ICAS

icas.unsam.edu.ar

Aprendizaje Automático

Probabilidad e Inferencia Bayesiana

Prof. Rodrigo Díaz
Lic. Manuel Szewc
Lic. Luis Agustín Nieto

UNSAM - 18 de marzo de 2021

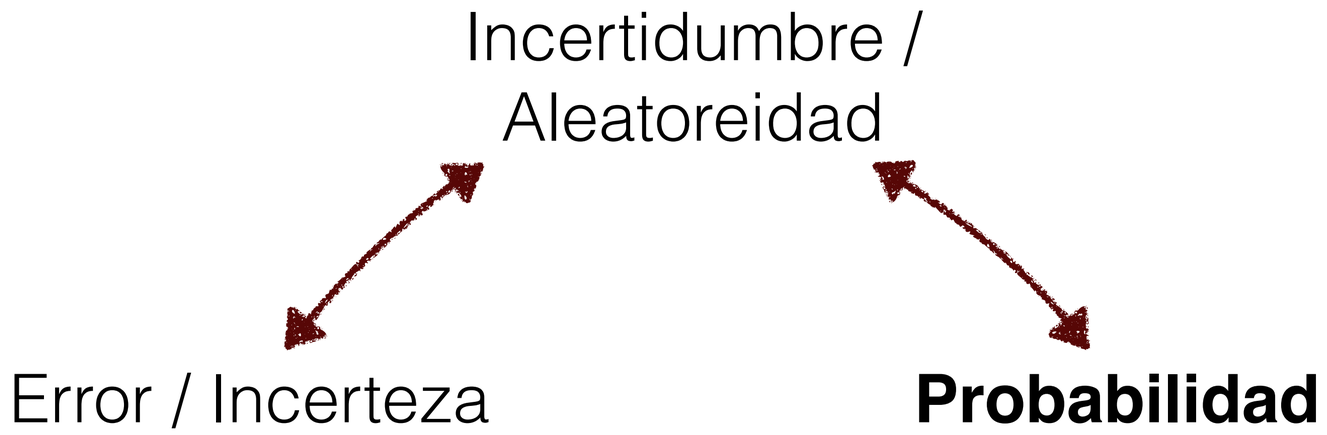
Probabilidad

¿Por qué tenemos que ver este tema?

ICAS
icas.unsam.edu.ar

- ¿Qué pasa si quiero describir con mi **modelo** eventos que no son deterministas (es decir, que tienen ciertas incertidumbre)?

- ¿Existe realmente alguna posibilidad de hacer un modelo determinista de los datos? Si reconocemos la existencia de incerteza, la aleatoreidad siempre está presente.



Probabilidad

Reglas de la probabilidad

ICAS
icas.unsam.edu.ar

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

Regla del producto

Teorema de Bayes

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

Regla de la probabilidad total

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

Probabilidad

Reglas de la probabilidad

ICAS
icas.unsam.edu.ar

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

Regla del producto

Teorema de Bayes

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

Regla de la suma

$$P(A + B) = P(A) + P(B) - P(A, B)$$

Probabilidad

Diferentes interpretaciones.

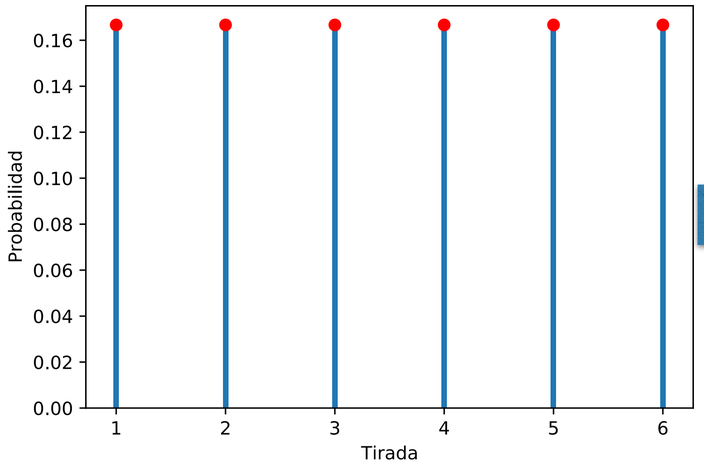
ICAS
icas.unsam.edu.ar

- Clásica (Laplace). Principio de indiferencia. Las probabilidades se reparten de forma igual entre todos los posibles resultados, suponiendo que pueden ser considerados iguales.
- Frecuentista (empírica). La probabilidad de un evento es igual a su frecuencia relativa (en relación con la cantidad de intentos posibles), cuando el número de ensayos aumenta.
- Bayesiana. También llamada probabilidad epistémica. Entiende a la probabilidad como un grado de incertidumbre.

