Getting the most out of particle physics experiments

Kyle Cranmer et al. (Sept. 2021). In: arXiv: 2109.04981 [hep–ph]

Andrew Fowlie

27 October 2021

Nanjing Normal University

Overview

1. Theory
2. Practical considerations and tools
3. Applications

Theory

Experiments

My background is beyond the Standard Model physics. An experiment to me could be e.g.

- Measurements of SM processes at the LHC
- Search for new particles at the LHC
- Search for dark matter in underground detectors
- Search for loop corrections from massive new particles in precision measurements

I know LHC also used for its hadron physics capabilities.

Why should we get the most out of them?

The experiments are time consuming and expensive, often once in a generation.

1. We've won the funding — about $10 billion for LHC.
2. We've built the experiment — about ten years for LHC.
3. We've collected the data — about twenty-five years for LHC.
4. What do we need to publish to maximise reuse of the results from the experiment?

Let's get every drop of science we can out of this time and money. That means, enabling other scientists to reuse the experimental results now and in the future.

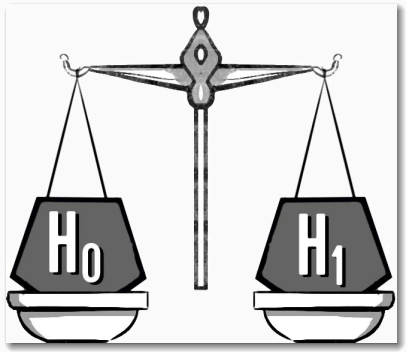
What is an experiment?

Let's focus on individual analyses (e.g. LHC experiment involves thousands of these). In an analysis, we are looking at some portion of the data **x** that could come from a data generating model, M.

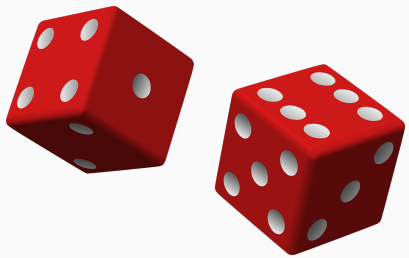
The model typically must describe

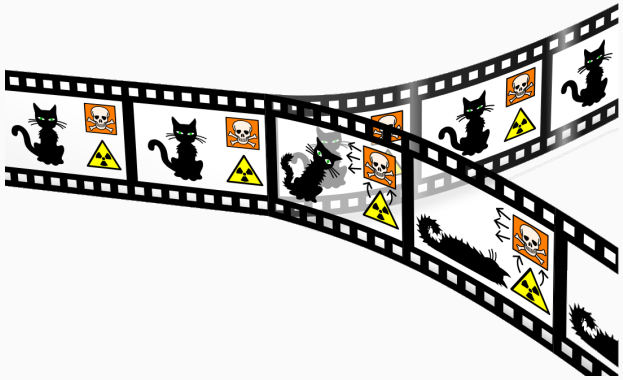
- Underlying physical process, e.g. the Standard Model or just QCD
- Detector and experimental effects

We often want to test models or measure their parameters.



What is an experiment?





The predictions from the model are uncertain because

- Physics fundamentally non-deterministic, e.g. quantum mechanics & quantum field theory
- Uncertainties in computing predictions from the model, e.g. finite order in perturbative calculations etc
- Uncertainties in detector and experimental measurements

What is an experiment?

Because of the uncertainties, we would end up with a random draw from a model if the model were true. We call the mathematics describing the relationship between the random observations and the model the statistical model.

So for our purposes, an experiment is a draw of data that could come from a statistical model of interest and that leads to statistical tests or estimates based on the statistical model

So publish the statistical model and the observed data!

What is a statistical model?

