Bayesian and frequentist approaches to resonance searches

arXiv:1902.03243 & arXiv:1712.05089

Andrew Fowlie

April 17, 2019

Nanjing Normal University

Table of contents

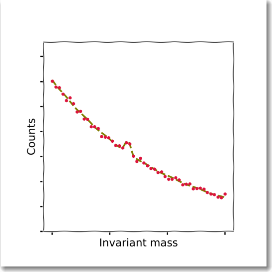
1. Background
2. Frequentist
3. Bayesian
4. Results from DAMPE
5. Results from toy Higgs search
6. Conclusions

Background

What is that?

A new particle? or just a fluctuation?

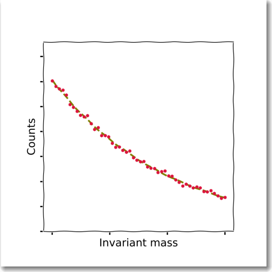
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

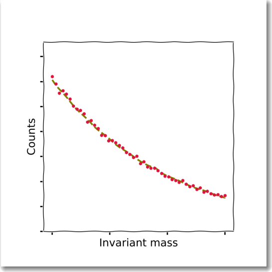
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

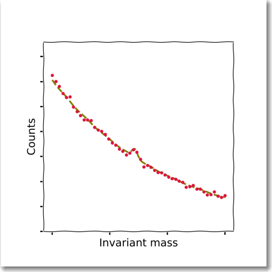
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

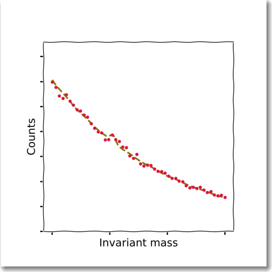
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

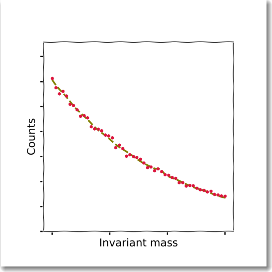
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

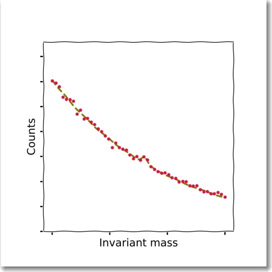
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

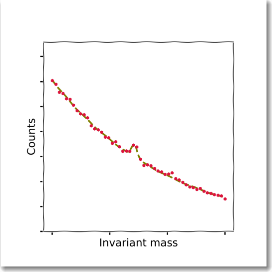
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

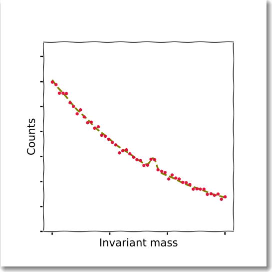
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

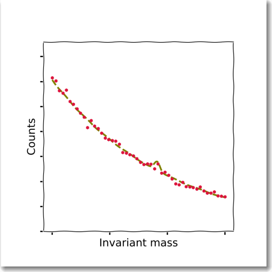
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

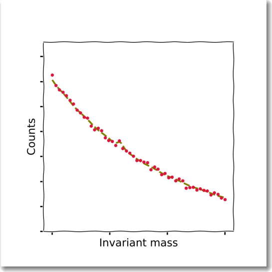
How can we characterise our uncertainty?



What is that?

A new particle? or just a fluctuation?

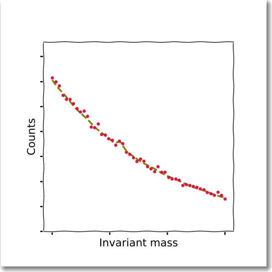
How can we characterise our uncertainty?



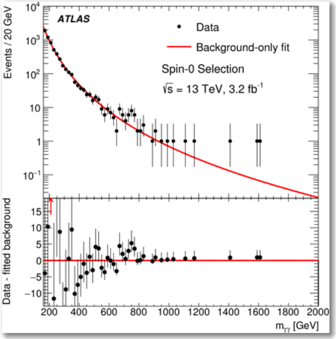
What is that?

