Herding cats? — Bayesian and frequentist contradictions and compromises

Andrew Fowlie

14 March 2022

Nanjing Normal University

Overview

1. Frequentist methods
2. Bayesian
3. Paradoxes and compromises

Testing and estimation

Roughly speaking, statistical tasks separate into

1. Model testing or comparison
2. Estimating or inferring the model's parameters

I will focus on first. In my opinion, first we should establish whether a phenomena exists, and then infer its parameters or properties.

Testing

Jeffreys and Fisher agree!

Jeffreys 1939

> *“ [I]n what circumstances do observations support a change of the form of the law itself? This question is really logically prior to the estimation of the parameters, since the estimation problem presupposes that the parameters are relevant ”*

Fisher 1925

> *“ It is a useful preliminary before making a statistical estimate . . . to test if there is anything to justify estimation at all ”*

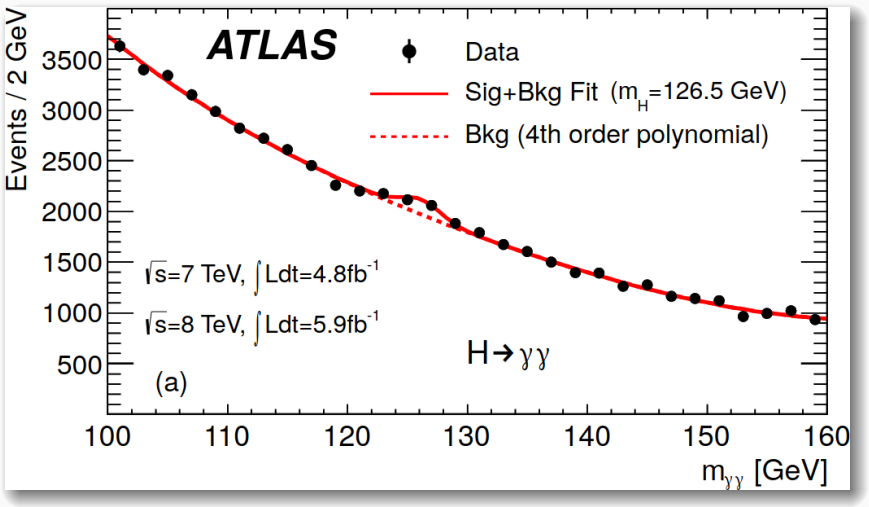
Discoveries!

Classic example. Higgs discovery in 2012.

How do we judge when the data indicates the presence of a new particle or phenomena?

Discoveries!

Di-photon spectrum contains a resonance (Aad et al. 2012).



Discovery was announced based on a particular choice of statistical methodology.

Methodology

We need a statistical methodology to judge evidence for a discovery. In the time available, let's consider

1. Frequentist; see e.g., Lyons 1989; Cowan 1998; James 2006; Behnke et al. 2013. Two schools
   - Error control
   - Evidential
2. Bayesian; see e.g., D'Agostini 2003; Gregory 2005; Sivia and Skilling 2006; Trotta 2008; Linden, Dose, and Toussaint 2014; Bailer-Jones 2017

Likelihood

Methods typically require at least the likelihood (see e.g., Cousins 2020)

$$\mathcal{L}(\boldsymbol{\Theta}) = p(D \mid M, \boldsymbol{\Theta})$$

This tells us the probability (density) of the observed data, D, given a particular model, M, and choice of parameters.

This is a function of the model's parameters, $\boldsymbol{\Theta}$, for fixed, observed data.

Frequentist methods

P-values

_P_-value (Wasserstein and Lazar 2016)

The p-value, p, is the probability of observing data as or more extreme than that observed, given the null hypothesis, H_0, i.e.,

$$p = P(\lambda \geq \lambda_{\text{Observed}} \mid H_0)$$

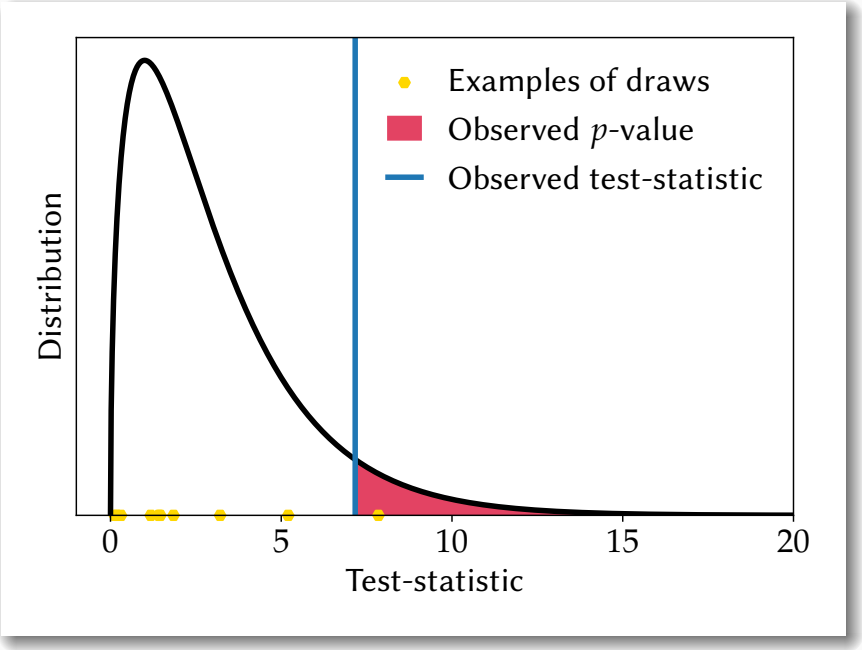
where λ is a test-statistic that summarises the data and defines extremeness, and H_0 specifies the distribution of λ

See Demortier 2008 for discussion about composite null hypotheses that don't uniquely specify the distribution of λ.

Test-statistic often based on (profiled) likelihood ratio (Neyman and Pearson 1933)

P-values

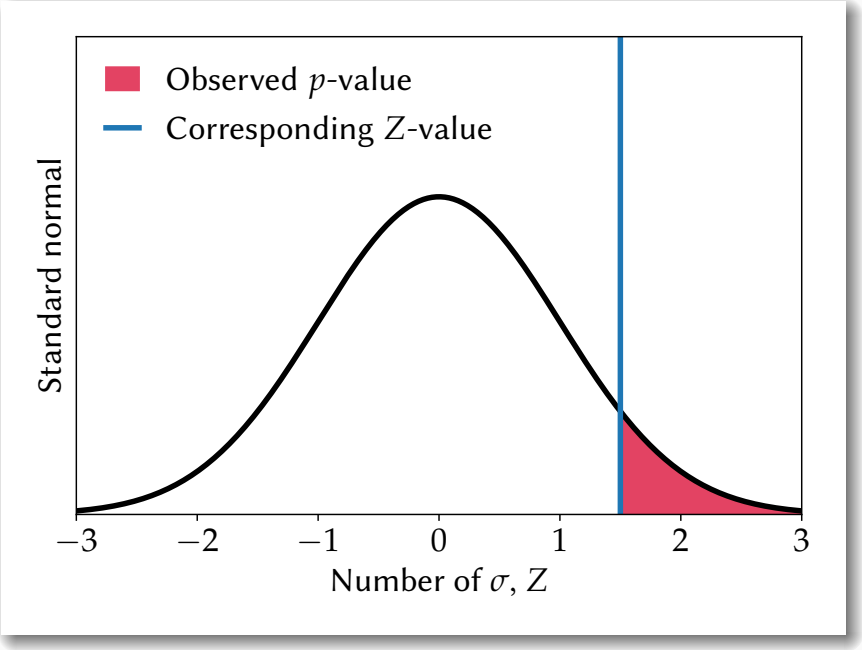
Thus p is a tail probability.