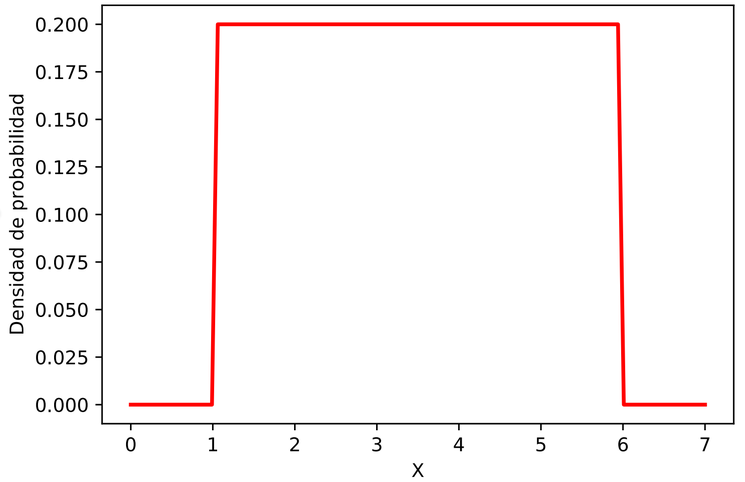
¿Cuál es la probabilidad de obtener un as al tirar un dado no cargado?

Funciones de densidad de probabilidad (PDF)

PMF



PDF



$$\forall i, f(x_i) \geq 0$$

$$\sum_i f(x_i) = 1$$

$$f(x) \geq 0 \qquad \int_{-\infty}^{\infty} f(x)\mathrm{d}(x) = 1$$

$$P(x \in (a, b)) = \int_a^b f(x)\mathrm{d}x$$

Probability

Estimation of the moments using samples

ICAS
icas.unsam.edu.ar

Si en lugar de la ley de probabilidad, tuvieramos una **muestra** (¡lo que ocurre a menudo!)

$$\{x_i\} \qquad i = 1 \dots N$$

$$\mathbb{E}_f[x] \longrightarrow \bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Ley de los grandes números

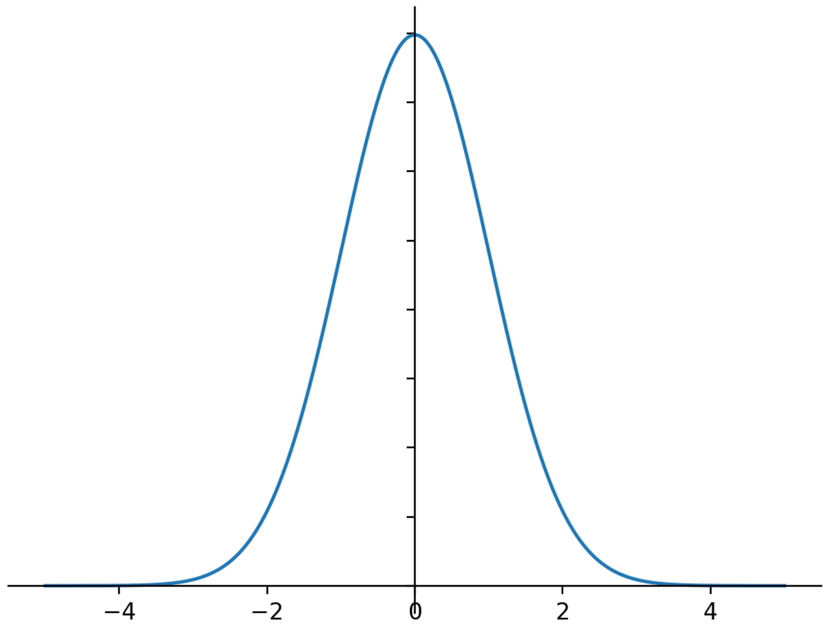
$$\mathrm{var}_f[x] \longrightarrow \bar{S} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{X})^2$$

Distribución Normal / Gaussiana

En una sola dimensión

ICAS
icas.unsam.edu.ar

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathbb{E}[x] = \mu$$

$$\mathrm{var}[x] = \sigma^2$$

Distribución de Bernoulli

- Idea.Tengo un proceso que tiene una probabilidad \mu de éxito, y solo dos resultados (éxito=1; fracaso=0). Ejemplo: tirar una moneda.
- El espacio de muestreo, entonces, es S = {0, 1}, y llamemos a la variable aleatoria X.
- ¿Podemos escribir la PMF, f, de este proceso?

$$P(X = 1) = f(1) = \text{???} \qquad P(X = 0) = f(0) = \text{???}$$

$$f(X) = \mu^X \cdot (1 - \mu)^{(1-X)}$$

Distribución de Bernoulli

¿Cuál es el valor de expectación? ¿Y la varianza?

Distribución de Bernoulli

$$f(X) = \mu^X \cdot (1-\mu)^{(1-X)}$$

Distribución de Bernoulli

¿Cuál es el valor de expectación? ¿Y la varianza?