Mathematically, the statistical model may be represented by a probability density function (pdf)

$$p(\mathbf{x} \mid M)$$

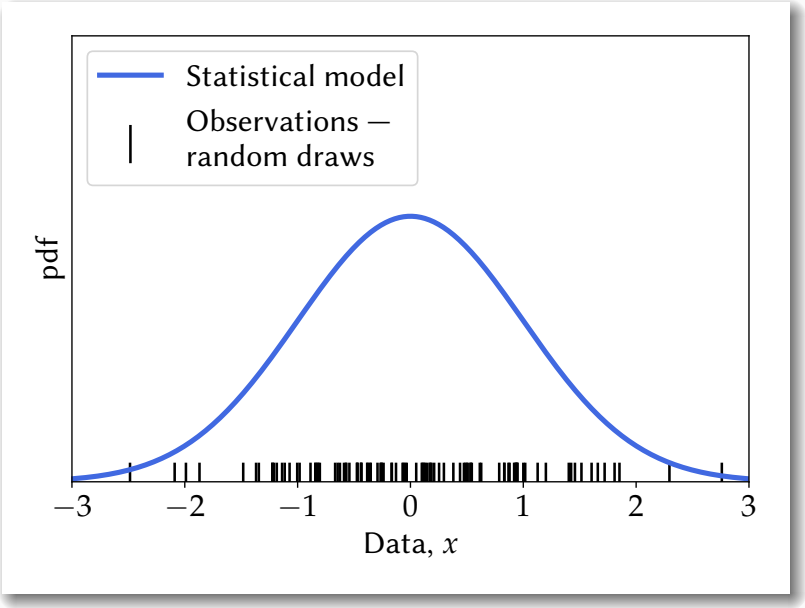
This tells us what data we could get from a model.

What is a statistical model?

E.g., a Gaussian for one-dimensional data

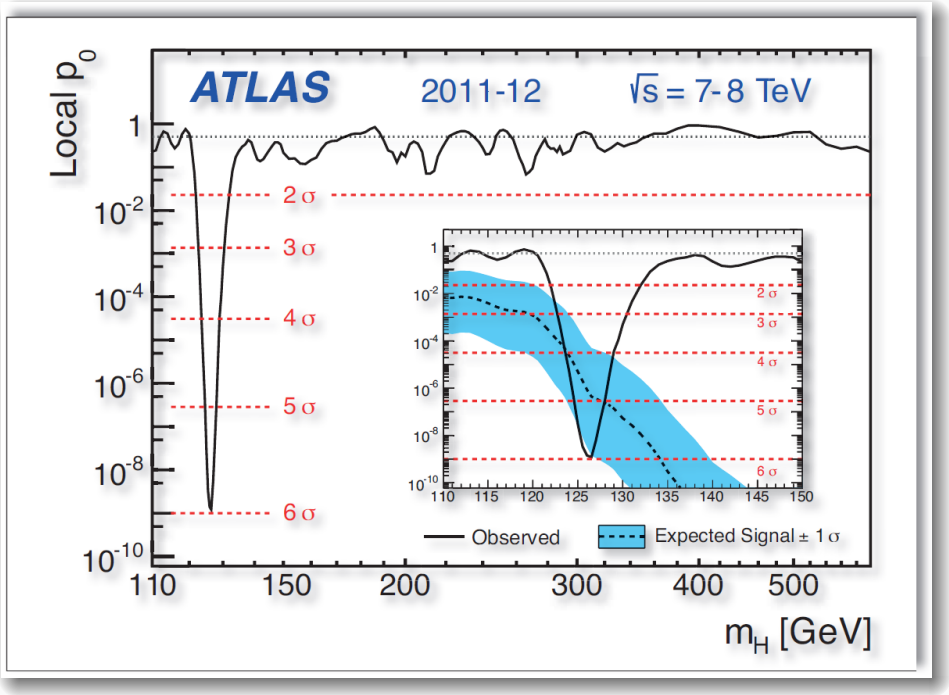
$$p(x \mid M) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

If that model were true, every repetition of the experiment would be another draw from this distribution.



What is an experiment?

Having got the data, we then make our inferences using our statistical models. E.g., computing p-values and confidence intervals are common (Cowan 1998; James 2006).



Higgs discovery p-value $< 10^{-7}$, $m_h = 125.7 \pm 0.7\,\text{GeV}$ etc.

Likelihood principle

Why do we need $p(\mathbf{x} \mid M)$ and $\mathbf{x}_{\text{Observed}}$? Perhaps the likelihood function (Cousins 2020)

$$\text{Likelihood} \equiv p(\mathbf{x} = \mathbf{x}_{\text{Observed}} \mid M)$$

would suffice? After all, we don't need to consider data that weren't observed, do we?

