P(\overline{A}/B) P(B) =$ 0,08

$P(A \cap \overline{B}) = P(A/\overline{B})\, P(\overline{B}) =$ 0,01

$P(\overline{A} \cap \overline{B}) = P(\overline{A}/\overline{B}) P(\overline{B}) =$ 0,72

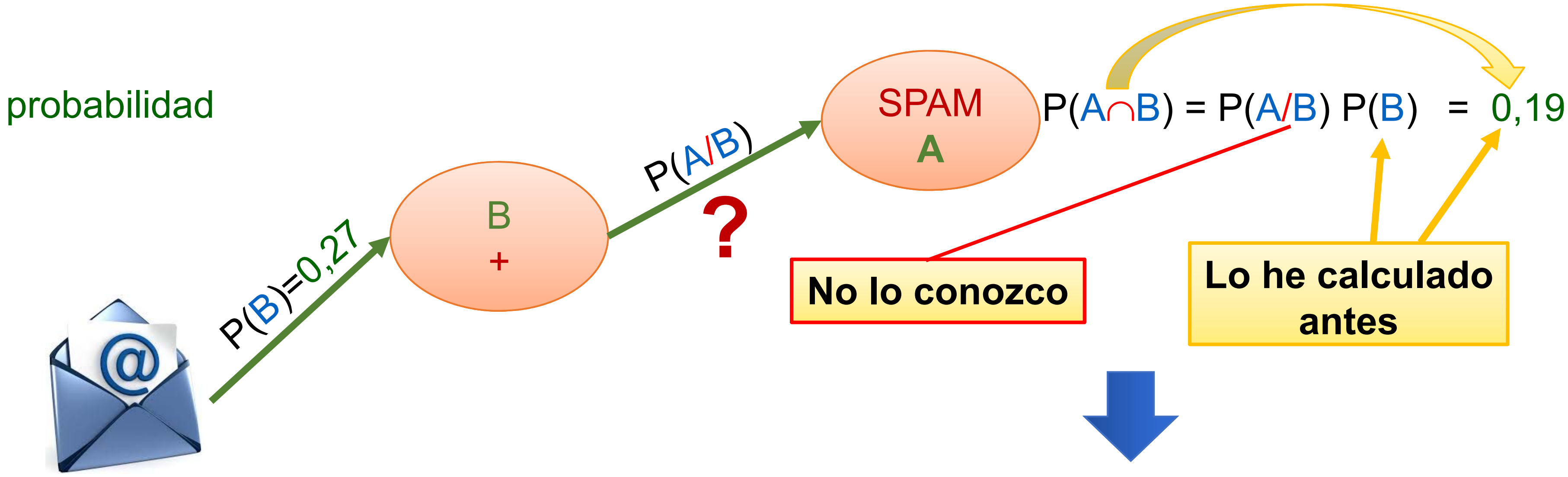
Árbol anterior

A = correo es SPAM

$\overline{A}$ = correo No SPAM

B = filtro da positivo

Ejemplo Teorema de Bayes



$$\underline{P(A/B) = P(A \cap B)/P(B)} = 0{,}19 \,/\, 0{,}27 = \underline{0{,}703}$$

$$\underline{P(A/B)} = P(A \cap B)/P(B) \underline{= P(B/A)\ P(A)/P(B)} =$$

$$0{,}95 \times 0{,}2 \,/\, 0{,}27 = \underline{0{,}703}$$

A = correo es SPAM

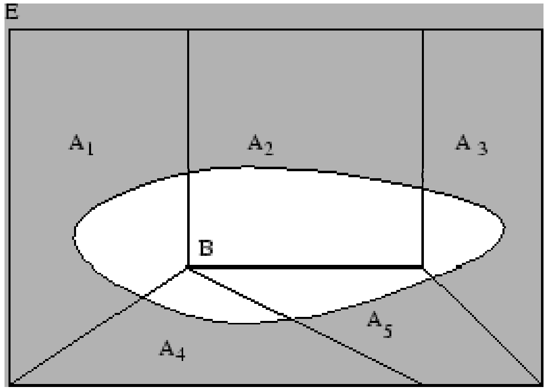
$\overline{A}$ = correo No SPAM

B = filtro da positivo

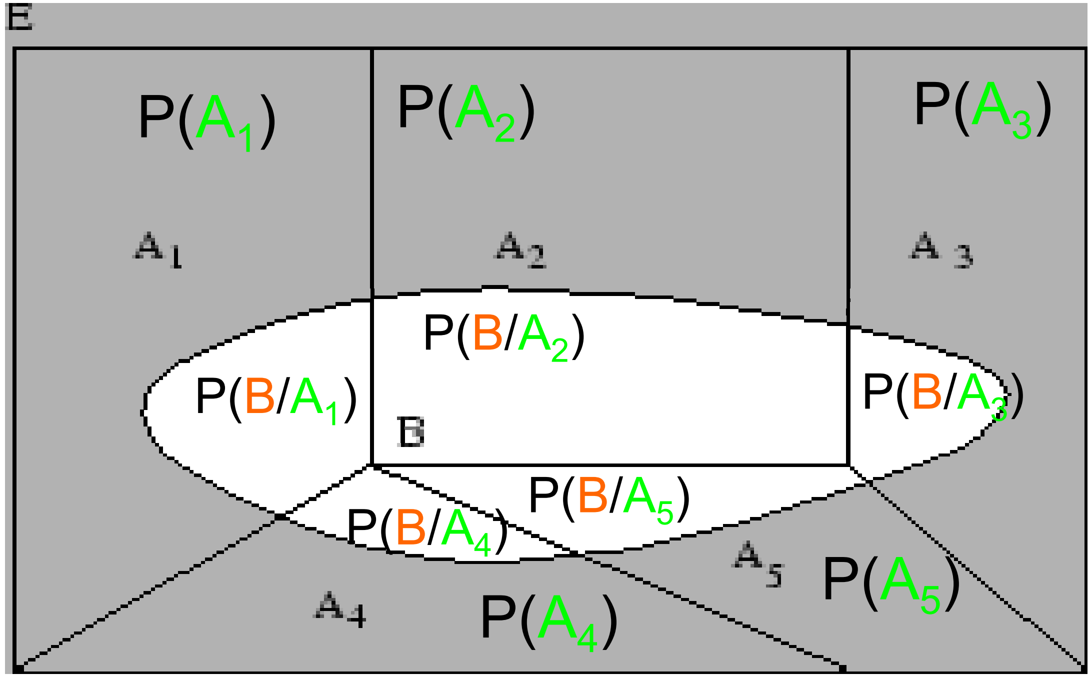
Teorema de Bayes

A_i mutuamente excluyentes

B asociado a todos ellos



$P(A_i)$ $\quad$ $P(B/A_i)$



¿$P(A_i/B)$?

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B/A_i)}{\sum_{j=1}^{N} P(A_j)P(B/A_j)}$$

Teorema de la Probabilidad Total

Glosario

Árbol de probabilidad
Casos favorables
Casos posibles
Condicional (probabilidad)
Contrarios o complementarios (sucesos)
Espacio muestral
Excluyentes (sucesos)
Imposible (suceso)
Independientes (sucesos)
Probabilidad
Producto o intersección (sucesos)
Regla de Laplace
Seguro (suceso)
Suceso
Suma o unión (sucesos)
Teorema de Bayes
Teorema Probabilidad Total

Herramientas Estadísticas para Big Data
Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

2- Probabilidad

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

www.upv.es

E. Vázquez
Dto. De Estadística e Investigación Operativa, Aplicadas y Calidad