$$\mathbb{E}_f[x] = \sum_x x f(x)$$

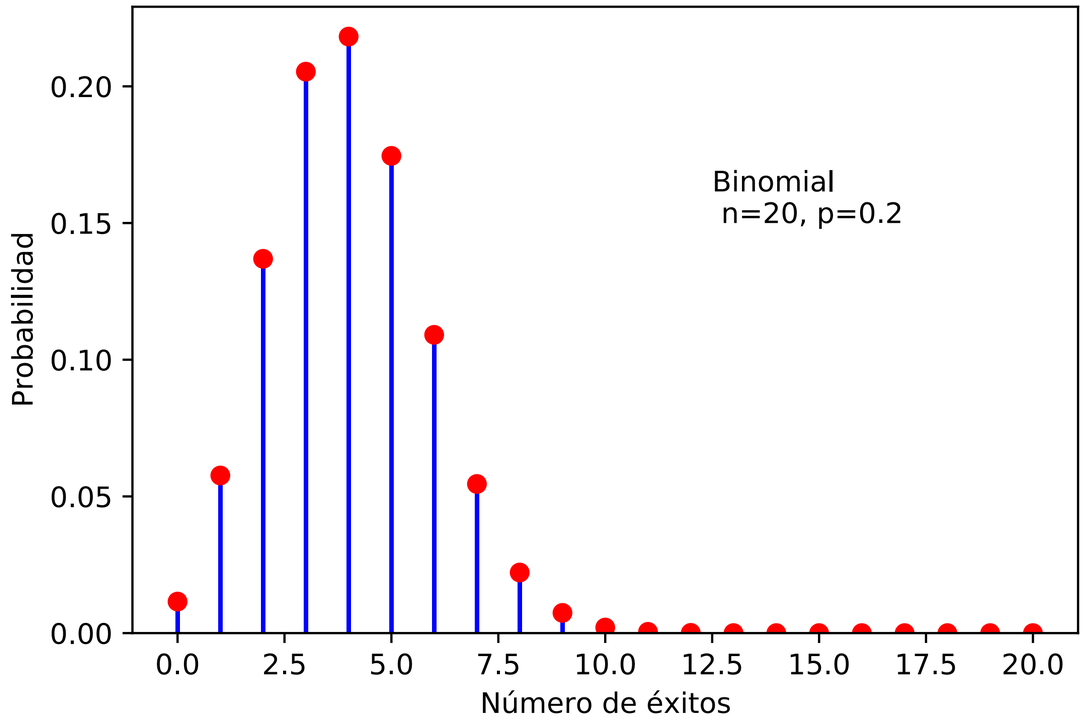
$$\text{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2$$

Respuesta:

Distribución Binomial

- Idea.Describe el resultado de *n* experimentos con una variable de Bernouilli de probabilidad mu

$$p(k|n,\mu) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$$



Probabilidad bayesiana

Thomas Bayes (1701 – 1761)

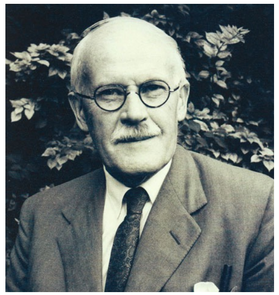
First appearance of the **product rule** (the base for the Bayes' theorem; *An Essay towards solving a Problem in the Doctrine of Chances*).

$$p(H_i|I,D) = \frac{p(D|H_i,I)}{p(D|I)} \cdot p(H_i|I)$$

Pierre-Simon Laplace (1749 – 1827)

Wide application of the **Bayes' rule.** Principle of insufficient reason (non-informative priors). Primitive version of the Bernstein–von Mises theorem.

Laplace's "inverse probability" is largely rejected for ~100 years. The reign of frequentist probability. Fischer, Pearson, etc.



Harold Jeffreys (1891 – 1989)

Objective Bayesian probability revived.
Jeffreys rule for priors.

(1940s - 1960s)
R. T. Cox
George Pólya
E. T. Jaynes

Plausible reasoning. Reasoning with uncertainty.
Probability theory as an extension of Aristotelian logic.
The product and sum rules deduced for basic principles.
MAXENT priors.

See E.T Jaynes. Probability Theory: The Logic of Science.

http://www-biba.inrialpes.fr/Jaynes/prob.html

The three desiderata

- Rules of deductive reasoning: strong syllogisms from Aristotelian logic.
- Brain works using plausible inference. Woman in wood cabin story.
- The rules for plausible reasoning are deduced from three simple desiderata on how the mind of a thinking robot should work (Cox-Póyla-Jaynes).