This is the likelihood principle (Berger and Wolpert 1984). However, that intuition and the likelihood principle are violated in frequentist statistics.

Statistical approaches

To allow Bayesian and frequentist approaches, we need the whole statistical model.

- Bayesian approach: condition only on observed data e.g.

$$P(M \mid \mathbf{x} = \mathbf{x}_{\text{Observed}}) \propto p(\mathbf{x} = \mathbf{x}_{\text{Observed}} \mid M)$$

  Automatically satisfies likelihood principle and likelihood alone enough.

- Frequentist approach: compute tail probabilities e.g.,

$$p\text{-value} = \int_{\lambda(\mathbf{x}) \geq \lambda_{\text{Observed}}} p(\mathbf{x} \mid H_0) \, \mathrm{d}\mathbf{x}$$

  for some choice of test-statistic $\lambda(\mathbf{x})$. Automatically violates it and requires whole statistical model.

Statistical approaches

Futhermore, to simulate experiments we need the whole statistical model. This enables

- Computation of p-value without asymptotic assumptions about the statistical model
- Simulation based inference in Bayesian setting, e.g., Approximate Bayesian Computation
- Simulation based cross-checks to validate statistical procedures

For which parameters?

We've agreed that we need to give a mathematical representation of a model that could generate the observed data. But models have unknown parameters!

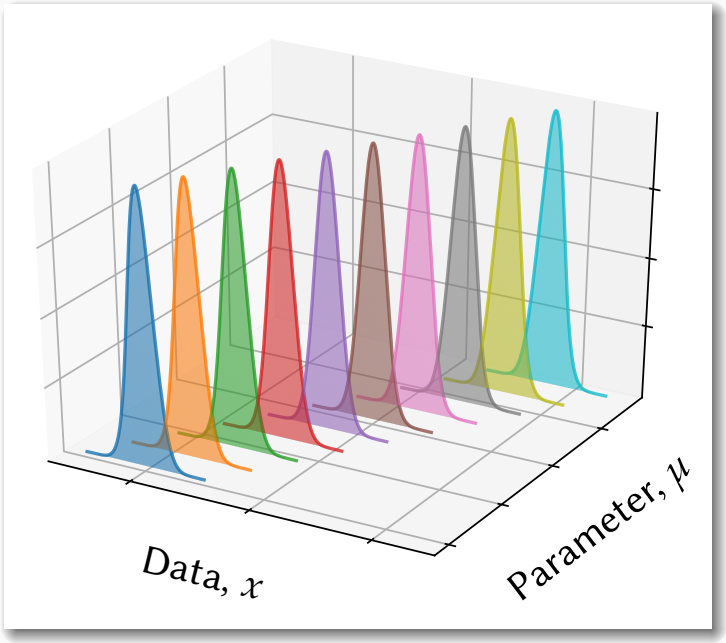
- Parameters of interest, $\boldsymbol{\Theta}$, e.g. the unknown masses and couplings of the particles.
- Nuisance parameters, $\boldsymbol{\Phi}$, describing detector and systematic effects

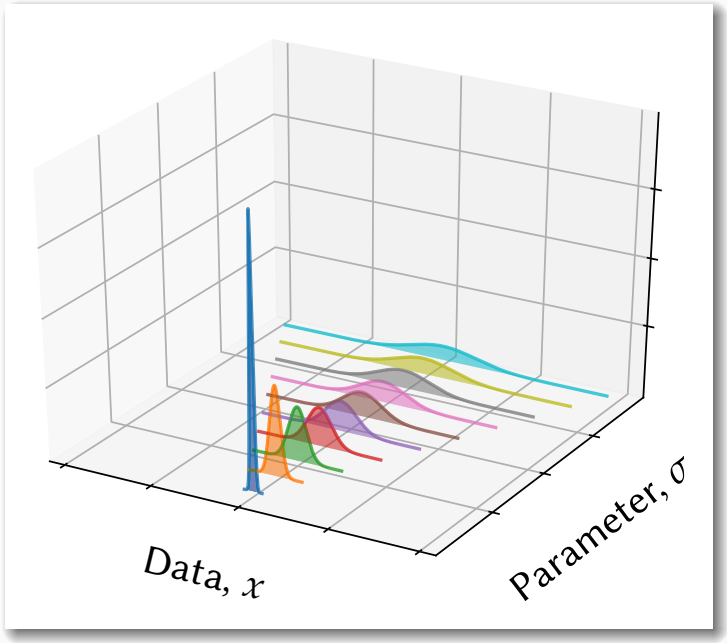
So let's in fact publish the dependence on those parameters too,

