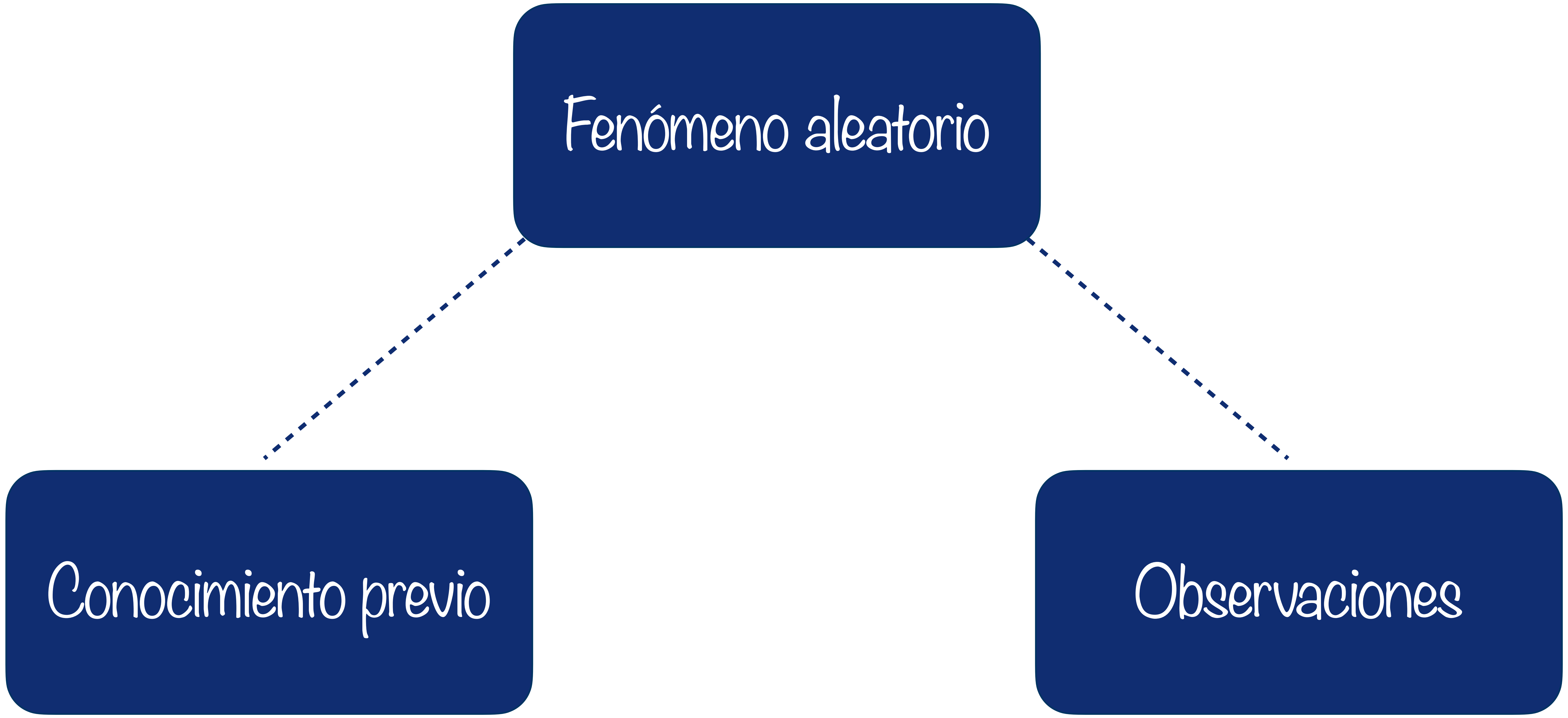
Introducción a la estadística bayesiana

José A. Perusquía Cortés

Estadística Bayesiana. Semestre 2024- II

Estadística

"La estadística se puede definir como el estudio de fenómenos aleatorios"



- Si se cuenta con toda la información bastará con un análisis exploratorio de lo contrario se necesitará un proceso de inferencia

Proceso de inferencia

- La única forma científica de hacer inferencia con base en una muestra es a partir de la probabilidad

$$x \longrightarrow X \longrightarrow \mathbb{P}(X = x)$$

- En ocasiones se puede suponer que $\mathbb{P}(X = x) = p(x \mid \theta)$, donde θ es desconocido (inferencia paramétrica)
- En otras ocasiones la forma funcional es desconocida (inferencia no paramétrica)

Enfoque clásico

- Se asume que θ es una constante fija que se aproxima a partir de un estimador $T(\mathbf{X})$ que se elige a partir de propiedades como:

  - Insesgado
  - Consistente
  - Minimal
  - Suficiente
  - Etc.