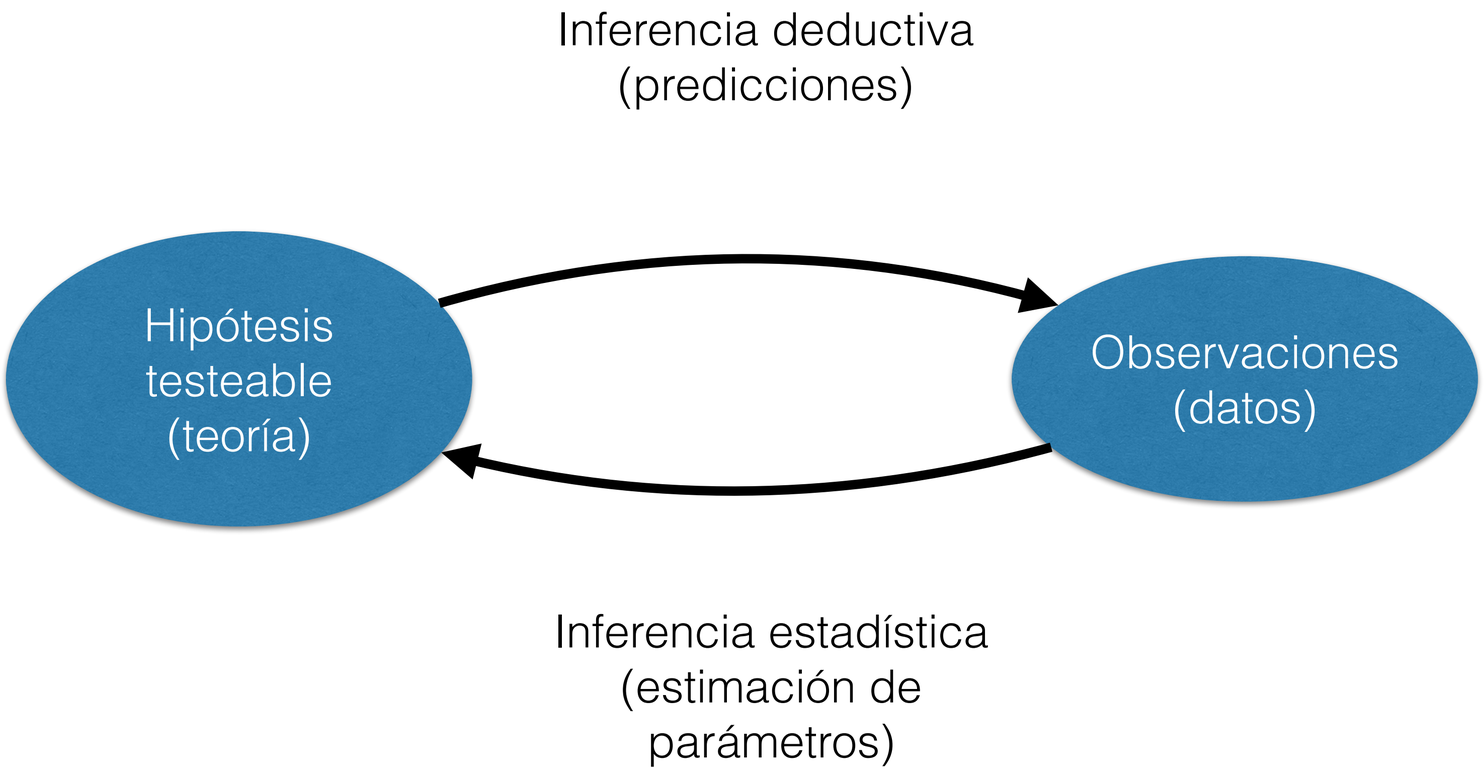
I. Degrees of Plausibility are represented by real numbers.

II. Qualitative Correspondence with common sense.

III. If a conclusion can be reasoned out in more than one way; then every possible way must lead to the same result.

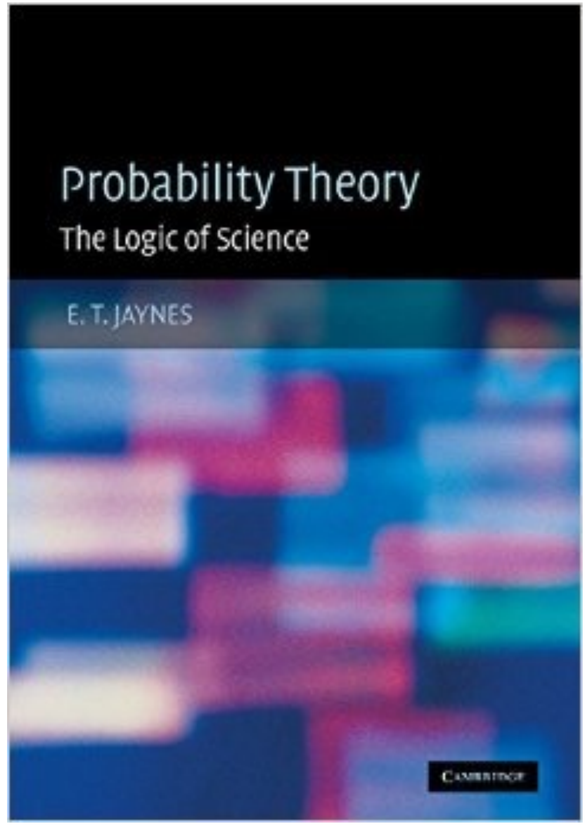
Una visión unificada

Fig. adapted from Gregory (2005)



