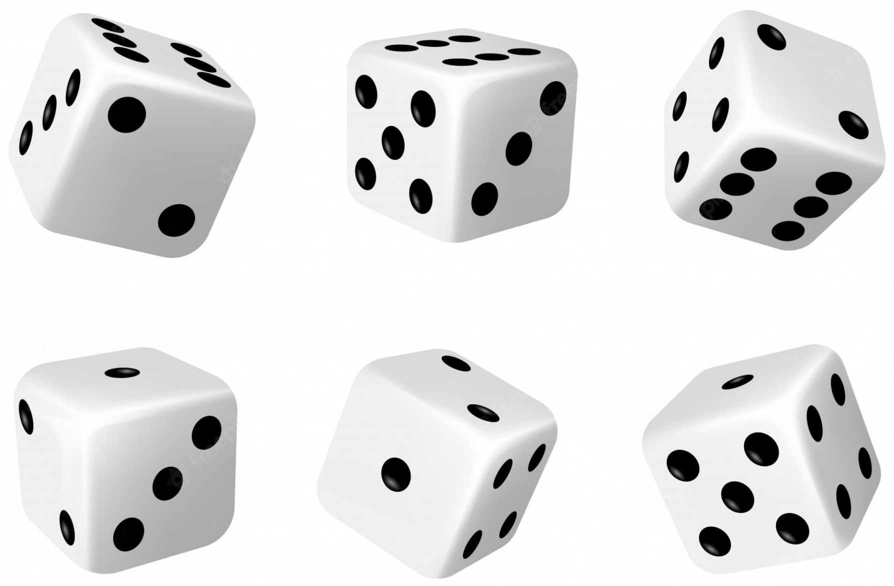
El estudio de los fenomenos aleatorios y sus interpretaciones

Inferencia Bayesiana vs frecuentista



Gino Marceca —

Objetivo de la charla

- Interpretacion Bayesiana vs Frecuentista de la probabilidad.
- La probabilidad como un estado del conocimiento.
- Descripcion y definicion de una variable o fenomeno aleatorio:
  - Valor medio, Varianza y Entropia
- Punto de quiebre en la interpretacion de la probabilidad: razonamiento deductivo vs inductivo
- La importancia de la combinatoria en la asignacion de probabilidades (distribucion binomial)
- Como las leyes de la probabilidad derivan de la logica humana y es en efecto el sentido comun reducido al calculo
- Como el sentido comun puede resultar engañoso y contraintuitivo muchas veces.
- **Gozar de sus sorprendentes predicciones basadas en las premisas mas simples.**

Un poco de la historia bayesiana

Como la logica deductiva aplicada a los juegos de azar podria ayudar a inferir problemas de la logica inductiva en la vida diaria?

Laplace re-descrubre el postulado de Bayes y hace uso del mismo de forma exitosa por ejemplo para estimar la masa de Saturno

Kolmogorov: Formalizacion de la teoria de probabilidades mediante axiomas

Jeffreys re-vive el teorema de Bayes y juega un rol importante en la interpretacion objetiva de la probabilidad bayesiana

1713 1763 1812 1933 1939 1946 1983 2003

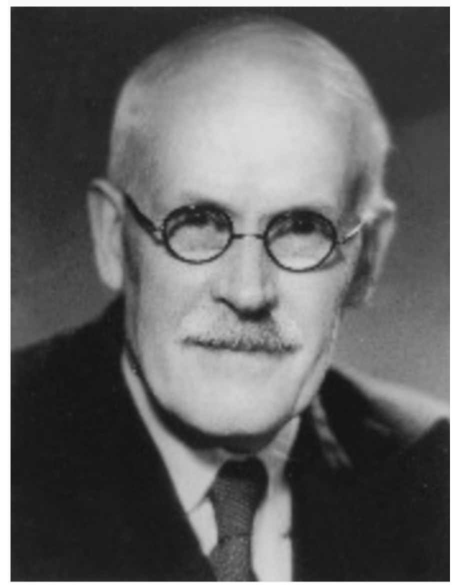
Jacob Bernoulli

Thomas Bayes

Pierre-Simon Laplace

Andrey Kolmogorov

Sir Harold Jeffreys



Richard Cox: Demuestra que todo conjunto de proposiciones logicas en donde se representen grados de posibi lidad mediante numeros reales, equivale a las reglas de Laplace de probabilidades: logica —> teoria de prob. bayesiana

E. T. Jaynes

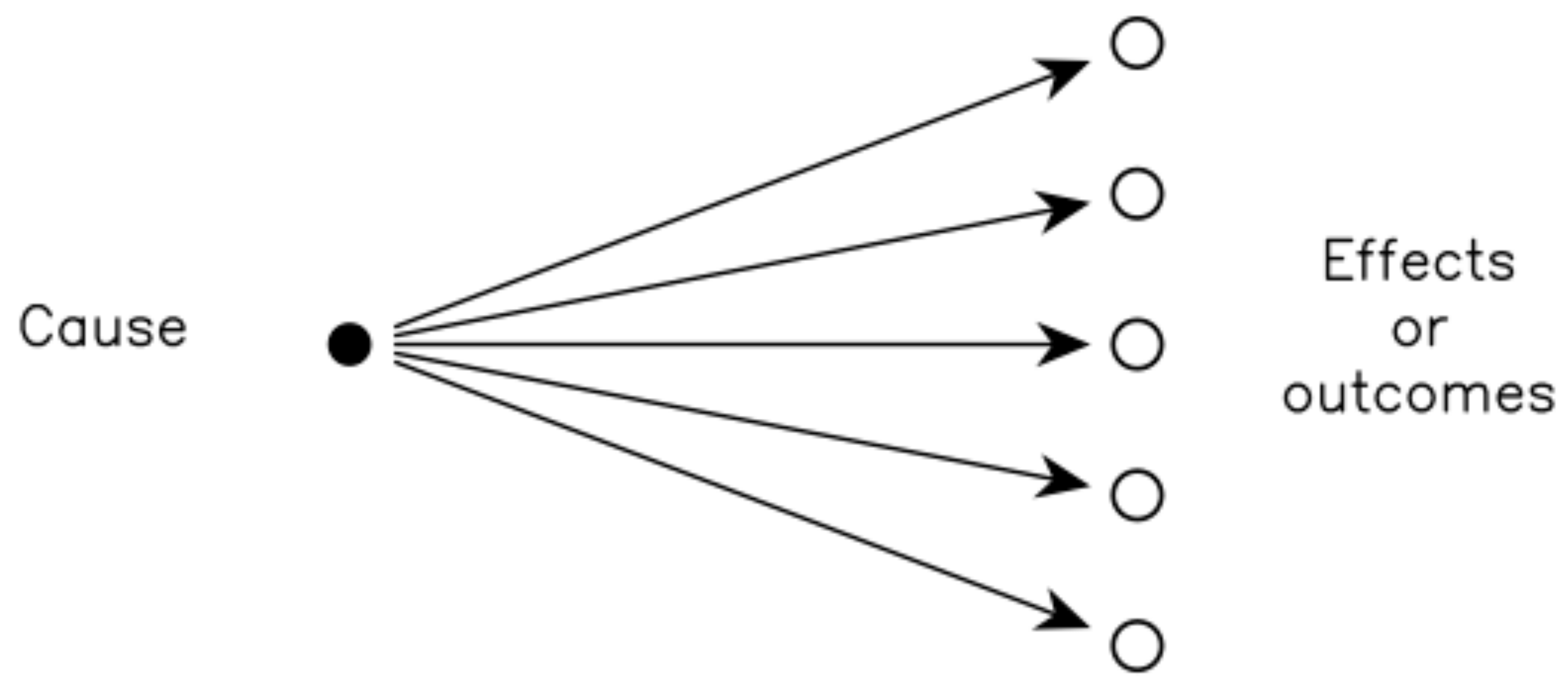
Idea de Bayes publicada (de forma anonima) en 1763. Conocido como el teorema de Bayes. Da una respuesta a Bernoulli.

A pesar de los numerosos exitos de Laplace, su desarrollo bayesiano fue desacreditado por muchos y olvidado luego de su muerte

Jaynes promociono y desarrollo la teoria de probabilidades como una extension de la logica. En particular la aplicacion del principio de maxima entropia para la asignacion de probabi lidades que mejor se ajustan al conocimiento anterior (priors).

Razonamiento deductivo vs inductivo

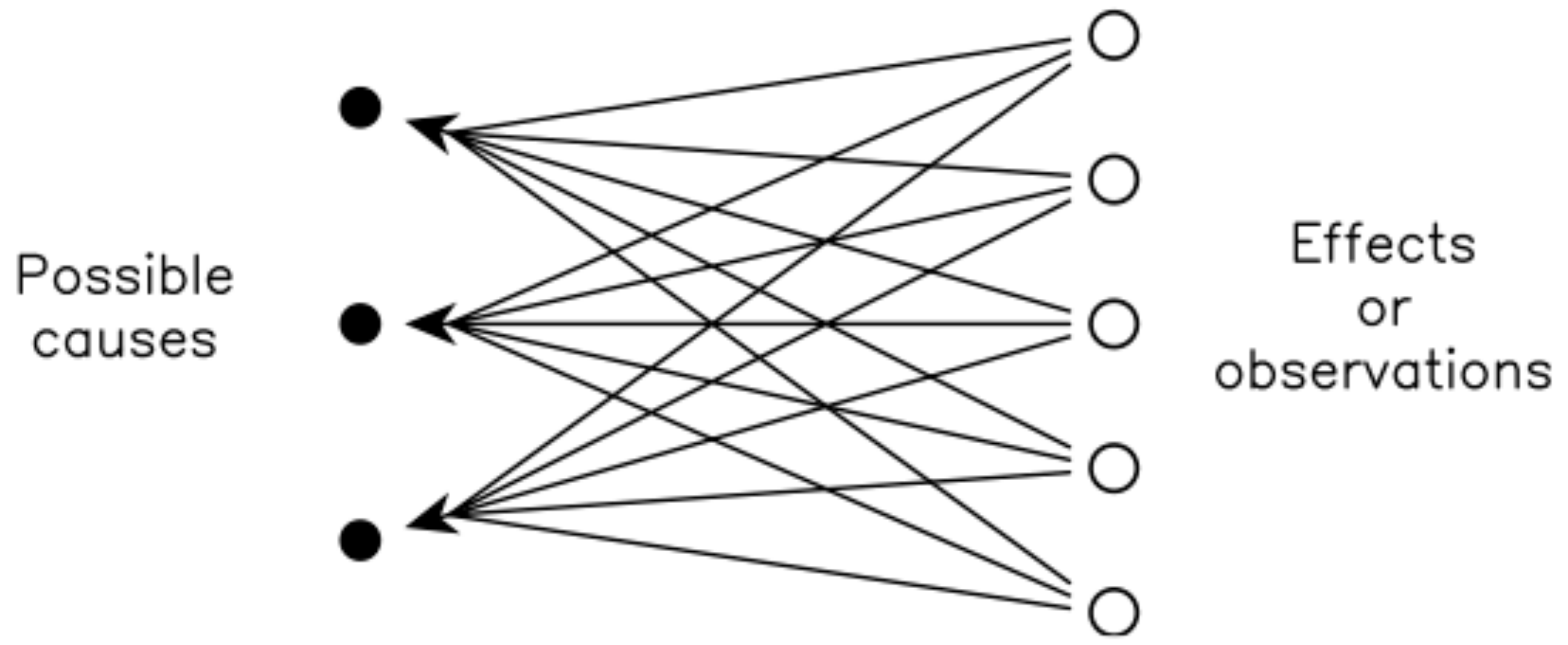
Deductivo



Matematica, juegos de azar, logica, computacion, …

Dada una moneda justa: cual es la probabilidad de tirarla 10 veces y que salga cara 9 veces?

Inductivo o razonamiento plausible



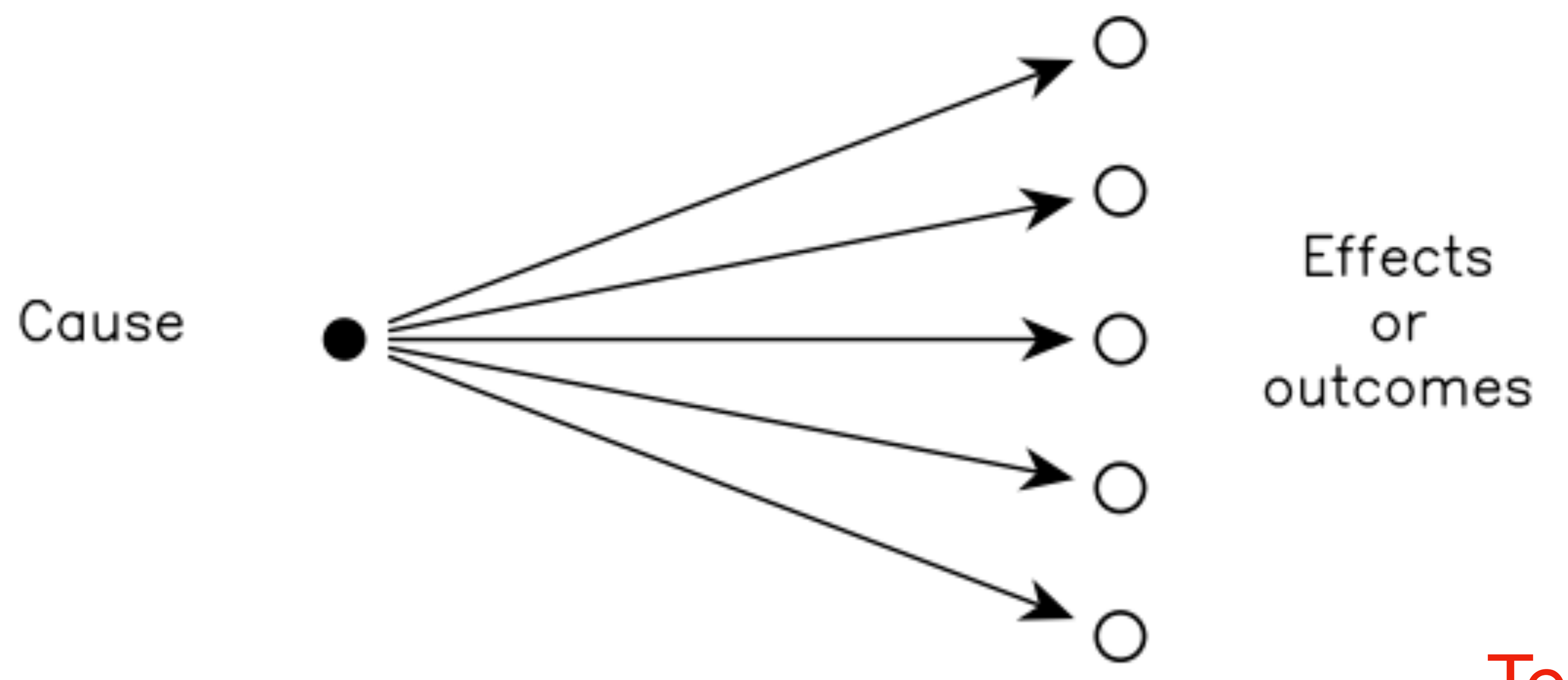
Fisica, biologia, la vida misma, …

Dado que observo 9 caras de 10 tiradas, cuales son sus causas?, es la moneda justa?

Estas preguntas son mucho mas dificiles de responder, e involucran inherentemente un grado de incertidumbre.

Razonamiento deductivo vs inductivo

Deductivo

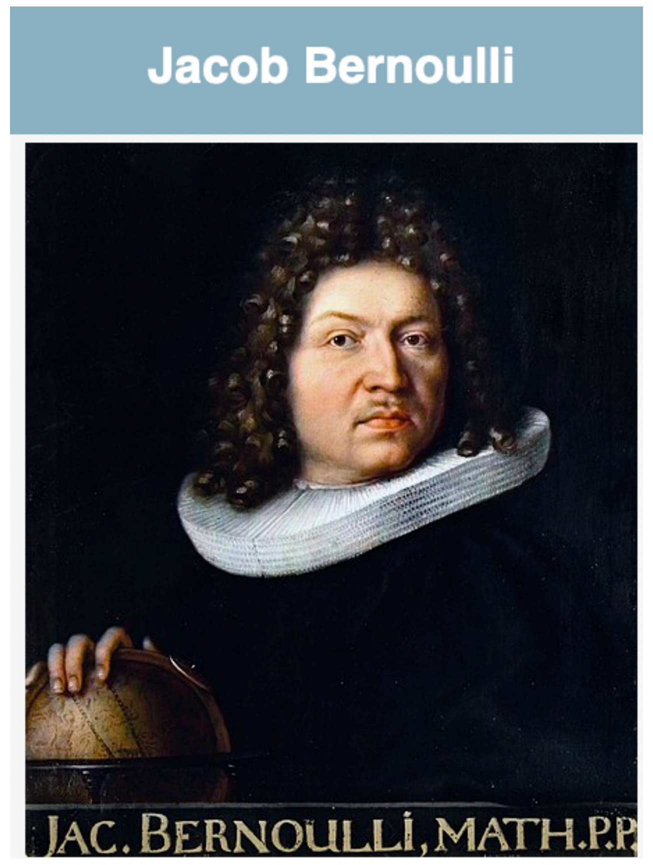


Matematica, juegos de azar, logica, computacion, …

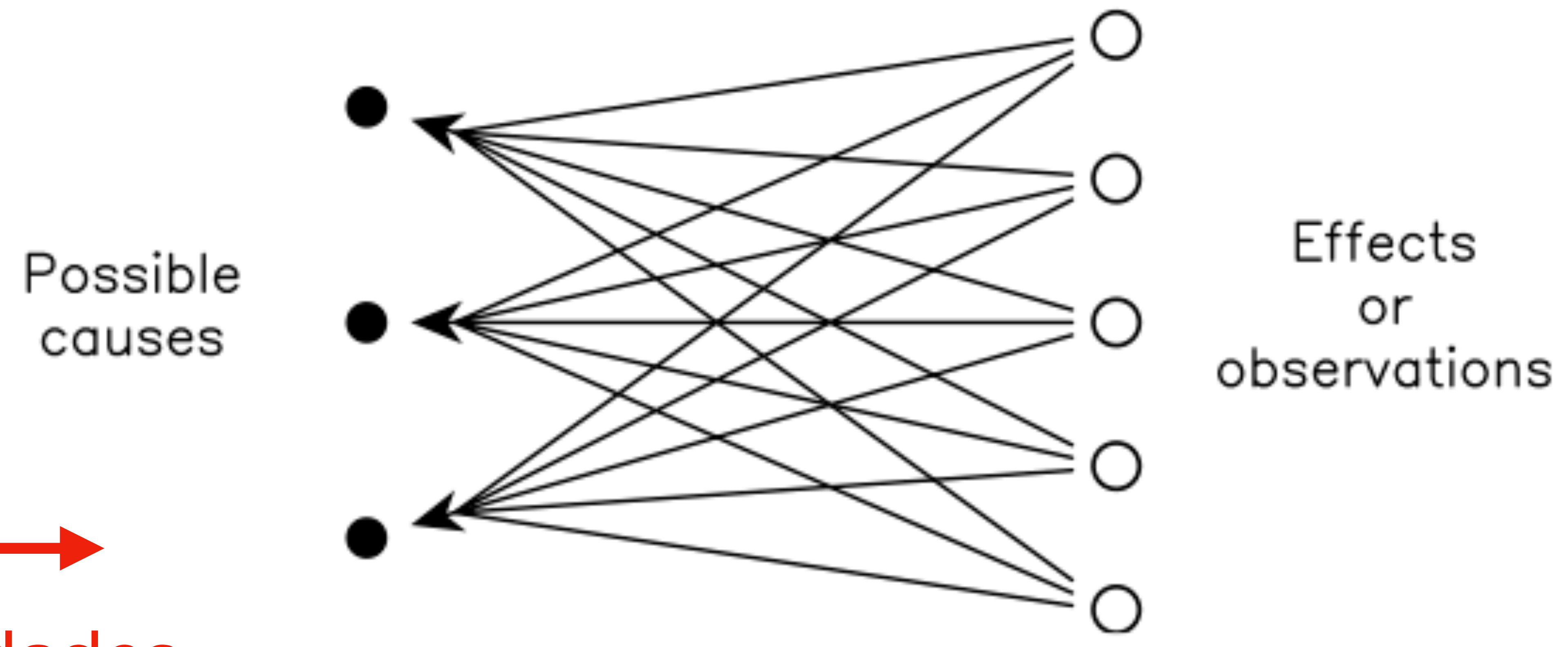
Dada una moneda justa: cual es la probabilidad de tirarla 10 veces y que salga cara 9 veces?

Teoria de probabilidades

1713



Inductivo o razonamiento plausible



Fisica, biologia, la vida misma, …

Dado que observo 9 caras de 10 tiradas, cuales son sus causas?, es la moneda justa?

Estas preguntas son mucho mas dificiles de responder, e involucran inherentemente un grado de incertidumbre.

Razonamiento bayesiano

… Lo máximo que podemos esperar es hacer la mejor inferencia
basada en los datos experimentales y en cualquier conocimiento previo que tengamos disponible,
reservándonos el derecho de revisar nuestra posición si sale a la luz nueva información …
[D.S. SIVIA, data analysis, a bayesian tutorial]

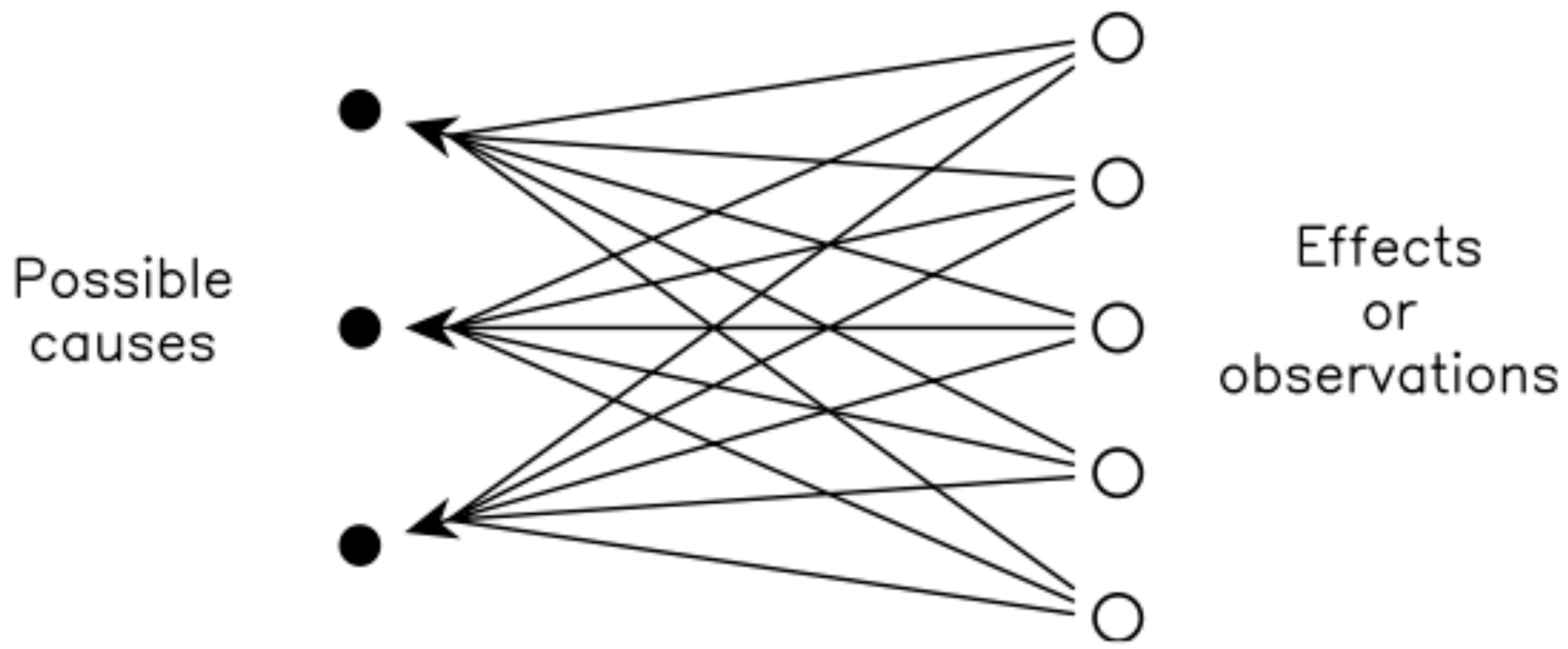
Una decisión era sabia, aunque llevara a consecuencias desastrosas, si la evidencia a mano indicaba que era la mejor que se podía tomar; y una decisión era tonta, aunque llevara a las consecuencias más felices posibles, si no era razonable esperar esas consecuencias.
[Herodotus, 500 A.C]

La teoría de la probabilidad no es más que el sentido común reducido al cálculo.
[Laplace, ~1812]

Nos centraremos en una rama particular del pensamiento bayesiano:
la interpretacion logica de la probabilidad o “bayesianismo objetivo”.

Frecuentista y bayesiano

Inductivo o razonamiento plausible



Escuela frecuentista

- Desarrollo formal de la epoca.
- Invencion de la Estadistica
- Es lo que se enseña usualmente en los libros de texto.
- Cramer, Fisher, Pearson, Neyman, …

Critica principal de los frecuentistas sobre el subjetivismo de los bayesianos

Escuela bayesiana

- Bayesianismo subjetivo
- Bayesianismo objetivo

- Vision mas moderna de la probabilidad (re-elaborada por Jeffrey y otros ~1939).
- Bernoulli, Bayes, Laplace, …

- Una de las interpretaciones de la mecanica cuantica "QBism"

Conceptos de probabilidad: fenomenos aleatorios

- Un sistema se dice que es aleatorio (random) cuando se desconocen sus condiciones iniciales o estas no pueden ser determinadas con 100% de certeza dada a la naturaleza del mismo.

- Ejemplo: lanzamiento de dados

Condiciones iniciales:

V_0, Ω_0, P_0, T_0

F = mg

Evolucion acorde a la Ley de Newton

Resultado impredecible

Las leyes de la fisica son deterministas: dadas las condiciones iniciales yo puedo saber donde va a estar el sistema en un tiempo determinado el cual evoluciona segun la 2da ley de Newton.

Podria entonces predecir la cara del dado en que va a caer?

Teoricamente SI pero practicamente NO ya que el fenomeno es muy sensible a las condiciones iniciales. Desde el punto de vista de la **teoria del caos**:

Pequeñas diferencias en las condiciones iniciales, como las debidas a errores en las mediciones o a errores de redondeo en el cálculo numérico, pueden dar lugar a resultados muy divergentes en estos sistemas dinámicos, lo que hace imposible la predicción a largo plazo de su comportamiento en general.
Esto puede ocurrir a pesar de que estos sistemas son deterministas, lo que significa que su comportamiento futuro sigue una evolución única y está totalmente determinado por sus condiciones iniciales, sin que intervengan elementos aleatorios.
[https://en.wikipedia.org/wiki/Chaos_theory]

La "aleatoriedad" es lo que llamamos nuestra incapacidad para predecir las cosas que, a su vez, refleja nuestra falta de conocimiento sobre el sistema de interés.

