Calculating p-values in Beyond-the-Standard-Model global fits

Ben Farmer
Imperial College London

Outline

- P-value revision
- Previous approaches to computing p-values in global fits
- Likelihood-based approach
- Summary

P-value revision

P-values

- Probability of data "more extreme", i.e. further from expectations, than that observed
- Basic idea: if the data looks really improbable under some model, then that model is not a good explanation/description of the data.

P-values

Example: sequence of “heads” and “tails”:
HTTHHTHTTHTHTTHHHTTHHTHTHHTTHTHHTTHHHHTHTHTHTHTHTTHHTTHHHTTTHTHTH
HTHTHHTHTHTHHHTHTHTHHTHTHHTHTTHTHTHTHTTHTHHHTHTHTHHTHTHTHTHTHTHTHT
HTHTHTTHTHTHTHTTTTTHHHTTTHTTHTHH

Question: Was this sequence of outcomes generated by fair flips of a fair coin?

P-values

Example: sequence of “heads” and “tails”:
HTTHHTHTTHTHTTHHHTTHHTHTHHTTHTHHTTHHHHTHTHTHTHTHTTHHTTHHHTTTHTHTH
HTHTHHTHTHTHHHTHTHTHHTHTHHTHTTHTHTHTHTTHTHHHTHTHTHHTHTHTHTHTHTHTHT
HTHTHTTHTHTHTHTTTTTHHHTTTHTTHTHH

Question: Was this sequence of outcomes generated by fair flips of a fair coin?

First test one thinks of: number of H vs T.

$$Pr(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

N=165, k=80 (H), N-k=85 (T)
Pr(X>=k) = 0.38
Pr(X<=k) = 0.68

Seems fine.

P-values

But wait… data has other features than just n(H), n(T).

Try number of *runs* in sequence (run is continuous sequence of either H or T, of any length).

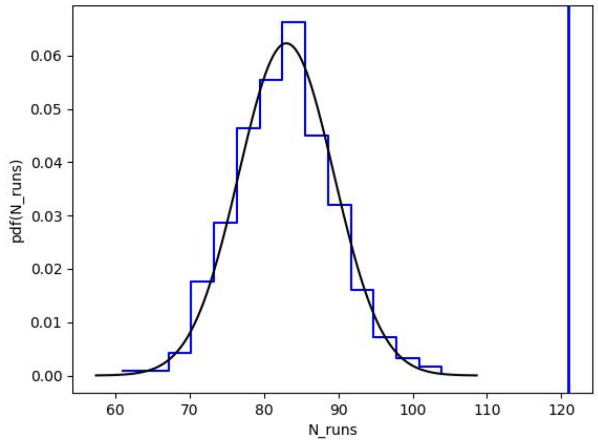
In our sequence (of length 165) there were R=121 runs.

P-values

But wait… data has other features than just n(H), n(T).

Try number of *runs* in sequence (run is continuous sequence of either H or T, of any length).

In our sequence (of length 165) there were R=121 runs.



E(R) = 1 + (n-1)2p(1-p) = 83

Var(R) = 2p(1-p)(2n-3-2p(1-p)(3n-5)) = (6.4)^2

Using Gaussian approximation for distribution of R we have

Pr(r>=R) ~ 1.5e-09 ~ 5.9 sigma

P-values

But wait… data has other features than just n(H), n(T).

Try number of *runs* in sequence (run is continuous sequence of either H or T, of any length).

In our sequence (of length 165) there were R=121 runs.

$E(R) = 1 + (n-1)2p(1-p) = 83$

$Var(R) = 2p(1-p)(2n-3-2p(1-p)(3n-5)) = (6.4)^2$

Using Gaussian approximation for distribution of R we have

$Pr(r>=R) \sim 1.5e-09 \sim 5.9$ sigma

Way too many runs, as it turns out! And I've been caught out, since I just manually generated the sequence by tapping H and T on my keyboard "randomly".

P-values

The point:

P-values are computing based on some *summary statistic*, which we can use to *order* the possible realisations of the data in some way, to define what is “surprising” and what isn’t.

But in general there are many ways to summarise the data!

P-values

The point:

P-values are computing based on some *summary statistic*, which we can use to *order* the possible realisations of the data in some way, to define what is "surprising" and what isn't.

But in general there are many ways to summarise the data!

Is there a "best" way? Sometimes yes:

Neyman-Pearson Lemma:

"For simple vs simple (i.e. no parameters), the likelihood ratio is the most powerful test"

$$\Lambda(x) = \frac{L(x|H_0)}{L(x|H_1)}$$

P-values

Neyman-Pearson Lemma:

“For simple vs simple (i.e. no parameters), the likelihood ratio is the most powerful test”

$$\Lambda(x) = \frac{L(x|H_0)}{L(x|H_1)}$$

P-values

Neyman-Pearson Lemma:

$$\Lambda(x) = \frac{L(x|H_0)}{L(x|H_1)}$$

"For simple vs simple (i.e. no parameters), the likelihood ratio is the most powerful test"

Power = probability to exclude H0 if H1 is true

Kind of intuitive: L(x|H1) is helping order the data specifically based on what we expect to see under H1.

P-values

Neyman-Pearson Lemma:

$$\Lambda(x) = \frac{L(x|H_0)}{L(x|H_1)}$$

"For simple vs simple (i.e. no parameters), the likelihood ratio is the most powerful test"

Power = probability to exclude H0 if H1 is true

Kind of intuitive: L(x|H1) is helping order the data specifically based on what we expect to see under H1.

