Using Bayes factors to understand anomalies at the LHC and Future Colliders

A. Fowlie, Eur. Phys. J. Plus 132, 46 (2017), arXiv:1607.06608 [hep-ph]

Andrew Fowlie

September 30 2017. Energy Frontier in Particle Physics: LHC and Future Colliders

Monash University

Table of contents

1. Nutshell
2. Theory
3. 750 GeV anomaly
4. Summary

Nutshell

Growing controversy in other sciences about p-values

American Statistical Association (2016) [2]:

- “The p-value was never intended to be a substitute for scientific reasoning”
- “Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.”
- “By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.”

Growing controversy in other sciences about p-values

Redefine Statistical Significance [3, 4]:

- “The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on “statistically significant” findings”

Abandon Statistical Significance [5]

- “we recommend abandoning the null hypothesis significance testing … p-values as just one of many pieces of information with no privileged role in scientific publication”

See also critiscism in Ref. [6].

Reproducibility crisis

There is a crisis in social sciences that ostensibly significant results cannot be reproduced.

The Economist
Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar
HOW SCIENCE GOES WRONG
99
Einsteinium

Suspicion that p-values are part of the problem. Suggestions to use alternative techniques or, pragmatically, to lower p-value threshold from 2σ.

Golden 5σ criteria

In particle physics, we are protected by 5σ criteria for discovery.

Golden 5σ criteria

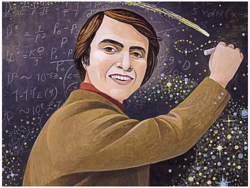
In particle physics, we are protected by 5σ criteria for discovery.

Degrades statistical power — small probability of rejecting false null hypothesis.

Golden 5σ criteria

In particle physics, we are protected by 5σ criteria for discovery.

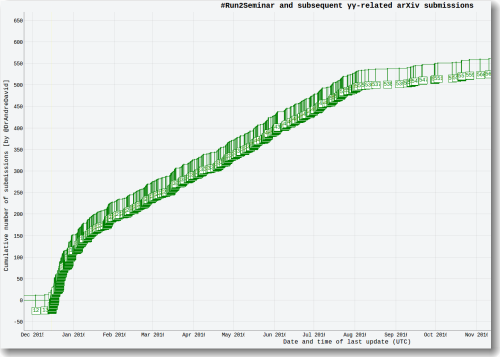
Equal evidential standards for plausible and implausible theories. Don't extraordinary claims require extraordinary evidence?



Golden 5σ criteria

In particle physics, we are protected by 5σ criteria for discovery.

In reality, no phenomenologist/theorist waits that long. We make hundreds of papers about 3σ and dozens about 2σ.



Golden 5σ criteria

In particle physics, we are protected by 5σ criteria for discovery.

p-values are frequently misinterpreted by physicists.

“The standard approach in teaching — of stressing the formal definition of a p-value while warning against its misinterpretation — has simply been an abysmal failure” Berger [7].

Golden 5σ criteria

In particle physics, we are protected by 5σ criteria for discovery.

Is 5σ enough anyway? D'Agostini predicts that the next 5σ claim from LHC will be a fluke [8].

Perhaps it is the combined anomalies in b-physics?

Theory

Current formalism (NHST)

Construct test-statistic, e.g. log-likelihood ratio for hypothesis

$$q \equiv -2 \ln \frac{\mathcal{L}(\mu = 0, \hat{\theta}_0)}{\mathcal{L}(\hat{\mu}, \hat{\theta})}$$

By Neyman-Pearson lemma, this maximises probability of rejecting null hypothesis if it is false.

Define the p-value,

$$\text{p-value} \equiv P(q \geq q_\text{observed} \,|\, \mu = 0)$$

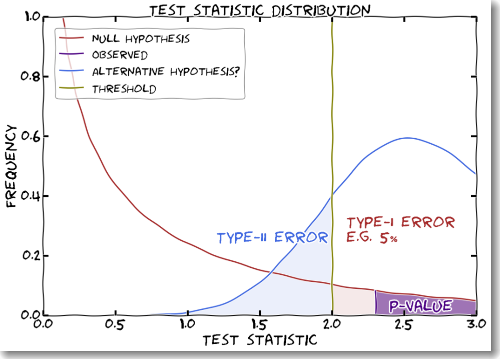
This may be difficult to calculate and interpret due to look-elsewhere effect.