Conceptos de probabilidad: fenomenos aleatorios

- Un sistema se dice que es aleatorio (random) cuando se desconocen sus condiciones iniciales o estas no pueden ser determinadas con 100% de certeza dada a la naturaleza del mismo.

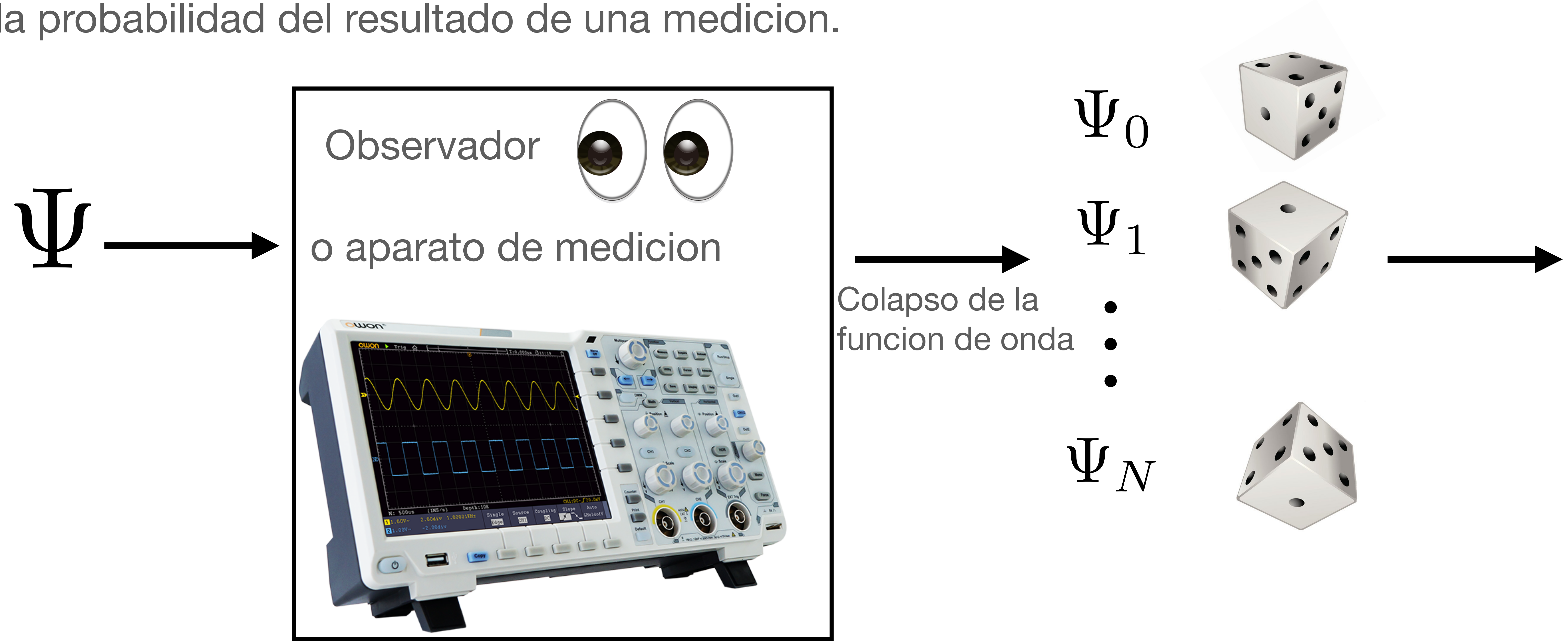
- La "aleatoriedad" es lo que llamamos nuestra incapacidad para predecir las cosas que, a su vez, refleja nuestra falta de conocimiento sobre el sistema de interés.

- Ejemplo: el estado de un sistema segun la mecanica cuantica

El estado de un sistema, e.g un electron, esta descripto por una funcion de onda.

$$\Psi$$

Esta es inaccesible por el experimento (no es observable). Solo podemos predecir la probabilidad del resultado de una medicion.



Fenomeno inherentemente aleatorio.
En este caso, por mas que tenga un aparato con 100% de precision, no podre nunca saber con exactitud el estado del sistema.
Segun la mecanica cuantica, la naturaleza es fundamentalmnente aleatoria.

Conceptos de probabilidad: fenomenos aleatorios

- Un sistema se dice que es aleatorio (random) cuando se desconocen sus condiciones iniciales o estas no pueden ser determinadas con 100% de certeza dada a la naturaleza del mismo.

- La "aleatoriedad" es lo que llamamos nuestra incapacidad para predecir las cosas que, a su vez, refleja nuestra falta de conocimiento sobre el sistema de interés.

- Ejemplo: errores de medicion

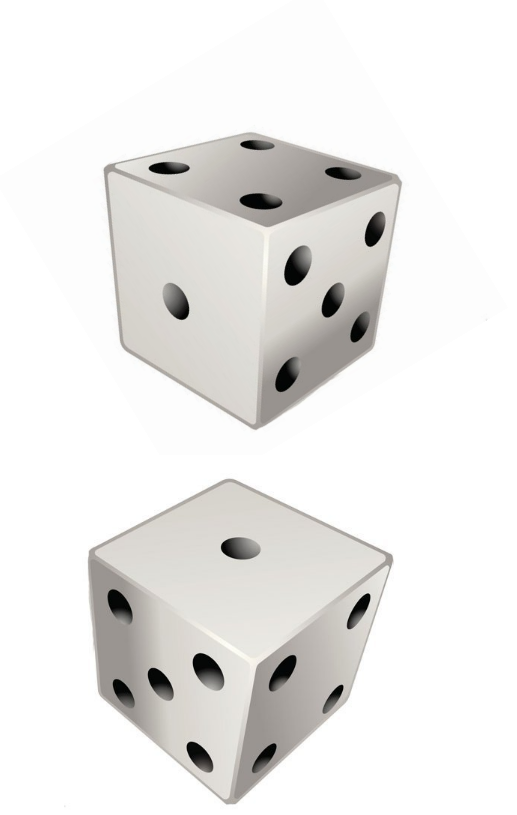
Aparato de medicion: reloj + persona que cronometra. Si repetimos el experimento los resultados van a ser distintos por mas que los corredores tarden lo mismo. Esto se debe al tiempo de reaccion y percepcion de la persona en el proceso de medicion. Los aparatos pueden tambien medir distinto dependiendo del ruido termico en la electronica.

Error estadistico: varia de una observacion a otra y este puede ser reducido al tomar el promedio de las mediciones

Error sistematico: se produce siempre, con el mismo valor, cuando utilizamos el instrumento de la misma manera y en el mismo caso

Conceptos de probabilidad: variable aleatoria

variable aleatoria X: funcion indicadora que mapea eventos a numeros reales. Esto es fundamental para poder hacer aritmetica. Es como convertir una rutina a 0 y 1 para que una computadora lo entienda.



$\xrightarrow{X}$ 4

$\xrightarrow{X}$ 1

Ω $\mathbb{R}$

Espacio de sampleo

Espacio mensurable (numeros reales)

$X \downarrow$ $X \downarrow$

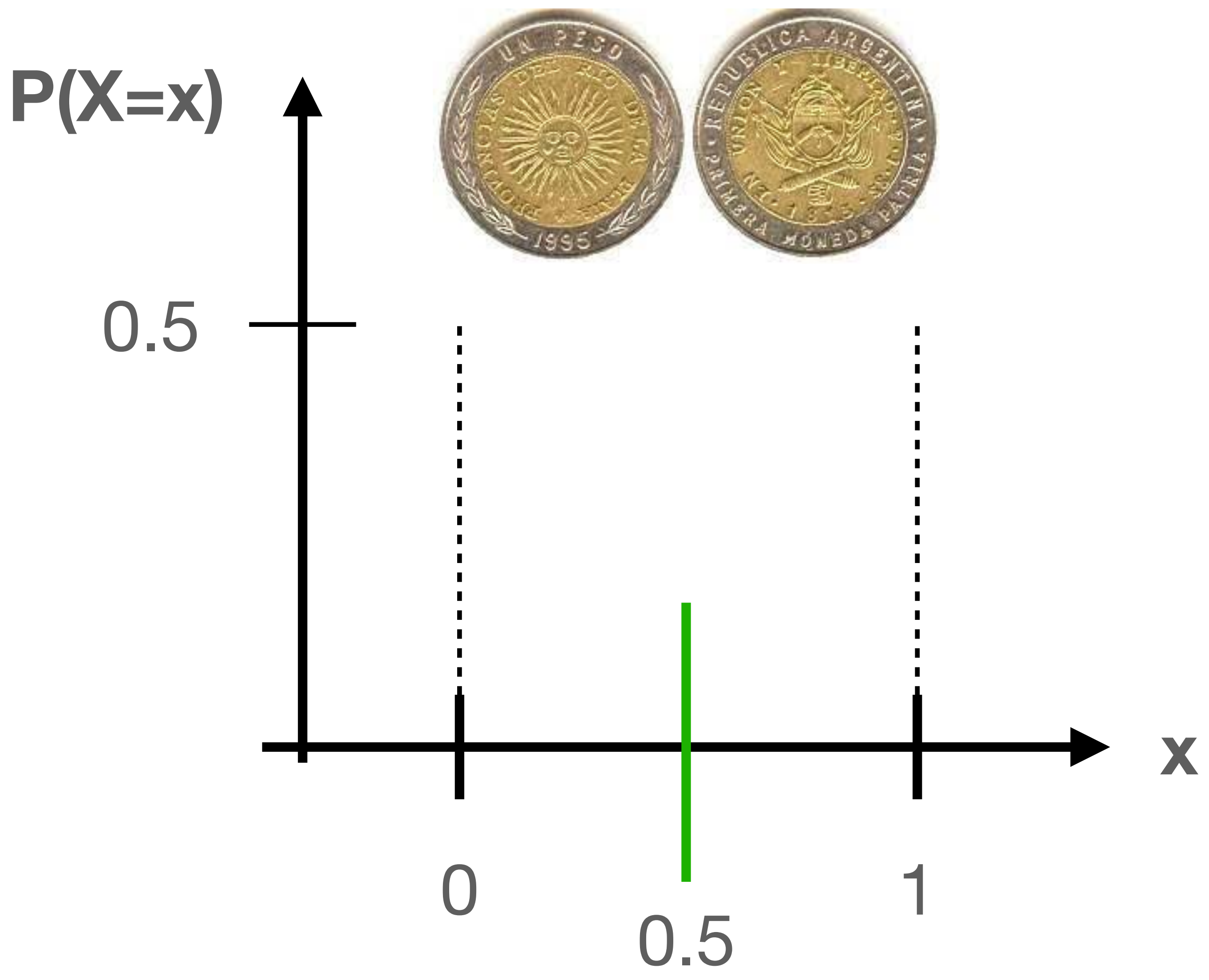
1 0

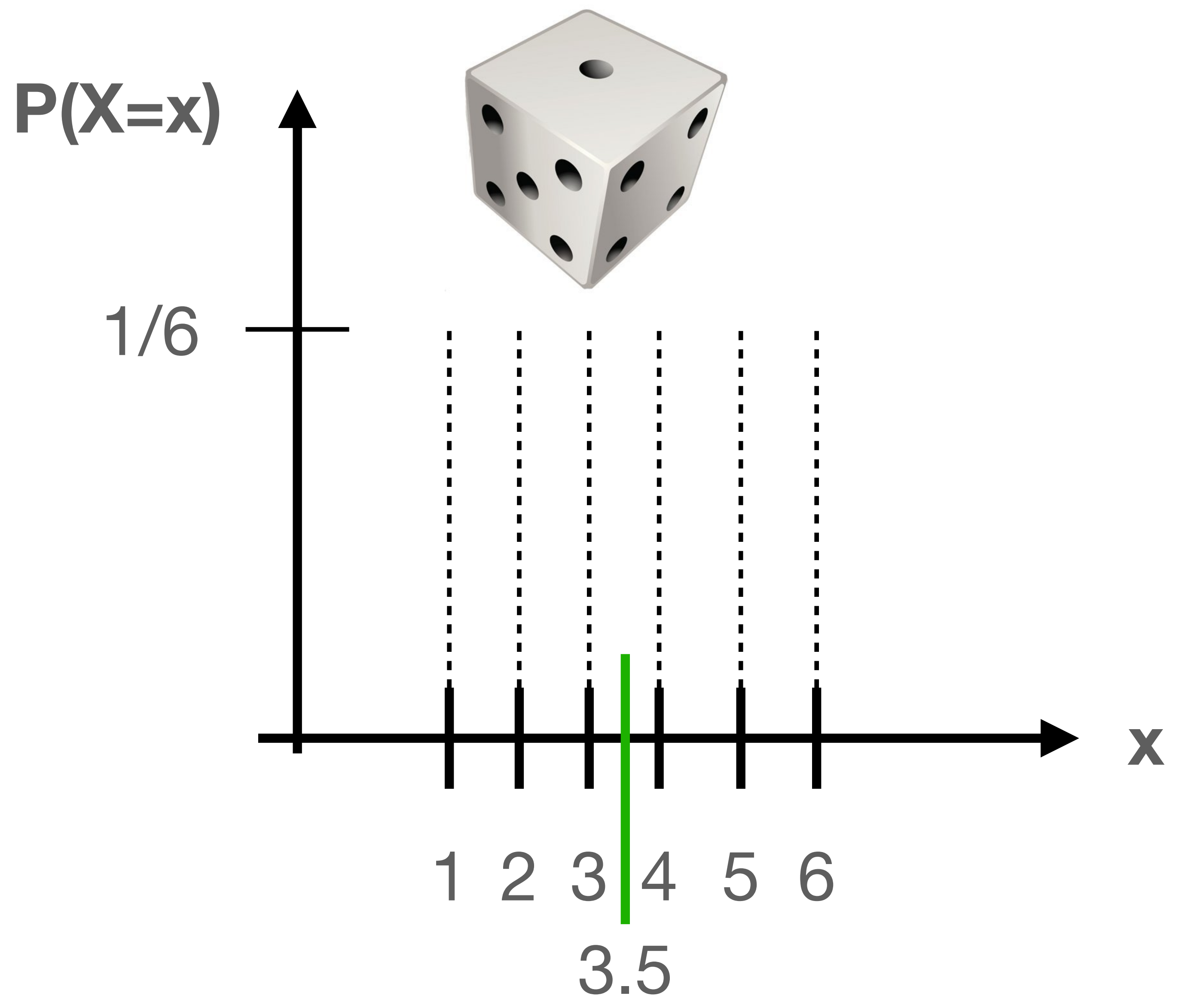
El grado de aleatoriedad de X se cuantifica con el concepto de probabilidad: P (X = x):

- Para una moneda justa, P (X=1) = 0.5
- Para un dado justo, P (X = 4) = 1/6

Descripcion de la variable X: distribucion y promedio

distribucion: Se dice que X tiene asociada una distribucion de probabilidad. Esta es una funcion que describe las distintas occurrencias de la variable aleatoria X:





Para simplificar notacion, de ahora en mas usaremos

$$P(X = x_i) = P(x_i)$$

promedio: Se define el promedio o valor esperado de X como:

$$E(X) = x_1 * P(x_1) + x_2 * P(x_2) + ... + x_n * P(x_n) = \sum_i^n x_i * P(x_i)$$

El promedio de una moneda justa: E(X) = 0*0.5 + 1*0.5 = 0.5

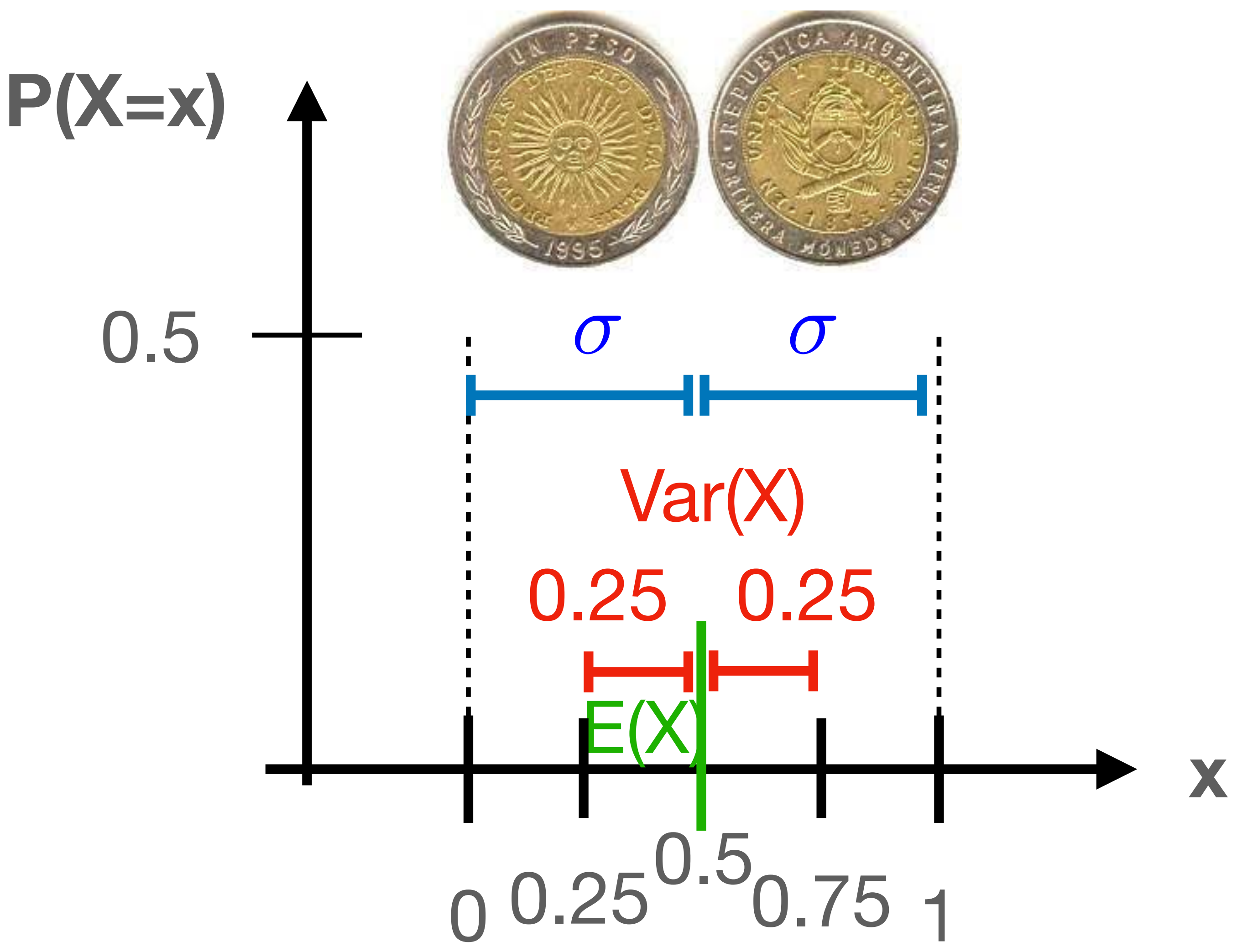
El promedio de un dado justo: E(X) = 1*1/6 + 2*1/6 + ... + 6*1/6 = 3.5

El promedio determina el valor central en el cual fluctua X (en una distribucion simetrica). Tambien podemos interpretarlo como, si tiro N veces la moneda, el valor esperado de exitos es la mitad.

Descripcion de la variable X: varianza

varianza: Se define la varianza de X como:

$$Var(X) = E[(X - E(X))^2] = \sum_{i}^{n} (x_i - E(X))^2 * P(x_i)$$

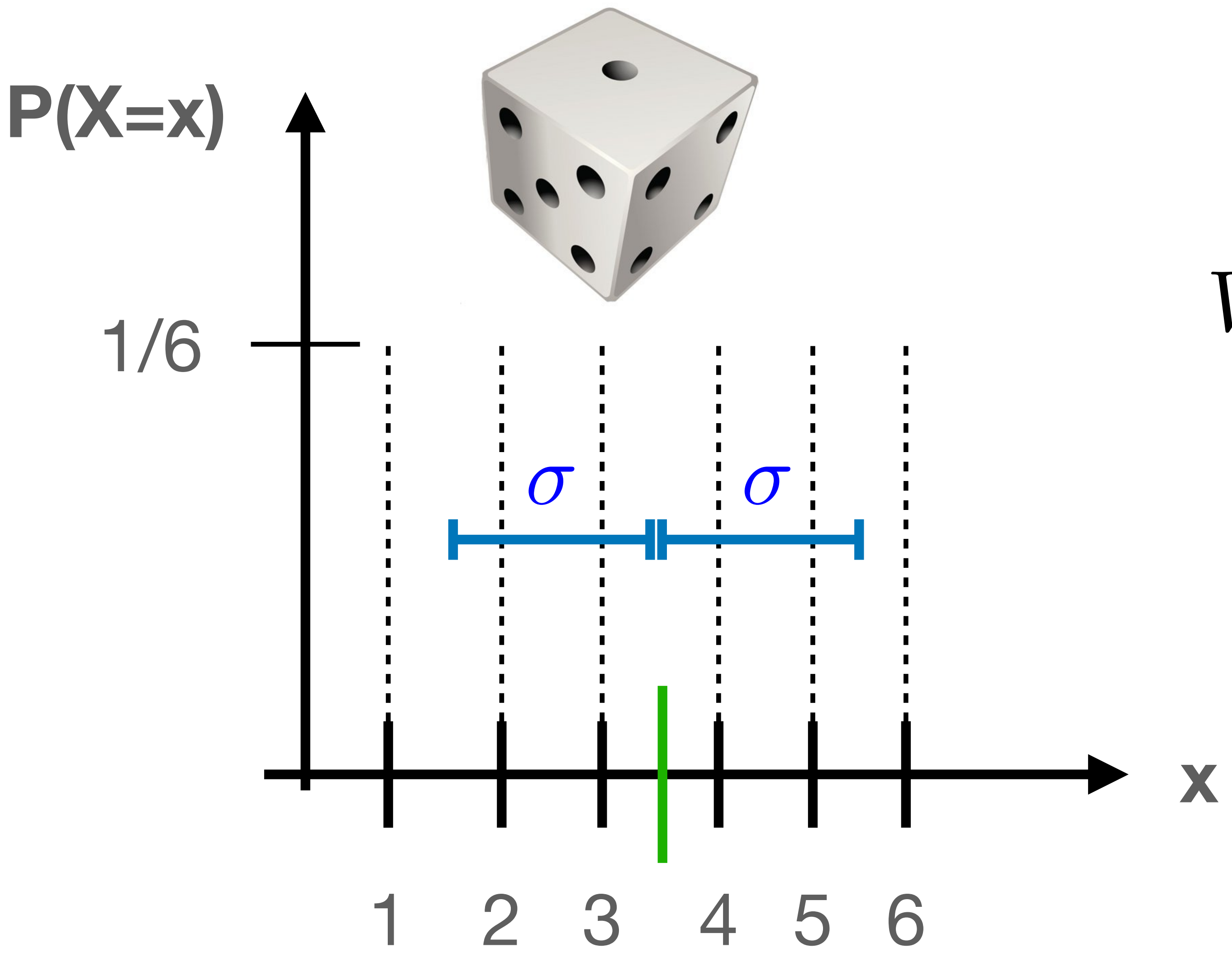


La varianza de una moneda justa:

$$Var(X) = E[(X - 0.5)^2] = (0 - 0.5)^2 * 0.5 + (1 - 0.5)^2 * 0.5 = 0.25$$

La varianza da una idea de la dispersion (o que tan alejada estan las realizaciones de X) respecto a su promedio. El desvio estandar es:

$$\sigma = \sqrt{Var(X)} = 0.5$$



La varianza de un dado justo:

$$Var(X) = E[(X - 3.5)^2] = (1 - 3.5)^2 * 1/6 + ... + (6 - 3.5)^2 * 1/6 = 2.916$$

$$\sigma = \sqrt{Var(X)} = 1.707$$

Un dado de 6 caras tiene mayor varianza que una moneda dado a que posee realizaciones mas alejadas de su promedio.
Un dado fluctua mas que una moneda ya que tiene mas posibilidades.

Descripcion de la variable X: entropia

entropia: Se define la entropia de X como:

$$H(X) = -\sum_{i}^{n} P(x_i) * \log P(x_i)$$

log en base 2: la entropia se expresa en **bits**

log en base e: la entropia se expresa en **nats**

La entropia de una moneda justa: 

$$H(X) = -(0.5 * \log_2(0.5) + 0.5 * \log_2(0.5)) = 1 \text{ bit}$$

La entropia de un dado justo: 

$$H(X) = -6 * 1/6 * \log_2(1/6) = 2.58 \text{ bits}$$

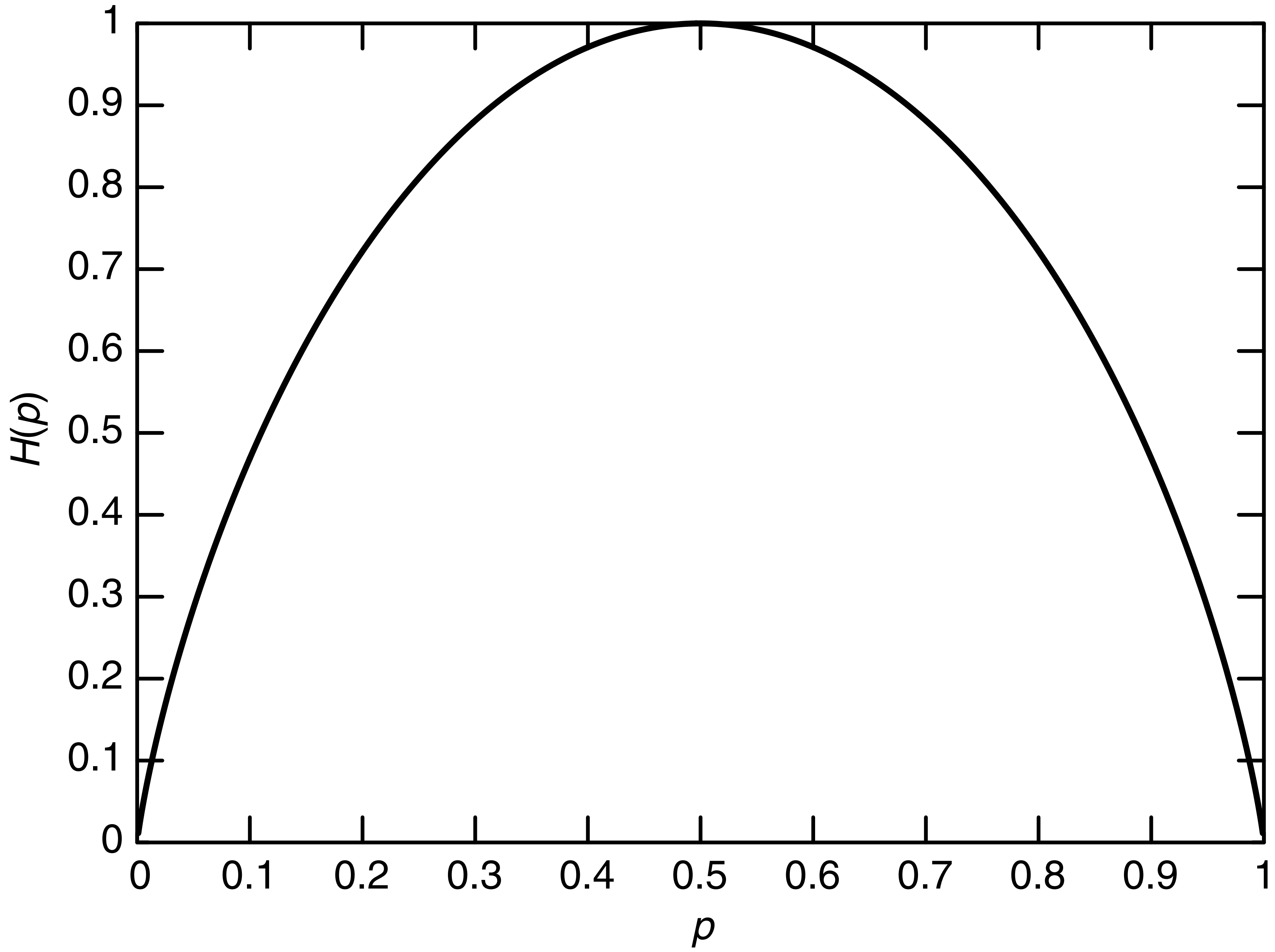
No es casual que la entropia tenga muchos significados, y en particular sea tan importante en la fisica, presente en la 2da ley de la termodinamica (como se vera mas adelante).