¡Los datos se consideran aleatorios a pesar de haber sido observados!

Ejemplo 1

- Para una serie de ensayos Bernoulli se quiere estimar θ^2
- El estimador máximo verosímil está dado por

$$\left(\frac{X}{n}\right)^2$$

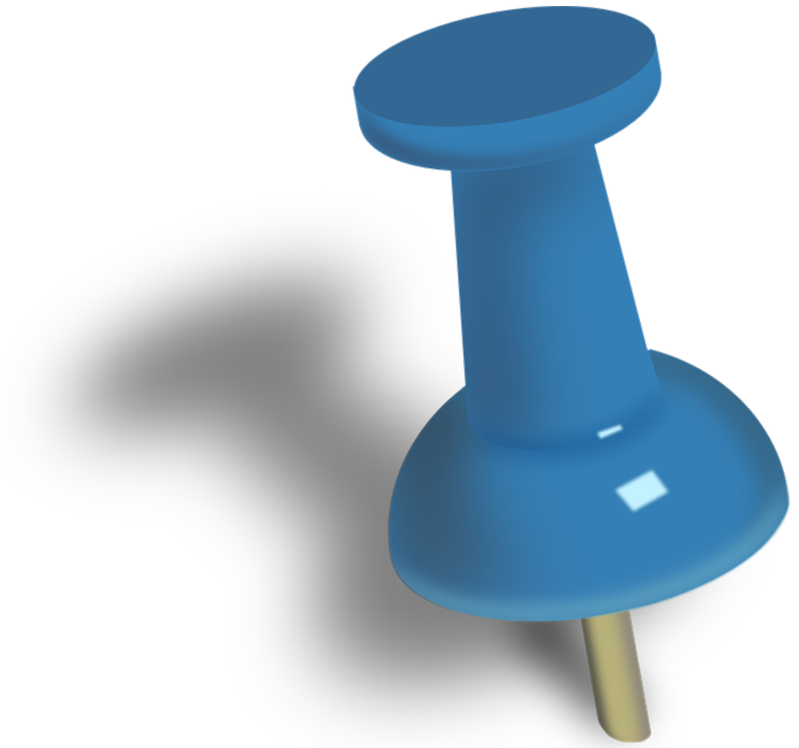
- Como no es insesgado se propone mejor

$$\left(\frac{X(X-1)}{n(n-1)}\right)^2$$

- Si se observa $x = 1$, ¡El estimador es igual a cero!

Ejemplo 2

- Suponer que se lanza una moneda n veces con x águilas
- Suponer que se lanza una tachuela n veces y se observa que cae de cabeza x veces



- **¡Los estimadores máximo verosímiles son los mismos!**
- **No se toman en cuenta los conocimientos previos del experimento**

Reflexión filosófica

- La inferencia clásica responde la pregunta

¿Qué valor de θ hace más plausibles los datos?

- Deberíamos de preguntarnos

¿Qué valor de θ es más probable dados los datos?

¿Habrá alguna forma más coherente y libre de contradicciones de hacer estadística?