A new particle? or just a fluctuation?

How can we characterise our uncertainty?



Another 750 GeV?



or something real? Should you write a paper about it? Announce a press conference? Start writing your Nobel prize speech?

How can we tell?

We need a statistical framework for treating our uncertainty.

1. Frequentist
2. Bayesian

Let's review them and compare results from them applied to realistic problems.

Frequentist

What is probability?

Probabilities are not degrees of certainty or belief.

Probabilities are frequencies at which events occur in identical repeat experiments.

$$P(A) = \lim_{N \to \infty} \frac{n_A}{N}$$

What can we do?

We cannot quantify our uncertainty about the resonance.

We can attempt to control the frequency at which we would make a type-1 error.

Type-1 error: Reject null hypothesis when it is true.

We must specify a null hypothesis, H_0, and a desired type-1 error rate, α. We reject H_0 at a pre-chosen significance α or we do not.

The rate α (implicitly) chosen to be about 10^{-7} (5σ) in particle physics.

Test-statistics

We construct a test-statistic that measures discrepancies between data and the null hypothesis, e.g. the log-likelihood ratio,

$$q \equiv -2\ln\frac{\max_{\boldsymbol{\theta}_1} P(D\,|\,M_1,\boldsymbol{\theta}_1)}{\max_{\boldsymbol{\theta}_2} P(D\,|\,M_0,\boldsymbol{\theta}_2)}$$

This involves numerical optimisation of the likelihood function over the models parameters $\boldsymbol{\theta}$.

In some settings, particular test-statistics can be shown to be the most powerful. The log-likelihood ratio is the most powerful one for comparing simple hypotheses.

Likelihood function

Probability (density) of data, D, given a particular model, M, with parameters $\boldsymbol{\theta}$

$$P\left(D \mid M, \boldsymbol{\theta}\right)$$

Typically well-defined and uncontroversial.

When used as a function of the parameters known as the likelihood function.

When used as a function of the data known as a sampling distribution.

Likelihood function for resonance search

Our data is binned. The likelihood is a product of Poissons, one for each bin

$$P(D|M,\boldsymbol{\theta}) = \prod_i \frac{e^{-\lambda_i}\lambda_i^{o_i}}{o_i!}$$

where the expected number of events depends on the model parameters, $\lambda = \lambda(\boldsymbol{\theta})$.

P-values

The probability of obtaining a test-statistic at least as extreme as the one we saw, if the null hypothesis was true

$$p\text{-value} = P\left(q \geq q_{\text{Observed}} \,\middle|\, H_0\right)$$

If p-value $< \alpha$, reject H_0

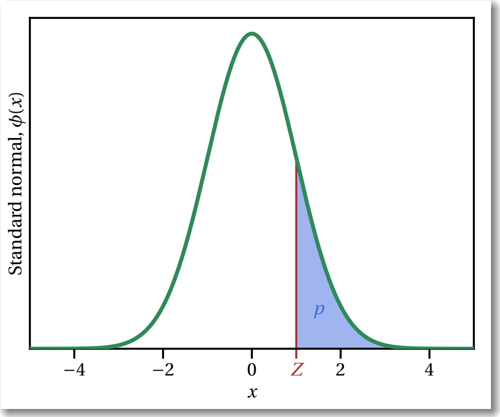
This is not a continuous measure of our confidence in H_0 — it was a means to controlling the type-1 error rate.

It is common nevertheless to interpret p as a measure of our confidence in H_0.

Conventional to convert p-value to Z-value (the number of sigma):

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

where Φ is the cumulative distribution function of a standard normal.



Global or local?

If the data had been different, we would have constructed a resonance model with a different mass to match the different data.

We would have looked elsewhere.

Global p-values account for this look-elsewhere effect.

Local p-values do not. They assume that we would only test particular parameters.