Thus p is uniformly distributed under H_0 (or dominated by uniform in discrete settings or composite null)

Z-values

In particle physics, it's common to translate p-values into Z-values. 5σ corresponds to about $p = 10^{-7}$. This is just a convention



through the equation

$$Z = \Phi^{-1}(1 - p)$$

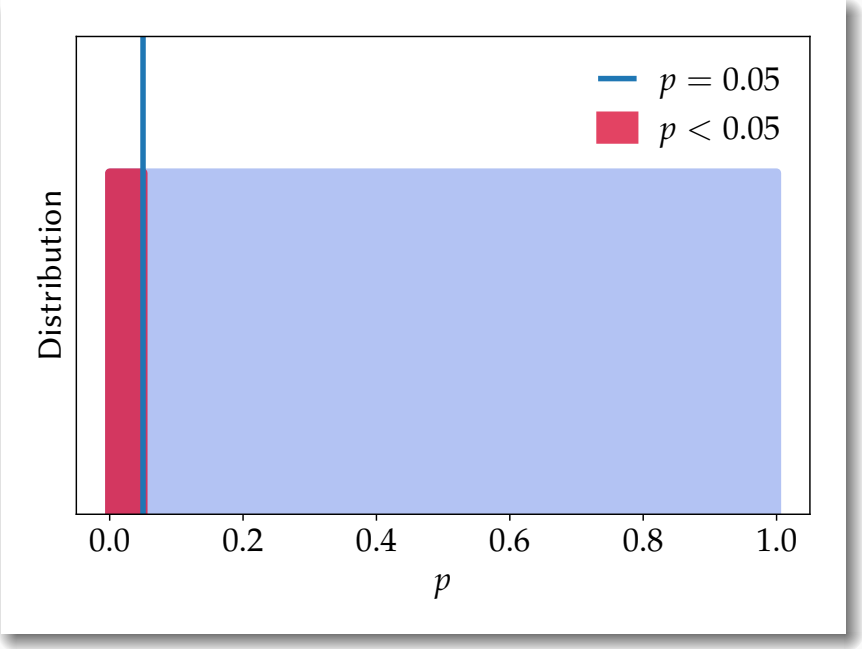
Interpreting *p*-values

P-values are popular in particle physics and elsewhere. Two possibly contradictory interpretations (Hubbard and Bayarri 2003):

- *P* is a measure of evidence against H_0 (Fisher 1925): small $p \Rightarrow$ H_0 implausible. See e.g., Hubbard and Lindsay 2008; Schervish 1996; Berger and Sellke 1987; Senn 2001; Murtaugh 2014
- *P* is a means to control error rate (Neyman and Pearson 1933): if we reject null when *p*-value ≤ 0.05, for example, becomes error theoretic approach with type 1 error rate $\alpha = 0.05$

Controlling type-1 error rate

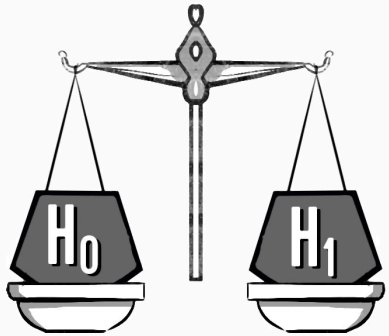
The p-value enables us to control type-1 error rate because it is uniformly distributed under the null



Placing a threshold $p < \alpha$ controls the type-one error rate to be α

Example from high-energy physics

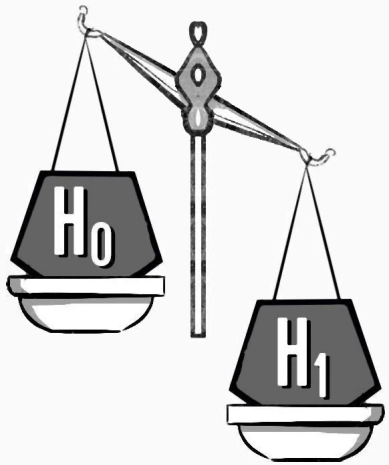
Original artwork Viktor Beekman and concepts Eric-Jan Wagenmakers



In high-energy physics, we want to discover new phenomena and new particles. Perform null hypothesis test:

- H_0 — Standard Model (SM) backgrounds only
- H_1 — SM + new physics, e.g. Higgs boson or supersymmetric particles

Example from high-energy physics

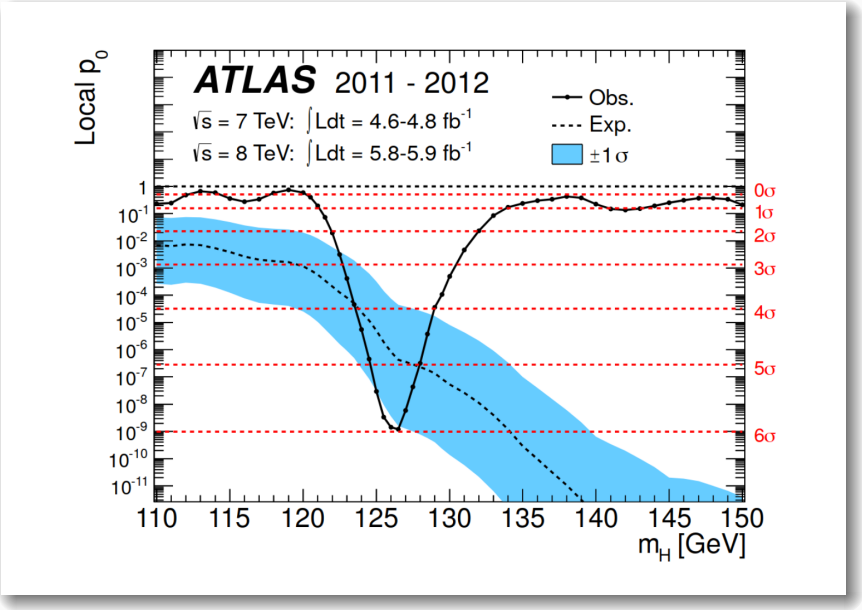


For a discovery we conventionally require a tiny global p-value of

$$p \lesssim 10^{-7}\,(5\sigma)$$

i.e., $\alpha \simeq 10^{-7}$ (Lyons 2013). Dual interpretation: threshold in evidence — extraordinary claims require extraordinary evidence — and imposes a 10^{-7} type-1 error rate.

Example from high-energy physics



Discovery of Higgs boson announced by ATLAS and CMS once significance greater than 5σ.

Some scares — e.g., 2015 diphoton excess (Strumia 2016) — but so far 5σ criterion prevented false discoveries (though think about flavor anomalies).

Misconceptions

Misconceptions galore by public and scientists, see e.g., Goodman 2008; Greenland et al. 2016

1. P is not probability of null hypothesis
2. P is not the probability that the data were produced by chance alone
3. P is not an error rate

Wagenmakers and al 2017

> *“ The fact that academics don’t know what p means is a symptom of the fact that p doesn’t tell anything worth knowing ”*

Though see Murtaugh 2014; Lakens et al. 2018; Cousins 2018; Lakens 2021; Mayo 2018.

Bayesian

Bayes factors

Forget long-run errors rates and data we don't have. Compute the change in plausibility of models in light of the data we have

- With this and priors for the models, we could compute the posterior plausibility of each model
- If you like, you can compute the probability that you are making an error in the case at hand (cf. long-run error rates that are independent of the observed data)
- We just apply probability theory to the problem (Jeffreys 1939). Simple in theory; in practice there are difficulties.

Bayes factors

The Bayes factor (Kass and Raftery 1995) relates the relative plausibility of two models after data to their relative plausibility before data;