Descripcion de la variable X: entropia

- Desde el punto de vista de la teoria de probabilidades, la entropia es una medida de la incerteza o "aleatoriedad" de una variable aleatoria X:

Supongamos una variable aleatoria X = 0 (con probabilidad p) y X = 1 (con probabilidad 1-p)

$$H(X) = -p \log p - (1-p) \log(p)$$



El valor de H(p) es 0 cuando p = 1 o p = 0. Esto tiene sentido dado que en este caso, la variable X no seria aleatoria y por ende no habria incerteza. Por ejemplo una moneda que tenga 2 soles en ambas caras. La probabilidad que salga sol es 1 y que salga cruz es 0. En esta moneda se puede hacer predicciones con 100% de certeza de que va a caer sol. Por el contrario, la entropia alcanza un maximo cuando p = 0.5. Esto corresponde a total incerteza en lo que va a salir. Un dado tiene mayor entropia que una moneda.

Descripcion de la variable X: entropia

- Desde el punto de vista de la teoria de la informacion, la entropia es una medida de la auto-informacion de una variable X. En particular, fue idea de Shannon en 1958, en definir y aplicar este concepto y revolucionar la teoria de las comunicaciones.

- La entropia segun Shannon representa un límite matemático absoluto para la compresión sin pérdidas de los datos de una fuente en un canal sin ruido (aplicaciones en CDs y DVDs).

- La entropia puede interpretarse tambien como el numero minimo de bits en promedio que se requieren para describir una variable aleatoria X.

Ejemplo: supogamos una variable X que puede tomar 8 valores al azar con igual probabilidad.

$$H(X) = -\sum_{i=1}^{8} p(i) \log p(i) = -\sum_{i=1}^{8} 1/8 \log 1/8 = 3 \text{ bits}$$

En efecto requiero de por lo menos 3 bits para describir una variable aleatoria de 8 posibles valores: 000, 001, 010, 100, 101, 111, 010, 011

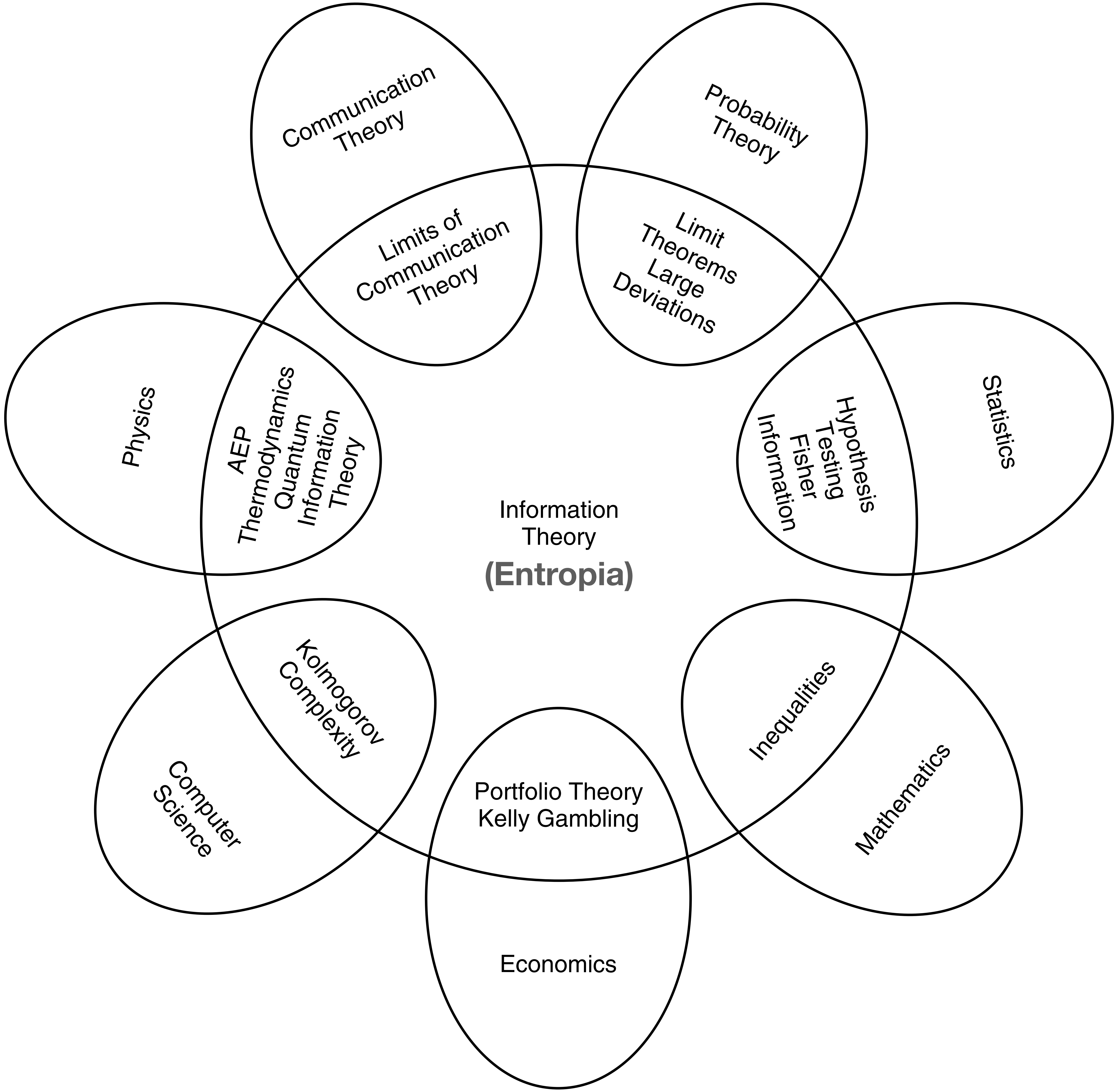
Ejemplo: supogamos una carrera de 8 caballos. La probabilidades de ganar de cada uno de los 8 caballos es: (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64). Supongamos que quiero enviar un mensaje indicando que caballo gano la carrera: una alternativa es enviar el indice del caballo que gano, e.g (1,2,3,4,5,6,7,8). Para esto se requieren 3 bits.
Pero dado que las probabilidades no son uniformes, podria ser mas eficiente y podria usar menos bits para los caballos mas probables y mas bits para los mas improbables: (0, 10, 110, 1110, 111100, 111101, 111110, 111111). Notar que esto es una secuencia valida, ya que el receptor podra descifrar univocamente los caballos ganadores en una secuencia dada:

$$H(X) = -1/2 \log 1/2 - 1/4 \log 1/4 - 1/8 \log 1/8 - 1/16 \log 1/16 - 4 * 1/64 \log 1/64 = 2 \text{ bits}$$

$$\underbrace{10}_{C2}\underbrace{110}_{C3}\underbrace{0}_{C1}\underbrace{111100}_{C5}\underbrace{0}_{C1}\underbrace{111111}_{C8}$$

En este caso, el numero de bits requeridos en promedio para describir el caballo ganador es:

$$1 * 1/2 + 2 * 1/4 + 3 * 1/8 + 4 * 1/16 + 4 * 6 * 1/64 = 2 \text{ bits} = H(X)$$



Communication Theory
Limits of Communication Theory
Probability Theory
Limit Theorems Large Deviations
Physics
AEP Thermodynamics Quantum Information Theory
Statistics
Hypothesis Testing Fisher Information
Information Theory
(Entropia)
Computer Science
Kolmogorov Complexity
Inequalities
Mathematics
Portfolio Theory Kelly Gambling
Economics

Interpretacion de la probabilidad frecuentista vs bayesiana

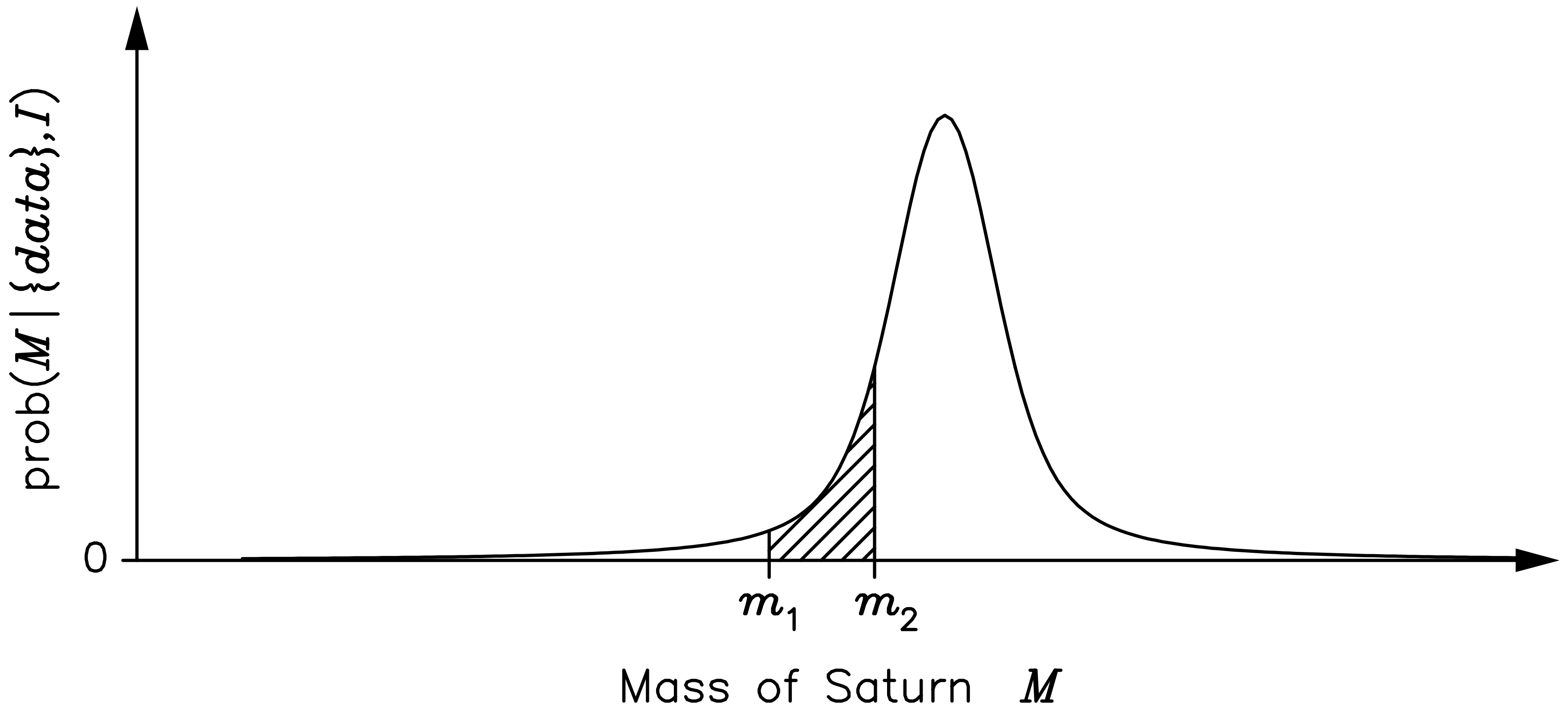
Interpretacion frecuentista: la probabilidad como frecuencia relativa limite:

$$P(X=x) = \lim_{n\to\infty} \frac{\text{numero de ocurrencias de X=x en n tiradas}}{n}$$

por ejemplo para una moneda justa, es esperable que haciendo muchas tiradas (n ~1000), la fraccion "numero de caras" / "numero de tiradas" ~ 0.5, y decimos que la probabilidad de obtener cara es P(x=0) = 0.5.

Esta interpretacion funciona bien para describir por ejemplo experimentos fisicos repetibles. Pero tiene su limitacion para eventos "unicos", como el origen del universo. O por ejemplo para estimar parametros, como la masa de un planeta. De acuerdo a esta postura, la masa de Saturno es una constante y no una variable aleatoria. No resulta muy practico imaginarse infinitos universos creados en que todo permanece constante salvo la masa de Saturno. Para resolver esta limitacion, los frecuentistas inventaron la estadistica!. Lamentablemente en estadistica no hay una receta unificadora para estimar parametros, lo cual conlleva a una serie de deficiencias a la hora de transmitir estos conocimientos en la estadistica ortodoxa (y que es lamentablemente lo que se enseña).

Interpretacion bayesiana: la probabilidad como grado de confianza o verosimilitud:
hasta que punto podemos pensar que algo es cierto, basondose en las pruebas disponibles.
Segun este punto de vista, la masa de Saturno es una variable aleatoria, y la probabilidad representa el grado de confianza de que el verdadero valor de la masa se encuentre en un rango determinado (m1, m2):



Veremos mas adelante como Laplace en 1812 re-descubrio el teorema de Bayes y lo aplico exitosamente para estimar la masa de Saturno, algo que los frecuentistas no podian hacer en esa epoca…
el valor actual de Saturno luego de 150 años de datos, comparado con el de Laplace difiere solo en un 0.63%!!

Derivacion de las leyes de la probabilidad

Formalizacion axiomatica segun Kolmogorov (1933), preferidos por los frecuentistas.

1. Para todo conjunto A de eventos posibles del espacio muestral S, P(A) > 0
2. Para dos subconjuntos de eventos A y B disjuntos: $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$
3. La probabilidad asignada al universo de sampleo S es P(S) = 1 (algo tiene que ocurrir)

Ejemplo: para un dado justo de 6 caras asigno P(X=i) = 1/6
(1) se cumple ya que P(X=i) > 0
(3) se cumple ya que las probabilidades suman 1: P(X=1) + P(X=2) + … + P(X=6) = 1
(2): dado dos eventos disjuntos, que salga el 1 o el 2 (no pueden salir los dos a la vez), se cumple que P(X=1 o X=2) (union) = P(X=1) + P(X=2) = 1/6 + 1/6 = 2/6

Propiedades derivadas:

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

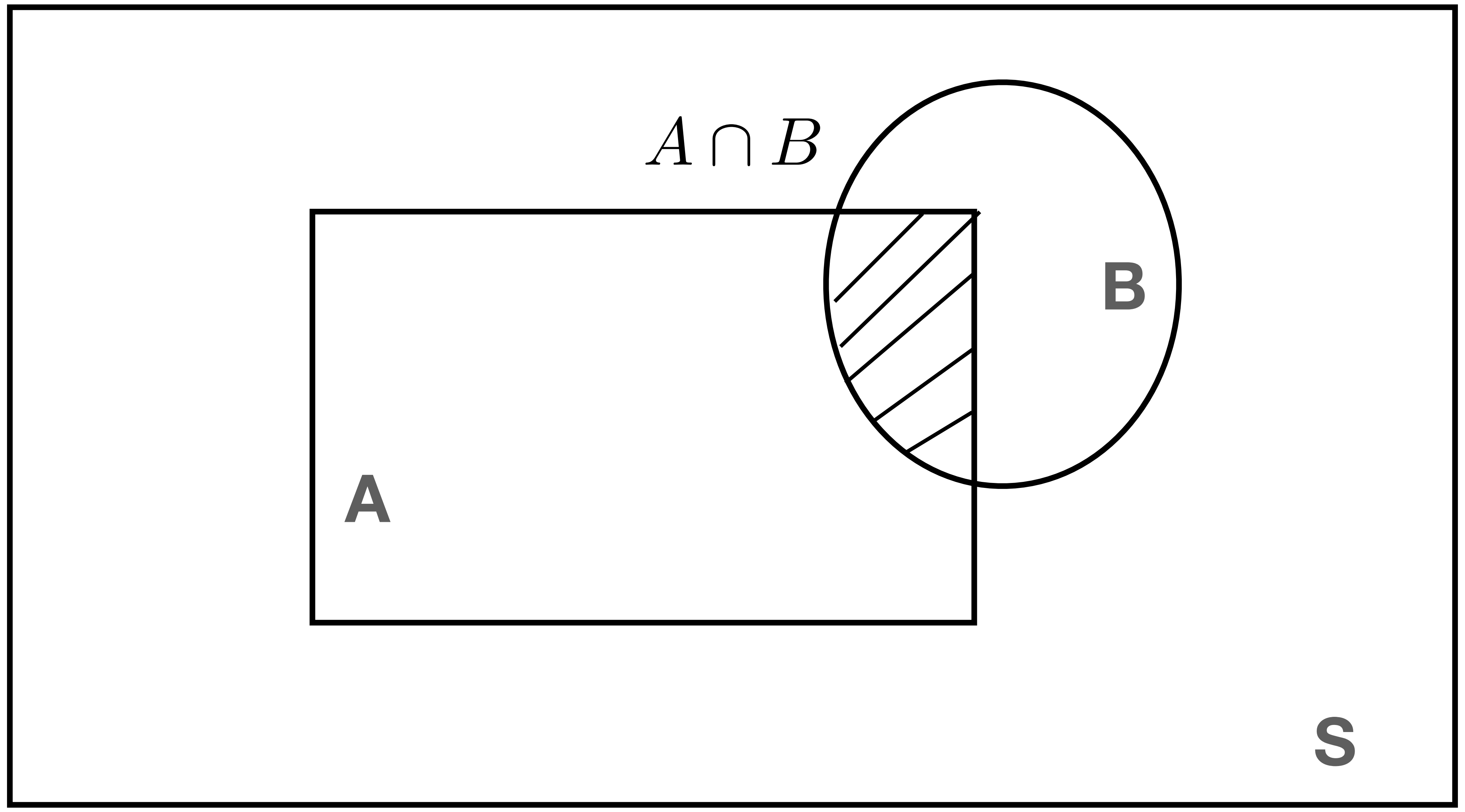
Demostracion:

$$P(A \cup B) = P(A \cup B \backslash A) = P(A) + P(B \backslash A) \quad (2)$$

$$P(A \cup B) = P(A) + P(B \backslash (A \cap B))$$

$$P(B) = P(B \backslash (A \cap B)) + P(A \cap B)$$

$$\implies P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Probabilidad condicional

Dado eventos A y B, donde P(B) > 0, se define la probabilidad condicional P(A | B): probabilidad de obtener A dado que sucedio B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{o} \quad P(A \cap B) = P(A|B)P(B) \qquad (1)$$

Notar que esta definicion es consistente con la nocion de frecuencia relativa como probabilidad:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \lim_{N \to \infty} \frac{\frac{\text{numero de ocurrencias X=A y X=B}}{N}}{\frac{\text{numero de ocurrencias X=B}}{N}} = \frac{\text{numero de ocurrencias X=A y X=B}}{\text{numero de ocurrencias X=B}}$$

Dos eventos A y B se dicen independientes, si:

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

En este caso, $P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$

Y la nocion de independencia resulta de esta forma intuitiva.

Por otro lado:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \qquad (2)$$

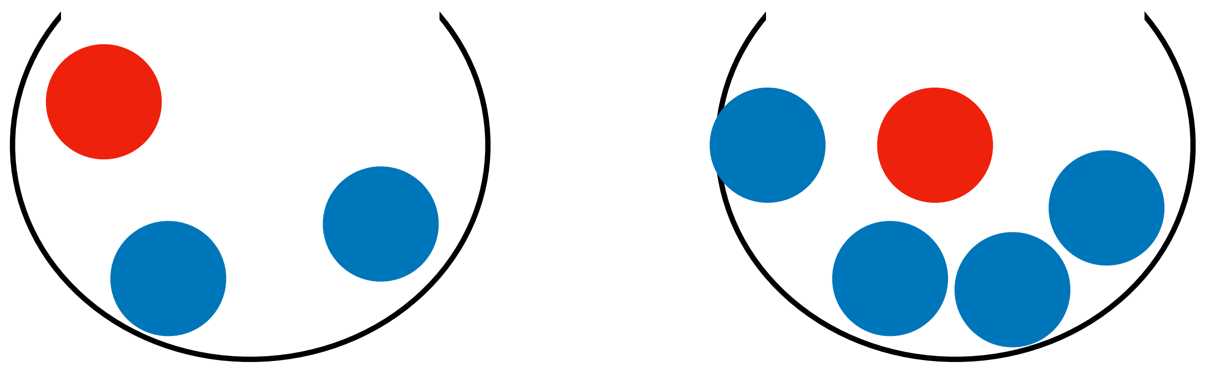
Juntando (1) y (2):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Formula de Bayes

Probabilidad condicional

Ejemplo:



C1 C2

P(roja) = 4 / 15 P(azul) = 11/15

	Rojo	Azul
C1	1	2
C2	1	4

P(rojo | C1) = 1/3

P(azul | C1) = 2/3

P(rojo | C2) = 1/5

P(azul | C2) = 4/5

P(C2 | roja) = ? P(C2 | azul) = ?

P(C1 | roja) = ? P(C1 | azul) = ?

Cual es la probabilidad de sacar una bola roja en la caja 1?

$$P(x = A) = \frac{\text{numero de eventos x=A}}{\text{numero de posibilidades totales}} \longrightarrow$$

Se justificara en un rato

$$P(\text{rojo}|C1) = \frac{\text{numero de bolas rojas en C1}}{\text{numero de bolas totales en C1}} = 1/3$$

Cual es la probabilidad de sacar una bola roja en la caja 2?

P(rojo|C2) = 1/5

Cual es la probabilidad de sacar una bola roja en la caja 1 o 2?

P(rojo) =P(rojo y C1) + P(rojo y C2) = P(rojo|C1)P(C1) + P(rojo|C2)P(C2) =

1/3*1/2 + 1/5*1/2 = 4/15

Cual es la probabilidad de que la caja es la 1 dado que saque una roja?

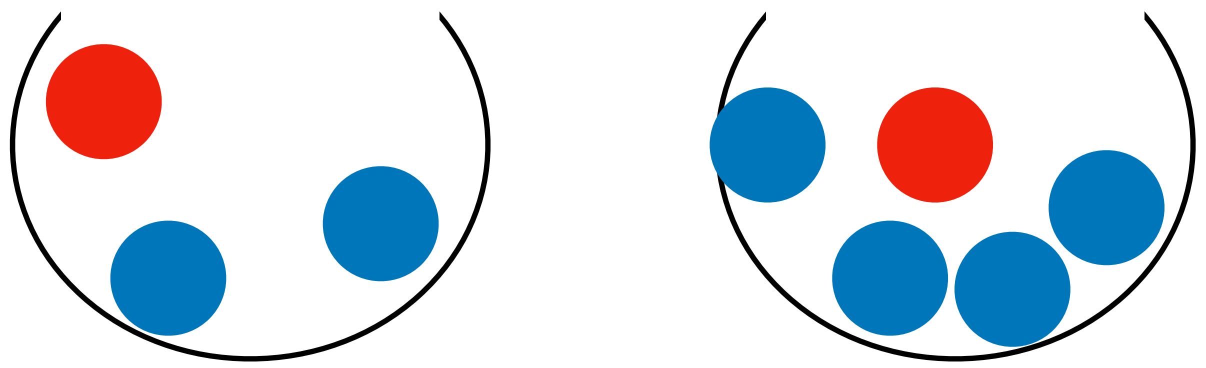
P(C1 | roja) = ?

Cual es la probabilidad de que la caja es la 1 dado que saque una azul?

P(C1 | azul) = ?

Probabilidad condicional

Ejemplo:



C1 C2

P(roja) = 4 / 15 Rojo Azul P(azul) = 11/15

	Rojo	Azul	
C1	1	2	P(rojo | C1) = 1/3 P(azul | C1) = 2/3
C2	1	4	P(rojo | C2) = 1/5 P(azul | C2) = 4/5

P(C2 | roja) = ? P(C2 | azul) = ?

P(C1 | roja) = ? P(C1 | azul) = ?

Cual es la probabilidad de sacar una bola roja en la caja 1?

$$P(A) = \frac{\text{numero de eventos favorables}}{\text{numero de posibilidades totales}} \longrightarrow$$ Se justificara en un rato

$$P(\text{rojo}|C1) = \frac{\text{numero de bolas rojas en C1}}{\text{numero de bolas totales en C1}} = 1/3$$

Cual es la probabilidad de sacar una bola roja en la caja 2?
P(rojo|C2) = 1/5

Cual es la probabilidad de sacar una bola roja en la caja 1 o 2?
P(rojo) =P(rojo y C1) + P(rojo y C2) = P(rojo|C1)P(C1) + P(rojo|C2)P(C2) =
1/3*1/2 + 1/5*1/2 = 4/15

Cual es la probabilidad de que la caja es la 1 dado que saque una roja?