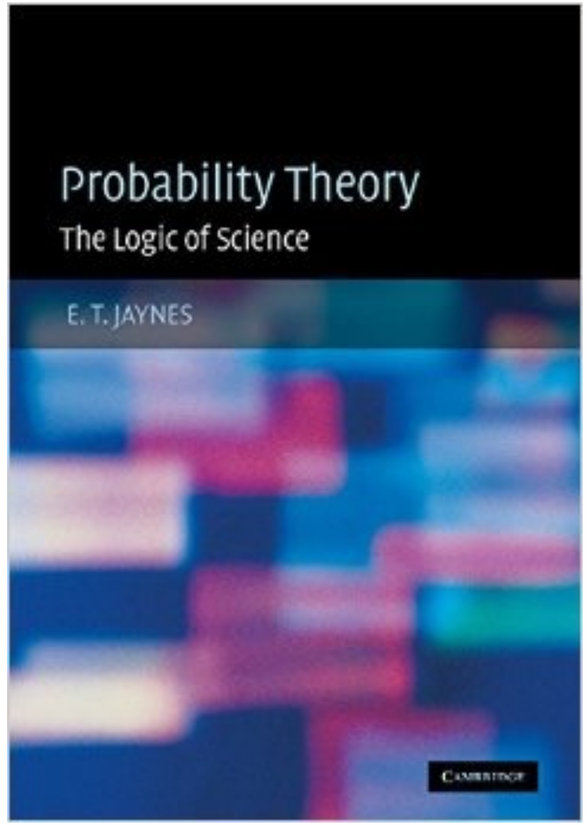
"…if degrees of plausibility are represented by real numbers, then there is a **uniquely determined set of quantitative rules for conducting inference**. That is, any other rules whose results conflict with them will necessarily violate an elementary –and nearly inescapable – desideratum of rationality or consistency.

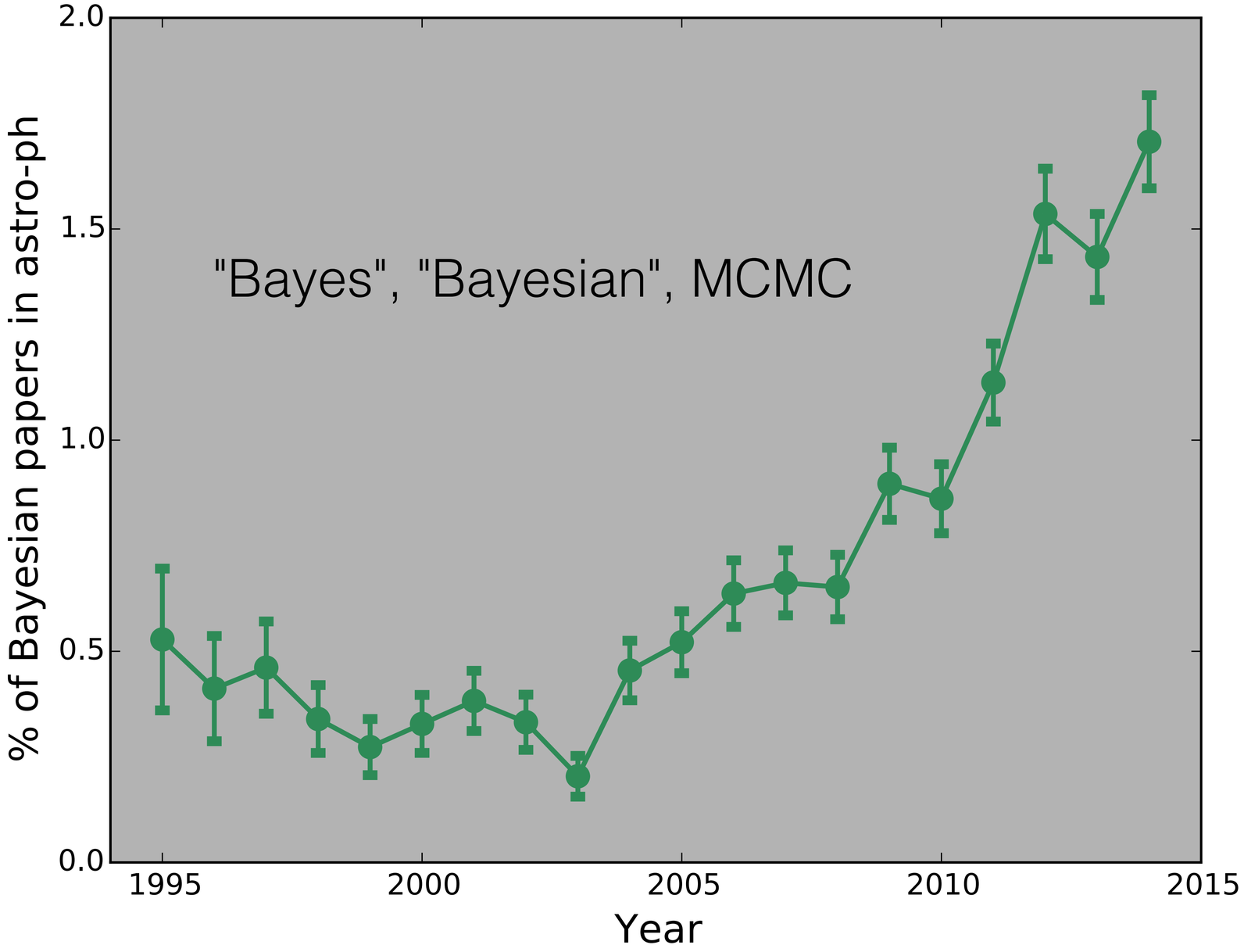
But the final result was just the **standard rules of probability theory**, given already by Daniel Bernoulli and Laplace; so why all the fuss? The important new feature was that these rules were **now seen as uniquely valid principles of logic in general, making no reference to 'chance' or 'random variables'**; so their range of application is vastly greater than had been supposed in the conventional probability theory that was developed in the early 20th century. As a result, **the imaginary distinction between 'probability theory' and 'statistical inference' disappears, and the field achieves not only logical unity and simplicity, but far greater technical power and flexibility in applications**."

