

CDMPP Computing Workshop

Feb. 2024

Frequentist and Bayesian inference



THE UNIVERSITY OF
WESTERN
AUSTRALIA



Australian
National
University



THE UNIVERSITY OF
MELBOURNE



THE UNIVERSITY OF
SYDNEY



THE UNIVERSITY
of ADELAIDE

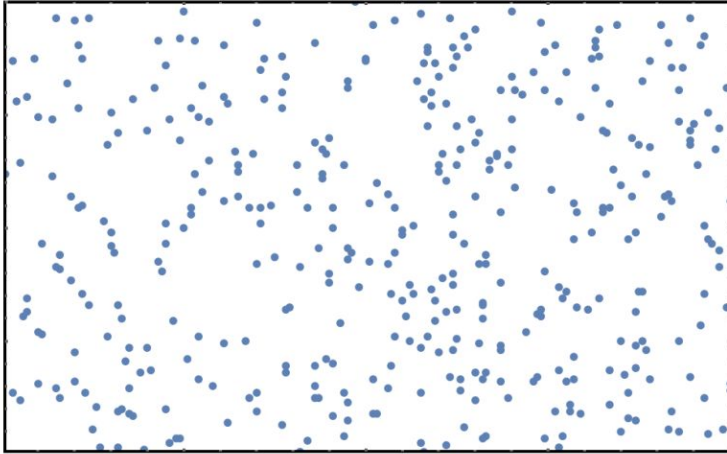
SWIN
BUR
NE

SWINBURNE
UNIVERSITY OF
TECHNOLOGY

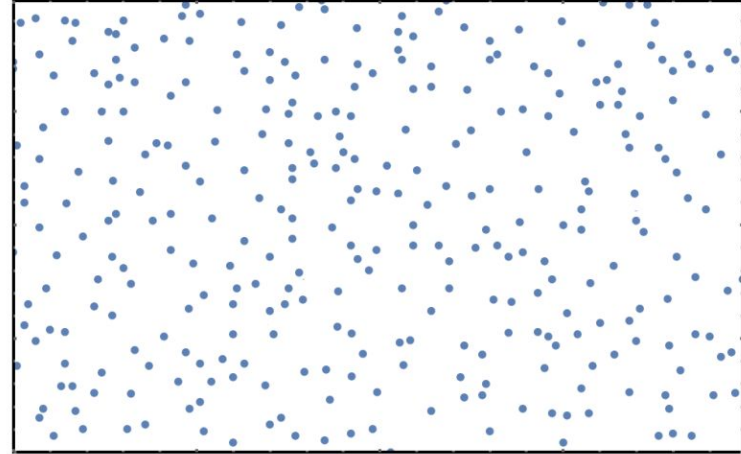
Why we need stats



A

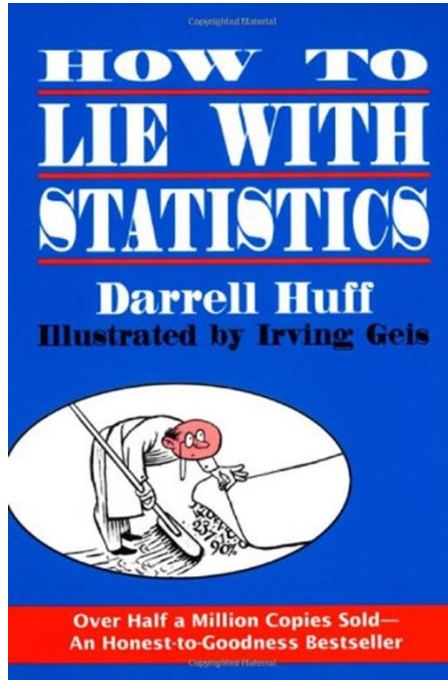


B



We are *too good* at pattern recognition

Applying statistics



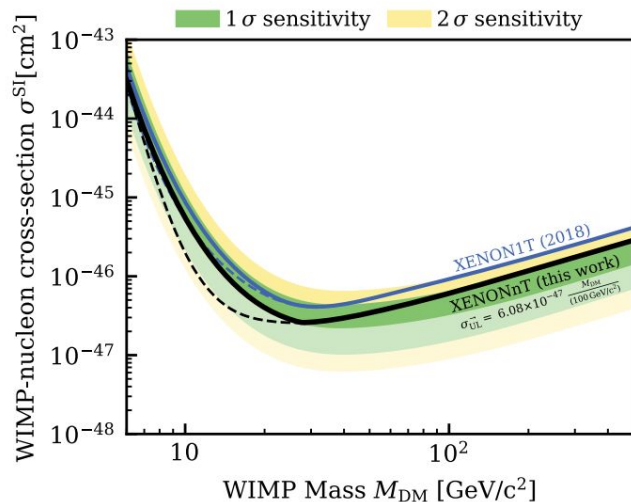
Even well meaning physicists run afoul

- OPERA ($>5\sigma$ → physical mistake)
- BICEP2 ($>5\sigma$ → wrong background model)
- 750 GeV diphoton excess ($>3\sigma$ - in Atlas and CMS → statistical fluctuation)
- Galactic centre excess (pulsars?)
- DAMA (12σ ??)

Common applications

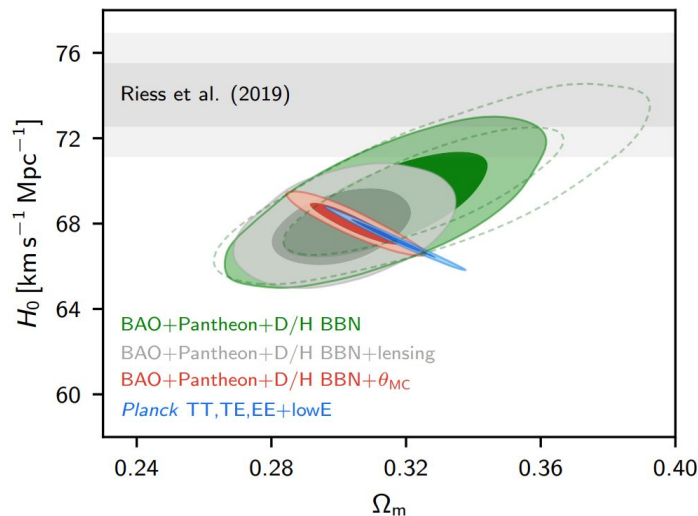
When you see these, what does it mean:

Upper limits:



XENONnT [arXiv:2303.14729](https://arxiv.org/abs/2303.14729)

Parameter estimation:



Planck [arXiv:1807.06209](https://arxiv.org/abs/1807.06209)

Review: what is probability?

Imagine set of *exclusive* events, which are each a potential random outcome:

E_1, E_2, \dots

Kolmogorov axioms:

1. $p(E_i) \geq 0$
2. $p(E_i \text{ or } E_j) = p(E_i) + p(E_j)$
3. $\sum_i p(E_i) = 1$

Foundations of the Theory of Probability, Kolmogorov, 1933

Any definition of probability must satisfy these (but the def. is not unique)

Bayes' theorem from conditional probability

From the Kolmogorov axioms it follows that:

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

$$p(B|A) = p(A|B)p(B)/p(A) \quad \longleftarrow \text{Bayes' theorem}$$

H = hypothesis, D = some data, I = prior information

$$P(H|D, I) = \frac{P(H \cap D, I)}{P(D|I)} \quad P(D|H, I) = \frac{P(D \cap H, I)}{P(H|I)}$$

$$\text{but: } P(H \cap D, I) = P(D \cap H, I)$$

$$P(H|D, I) P(D|I) = P(D|H, I) P(H|I)$$

$$P(H|D, I) = \frac{P(D|H, I)P(H|I)}{P(D|I)}$$

Bayes' theorem: quick example

You take test for a rare and terrible disease and it comes back positive, what was the probability you have the disease?

Given:

- False negative rate of 0%
- False positive rate of 1%
- Community prevalence of 0.1%

$$p(B|A) = p(A|B)p(B)/p(A)$$

$$p(\text{disease}|\text{positive}) = p(\text{positive}|\text{disease}) p(\text{disease}) / p(\text{positive})$$

$$p(\text{positive}) = p(\text{false positive}) + p(\text{true positive})$$

$$p(\text{disease}|\text{positive}) = 1 * 0.001 / (1*0.001 + 0.01*.999) = 0.09$$

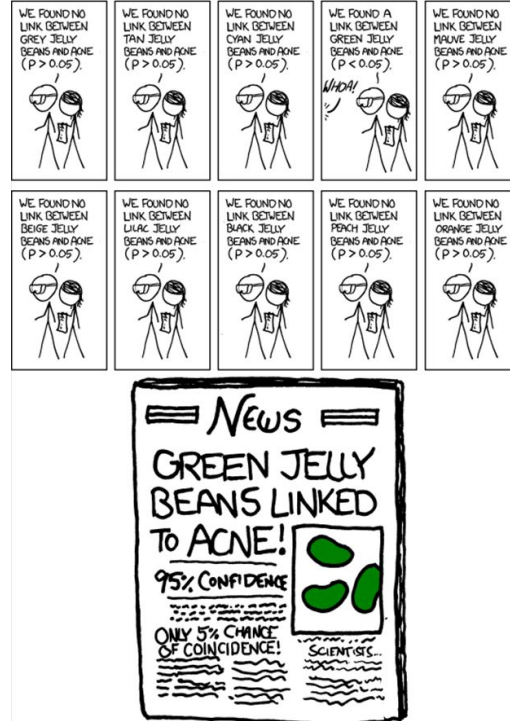
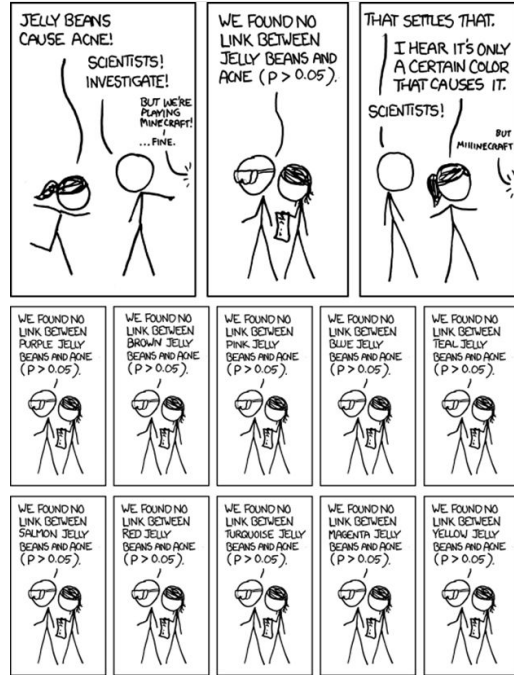
Frequentist vs. Bayesian philosophies

	Frequentist	Bayesian
Interpretation of probability	Frequency	Odds
Inference	P-values, coverage	Credence, degree of belief

- Much hay has been made of these differences, in short: they can answer different questions
- Different problems may have a more suitable approach