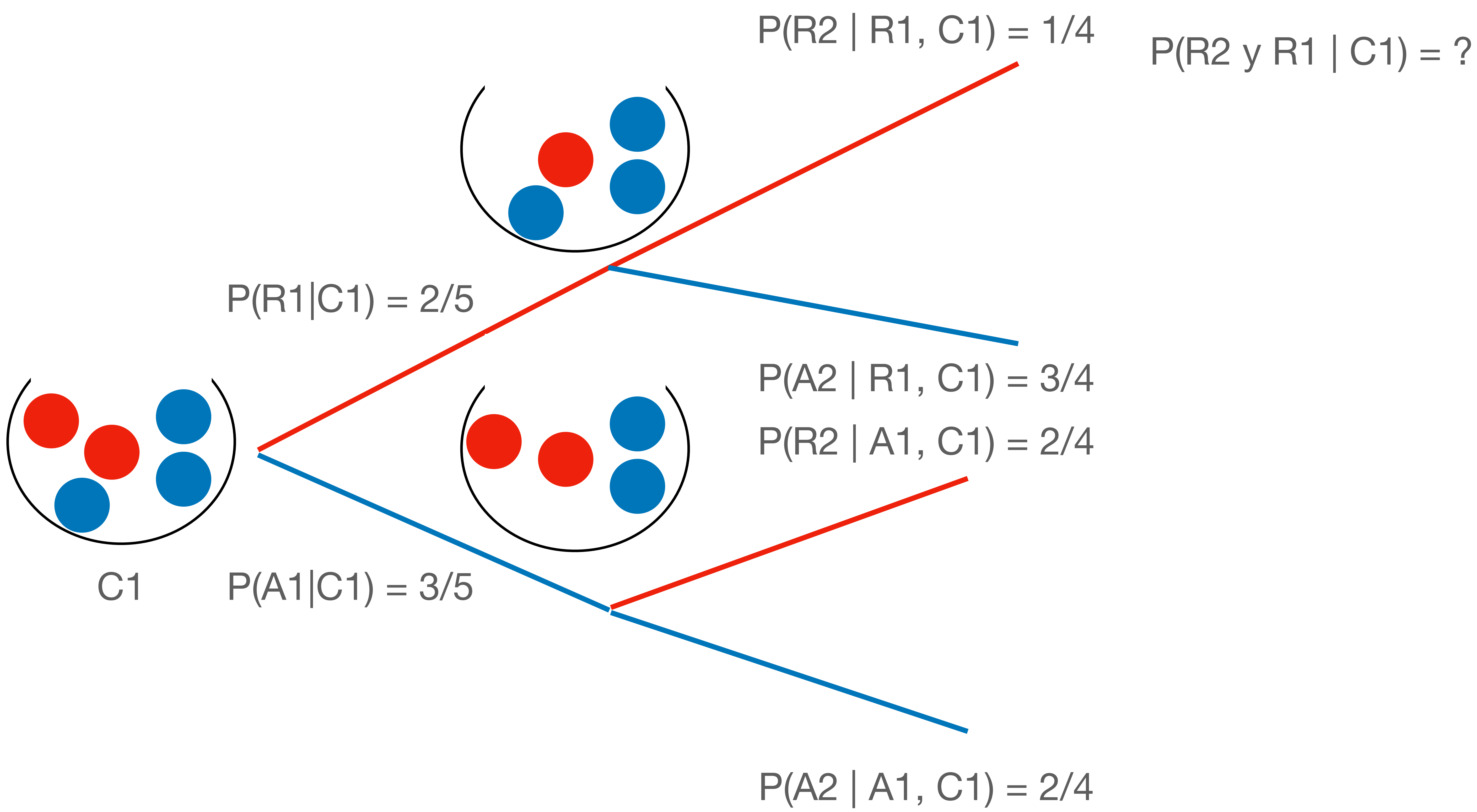
P(C1 | roja) = P(roja | C1) P(C1) / P(roja)

Cual es la probabilidad de que la caja es la 1 dado que saque una azul?

P(C1 | azul) = P(azul | C1) P(C1) / P(azul)

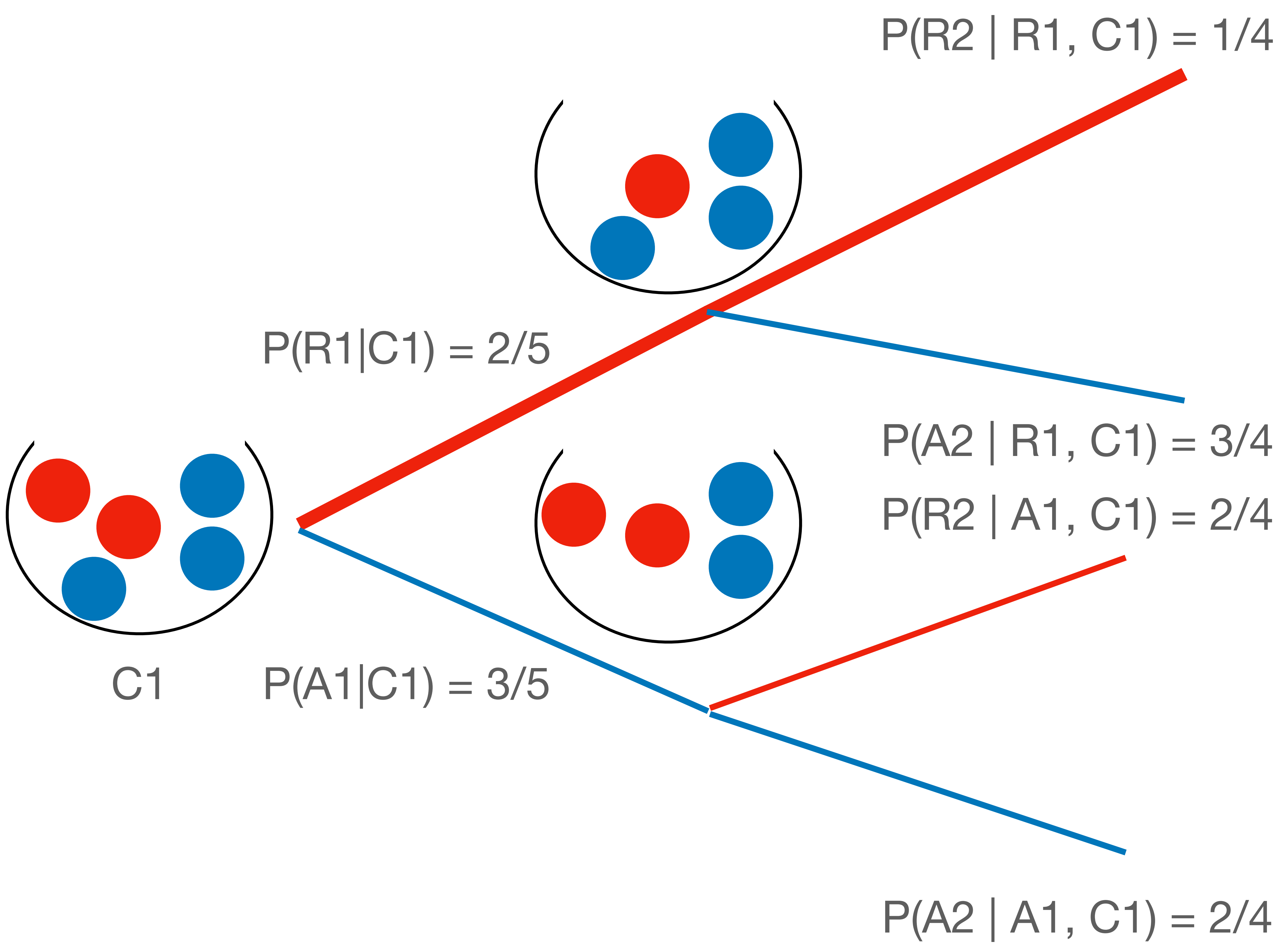
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Probabilidad condicional



Probabilidad condicional

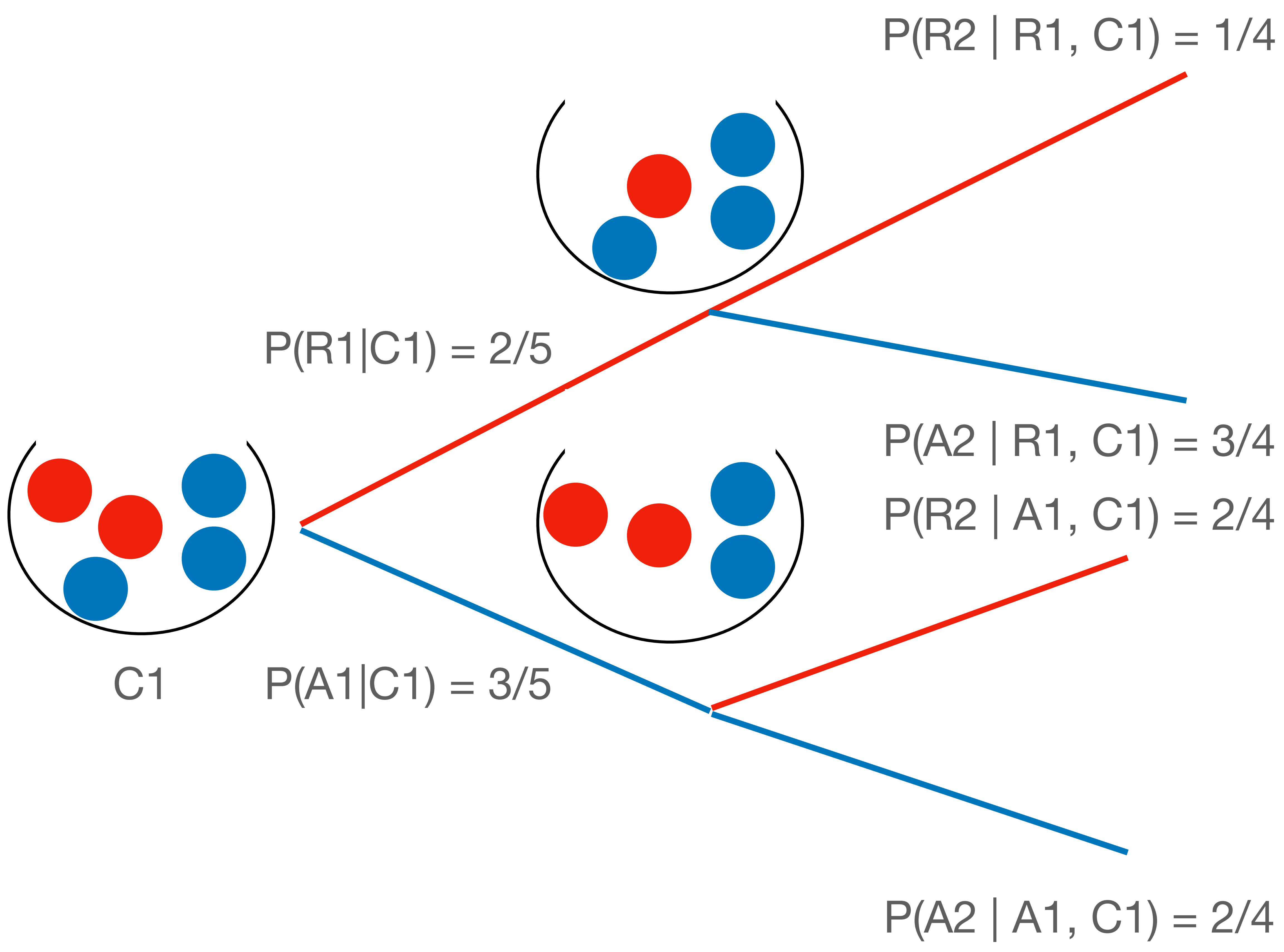
$$P(A \cap B) = P(A|B)P(B)$$



P(R2 y R1 | C1) = ?

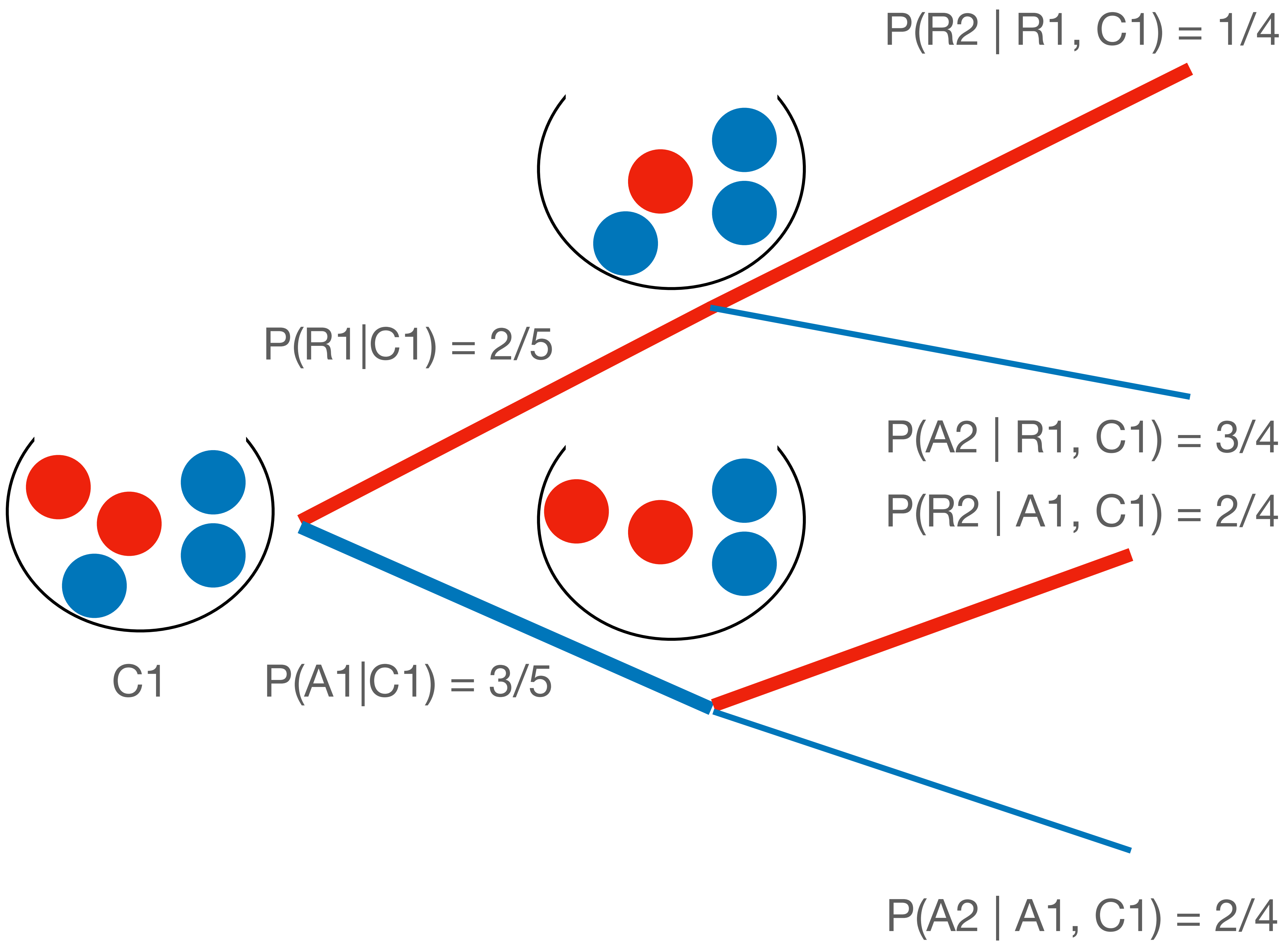
P(R2 y R1 | C1) = P(R2|R1, C1) P(R1|C1) = 1/4*2/5 = 1/10

Probabilidad condicional



P(R2 | C1) = ?

Probabilidad condicional



P(R2 | C1) = ?

P(R2 | C1) = P(R2 y R1 | C1) + P(R2 y A1 | C1)
= 1/10 + 3/10 = 4/10

La probabilidad como un estado de conocimiento

Frecuentistas: Aleatoriedad, repeticion de experimentos

Bayesianos: Incerteza, estado de conocimiento

La critica de los frecuentistas esta en que el estado de conocimiento o incerteza que uno puede tener acerca de un problema resulta subjetivo, mientras que la aleatoriedad proveniente de la repeticion de experimentos es inherente a la naturaleza y por ende es una descripcion objetiva mas cerca de la “verdad”. **Pero esta el concepto de aleatoriedad bien definido?**

Podemos pensar que la aleatoriedad resulta en realidad de nuestra ignorancia e incapacidad de conocer todas las variables.

Una computadora puede generar aproximadamente numeros aleatorios mediante un algoritmo conocido como *rand* que genera numeros random reproducibles con una *semilla.* Que pasa si yo conociera el algoritmo de *rand* y su *semilla?,* podria de esta forma predecir lo que la computadora va a generar.

Uno podria objetar que en la naturaleza existen sistemas puramentes aleatorios, por ejemplo el **caos** (como la tirada de un dado) o **sistemas cuanticos**. Pero en el 1er caso justamente la aparente aleatoriedad en el largo plazo se produce por no poder conocer 100% las condiciones iniciales, la evolucion del sistema es puramente determinista mediante la 2da ley de Newton.

Podemos de esta forma pensar a la **aleatoriedad** como la incapacidad de predecir cosas que, a su vez, refleja nuestra falta de conocimiento sobre el sistema de interés.

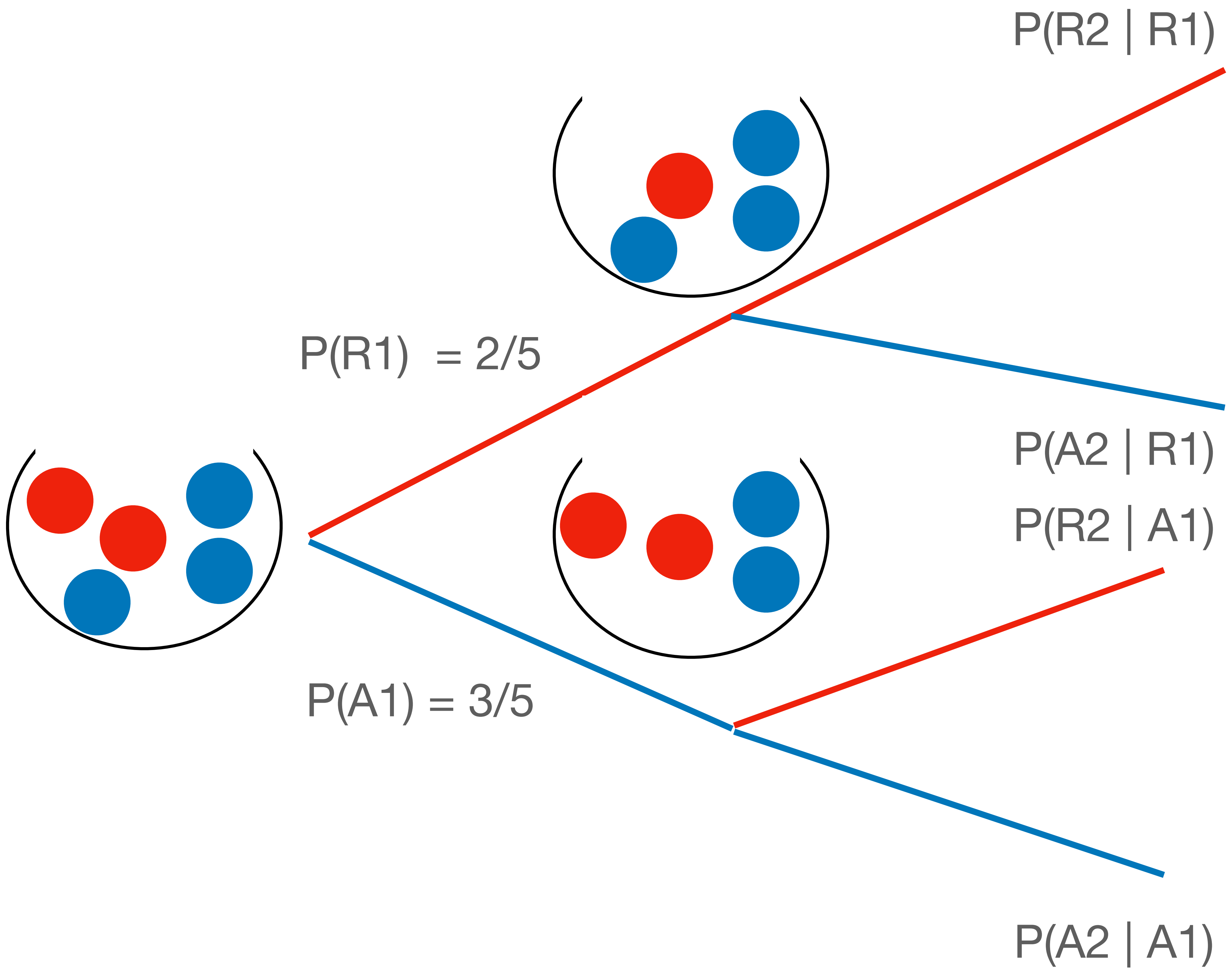
Segun los bayesianos, la probabilidad representa un estado de conocimiento (cuantificacion de nuestra ignorancia) más que una entidad físicamente real.

Bayesiano o Frecuentista?

https://www.youtube.com/watch?v=GEFxFVESQXc&t=247s

La probabilidad como un estado de conocimiento

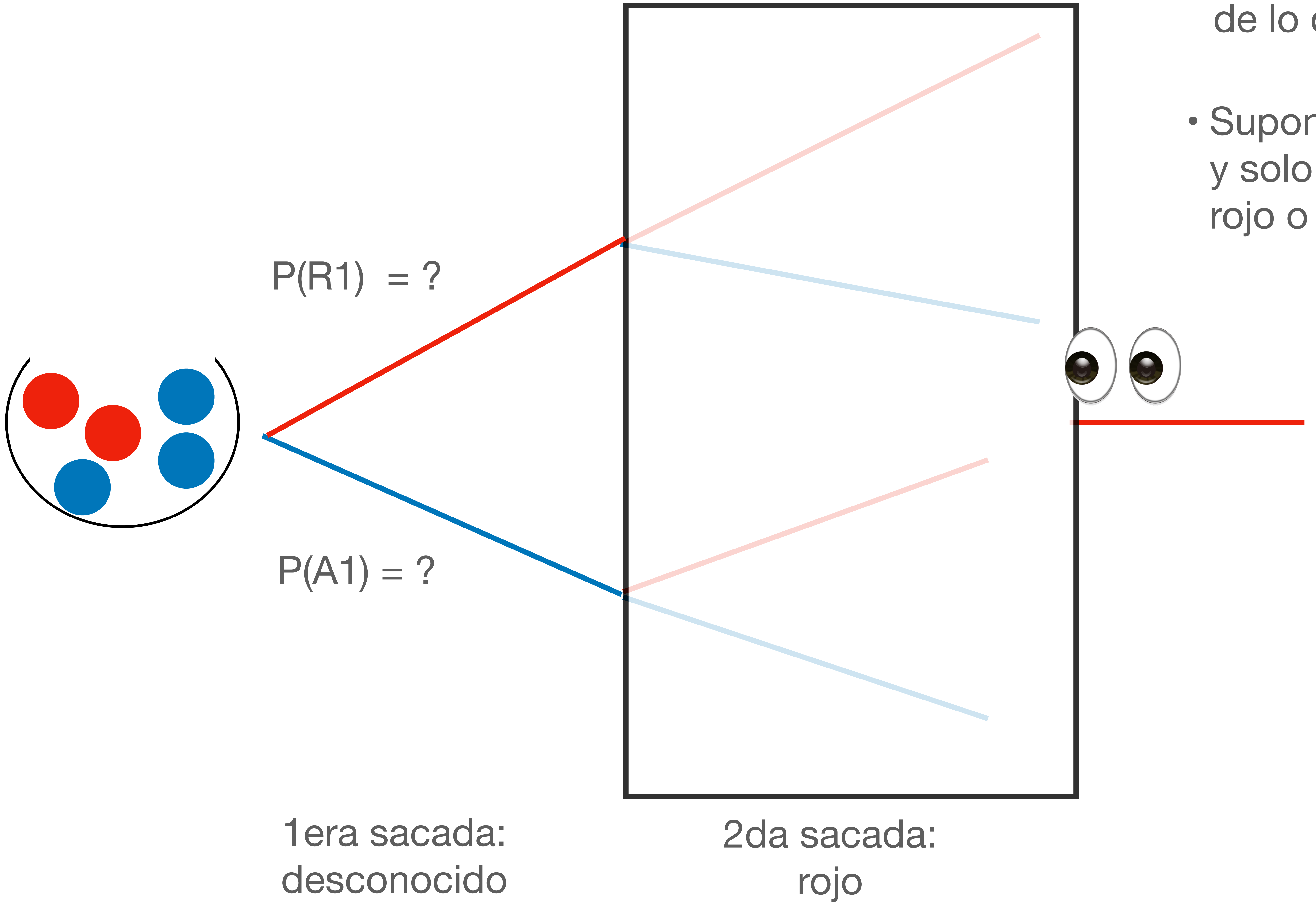
Ejemplo (Jaynes 1989):



- La probabilidad de obtener rojo o azul en la 2da sacada depende de lo que salio en la primera.

La probabilidad como un estado de conocimiento

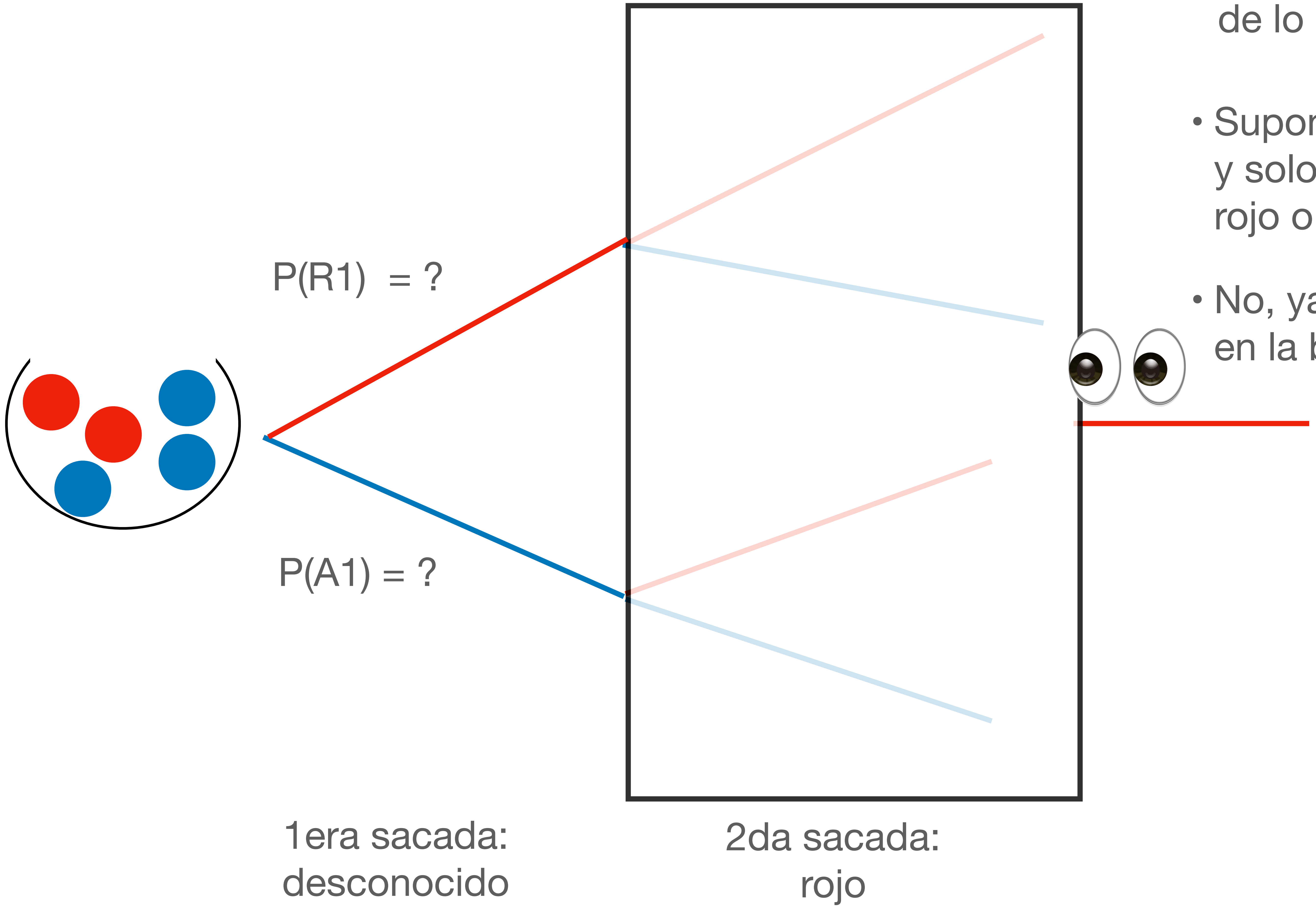
Ejemplo (Jaynes 1989):



- La probabilidad de obtener rojo o azul en la 2da sacada depende de lo que salio en la primera.
- Supongamos ahora que no sabemos que salio en la 1era sacada y solo sabemos lo que salio en la 2da. La probabilidad haber sacado rojo o azul en la 1era cambia?