Current formalism (NHST)

Construct test-statistic, e.g. log-likelihood ratio for hypothesis

$$q \equiv -2 \ln \frac{\mathcal{L}(\mu = 0, \hat{\hat{\theta}}_0)}{\mathcal{L}(\hat{\mu}, \hat{\theta})}$$

By Neyman-Pearson lemma, this maximises probability of rejecting null hypothesis if it is false.



Current formalism (NHST)

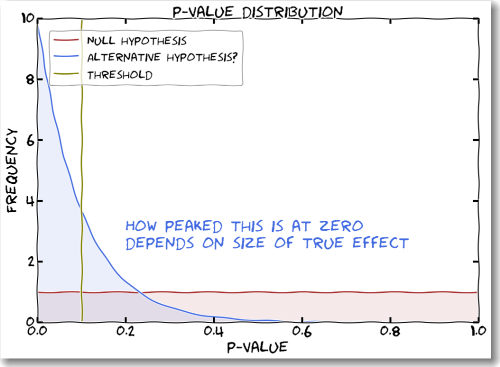
Conventional to convert this into a Z-value under one-tailed Gaussian

$$\text{p-value} = \frac{1}{2}\left[1 - F_{\chi^2}(Z^2)\right]$$

We then reject null hypothesis if Z-value greater than a threshold of 5.

What does it mean?

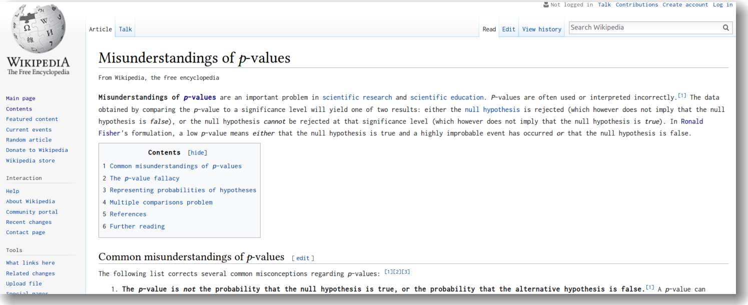
All p-values equally probable under null hypothesis. Why are small p-values special?



Because peaked around zero under alternative hypothesis. Evidence depends upon the size of peak!

p-value fallacies

There are many p-value fallacies. It has a wiki page.



Misunderstandings of *p*-values

From Wikipedia, the free encyclopedia

Misunderstandings of *p*-values are an important problem in scientific research and scientific education. *P*-values are often used or interpreted incorrectly.[1] The data obtained by comparing the *p*-value to a significance level will yield one of two results: either the null hypothesis is rejected (which however does not imply that the null hypothesis is *false*), or the null hypothesis *cannot* be rejected at that significance level (which however does not imply that the null hypothesis is *true*). In Ronald Fisher's formulation, a low *p*-value means *either* that the null hypothesis is true and a highly improbable event has occurred *or* that the null hypothesis is false.

Contents [hide]

Common misunderstandings of *p*-values [edit]

The following list corrects several common misconceptions regarding *p*-values: [1][2][3]

1. **The *p*-value is *not* the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.**[1] A *p*-value can

p-value is not related by any general formula to $p(H_0 \mid D)$.

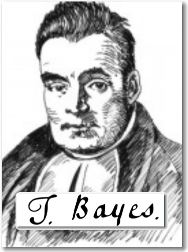
Public and physicists often confused.

Bayesian approach

Plausibility represented by probabilities.

A calculus of beliefs.

Directly calculate change in relative plausibility of two hypothesis in light of data.



Developed by Bayes, Laplace and Jeffreys.

Bayes factor

This is as simple as finding

$$\text{Bayes factor} = \frac{\text{Relative plausibility after data}}{\text{Relative plausibility before data}}$$

in math, by Bayes' theorem,

$$\text{Bayes factor} = \underbrace{\frac{p(D \mid M_a)}{p(D \mid M_b)}}_{\text{Calculate this ratio}} = \frac{\overbrace{\dfrac{p(M_a \mid D)}{p(M_b \mid D)}}^{\text{Posterior odds — output}}}{\underbrace{\dfrac{p(M_a)}{p(M_b)}}_{\text{Prior odds — input}}}$$

for models a and b, and data D.