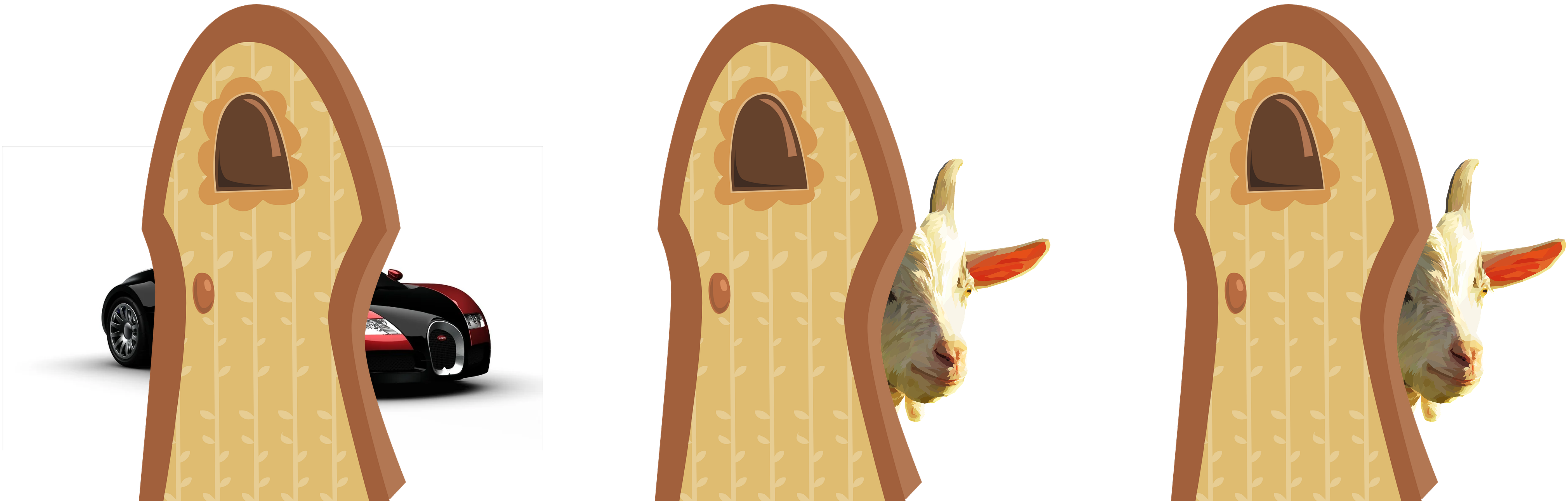
Teorema de Bayes

Dado un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$, $\{E_n\}_n$ una partición finita o numerable de Ω y $A \in \mathscr{F}$ tal que $\mathbb{P}(A) > 0$. Entonces,

$$\mathbb{P}(E_n \,|\, A) = \frac{\mathbb{P}(A \,|\, E_n) \cdot \mathbb{P}(E_n)}{\sum_m \mathbb{P}(A \,|\, E_m) \cdot \mathbb{P}(E_m)}$$

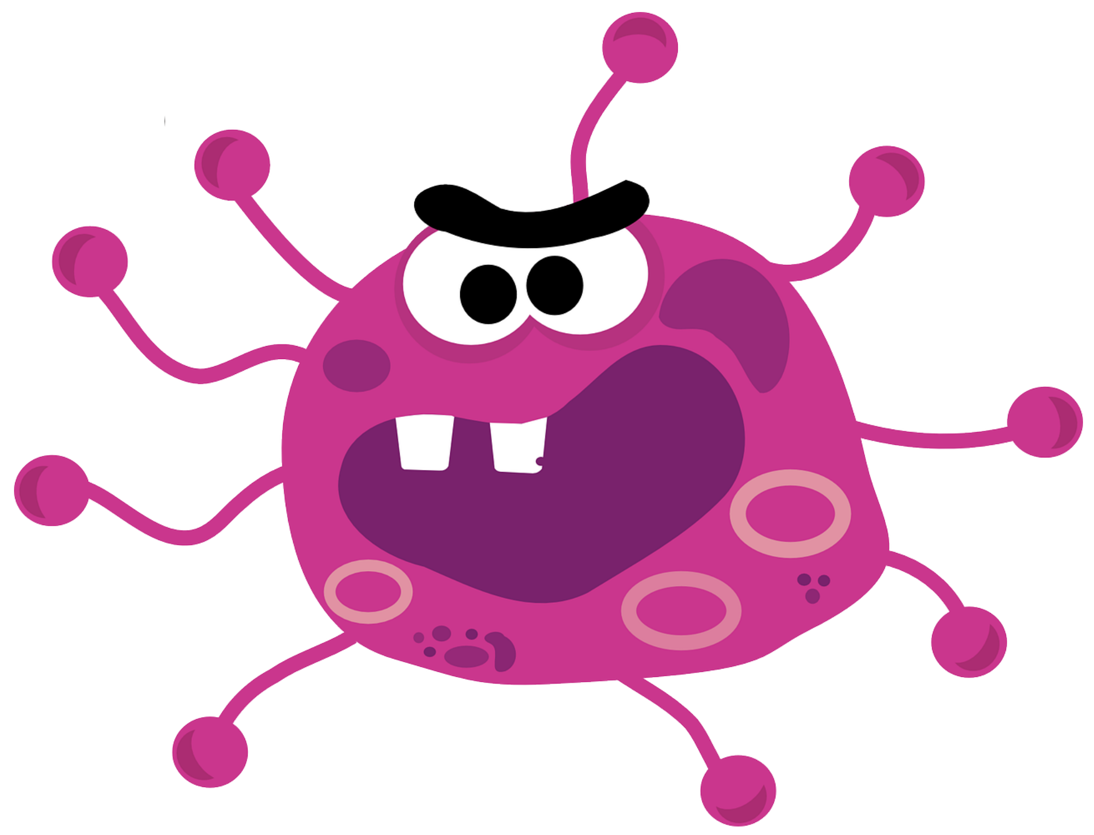
Ejemplo 3. Problema de Monty Hall.