Downside: only works for simple hypotheses. In general there is no "uniformly most powerful test" (essentially because the best test varies depending on the true parameter values)

But likelihood ratio tests turn out to have other useful properties, so they are commonly used despite not being always most optimal for composite hypotheses. They are a good general-purpose tool.

BSM global fits

Goal: fit the parameters of some new physics model to all relevant experimental data

Compute: Joint likelihood function:

$$L(x, y, z|\theta) = L(x|\theta) \times L(y|\theta) \times L(z|\theta)$$

“Look-elsewhere effect”

Related concepts: “trial correction”, “p-hacking”, “data-dredging”, “cherry-picking”

Basic idea:

If you do many different sorts of p-value calculation, and then pick the one with the lowest p-value after looking at the data, you have just screwed up the frequentist properties of your test procedure and your p-value isn’t the right number anymore.

"Look-elsewhere effect"

Coin example: In each trial, take the lowest p-value out of the "H/T" test and the "N_runs" test.

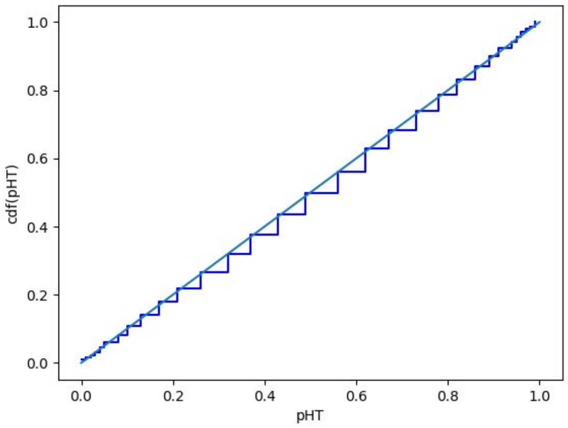
I.e $$p = \min(p_{H/T}, p_{N_{\mathrm{runs}}})$$

“Look-elsewhere effect”

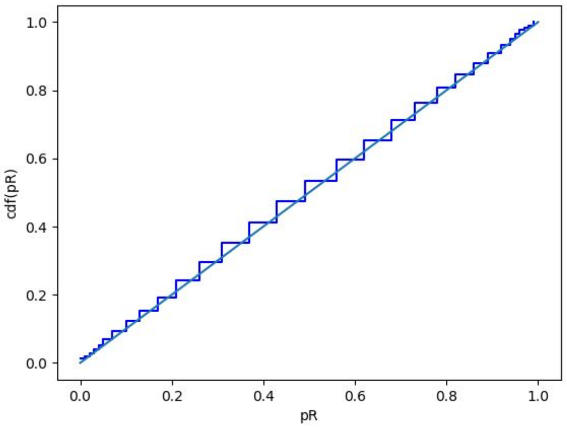
Coin example: In each trial, take the lowest p-value out of the “H/T” test and the “N_runs” test.

I.e $$p = \min(p_{H/T}, p_{N_{\text{runs}}})$$

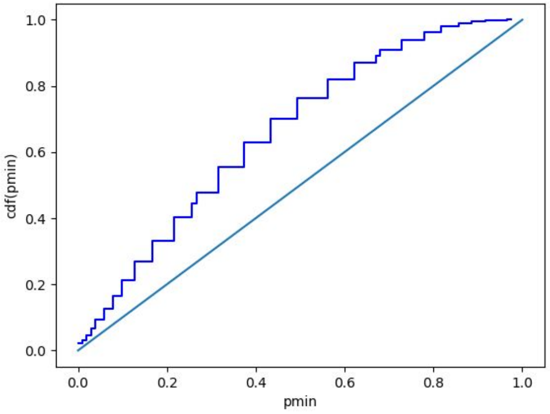
CDF H/T p-value



CDF runs p-value



CDF min p-value



BSM global fits

BSM global fits

Goal: fit the parameters of some new physics model to all relevant experimental data

Compute: Joint likelihood function:

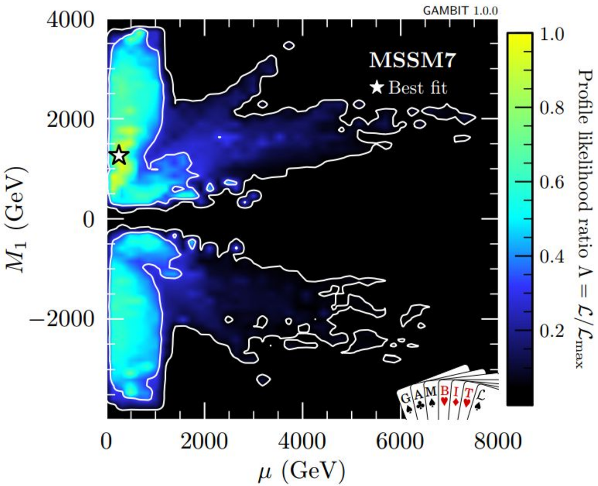
$$L(x, y, z|\theta) = L(x|\theta) \times L(y|\theta) \times L(z|\theta)$$