La probabilidad como un estado de conocimiento

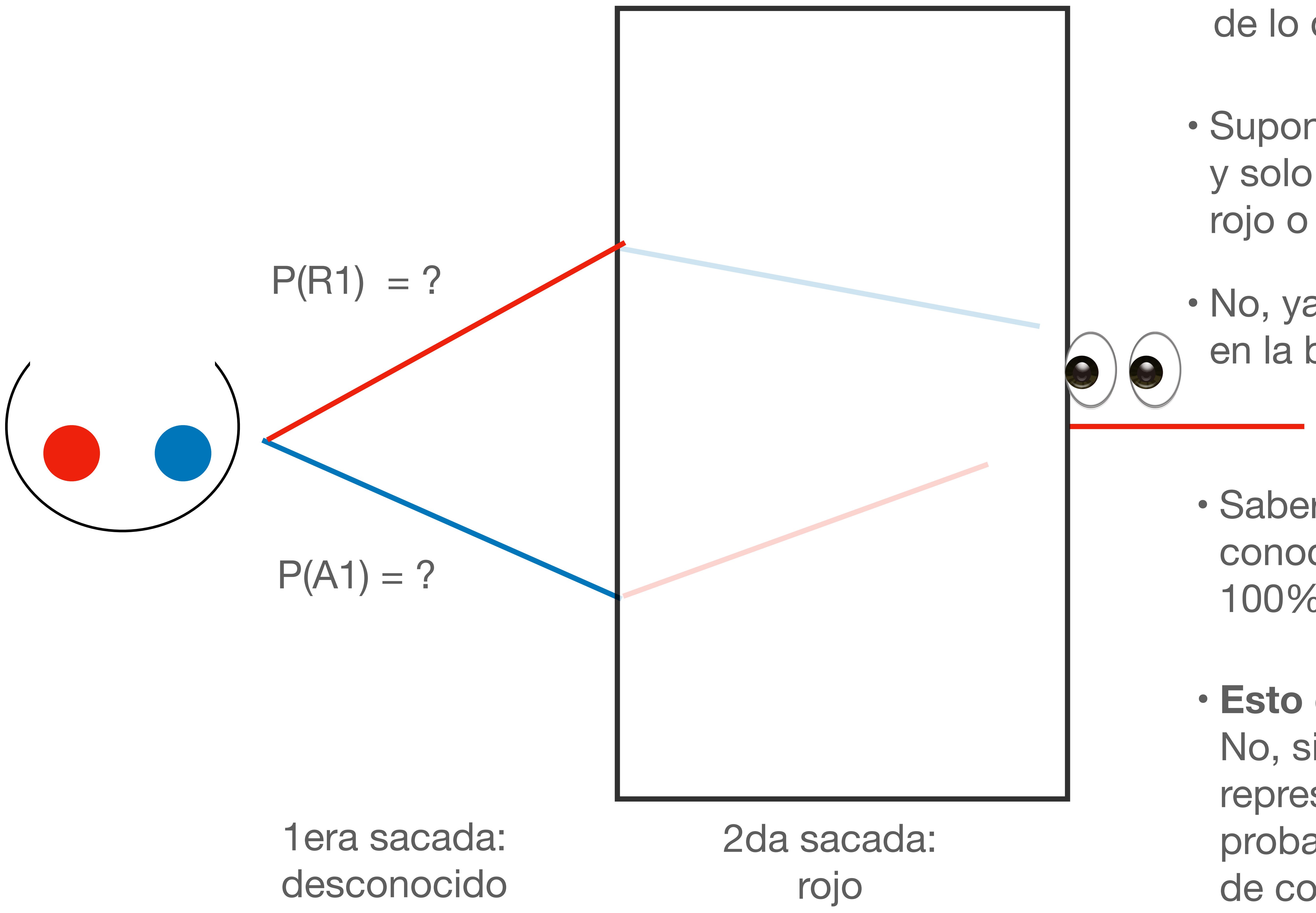
Ejemplo (Jaynes 1989):



- La probabilidad de obtener rojo o azul en la 2da sacada depende de lo que salio en la primera.
- Supongamos ahora que no sabemos que salio en la 1era sacada y solo sabemos lo que salio en la 2da. La probabilidad haber sacado rojo o azul en la 1era cambia?
- No, ya que al momento de la 1era sacada habian 2 rojas y 3 azules en la bolsa, por ende la probabilidad es P(R1) = 2/5 y P(A1) = 3/5.

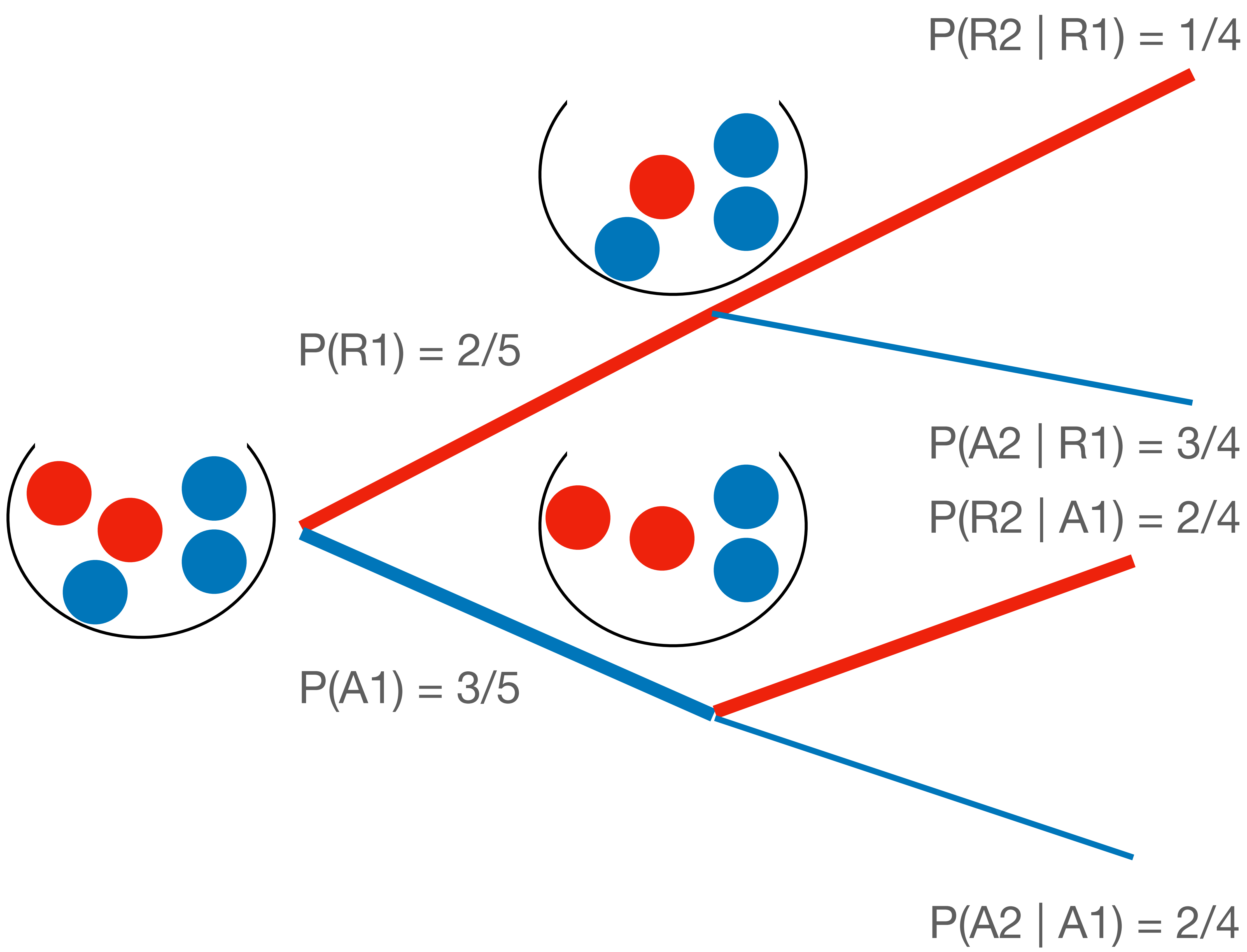
La probabilidad como un estado de conocimiento

Ejemplo (Jaynes 1989):



- La probabilidad de obtener rojo o azul en la 2da sacada depende de lo que salio en la primera.
- Supongamos ahora que no sabemos que salio en la 1era sacada y solo sabemos lo que salio en la 2da. La probabilidad haber sacado rojo o azul en la 1era cambia?
- ~~No, ya que al momento de la 1era sacada habian 2 rojas y 3 azules en la bolsa, por ende la probabilidad es P(R1) = 2/5 y P(A1) = 3/5.~~
- Saber lo que salio en la 2da sacada, por ejemplo rojo, modifica mi conocimiento acerca de lo que salio en la 1era, se con probabilidad 100% que la 1era sacada tuvo que ser azul.
- **Esto quiere decir que el presente modifica el pasado?** No, simplemente que las probabilidades condicionales representan conexiones lógicas más que causales y que la probabilidad puede interpretarse mas bien como un estado de conocimiento.

Probabilidad marginal



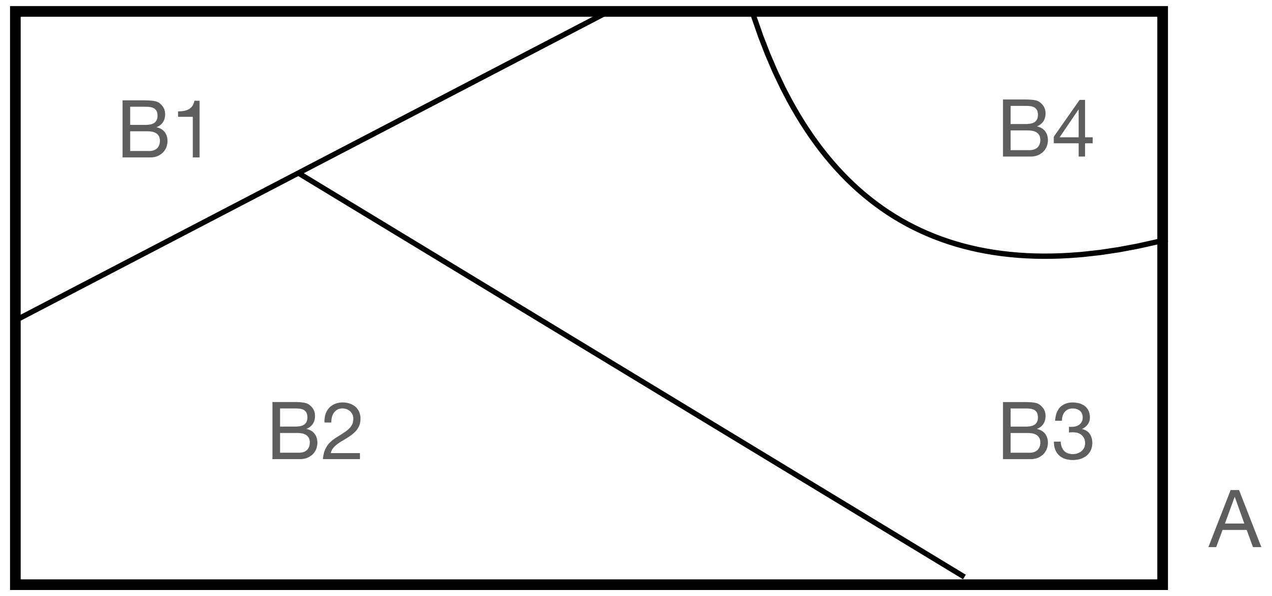
$$P(A \cap B) = P(A|B)P(B)$$

P(R2) = ?

P(R2) = P(R2 y R1) + P(R2 y A1)
= P (R2|R1)* P(R1) + P (R2|A1)* P(A1)
= 1/10 + 3/10 = 4/10

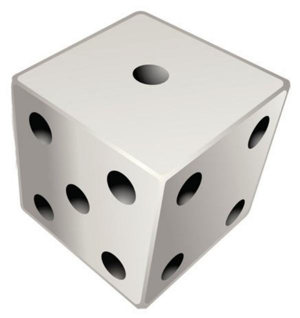
En general se cumple que:

$$P(A) = \sum_i^n P(A \cap B_i) = \sum_i^n P(A|B_i)P(B_i)$$



(dado un subconjunto B tal que)

$$\sum_i^n P(B_i) = 1$$



1	2	3
4	5	6

Impar	Par

P(2) = 1/6 = P(2|par)P(par) + P(2|impar)P(impar) = 1/3 * 1/2 + 0 * 1/2 = 1/6

P(par) = 1/2 = P(par|1)P(1) + … + P(par|6)P(6) = 3*1/6 = 1/2

Aplicacion de Bayes: test de enfermedad

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Formula de Bayes $$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i^n P(B|A_i)P(A_i)}$$

$$P(E|+) = \frac{P(+|E)\overbrace{P(E)}^{\text{Prior}}}{\underbrace{P(+|E)}_{\substack{\text{Sensibilidad} \\ \text{(falsos negativos)}}}P(E) + \underbrace{P(+|-E)}_{\substack{\text{1-Especificidad} \\ \text{(falsos positivos)}}}P(-E)}$$

$$P(E) = \frac{\text{cantidad de casos con E}}{\text{poblacion total}}$$

Test que da siempre negativo tiene **especificidad** de 100%. Pero esto tendria muchos **falsos negativos**.
Test que da siempre positivo tiene **sensibilidad** de 100%. Pero esto tendria muchos **falsos positivos**.

Un buen test tiene que tener ambos alta especificidad y sensibilidad

Supongamos que quiero evaluar una enfermad poco usual en la poblacion con un test de sensibilidad y especificidad del 99%.
P(E) = 0.001, P(+|E) = 0.99, P(+|-E) = 0.01

$$P(E|+) = \frac{0.99 * 0.001}{0.99 * 0.001 + 0.01 * 0.999} = 9\%$$

Aplicacion de Bayes: test de enfermedad I

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Formula de Bayes $$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i^n P(B|A_i)P(A_i)}$$

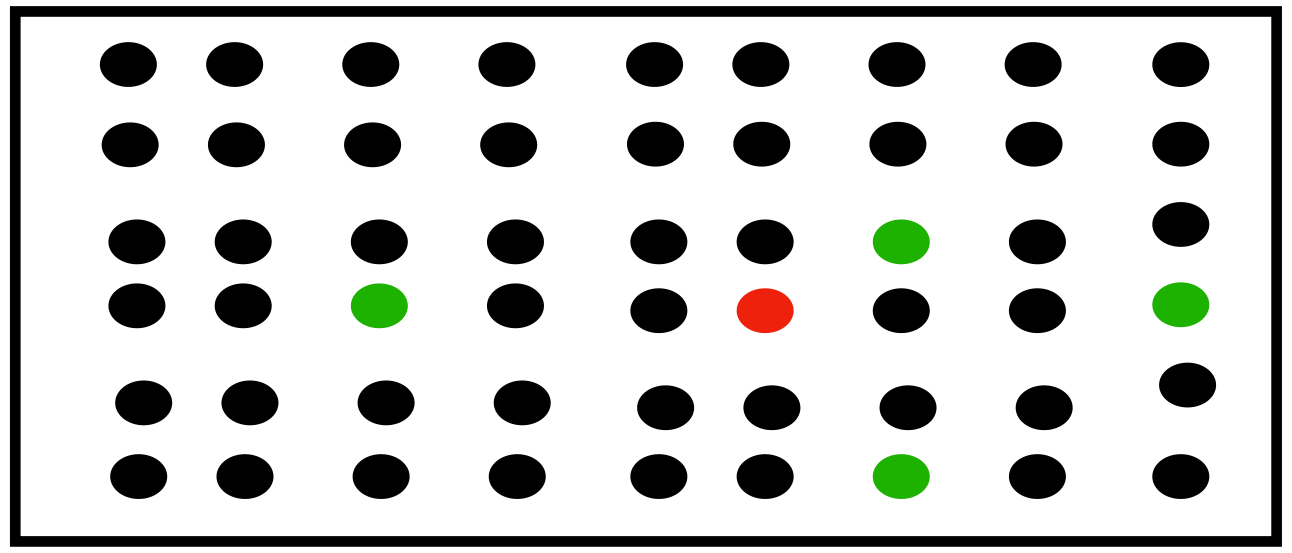
$$P(E|+) = \frac{P(+|E)\overbrace{P(E)}^{\text{Prior}}}{\underbrace{P(+|E)}_{\substack{\text{Sensibilidad}\\\text{(falsos negativos)}}}P(E) + \underbrace{P(+|-E)}_{\substack{\text{1-Especificidad}\\\text{(falsos positivos)}}}P(-E)} \qquad P(E) = \frac{\text{cantidad de casos con E}}{\text{poblacion total}}$$

Test que da siempre negativo tiene **especificidad** de 100%. Pero esto tendria muchos **falsos negativos**.
Test que da siempre positivo tiene **sensibilidad** de 100%. Pero esto tendria muchos **falsos positivos**.

Un buen test tiene que tener ambos alta especificidad y sensibilidad

Supongamos que quiero evaluar una enfermad poco usual en la poblacion con un test de sensibilidad y especificidad del 99%.
P(E) = 0.001, P(+|E) = 0.99, P(+|-E) = 0.01

$$P(E|+) = \frac{0.99 * 0.001}{0.99 * 0.001 + 0.01 * 0.999} = 9\%$$



En una poblacion de 1000 el test me va a dar un positivo verdadero y 10 falsos positivos.

P(E|+) = 9%

Actualizacion del conocimiento segun Bayes

$$\underbrace{P(\text{hipotesis}|\text{data}, I)}_{\text{Conocimiento posterior}} = \frac{\overbrace{P(\text{data}|\text{hipotesis}, I)}^{\text{Verosimilitud}}\overbrace{P(\text{hipotesis}|I)}^{\text{Conocimiento anterior}}}{\underbrace{P(\text{data}|I)}_{\text{Evidencia}}} \propto P(\text{data}|\text{hipotesis}, I)P(\text{hipotesis}|I)$$

Dado un conjunto de datos {Dk}, puedo usar todos los datos a la vez y aplicar el teorema de Bayes en el conjunto total:

$$P(H|\{D_k\}) \propto P(\{D_k\}|H)P(H)$$

Esto es equivalente a aplicar el teorema de un dato por vez de forma secuencial y actualizar mi conocimiento anterior de mi 2da iteracion por el posterior de mi 1era iteracion (asumiendo que los datos son *independientes*):

Metodo secuencial

$$P(H|D_1) \propto P(D_1|H)P(H)$$

$$P(H|D_1, D_2) \propto P(D_2|H)P(H|D_1)$$

Metodo colectivo

$$P(H|D1, D2) \propto P(D1, D2|H)P(H)$$

Actualizacion del conocimiento

$$P(A|B) \propto P(B|A)P(A)$$

Analisis de los datos (inferencia)

Cual metodo usar dependera de la naturaleza de los datos, si los mismos estan todos disponibles al momento del analisis o si los mismos arriban de forma secuencial siguiendo un periodo de tiempo.

Aplicacion de Bayes: test de enfermedad II

$$P(E|+) = \frac{P(+|E)\overbrace{P(E)}^{\text{Prior}}}{\underbrace{P(+|E)}_{\substack{\text{Sensibilidad} \\ \text{(falsos negativos)}}}P(E) + \underbrace{P(+|-E)}_{\substack{\text{1-Especificidad} \\ \text{(falsos positivos)}}}P(-E)}$$

$$P(E) = \frac{\text{cantidad de casos con E}}{\text{poblacion total}}$$

Resultado del 1er test:
P(E) = 0.001, P(+|E) = 0.99, P(+|-E) = 0.01

$$P(E|+) = \frac{0.99 * 0.001}{0.99 * 0.001 + 0.01 * 0.999} = 9\%$$

Supongamos que me hago un 2do test en otro centro y me da tambien positivo, cual es la probabilidad ahora de tener la enfermedad?
P(E |+1) = 0.09, P(+2|E) = 0.99, P(+2|-E) = 0.01, P(-E | +1) = 0.91

$$P(E|+_1, +_2) = \frac{P(+_2|E)P(E|+_1)}{P(+_2|E)P(E|+_1) + P(+_2|-E)P(-E|+_1)} = \frac{0.99 * 0.09}{0.99 * 0.09 + 0.01 * 0.91} = 91\%$$

Asignacion de probabilidad y el principio de indiferencia

Principio de razon insuficiente (Bernoulli 1713) o de indiferencia (Keynes 1921):
Si podemos enumerar un conjunto de posibilidades básicas, mutuamente excluyentes, y no tenemos ninguna razón para creer que alguna de ellas sea más probable que otra, entonces debemos asignar la misma probabilidad a todas.

Ejemplo:
Un dado: tiene sentido en principio asignarle la misma probabilidad a cada numero:
$P(X=i) = 1/6$

- **Hay una forma mas fundamental de justificar esto?, que se quiere decir con “*no tener ninguna razón para creer que alguna de ellas sea más probable que otra*”?**

- **Supongamos una moneda en el cual nos dicen que tiene dos caras: H y T.
Informacion I: tengo una moneda de 2 caras H y T.**

Problema 1: cual es la probabilidad de que salga cara $P(H \mid I)$?
Problema 2: cual es la probabilidad de que salga cruz $P(T \mid I)$?

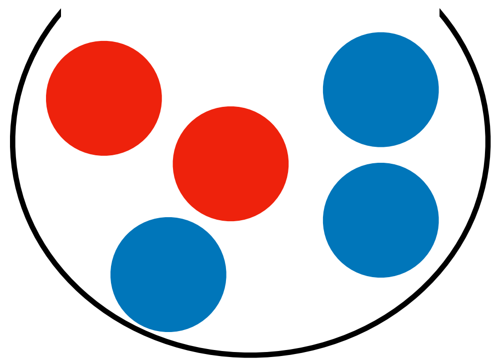
- **Si ahora hacemos el siguiente intercambio H <—> T, el párrafo anterior podría reescribirse intercambiando "Cara" y "Cruz", y "H" y "T", y el problema quedaria igual. Consistencia implica pues que $P(H \mid I) = P(T \mid I)$. A su vez, $P(H \mid I) + P(T \mid I) = 1$. Lo que implica que $P(T \mid I) = P(H \mid I) = 0.5$. En caso contrario de asignarle probabilidades distintas a H y T, estaria asignando probabilidades distintas dando la misma informacion I. Coherencia implica pues equiprobabilidad dada informacion limitada solo de la existencia de dos caras.**

- **Simetria en cuestion: permutacion de las posibilidades $X1 \longleftrightarrow X2 \longleftrightarrow \ldots \longleftrightarrow Xn$**

- **Lo que implica $P(Xi)$ = constante**

Este tipo de simetrias (transformaciones de grupo) que implican la conservacion de ciertas cantidades es la base de la fisica fundamental: El Modelo Estandar, conocido como el teorema de Noether. Aunque este ultimo se aplica a sistemas fisicos y no logicos.

Principio de indiferencia + reglas de probabilidad



- Supongamos que tengo una urna con R bolas rojas y A azules —> informacion existente del problema
- Principio de indiferencia: probabilidad de que cualquier bola sea elegida P(j | I) = constante = 1 / (R + A)
- Aplicando la regla de marginalizacion:

$$P(A) = \sum_i^n P(A \cap B_i) = \sum_i^n P(A|B_i)P(B_i)$$

$$P(R|I) = \sum_{j=1}^{R+A} P(R,j|I) = \sum_{j=1}^{R+A} P(R|j,I)P(j|I) = \frac{1}{R+A}\sum_{j=1}^{R+A} P(R|j,I)$$

- El termino P (R | j, I) = probabilidad de sacar bola roja dado que saque una bola = 1 si la bola j es roja y 0 si es azul.
- Por lo tanto, dado que sabemos que hay R bolas rojas, obtenemos que:

$$P(R|I) = \frac{1}{R+A}\sum_{j=1}^{R} 1 = \frac{R}{A+R}$$

- Este resultado justifica el resultado de asignar la probabilidad a un evento A como:

$$P(A) = \frac{\text{numero de eventos favorables}}{\text{numero de posibilidades totales}}$$

Combinatoria y probabilidad

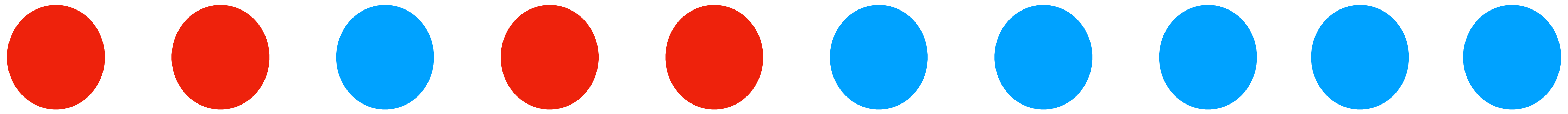
$$P(A) = \frac{\text{numero de eventos favorables}}{\text{numero de posibilidades totales}}$$

- En casos mas complicados, el numerador de este termino puede resultar dificil de calcular, y se requiere el uso de la **combinatoria** para calcular todos los “numeros de eventos favorables”.
- Supongamos por ejemplo que en vez de sacar 1 bola sacamos 10 (asumiendo que cada vez que saco una agrego otra igual a la urna para mantener siempre la misma proporcion) y me pregunto cual es la probabilidad de sacar r rojas en N=10 sacadas?

