Noise vs. Signal
&
Frequentist hypothesis testing

Why are our HEP-friends frequentists?
And what happens if we steal their methods and apply them in astronomy?

ICIC Data Analysis Workshop 2016

Ln(a) Sellentin
Imperial College London
&
Université de Genève

email: elena.sellentin@posteo.de

Some cases are clearly Bayesian

Some cases are clearly frequentist

- Medical tests:
  - Allergy tests
  - Pregnancy tests
  - Blood tracers for cancer types
  - Effictiveness of medication

True positives	false positives
True negatives	false negatives

Bayesian statistics in a nutshell

- Parameter estimation:

$$P(\boldsymbol{\theta}_M|\boldsymbol{X}) = \frac{P(\boldsymbol{\theta}_M)P(\boldsymbol{X}|\boldsymbol{\theta}_M)}{P(\boldsymbol{X})}$$

- Model comparison:

$$\frac{L(M_1|\boldsymbol{X})}{L(M_2|\boldsymbol{X})} = \frac{\mathcal{P}(M_1)}{\mathcal{P}(M_2)}\frac{\varepsilon_1}{\varepsilon_2}$$

Frequentist statistics in a nutshell

- Frequentist comes from 'frequency'.
- Rely on an actual or hypothetical repetition of an experiment.
- Friends of limit theorems and asymptotics: $for\ N \to \infty$
- Mindset:

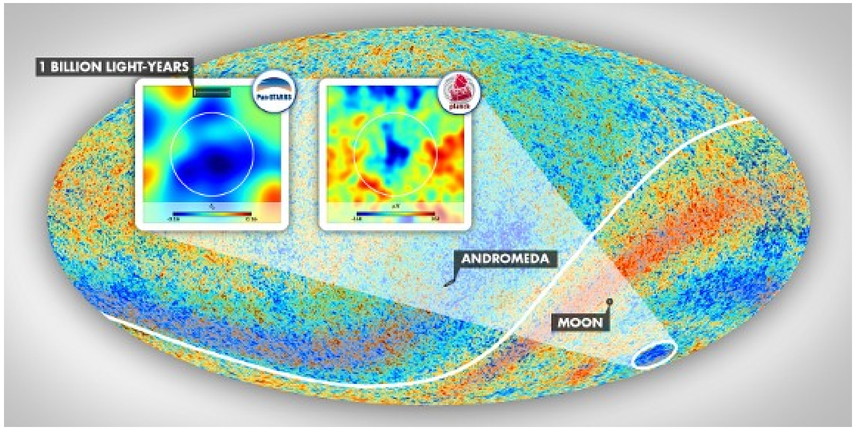
"I have measured the mass of the proton 1 million times.

I always get $1.672621898(21)\times10^{-27}$ kg.

I think if I measure once more, I'll again get $1.672621898(21)\times10^{-27}$ kg."

Frequentist questions to Bayesians

- How exactly do you get these priors?
- Do you really just fit a model, without checking previously that your 'signal' isn't just noise?
- You do know that each time you fit, it is guaranteed that you get an answer? Even if it was just noise?
- How do you get rid of a bad model? Without replacement?



Broadly speaking

- Priors:
  - → null-hypothesis + sampling distribution of test statistics T

$$if\ x_i\ \sim \mathcal{D}(x|\vec{\theta}),\quad then\ T(x) \sim ?,\ hence\ T(x_{obs})...$$

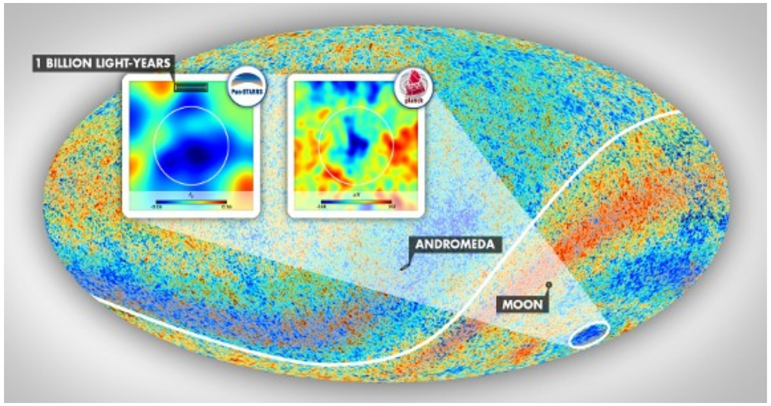
- Model comparisons:
  - → hypothesis rejection & p-values
  - → Likelihood-ratio tests, $\Delta\chi^2$
- Parameter estimation:
  - – quite similar! ML-estimators, LS-estimator & sample estimators

