Probabilidad

¿Cuál es la probabilidad de aprobar Minería de Datos?

¿Cuál es la probabilidad de no encontrarme un trancon cuando voy a clase?

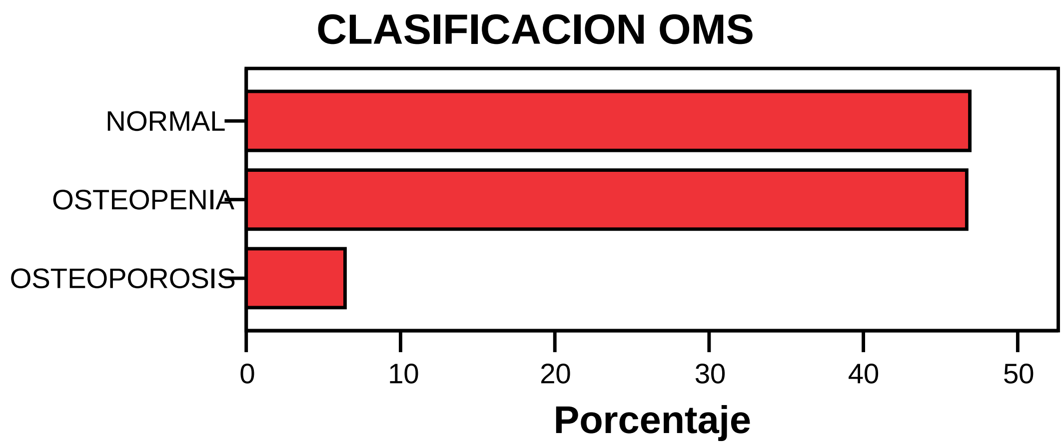
Todos los días nos hacemos preguntas sobre probabilidad e incluso los que hayáis visto poco de la materia en cursos anteriores, tenéis una idea intuitiva lo suficientemente correcta para lo que necesitamos de ella en este curso.

Nociones de probabilidad

- Frecuentista (objetiva): Probabilidad de un suceso es la frecuencia relativa (%) de veces que ocurriría el suceso al realizar un experimento repetidas veces.

CLASIFICACION OMS

		Frecuencia	Porcentaje
Válidos	NORMAL	469	46,9%
	OSTEOPENIA	467	46,7%
	OSTEOPOROSIS	64	6,4%
	Total	1000	100,0



- Subjetiva (bayesiana): Grado de certeza que se posee sobre un suceso. Es personal.

En ambos tipos de definiciones aparece el concepto de suceso. Vamos a ver qué son y algunas operaciones que se pueden realizar con sucesos.

Sucesos

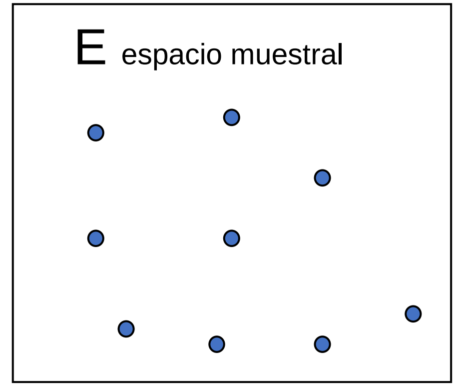
- Cuando se realiza un experimento aleatorio diversos resultados son posibles. El conjunto de todos los resultados posibles se llama espacio muestral (E).

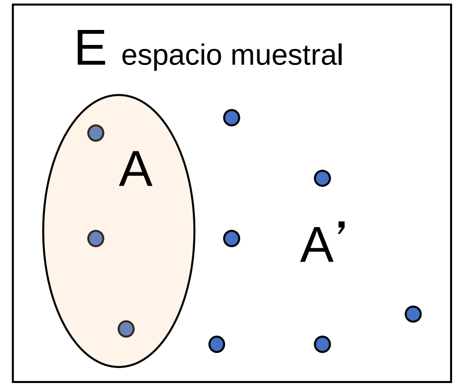
- Se llama suceso a un subconjunto de dichos resultados.

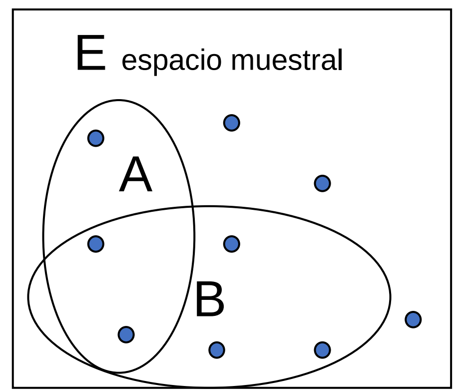
- Se llama suceso contrario (complementario) de un suceso A, A', al formado por los elementos que no están en A

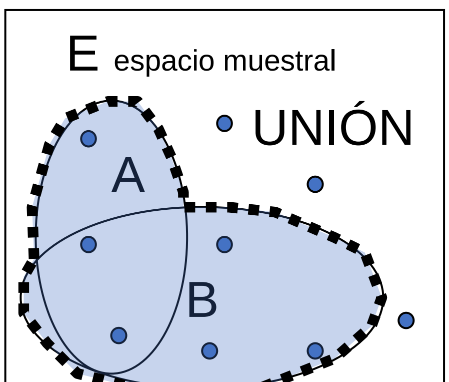
- Se llama suceso unión de A y B, AUB, al formado por los resultados experimentales que están en A o en B (incluyendo los que están en ambos.

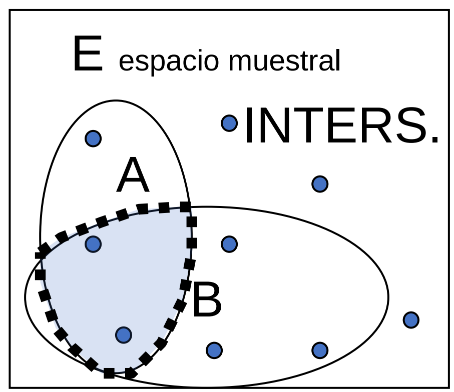
- Se llama suceso intersección de A y B, A∩B o simplemente AB, al formado por los elementos que están en A y B







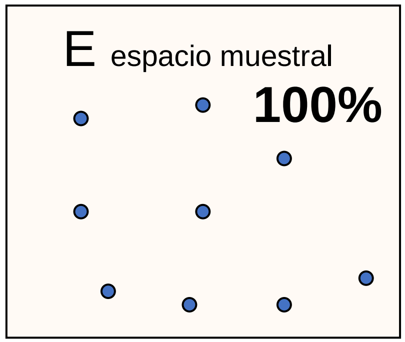




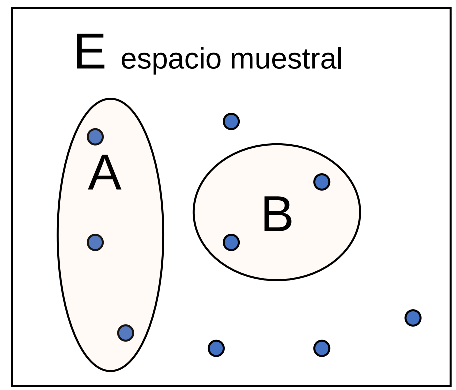
Definición de probabilidad

- Se llama probabilidad a cualquier función, P, que asigna a cada suceso A un valor numérico P(A), verificando las siguientes reglas (axiomas)

  - $P(E)=1$

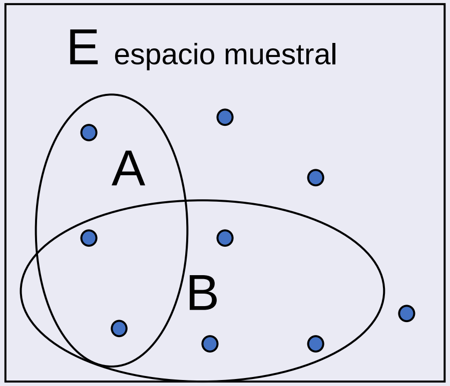


  - $0 \leq P(A) \leq 1$
  - $P(A \cup B)=P(A)+P(B)$ si $A \cap B=\emptyset$
    - $\emptyset$ es el conjunto vacío.

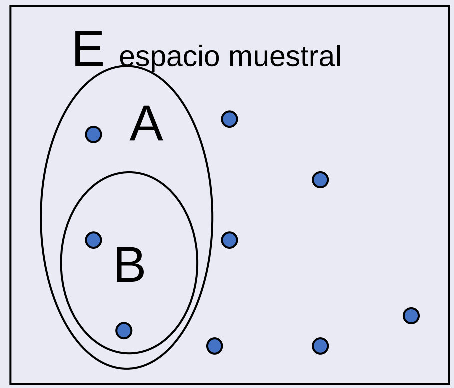


- Podéis imaginar la probabilidad de un subconjunto como el tamaño relativo con respecto al total (suceso seguro)

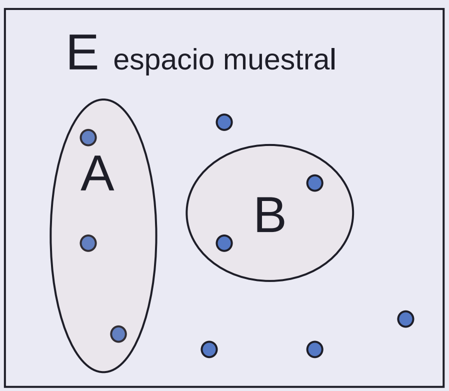
EJEMPLOS



P(A)=?
P(B)=?
P(AUB)=?
P(AB)=?
P(A')=?
P(B')=?



P(A)=?
P(B)=?
P(AUB)=?
P(AB)=?
P(A')=?
P(B')=?



