Bayesian hypothesis testing

Stefan Czesla

The unity of all science consists alone in its method,
not in its material

Karl Pearson (1892)

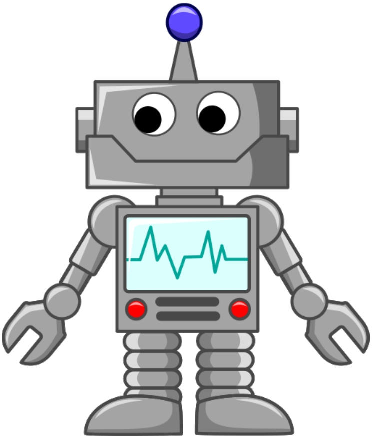
The reasoning robot

Jaynes 2003, The logic of science

The robot shall reason about Aristotelian propositions:

$$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \ldots$$

What are the rules of reasoning?



Logic: Propositional calculus

All logic functions can be represented by negation and conjunction:

Negation: $\bar{a}$
True if a is false

Conjunction: $c = ab$
True iff a and b are true

For convenience, we also define the disjunction

Disjunction: $d = a + b \quad (= \overline{\bar{a}\bar{b}})$

Unfortunately, certainty is rare. What then?

Cox's theorem

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two propositions and

$$\boldsymbol{b}|\boldsymbol{a}$$

be a measure[1] of reasonable credibility in $\boldsymbol{b}$ given $\boldsymbol{a}$ is true.

> **Desideratum**: $\boldsymbol{b}|\boldsymbol{a}$ is represented by a **real number**.
> xxxxxxxxxxxxxx Greater credibility $\rightarrow$ **larger** number

Immediate consequence: Comparability

How does this measure transform?

Cox 1946; Jaynes 2003 (The logic of science); Van Horne 2003

[1] Cox calls $\boldsymbol{b}|\boldsymbol{a}$ the *likelihood*

Cox's theorem

Cox's first assumption:

$$\boldsymbol{c} \cdot \boldsymbol{b} | \boldsymbol{a} = F(\boldsymbol{b} | \boldsymbol{a},\ \boldsymbol{c} | \boldsymbol{b} \cdot \boldsymbol{a})$$

with continuous, strictly monotonic function F.

Cox's example

$\boldsymbol{b}$: A sprinter can run from A to B
$\boldsymbol{c}$: The sprinter can run A–B–A
$\boldsymbol{a}$: Landscape, course, etc.

Cox's theorem

The solution reads

$$w(\boldsymbol{c} \cdot \boldsymbol{b} | \boldsymbol{a}) = w(\boldsymbol{b}|\boldsymbol{a})\, w(\boldsymbol{c}|\boldsymbol{b} \cdot \boldsymbol{a})$$

with continuous, monotonic function w.

Letting $\boldsymbol{c} = \boldsymbol{b}$, we obtain

$$\begin{aligned} w(\boldsymbol{b} \cdot \boldsymbol{b}|\boldsymbol{a}) &= w(\boldsymbol{b}|\boldsymbol{a})\, w(\boldsymbol{b}|\boldsymbol{b} \cdot \boldsymbol{a}) \\ w(\boldsymbol{b}|\boldsymbol{a}) &= w(\boldsymbol{b}|\boldsymbol{a})\, w(\boldsymbol{b}|\boldsymbol{b} \cdot \boldsymbol{a}) \\ w(\boldsymbol{b}|\boldsymbol{b} \cdot \boldsymbol{a}) &= 1 \quad \text{certainty} \end{aligned}$$

Cox's theorem

Second assumption:

$$w(\sim \boldsymbol{b}|\boldsymbol{a}) = S(w(\boldsymbol{b}|\boldsymbol{a}))$$

with some function S.

$$S(x) = (1 - x^m)^{1/m} \quad \text{and} \quad 0 < m < \infty$$

Solution

$$w^m(\boldsymbol{b}|\boldsymbol{a}) + w^m(\sim \boldsymbol{b}|\boldsymbol{a}) = 1$$