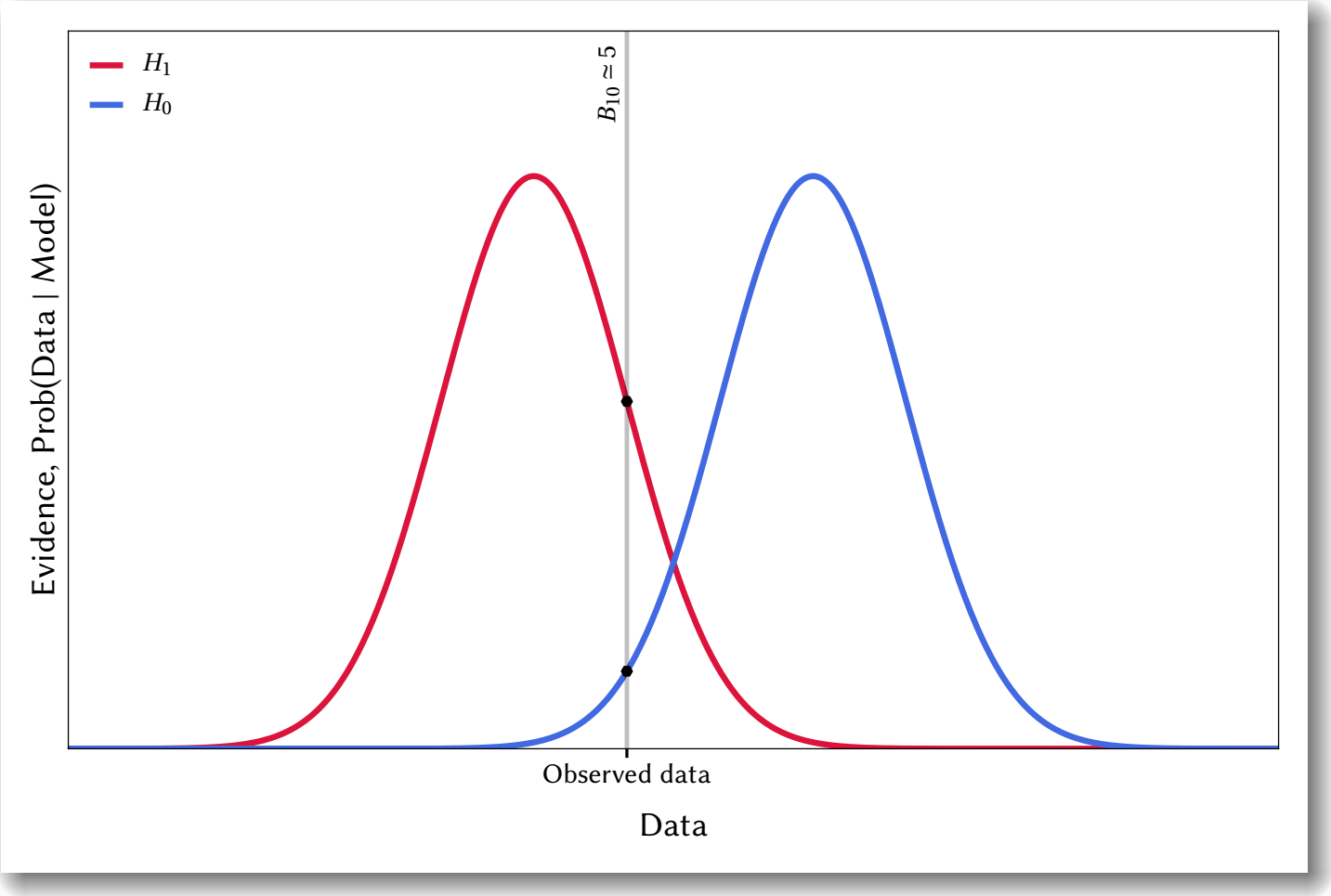
$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}$$

where

$$\text{Bayes factor} = \frac{p(\text{Observed data} \mid \text{Model } a)}{p(\text{Observed data} \mid \text{Model } b)}$$

A nice result — by applying laws of probability, we see that models should be compared by nothing other than their ability to predict the observed data.

Bayes factors



Bayesian evidence

The factors in the ratio are Bayesian evidences

$$\mathcal{Z} \equiv p(D \,|\, M) = \int_{\Omega_{\boldsymbol{\Theta}}} \mathcal{L}(\boldsymbol{\Theta})\, \pi(\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta},$$

where D is the observed data, $\mathcal{L}(\boldsymbol{\Theta}) = p(D \,|\, \boldsymbol{\Theta}, M)$ is the likelihood and $\pi(\boldsymbol{\Theta}) = P(\boldsymbol{\Theta} \,|\, M)$ is our prior, and $\boldsymbol{\Theta}$ are the model's parameters.

The prior describes what we knew about the parameters before seeing the data

The evidence is the likelihood averaged over the prior — the averaging penalises fine-tuned models

Sensitivity to priors

Evidences are the likelihoods averaged over priors.

Many consider the resulting dependence of the Bayes factor on the priors to be a major and perhaps fatal problem; see e.g., Berger and Pericchi 2001; Cousins 2008

- No priors, no predictions. I need to compare your model's predictions with data. If you don't tell the plausible parameters, how am I to know what it predicts?
- Sensitive to arbitrary choices. If the inference changes dramatically within a class of reasonable priors, we can't draw reliable conclusions.

Sensitivity to priors

Evidences are the likelihoods averaged over priors.

Many consider the resulting dependence of the Bayes factor on the priors to be a major and perhaps fatal problem; see e.g., Berger and Pericchi 2001; Cousins 2008

Paraphrasing Hill 1975

“ the lack of a concrete theory for choosing priors no more implies that one should not use Bayesian statistics than does the lack of a theory that tells us the right price to pay for groceries implies we should not use money ”

Subjective & Objective

There are different approaches to constructing priors, leading to different flavors of Bayesian inference

Subjective

Priors reflect state of knowledge and could be constructed by e.g., consulting experts (see e.g., Goldstein 2006; Mikkola et al. 2021)

Dictated by state of knowledge

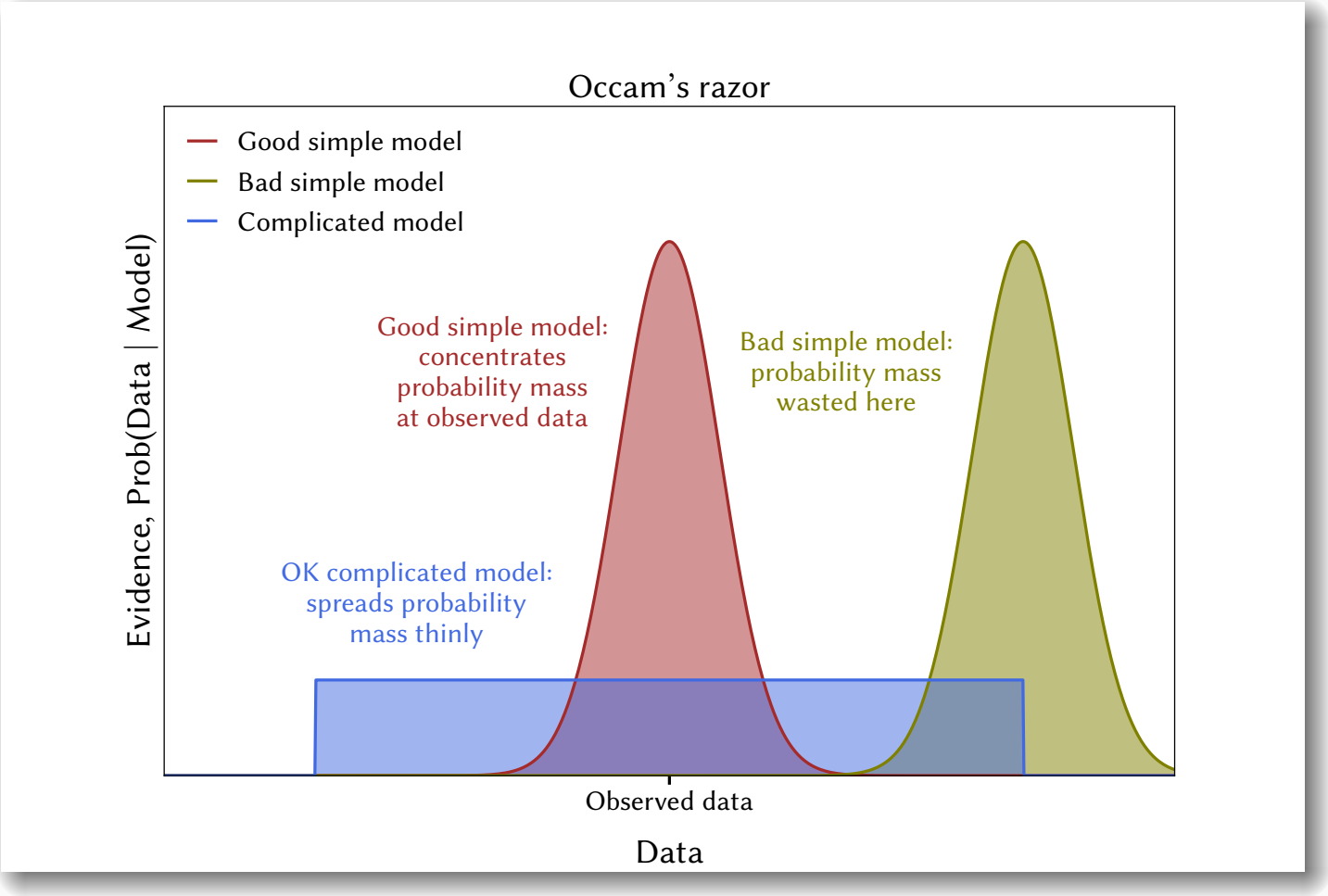
Priors could be dictated by e.g., a symmetry (Jaynes 1968)