Bootstrap

We could calculate p-values by bootstrap:

1. Perform a toy experiment — sample data from the null hypothesis
2. Calculate our test-statistic — this requires maximization of the likelihood function
3. Find fraction for which $q > q_{\text{Observed}}$

This could be numerically challenging for small p-values, as the probability that $q > q_{\text{Observed}}$ would be small! We would need about $\mathcal{O}(1/p)$ toy experiments.

Maximizing the likelihood function could be numerically expensive for models with many parameters.

We calculated global p-values with Gross-Vitells [1] — a powerful semi-analytic technique.

It permits us to instead look for $q > u$, where we may choose u ourselves. This avoids the $\mathcal{O}(1/p)$ scaling of the number of toy experiments.

For resonance searches with an unknown mass and strength,

$$\text{Global } p\text{-value} \approx \frac{1}{2} P\left(\chi^2_1 > q\right) + N e^{-q/2}$$

where N must be found using toy experiments. The formula must be modified if the width was also unknown.

Asymptotic formulae

For resonance searches, there is an asymptotic formula for the local p-value that neglects look-elsewhere effects.

The formula makes use of Wilks'/Chernoff's theorem. Assuming only positive signals [2]

$$\text{Local } p\text{-value} = 1 - \Phi\left(\sqrt{q}\right)$$

where Φ is the cumulative distribution function of a standard normal distribution.

Bayesian

What is probability?

Probabilities are degrees of belief about any proposition.

There is a unique rule for updating them in light of information — Bayes' theorem.

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

Bayesian statistics $\Leftrightarrow$ probability theory

What can we do?

We can simply update our relative belief in models in light of data. For resonance searches, we update our belief in the signal + background model relative to the background only model.

The factor that updates our belief is a Bayes factor [3].

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

$$B = \frac{P(D \mid M_1)}{P(D \mid M_2)}$$

Don’t need to specify priors for the model

The Bayes factor updates our prior odds

$$\frac{P\left(M_1 \mid D\right)}{P\left(M_2 \mid D\right)} = B \times \frac{P\left(M_1\right)}{P\left(M_2\right)}$$

Ordinarily, we don’t specify them — let the reader perform the final multiplication. To compare with the p-value, though, we assume equal prior odds and find

$$P\left(M_0 \mid D\right) = \frac{1}{1 + B}$$

This is the plausibility of the background model in light of data.

How do we do it?

The numerator and denominator are so-called Bayesian evidences. For a model with parameters $\boldsymbol{\theta}$,

$$P(D|M) = \int P(D|M,\boldsymbol{\theta})\, p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta}$$

The factor $p(\boldsymbol{\theta}|M)$ is our prior for the model's parameters.

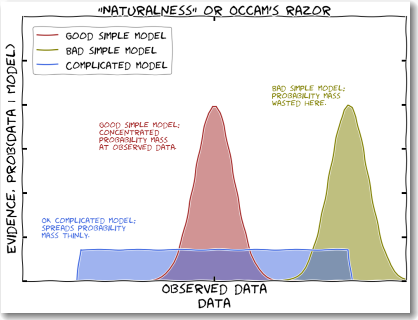
Efficient integration

That integral could be difficult, especially if the model contains many parameters

For a few parameters, (adaptive) quadrature might suffice, especially if any modes in the integrand are treated specially. In general, a dedicated algorithm might be necessary — e.g., nested sampling [4].

Occam's razor

The Bayesian evidence contains an automatic Occam's razor. Consider one-dimensional continuous data. $\int P(D|M)\,dD = 1$.



Complicated models make diffuse predictions. They squander their probability mass away from the observed data and are penalized.

What prior should I pick [5]?

Prior densities transform covariantly. A flat prior doesn't necessarily reflect ignorance — flat in which parameterisation!

$$p\left(x\right) = \text{const.} \Rightarrow p\left(x^2\right) \propto \frac{1}{x}$$

Priors

What prior should I pick [5]?

Subjective — anything goes

The prior represents your belief. That's it. There are no logical constraints on priors [6].