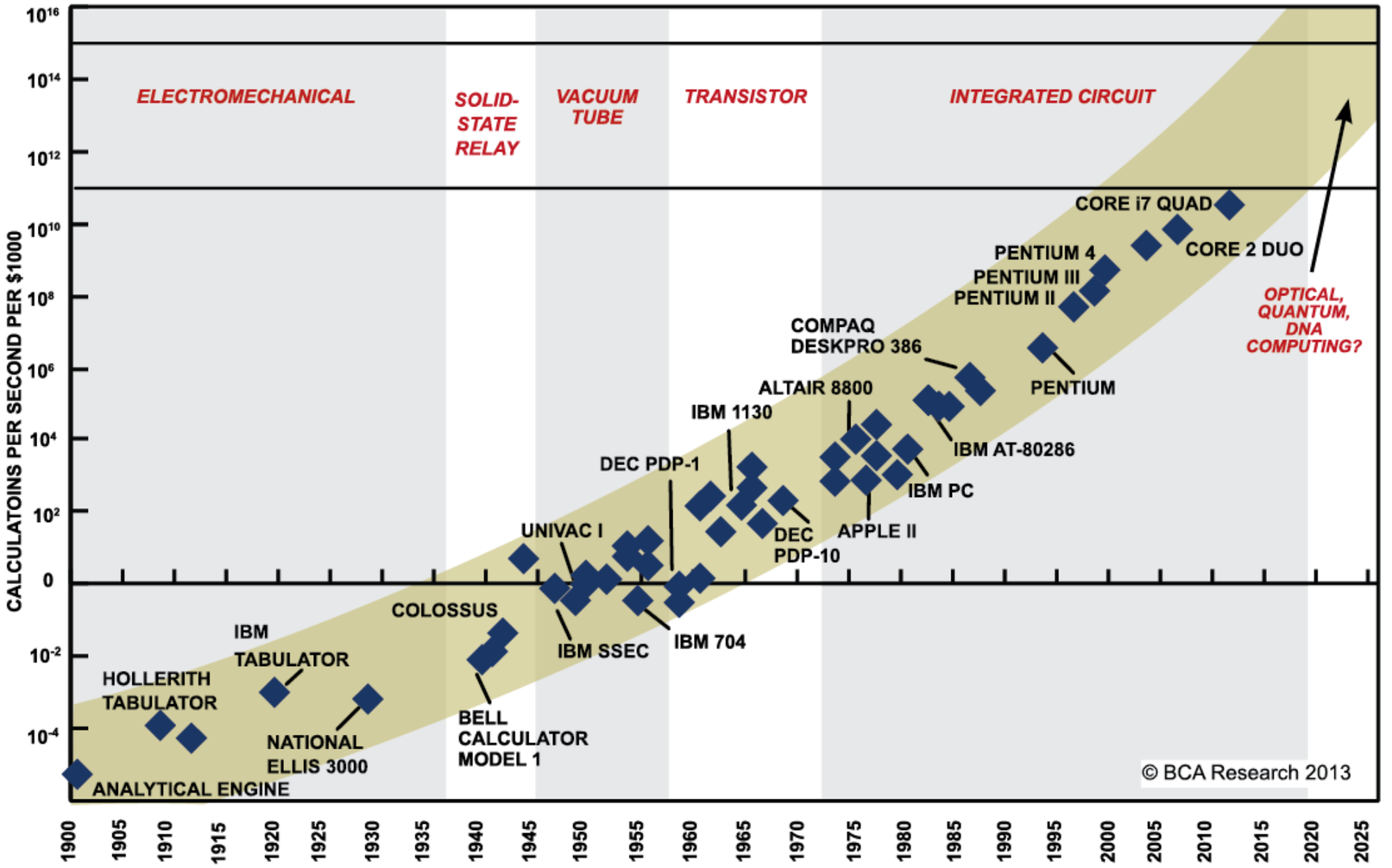


"…if degrees of plausibility are represented by real numbers, then there is a **uniquely determined set of quantitative rules for conducting inference**. That is, any other rules whose results conflict with them will necessarily violate an elementary –and nearly inescapable – desideratum of rationality or consistency.

But the final result was just the **standard rules of probability theory**, given already by Daniel Bernoulli and Laplace; so why all the fuss? The important new feature was that these rules were **now seen as uniquely valid principles of logic in general, making no reference to 'chance' or 'random variables'**; so their range of application is vastly greater than had been supposed in the conventional probability theory that was developed in the early 20th century. As a result, **the imaginary distinction between 'probability theory' and 'statistical inference' disappears, and the field achieves not only logical unity and simplicity, but far greater technical power and flexibility in applications.**"




"Bayes", "Bayesian", MCMC
% of Bayesian papers in astro-ph
Year
0.0
0.5
1.0
1.5
2.0
1995
2000
2005
2010
2015



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

Modelos

$$d_i = m_i + e_i \qquad i = \{1, \ldots, N\}$$

data **model** **error**

$$d_i = \{x_i, t_i\} \qquad m_i = g(x_i|\vec{\theta})$$

Two basic tasks of statistical inference

`Learning process`

(parameter estimation)