Formal rules for selecting priors

Construct priors that e.g., maximise what we expect to learn about a model's parameters (Kass and Wasserman 1996; Consonni et al. 2018)

Occam's razor

Evidence automatic Occam razor (MacKay 1992; Jefferys and Berger 1992)



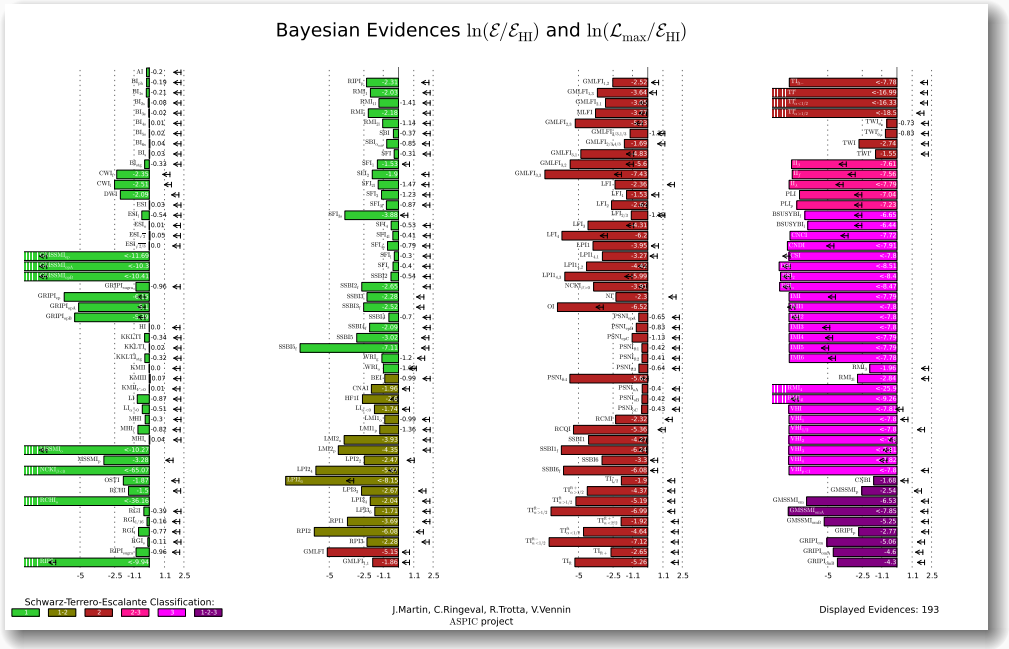
Example from cosmology

Best models of inflation after Planck 2013 (Martin et al. 2014; Martin, Ringeval, and Vennin 2014)

- Collate about 200 models of slow-roll inflation
- Formulate suitable priors
- Formulate likelihoods for Planck 2013 data
- Compute Bayesian evidences through a Monte Carlo integration method

What's left?

About 30 models were disfavoured and about 70 strongly so.



No stand-out preferred model.

Paradoxes and compromises

Likelihood principle

Originated by considering stopping rules (Barnard 1949). Proven by Birnbaum 1962

Berger and Wolpert 1988

“ all evidence, which is obtained from an experiment, about an unknown quantity θ, is contained in the likelihood function of θ for the given data $[\mathcal{L}(\theta)]$ ”

- Forbids evidential interpretations of frequentist statistics and *p*-values — since they depend on considering data other than that observed
- Implicitly obeyed by Bayesian statistics (though violated by reference priors)

Intuition for likelihood principle

Pratt 1962

“ An engineer draws a random sample of electron tubes and measures the plate voltage under certain conditions with a very accurate volt-meter …

A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4 … ”

Intuition for likelihood principle

Pratt 1962

“ Later he visits the engineer’s laboratory, and notices that the volt meter used reads only as far as 100, so the population appears to be “censored.” This necessitates a new analysis, if the statistician is orthodox. ”

Intuition for likelihood principle

Pratt 1962

“However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all.”

Intuition for likelihood principle

Pratt 1962

“ But the next day the engineer telephones and says: “I just discovered my high-range volt-meter was not working the day I did the experiment you analyzed for me.” The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. ”

Intuition for likelihood principle

Pratt 1962

> *“ . . . “Next you’ll be asking me about my oscilloscope.” ”*

Status of likelihood principle

At the time, considered profound by some though not universally accepted

Savage 1962

> *“ Without any intent to speak with exaggeration or rhetorically, it seems to me that this is really a historic occasion. ”*

Over time limited practical impact: ignored by Bayesian because it’s automatically satisfied; ignored by frequentists because it’s automatically violated.

Posterior of null versus p

Of course,

$$\text{Posterior of null} \neq p\text{-value}$$

However, well-known that typically

$$\text{Posterior of null} \gg p\text{-value}$$

for broad classes of priors. P typically overstates the evidence against the null

Posterior of null versus p

Bounds

Famous bound (Vovk 1993; Sellke, Bayarri, and Berger 2001) that under mild assumptions

$$B_{10} \leq \frac{1}{-ep \ln p}$$

With equal priors for null and alternative, $p = 0.05$ corresponds to at least about 30% posterior probability of the null (see e.g., Berger and Sellke 1987; Berger and Delampady 1987; Benjamin et al. 2018)

Examples from high-energy physics

I see this in high-energy physics. B for some Z (read papers for discussion of priors etc)

- Higgs discovery — posterior of null about 100 times greater than p (Fowlie 2019)
- ATLAS 2015 diphoton — $Z = 2.1\sigma$ and $B \simeq 7$ (Fowlie 2017)
- DAMPE — $Z = 2.3\sigma$ and $B \simeq 2$ (Fowlie 2018)
- 2020 XENON — $Z = 3.5\sigma$ and $B \simeq 3$ (Athron et al. 2021)

Jeffreys-Lindley paradox

- Folk theorem that differences between Bayesian and frequentist methods vanish once sufficient data collected
- Jeffreys-Lindley paradox (Jeffreys 1939; Lindley 1957) destroyed that misconception

See e.g., Robert 2014; Cousins 2017; Wagenmakers and Ly 2021 for reviews.

Jeffreys-Lindley paradox

- Take n measurements of an unknown mean, θ, and test whether $\theta = 0$
- The p-value depends on the t-statistic

$$t = \frac{\sqrt{n}\bar{x}}{\sigma} \quad \text{through} \quad p = 2(1 - \Phi(t))$$

- The Bayes factor in favour of $\theta = 0$, though, depends explicitly on n and t,

$$B \approx e^{-\frac{1}{2}t^2} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}$$

Thus for fixed (and e.g., tiny) p, B may favour null by arbitrary factor by sufficiently large n!

Blame

- No blame — Bayesian and frequentist testing answer different questions. Nevertheless, paradoxes help us understand foundations e.g., ladder paradox, Maxwell's demon, and Schrodinger's cat
- Blame p — evidence should depend on n, p-value contraindicated
- Blame fixed significance level — α should be a function of n
- Blame B — cannot be right to accept null even when p arbitrarily tiny
- Blame null — testing point null hypothesis (i.e., testing $\theta = 0$) problematic
- Blame setting — fixing t but increasing n unrealistic (Fowlie 2020)

Statistical cocktail

What if we could combine

Evidential aspect of Bayes + Error theoretic aspect of frequentist

Would we obtain a wonderful resolution? or something that everyone hated? Two contentious tastes that taste contentious together?

Good's compromise

A compromise (Good 1957; Good 1961; Good 1992)

- Compute B as though you were a Bayesian
- Compute p as though you were a frequentist using B as a test-statistic!

$$p = P(B \geq B_{\text{Observed}} \mid H_0)$$

- Report B and p. B for evidence, p for error control

What if $B \ggg 1$ but p insignificant or vice-versa?