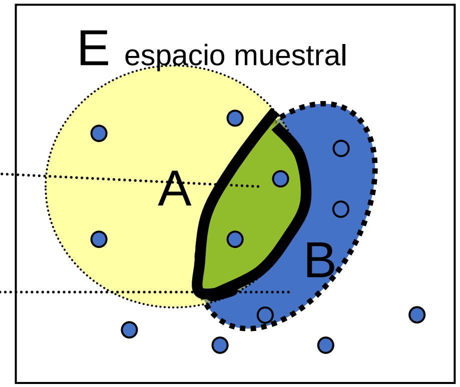
P(A)=?
P(B)=?
P(AUB)=?
P(AB)=?
P(A')=?
P(B')=?

Probabilidad condicionada

- Se llama probabilidad de A condicionada a B, o probabilidad de A sabiendo que pasa B:

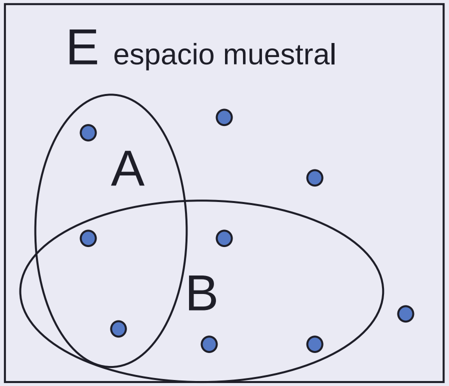
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

"tamaño" de uno respecto al otro



- Error frecuentíiiiiiisimo:
  - No confundir probabilidad condicionada con intersección.
  - En ambos medimos efectivamente la intersección, pero…
    - En P(A∩B) con respecto a P(E)=1
    - En P(A|B) con respecto a P(B)

EJEMPLOS



P(A)=3/9=1/3

P(B)=5/9

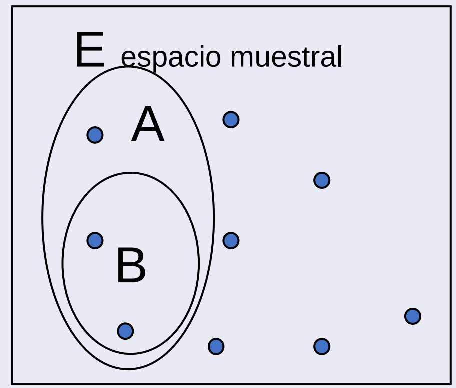
P(AUB)=6/9=2/3

P(AB)=2/9

P(A’)=6/9=2/3

P(B’)=4/9

P(A|B)=? P(B|A)=?



P(A)=3/9=1/3

P(B)=2/9

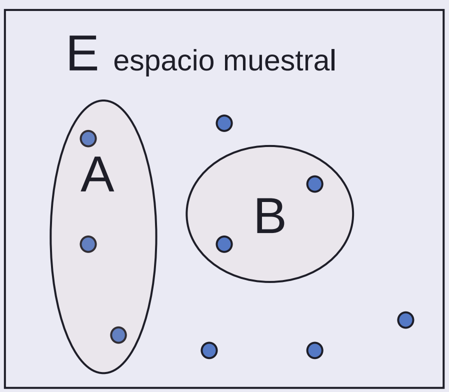
P(AUB)=3/9=1/3

P(AB)=2/9

P(A’)=6/9=2/3

P(B’)=7/9

P(A|B)=? P(B|A)=?



P(A)=3/9=1/3

P(B)=2/9

P(AUB)=5/9

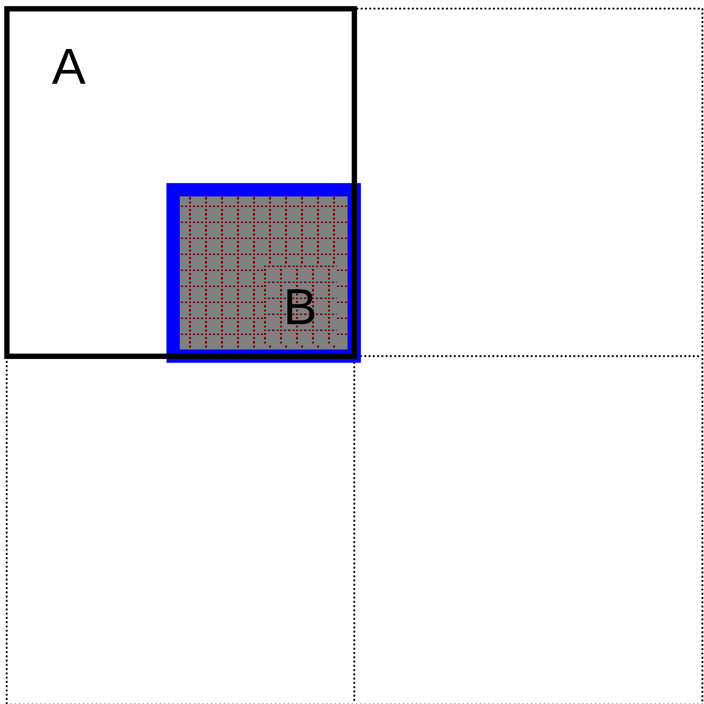
P(AB)=0

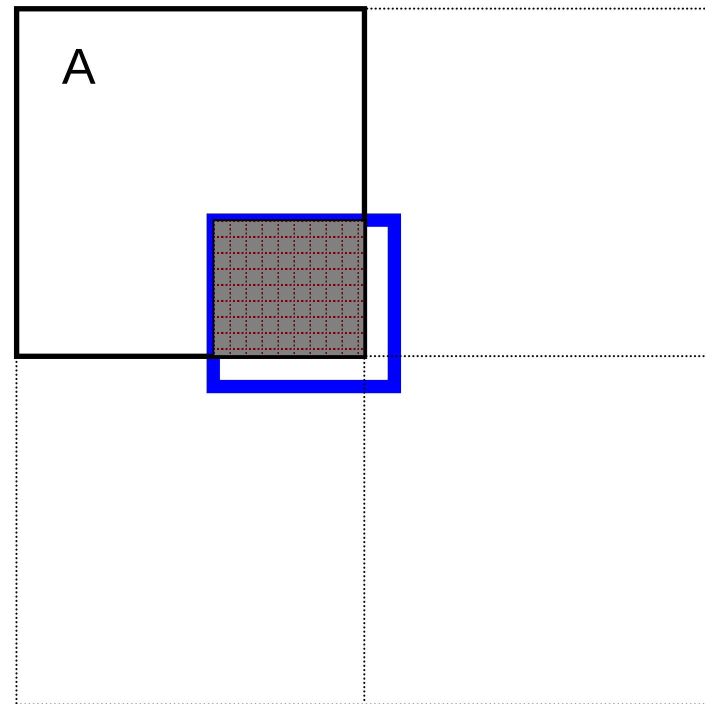
P(A’)=6/9=2/3

P(B’)=7/9

P(A|B)=? P(B|A)=?

Intuir la probabilidad condicionada





$P(A) = 0{,}25$
$P(B) = 0{,}10$
$P(A \cap B) = 0{,}10$

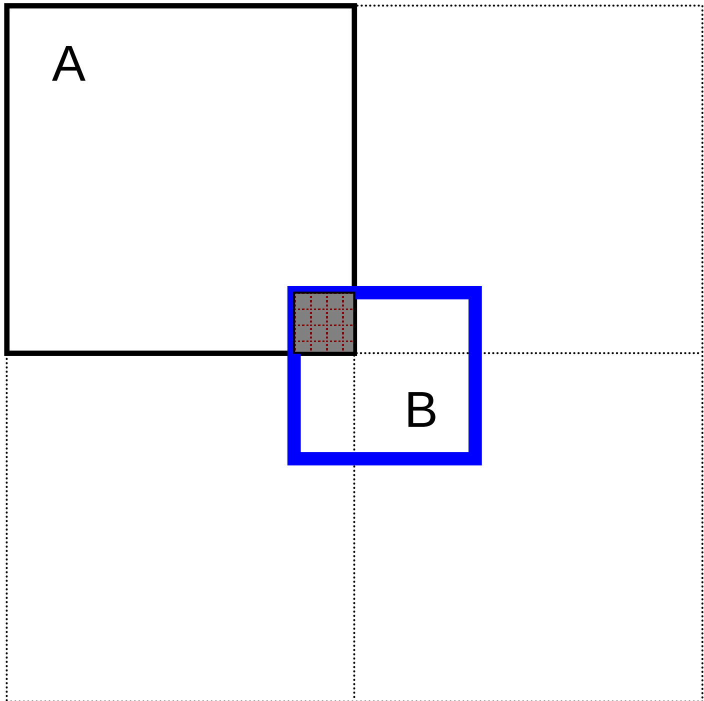
$P(A) = 0{,}25$
$P(B) = 0{,}10$
$P(A \cap B) = 0{,}08$

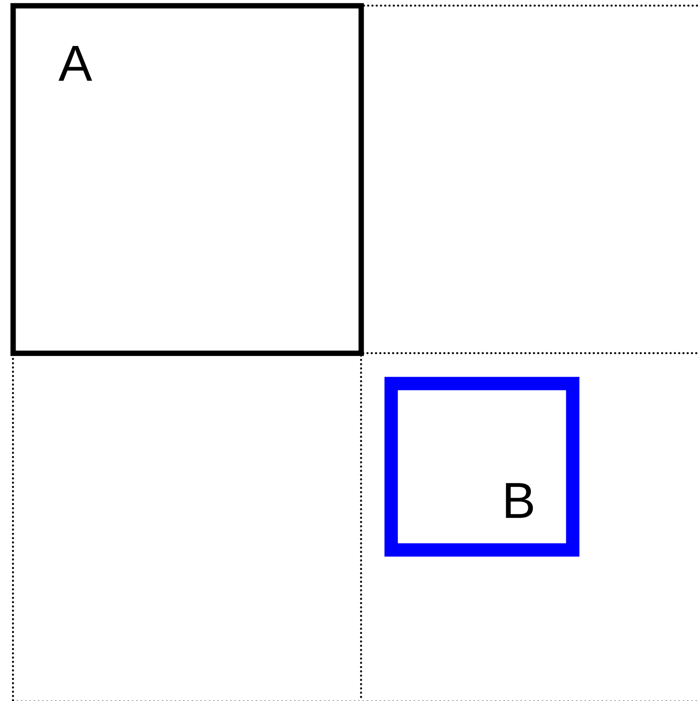
¿Probabilidad de A sabiendo que ha pasado B?