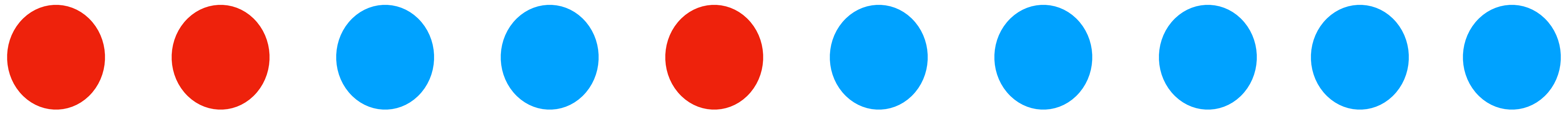
$$P(r|N,I) = \frac{\text{cantidad de combinaciones que dan r rojas en N sacadas}}{2^N}$$

Ejemplo: r = 4 y N = 10

Caso favorable   (Secuencia S1)



Caso no favorable   (Secuencia S2)



- Para calcular la probabilidad de sacar r rojas de N sacadas, aplico la regla de marginalizacion en todas las posibles secuencias:

$$P(r|N,I) = \sum_{k=1}^{2^N} P(r, S_k|N,I) = \sum_{k=1}^{2^N} \underbrace{P(r|S_k,N,I)}_{\substack{\text{= 1 si Sk tiene} \\ \text{r rojas, de lo} \\ \text{contrario = 0}}} P(S_k|N,I)$$

$\longrightarrow$ Solo debo calcular P(Sk | N, I) y sumar para las secuencias en que hay exactamente r rojas

Combinatoria y probabilidad

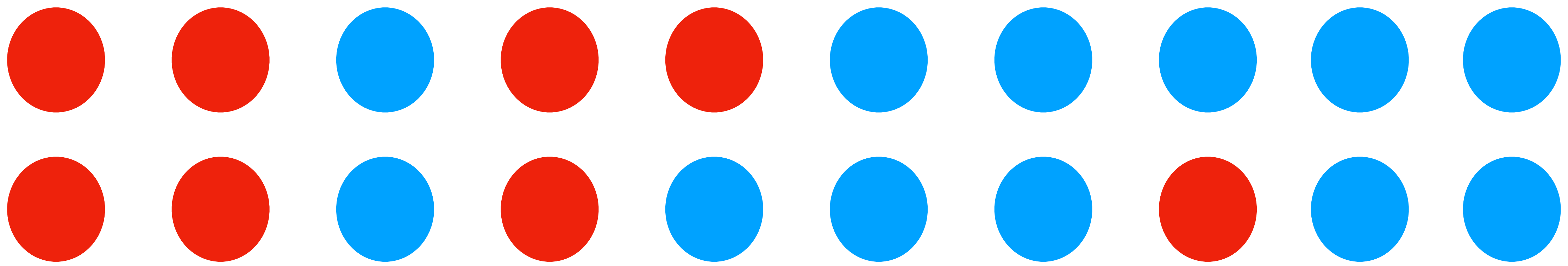
$$P(r|N,I) = \sum_{k=1}^{2^N} P(r, S_k|N,I) = \sum_{k=1}^{2^N} \underbrace{P(r|S_k,N,I)}_{} P(S_k|N,I)$$

= 1 si Sk tiene r rojas, de lo contrario = 0 $\longrightarrow$ Solo debo calcular P(Sk | N, I) y sumar para las secuencias en que hay exactamente r rojas

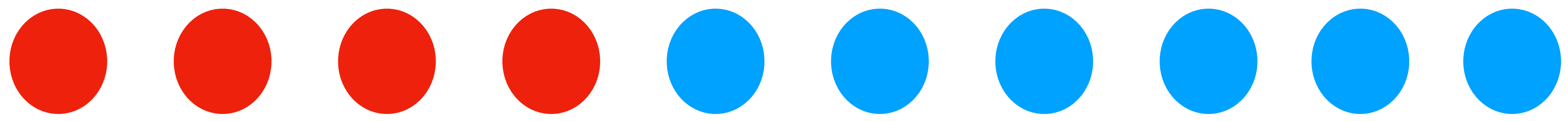
$P(S_{r,k}|N,I) = ?$ Probabilidad de sacar una secuencia de r bolas rojas y N-r azules.

Supongamos que r = 4 y N = 10, a = 6

- Por el principio de indiferencia, asignamos la misma probabilidad a todas las secuencias que tengan r rojas y a azules.



- Nos centramos por ende en una secuencia en particular con r rojas: cual es la probabilidad de sacar esta secuencia?



- Los eventos son independientes, es decir que sacar una bola no influye en lo que pueda sacar despues.

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

$$P(1_r \text{ y } 2_r \text{ y } \ldots 4_r \text{ y } 1_a \ldots \text{y } 6_a) = P(1_r)P(2_r)\ldots P(4_r)P(1_a)\ldots P(6_a) = P(R)^4 P(A)^{10-4}$$

$$\boxed{P(S_{r,k}|N,I) = P(R)^r P(A)^{N-r}}$$

Combinatoria y probabilidad

$$P(r|N,I) = \sum_{k=1}^{2^N} P(r,S_k|N,I) = \sum_{k=1}^{2^N} \underbrace{P(r|S_k,N,I)}_{} \, P(S_k|N,I)$$

= 1 si Sk tiene r rojas, de lo contrario = 0 → Solo debo calcular P(Sk | N, I) y sumar para las secuencias en que hay exactamente r rojas

$$P(S_{r,k}|N,I) = P(R)^r P(A)^{N-r}$$

- Vimos que la probabilidad de sacar una bola roja es: $P(R) = \dfrac{R}{A+R}$
- Y de sacar una azul es: $P(A) = \dfrac{A}{A+R}$
- Por lo tanto, la prob de sacar **una secuencia particular** de r rojas y N-r azules es: $\boxed{P(S_{r,k}|N,I) = \dfrac{R^r A^{N-r}}{(R+A)^N}}$

- La prob entonces de sacar **alguna secuencia** de r bolas rojas $P(r|N,I)$ se reduce a este ultimo termino $P(S_{r,k}|N,I)$ multiplicado por el número de posibles secuencias de N sacadas que contienen exactamente r rojas.

Combinatoria y permutaciones

- Deseo calcular el número de posibles secuencias de N sacadas que contienen exactamente r rojas.

1. En cuantas formas posibles puedo ordenar n elementos en una linea recta?

1	2	3	4	5	(n=5)
2	3	4	5		
3	4	5			
4	5				
5					

n x (n-1) x (n-2) x … x 3 x 2 x 1 = n! (factorial de n)

2. En cuantas formas posibles puedo tomar m elementos de n distintos?. Podemos ver esto como un paso intermedio en el calculo de arriba y frenar el producto cuando hayamos agarrado m elementos:

$$n \times (n-1) \times (n-2) \times \cdots \times (n-m+2) \times (n-m+1) = {}^{n}P_{m}$$ (**permutador**)

$${}^{n}P_{m} = \frac{n!}{(n-m)!}$$

- Ejemplo: numero de formas posibles de tomar 3 elementos de 5?

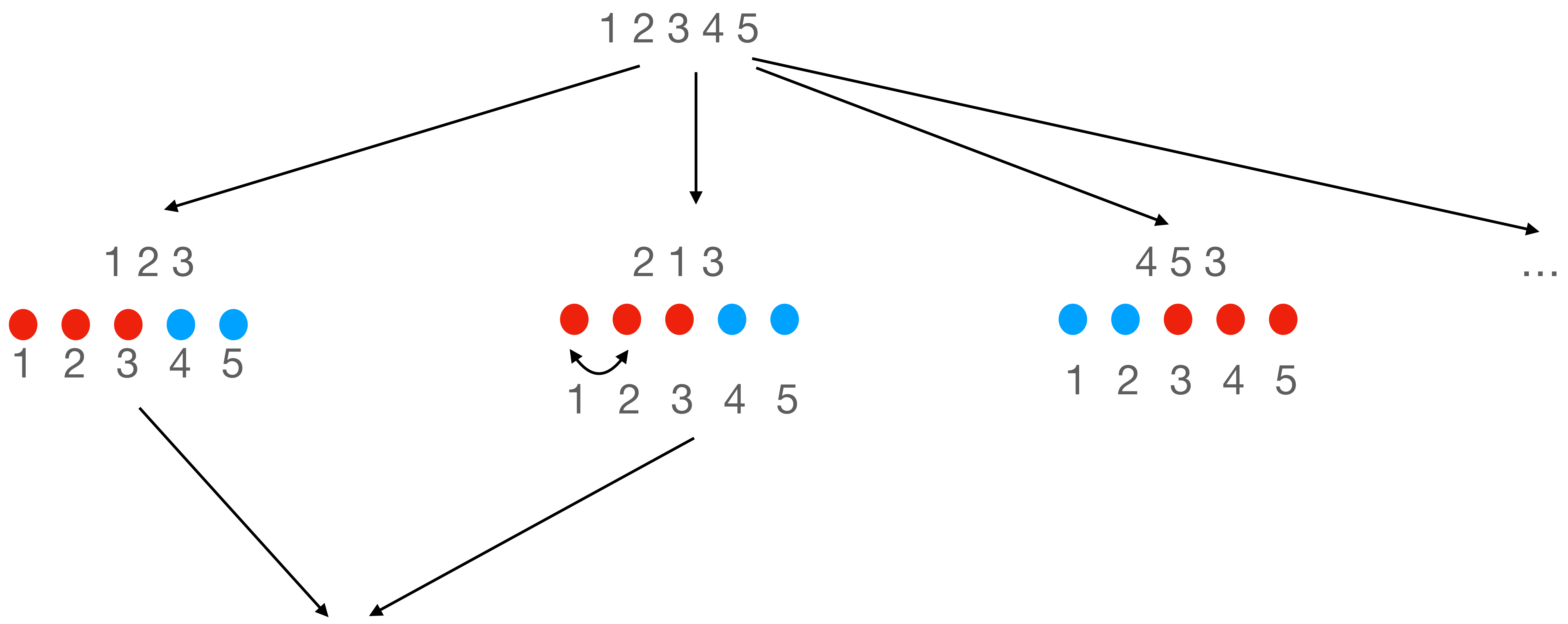
$${}^{5}P_{3} = \frac{5 \times 4 \times 3 \times \cancel{2 \times 1}}{\cancel{2 \times 1}} = 5 \times 4 \times 3 = 60$$

1 2 3 4 5

1 2 3 2 1 3 4 5 3

Combinatoria y permutaciones

- Volviendo a las bolas, en nuestro caso queremos saber cuantas formas posibles puedo seleccionar r bolas de N, pero sin importar el orden en el que estas son elegidas!



El resultado de esta permutacion da lugar a la misma secuencia de bolas rojas. Esto significa que no me importa el orden.

Divido el permutador por el numero de veces m elementos pueden permutarse. Que como vimos este es el factorial de m (m!)

3. El numero de formas posibles de elegir r bolas de N sin importar el orden es por ende el **combinatorio**:

$$^{n}C_{m} = \frac{^{n}P_{m}}{m!} = \frac{n!}{m!(n-m)!}$$

Distribucion binomial

- La prob entonces de sacar **alguna secuencia** de r bolas rojas de N es:

$$P(r|N,I) = {}^N C_r \frac{R^r A^{N-r}}{(R+A)^N} = \frac{N!}{r!(N-r)!} \frac{R^r A^{N-r}}{(R+A)^N}$$

- El valor esperado de r/N es:

$$E(\frac{r}{N}) = \sum_{r=0}^{N} \frac{r}{N} P(r|N,I) = P(R) = \frac{R}{A+R}$$

- Y la varianza de r/N:

$$Var\left(\frac{r}{N}\right) = E\left((\frac{r}{N} - P(R))^2\right) = \frac{P(R)P(A)}{N}$$

- La varianza se hace cero cuando N es muy grande, y la variable aleatoria r/N en el limite de N grande da P(R) = probabilidad de sacar una roja.

$$P(R) = \underbrace{\lim_{N\to\infty} \frac{\text{numero de rojas}}{N}}_{\text{Interpretacion frecuentista: asignacion de probabilidad basado en un numero infinito de experimentos}} = \underbrace{\frac{\text{numero de casos favorables}}{\text{numero de casos totales}}}_{\text{asignacion de probabilidad basado en el principio de indiferencia}} = \frac{R}{R+A}$$

Interpretacion frecuentista: asignacion de probabilidad basado en un numero infinito de experimentos

asignacion de probabilidad basado en el principio de indiferencia

Jacob Bernoulli

- Qué se podría decir sobre la probabilidad de obtener una bola roja, en un solo sorteo, dado un número **finito** de resultados observados? sin conocer la proporcion R y A?

—> Bayes y Laplace

Distribucion binomial

- La prob entonces de sacar **alguna secuencia** de r bolas rojas de N es:

$$P(r|N,I) = {}^{N}C_{r}\frac{R^{r}A^{N-r}}{(R+A)^{N}} = \frac{N!}{r!(N-r)!}\frac{R^{r}A^{N-r}}{(R+A)^{N}} = \frac{N!}{r!(N-r)!}p^{r}(1-p)^{N-r}$$

$$p = \text{probabilidad de sacar una roja} = \frac{R}{R+A}$$

- Sacar una bola es una variable aleatoria que puede tener dos valores: 0 o 1, la bola es roja o no lo es.
  Cuando los eventos de sacar una bola son independientes e identicos (es decir la probabilidad p no cambia de una bola a otra) la probabilidad de sacar r rojas de N sacadas obedece a la **distribucion binomial:**

$$B(k|N,p) = \frac{N!}{k!(N-k)!}p^{k}(1-p)^{N-k}$$

k = numero de exitos
p = probabilidad de un exito
N = cantidad de experimentos

- Para el caso particular de N = 1, tengo la distribucion de Bernoulli:

$$Be(k|p) = p^{k}(1-p)^{1-k}$$

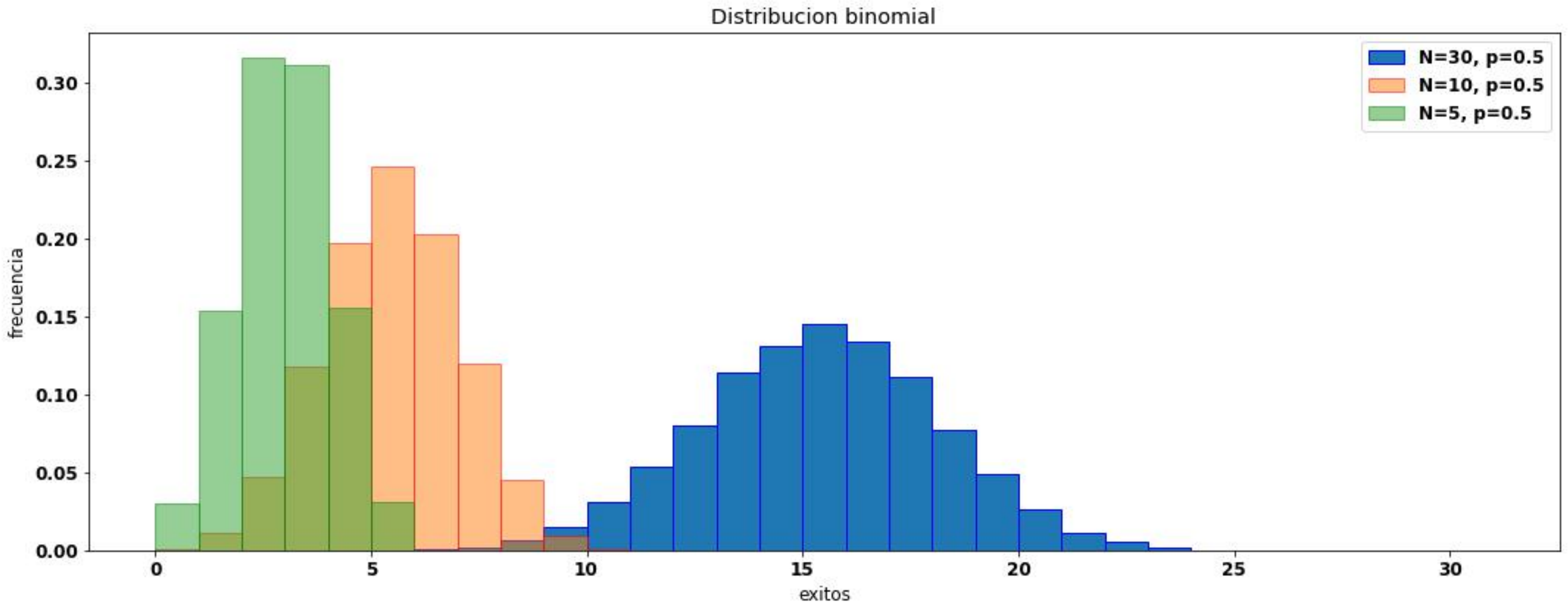
- Por ejemplo tirar una moneda. Asumiendo que p = 0.5:
  La probabilidad que salga cara = Be(cara | 0.5) = p = 0.5
  La probabilidad que salga cruz = Be(cruz | 0.5) = 1 - p = 0.5

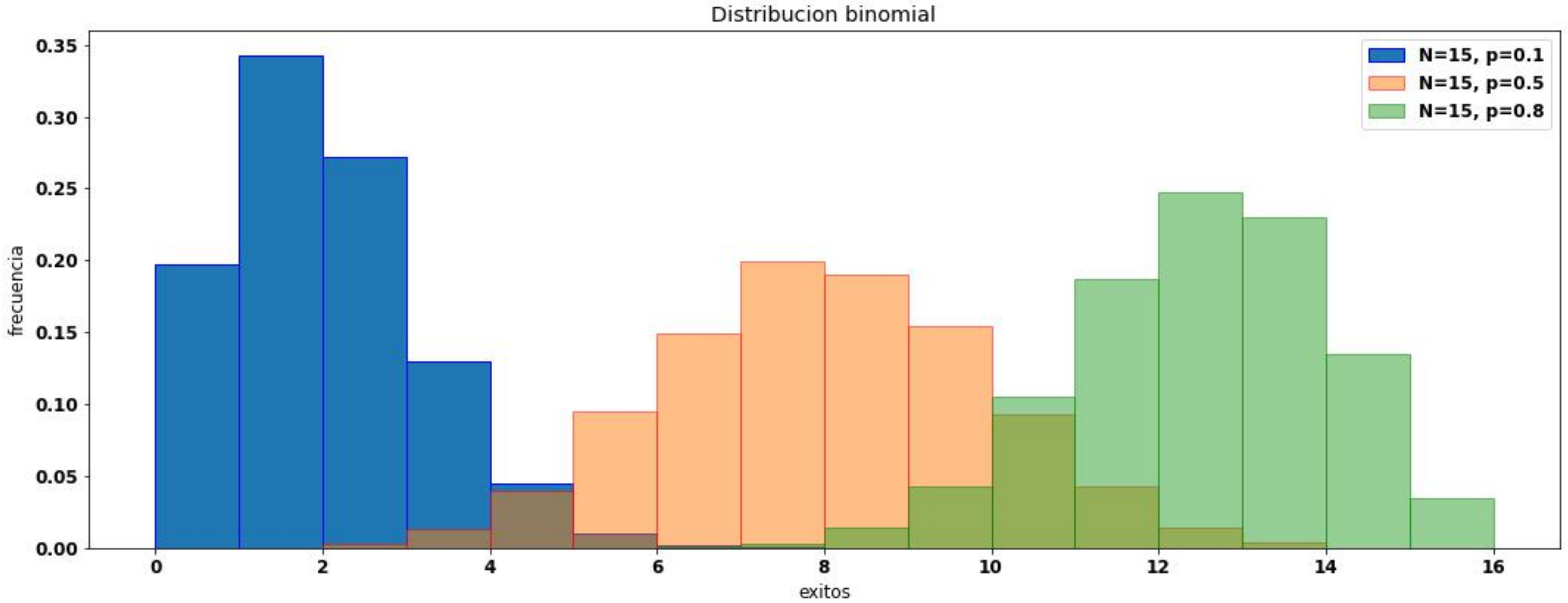
Distribucion binomial

- Una variable aleatoria que puede tener dos valores: 0 o 1.
  Cuando los eventos son independientes e identicos (es decir la probabilidad p de un evento singlular no cambia) la probabilidad de obtener k exitos de N obedece a la **distribucion binomial:**

$$B(k|N,p) = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$$

k = numero de exitos
p = probabilidad de un exito
N = cantidad de experimentos



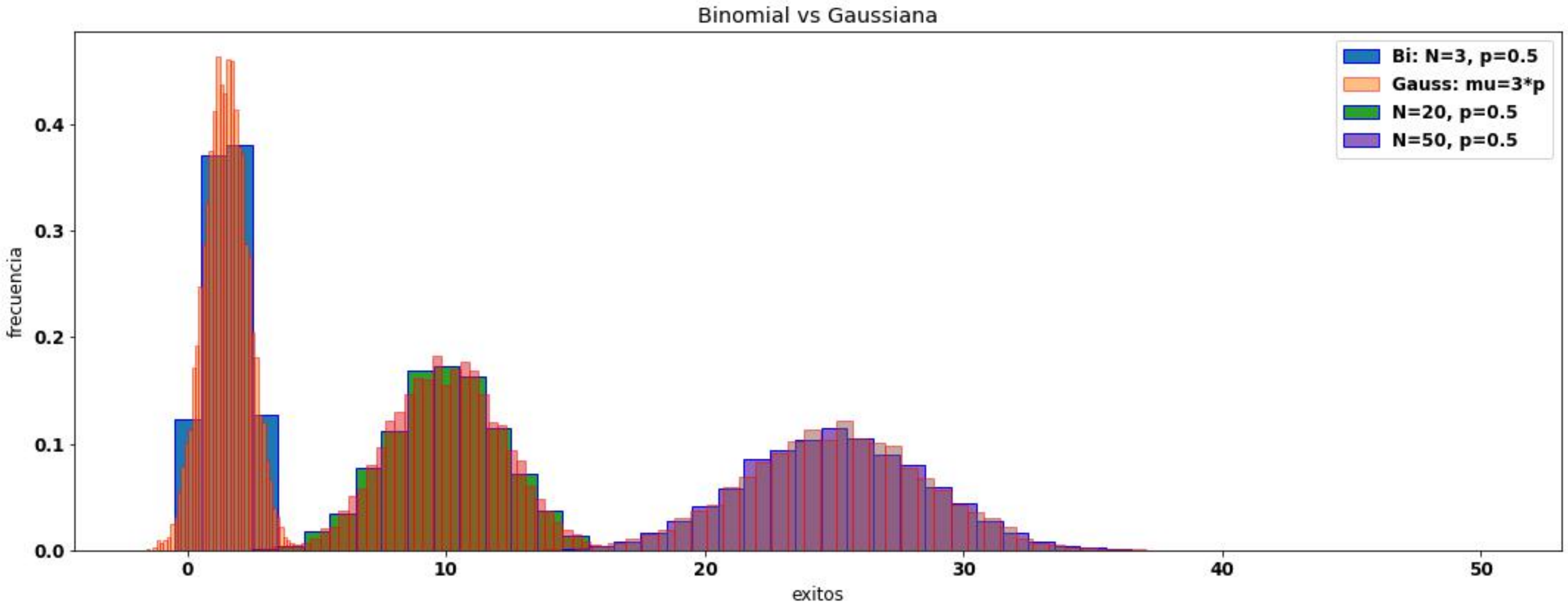


Distribucion binomial vs Gauss

- Si N grande, la asimetria de la binomial no es grande y se aproxima a la distribución normal. La aproximacion es mejor a medida que aumenta N (al menos 20) y es mejor cuando p no está cerca de 0 o 1.

$$B(k|N,p) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k} \qquad B(k|N,p) \sim \mathrm{Gauss}(x|\mu = Np, \sigma = \sqrt{Np(1-p)})$$

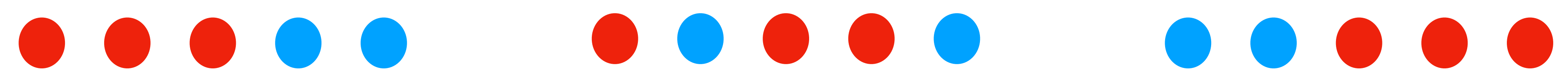
- La binomial es la suma de N variables aleatorias independientes que obedecen a la distribucion de Bernoulli cada una con probabilidad p. **El teorema central del limite** (veremos mas adelante) nos dice que esta suma aproxima a Gauss.



Un poquito mas de combinatoria

3. El numero de formas posibles de elegir r bolas de N sin importar el orden es por ende el **combinatorio**:

$$ {}^{n}C_{m} = \frac{{}^{n}P_{m}}{m!} = \frac{n!}{m!(n-m)!} $$



- **El problema del cumpleaños:**
  Cual es la probabilidad de que en un grupo de n personas hayan al menos 2 que cumplan el mismo dia?

Un poquito mas de combinatoria

3. El numero de formas posibles de elegir r bolas de N sin importar el orden es por ende el **combinatorio**:

$$^{n}C_{m} = \frac{^{n}P_{m}}{m!} = \frac{n!}{m!(n-m)!}$$

- **El problema del cumpleaños:**
  Cual es la probabilidad de que en un grupo de n personas hayan al menos 2 que cumplan el mismo dia?
  para un grupo de 23 personas la probabilidad es ~51%!!

  Intuitivamente se razona asi:
  Para buscar un match en un grupo de 23 personas tengo que considerar todas las combinaciones posibles de pares.
  Esto es justamente el combinatorio:

$$^{23}C_{2} = \frac{23!}{2!(23-2)!} = \frac{23 \times 22}{2} = 253$$

  253 es mayor a la mitad de dias que tiene un año, y por lo tanto es muy probable encontrar un match dadas las cantidades de combinaciones.

  Calculo de la probabilidad:
  Probabilidad A de que en un grupo de 23 personas no haya algun dia de cumpleaños repetido:

$$P(A) = \frac{\text{numero de casos posibles de A}}{\text{numero total de posibilidades}}$$

El problema del cumpleaños

Probabilidad A de que en un grupo de 23 personas no haya algun dia de cumpleaños repetido:

$$P(A) = \frac{\text{numero de casos posibles de A}}{\text{numero total de posibilidades}}$$

$$\text{numero total de posibilidades} = 365^{23}$$

$$\text{numero de casos posibles de tener 23 fechas distintas} = ?$$

Ejemplo, para un grupo de 2 personas:
tres posibles casos de fechas distintas son: {{01/02, 02/02}, {02/02, 01/02}, {07/02, 04/11}}
tres posibles casos totales son: {{01/02, 01/02}, {08/12, 01/02}, {09/01, 07/11}}

La pregunta es equivalente a: cuantas formas posibles puedo seleccionar 23 elementos distintos de 365

- Ejemplo: numero de formas posibles de tomar 3 elementos de 5?