En un programa televiso un concursante elige una puerta con la posibilidad de ganar un auto. El presentador abrirá inmediatamente una puerta conteniendo una cabra y le dará la oportunidad de cambiar puerta al concursante. ¿Cuál es la mejor decisión para el concursante?



Ejemplo 4. Aplicación a epidemiología.

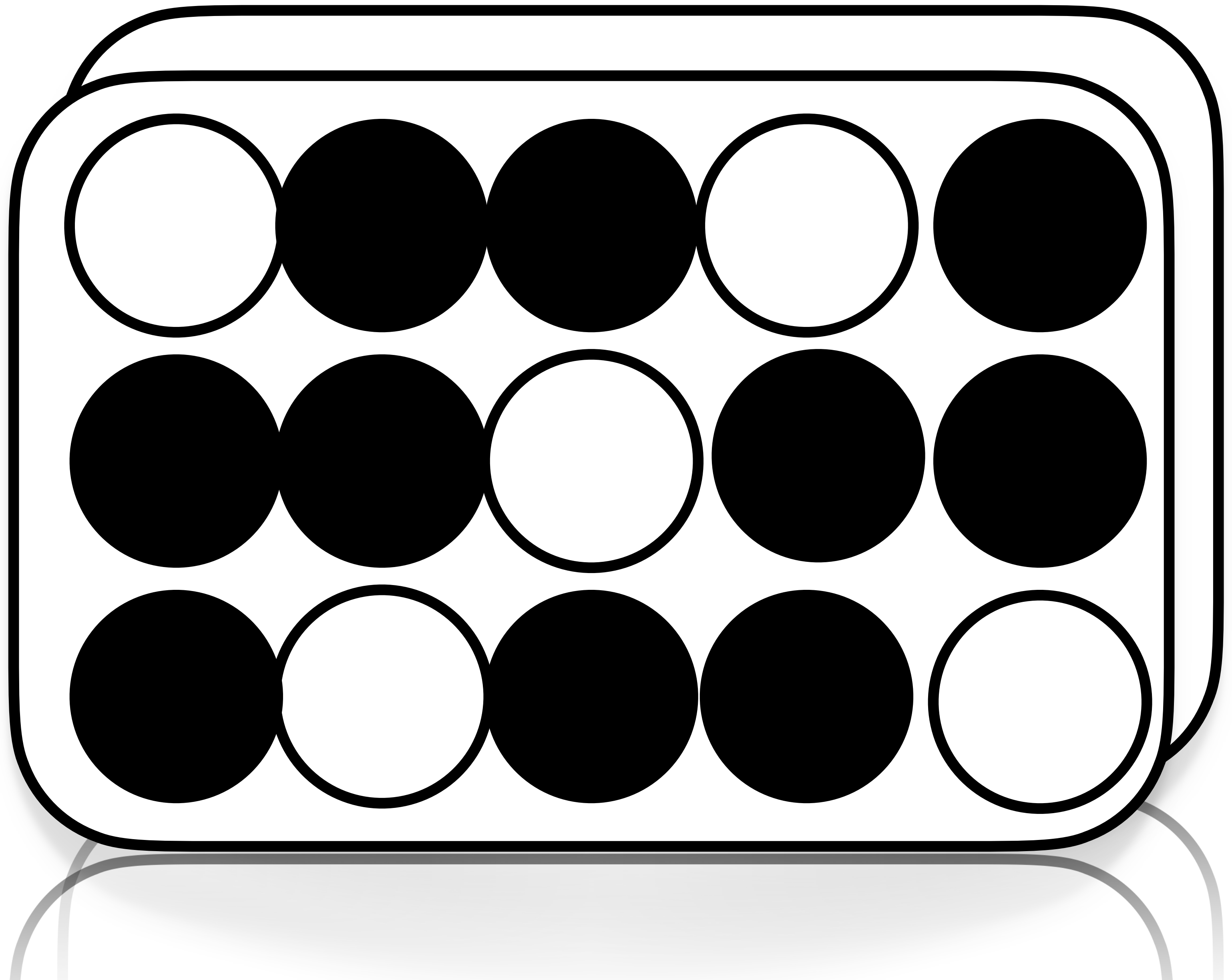
Una cierta enfermedad tiene una incidencia del 2%. Si la tasa de falsos negativos de una prueba es del 10% y la tasa de faltos positivos es del 1%, ¿Cuál es la probabilidad de que una persona que tenga un test positivo realmente tenga la enfermedad?