Priors

What prior should I pick [5]?

Subjective — elicit priors from experts

There are no rules but not everyone's prior is equal. Consult experts — who draw upon their knowledge and experience — to construct a prior [7].

Priors

What prior should I pick [5]?

Objective — Jaynesian

Jaynes' robot [8]. Priors are uniquely determined by your state of knowledge. Thus scientists with the same background knowledge construct the same priors.

There are particular rules — e.g., the principle of indifference, symmetry groups and maximum entropy.

Priors

What prior should I pick [5]?

Objective — default/reference priors

“Ignorance” is defined with respect to what could be learned in a particular experiment.

Priors are constructed that express ignorance by maximizing what you expect to learn in that experiment [9].

Priors

What prior should I pick [5]?

Robust analysis

Sympathetic to Jaynesian approach. Our prior knowledge isn't sufficient to uniquely determine a prior.

Check sensitivity to a class of priors that could reasonably be in agreement with our prior knowledge [10].

Can more data help?

A few results about the role of priors in the asymptotic limit:

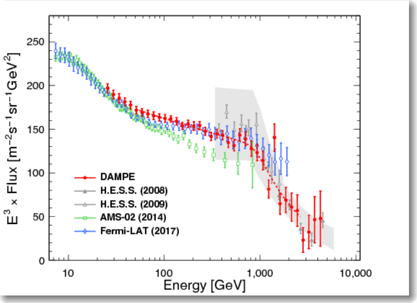
1. Under mild assumptions, there are theorems demonstrating that the posterior for the true model converges to one [11]
2. The impact of the breadth of the prior doesn't necessarily diminish as we collect data (Bartlett-Lindley paradox [12]).
   If a model contains a flat prior for a parameter on 0 to L, the evidence is typically penalized by
   $$P(D \,|\, M) \propto \frac{1}{L}$$
3. The Jeffreys-Lindley paradox [13, 14] shows Bayesian and frequentist results conflict even in asymptotic limit

From quantum mechanics, we learned an antidote to disputes about interpretations.

Shut up and calculate.

Results from DAMPE

What is that at 1.4 TeV?

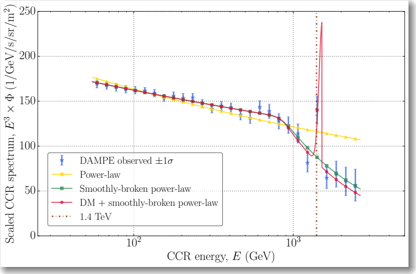


You know this well [15]. Let's turn the (statistical) handle!

Models

1. A single power law
2. A smoothly-broken power law
3. A smoothly-broken power law and with a signal from annihilating DM particles of mass m_χ, predicting a half-Gaussian feature of amplitude A and width σ

Best-fits



Which models are preferred?

Results

Very difficult to argue that there was strong evidence for DM.

1. Smoothly-broken power law $\ggg$ single power law by Bayesian and frequentist measures — $B \approx 10^{10}$ and a tiny p-value
2. Smoothly-broken power law $\simeq$ smoothly-broken power law + DM

For the latter, we found $B \approx 2$, but was sensitive to our choices of prior. The maximum possible Bayes factor was $B \approx 500$.

The global Z-value was about 2.3σ, whereas local was about 3.6σ.

Results from toy Higgs search

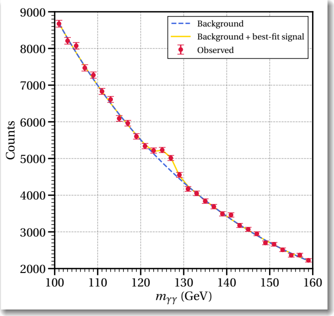
Higgs discovery

The most famous resonance search of them all!

In 2012 ATLAS and CMS observed a new boson in several resonance searches.

Toy problem

Let's use the search for the Higgs in the diphoton channel by ATLAS with 25/fb [16] as a toy problem.