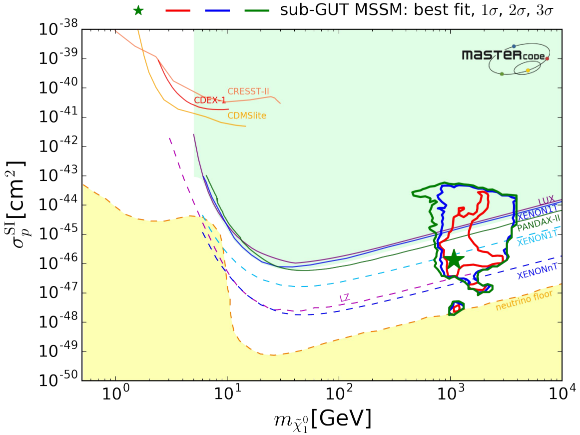
BSM global fits

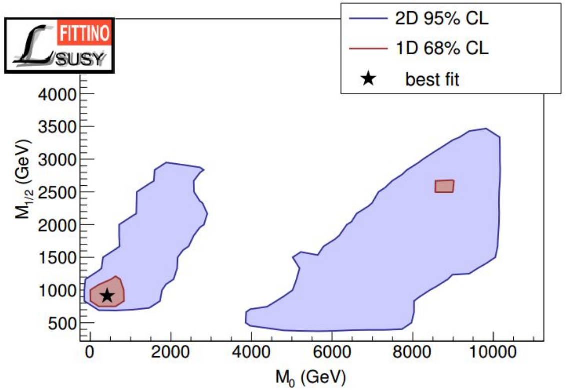
Goal: fit the parameters of some new physics model to all relevant experimental data

Compute: Joint likelihood function:

$$L(x, y, z|\theta) = L(x|\theta) \times L(y|\theta) \times L(z|\theta)$$







GAMBIT arxiv:1705.07917, MasterCode arXiv:1711.00458, Fittino arxiv:1508.05951

BSM global fits

$$L(x, y, z|\theta) = L(x|\theta) \times L(y|\theta) \times L(z|\theta)$$

Confidence intervals computed via (composite) likelihood ratio test:

$$\Lambda(x, y, z) = \frac{L(x, y, z|\theta_0, \hat{\eta})}{L(x, y, z|\hat{\hat{\theta}}, \hat{\hat{\eta}})}$$

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

BSM global fits

$$L(x,y,z|\theta) = L(x|\theta) \times L(y|\theta) \times L(z|\theta)$$

Confidence intervals computed via (composite) likelihood ratio test:

$$\Lambda(x,y,z) = \frac{L(x,y,z|\theta_0,\hat{\eta})}{L(x,y,z|\hat{\hat{\theta}},\hat{\hat{\eta}})}$$

Wilks' theorem: under certain regularity conditions, -2log(Λ) is asymptotically distributed a χ^2_ν , where the degrees of freedom are equal to the dimension of θ_0 .

Likelihood ratio -> p-values -> confidence regions.

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

BSM global fits

Problems:

- Only asymptotically valid, and regularity conditions can be violated (leads to coverage problems, e.g. Bridges et. al. arxiv:1011.4306, Akrami et. al. 1011.4297)

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

BSM global fits

Problems:

- Only asymptotically valid, and regularity conditions can be violated (leads to coverage problems, e.g. Bridges et. al. arxiv:1011.4306, Akrami et. al. 1011.4297)
- Says nothing about “absolute” goodness-of-fit (“goodness” measured against best fit, but best fit may not be a “good” fit)

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

BSM global fits

Problems:

- Only asymptotically valid, and regularity conditions can be violated (leads to coverage problems, e.g. Bridges et. al. arxiv:1011.4306, Akrami et. al. 1011.4297)
- Says nothing about "absolute" goodness-of-fit ("goodness" measured against best fit, but best fit may not be a "good" fit)
- Not well suited for "discovery" tests (i.e. for excluding Standard Model), since usually the Standard Model is non-nested with respect to the new physics.

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

BSM global fits

Problems:

- Only asymptotically valid, and regularity conditions can be violated (leads to coverage problems, e.g. Bridges et. al. arxiv:1011.4306, Akrami et. al. 1011.4297)
- Says nothing about “absolute” goodness-of-fit (“goodness” measured against best fit, but best fit may not be a “good” fit)
- Not well suited for “discovery” tests (i.e. for excluding Standard Model), since usually the Standard Model is non-nested with respect to the new physics.

All three of these problems can be attacked via improvements to p-value calculations.