Good's compromise

Good 1992

“ [T]he pure Bayesian throws away the use of P-values, naive or otherwise. But because clients often want answers having the veneer of objectivity, the use of P-values is somewhat justifiable ”

Good 1992

“ I don't think epistemic probabilities have sharp values. When they are very vague, you might have to fall back either on P-values ”

Good 1992

“ [Results from significance test] correct in the long run in a certain proportion of cases, thus protecting the statistician's rear end to some extent, but the client's less so ”

Likelihood ratios and Bayes factors

Good 1992

> *“ the ratio of maximum likelihoods can be regarded as a (very poor) approximation to a Bayes factor . . . Thus Neyman and Pearson perhaps were unconscious Bayes/non-Bayes compromisers ”*

Indeed, the Neyman-Pearson lemma shows that Bayes factor most powerful test-statistic for simple hypotheses.

For non-simple ones, the Bayes factor most expected power (Zhang 2016; Fowlie 2021)

Frequentist properties of Bayes factors

Kerridge's theorem

Kerridge's simple theorem demonstrates that (Kerridge 1963)

$$P(B_{10} \geq t \mid H_0) \leq 1/t$$

That is, the chance that the Bayes factor favours H_1 by at least t when H_0 is true must be less than $1/t$.

This enables Bayes factors to be used in so-called universal inference (Wasserman, Ramdas, and Balakrishnan 2020). Bayes factors many convenient frequentist-style properties.

See also Berger 2003; Bayarri and Berger 2004; Bayarri et al. 2016

Turing's theorems and Gibbs' inequality

Turing via Good (see Good 1965; Good 1960)

Expectation of logarithm of Bayes factor indictes correct model.

$$\langle \log B_{10} \rangle_1 \geq 0$$
$$\langle \log B_{10} \rangle_0 \leq 0$$

Identical to Gibbs' inequality.

Intuitive — data on average must indicate correct model through logarithm of Bayes factor (weight of evidence) or at worst be irrelevant

See also Etz and Wagenmakers 2017 for some history of Turing's involvement here.

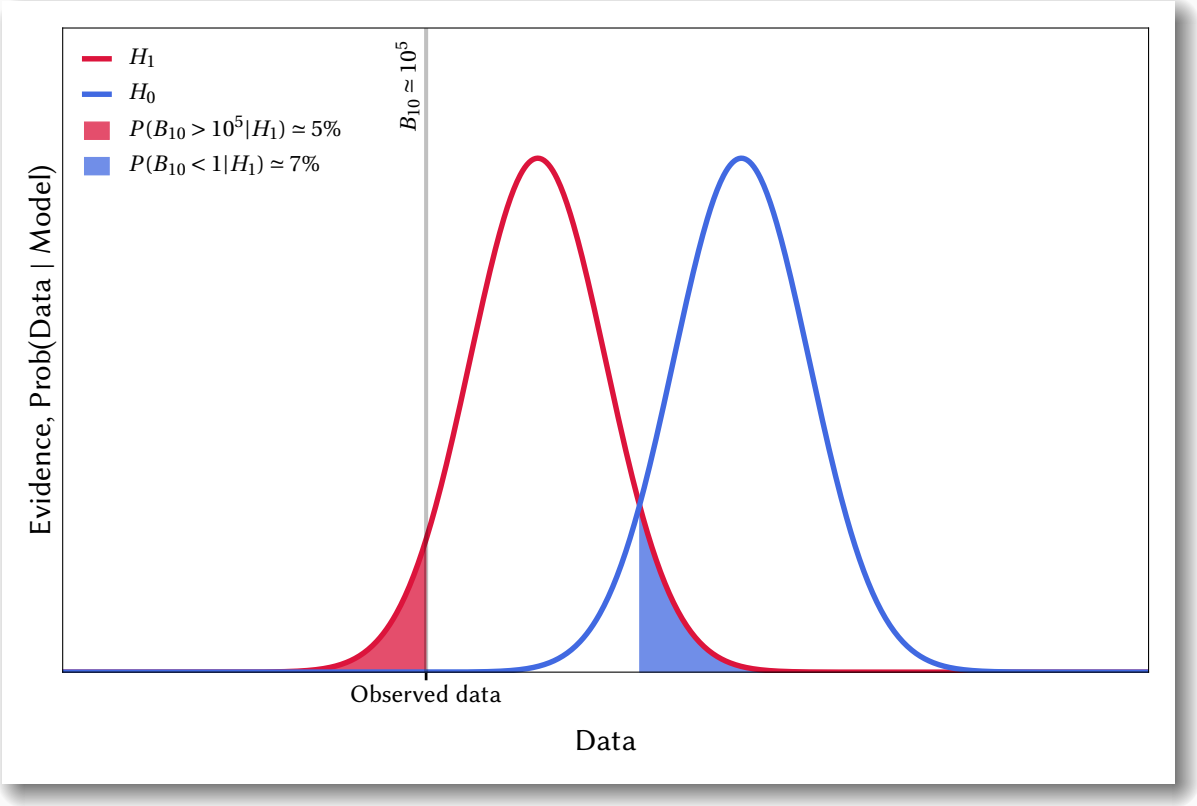
Use in cosmology

Curiously, this compromise recently used in cosmology in Joachimi et al. 2021

- Tension between measurements of the Hubble constant (Verde, Treu, and Riess 2019)
- Tension quantified using Bayesian evidences or statistics involving them
- What about their frequentist properties?

Bayes factors by chance

Chances of obtaining such a sizeable Bayes factor (10^5) under replication, assuming indicated model, could be small (5%)



Could even be appreciable chance of Bayes factor favouring a different model (7%). Does it matter?

Bayes factors by chance

Verde, Treu, and Riess 2019 consider chances of obtaining such a sizeable Bayes factor were the indicated model and some estimator of its parameters. Similar to

$$P(B \geq B_{\text{Obs.}} \mid D, H_1)$$

Considering this reduces tension between discrepant Hubble measurements. Further suggest reducing noise in evidence estimate by compressing data.

Computational challenges

Bayesian and frequentist techniques pose computational challenges

Bayesian — integration

The evidence is a challenging multi-dimensional integral

$$\mathcal{Z} \equiv p(D \,|\, M) = \int_{\Omega_{\boldsymbol{\Theta}}} \mathcal{L}(\boldsymbol{\Theta})\, \pi(\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta},$$

Usually impossible analytically.

Computational challenges

Bayesian and frequentist techniques pose computational challenges

_P_-values — compression

We need to compute p as small as about 10^{-7} through

$$p = P(\lambda \geq \lambda_{\text{Observed}} \mid H_0)$$

The tiny region of sampling space for which $\lambda \geq \lambda_{\text{Observed}}$

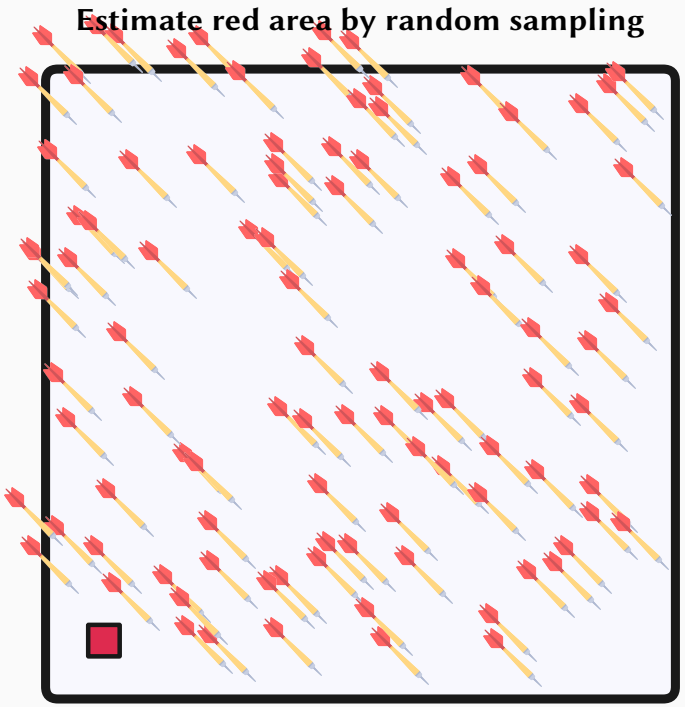
Common challenge

In fact, these are the same problem — compression

- Compression from the whole sampling space to the tiny region corresponding to p
- Compression from the whole prior to the region of significant likelihood that contributes to the integral