Evidence

A Bayes factor is itself a ratio of evidences. An evidence may be calculated by the integral

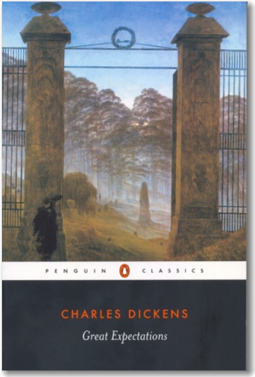
$$p(D \mid M) = \int p(D \mid M, x) \cdot p(x \mid M) \, dx$$

The integration is over the model's parameters x. The integration may be computationally challenging.

The integrand is a product of likelihood and prior.

Likelihood could be e.g. a Gaussian for Higgs mass measurement or a Poisson for a counting experiment.

Evidence



Take nothing on its looks; take everything on BAYESIAN evidence. There's no better rule.

Automatic Occam's razor

The Bayes factor automatically penalises models that make wrong or diffuse predictions.

Each model has a finite probability mass to spend:

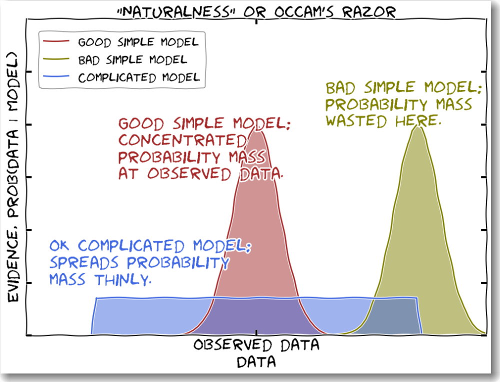
$$\int p(D \mid M)\, dD = 1$$

Good models spend it wisely about the observed data such that their evidence is big.

Bad models either make diffuse predictions — with probability spread thinly — or bad predictions.

Automatic Occam’s razor

The Bayes factor automatically penalises models that make wrong or diffuse predictions.



This automatic Occam’s razor, incidentally, penalises fine-tuning associated with hierarchy problem.

p-value fallacy

The difference between Bayesian model selection and p-values is just the prior odds, such that

$$\text{Posterior odds} = \text{p-value} \times \text{prior odds}$$

Since we are objective scientists, we only report the p-value and don’t mention any priors!

p-value fallacy

~~The difference between Bayesian model selection and p-values is just the prior odds, such that~~

~~$\text{Posterior odds} = \text{p-value} \times \text{prior odds}$~~

~~Since we are objective scientists, we only report the p-value and don't mention any priors!~~

F

p-value fallacy

There is no general equation linking a p-value with plausibility of null hypothesis or any other product of Bayesian analysis.

In fact,

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

Recommended practice is to report Bayes factors and permit a reader to supply prior odds.

Priors for parameters

Reflect beliefs/ignorance about parameters prior to data.

Parameter inference: shape and diffuseness washed out by sufficient data. Even possible that an improper prior leads to a proper posterior.

Model selection: difficulties. Shape and diffuseness cannot be washed out.

Priors for parameters

Reflect beliefs/ignorance about parameters prior to data.

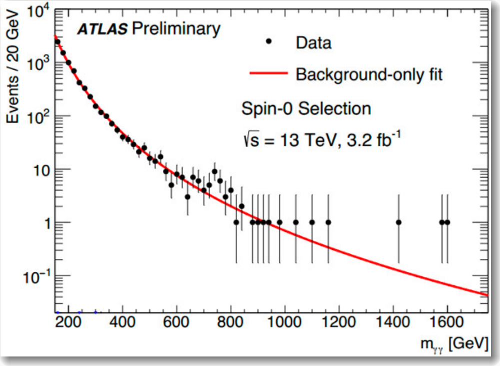
Often, prior should reflect our ignorance by an invariance under a symmetry.

If we are ignorant of scale, our prior should be invariant under rescaling. This is a logarithmic prior, $p(\ln x) =$ const.

Must check sensitivity to variations in prior shape and breadth that are somewhat compatible with our prior knowledge.

750 GeV anomaly

Let’s go back in time!



An anomaly that can be seen with the eye! How exciting! That must be evidence for new physics. But how much?

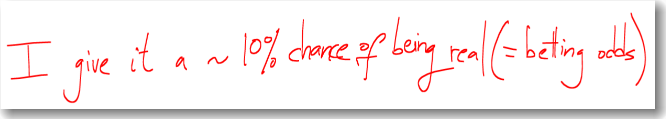
Interpretations

No one was sure, but many people liked to articulate their belief in the digamma with a probability.

Though this was mainly anecdotal, rarely in talks or papers.

Interpretations

I found an example.