`Decision making`

(model comparison)

Learning process

Bayesian **probability** represents a state of

knowledge

$$p(\bar{\theta}|H_i, I) \longrightarrow p(\bar{\theta}|D, H_i, I)$$

Prior → Posterior

$\bar{\theta}$: parameter vector
H_i: hypothesis
I: information
D: data

Discrete space
(hypothesis space)

$$p(H_i|I) \longrightarrow p(H_i|I, D)$$

Learning process

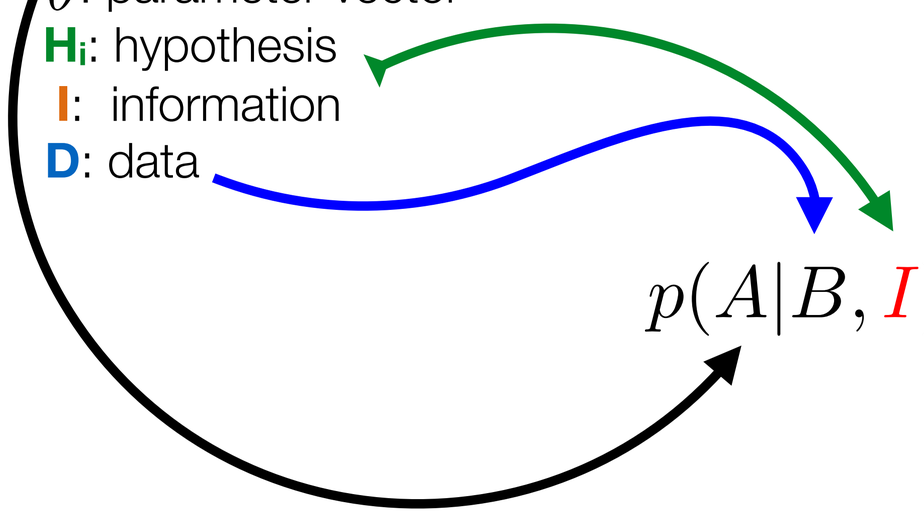
Bayesian **probability** represents a state of

knowledge

$$p(\bar{\theta}|H_i, I) \longrightarrow p(\bar{\theta}|D, H_i, I)$$

Prior → Posterior

$\bar{\theta}$: parameter vector
H_i: hypothesis
I: information
D: data



$$p(A|B, I) = \frac{p(B|A, I)\, p(A|I)}{p(B|I)}$$

Learning process

$\bar{\theta}$: parameter vector
H_i: hypothesis
I: information
D: data

Entra la función verosimilitud (**likelihood**)

$$p(\bar{\theta}|H_i, I, D) = \frac{p(D|\bar{\theta}, H_i, I)}{p(D|H_i, I)} \cdot p(\bar{\theta}|H_i, I)$$