Random sampling

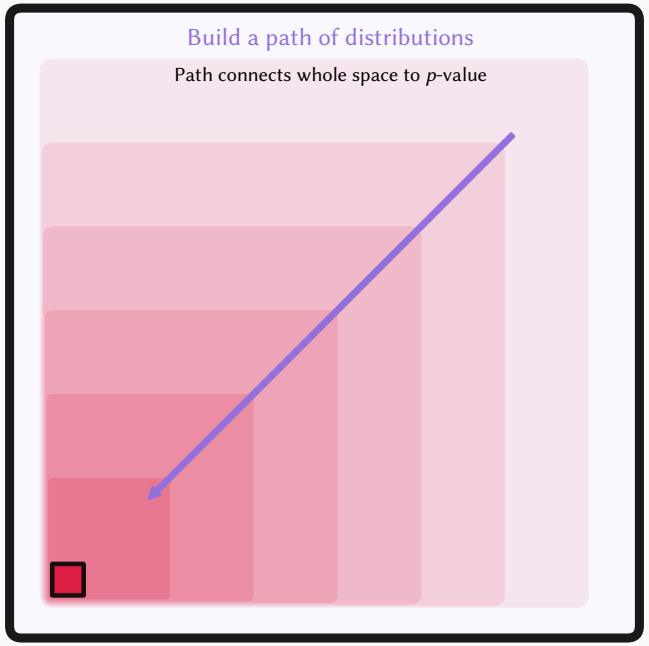
This won't be efficient



Nested sampling

Compress through a path of distributions

Solution — path sampling



Nested sampling

Nested sampling (Skilling 2006) estimates compression along path statistically. The gist:

- Draw n live points
- Keep half with greatest test-statistic
- We just compressed by 1/2!

This enables NS to estimate compression required for Bayesian computation and p-values (Fowlie, Hoof, and Handley 2022)

Summary

- Two motivations and interpretations of p-value: error control and evidence
- Latter lacks foundation
- Bayes factors quantify change in plausibility, but sensitive to priors
- Paradoxes about evidence implied by them and attempts at resolution
- Despite century of debate, matters remain unsettled!
- Nested sampling rises to universal computational challenge of compression

Backup

Estimation

Though see e.g., Cumming 2013 for disagreement.

Cumming 2013

> *“Suppose you read in the news that “support for Proposition X is 53%, in a poll with an error margin of 2%.” . . . more informative than stating that support is “statistically significantly greater than 50%, $p < .01$.”*

Nevertheless, here we focus on testing.

Letting the data speak for itself

Gould 1981

“ inanimate data can never speak for themselves, and we always bring to bear some conceptual framework, either intuitive and ill-formed, or tightly-formed and structured, to the task of investigation, analysis and interpretation ”

Tukey et al. 1977

“ No body of data tells us all we need to know about its own analysis ”

Jaynes 2003

“ The data cannot speak for themselves; and they never have, in any real problem of inference ”

Choice of test statistic

Conventionally based on profiled likelihood ratio

$$\lambda = \frac{p\left(D \mid \hat{\boldsymbol{\Theta}}_1, H_1\right)}{p\left(D \mid \hat{\boldsymbol{\Theta}}_0, H_0\right)}$$

where $\hat{\boldsymbol{\Theta}}_0$ are the best-fit parameters under H_0 etc and D are the data.

Optimal in simple cases (Neyman and Pearson 1933) and some slightly non-simple cases (Karlin-Rubin theorem).

Won't dwell on choice of test-statistic in this talk or how to compute it from a given dataset, which could involve multi-dimensional optimisation.

Objectivity

Gelman and Robert 2013

> *“ Strain on the gnat of prior while swallowing the camel that is the likelihood ”*

The *p*-value depends on researcher’s entire analysis plan and intentions (Wagenmakers 2007)

- How many tests did they perform?
- What would they have done were the data different?
- Why did they stop collecting data?

May be specified ahead of time e.g., registered reports gaining popularity in other fields (Kiyonaga and Scimeca 2019)

Backlash and defence

Ioannidis 2005 argued that most research findings were false

The subsequent replication crisis led to doubts about *p*-values, thresholds and testing (Benjamin et al. 2018; McShane et al. 2019).

Though see Murtaugh 2014; Lakens et al. 2018; Cousins 2018; Lakens 2021; Mayo 2018 for defence of *p*. Renaissance in Mayo's severe testing framework (Mayo 2018), though see Gelman et al. 2019

Is this a Bayesian approach?

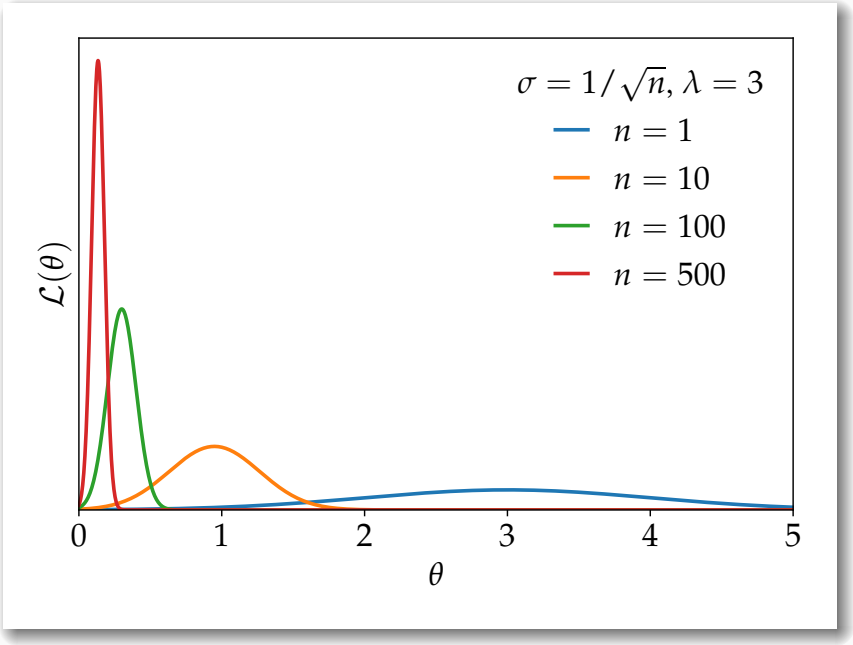
Dispute about whether Bayes factors belong in Bayesian paradigm (Gelman and Shalizi 2013; Robert 2016)

Robert 2016

> *“ I see [the Bayes factor] as a child of its time … Bayesian model comparison should abstain from automated and hard decision making. Looking at the marginal likelihood of a model as evidence makes it harder to refrain from setting decision bounds when compared with returning a posterior distribution ”*

Jeffreys-Lindley paradox

For a fixed p but increasing n, the likelihood function increasingly centred at $\theta = 0$ even though discrepancy from $\theta = 0$ is fixed



Fixed t but changing n is arguably change of state of knowledge of origin of t (Fowlie 2020)

References

References

Aad, Georges et al. (2012). "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." In: Phys. Lett. B 716, pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

Athron, Peter et al. (July 2021). "Global fits of axion-like particles to XENON1T and astrophysical data." In: JHEP 05, p. 159. DOI: 10.1007/JHEP05(2021)159. arXiv: 2007.05517 [astro-ph.CO].

Bailer-Jones, C.A.L. (2017). Practical Bayesian Inference: A Primer for Physical Scientists. Cambridge University Press. ISBN: 9781108127677.

References ii

Barnard, G. A. (July 1949). "Statistical Inference." In: Journal of the Royal Statistical Society: Series B (Methodological) 11.2, pp. 115–139. DOI: 10.1111/j.2517-6161.1949.tb00028.x. URL: https://doi.org/10.1111/j.2517-6161.1949.tb00028.x.

Bayarri, M Jésus and James O Berger (2004). "The interplay of Bayesian and frequentist analysis." In: Statistical Science 19.1, pp. 58–80.

Bayarri, M.J. et al. (June 2016). "Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses." In: Journal of Mathematical Psychology 72, pp. 90–103. DOI: 10.1016/j.jmp.2015.12.007. URL: https://doi.org/10.1016/j.jmp.2015.12.007.