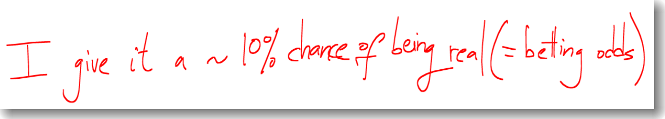
I give it a ~ 10% chance of being real (= betting odds)

Is this analysis rational/scientific? What is the origin of 10% or its relation to p-values?

Why are we just making up numbers?

Interpretations

I found an example.



I give it a ~ 10% chance of being real (= betting odds)

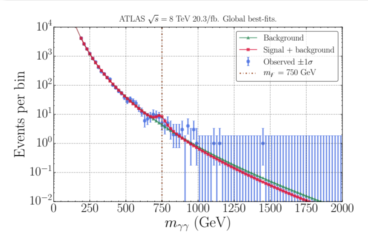
Is this analysis rational/scientific? What is the origin of 10% or its relation to p-values?

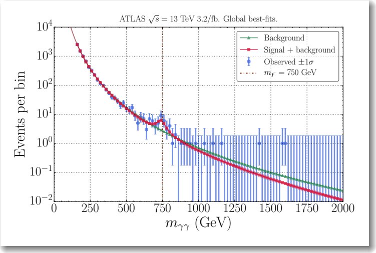
Why are we just making up numbers?

Because we wanted to express degrees of belief! Statistics with which we described 750 GeV anomaly — p-values and sigmas — weren't that useful.

Data — reminder

Likelihood function was a Poisson for predicted and observed numbers of events in each bin in ATLAS data.





There were three ATLAS datasets [9, 10]:

- 20.3/fb at 8 TeV. Anomaly about 2σ local
- 3.2/fb at 13 TeV. Anomaly 3.9σ local, 2.1σ global
- 15.4/fb at 13 TeV. Anomaly 2.6σ local, 0.8σ global

Models for 750 GeV anomaly

Modelled digamma by a Breit-Wigner resonance with an unknown mass and width:

$$p(m_{\gamma\gamma}) \propto \frac{1}{(m_{\gamma\gamma}^2 - m_F^2)^2 + \Gamma_F^2 m_F^2}$$

Modelled backgrounds with parametric form presented by ATLAS:

$$p(m_{\gamma\gamma}) \propto \left[1 - \left(\frac{m_{\gamma\gamma}}{\sqrt{s}}\right)^{\frac{1}{3}}\right]^b \left(\frac{m_{\gamma\gamma}}{\sqrt{s}}\right)^a .$$

Specified expected numbers of background and signal events.

Priors

Since SM and SM + F are composite models, limited sensitivity to diffuseness of priors for parameters in SM background.

Match prior ranges of mass and width to that which was searched for:

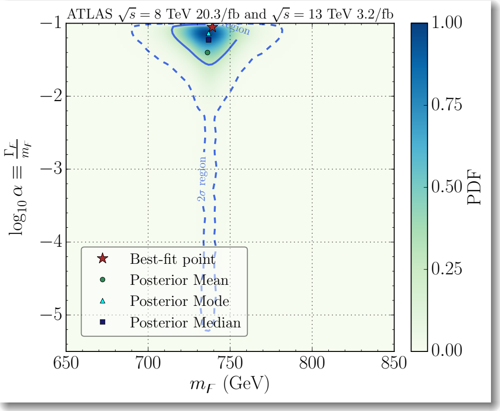
$$200\,\text{GeV} \leq m_F \leq 2\,\text{TeV}$$
$$5 \times 10^{-6} \leq \Gamma_F / m_F \leq 0.1$$

Picked reasonable range of number of events.

Checked prior sensitivity by considering changing shape and breadth of priors.

Posterior for mass and width

There were no surprises in the posterior — it favoured a substantial width and mass about 750 GeV.



Bayes factor

Found that digamma increased in plausibility by about 8 relative to SM in light of ATLAS data at height of excitement.

$$\text{Posterior odds} \approx 8 \times \text{prior odds}$$

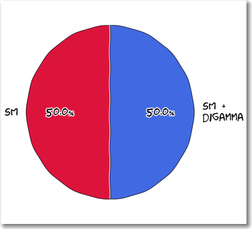
This is “substantial” evidence, lying between “barely worth a mention” and “strong”.

Not particularly sensitive to prior.