1 2 3 4 5

1 2 3   2 1 3   4 5 3

Esto es el permutador que vimos:

$$\text{numero de casos posibles de tener 23 fechas distintas} = {}^{365}P_{23}$$

$$P(B) = 1 - P(A) = 1 - \frac{{}^{365}P_{23}}{365^{23}} \sim 51\%$$

El problema del cumpleaños

- Cual es la probabilidad de que en un grupo de n personas hayan al menos 2 que cumplan el mismo dia?
  para un grupo de 23 personas la probabilidad es ~51%!!

- Cual es la probabilidad de que en un grupo de 23 personas hayan al menos 2 que cumplan en un dia en particular del año?

$$B(k >= 2|23, 1/365) = \sum_{k=2}^{23} {}^{23}C_k \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{23-k} = 0.182\%$$

- Cual es la probabilidad de que en un grupo de 23 personas haya al menos 1 que cumplan en un dia en particular del año?

$$B(k >= 1|23, 1/365) = 1 - \left(\frac{364}{365}\right)^{23} = 6.114\%$$

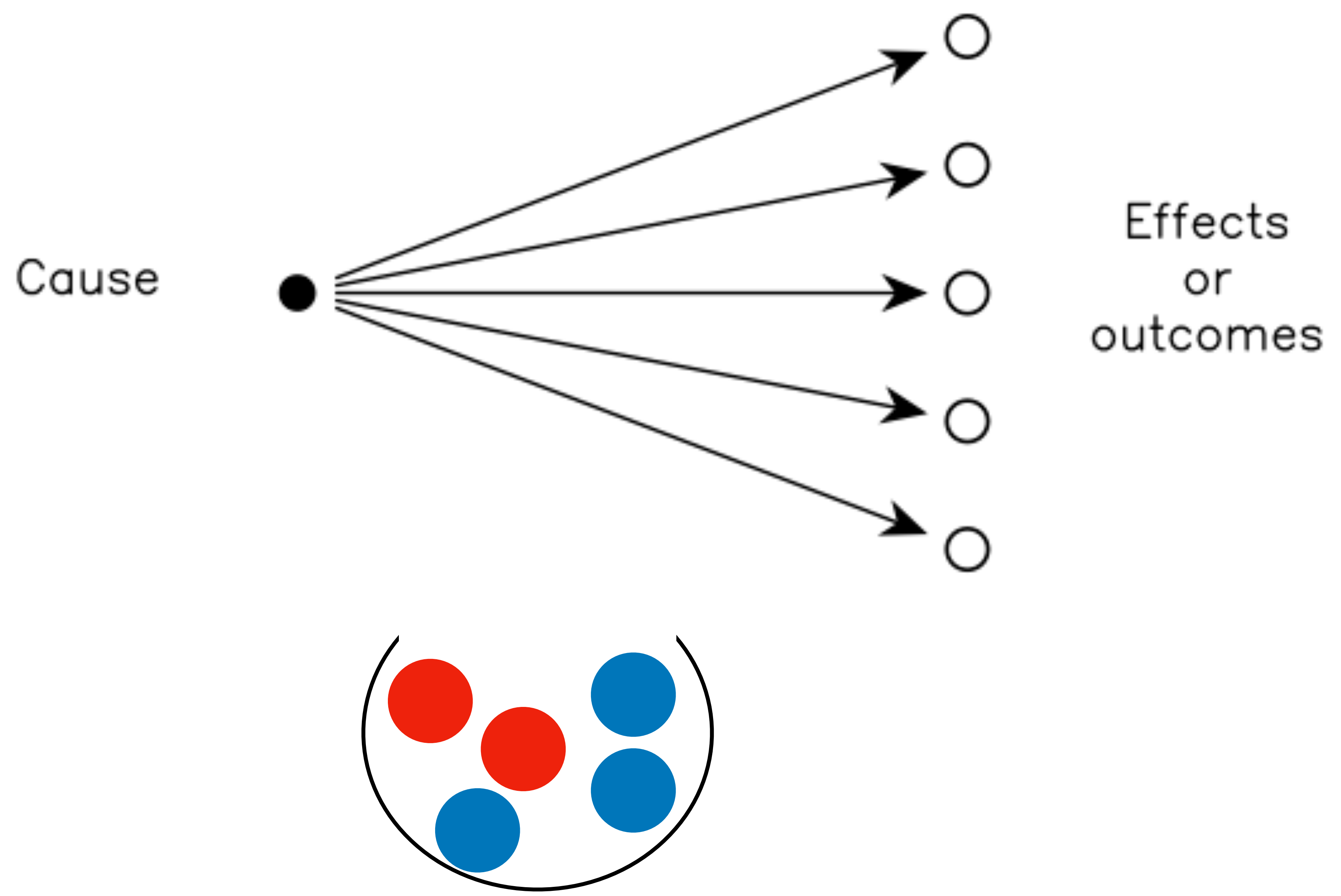
- El numero minimo de personas de un grupo para que haya al menos una que cumpla en un dia particular del año, e.g 5 de Noviembre, y que la probabilidad sea >= 50% es de N = 253

$$B(k >= 1|253, 1/365) = 1 - \left(\frac{364}{365}\right)^{253} = 50\%$$

- Pueden hacer el experimento y chequear con sus amigos facebook!

Razonamiento deductivo vs inductivo

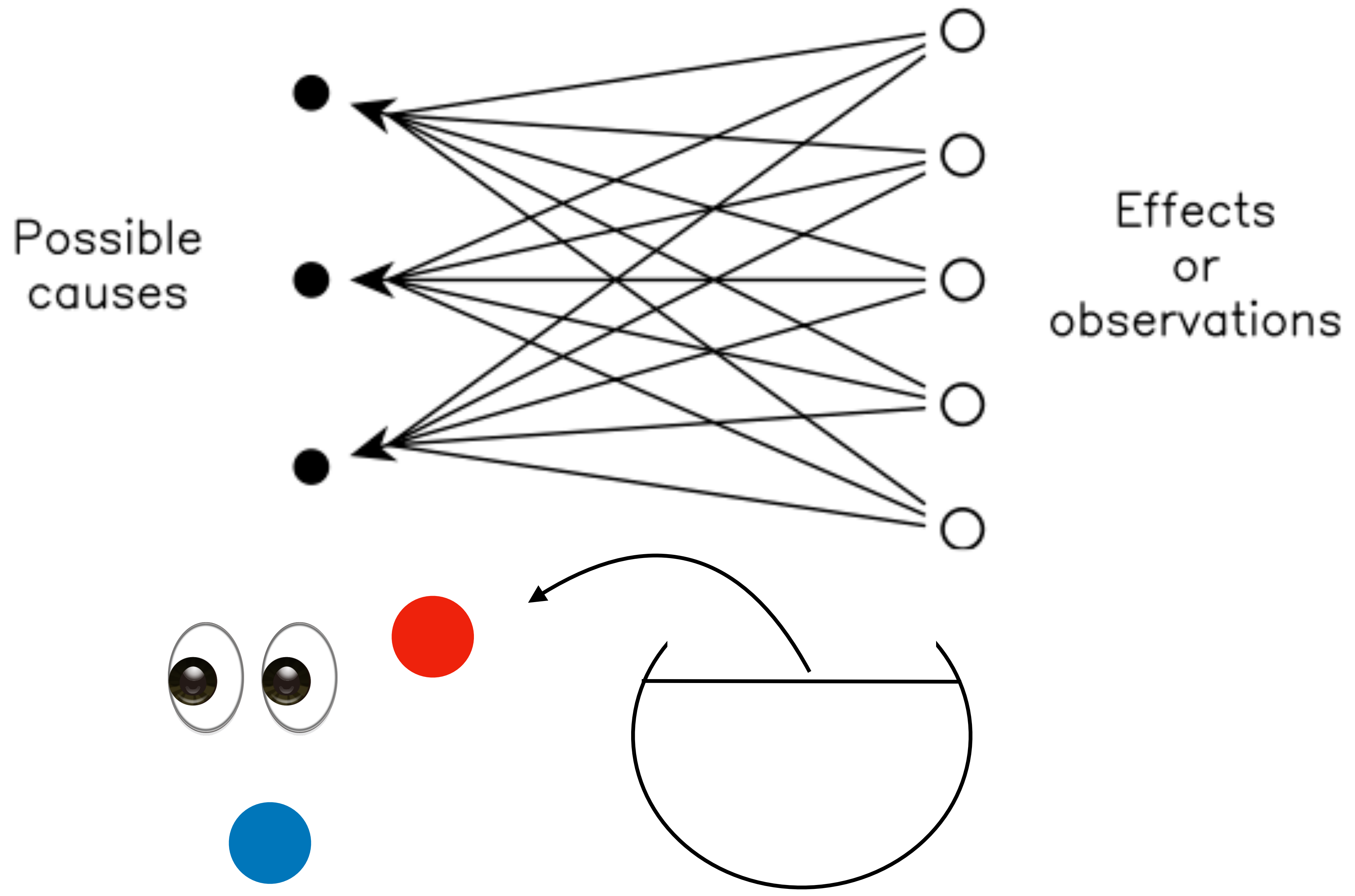
Deductivo



Dada una urna con 20% rojas y 80% azules, que probabilidad hay que sacando 10 de obtener 8 rojas?

—> Distribucion binomial / Bernoulli

Inductivo o razonamiento plausible



Dado que observo 10 rojas de 23 sacadas, que puedo decir de la proporcion de rojas/azules en la urna?

—> Bayes y Laplace

Determinacion del sesgo de una moneda

- Qué se podría decir sobre la probabilidad de obtener una bola roja, en un solo sorteo, dado un número **finito** de resultados observados? sin conocer la proporcion R y A?
- Dado que observo 10 rojas de 23 sacadas, que puedo decir de la proporcion de rojas/azules en la urna?
- Como podria determinar el sesgo de una moneda?. Es la moneda (o el lanzador) justa?

- Queremos determinar la imparcialidad de una moneda y el grado de certeza de ello. Una moneda es parcial o justa cuando la probabilidad de que salga cara o cruz es p = 50%. El grado de imparcialidad o sesgo puede ser: 0 <= p <=1.
- Determinar la certeza de que p este en un rango determinado, basado en los datos:

$$\mathrm{prob}(p|datos, I) = ?$$

datos = tiro N veces la moneda y observo cuantas caras salieron (exitos k)

- Aplicando la regla de Bayes para la determinacion del “parametro” “p” (sesgo):

$$\underbrace{\mathrm{prob}(p|datos, I)}_{\text{Conocimiento posterior}} = \frac{\mathrm{prob}(datos|p, I)\mathrm{prob}(p|I)}{\mathrm{prob}(datos)} \propto \underbrace{\mathrm{prob}(datos|p, I)}_{\text{Verosimilitud}}\underbrace{\mathrm{prob}(p|I)}_{\text{Conocimiento anterior}}$$

- La verosimilitud es una medida de la prob de haber obtenido los datos que realmente observamos, si el valor del sesgo estuviera dado
- La verosimilitud es un **razonamiento deductivo:** asumiendo que conozco el sesgo “p”, que puedo decir de los datos?
- El conocimiento posterior es un **razonamiento inductivo:** basado en los datos, que puedo de decir del sesgo “p”?, lo desconocido?

Verosimilitud —> distirbucion Binomial: $\mathrm{prob}(datos|p, I) \propto p^k(1-p)^{N-k}$

Determinacion del sesgo de una moneda

datos = tiro N veces la moneda
y observo cuantas caras
salieron (exitos k)

$$\underbrace{\mathrm{prob}(p|datos, I)}_{\text{Conocimiento posterior}} = \frac{\mathrm{prob}(datos|p, I)\mathrm{prob}(p|I)}{\mathrm{prob}(datos)} \propto \underbrace{\mathrm{prob}(datos|p, I)}_{\text{Verosimilitud}}\underbrace{\mathrm{prob}(p|I)}_{\text{Conocimiento anterior}}$$

Verosimilitud —> distirbucion Binomial: $\mathrm{prob}(datos|p, I) \propto p^k(1-p)^{N-k}$

Conocimiento anterior —> asumo a priori una distribucion uniforme
(ignorancia total sobre la naturaleza de la moneda): $\mathrm{prob}(p, I) = U[0, 1]$

Conocimiento posterior: $\mathrm{prob}(p|datos, I) \propto p^k(1-p)^{N-k}U[0, 1] = p^k(1-p)^{N-k}$

Determinacion del sesgo de una moneda

datos = tiro N veces la moneda
y observo cuantas caras
salieron (exitos k)

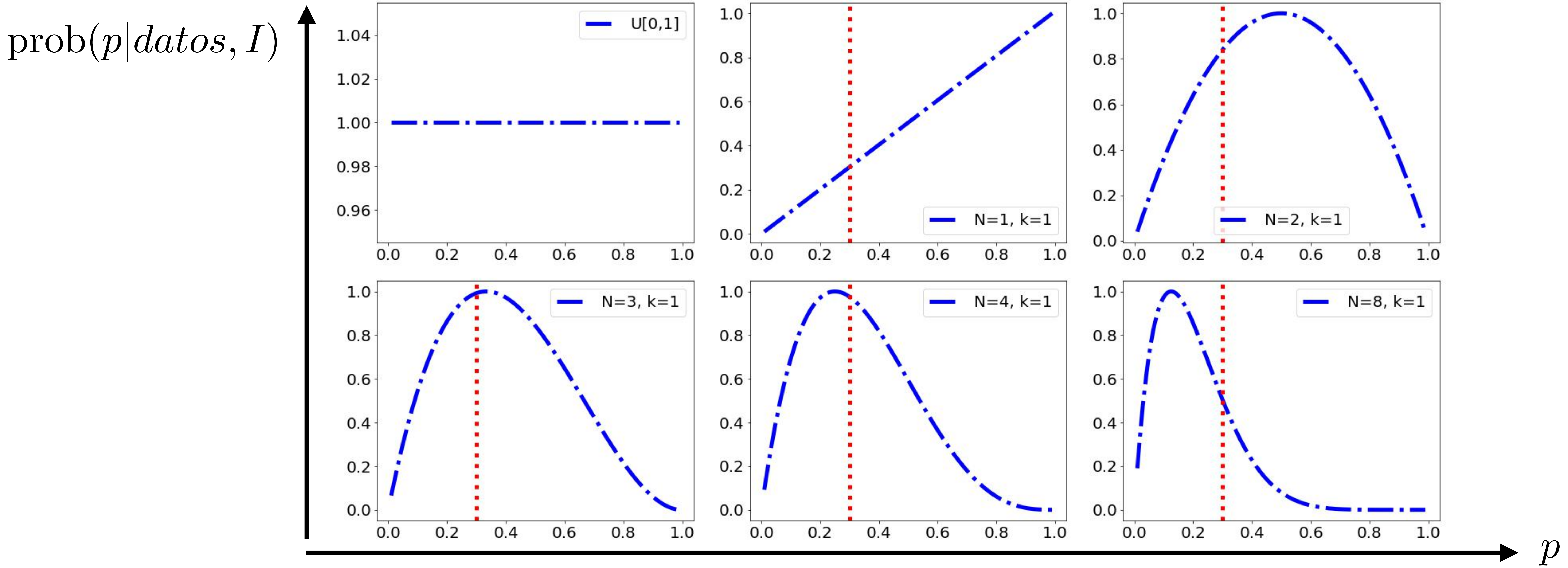
$$\underbrace{\text{prob}(p|datos, I)}_{\text{Conocimiento posterior}} = \frac{\text{prob}(datos|p, I)\text{prob}(p|I)}{\text{prob}(datos)} \propto \underbrace{\text{prob}(datos|p, I)}_{\text{Verosimilitud}}\underbrace{\text{prob}(p|I)}_{\text{Conocimiento anterior}}$$

Verosimilitud —> distirbucion Binomial: $\text{prob}(datos|p, I) \propto p^k(1-p)^{N-k}$

Conocimiento anterior —> asumo a priori una distribucion uniforme
(ignorancia total sobre la naturaleza de la moneda): $\text{prob}(p, I) = U[0, 1]$

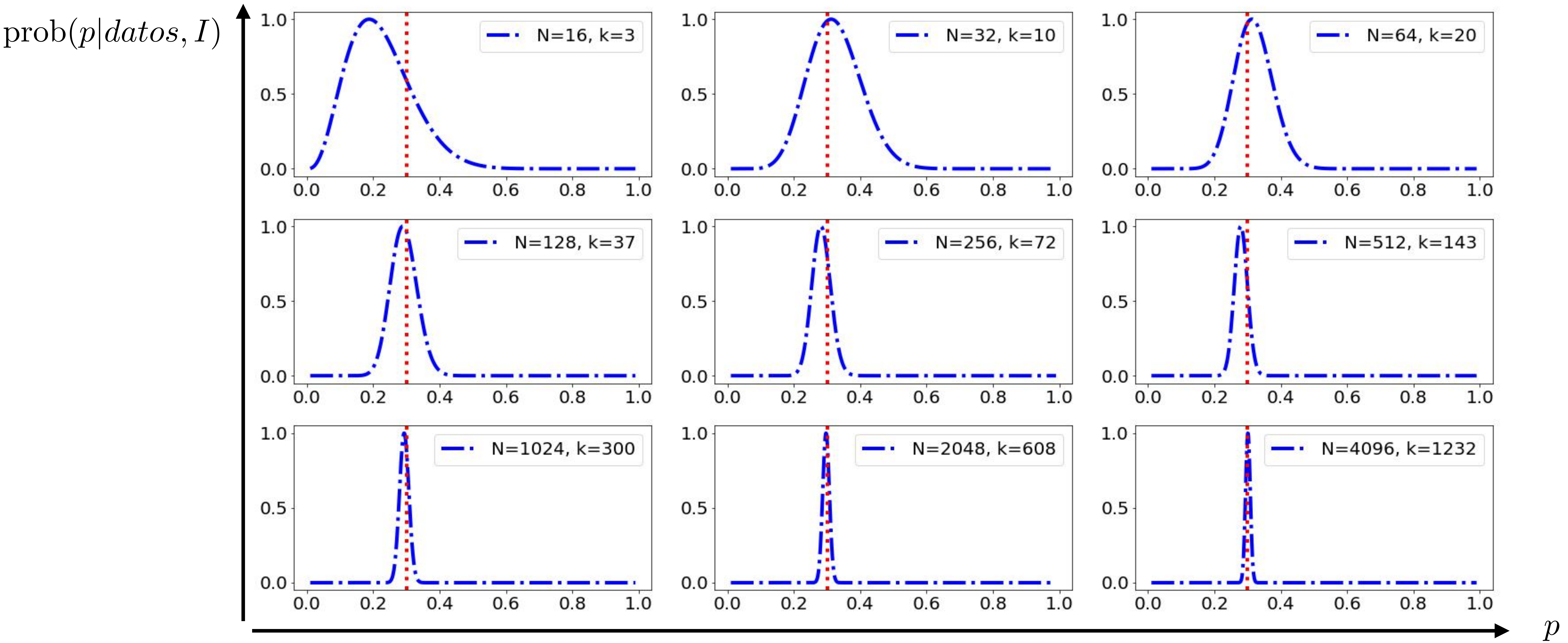
Conocimiento posterior: $\text{prob}(p|datos, I) \propto p^k(1-p)^{N-k}U[0, 1] = p^k(1-p)^{N-k}$

Evolucion del conocimiento posterior a medida que tiro la moneda



Determinacion del sesgo de una moneda

Evolucion del conocimiento posterior a medida que tiro la moneda



- A medida que tiro la moneda el conocimiento posterior evoluciona. El maximo empieza a fluctuar cada vez menos alrededor del valor verdadero (p=0.3) y el ancho de la campana empieza a disminuir (mi grado de confianza es cada vez mayor). Esto fue hecho con una simulacion, donde p = 0.3, una moneda de tres caras: 1 cara y dos cruz!.

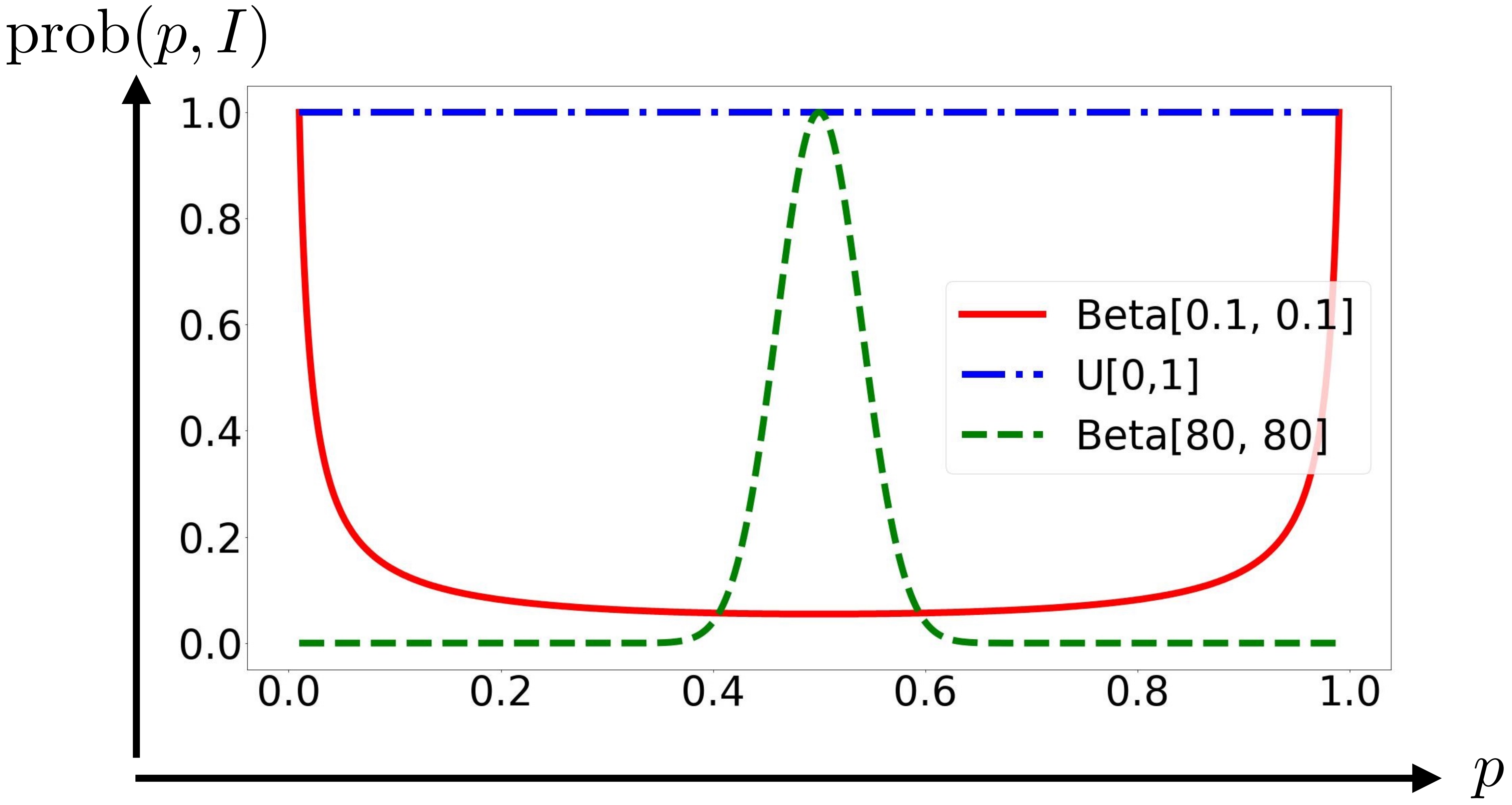
Determinacion del sesgo de una moneda

Que pasa si asumo otros conocimientos anteriores?

Verosimilitud —> distribucion Binomial: $\text{prob}(datos|p, I) \propto p^k (1-p)^{N-k}$

Conocimiento anterior $\text{prob}(p, I)??$

- Distintas personas podran asumir distintas formas para **prob(p)** en base a su experiencia y conocimiento antes de realizar el experimento, esto es la base de la discordia en los seres humanos!!
  - Algunos asumiran ignorancia total: U[0, 1]
  - Otros diran con firmeza que es muy probable que la moneda sea justa.
  - Y otros… que hay pocas chances que sea justa y que seguro esta cargada sea para la cara o cruz.