References iii

Behnke, O. et al. (2013). Data Analysis in High Energy Physics: A Practical Guide to Statistical Wiley. ISBN: 9783527653430.

Benjamin, Daniel J. et al. (Jan. 2018). "Redefine statistical significance." In: Nature Human Behaviour 2.1, pp. 6–10. ISSN: 2397-3374. DOI: 10.1038/s41562-017-0189-z.

Berger, J.O. and R.L. Wolpert (1988). The Likelihood Principle. second. Vol. 6. Lecture notes – monographs series. Institute of Mathematical Statistics. ISBN: 9780940600133.

Berger, James O (2003). "Could Fisher, Jeffreys and Neyman have agreed on testing?" In: Statistical Science 18.1, pp. 1–32.

References iv

Berger, James O and Luis R Pericchi (2001). "Objective Bayesian methods for model selection: Introduction and comparison." In: IMS Lecture Notes – Monograph Series 38, pp. 135–207. DOI: 10.1214/lnms/1215540968.

Berger, James O. and Mohan Delampady (Aug. 1987). "Testing Precise Hypotheses." In: Statist. Sci. 2.3, pp. 317–335. DOI: 10.1214/ss/1177013238.

Berger, James O. and Thomas Sellke (1987). "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence." In: J. Am. Stat. Assoc. 82.397, pp. 112–122. DOI: 10.1080/01621459.1987.10478397.

Birnbaum, Allan (1962). "On the Foundations of Statistical Inference." In: J. Am. Stat. Assoc. 57.298, pp. 269–306. DOI: 10.1080/01621459.1962.10480660.

References v

Consonni, Guido et al. (June 2018). "Prior Distributions for Objective Bayesian Analysis." In: Bayesian Analysis 13.2, pp. 627–679. DOI: 10.1214/18-BA1103.

Cousins, Robert D. (2008). "Comment on 'Bayesian Analysis of Pentaquark Signals from CLAS Data', with Response to the Reply by Ireland and Protopopsecu." In: Phys. Rev. Lett. 101, p. 029101. DOI: 10.1103/PhysRevLett.101.029101. arXiv: 0807.1330 [hep-ph].

– (2017). "The Jeffreys-Lindley paradox and discovery criteria in high energy physics." In: Synthese 194.2. [Erratum: Synthese (2015)], pp. 395–432. DOI: 10.1007/s11229-014-0525-z,10.1007/s11229-015-0687-3. arXiv: 1310.3791 [physics.data-an].

References vi

Cousins, Robert D. (July 2018). "Lectures on Statistics in Theory: Prelude to Statistics in Practice." In: arXiv e-prints. See Sec. 7.4. arXiv: 1807.05996 [physics.data-an].

– (Oct. 2020). "What is the likelihood function, and how is it used in particle physics?" In: arXiv preprint. CERN EP Newsletter. arXiv: 2010.00356 [physics.data-an].

Cowan, G. (1998). Statistical Data Analysis. Clarendon Press. ISBN: 9780191583346.

Cumming, Geoff (Nov. 2013). "The New Statistics." In: Psychological Science 25.1, pp. 7–29. DOI: 10.1177/0956797613504966.

D'Agostini, Giulio (2003). Bayesian Reasoning In Data Analysis: A Critical Introduction. World Scientific Publishing Company. ISBN: 9789814486095.

Demortier, Luv (2008). “P Values and Nuisance Parameters.” In: DOI: 10.5170/CERN-2008-001.23. URL: https://cds.cern.ch/record/1099967.

Etz, Alexander and Eric-Jan Wagenmakers (2017). “JBS Haldane’s contribution to the Bayes factor hypothesis test.” In: Statistical Science, pp. 313–329.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver and Boyd.

Fowlie, Andrew (2017). “Bayes factor of the ATLAS diphoton excess: Using Bayes factors to understand anomalies at the LHC.” In: Eur. Phys. J. Plus 132.1, p. 46. DOI: 10.1140/epjp/i2017-11340-1. arXiv: 1607.06608 [hep–ph].

References viii

Fowlie, Andrew (2018). "DAMPE squib? Significance of the 1.4 TeV DAMPE excess." In: Phys. Lett. B 780, pp. 181–184. DOI: 10.1016/j.physletb.2018.03.006. arXiv: 1712.05089 [hep-ph].

– (2019). "Bayesian and frequentist approaches to resonance searches." In: JINST 14.10, P10031. DOI: 10.1088/1748-0221/14/10/P10031. arXiv: 1902.03243 [hep-ph].

– (Dec. 2020). "Objective Bayesian approach to the Jeffreys-Lindley paradox." In: Comm. Statist. Theory Methods. DOI: 10.1080/03610926.2020.1866206. arXiv: 2012.04879 [stat.ME].

– (Oct. 2021). "Neyman-Pearson lemma for Bayes factors." In: DOI: 10.1080/03610926.2021.2007265. arXiv: 2110.15625 [math.ST].

Fowlie, Andrew, Sebastian Hoof, and Will Handley (May 2022). "Nested Sampling for Frequentist Computation: Fast Estimation of Small p-Values." In: Phys. Rev. Lett. 128.2, p. 021801. DOI: 10.1103/PhysRevLett.128.021801. arXiv: 2105.13923 [physics.data-an].

Gelman, Andrew and Christian P Robert (2013). ""Not only defended but also applied": The perceived absurdity of Bayesian inference." In: The American Statistician 67.1, pp. 1–5.

Gelman, Andrew and Cosma Rohilla Shalizi (2013). "Philosophy and the practice of Bayesian statistics." In: British Journal of Mathematical and Statistical Psychology 66.1, pp. 8–38.

References x

Gelman, Andrew et al. (May 2019). "Many perspectives on Deborah Mayo's "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars"." In: arXiv e-prints, arXiv:1905.08876, arXiv:1905.08876. arXiv: 1905.08876 [stat.OT].

Goldstein, Michael (Sept. 2006). "Subjective Bayesian Analysis: Principles and Practice." In: Bayesian Analysis 1.3. DOI: 10.1214/06-ba116. URL: https://doi.org/10.1214/06-ba116.

Good, I. J. (1957). "Saddle-point Methods for the Multinomial Distribution." In: The Annals of Mathematical Statistics 28.4, pp. 861–881. DOI: 10.1214/aoms/1177706790.

– (1961). "Weight of evidence, causality and false-alarm probabilities." In: Information Theory: Fourth London Symposium. Butterworth, London, pp. 125–136.

Good, I. J. (1992). "The Bayes/Non-Bayes Compromise: A Brief Review." In: Journal of the American Statistical Association 87.419, pp. 597–606. ISSN: 01621459.

– (1960). "Weight of evidence, corroboration, explanatory power, information and the utility of experiments." In: Journal of the Royal Statistical Society: Series B (Methodological) 22.2, pp. 319–331.

– (1965). "A list of properties of Bayes-Turing Factors." In: NSA Technical Journal 10.2, pp. 1–6.

Goodman, Steven (2008). "A Dirty Dozen: Twelve P-Value Misconceptions." In: Seminars in Hematology 45.3, pp. 135–140. ISSN: 0037-1963. DOI: 10.1053/j.seminhematol.2008.04.003.

Gould, Peter (1981). "Letting the Data Speak for Themselves." In: Annals of the Association of American Geographers 71.2, pp. 166–176.

Greenland, Sander et al. (Apr. 2016). "Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations." In: European Journal of Epidemiology 31.4, pp. 337–350. ISSN: 1573-7284. DOI: 10.1007/s10654-016-0149-3.

Gregory, P. (2005). Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press. ISBN: 9780521841504.

Hubbard, Raymond and M. J Bayarri (2003). "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing." In: Am. Stat. 57.3, pp. 171–178. DOI: 10.1198/0003130031856.

Hubbard, Raymond and R. Murray Lindsay (Feb. 2008). "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing." In: Theory & Psychology 18.1, pp. 69–88. DOI: 10.1177/0959354307086923.

Ioannidis, John P. A. (Aug. 2005). "Why Most Published Research Findings Are False." In: PLoS Medicine 2.8, e124. DOI: 10.1371/journal.pmed.0020124.

James, F. (2006). Statistical Methods in Experimental Physics. World Scientific. ISBN: 9789812567956.

Jaynes, Edwin T (2003). Probability theory: The logic of science. Cambridge University Press.