Ejemplo 5. Problema de urnas

Suponer que se tiene un urna con 100 bolas de las cuales $100-n$ son blancas y n son negras, en donde n se distribuye uniformemente en $\{0,1,\ldots,100\}$. Si en el primer intento se saca una bola blanca y no se regresa a la urna, entonces la segunda bola:

1. Es más probable que salga blanca.
2. Es más probable que salga negra.
3. Son equiprobables.



Naive Bayes

- Un conjunto de modelos bayesianos de aprendizaje supervisado y utilizado en gran medida para la clasificación de textos

- Dado un conjunto de observaciones $\{\mathbf{x}_i\}_{i=1}^{N}$ para los cuales **se conoce el grupo al que pertenecen**, denotado por $\{y_i\}_{i=1}^{N}$ se asume que:

$$\mathbb{P}(\mathbf{x}_i \mid y_i) = \prod_{j=1}^{M} \mathbb{P}(x_{ij} \mid y_i)$$

- En otras palabras, dado el grupo los atributos son **condicionalmente independientes**

Naive Bayes

- Para una nueva observación $\mathbf{x}_{N+1}$ se busca encontrar el grupo al que pertenece, calculando para toda j

$$\mathbb{P}(y_j \mid \mathbf{x}_{N+1}) \propto \mathbb{P}(\mathbf{x}_{N+1} \mid y_j)\mathbb{P}(y_j)$$