arxiv:1705.07917, arXiv:1711.00458, arxiv:1508.05951

A brief history of p-value calculations in SUSY global fits

Baer, Balazs - arxiv:0303114

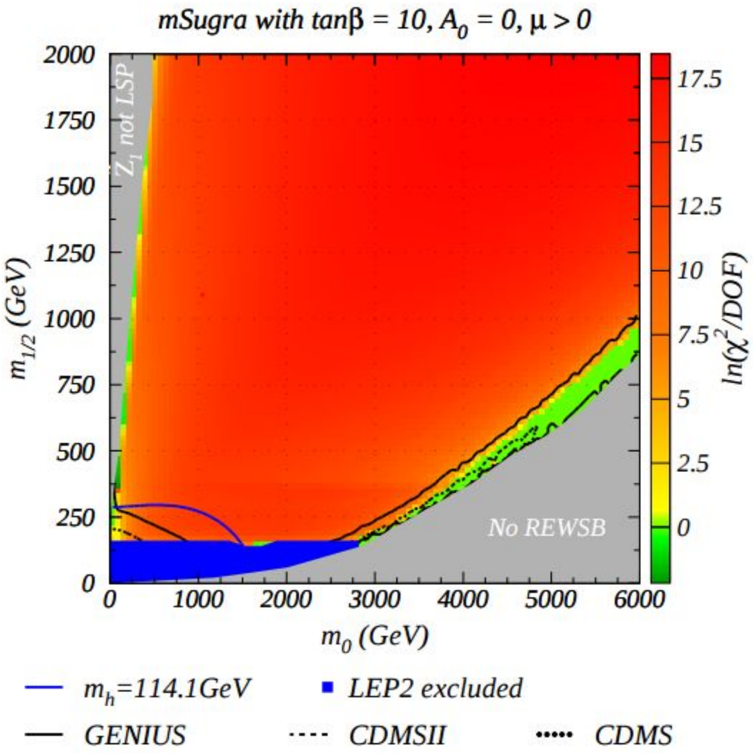
Not explicitly a p-value calculation, but equivalent.

- N measurements of different observables, e.g.

$$BF(b \to s\gamma) = (3.25 \pm 0.54) \times 10^{-4}$$

- Treat measurements as Normal random variables
- Sum of N normal random variables Q $\sim \chi^2_\nu$ with N DOF.

$$Q = \sum_{i=1}^{N} \left(\frac{\mu_i(\theta) - X_i}{\sigma_i} \right)^2$$



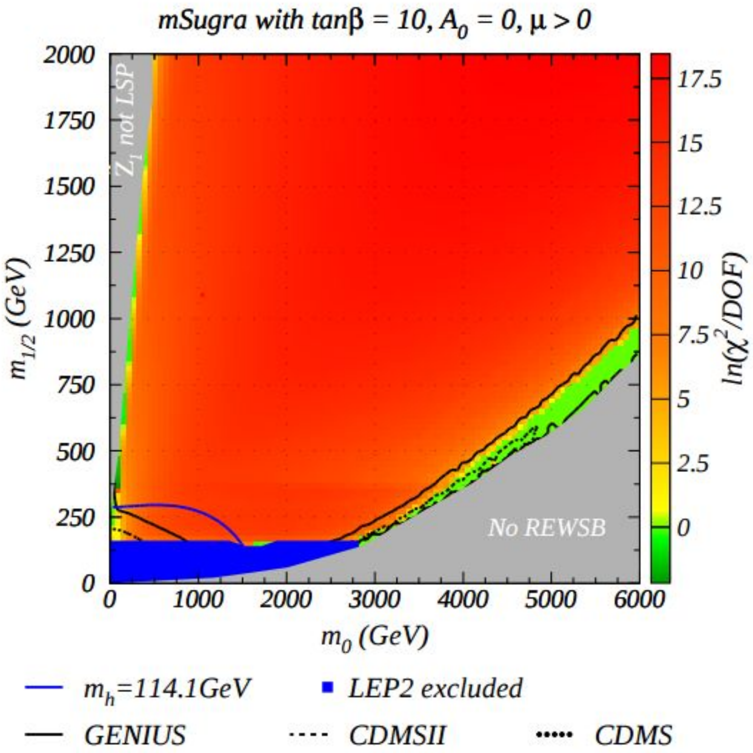
A brief history of p-value calculations in SUSY global fits

- Only thing missing was explicit calculation of a p-value from this.

Problems:

- Only works for normal random variables
  - E.g. limits don't work (bound on dark matter relic density, bounds on sparticle masses from LEP, etc)
  - Cannot deal with more detailed descriptions of experimental data ("proper" likelihoods).
  - Implicitly relies on asymptotic normality of MLEs.
  - Can't really include systematic uncertainties properly.

- "Local" only (no look-elsewhere correction)



A brief history of p-value calculations in SUSY global fits

This basic framework is still used today, e.g. MasterCode arxiv.org:1711.0045

χ^2 without HiggsSignals	28.86	18.02
Number of d.o.f.	24	23
p-value	23%	76%

Table 4

The spectra at the best-fit points including the LHC 13-TeV data and including (left column) or dropping (right column) the $(g-2)_\mu$ constraint. The masses are quoted in GeV. The three bottom lines give the values of the χ^2 function dropping HiggsSignals, *the numbers of degrees of freedom (d.o.f.) and the corresponding p-values.*

A brief history of p-value calculations in SUSY global fits

A fairly heroic attempt to do better was arxiv:1508.05951 (Fittino group, “Killing the cMSSM softly”)

Procedure was:

- Do a global fit via a chi^2 function as previously.

A brief history of p-value calculations in SUSY global fits

A fairly heroic attempt to do better was arxiv:1508.05951 (Fittino group, “Killing the cMSSM softly”)

Procedure was:

- Do a global fit via a chi^2 function as previously.
- Under the hypothesis that the best fit point found is the “true model”, generate pseudo-data (i.e. simulate) the input observables many times.

A brief history of p-value calculations in SUSY global fits

A fairly heroic attempt to do better was arxiv:1508.05951 (Fittino group, “Killing the cMSSM softly”)

Procedure was:

- Do a global fit via a chi^2 function as previously.
- Under the hypothesis that the best fit point found is the “true model”, generate pseudo-data (i.e. simulate) the input observables many times.
- For all model points previously sampled, recompute chi^2 for all the pseudo-data, and find the smallest.