Cox's theorem

The sum and product rule (to the m^{th}) power:

$$\begin{aligned} 1 &= w^m(\boldsymbol{b}|\boldsymbol{a}) + w^m(\sim \boldsymbol{b}|\boldsymbol{a}) \\ w^m(\boldsymbol{c}\cdot\boldsymbol{b}|\boldsymbol{a}) &= w^m(\boldsymbol{b}|\boldsymbol{a})\, w^m(\boldsymbol{c}|\boldsymbol{b}\cdot\boldsymbol{a}) \end{aligned}$$

With $P(x) = w^m(x)$ we obtain the rules of **probability theory**

$$\begin{aligned} 1 &= P(\boldsymbol{b}|\boldsymbol{a}) + P(\sim \boldsymbol{b}|\boldsymbol{a}) && \sim \text{negation} \\ P(\boldsymbol{c}\cdot\boldsymbol{b}|\boldsymbol{a}) &= P(\boldsymbol{b}|\boldsymbol{a})\, P(\boldsymbol{c}|\boldsymbol{b}\cdot\boldsymbol{a}) && \sim \text{conjunction} \end{aligned}$$

Theories in accordance with the assumptions are **isomorphic** to probability theory.

Bayes theorem

Data, model, and Bayes' theorem

$$P(\boldsymbol{a}|\boldsymbol{bc}) = \frac{P(\boldsymbol{a}|\boldsymbol{c})\,P(\boldsymbol{b}|\boldsymbol{ac})}{P(\boldsymbol{b}|\boldsymbol{c})}$$

Common situation

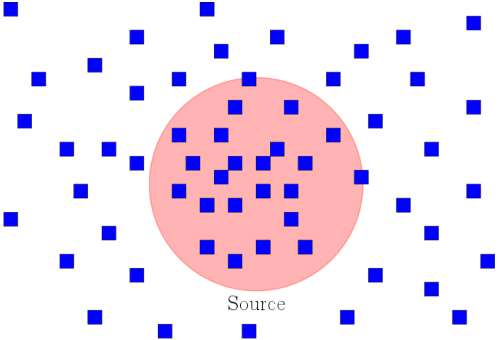
- Data D
- Model $f(\vec{\theta})$ depending on parameters $\vec{\theta} = (\theta_1, \theta_2, \ldots)$
- Other available information, I

$$P(\vec{\theta}|D, f\,I) = \frac{P(\vec{\theta}|f\,I)\,P(D|\vec{\theta}, f\,I)}{P(D|f\,I)}$$

Prior, likelihood, and posterior (inverse probability)

Setting up a problem

Source region, known position, Poisson process ($\mathcal{P}$)
Known BG count rate: λ_b, but **unknown** source count rate λ_s



n_s counts in source region. What about λ_s?

Parameter estimation

Use Bayes' theorem $I_{\mathcal{P}} = \{\mathcal{P}, \lambda_b, \text{location}, \ldots\}$:

$$P(\lambda_s|n_s, I_{\mathcal{P}}) = \frac{P(\lambda_s|I_{\mathcal{P}})P(n_s|\lambda_s, I_{\mathcal{P}})}{P(n_s|I_{\mathcal{P}})}$$

The likelihood

$$P(n_s|\lambda_s, I_{\mathcal{P}}) = \sum_{i=0}^{n_s} \mathcal{P}(i|\lambda_s)\mathcal{P}(n_s - i|\lambda_b)$$

What about the prior?

$$P(\lambda_s|I_{\mathcal{P}}) = \mathcal{C}/\lambda_s \quad \text{with} \quad \mathcal{C} > 0$$