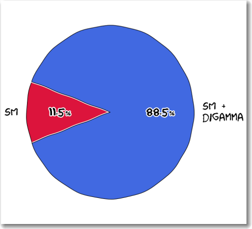
Linear priors increase Bayes factor, as they increase preference for $\Gamma_F / m_F \approx 0.05$, which was if anything a priori implausible.

Increasing breadth of priors of digamma model only decreases Bayes factor.

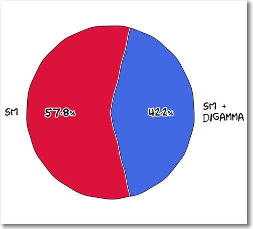
Relative plausibility between SM + digamma and SM



(I) Equal prior odds



(II) Peak excitement



(III) All told

Evidence from ATLAS was never particularly impressive. The 2σ + 3.9σ local anomalies were not strong evidence.

Bayes factor of less than 8 less than Bayes factor for double-headed coin versus fair coin if 3 heads in a row.

Combination of ATLAS and CMS

Difficult to model, as slightly different ansatz for background and different selection efficiencies.

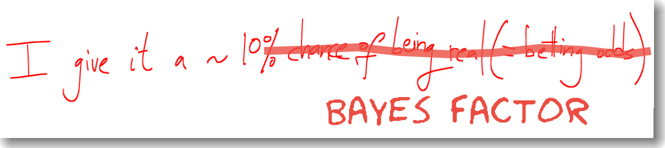
Difficulties present for frequentist and Bayesian analysis.

We'll probably never know any of the important numbers.

Summary of 750 GeV analysis

Frequentist — Probability of observing a test statistic anywhere in (m_F, Γ_F) so extreme were the null hypothesis true was about 0.02. For the best-fitting (m_F, Γ_F), it was about 5×10^{-5}.

Bayesian — In light of data, digamma model about 10 times more plausible than it was relative to Standard Model.



I give it a ~ 10% ~~chance of being real (= betting odds)~~

BAYES FACTOR

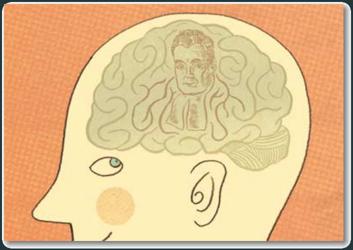
The latter appears to be closer to how theorists think about anomalies.

Summary

Summary

- Bayes factor directly measures relative change in plausibility of hypothesis
- No issues with their interpretation. cf. look-elsewhere effects and p-value fallacies
- Danger that future colliders become factories for 750 GeV style bonanzas. Remember D'Agostini's prediction: the next 5σ will be a fluke
- Present p-values and Bayes factors to fully describe an anomaly
- Latter appear to be closer to the way theorists/phenomenologists think about anomalies

Questions?



References

[1] A. Fowlie, “Bayes factor of the ATLAS diphoton excess: Using Bayes factors to understand anomalies at the LHC,” Eur. Phys. J. Plus 132, 46 (2017), arXiv:1607.06608 [hep-ph].

[2] R. L. Wasserstein and N. A. Lazar, “The ASA’s Statement on p-values: Context, Process, and Purpose,” The American Statistician 70, 129–133 (2016).

[3] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. Wagenmakers, and et al, “Redefine statistical significance,” Human Nature Behavior (2017).

References ii

[4] E. Wagenmakers and et al, “Bayesian spectacles,” https://www.bayesianspectacles.org/ (2017).

[5] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, “Abandon statistical significance,” (2017), arXiv:1709.07588 [stat.ME].

[6] D. Lakens and et al, “Justify Your Alpha: A Response to “Redefine Statistical Significance”,” (2017).

[7] T. Sellke, M. J. Bayarri, and J. O. Berger, “Calibration of p-values for Testing Precise Null Hypotheses,” The American Statistician 55, 62–71 (2001).

References iii

[8] G. D'Agostini, "The Waves and the Sigmas (To Say Nothing of the 750 GeV Mirage)," (2016), arXiv:1609.01668 [physics.data-an].

[9] M. Aaboud et al., "Search for resonances in diphoton events at $\sqrt{s} = 13$ TeV with the ATLAS detector," JHEP 09, 001 (2016), arXiv:1606.03833 [hep-ex].

[10] M. Aaboud et al., "Search for new phenomena in high-mass diphoton final states using 37 fb^{-1} of proton–proton collisions collected at $\sqrt{s} = 13$ TeV with the ATLAS detector," Submitted to: Phys. Lett. (2017), arXiv:1707.04147 [hep-ex].