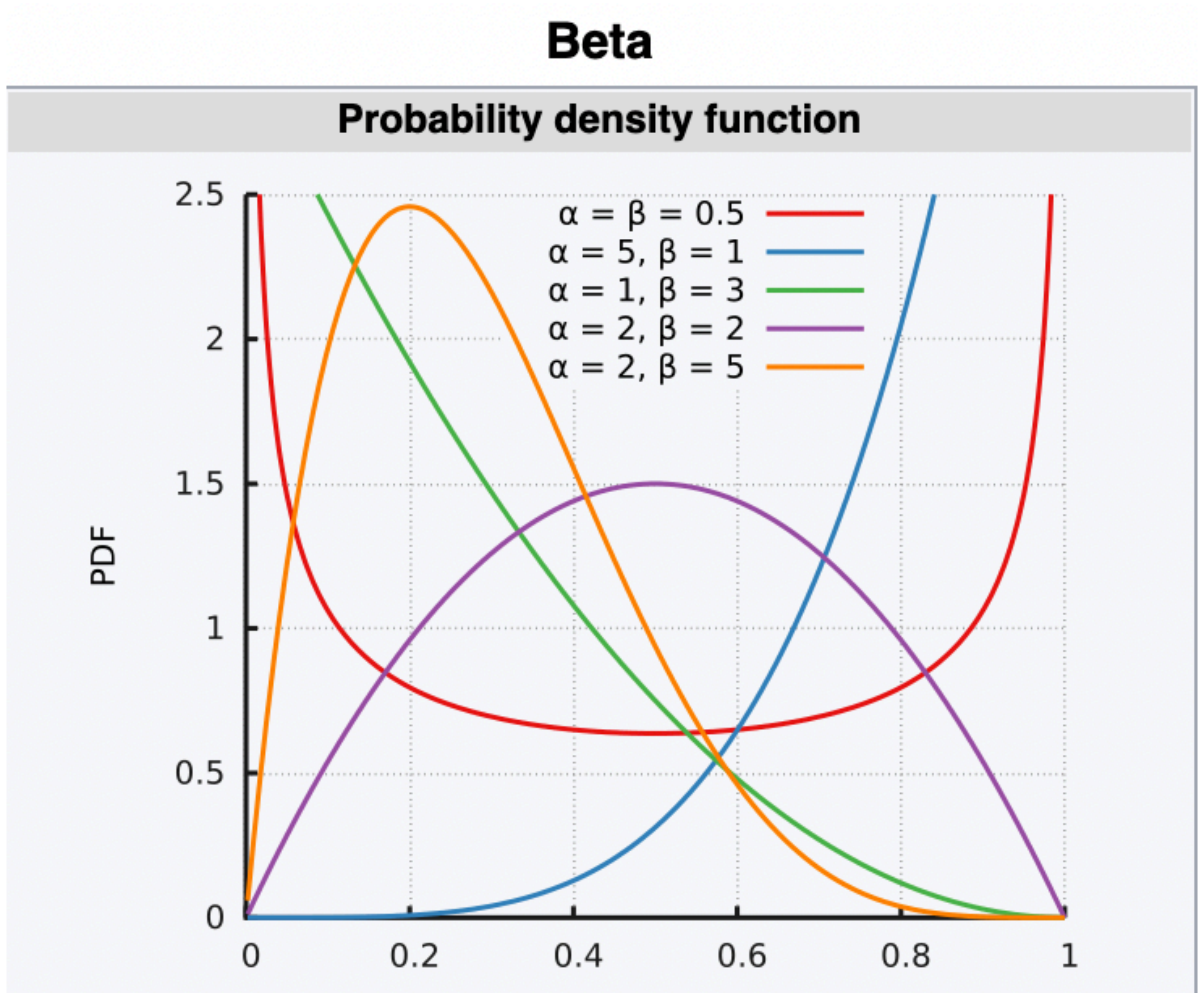
Determinacion del sesgo de una moneda

La distribucion Beta

Verosimilitud —> distribucion Binomial: $\mathrm{prob}(datos|p, I) \propto p^k(1-p)^{N-k}$

Conocimiento anterior $\mathrm{prob}(p, I)??$

- La distribucion Beta es una buena eleccion para modelar las distintas suposiciones a priori.
- Son una familia de distribuciones muy flexibles que dependen de dos parametros: $\mathrm{Beta}(p|\alpha, \beta)$



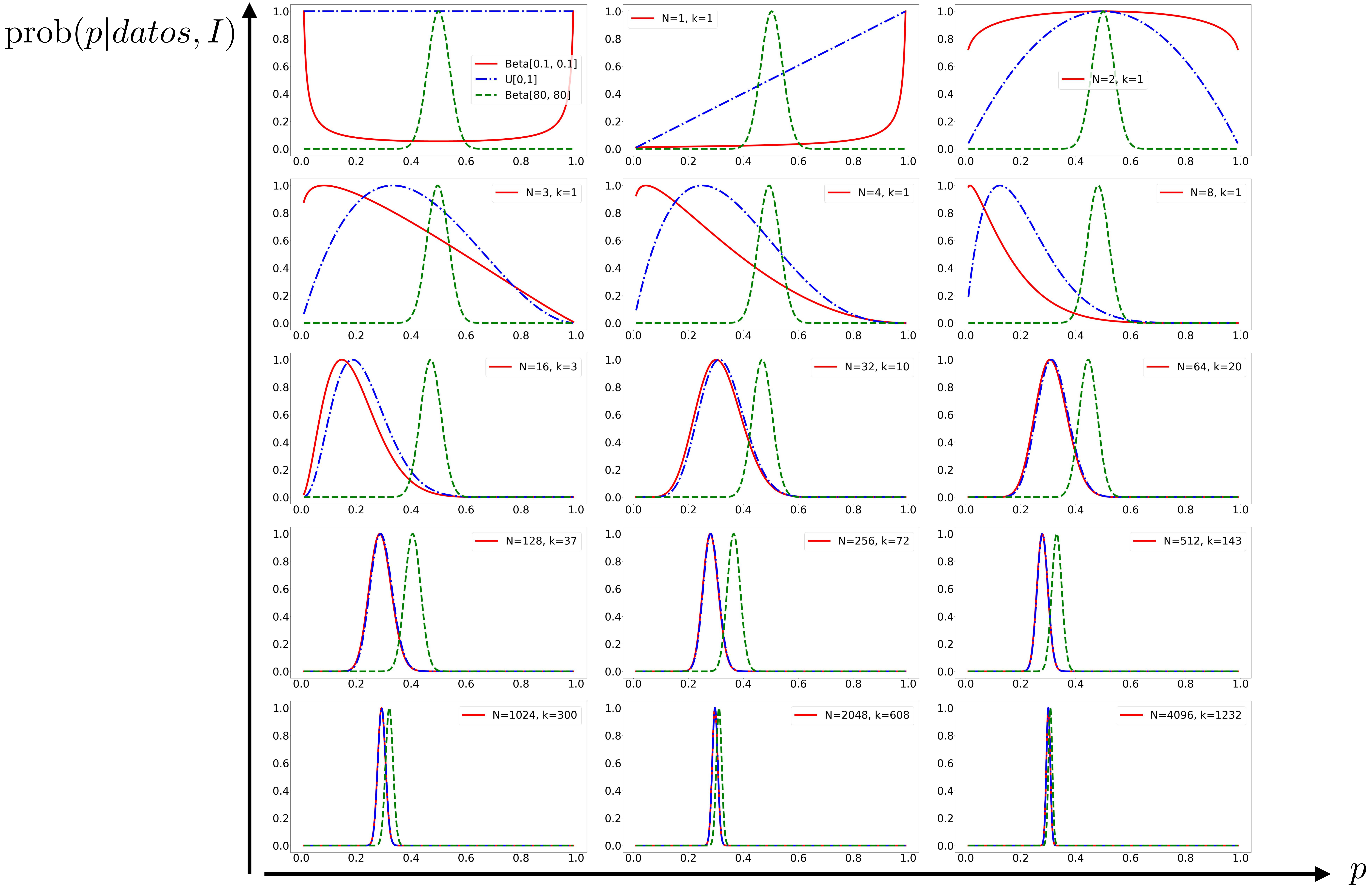
- Resultan ventajosas al combinarlas con la distribucion Binomial, ya que el producto de una binomial con una Beta da una Beta:

$$\mathrm{prob}(p|datos, I) \propto \overbrace{p^k(1-p)^{N-k}}^{\text{Verosimilitud}}\overbrace{\mathrm{Beta}(p|\alpha, \beta)}^{\text{Conocimiento anterior}} \propto \underbrace{\mathrm{Beta}(p|\alpha + k, \beta + N - k)}_{\text{Conocimiento posterior}}$$

- Esto es lo que se conoce como <u>priors conjugados</u>

Determinacion del sesgo de una moneda

Evolucion del conocimiento posterior <u>con distintos priors</u> a medida que tiro la moneda



Determinacion del sesgo de una moneda

Conclusiones

- Cuando hay pocos datos, los **conocimientos posteriores** resultantes muy diferentes y depende de lo que sabiamos o suponíamos antes de los resultados.

- A medida que aumenta la **evidencia empírica**, llegamos a las mismas conclusiones independientemente de nuestras **creencias iniciales**; la posterior está entonces dominada por la **función de verosimilitud**, y la **elección previa** se vuelve irrelevante.

- Necesitamos hacer ~1000 tiradas de una moneda para poder determinar con un grado de certeza aceptable si esta sesgada en este caso entre 0.25 y 0.35.

- El numero de tiradas necesarias para determinar el sesgo dependera de que tan sesgada este la moneda. Cuanto mas sesgada este, mas rapido vamos a poder concluir su sesgo.

- La suposicion anterior de que la moneda sea justa afirma estar bien informado sobre el caracter de la moneda independiente mente de los datos. Por lo tanto, se necesita mucho más para converncerse de que la moneda no es justa.
  Se necesito ~4000 tiradas para arrastrar su creencia (a patadas y gritos) que la moneda no es en efecto justa.

Derivacion de las leyes de la probabilidad v2

Derivacion de Cox: Logica $\Longrightarrow$ Leyes de la probabilidad

- En 1946 Richard Cox se cuestiono acerca de cuantificar las reglas que definen la logica y el razonamiento consistente.
- En particular se centro en aquellas proposiciones en las cuales podemos expresarlas como un grado de certeza en su verdad.
  - “Va a llover mañana”, “esta moneda es justa”, “el emperador de Egipto murio de un flechazo”
- **Sus postulados de partida fueron:**

1. Transitividad: dado el contexto K, si tenemos una mayor creencia en A que en B y mayor en B que en C —> tenemos mayor creencia en A que en C.
Esto implica que la forma de cuantificar la creencia de algo debe ser con numeros reales (es decir ordenados)

$$\left.\begin{array}{l}\pi(A|K) > \pi(B|K) \\ \pi(B|K) > \pi(C|K)\end{array}\right\} \Longrightarrow \pi(A|K) > \pi(C|K)$$

2. Plausibilidad del conjunto: saber acerca de la proposicion A y tambien de B dado el conocimiento de A, sujetos a un contexto K es suficiente para saber todo acerca de AB|K:

$$a = \pi(A|K),\ b = \pi(B|A,K),\ \pi(A,B|K) = F(a,b)$$

3. Consistencia logica (aristotelica): proposiciones logicamente equivalentes deben tener el mismo grado de verdad.

- (2 y 3) —> **Asociatividad logica:** $(P \wedge (Q \wedge S)) \leftrightarrow ((P \wedge Q) \wedge S)$

4. Plausibilidad de la negacion: el grado de creencia acerca de A define el grado de creencia acerca de su negacion Ā

Leyes de la probabilidad: $\begin{cases} Pr(A|B) + Pr(\text{not } A|B) = 1 \\ Pr(A,B|C) = Pr(A|B,C)Pr(B|C) \end{cases}$

Segun Jaynes (2003), esto fue el avance conceptual mas importante en la formulacion de la teoria de probabilidades desde Laplace…

Interpretacion de la probabilidad frecuentista vs bayesiana (continuacion)

- La critica popular de los frecuentistas es que si una probabilidad representa un grado de creencia, entonces debe ser subjetiva porque mi creencia puede ser diferente a la del otro.
  (Uso de distintos conocimientos anteriores determinan distintos grados de confianza).

- Esto es cierto, pero realmente implica subjetividad o es mas una limitacion a como asignar probabilidades en base a la informacion disponible?

- Dos personas que estan condicionados a la misma informacion, asignaran la misma probabilidad a un dado suceso. Tiene esto algo de subjetivo?

$$P(A|I_1) = P(A|I_2) \Leftrightarrow I_1 = I_2$$

- Cox demonstro que las leyes de probabilidad no solo aplican a las frecuencias, sino que constituyen el calculo basico de la logica y el razonamiento plausible

Laplace y la probabilidad como medida de nuestra ignorancia

- **Principio de razon suficiente**: afirma que todo debe tener una razón o una causa (Leibniz)

(Laplace, tratado filosofico sobre la probabilidad ~1812)

Debemos, pues, considerar el estado actual del universo como el efecto de su estado anterior y como la causa del que ha de seguir. Una inteligencia que, en un instante dado, pudiera comprender todas las fuerzas por las que está animada la naturaleza y la situación respectiva de los seres que la componen, si además fuera lo suficientemente vasta como para someter estos datos al análisis, abarcaría en la misma fórmula los movimientos de los cuerpos más grandes del universo y los de los átomos más ligeros. Para tal inteligencia nada sería incierto, y el futuro, como el pasado, estaría abierto a sus ojos.

…

La trayectoria de una simple molécula de aire o de vapor está regulada de manera tan cierta como la de las órbitas planetarias; la única diferencia entre ellas es la que aporta nuestra ignorancia

…

La probabilidad es relativa en parte a esta ignorancia y en parte a nuestro conocimiento

…

La teoría de las probabilidades consiste en reducir todos los acontecimientos del mismo tipo a un cierto número de casos igualmente posibles, es decir, a casos cuya existencia es igualmente incierta, y en determinar el número de casos favorables al acontecimiento cuya probabilidad se busca.

$$P(A) = \frac{\text{numero de eventos favorables}}{\text{numero de posibilidades totales}}$$

Laplace y la probabilidad como medida de nuestra ignorancia

… En asuntos que sólo son probables, la diferencia en los datos que cada hombre tiene sobre ellos es una de las principales causas de la diversidad de opiniones que se encuentran para sostener en tales asuntos …

El dilema de Perseo

A

B

C

- Perseo tiene que salvar a la bella Andromeda, la cual se encuentra detras de una de las puertas. En dos de ellas habitan monstruosas gorgonas, y Perseo elige la compuerta B.
- Al elegir, Pegasa, la urraca mítica, interviene, y le dice que en la compuerta A hay una gorgona. Que le conviene a Perseo, cambiar su eleccion o no?

Laplace y la probabilidad como medida de nuestra ignorancia

… En asuntos que sólo son probables, la diferencia en los datos que cada hombre tiene sobre ellos es una de las principales causas de la diversidad de opiniones que se encuentran para sostener en tales asuntos …

El dilema de Perseo

A **B** **C**

- Perseo tiene que salvar a la bella Andromeda, la cual se encuentra detras de una de las puertas. En dos de ellas habitan monstruosas gorgonas, y Perseo elige la compuerta B.
- Al elegir, Pegasa, la urraca mítica, interviene, y le dice que en la compuerta A hay una gorgona. Que le conviene a Perseo, cambiar su eleccion o no?
- RTA: Me conviene cambiar de eleccion!!
  - Originalmente, la probabilidad de que Andromeda este en B es: P(B) = 1/3 y de que este en A o C, P(A o C) = 2/3.
    - Pero dada la nueva evidencia I: “en A hay una gorgona”, ahora debo actualizar mi estado de conocimiento:
    - P(A | I) = 0 (ya que me aseguraron que hay una gorgona), y por ende P(C | I) = 2/3.
    - Entonces, si cambio de eleccion a C voy a duplicar mis chances de exito de 1/3 a 2/3.

El dilema de Perseo

A

B

C

- Supongamos que inicialmente elijo la compuerta B y que luego me aseguran que en la A hay una gorgona.

- A: Andromeda esta en la compuerta A
- B: Andromeda esta en la compuerta B
- C: Andromeda esta en la compuerta C
- $I_{A|B}$: informacion de que la gorgona esta en A luego de haber elegido B

- Voy a calcular usando Bayes la probabilidad de que Andromeda este en B dada la informacion $I_{A|B}$ y compararla con que este en C:

$P(B|I_{A|B})$ vs $P(C|I_{A|B})$ $\longrightarrow$ Me conviene seguir con B o cambiar a C luego de $I_{A|B}$?

$$P(B|I_{A|B}) = \frac{P(I_{A|B}|B)P(B)}{P(I_{A|B})} = \frac{\overbrace{P(I_{A|B}|B)}^{1/2}\overbrace{P(B)}^{1/3}}{\underbrace{P(I_{A|B}|B)}_{1/2}\underbrace{P(B)}_{1/3} + \underbrace{P(I_{A|B}|A)}_{0}\underbrace{P(A)}_{1/3} + \underbrace{P(I_{A|B}|C)}_{1}\underbrace{P(C)}_{1/3}}$$

= 1/6 / (1/6 + 1/3) = 1/3

El dilema de Perseo

A

B

C

- Supongamos que inicialmente elijo la compuerta B y que luego me aseguran que en la A hay una gorgona.

- A: Andromeda esta en la compuerta A
- B: Andromeda esta en la compuerta B
- C: Andromeda esta en la compuerta C
- $I_{A|B}$: informacion de que la gorgona esta en A luego de haber elegido B

- Voy a calcular usando Bayes la probabilidad de que Andromeda este en B dada la informacion $I_{A|B}$ y compararla con que este en C:

$P(B|I_{A|B})$ vs $P(C|I_{A|B})$ $\longrightarrow$ Me conviene seguir con B o cambiar a C luego de $I_{A|B}$?

$$P(C|I_{A|B}) = \frac{P(I_{A|B}|C)P(C)}{P(I_{A|B})} = \frac{\overbrace{P(I_{A|B}|C)}^{1}\overbrace{P(C)}^{1/3}}{\underbrace{P(I_{A|B}|B)}_{1/2}\underbrace{P(B)}_{1/3} + \underbrace{P(I_{A|B}|A)}_{0}\underbrace{P(A)}_{1/3} + \underbrace{P(I_{A|B}|C)}_{1}\underbrace{P(C)}_{1/3}}$$

= 1/3 / (1/6 + 1/3) = 2/3

El dilema de Perseo

A

B

C

- Supongamos que inicialmente elijo la compuerta B y que luego me aseguran que en la A hay una gorgona.

- A: Andromeda esta en la compuerta A
- B: Andromeda esta en la compuerta B
- C: Andromeda esta en la compuerta C
- $I_{A|B}$: $\text{informacion de que la gorgona esta en A luego de haber elegido B}$

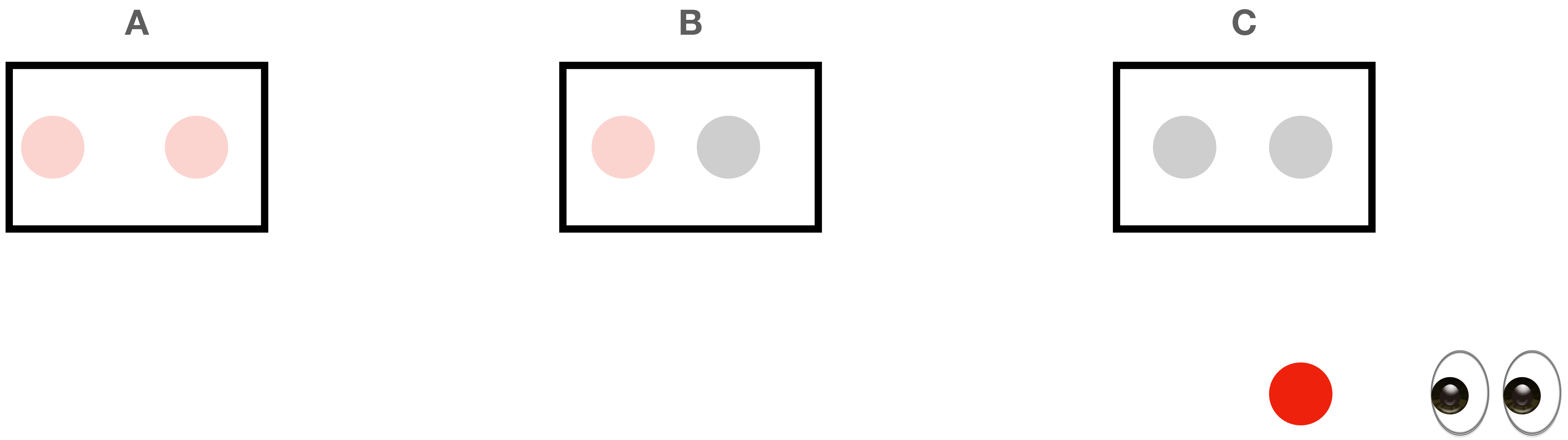
- Voy a calcular usando Bayes la probabilidad de que Andromeda este en B dada la informacion $I_{A|B}$ y compararla con que este en C:

$P(B|I_{A|B})$ vs $P(C|I_{A|B}) \longrightarrow$ Me conviene seguir con B o cambiar a C luego de $I_{A|B}$?

$$\boxed{P(C|I_{A|B}) = 2/3} \qquad P(B|I_{A|B}) = 1/3$$

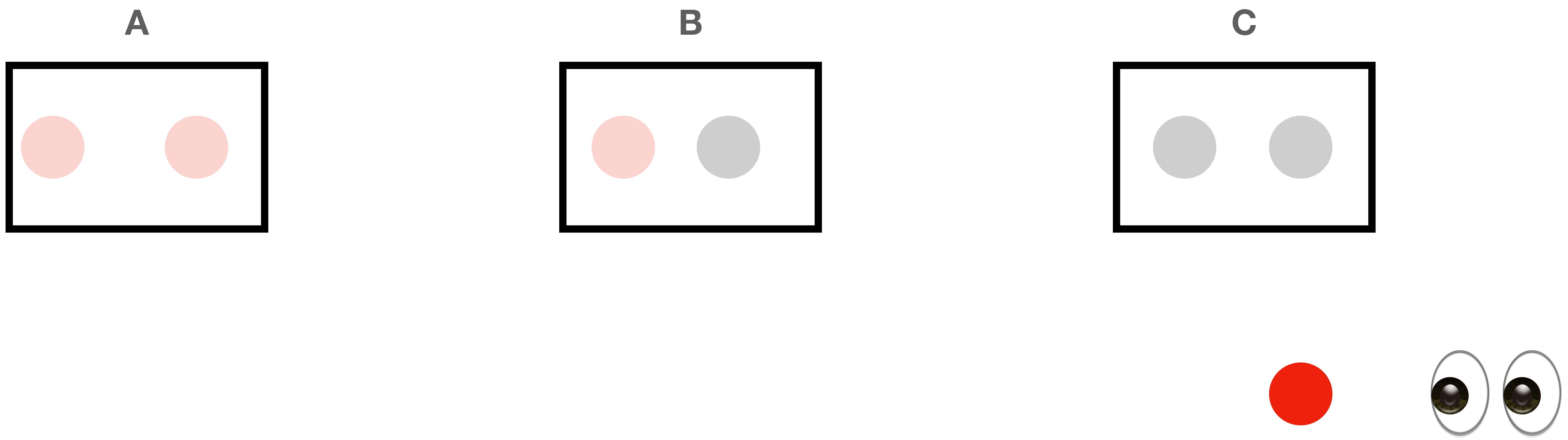
Dada la informacion $I_{A|B}$, me conviene cambiar de eleccion a C dado que es el doble mas probable de acertar

Laplace y la probabilidad como medida de nuestra ignorancia



- Cual es la probabilidad de que la otra bola de la misma caja sea roja?

Laplace y la probabilidad como medida de nuestra ignorancia

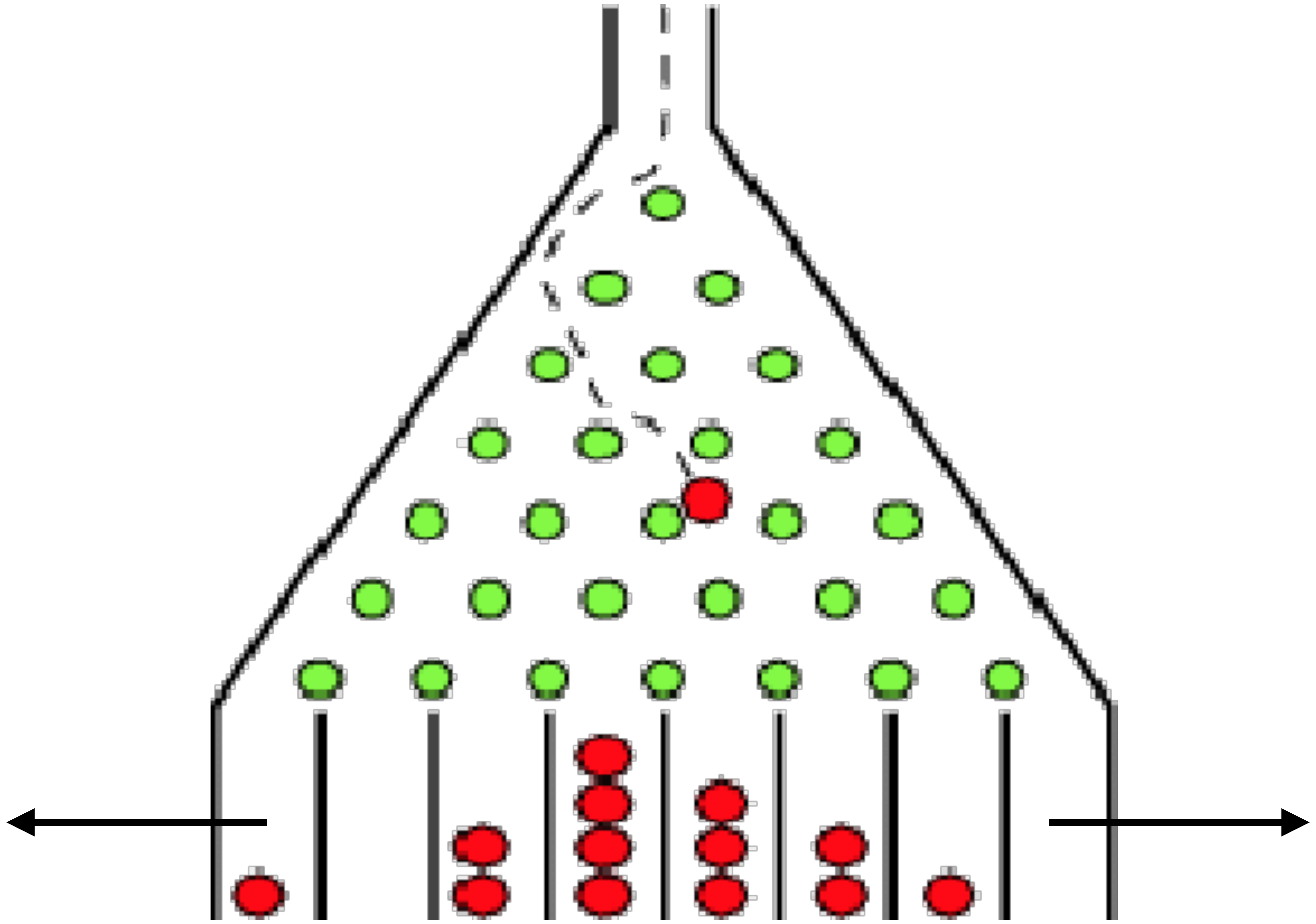


- Cual es la probabilidad de que la otra bola de la misma caja sea roja?

$$P(A|\text{roja}) = \frac{P(\text{roja}|A)P(A)}{P(\text{roja})} = 2/3$$

Experiencia de Galton

La bola termina aqui solo si tuve N exitos de N lanzamientos



La bola termina aqui solo si tuve N fracasos de N lanzamientos

- Exito: si la bola al chocar sale para la izquierda
- Fracaso: si la bola al chocar sale hacia la derecha
- Podemos considerar cada hilera como el lanzamiento de una moneda de dos caras.
- Tenemos pues N hileras de lanzamientos.
- Cuantos exitos o fracasos tendre en N lanzamientos estara dado por una distirbucion binomial

La teoría de la probabilidad no es más que el sentido común reducido al cálculo.
[Laplace, ~1812]

“Es notable que una ciencia, que comenzó con una consideración de los juegos de azar, se eleve al rango de los temas más importantes del conocimiento humano”

Bibliografia

- Data Analysis, A Bayesian Tutorial 2nd edition, D. S. Sivia with J. Skilling
- Elements of Information Theory, 2nd edition, Thomas M. Cover, Joy A. Thomas
- Statistical data analysis, Glen Cowan
- Théorie Analytique des Probabilités, Laplace 1812.
- Essai philosophique sur les Probabilités, Laplace 1820. http://www.probabilityandfinance.com/pulskamp/Laplace/index.html
- An essay towards solving a problem in the doctrine of chances, Mr. Bayes 1763 https://royalsocietypublishing.org/doi/epdf/10.1098/rstl.1763.0053

Agradecimientos

- A mi novia Kari por haberme ayudado con la maqueta y tambien por escuchar e intercambiar opiniones sobre la charla
- A mi abuelo Jose por ayudarme tambien con la maqueta