P-values as a criteria for discovery

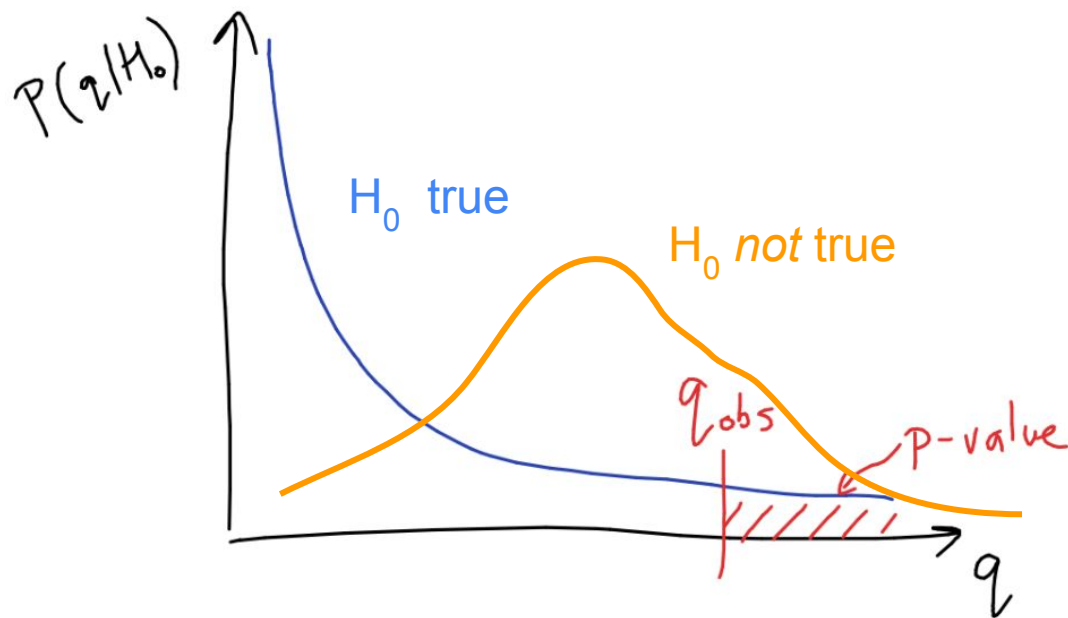


(barely) not statistically significant ($p=0.052$)
a barely detectable statistically significant difference ($p=0.073$)
a borderline significant trend ($p=0.09$)
a certain trend toward significance ($p=0.08$)
a clear tendency to significance ($p=0.052$)
a clear trend ($p<0.09$)
a clear, strong trend ($p=0.09$)
a considerable trend toward significance ($p=0.069$)
a decreasing trend ($p=0.09$)
a definite trend ($p=0.08$)
a distinct trend toward significance ($p=0.07$)
a favorable trend ($p=0.09$)
a favourable statistical trend ($p=0.09$)
a little significant ($p<0.1$)
a margin at the edge of significance ($p=0.0608$)
a marginal trend ($p=0.09$)
a marginal trend toward significance ($p=0.052$)
a marked trend ($p=0.07$)
a mild trend ($p<0.09$)
a moderate trend toward significance ($p=0.068$)

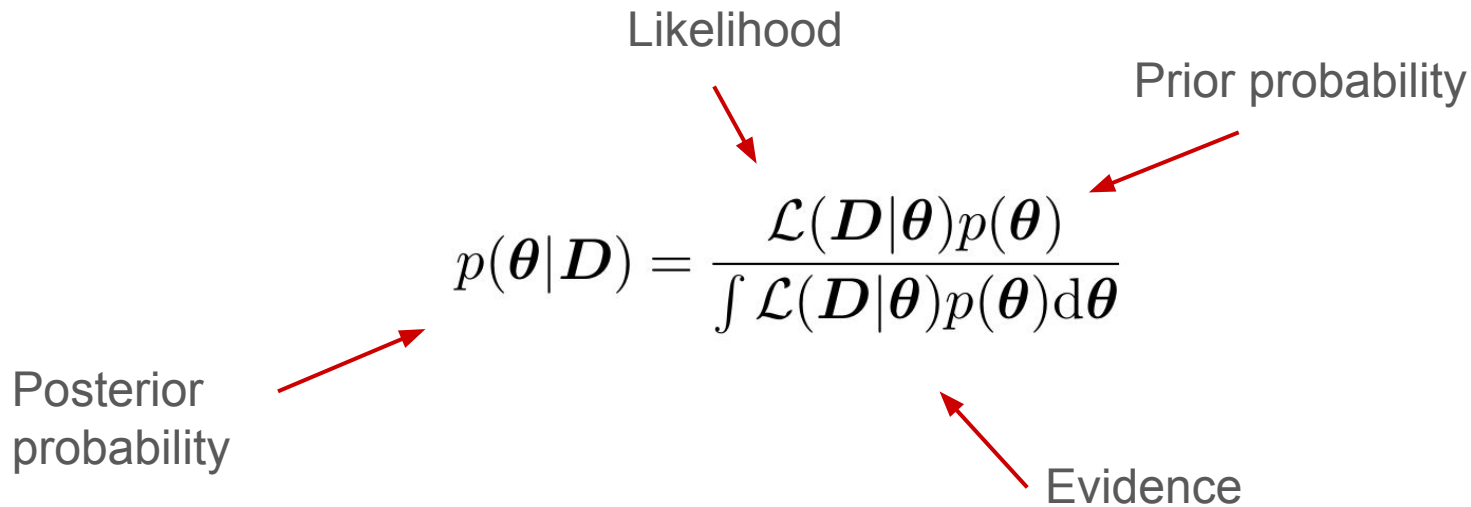
From [blog by Matthew Hankins](#)

P-values as a criteria for discovery

Define a test statistic, q , which has very different distribution under competing hypotheses



Revisiting Bayes' theorem



A diagram illustrating Bayes' theorem. The central equation is $p(\theta|D) = \frac{\mathcal{L}(D|\theta)p(\theta)}{\int \mathcal{L}(D|\theta)p(\theta)d\theta}$. Four red arrows point from descriptive labels to parts of the equation: 'Likelihood' points to the numerator's first term, 'Prior probability' points to the numerator's second term, 'Evidence' points to the denominator, and 'Posterior probability' points to the left side of the equation.