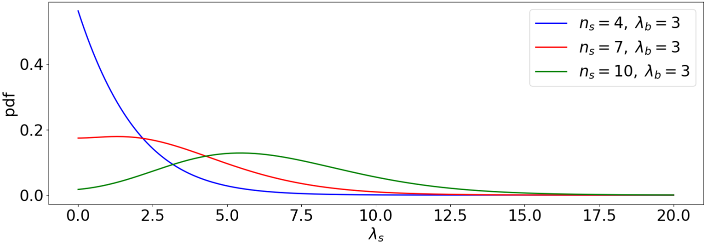
Improper! Defined up to a constant (typical for ignorance prior)

Parameter estimation

The normalization

$$P(n_s|I_{\mathcal{P}}) = \int_0^\infty P(n_s, \lambda_s|I_{\mathcal{P}})d\lambda_s = \int_0^\infty P(\lambda_s|I_{\mathcal{P}})P(n_s|\lambda_s, I_{\mathcal{P}})d\lambda_s$$

$$P(\lambda_s|n_s, I_{\mathcal{P}}) = \frac{\mathcal{C}/\lambda_s \sum_{i=0}^{n_s} \mathcal{P}(i|\lambda_s)\mathcal{P}(n_s - i|\lambda_b)}{\int_0^\infty \mathcal{C}/\lambda_s \sum_{i=0}^{n_s} \mathcal{P}(i|\lambda_s)\mathcal{P}(n_s - i|\lambda_b)d\lambda_s}$$



Hypothesis testing

$\mathrm{H}_0 : \lambda_\mathrm{s} \leq \lambda_0$ and $\mathrm{H}_1 : \lambda_\mathrm{s} > \lambda_0$

Calculate probability **for** (not against) the hypotheses:

$$P(H_0|n_s, I_\mathcal{P}) = \frac{P(H_0|I_\mathcal{P})P(n_s|H_0, I_\mathcal{P})}{P(n_s|I_\mathcal{P})}$$

$$P(H_1|n_s, I_\mathcal{P}) = \frac{P(H_1|I_\mathcal{P})P(n_s|H_1, I_\mathcal{P})}{P(n_s|I_\mathcal{P})}$$

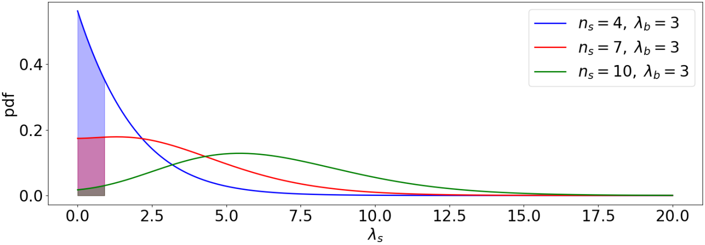
$$\frac{P(H_0|n_s, I_\mathcal{P})}{P(H_1|n_s, I_\mathcal{P})} = \frac{P(H_0|I_\mathcal{P})}{P(H_1|I_\mathcal{P})} \times \frac{P(n_s|H_0, I_\mathcal{P})}{P(n_s|H_1, I_\mathcal{P})}$$

Posterior odds = Prior odds × Bayes factor

Hypothesis testing

$\mathrm{H}_0 : \lambda_s \leq \lambda_0$ and $\mathrm{H}_1 : \lambda_s > \lambda_0$

Assume: $\lambda_0 = 1$ and prior odds $= 1/2 : 1/2$



$$\frac{P(H_0|n_s, I_{\mathcal{P}})}{P(H_1|n_s, I_{\mathcal{P}})} = 0.69(n_s = 4)\ , \quad 0.19(n_s = 7)\ , \quad 0.02(n_s = 10)$$

But, is there evidence for $\lambda_s > 0$ at all?

Point hypotheses testing

$\mathrm{H}_0 : \lambda_\mathrm{s} = 0$ and $\mathrm{H}_1 : \lambda_\mathrm{s} > 0$

$$\lim_{\lambda_0 \to 0} \frac{P(H_0|n_s, I_\mathcal{P})}{P(H_1|n_s, I_\mathcal{P})} = 0 \quad ???$$

On $I_\mathcal{P}$, the probability is zero.