A brief history of p-value calculations in SUSY global fits

A fairly heroic attempt to do better was arxiv:1508.05951 (Fittino group, “Killing the cMSSM softly”)

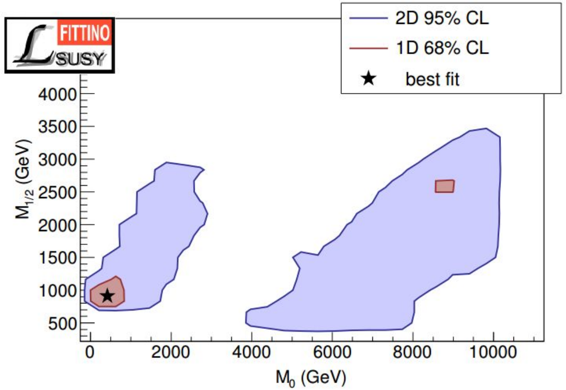
Procedure was:

- Do a global fit via a chi^2 function as previously.
- Under the hypothesis that the best fit point found is the “true model”, generate pseudo-data (i.e. simulate) the input observables many times.
- For all model points previously sampled, recompute chi^2 for all the pseudo-data, and find the smallest.
- Define p-value as:

$$Pr(Q_{\mathrm{min}} < Q_{\mathrm{min,obs}}) \approx \frac{n}{N_{\mathrm{trials}}}$$

where n is the number of pseudo-data trials in which Q_min < Q_min_obs

$$Q = \sum_{i=1}^{N} \left(\frac{\mu_i(\theta) - X_i}{\sigma_i} \right)^2$$



A brief history of p-value calculations in SUSY global fits

Pros:

- Numerical determination of test-statistic distribution
- “Global” goodness of fit test (“calibrates” the distribution of minimum local p-value found in scan: c.f. the minimum p-value test in coin-flipping example)

A brief history of p-value calculations in SUSY global fits

Pros:

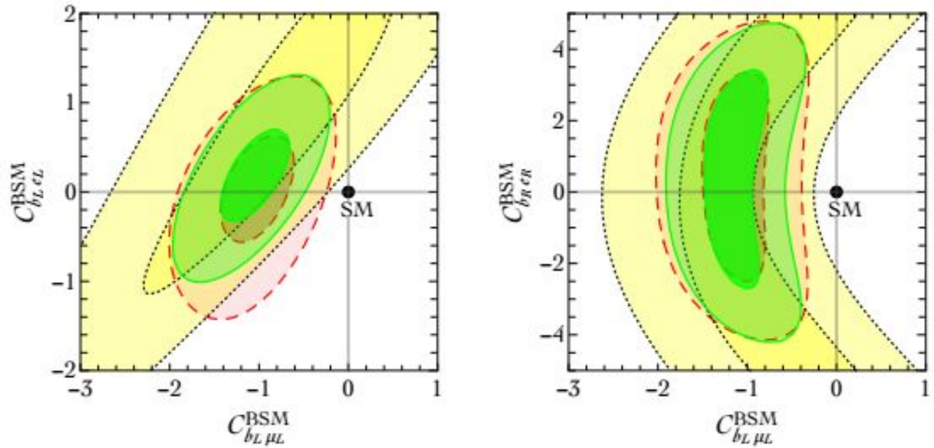
- Numerical determination of test-statistic distribution
- “Global” goodness of fit test (“calibrates” the distribution of minimum local p-value found in scan: c.f. the minimum p-value test in coin-flipping example)

Cons:

- Not true re-fit; only Markov Chain samples from original fit are used in the re-fit
- Extremely computationally expensive (~850 million MCMC samples collected for 5 parameter model)
    - Required in order to get good enough coverage for pseudo-data refitting
- Still no detailed control of systematics (Gaussian assumptions)
- Still not easy to include “proper” experimental likelihood functions
- Not going to work for e.g. signal discovery where we want to simulate data under some Standard Model or “background-only” hypothesis, because best fits of that data will not be near the well-sampled parameter region.

A brief history of p-value calculations in SUSY global fits

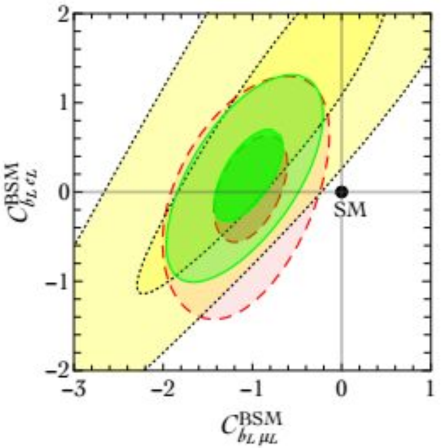
- So far we have only looked at goodness-of-fit tests.
- But there are other kinds of tests one can do. Importantly, “signal searches”! I.e. attempt to exclude some Standard Model hypothesis in favour of New Physics.
- This is a bit tricky in SUSY fits and has not been done. But has been done (and is quite easy to do) in e.g. Flavour global fits:

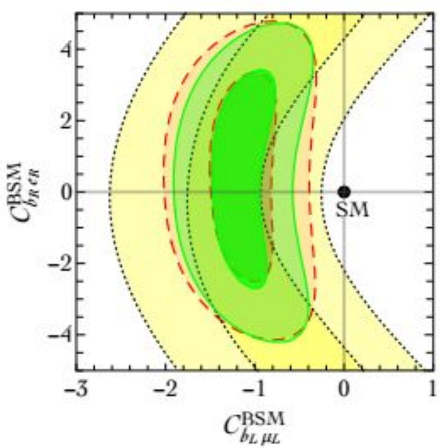


Flavour - arxiv:1704.05438

A brief history of p-value calculations in SUSY global fits

- So far we have only looked at goodness-of-fit tests.
- But there are other kinds of tests one can do. Importantly, “signal searches”! I.e. attempt to exclude some Standard Model hypothesis in favour of New Physics.
- This is a bit tricky in SUSY fits and has not been done. But has been done (and is quite easy to do) in e.g. Flavour global fits:





- Existence of explicit “SM” point in parameter space means that same LR test used to construct confidence intervals can also be used to try and exclude the Standard Model.
- No such point in SUSY models; non-nested.

(though note; still relying on asymptotics here)

Flavour - arxiv:1704.05438

What would be “optimal”?

Some desiderata:

- Full likelihood framework (i.e. handle any pdf for experimental data, deal with systematic uncertainties as the experimentalists do (profiling, marginalising, etc.)
- Goodness-of-fit
- Signal-discovery
- Global significance
- Numerical determination/validation of test statistic distributions
- Asymptotic understanding would be nice
- Numerical feasibility

Test statistic construction

- To start with, we need a test statistic that has power to rule out our null hypothesis when some other hypothesis is true.
- The profile likelihood we use for confidence intervals loses power when the null hypothesis = best fit point
- The "old-school" chi-squared statistic doesn't, but it only works for Gaussian likelihoods

$$\Lambda(x, y, z) = \frac{L(x, y, z|\theta_0, \hat{\eta})}{L(x, y, z|\hat{\hat{\theta}}, \hat{\hat{\eta}})}$$

$$Q = \sum_{i=1}^{N} \left(\frac{\mu_i(\theta) - X_i}{\sigma_i} \right)^2$$

Can we construct a likelihood ratio test that "works" similarly to the "old-school" chi-squared test statistics?

Test statistic construction

Yes! Though not entirely uniquely.

One good candidate: consider joint likelihood for independent normal random variables:

$$L(\boldsymbol{x};\boldsymbol{\mu}) = \prod_{i=1}^{N} \mathcal{N}(x_i;\mu_i,\sigma_i)$$

Consider the profile likelihood ratio:

$$q = -2\left[\sum_{i=1}^{N} \log L(x_i;\mu_{i,0};\sigma_i) - \log L(x_i;\hat{\hat{\mu}}_i;\sigma_i)\right]$$

This turns out to be exactly equivalent to the "old-school" chi-square statistic:

$$q = \sum_{i=1}^{N} \log(2\pi\sigma_i^2) + \frac{(x_i-\mu_{i,0})^2}{\sigma_i^2} - \log(2\pi\sigma_i^2) - \frac{(x_i-\hat{\hat{\mu}}_i)^2}{\sigma_i^2}$$

$$= \sum_{i=1}^{N} \frac{(x_i-\mu_{i,0})^2}{\sigma_i^2}$$

(since $\hat{\hat{\mu}}_i = x_i$)

Test statistic construction

So we can consider extending this sort of likelihood ratio to non-Gaussian likelihoods.

Example: Experiment with multiple correlated Poisson bins (e.g. CMS 2OS lepton search at 13 TeV*)

$$L(\boldsymbol{n},\boldsymbol{X};\boldsymbol{\lambda},\boldsymbol{\theta}) = \left[\prod_{i=1}^{N}\frac{\lambda_i e^{-\lambda_i}}{n_i!}\right]\cdot\mathcal{N}^N(\boldsymbol{X};\boldsymbol{\theta},\Sigma)$$

$$\lambda_i = \mu\cdot s_i + b_i + \theta_i$$

Test statistic:

$$q = -2\log\left(\frac{L(\boldsymbol{n},\boldsymbol{X};\boldsymbol{s},\hat{\boldsymbol{\theta}})}{L(\boldsymbol{n},\boldsymbol{X};\hat{\hat{\boldsymbol{s}}},\hat{\hat{\boldsymbol{\theta}}})}\right)$$

Asymptotic properties: 7 bins, so 7 “s” parameters. Wilks’ theorem -> chi^2 with DOF=7.

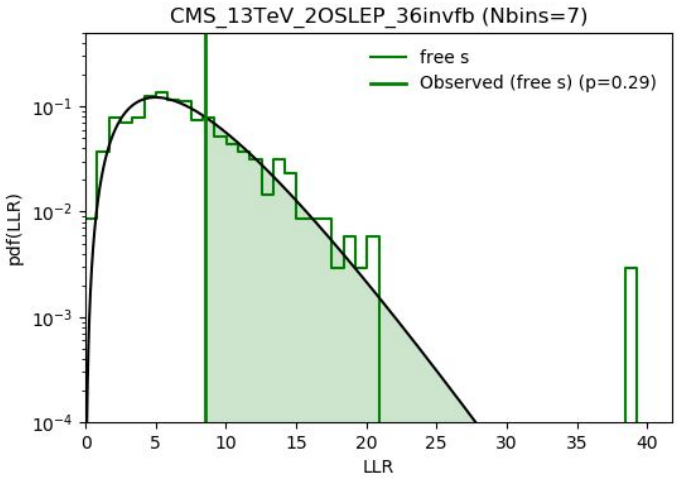
* “Search for new phenomena in final states with two opposite-charge, same-flavor leptons, jets, and missing transverse momentum in pp collisions at s 13 TeV,” JHEP 03 (2018) 076, arXiv:1709.08908