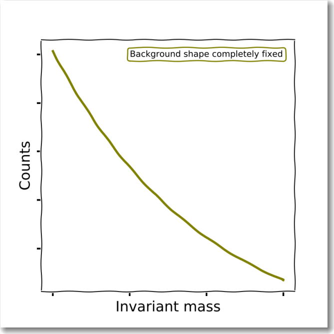


An important search for the discovery of the Higgs.

Background model

There is a monotonically falling background.

We could describe it by a basis of polynomials (e.g. Bernstein) but so that we can perform many calculations, we just use a fixed background and neglect parametric uncertainties in it.



Signal model i

We model the signal predicted by a Higgs as a Gaussian centred at m_h.

The width was the experimental resolution of about 1.5 GeV.

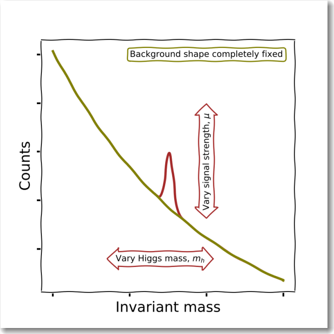
We specified the strength relative to the Standard Model prediction (at 125 GeV),

$$\mu = \frac{\text{efficiency} \times \text{cross section}}{(\text{efficiency} \times \text{cross section})_{\text{SM @ 125 GeV}}}$$

This is an approximation as we did not model dependence of efficiency or cross section as functions of Higgs mass.

Signal model ii

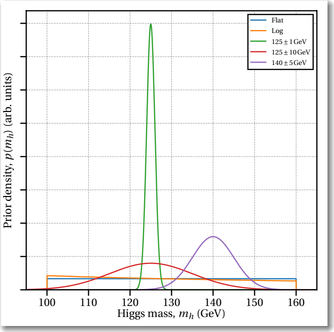
There were thus two unknown parameters describing the location and strength of the resonance, m_h and μ.



Priors

For our Bayesian calculations, we must place priors on m_h and μ. We experiment with several choices.

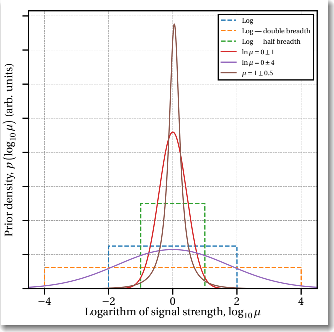
Priors



Broad priors (log and flat) and narrow ones representing specific prior knowledge.

Going beyond the range searched for the experiment (100 – 160 GeV) could represent our belief but only dilutes evidence for a signal.

Priors



We vary the breadth of the log prior for the signal strength, and the shape of the prior.

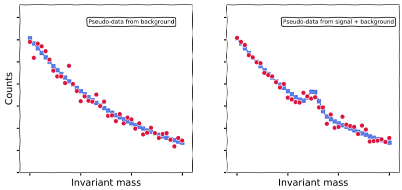
Data and pseudo-data

We use the real 25/fb collected by ATLAS [16].

We sample our own pseudo-data from the background model and the signal + background model with $\mu = 1$, $m_h = 125\,\text{GeV}$.

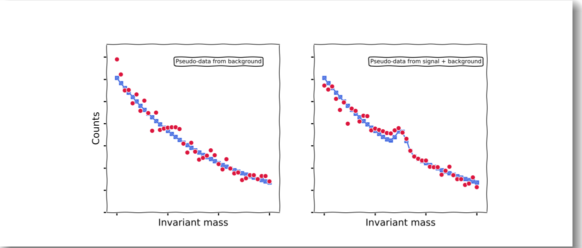
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



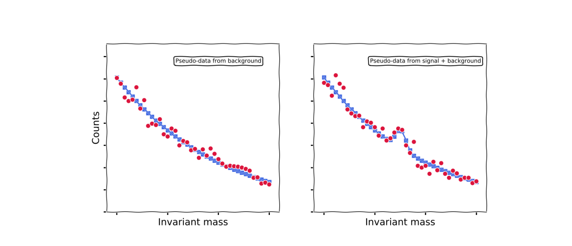
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