$$p(\mathbf{x} \mid M, \boldsymbol{\Theta}, \boldsymbol{\Phi})$$

For which parameters?

Different distribution for each choice of parameters. Need to preserve whole dependence $p(\mathbf{x} \mid M, \boldsymbol{\Theta}, \boldsymbol{\Phi})$





E.g. Gaussian of unknown mean and standard deviation $\mathcal{N}(\mu, \sigma^2)$.

Statistical model describes relationship between potential observations and the unknown parameters of the model.

Which statistical model?

Which statistical model? Physicists are building new models with new particles all the time.

- Could publish more than one statistical model
- Publish "model fragments". Break the model down into distinct components to allow it to be modified to describe new models
- Express the statistical model in a convenient parameterization

Hopefully, we can then go from model a to model model b in the future

$$p(\mathbf{x} \mid M_a, \mathbf{\Theta}_a, \mathbf{\Phi}) \to p(\mathbf{x} \mid M_b, \mathbf{\Theta}_b, \mathbf{\Phi})$$

This is known as recasting or reinterpretation (Abdallah et al. 2020).

Parameterization

In order to aid reinterpretation, we should chose a convenient parameterization.

For many physics models, the statistical model depends on the model and parameters only through several pseudo observables

$$p(\mathbf{x} \mid M, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = p(\mathbf{x} \mid m(\boldsymbol{\Theta}), \sigma(\boldsymbol{\Theta}), \Gamma(\boldsymbol{\Theta}), \boldsymbol{\Phi})$$

E.g., here the pseudo observables are masses m, cross-sections σ and decay widths Γ. So publish

$$p(\mathbf{x} \mid m, \sigma, \Gamma, \boldsymbol{\Phi})$$

Future users can map from their model to these pseudo observables.

Practical considerations and tools

Serialising a statistical model

We need to write the statistical model to disk in a format that can be loaded and reused in the future. There are tools

- Ideally, machine and human readable with an unambiguous declarative specification
- If you work with `root`, publish a `RooWorkspace`
- If you work with binned data, follow the `HistFactory` (Cranmer et al. 2012) specification using e.g. `pyhf` (Heinrich et al. 2021)

See our paper (Cranmer et al. 2021) for more discussion of tools and formats.

Depositing observed data

Data may be stored in online repository (Roche et al. 2014) e.g. HEPData (Whalley 1989; Maguire, Heinrich, and Watt 2017) or Zenodo.

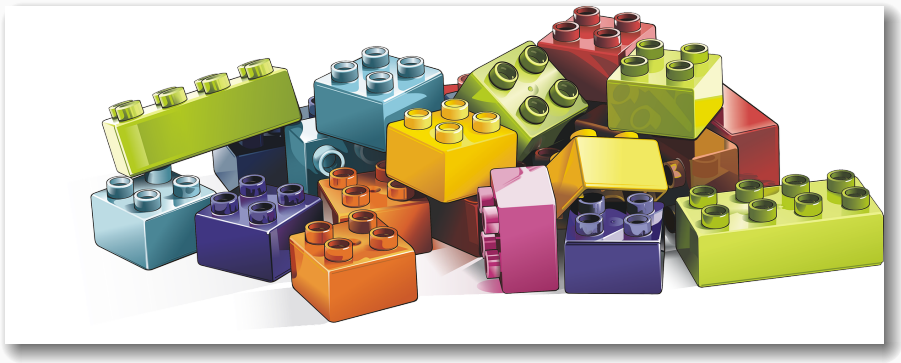
This is now standard practice in high-energy physics. Whatever tools/platforms you use, try to be fair: Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