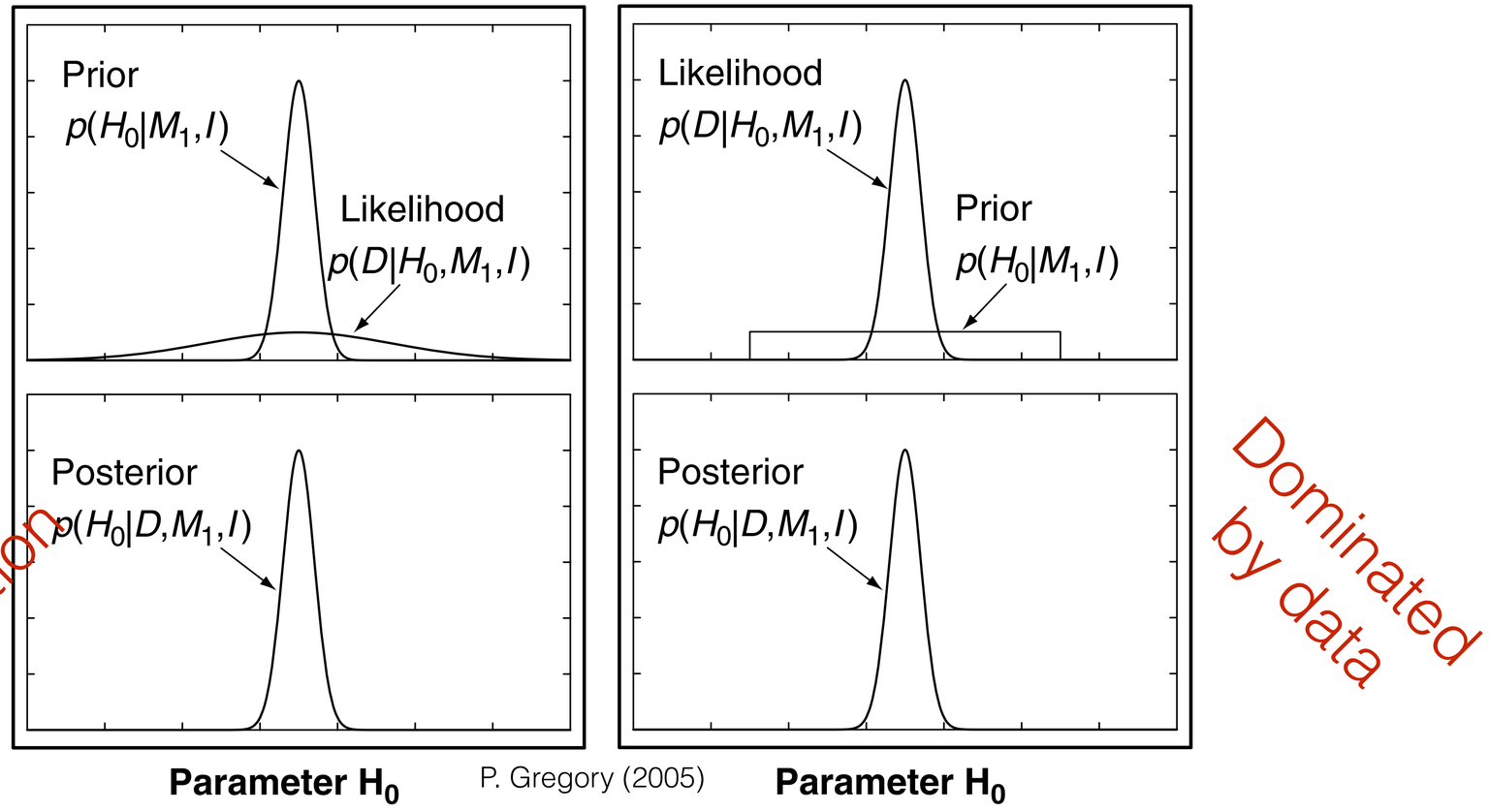
Posterior — Prior

$$p(\bar{\theta}|D, H_i, I) \propto \mathcal{L}_\theta(H_i) \cdot p(\bar{\theta}|H_i, I)$$

- Vista como una función de los parámetros, la función de distribución de los datos, D, se llama **verosimilitud**.
- La constante de proporcionalidad tiene muchos nombres: verosimilitud marginalizada, evidencia, *prior predictive*, etc. Es difícil de calcular. Central para la comparación de modelos.

Optimising the learning process

- The likelihood needs to be selective for the learning process to be effective.



P. Gregory (2005)

$$\underset{\textbf{data}}{d_i} = \underset{\textbf{model}}{m_i} + \underset{\textbf{error}}{e_i}$$

Constructing the likelihood

The data:

$$D = D_1\, D_2 \ldots D_n = \{D_i\}$$

D_i: the *i-th* measurement is in the infinitesimal range y_i to $y_i + \mathrm{d}y_i$

The errors:

E_i: the *i-th* error is in the infinitesimal range e_i to $e_i + \mathrm{d}e_i$

$p(E_i|\theta, H, I) = f_E(e_i)$ The probability distribution of statement E_i

Most used f_E

$$f_E(e_i) = N(0, \sigma_i^2)$$

The model:

M_i: the *i-th* error is in the infinitesimal range m_i to $m_i + \mathrm{d}m_i$

$p(M_i|\theta, H, I) = f_M(m_i)$ The probability distribution of statement M_i

Constructing the likelihood

The data:

$$D = D_1\, D_2 \ldots D_n = \{D_i\}$$

We want to build the probability distribution:

$$p(D|\theta, H, I) = p(D_1, D_2, ..., D_n|\theta, H, I)$$

Write: $y_i = m_i + e_i$

It can be shown that:

$$p(D_i|\theta, H, I) = \int \mathrm{d}m_i\, f_M(m_i)\, f_E(y_i - m_i)$$

Convolution equation

Constructing the likelihood

$$p(D_i|\theta, H, I) = \int \mathrm{d}m_i \, f_M(m_i) \, f_E(y_i - m_i)$$

But for a **deterministic model**, m_i is obtained from a (usually analytically) function f without any uncertainty (say, a Keplerian curve for RV measurements)

$$m_i = f(x_i|\theta)$$
$$f_M(m_i) = \delta(m_i - f(x_i|\theta))$$

Then:
$$p(D_i|\theta, H, I) = \int \mathrm{d}m_i \, \delta(m_i - f(x_i|\theta)) \, f_E(y_i - m_i)$$
$$= f_E(y_i - f(x_i|\theta) = p(E_i|\theta, H, I)$$

$$p(D|\theta, H, I) = p(D_1, D_2, ..., D_n|\theta, H, I) = p(E_1, E_2, ..., E_n|\theta, H, I)$$

Constructing the likelihood

$$p(D|\theta, H, I) = p(D_1, D_2, ..., D_n|\theta, H, I) = p(E_1, E_2, ..., E_n|\theta, H, I)$$

Now, for **independent errors**

$$\begin{aligned} p(D|\theta, H, I) &= p(E_1, E_2, ..., E_n|\theta, H, I) \\ &= p(E_1|\theta, H, I)...p(E_n|\theta, H, I) \\ &= \prod_{i=1}^{n} p(E_i|\theta, H, I) \end{aligned}$$

And, if in addition, the errors are **distributed normally:**

$$p(D|\theta, H, I) \overset{indep., gauss.}{\propto} \exp -\frac{\chi^2_\theta}{2}$$