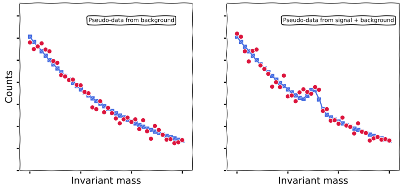
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



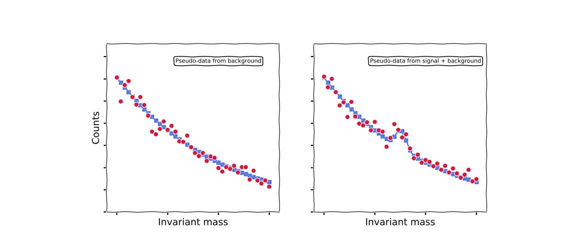
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



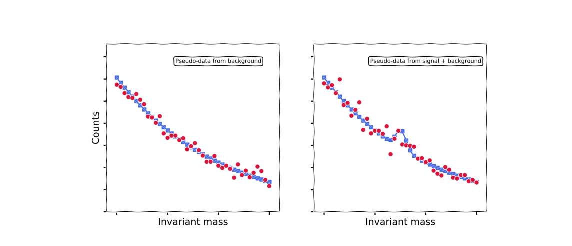
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



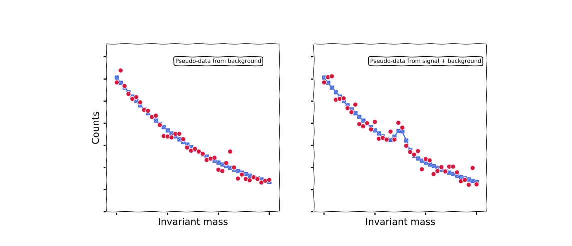
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



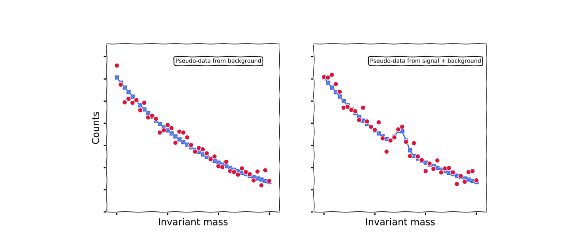
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



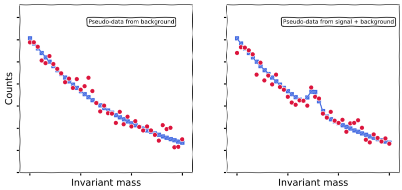
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



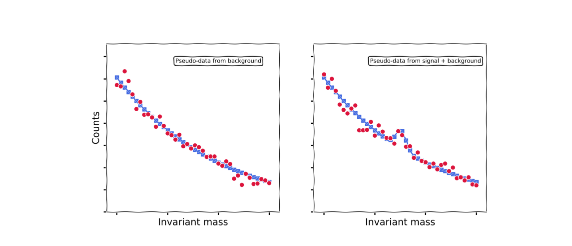
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



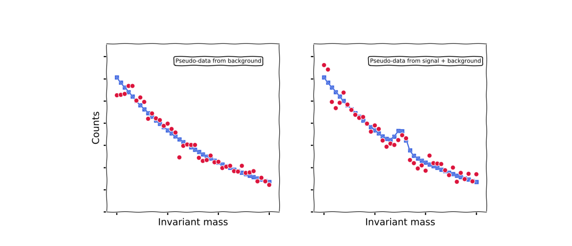
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.



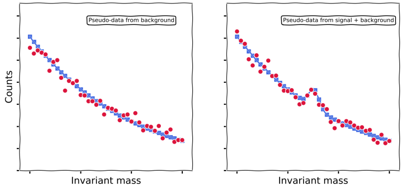
The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.