Open or closed

Technical consideration in representing statistical models:

- Open world: anything goes, build it how you like.
- Closed world: allow models to be built only from a finite set of building blocks. Simple e.g. the model is a linear combination of Gaussian distributions

Open world



Closed world

Applications

Applications

We discuss many applications of published statistical models familiar to us in our paper (Cranmer et al. 2021):

- Parton distribution functions
- Higgs boson measurements at the LHC
- Searches for new particles at the LHC
- Heavy flavor physics
- Searches for dark matter
- World averages
- Global fits

Applications

We discuss many applications of published statistical models familiar to us in our paper (Cranmer et al. 2021):

- Parton distribution functions
- Higgs boson measurements at the LHC
- Searches for new particles at the LHC
- Heavy flavor physics
- Searches for dark matter
- World averages
- Global fits $\rightarrow$ look at this one since I am most familiar with it

Global fits

In a global fit, we want to

- Collate many experimental measurements and searches
- Re-interpret them in particular models of interest
- Find which models and parameters are favoured by data
- Find what the models predict for future experiments

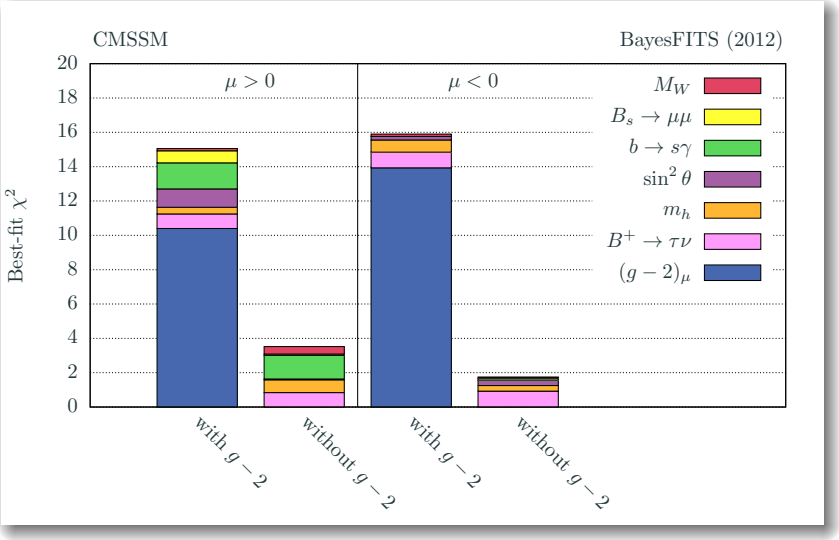
Example of global fit

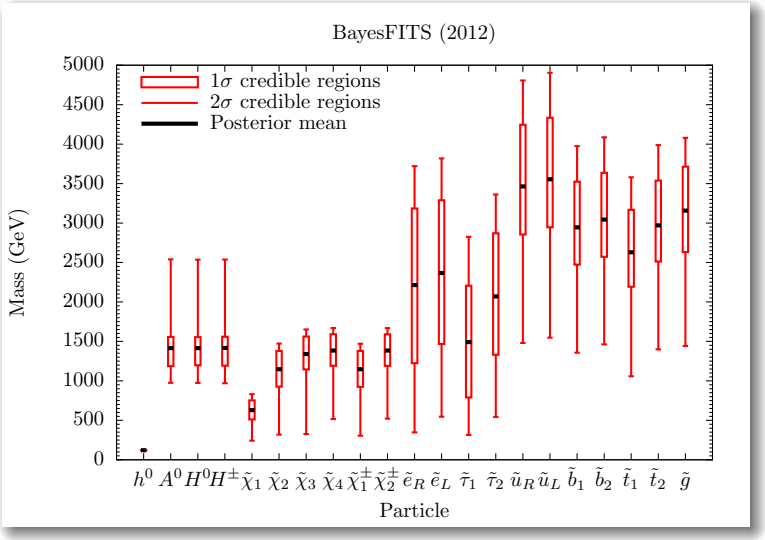
Take the minimal supersymmetric model. Build statistical model from

- Higgs measurements at the LHC
- Direct searches for supersymmetric particles at the LHC
- Dark matter density
- Null results of direct and indirect searches for dark matter
- Electroweak precision observables
- Flavor observables

Fit model to it. Find allowed parameters, masses, etc.