$P(A|B)=1$

$P(A|B)=0{,}8$

Intuir la probabilidad condicionada





P(A) = 0,25
P(B) = 0,10
P(A∩B) = 0,005

P(A) = 0,25
P(B) = 0,10
P(A∩B) = 0

¿Probabilidad de A sabiendo que ha pasado B?

P(A|B)=0,05

P(A|B)=0

Algunas reglas de cálculo prácticas

- Cualquier problema de probabilidad puede resolverse en teoría mediante aplicación de los axiomas. Sin embargo, es más cómodo conocer algunas reglas de cálculo:

  - $P(A') = 1 - P(A)$

  - $P(AUB) = P(A) + P(B) - P(AB)$

  - $P(AB) = P(A)\, P(B|A)$

    $= P(B)\, P(A|B)$

    - Prob. de que pasen A y B es la prob. de A y que también pase B sabiendo que pasó A.

Ejemplo (I)

Recuento

		MENOPAUSIA		
		NO	SI	Total
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

- Se ha repetido en 1000 ocasiones el experimento de elegir a una mujer de una población muy grande. El resultado está en la tabla.
  - ¿Cuál es la probabilidad de que una mujer tenga osteoporosis?
    - P(Osteoporosis)=64/1000=0,064=6,4%
      - Noción frecuentista de probabilidad
  - ¿Cuál es la probabilidad de que una mujer no tenga osteoporosis?
    - P(No Osteoporosis)=1-P(Osteoporsis)=1-64/1000=0,936=93,6%

Ejemplo (II)

Recuento

		MENOPAUSIA		
		NO	SI	Total
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

- ¿Probabilidad de tener osteopenia u osteoporosis?
  - P(Osteopenia∪Osteoporosis)=P(Osteopenia)+P(Osteoporosis)-P(Osteopenia∪Osteoporosis)=467/1000+64/1000=0,531
    - Son sucesos disjuntos
    - Osteopenia ∩ Osteoporosis=Ø
- ¿Probabilidad de tener osteoporosis o menopausia?
  - P(Osteoporosis∪Menopausia)=P(Osteoporosis)+P(Menopausia)-P(Osteoporosis ∩ Menopausia)=64/1000+697/1000-58/1000=0,703
    - No son sucesos disjuntos
- ¿Probabilidad de una mujer normal?
  - P(Normal)=469/1000=0,469
  - P(Normal)=1-P(Normal')=1-P(Osteopenia∪Osteoporosis) =1-0,531=0,469

Ejemplo (III)

Recuento

		MENOPAUSIA		Total
		NO	SI	
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

- Si es menopáusica… ¿probabilidad de osteoporosis?
  - P(Osteoporosis|Menopausia)=58/697=0,098

- ¿Probabilidad de menopausia y osteoporosis?
  - P(Menop ∩ Osteoporosis) = 58/1000=0,058
    - Otra forma:

$$P(Menop \cap Osteoporosis) = P(Menop) \times P(Osteoporosis \mid Menop) =$$
$$= \frac{\cancel{697}}{1000} \times \frac{58}{\cancel{697}} = 58/1000 = 0{,}058$$

Ejemplo (III)

Recuento

		MENOPAUSIA		Total
		NO	SI	
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

- Si tiene osteoporosis… ¿probabilidad de menopausia?
  - P(Menopausia|Osteoporosis)=58/64=0,906
- ¿Probabilidad de menopausia y no osteoporosis?
  - P(Menop ∩ No Osteoporosis) = 639/1000=0,639
- Si tiene no tiene osteoporosis… ¿probabilidad de no menopausia?
  - P(No Menopausia|NoOsteoporosis)=297/936=0,317

Independencia de sucesos

- Dos sucesos son independientes si el que ocurra uno, no añade información sobre el otro.

  - A es independiente de B

    $\Leftrightarrow$ $P(A|B) = P(A)$

    $\Leftrightarrow$ $P(AB) = P(A)\,P(B)$

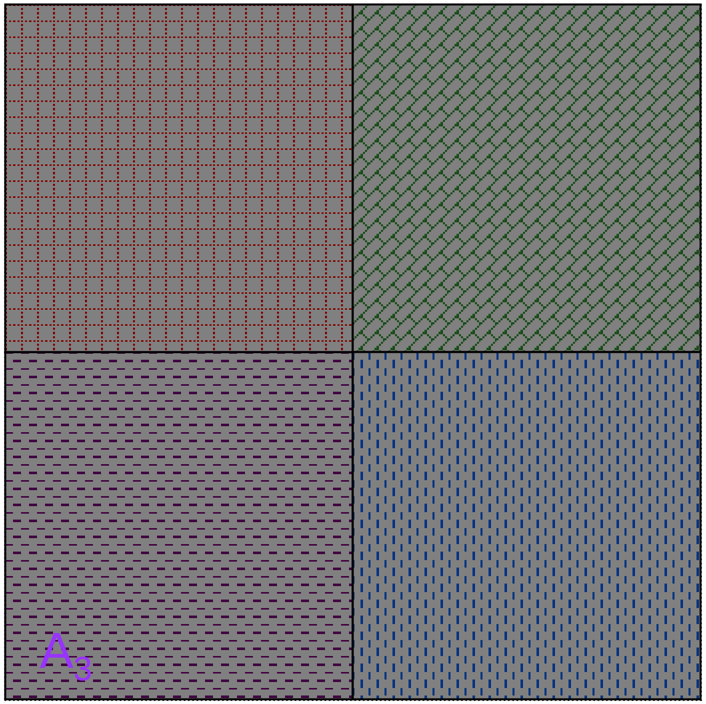
Ejemplo (IV)

Recuento

		MENOPAUSIA		Total
		NO	SI	
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

- ¿Son independientes menopausia y osteoporosis?
  - Una forma de hacerlo
    - P(Osteoporosis)=64/1000=0,064
    - P(Osteoporosis|Menopausia)=58/697=0,098
      - La probabilidad de tener osteoporosis es mayor si ha pasado la menopausia. Añade información extra. ¡No son independientes!
  - ¿Otra forma?
    - P(Menop ∩ Osteoporosis) = 58/1000 = 0,058
    - P(Menop) P(Osteoporosis)= (697/1000) x (64/1000) = 0,045
      - La probabilidad de la intersección no es el producto de probabilidades. No son independientes.

Sistema exhaustivo y excluyente de sucesos

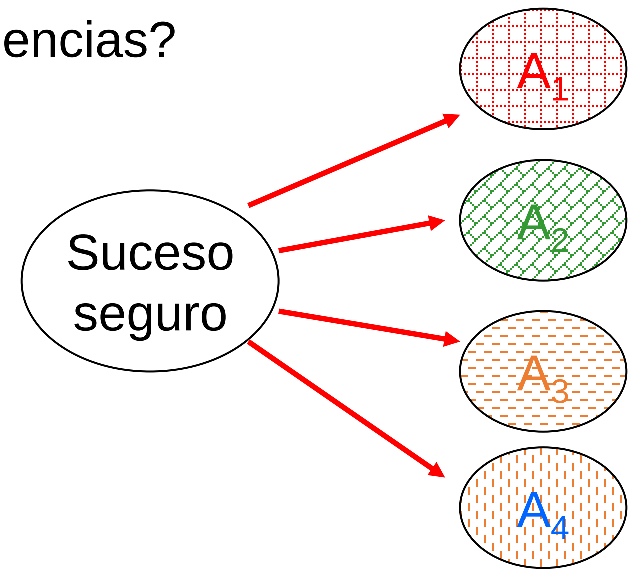


Son una colección de sucesos

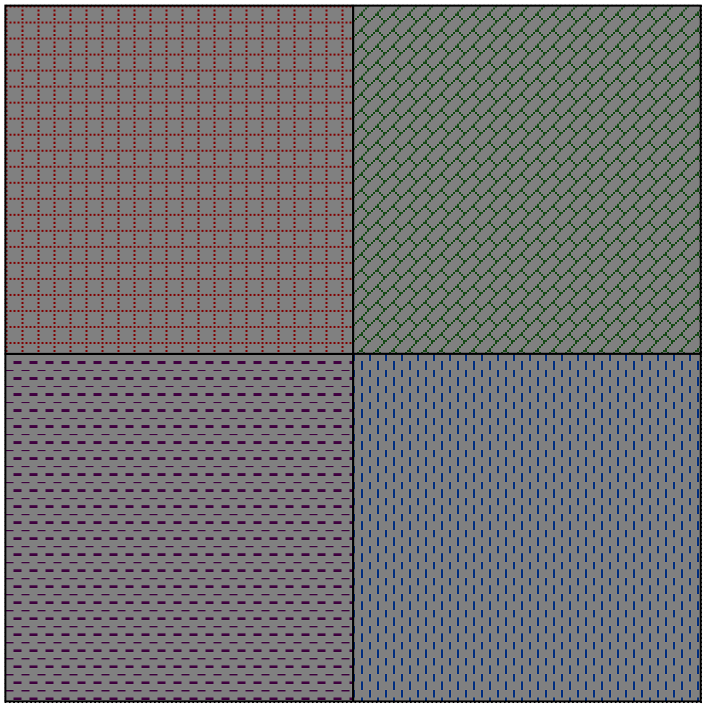
$A_1, A_2, A_3, A_4 \ldots$

Tales que la unión de todos ellos forman el espacio muestral, y sus intersecciones son disjuntas.

¿Recordar cómo formar intervalos en tablas de frecuencias?