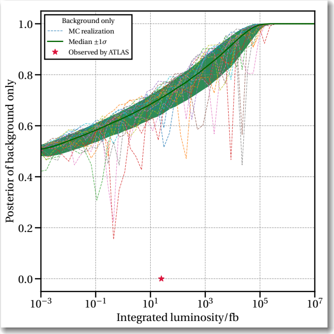


The models tell us the expected number of counts in each bin for a particular integrated luminosity.

We sample pseudo-data at many integrated luminosities by drawing counts from Poisson distributions in each bin.

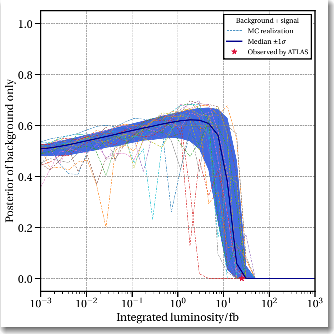


Evolution of p-value and posterior as we collect data



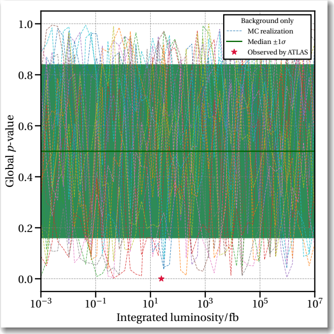
The posterior slowly approaches 1 when the background model is correct

Evolution of p-value and posterior as we collect data



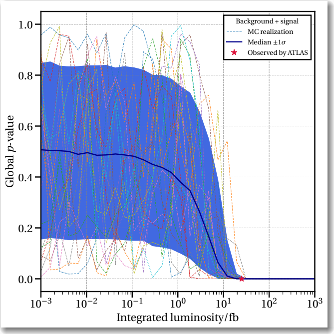
and zero when the signal model is correct, though in this case there is an extremely mild preference for the background model until about 10/fb.

Evolution of p-value and posterior as we collect data



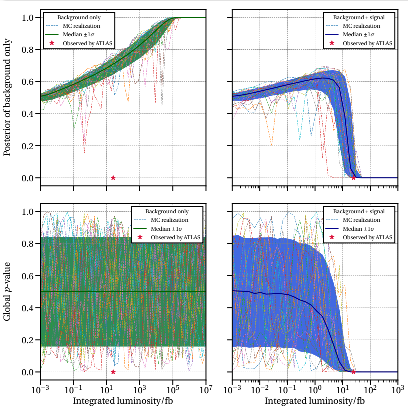
The p-value makes a random walk between 0 and 1 when the background model is correct

Evolution of p-value and posterior as we collect data



and when the signal model is correct, it makes a (noisy) walk towards zero.

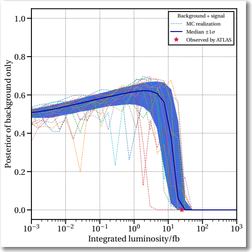
Evolution of p-value and posterior as we collect data



Bayesian (top)/frequentist (bottom). Background model true (left)/signal model true (right).

Preference for the wrong model?

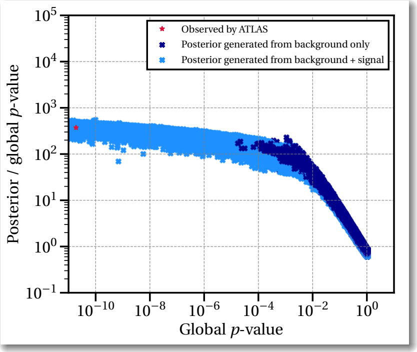
The signal model was the true one but posterior rose from 0.5 to favor the background!



Poisson fluctuations in the background are an economical explanation of signals as $s \lesssim \sqrt{b}$. The signal model requires tuning. Thus mild preference for background model.