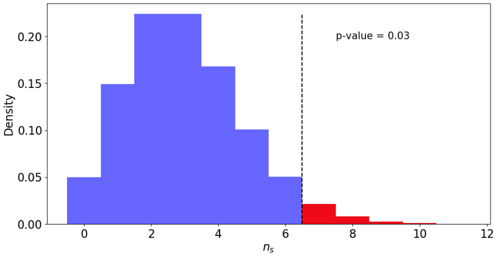
What about a **classical test of significance?**

A classical test of significance

$\mathrm{H}_0 : \lambda_\mathrm{s} = 0$ (to be nullified)

Test statistic (T): Number of photons in source region.

Determine p(robability)-value: $p = P(T \geq n_s | H_0, \lambda_b = 3)$



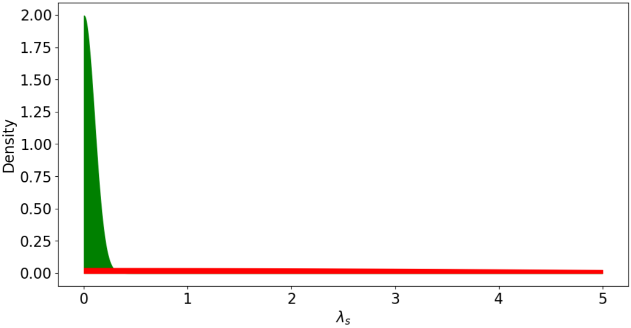
Reject H_0 if p is sufficiently small (e.g., 0.05)
but $P(D|H_0) \neq P(H_0|D)$

A Bayesian point hypotheses test

Introduce **new, sharply peaked prior:**

π_0 on $\lambda_s = 0$ and $(1 - \pi_0)$ distributed over $\lambda_s > 0$

$\rightarrow$ Two models (with and without λ_s)



Sketch of the prior

Point hypotheses testing

Calculate probability of H_0:

$$P(H_0|n_s, I_\pi) = \frac{P(H_0|I_\pi)P(n_s|H_0, I_\pi)}{P(n_s|I_\pi)}$$

$$P(H_0|n_s, I_\pi) = \frac{\pi_0 \mathcal{P}(n_s|\lambda_b, I_\pi)}{\pi_0 \mathcal{P}(n_s|\lambda_b, I_\pi) + (1-\pi_0)\int P(\lambda_s|I_\pi) P(n_s|\lambda_s, I_\pi) d\lambda_s}$$

$P(\lambda_s|I_\pi) = \mathcal{C}/\lambda_s$?

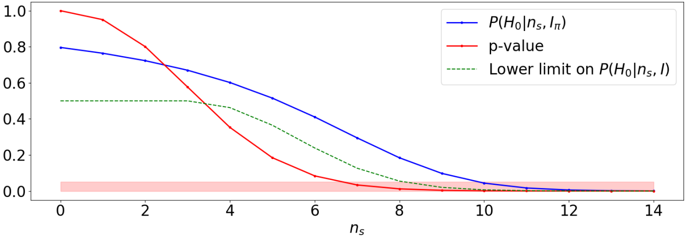
We need a proper (normalizable) prior

Point hypotheses testing

Jeffreys argues for a Cauchy distribution:

$$P(\lambda_s|I_\pi) = \frac{2}{\pi(\gamma - \lambda_s^2)}$$

How do we choose γ? I argue for $\gamma = \sqrt{\lambda_b}$ (scale of the problem)



p-value vs. probability of H_0

at $n_s = 7$: $p = 0.03$ but $P(H_0|n_s, I_\pi) = 0.29$ **(!)**

Summary

- Cox's theorem
- Parameter estimation
- Hypothesis testing
- null hypothesis testing