Example of global fit





Goodness of fit and preferred masses in a minimal supersymmetric model (Fowlie et al. 2012) from statistical model built from public data.

Benefit of public statistical models and data

If we must construct statistical models ourselves

- ⚠ Make approximations — could be crude, lose power of data
- ⚠ Make errors — hopefully not but somewhat inevitable, results can be hard to validate
- ⚠ Time consuming — challenging undertaking performed by big teams — see e.g. `GAMBIT` (Athron et al. 2017)

Benefit of public statistical models recognised as early as 2006 (Ruiz de Austri, Trotta, and Roszkowski 2006)

> … it would be easy to incorporate the full likelihood functions from various experimental measurements if they were available. However, even though the actual measurements contain much more useful information, most measurements in particle physics experiments are presented only by the mean and the standard deviation …

Summary

- Let's maximise the impact and future re-use of expensive experiments!
- Statistical models and the observed data should be preserved and published
- There are technical solutions to achieve this
- The parameterization and fragments of the statistical model should be chosen carefully to maximise reuse

References

Abdallah, Waleed et al. (2020). "Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2." In: SciPost Phys. 9.2, p. 022. arXiv: 2003.07868 [hep–ph].

Athron, Peter et al. (2017). "GAMBIT: The Global and Modular Beyond-the-Standard-Model Inference Tool." In: Eur. Phys. J. C 77.11. [Addendum: Eur.Phys.J.C 78, 98 (2018)], p. 784. arXiv: 1705.07908 [hep–ph].

Berger, James and Robert Wolpert (1984). The likelihood principle. English. Vol. 6. Hayward, CA: IMS, Institute of Mathematical Statistics, pp. xi + 206. ISBN: 0-940600-06-4.

References ii

Cousins, Robert D. (Oct. 2020). "What is the likelihood function, and how is it used in particle physics?" In: arXiv: 2010.00356 [physics.data-an].

Cowan, G. (1998). Statistical data analysis. Clarendon Press. ISBN: 978-0-19-850156-5.

Cranmer, Kyle et al. (Sept. 2021). "Publishing statistical models: Getting the most out of particle physics experiments." In: arXiv: 2109.04981 [hep-ph].

Cranmer, Kyle et al. (Jan. 2012). "HistFactory: A tool for creating statistical models for use with RooFit and RooStats." In: CERN-OPEN-2012-016. https://cds.cern.ch/record/1456844.

Fowlie, Andrew et al. (2012). "The CMSSM Favoring New Territories: The Impact of New LHC Limits and a 125 GeV Higgs." In: Phys. Rev. D 86, p. 075010. arXiv: 1206.0264 [hep-ph].

References iii

Heinrich, Lukas et al. (2021). "pyhf: pure-Python implementation of HistFactory statistical models." In: J. Open Source Softw. 6.58, p. 2823.

James, Fred (Jan. 2006). Statistical Methods in Experimental Physics. World Scientific. ISBN: 981-256-795-X.

Maguire, Eamonn, Lukas Heinrich, and Graeme Watt (2017). "HEPData: a repository for high energy physics data." In: J. Phys. Conf. Ser. 898.10. Ed. by Richard Mount and Craig Tull, p. 102006. arXiv: 1704.05473 [hep-ex].

Roche, Dominique G. et al. (Jan. 2014). "Troubleshooting Public Data Archiving: Suggestions to Increase Participation." In: PLOS Biology 12.1, pp. 1–5. URL: https://doi.org/10.1371/journal.pbio.1001779.

References iv

Ruiz de Austri, Roberto, Roberto Trotta, and Leszek Roszkowski (2006). "A Markov chain Monte Carlo analysis of the CMSSM." In: JHEP 05, p. 002. arXiv: hep-ph/0602028.

Whalley, M. R. (1989). "The Durham-RAL high-energy physics databases: HEPDATA." In: Comput. Phys. Commun. 57, pp. 536–537.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." In: Scientific Data 3. ISSN: 2052-4463.