Test statistic construction

This is a slight generalisation of the goodness of fit test proposed by Baker and Cousins (1984)* for fits to histograms.

$$q = -2\log\left(\frac{L(\boldsymbol{n},\boldsymbol{X};\boldsymbol{s},\hat{\hat{\boldsymbol{\theta}}})}{L(\boldsymbol{n},\boldsymbol{X};\hat{\boldsymbol{s}},\hat{\boldsymbol{\theta}})}\right)$$

Has nice, known, asymptotic properties:



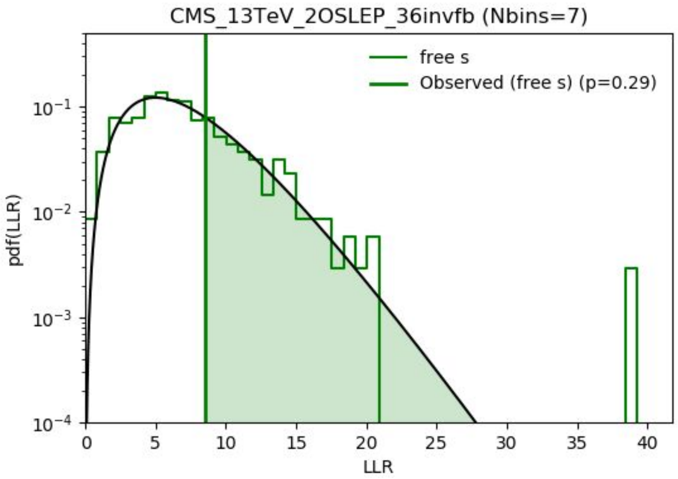
*Baker and Cousins, "Clarification of the use of CHI-square and likelihood functions in fits to histograms" **Nucl.Instrum.Meth. 221** (1984)

Test statistic construction

$$q = -2\log\left(\frac{L(n,X;s,\hat{\theta})}{L(n,X;\hat{s},\hat{\hat{\theta}})}\right)$$

This is a slight generalisation of the goodness of fit test proposed by Baker and Cousins (1984)* for fits to histograms.

Has nice, known, asymptotic properties:



Downside:

- “local” p-value only (doesn’t “know” about space of possible signals in “real” parameter space; we are only testing one fixed “s” hypothesis)
  - But actually, this is exactly the same as the “old-school” GOF test statistic. And we could numerically “correct” it in exactly the same way Fittino did (though at the same computational expense!)

*Baker and Cousins, “Clarification of the use of CHI-square and likelihood functions in fits to histograms” **Nucl.Instrum.Meth. 221** (1984)

Test statistic construction

$$\lambda_i = \mu \cdot s_i + b_i + \theta_i$$

Other options? Well, we could consider a fixed signal *shape*, and vary only a single scaling parameter:

$$q = -2\log\left(\frac{L(\boldsymbol{n},\boldsymbol{X};\mu=0,\hat{\hat{\theta}})}{L(\boldsymbol{n},\boldsymbol{X};\hat{\mu},\hat{\theta})}\right)$$

We could consider testing either mu=1 or mu=0 as our null hypothesis. (chi^2 with DOF=1 in either case)

mu=1 would be another GOF test, but where we consider the GOF of a specific signal shape. Currently I am not sure of the difference this makes relative to the previous case.

mu=0, on the other hand, is quite a different sort of test. This is an attempt to exclude a background-only hypothesis, which hasn't been attempted in SUSY global fits before.

Test statistic construction

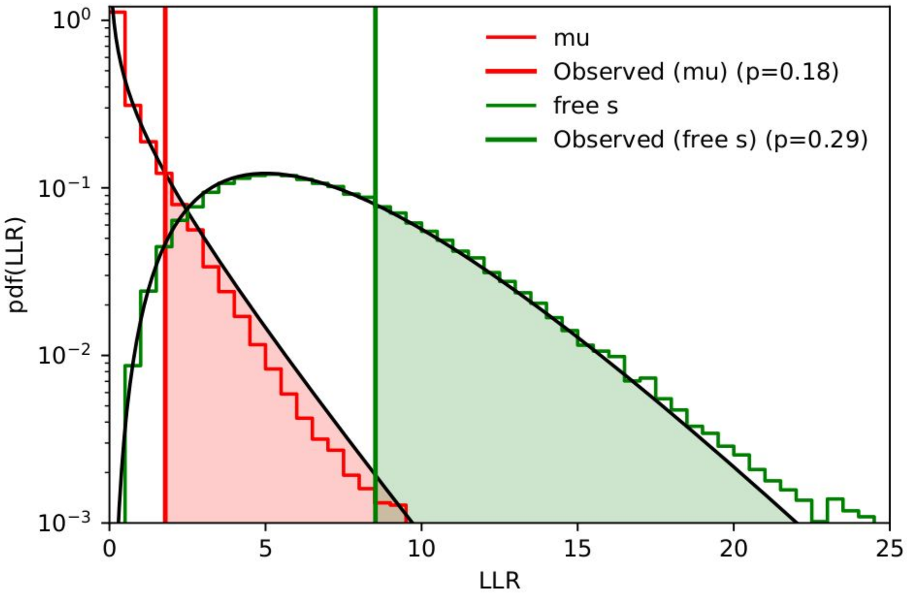
$$\lambda_i = \mu \cdot s_i + b_i + \theta_i$$

Other options? Well, we could consider a fixed signal *shape*, and vary only a single scaling parameter:

$$q = -2\log\left(\frac{L(\boldsymbol{n}, \boldsymbol{X}; \mu = 0, \hat{\theta})}{L(\boldsymbol{n}, \boldsymbol{X}; \hat{\hat{\mu}}, \hat{\hat{\theta}})}\right)$$

Note, this test has the potential to discover signals that the experiments themselves haven't seen!