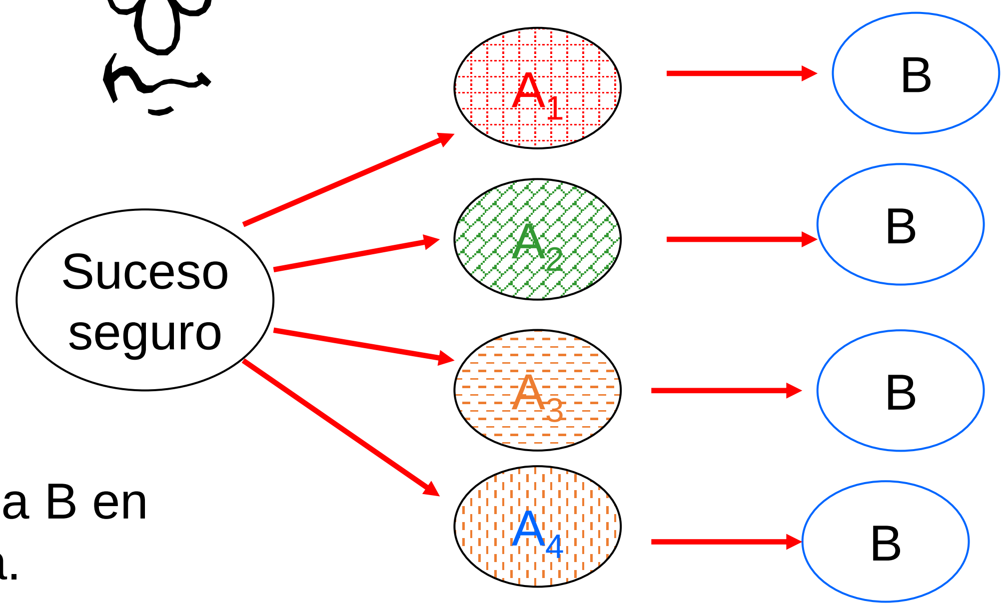


Divide y vencerás



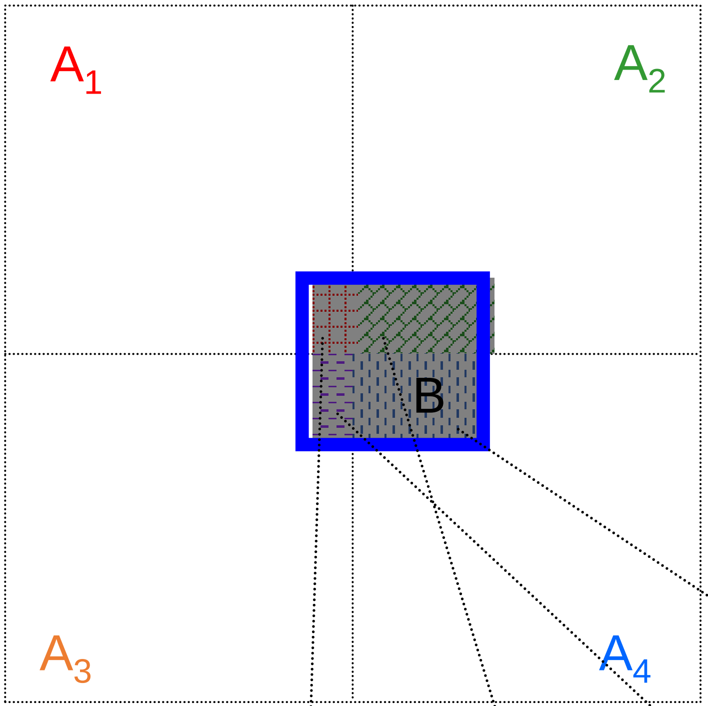
Todo suceso B, puede ser descompuesto en componentes de dicho sistema.

$$B = (B\cap A_1)\ U\ (B\cap A_2)\ U\ (B\cap A_3)\ U\ (B\cap A_4)$$



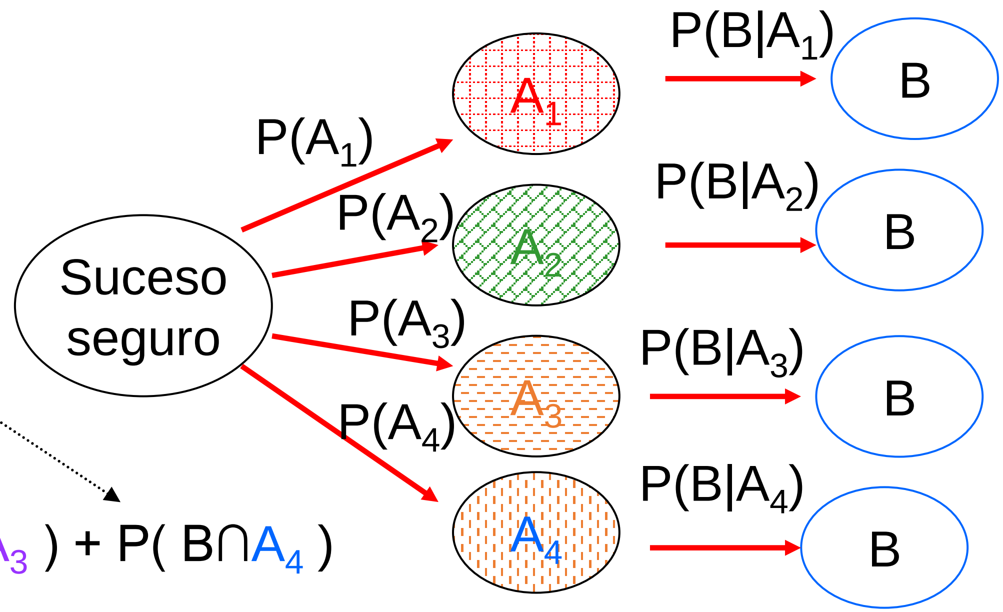
Nos permite descomponer el problema B en subproblemas más simples. Funciona.

Teorema de la probabilidad total



Si conocemos la probabilidad de B en cada uno de los componentes de un sistema exhaustivo y excluyente de sucesos, entonces…

… podemos calcular la probabilidad de B.



$$P(B) = P(B\cap A_1) + P(B\cap A_2) + P(B\cap A_3) + P(B\cap A_4)$$

$$=P(A_1)\, P(B|A_1) + P(A_2)\, P(B|A_2) + \ldots$$

Ejemplo (I): En este aula el 70% de los alumnos son mujeres. De ellas el 10% son fumadoras. De los hombres, son fumadores el 20%.

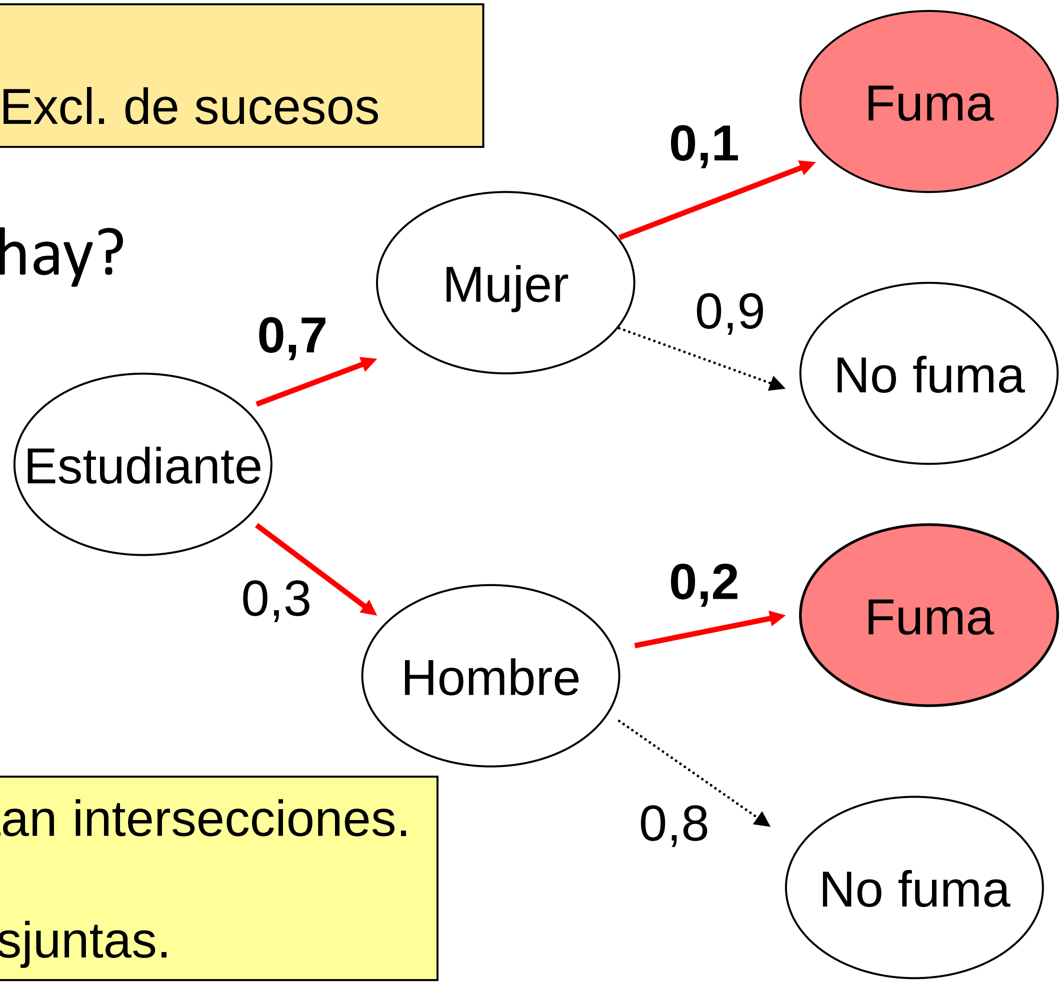
T. Prob. Total.
Hombres y mujeres forman un sist. Exh. Excl. de sucesos

- ¿Qué porcentaje de fumadores hay?
  - P(F) = P(M∩F) + P(H∩F)

    = P(M)P(F|M) + P(H)P(F|H)

    =0,7 x 0,1 + 0,3 x 0,2

    = 0,13 =13%



•Los caminos a través de nodos representan intersecciones.