- Se elige el grupo que tenga la **mayor probabilidad**

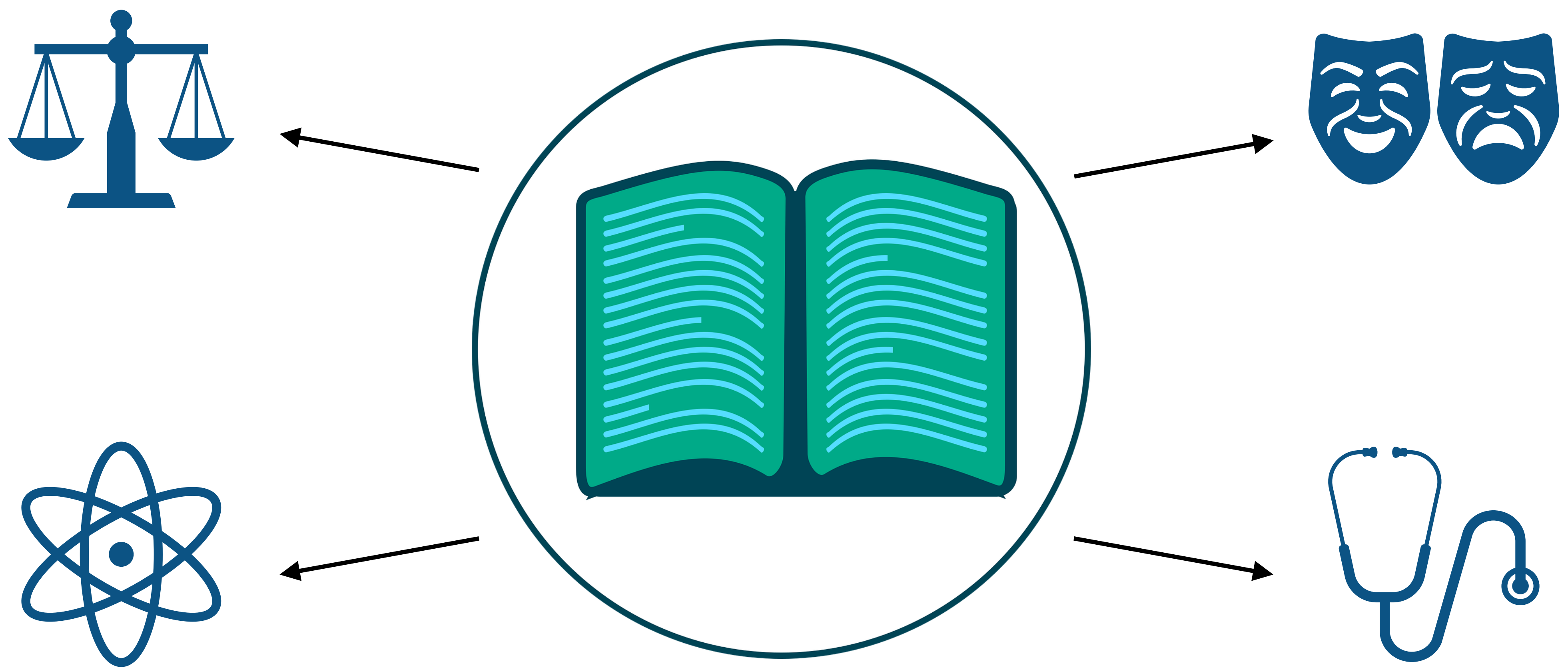
- Para evitar posible problemas numéricos se considera en su lugar una transformación logarítmica, esto es,

$$\log(\mathbb{P}(y_j \mid \mathbf{x}_{N+1})) \propto \log(\mathbb{P}(\mathbf{x}_{N+1} \mid y_j)) + \log(\mathbb{P}(y_j))$$

Naive Bayes

- Dependiendo del tipo de atributos se tiene:
  - **Gaussian naive Bayes** para datos continuos.
  - **Multinomial naive Bayes** para frecuencias.
  - **Bernoulli naive Bayes** para atributos dicotómicos.

- Para la clasificación de textos nos centramos (principalmente) en el clasificador multinomial o Bernoulli.



Ejemplo 6. Detección de spam

- El spam es cualquier correo electrónico no solicitado con una gran variedad de fines, desde:
  - Propaganda comercial
  - Robo de identidad
  - Difusión de virus informáticos

- Dado un conjunto de entrenamiento se obtienen a partir de las frecuencias:
  - $\mathbb{P}(\text{Spam})$ y $\mathbb{P}(\text{Ham})$
  - $\mathbb{P}(\omega_i \mid \text{Spam})$ y $\mathbb{P}(\omega_i \mid \text{Ham})$ para cada palabra ω_i en el vocabulario
  - Palabras que no se hayan visto en el entrenamiento se omiten

Ejemplo 6. Detección de spam

- Considerando el modelo Bernoulli se tiene que

$$\mathbb{P}(\omega_i \,|\, \text{Spam}) = \frac{\sum_k 1(\omega_i, \text{Spam}_k)}{\#\text{Spam}}$$

- Para evitar probabilidades nulas se realiza el **suavizamiento de Laplace**

$$\mathbb{P}(\omega_i \,|\, \text{Spam}) = \frac{\sum_k 1(\omega_i, \text{Spam}_k) + 1}{\#\text{Spam} + 2}$$

Ejemplo 6. Detección de spam

- Considerando el modelo multinomial se tiene que

$$\mathbb{P}(\omega_i \,|\, \text{Spam}) = \frac{\#(\omega_i, \text{Spam})}{\sum_i \#(\omega_i, \text{Spam})}$$

- Para evitar probabilidades nulas se realiza el **suavizamiento de Laplace**

$$\mathbb{P}(\omega_i \,|\, \text{Spam}) = \frac{\#(\omega_i, \text{Spam}) + \alpha}{\sum_i (\#(\omega_i, \text{Spam}) + \alpha)}$$

- Por lo general $\alpha = 1$