Why? Power is increased by having explicit alternate hypothesis of given signal shape, e.g. from scan best fit. Looks directly for this signal in data across many experiments.



Test statistic construction

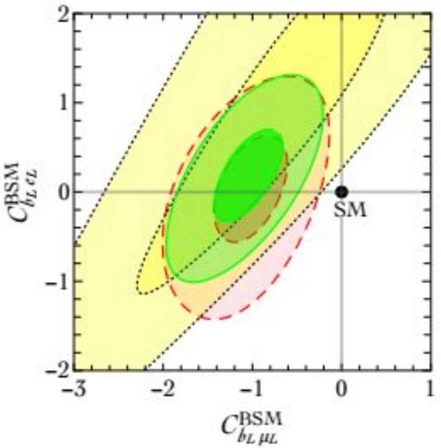
$$\lambda_i = \mu \cdot s_i + b_i + \theta_i$$

Other options? Well, we could consider a fixed signal *shape*, and vary only a single scaling parameter:

$$q = -2\log\left(\frac{L(\boldsymbol{n}, \boldsymbol{X}; \mu = 0, \hat{\theta})}{L(\boldsymbol{n}, \boldsymbol{X}; \hat{\hat{\mu}}, \hat{\hat{\theta}})}\right)$$

But again, only a local p-value. It will be too strong, because we "cherry-pick" the signal shape from our global fit which has "seen" the data.

Correction is even harder this time though, because samples from original global fit are not near mu=0 hypothesis, they are near mu=1. Would need to do (at least one) intensive global fit of background-only data!



Summary

- We can construct test statistics with nice asymptotic properties
    - In a full likelihood-based framework (deal with any sort of experimental pdf, and systematics)
    - For both “goodness of fit” tests and “signal search” tests
- BUT, these are just “local” p-values (although these are all that exist in most of the literature!)
- “Brute force” trial/look-elsewhere corrections are possible (ala Fittino), but extremely expensive numerically

Summary

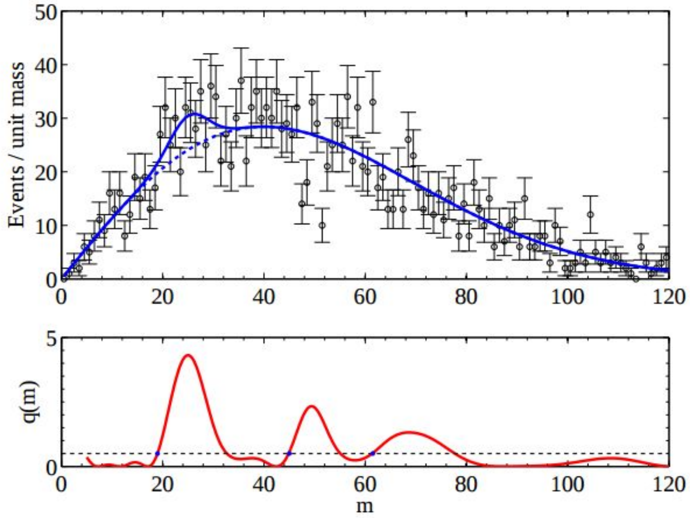
- We can construct test statistics with nice asymptotic properties
  - In a full likelihood-based framework (deal with any sort of experimental pdf, and systematics)
  - For both “goodness of fit” tests and “signal search” tests
- BUT, these are just “local” p-values (although these are all that exist in most of the literature!)
- “Brute force” trial/look-elsewhere corrections are possible (ala Fittino), but extremely expensive numerically

As a first step, I think we will just proceed with computing local p-values, and perhaps consider a less thorough version of the Fittino trial correction.

Future -> easier global p-values?

Epilogue: Speculations on trial corrections

- There exists plenty of literature in experimental HEP on the look-elsewhere effect
- But it is mainly concerned with fairly simple parameter spaces (e.g. search for Higgs boson or WIMP, vary mass). Some tricks exist to compute “trial-correction” factors in these sort of cases:



Gross and Vitells: arxiv:1005.1891 “Trial factors for the look elsewhere effect in high energy physics”

Also some nice discussions in Algeri, van Dyk, Conrad, Anderson, arxiv:1602.03765 “On methods for correcting for the look-elsewhere effect in searches for new physics”

Unfortunately this only works with one free parameter (I think), and requires thorough exploration of the whole parameter space (going to be a problem in higher dimensions)

Epilogue: Speculations on trial corrections

In fact, there exists a lot of statistics literature on this sort of thing, but I have yet to find something that I can see how to apply in our case.

Probably need to consult with some statisticians! I suspect that there is no easy and accurate solution, but there might exist approximate/conservative solutions.