•Las bifurcaciones representan uniones disjuntas.

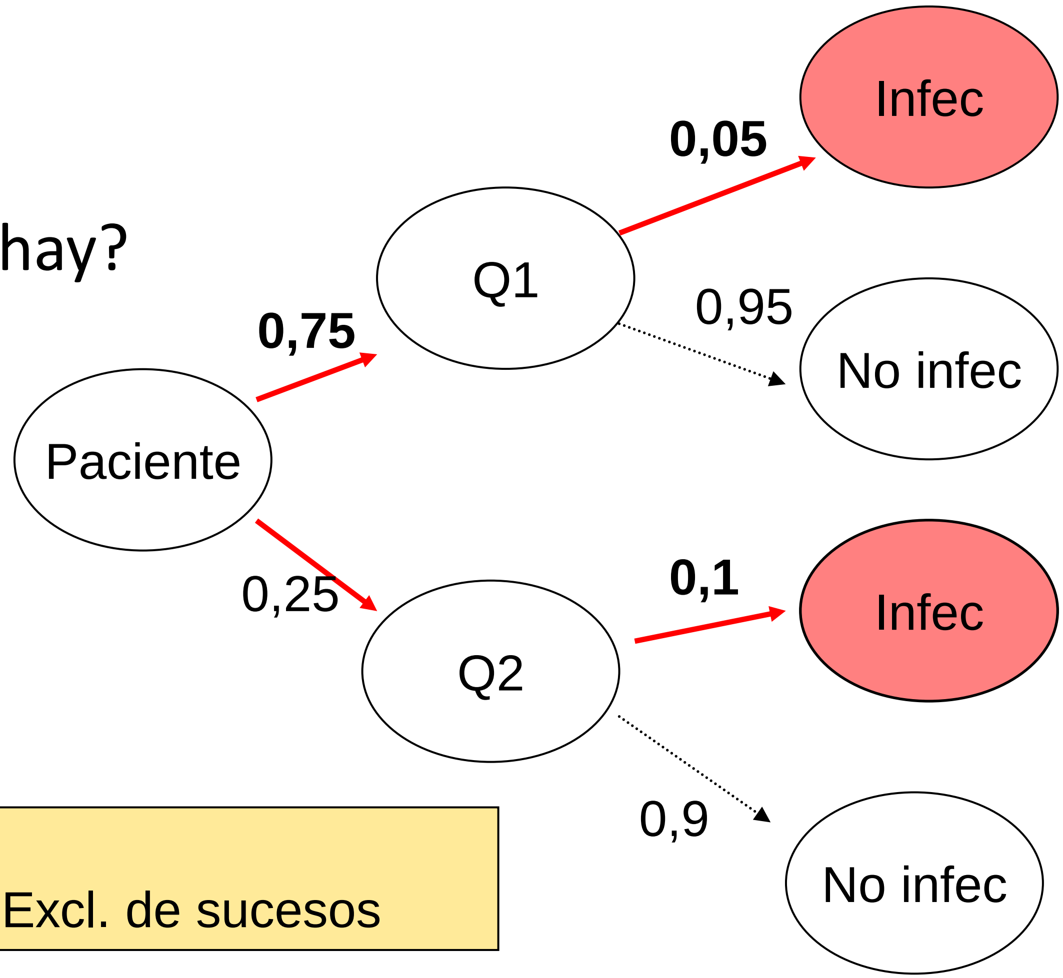
Ejemplo (II): En un centro hay dos quirófanos. El 1º se usa el 75% de veces para operar. En el 1º la frec. de infección es del 5% y en el 2º del 10%.

- ¿Qué probabilidad de infección hay?
  - $P(I) = P(Q1 \cap I) + P(Q2 \cap I)$

    $= P(Q1)P(I|Q1) + P(Q2)P(I|Q2)$

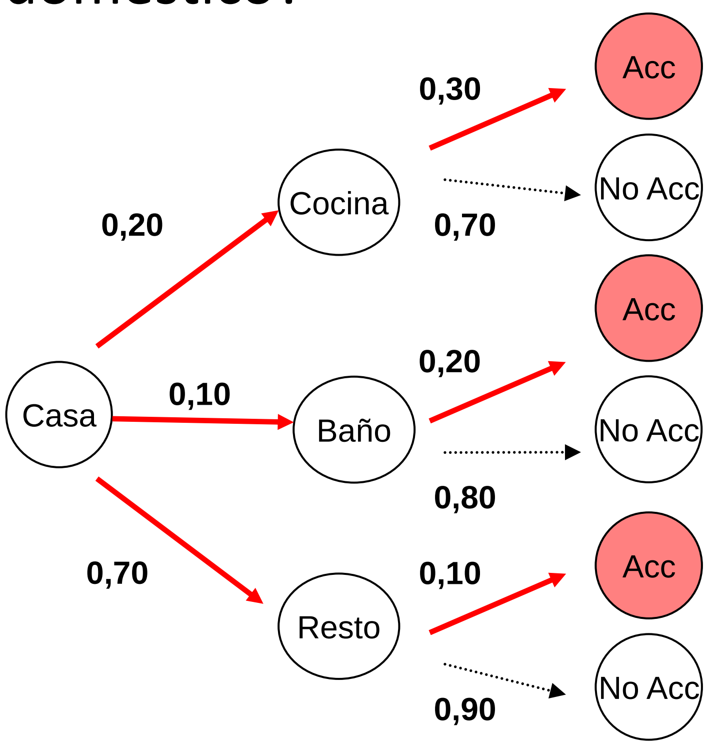
    $= 0,75 \times 0,05 + 0,25 \times 0,1$

    $= 0,0625$



T. Prob. Total.
Los dos quirófanos forman un sist. Exh. Excl. de sucesos

Ejemplo (III): El 20% del tiempo que se está en una casa transcurre en la cocina, el 10% en el baño y el resto entre el salón y el dormitorio. Por otro lado la probabilidad de tener un accidente doméstico estando en la cocina es de 0,30 de tenerlo estando en el baño es de 0,20 y de tenerlo fuera de ambos de 0,10. ¿Cuál es la probabilidad de tener un accidente doméstico?



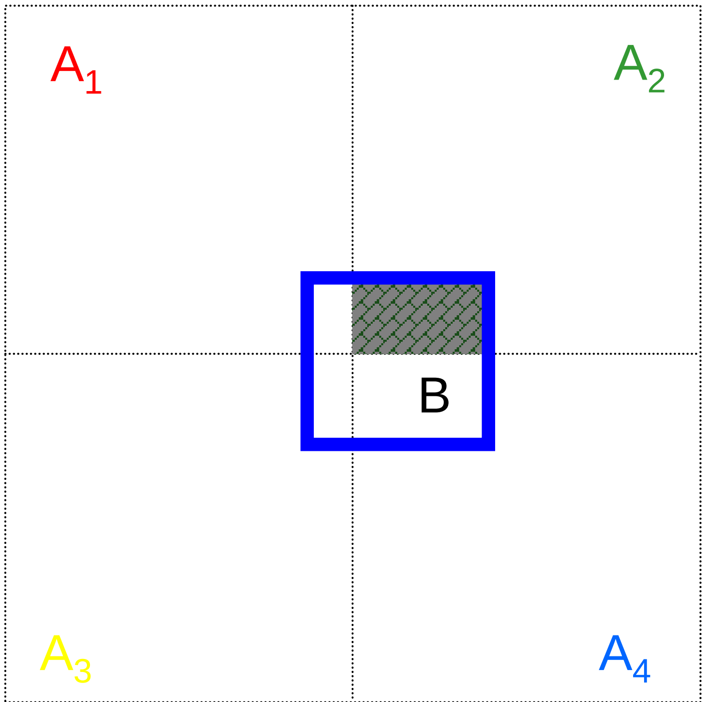
$P(A) = P(A \cap C) + P(A \cap B) + P(A \cap R) =$

$P(C)P(A|C) + P(B)P(A|B) + P(R)P(A|R)$

$= 0{,}2 \times 0{,}3 + 0{,}1 \times 0{,}2 + 0{,}7 \times 0{,}1 = 0{,}15 =$

15%

Teorema de Bayes



Si conocemos la probabilidad de B en cada uno de los componentes de un sistema exhaustivo y excluyente de sucesos, entonces…

…si ocurre B, podemos calcular la probabilidad (*a posteriori*) de ocurrencia de cada A_i.

$$P(Ai\,|\,B)=\frac{P(B\ Ai)}{P(B)}$$

donde P(B) se puede calcular usando el teorema de la probabilidad total:

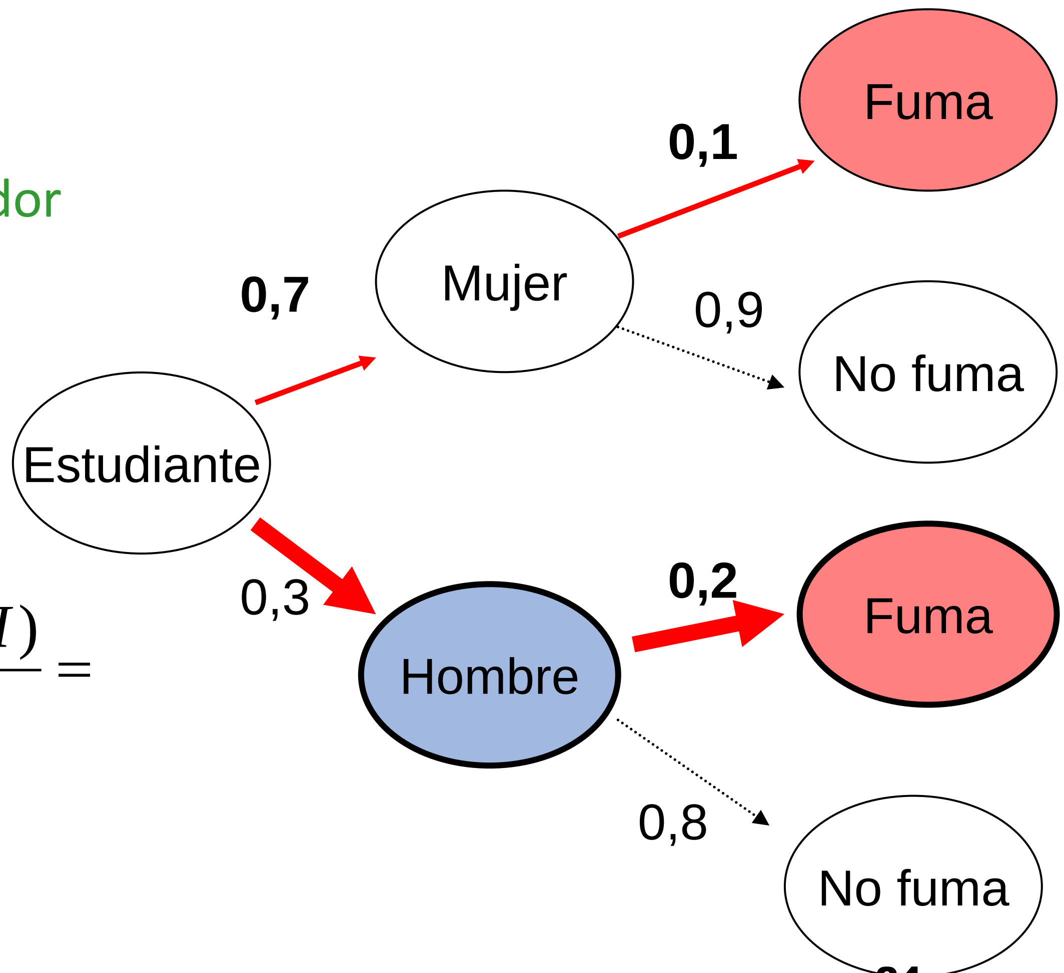
$$P(B)=P(B\cap A_1) + P(B\cap A_2\ ) + P(\ B\cap A_3\ ) + (\ B\cap A_4\ )$$

$$=P(B|A_1)\ P(A_1) + P(B|A_2)\ P(A_2) + \ldots$$

Ejemplo (IV): En este aula el 70% de los alumnos son mujeres. De ellas el 10% son fumadoras. De los varones, son fumadores el 20%.

- ¿Qué porcentaje de fumadores hay?
  - P(F) = =0,7 x 0,1 + 0,3 x 0,2 = 0,13
    - (Resuelto antes)

- Se elije a un individuo al azar y es… fumador ¿Probabilidad de que sea un hombre?



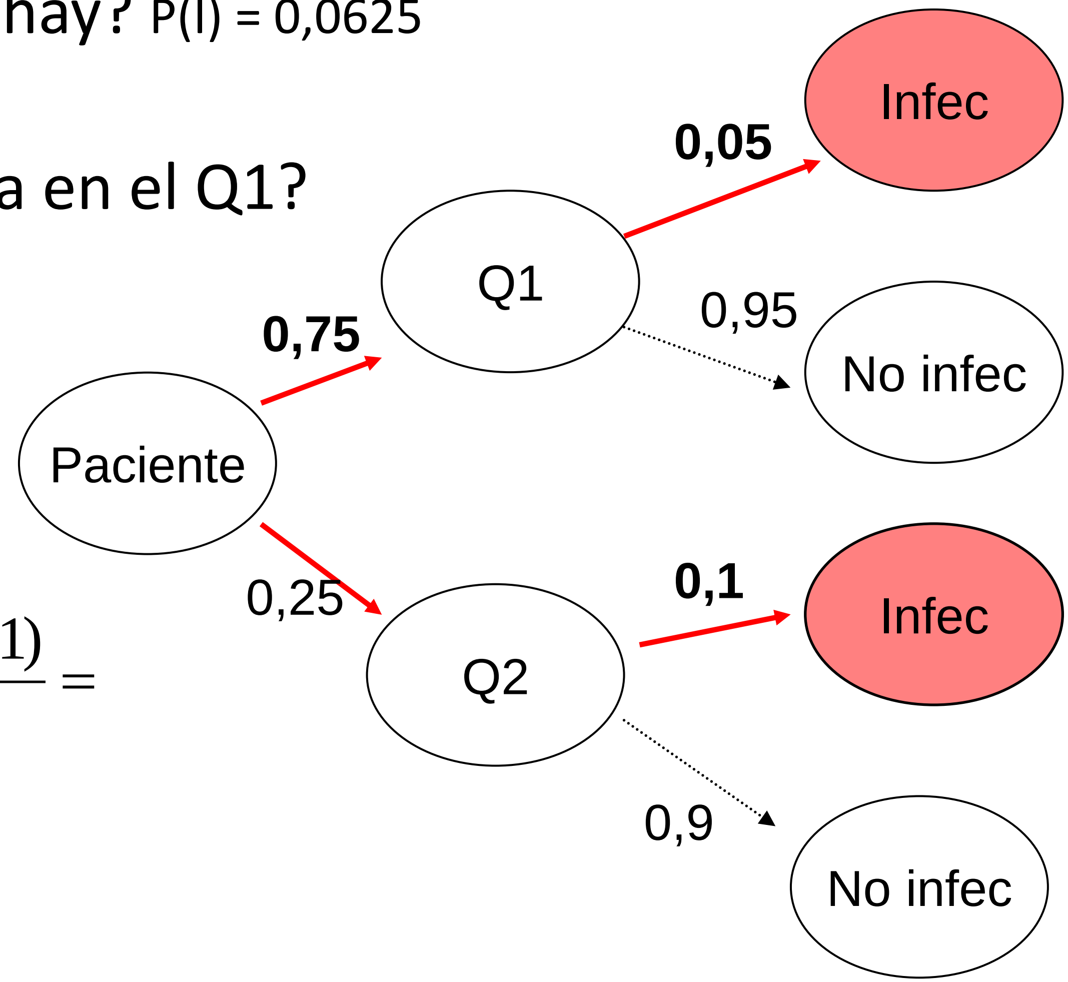
$$P(H \mid F) = \frac{P(H \cap F)}{P(F)} = \frac{P(H) \cdot P(F \mid H)}{P(F)} =$$

$$= \frac{0{,}3 \times 0{,}2}{0{,}13} = 0{,}46$$

Ejemplo (V): En un centro hay dos quirófanos. El 1º se usa el 75% de veces para operar. En el 1º la frec. de infección es del 5% y en el 2º del 10%.

- ¿Qué probabilidad de infección hay? P(I) = 0,0625
- Se ha producido una infección.

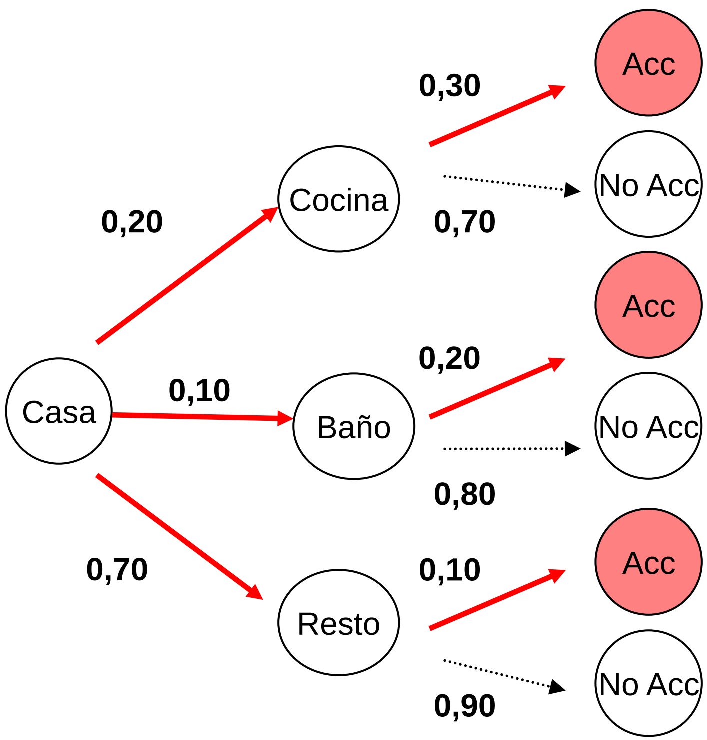
¿Qué probabilidad hay de que sea en el Q1?



$$P(Q1 \mid I) = \frac{P(Q1 \cap I)}{P(I)} = \frac{P(Q1) \cdot P(I \mid Q1)}{P(I)} =$$

$$= \frac{0{,}75 \times 0{,}05}{0{,}0625} = 0{,}6$$

Ejemplo (VI): El 20% del tiempo que se está en una casa transcurre en la cocina, el 10% en el baño y el resto entre el salón y el dormitorio. Por otro lado la probabilidad de tener un accidente doméstico estando en la cocina es de 0,30 de tenerlo estando en el baño es de 0,20 y de tenerlo fuera de ambos de 0,10. Se ha producido un accidente, ¿cuál es la probabilidad de que haya sido en la cocina?



P(A) = 0,15 (ya calculado)

$$P(C \mid A) = \frac{P(C \cap A)}{P(A)} = \frac{P(C) \cdot P(A \mid C)}{P(A)} =$$

$$= \frac{0{,}20 \times 0{,}30}{0{,}15} = 0{,}4$$

Ejemplo de prueba diagnósticas: Diabetes

- Los carbohidratos ingeridos terminan como glucosa en la sangre. El exceso se transforma en glucógeno y se almacena en hígado y músculos. Este se transforma entre comidas de nuevo en glucosa según necesidades.