Likelihood

Prior probability

$$p(\theta|D) = \frac{\mathcal{L}(D|\theta)p(\theta)}{\int \mathcal{L}(D|\theta)p(\theta)d\theta}$$

Posterior probability

Evidence

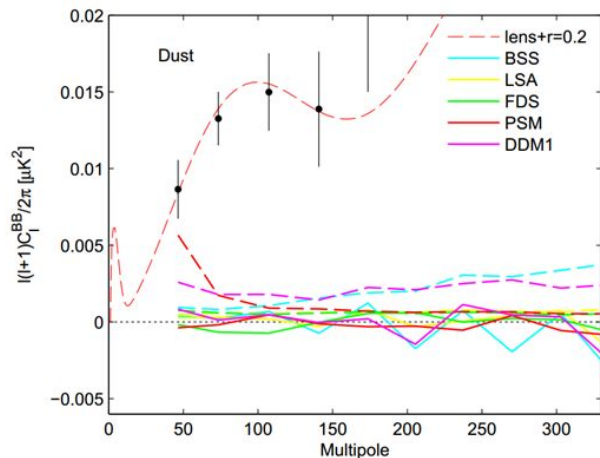
Likelihoods

$$\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}) = \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$$

- The probability that a set of parameters, $\boldsymbol{\theta}$, reproduces the data, \mathbf{D}
- For a fixed dataset \mathbf{D} , if we vary $\boldsymbol{\theta}$, the sum of all likelihood values for different choices of θ is not 1
- We get a collection of values that indicate how likely the data is under each possible θ .

Prior probability

- People find them unnatural, but they neglect the fact that they are using them
- “The probability that each of these models reflects reality is hard to assess” BICEP2 arXiv:1403.3985v1



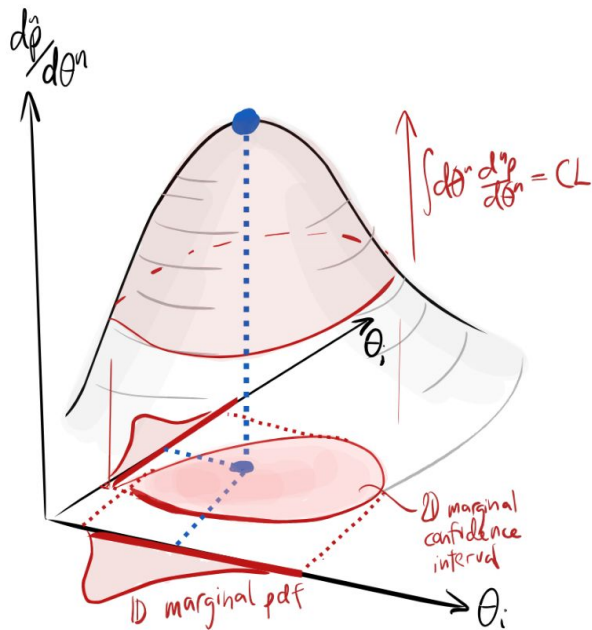
Evidence

- Fully marginalised posterior (integrated), normalises the probability
- Can be used for model comparison (wont do this here)

Presenting credible regions

$$p(\theta_1|\mathbf{D}) = \int p(\boldsymbol{\theta}|\mathbf{D})d\theta_2d\theta_3\dots d\theta_n$$

$$\int_{\theta_i \in \text{CrI}} p(\theta_i|\mathbf{D})d\theta_i = \text{CL}$$



How do we get the samples?

We could randomly or systematically (grid) scan over θ values and compute the posterior at each point - this gets slow with multi-dimensional data.

Standard practice is to use a sampler like MultiNest

Let's do some examples

Open the notebook..

National Partners



Australian
National
University



THE UNIVERSITY OF
SYDNEY



Australian Government
Department of Defence



Australian Government



International Partners



UvA



Stockholm
University



The
University
Of
Sheffield.