Comparison between p-value and posterior

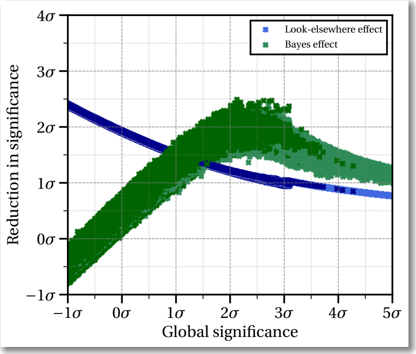
We performed about a million pseudo-experiments.



The posterior of the background model about 10^2 – 10^3 times greater than global p-value!

The Bayes effect

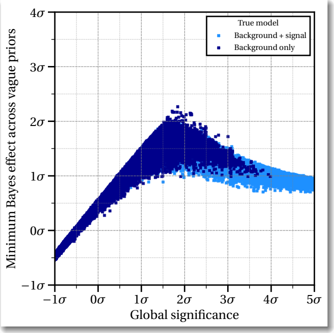
The magnitude of the effect greater than the well-known look-elsewhere effect.



Global significances reduced by $1 - 2\sigma$.

Prior dependence

We checked many priors. The effect could be reduced but remained important.



See paper [17] for full discussion about prior dependence of this effect.

Conclusions

1. Weak evidence for DM from DAMPE
2. Detailed comparison of Bayesian and frequentist methods in resonance searches in toy experiments
3. Posterior ultimately converged to 0 or 1; p-value makes random walk if H_0 correct
4. p-values overstate evidence against the null! p-value $\lll$ posterior of background model
5. Checked that the effect was robust with respect to several choices of prior
6. When looking at an anomaly, we must remember the look-elsewhere effect and the Bayes effect

Bibliography i

[1] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics," Eur. Phys. J. **C70**, 525–530 (2010), arXiv:1005.1891.

[2] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," Eur. Phys. J. **C71**, [Erratum: Eur. Phys. J. **C73**, 2501 (2013)], 1554 (2011), arXiv:1007.1727.

[3] R. E. Kass and A. E. Raftery, "Bayes factors," J. Am. Stat. Assoc. **90**, 773–795 (1995).

[4] J. Skilling, "Nested sampling for general Bayesian computation," Bayesian Anal. **1**, 833–859 (2006).

[5] R. E. Kass and L. Wasserman, “The selection of prior distributions by formal rules,” J. Am. Stat. Assoc. **91**, 1343–1370 (1996).

[6] B. De Finetti, *Theory of Probability. A Critical Introductory Treatment*, Probability and Statistics (Wiley, 1979).

[7] L. J. Savage, *The Foundations of Statistics*, (Dover, 1972).

[8] E. T. Jaynes, *Probability Theory: The Logic of Science*, (Cambridge University Press, 2003).

[9] J. M. Bernardo, “Reference Posterior Distributions for Bayesian Inference,” J. Royal Stat. Soc. Series B (Methodological) **41**, 113–147 (1979).

[10] J. O. Berger et al., “An overview of robust Bayesian analysis,” Test **3**, 5–124 (1994).

[11] S. Chib and T. A. Kuffner, "Bayes factor consistency," (2016), arXiv:1607.00292.

[12] M. S. Bartlett, "A Comment on D. V. Lindley's Statistical Paradox," Biometrika **44**, 533–534 (1957).

[13] H. Jeffreys, *The Theory of Probability*, Oxford Classic Texts in the Physical Sciences (Oxford University Press, 1939).

[14] D. V. Lindley, "A statistical paradox," Biometrika **44**, 187–192 (1957).

[15] G. Ambrosi et al., "Direct detection of a break in the teraelectronvolt cosmic-ray spectrum of electrons and positrons," Nature **552**, 63–66 (2017), arXiv:1711.10981.

[16] G. Aad et al., "Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC," Phys. Lett. **B726**, [Erratum: Phys. Lett. **B734**, 406 (2014)], 88–119 (2013), arXiv:1307.1427.

[17] A. Fowlie, "Bayesian and frequentist approaches to resonance searches," (2019), arXiv:1902.03243.