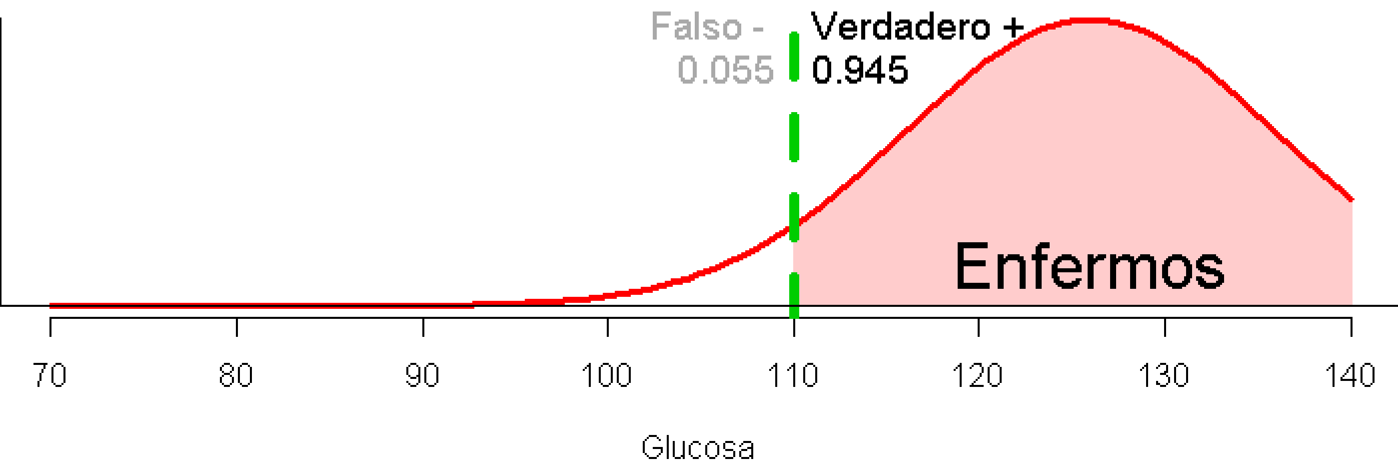
- La principal hormona que regula su concentración es la insulina. La diabetes provoca su deficiencia o bien la insensibilidad del organismo a su presencia. Es una enfermedad muy común que afecta al 2% de la población (prevalencia)

- Una prueba común para diagnosticar la diabetes, consiste en medir el nivel de glucosa. En individuos sanos suele variar entre 64 y 110mg/dL.
  - El cambio de color de un indicador al contacto con la orina suele usarse como indicador (resultado del test positivo)

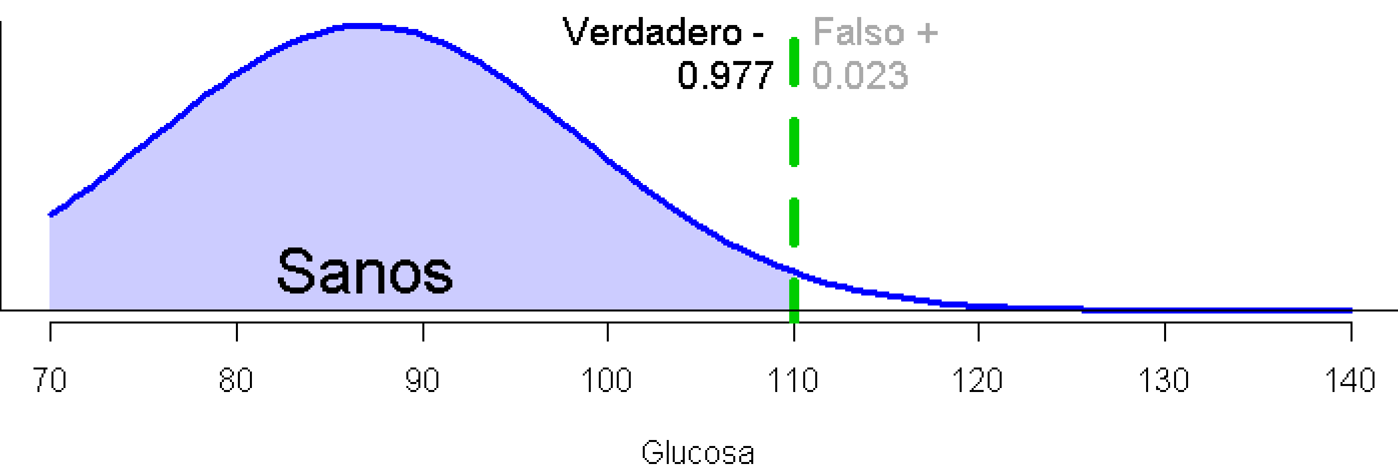
- Valores por encima de 110 mg/dL se asocian con un posible estado pre-diabético.
  - Pero no es seguro. Otras causas podrían ser: hipertiroidismo, cancer de páncreas, pancreatitis, atracón reciente de comida…
- Supongamos que los enfermos de diabetes, tienen un valor medio de 126mg/dL.

Funcionamiento de la prueba **diagnóstica** de glucemia

- **Valor límite: 110mg/dL**
  - Superior: test positivo.
  - Inferior: test negativo.

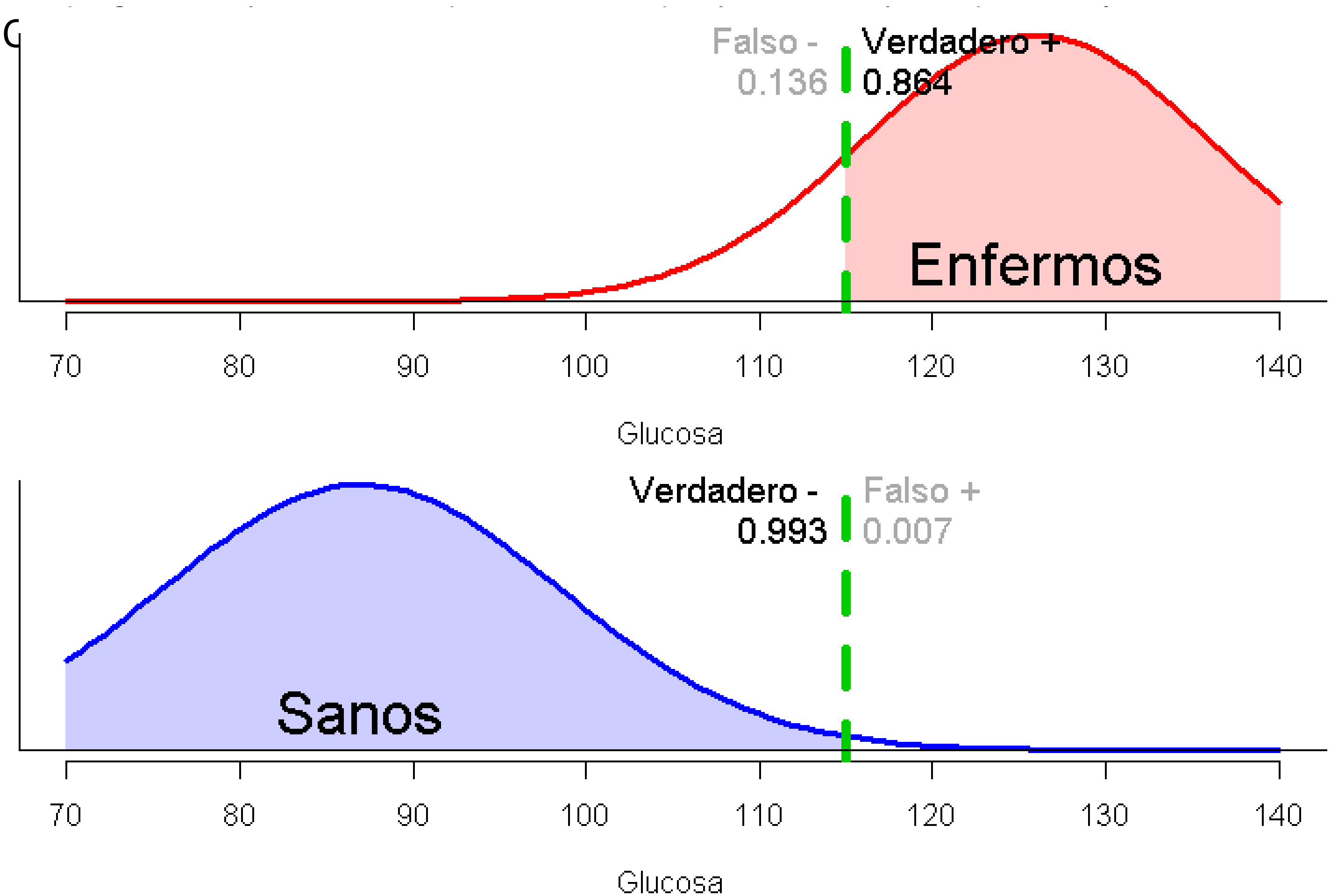


obabilidad de acierto:
- Para enfermos
  - Verdadero positivo (sensibilidad)
- Para sanos
  - Verdadero negativo (especificidad)



obabilidad de error
- Para enfermos
  - Falso –
- Para sanos
  - Falso +

¿Cómo



No es simple. **No es posible aumentar sensibilidad y especificidad al mismo tiempo**. Hay que elegir una solución de compromiso: Aceptable sensibilidad y especificidad.