– (1968). "Prior Probabilities." In: IEEE Transactions on Systems Science and Cybernetics 4.3, pp. 227–241. DOI: 10.1109/TSSC.1968.300117.

Jefferys, William H and James O Berger (1992). "Ockham's razor and Bayesian analysis." In: American Scientist 80.1, pp. 64–72.

Jeffreys, Harold (1939). The Theory of Probability. Oxford Classic Texts in the Physical Sciences. Oxford University Press. ISBN: 9780198503682.

Joachimi, B. et al. (2021). "When tension is just a fluctuation: How noisy data affect model comparison." In: Astron. Astrophys. 647, p. L5. DOI: 10.1051/0004-6361/202039560. arXiv: 2102.09547 [astro-ph.CO].

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors." In: J. Am. Statist. Assoc. 90.430, pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

Kass, Robert E. and Larry Wasserman (1996). "The Selection of Prior Distributions by Formal Rules." In: Journal of the American Statistical Association 91.435, pp. 1343–1370. DOI: 10.1080/01621459.1996.10477003.

Kerridge, D. (Sept. 1963). "Bounds for the Frequency of Misleading Bayes Inferences." In: Ann. Math. Statist. 34.3, pp. 1109–1110. DOI: {10.1214/aoms/1177704038}.

Kiyonaga, Anastasia and Jason M. Scimeca (Sept. 2019). "Practical Considerations for Navigating Registered Reports." In: Trends in Neurosciences 42.9, pp. 568–572. DOI: 10.1016/j.tins.2019.07.003.

Lakens, Daniël (2021). "The Practical Alternative to the *p* Value Is the Correctly Used *p* Value." In: Perspectives on Psychological Science. DOI: 10.1177/1745691620958012.

Lakens, Daniel et al. (Mar. 2018). "Justify your alpha." In: Nature Human Behaviour 2.3, pp. 168–171. ISSN: 2397-3374. DOI: 10.1038/s41562-018-0311-x.

Linden, W. von der, V. Dose, and U. von Toussaint (2014). Bayesian Probability Theory: Applications in the Physical Sciences. Cambridge University Press. ISBN: 9781107035904.

Lindley, D. V. (1957). "A Statistical Paradox." In: Biometrika 44.1/2, pp. 187–192. ISSN: 00063444. DOI: 10.1093/biomet/44.1-2.187.

Lyons, L. (1989). Statistics for Nuclear and Particle Physicists. Cambridge University Press. ISBN: 9781316101636.

Lyons, Louis (Oct. 2013). "Discovering the Significance of 5 sigma." In: arXiv preprint. arXiv: 1310.1284 [physics.data-an].

MacKay, David J.C. (1992). "Bayesian methods for adaptive models." en. Doctoral dissertation. DOI: 10.7907/H3A1-WM07.

Martin, Jérôme, Christophe Ringeval, and Vincent Vennin (2014). "Encyclopædia Inflationaris." In: Phys. Dark Univ. 5-6, pp. 75–235. DOI: 10.1016/j.dark.2014.01.003. arXiv: 1303.3787 [astro-ph.CO].

Martin, Jérôme et al. (2014). "The Best Inflationary Models After Planck." In: JCAP 1403, p. 039. DOI: 10.1088/1475-7516/2014/03/039. arXiv: 1312.3529 [astro-ph.CO].

Mayo, Deborah G (2018). Statistical inference as severe testing. Cambridge University Press.

McShane, Blakeley B. et al. (2019). "Abandon Statistical Significance." In: The American Statistician 73.sup1, pp. 235–245. DOI: 10.1080/00031305.2018.1527253. arXiv: 1709.07588 [stat.ME].

Mikkola, Petrus et al. (Dec. 2021). "Prior knowledge elicitation: The past, present, and future." In: arXiv: 2112.01380 [stat.ME].

Murtaugh, Paul A. (Mar. 2014). "In defense of P values." In: Ecology 95.3, pp. 611–617. DOI: 10.1890/13-0590.1.

Neyman, J. and E. S. Pearson (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses." In: Philos. Trans. Roy. Soc. London Ser. A 231, pp. 289–337. ISSN: 02643952. DOI: 10.1098/rsta.1933.0009.

Pratt, John W. (1962). "On the foundations of statistical inference: Discussion." In: Journal of the American Statistical Association 57.298, pp. 307–326. DOI: 10.1080/01621459.1962.10480661.

Robert, Christian P (2016). "The expected demise of the Bayes factor." In: Journal of Mathematical Psychology 72, pp. 33–37.

– (2014). "On the Jeffreys-Lindley Paradox." In: Philos. Sci. 81.2, pp. 216–232. DOI: 10.1086/675729. arXiv: 1303.5973 [stat.ME].

Savage, L.J. (1962). "On the foundations of statistical inference: Discussion." In: Journal of the American Statistical Association 57.298, pp. 307–326. DOI: 10.1080/01621459.1962.10480661.

Schervish, Mark J. (1996). "P Values: What They are and What They are Not." In: Am. Stat. 50.3, pp. 203–206. DOI: 10.1080/00031305.1996.10474380.

References xx

Sellke, Thomas, M. J Bayarri, and James O Berger (2001). "Calibration of p-values for Testing Precise Null Hypotheses." In: The American Statistician 55.1, pp. 62–71. DOI: 10.1198/000313001300339950.

Senn, S (Mar. 2001). "Two cheers for P-values?" In: Journal of Epidemiology and Biostatistics 6.2, pp. 193–204. DOI: 10.1080/13595220175317295 3.

Sivia, D. and J. Skilling (2006). Data Analysis: A Bayesian Tutorial. Oxford University Press. ISBN: 9780198568322.

Skilling, John (Dec. 2006). "Nested sampling for general Bayesian computation." In: Bayesian Analysis 1.4, pp. 833–859. DOI: 10.1214/06-BA127.

Strumia, Alessandro (2016). "Interpreting the 750 GeV digamma excess: a review." In: 51st Rencontres de Moriond on EW Interactions and Unified Theories ARISF, pp. 407–426. arXiv: 1605.09401 [hep-ph].

Trotta, Roberto (2008). "Bayes in the sky: Bayesian inference and model selection in cosmology." In: Contemp. Phys. 49, pp. 71–104. DOI: 10.1080/00107510802066753. arXiv: 0803.4089 [astro-ph].

Tukey, John W et al. (1977). Exploratory data analysis. Vol. 2. Reading, MA.

Verde, L., T. Treu, and A. G. Riess (July 2019). "Tensions between the Early and the Late Universe." In: Nature Astron. 3, p. 891. DOI: 10.1038/s41550-019-0902-0. arXiv: 1907.10625 [astro-ph.CO].

Vovk, V. G. (1993). "A Logic of Probability, with Application to the Foundations of Statistics." In: J. Royal Stat. Soc. B55.2, pp. 317–341. DOI: 10.1111/j.2517-6161.1993.tb01904.x.

Wagenmakers, E. and et al (2017). "Bayesian Spectacles." In: https://www.bayesianspectacles.org/.

Wagenmakers, Eric-Jan (Oct. 2007). "A practical solution to the pervasive problems of *p* values." In: Psychonomic Bulletin & Review 14.5, pp. 779–804. ISSN: 1531-5320. DOI: 10.3758/BF03194105.

Wagenmakers, Eric-Jan and Alexander Ly (2021). "History and Nature of the Jeffreys-Lindley Paradox." In: DOI: 10.48550/arXiv.2111.10191. arXiv: 2111.10191 [stat.ME].

Wasserman, Larry, Aaditya Ramdas, and Sivaraman Balakrishnan (2020). "Universal inference." In: Proceedings of the National Academy of Sciences 117.29, pp. 16880–16890. arXiv: 1912.11436 [math.ST].

Wasserstein, Ronald L. and Nicole A. Lazar (Apr. 2016). "The ASA Statement on p-Values: Context, Process, and Purpose." In: The American Statistician 70.2, pp. 129–133. DOI: 10.1080/00031305.2016.1154108. URL: https://doi.org/10.1080/00031305.2016.1154108.

Zhang, Jin (Nov. 2016). "Bayesian (mean) most powerful tests." In: Australian & New Zealand Journal of Statistics 59.1, pp. 43–56. DOI: 10.1111/anzs.12171. URL: https://doi.org/10.1111/anzs.12171.