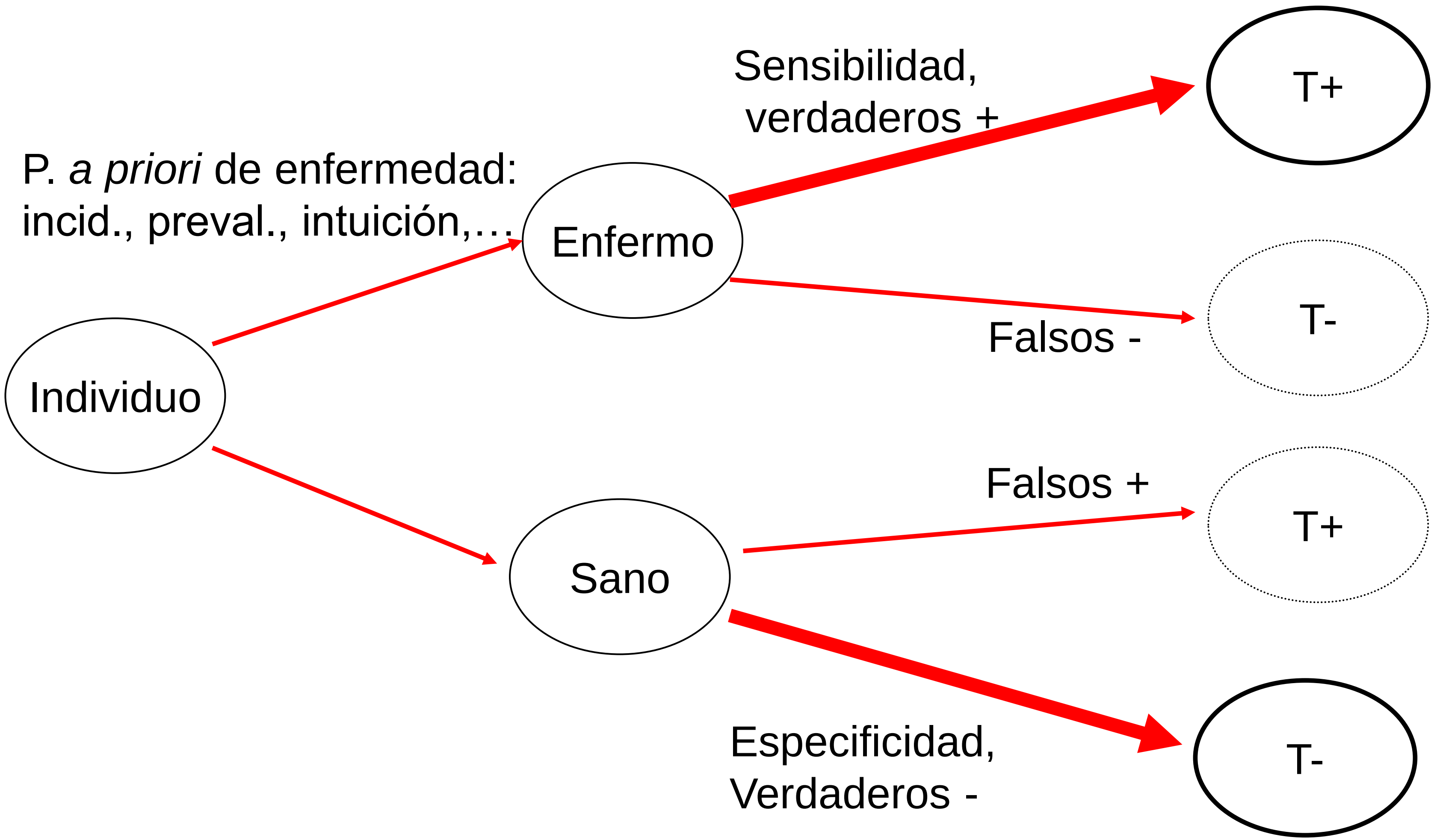
Una prueba diagnóstica ayuda a mejorar una estimación de la probabilidad de que un individuo presente una enfermedad.

- En pricipio tenemos una idea subjetiva de P(Enfermo). Nos ayudamos de…
  - Incidencia: Porcentaje de nuevos casos de la enfermedad en la población.
  - Prevalencia: Porcentaje de la población que presenta una enfermedad.

- Para confirmar la sospecha, usamos una prueba diagnóstica. Ha sido evaluada con anterioridad sobre dos grupos de individuos: sanos y enfermos. Así de modo frecuentista se ha estimado:
  - P(+ | Enfermo)= Sensibilidad (verdaderos +)= Tasa de acierto sobre enfermos.
  - P(- | Sano) = Especificidad (verdaderos -)= Tasa de acierto sobre sanos.

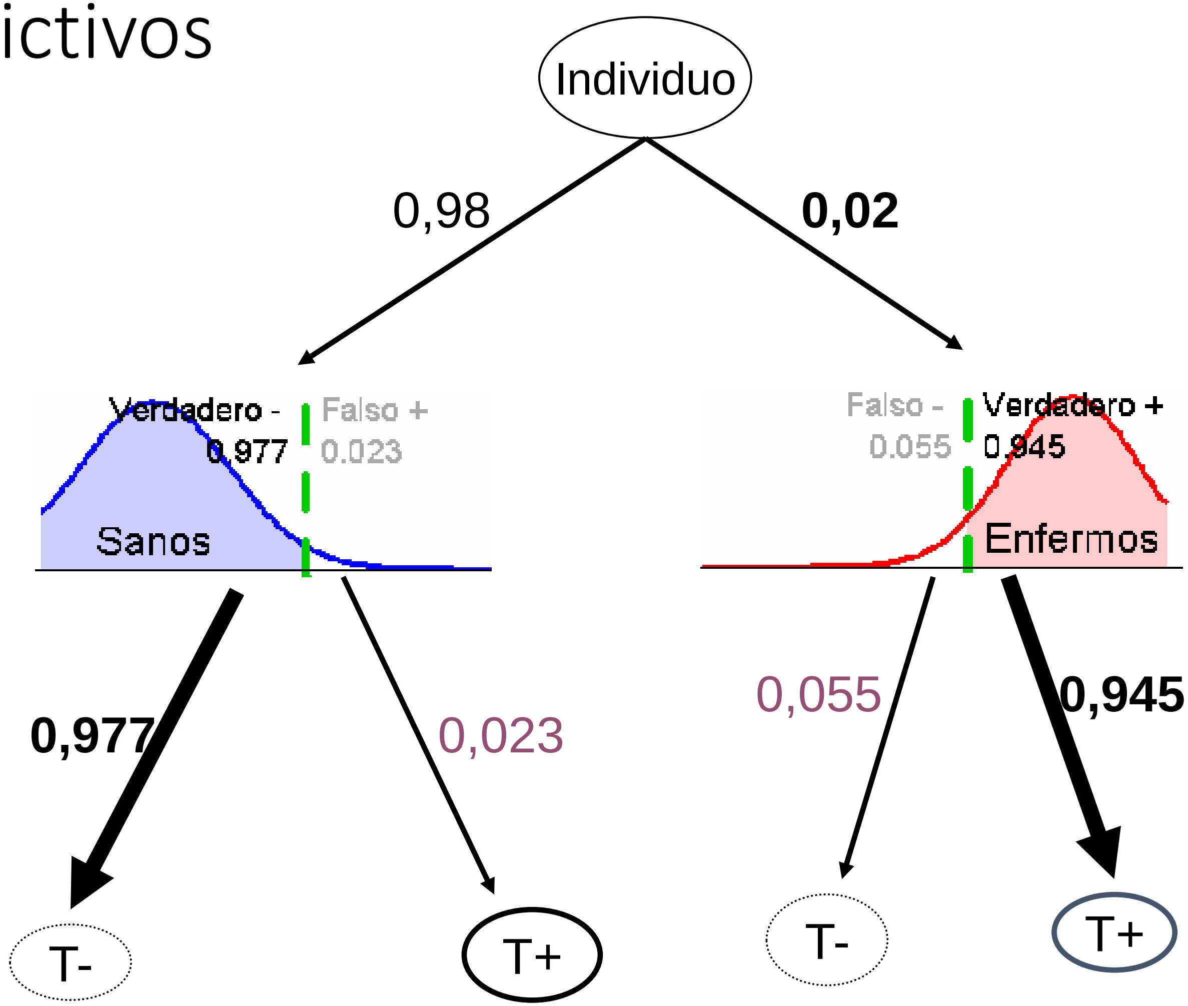
- A partir de lo anterior y usando el teorema de Bayes, podemos calcular las probabilidades *a posteriori* (en función de los resultados del test): Índices predictivos
  - P(Enfermo | +) = Índice predictivo positivo
  - P(Sano | -) = Índice predictivo negativo

Pruebas diagnósticas: aplicación T. Bayes.



Ejemplo: Índices predictivos

- La diabetes afecta al 2% de los individuos.
- La presencia de glucosuria se usa como indicador de diabetes.
- Su sensibilidad es de 0,945.
- La especificidad de 0,977.
- Calcular los índices predictivos.



$$P(Sano|T-)=\frac{P(Sano\cap T-)}{P(T-)}=\frac{P(Sano)P(T-|Sano)}{P(Sano)P(T-|Sano)+P(Enf)P(T-|Enf)}$$
$$=\frac{0{,}98\cdot 0{,}977}{0{,}98\cdot 0{,}977+0{,}02\cdot 0{,}055}=0{,}999$$

$$P(Enf\,|T+)=\frac{P(Enf\cap T+)}{P(T+)}=\frac{P(Enf)P(T+|Enf)}{P(Sano)P(T+|Sano)+P(Enf)P(T+|Enf)}$$
$$=\frac{0{,}02\cdot 0{,}945}{0{,}02\cdot 0{,}945+0{,}98\cdot 0{,}023}=0{,}456$$

Observaciones

-¿Qué probabilidad tengo de estar enfermo?

- En principio un 2%. Le haremos unas pruebas.

- En el ejemplo anterior, al llegar un individuo a la consulta tenemos una idea *a priori* sobre la probabilidad de que tenga una enfermedad.

- A continuación se le pasa una prueba diagnóstica que nos aportará nueva información: Presenta glucosuria o no.

- En función del resultado tenemos una nueva idea (*a posteriori*) sobre la probabilidad de que esté enfermo.
  - Nuestra opinión a priori ha sido modificada por el resultado de un experimento.

- Presenta glucosuria. La probabilidad ahora es del 45,6%.