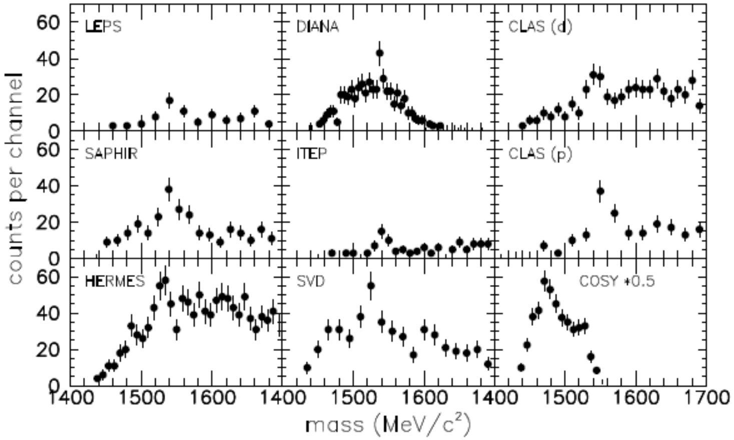
$$\bar{x} = 1/N \sum_i x_i, \qquad \hat{\theta} = argmax[L(\vec{x}|\theta)], \qquad minim[\chi^2]$$

- Inversion of the workflow:
  - – Order of parameter estimation & model/hypothesis selection

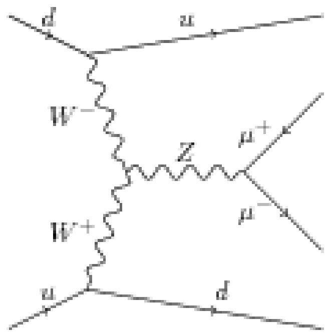
Workflows

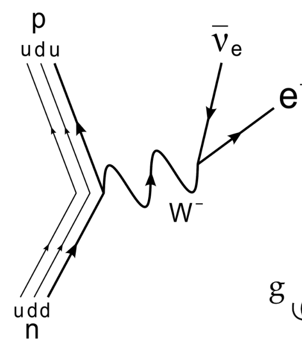
- Astro:
  - 0.) Get data = true signal + noise
  - 1.) select parametric model (decides which 'signal' is in the data)
  - 2.) estimate the model parameters
  - 3.) doubt model, compare it to a competitor model (evidences)
- HEP:
  - 0.) Get data. H_0: no prejudice about potentially hidden signals.
  - 1.) non-parametric model checks: is it maybe still noise? (p-values)
  - 2.) It's not noise!
  - 3.) Select model and estimate its parameters.

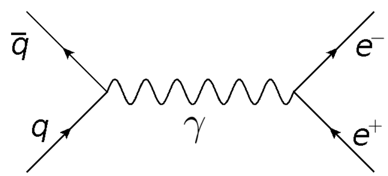


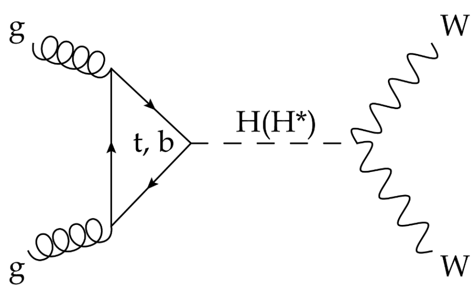


Particle creation **is** frequentist by nature



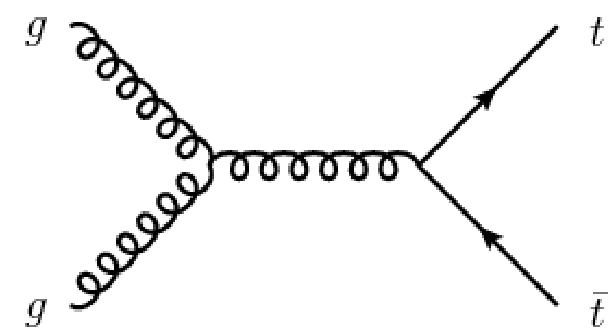


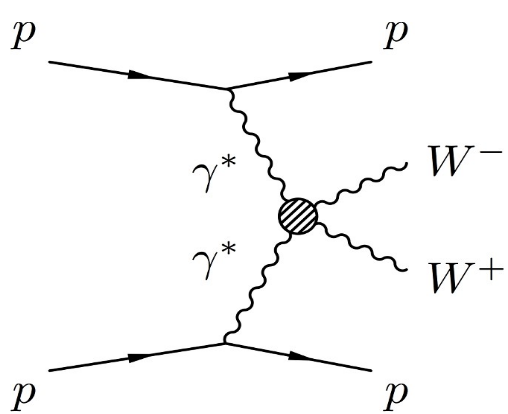


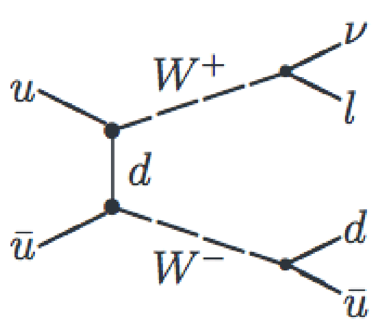


p+p → a lot!

'→': Transition **probabilities**

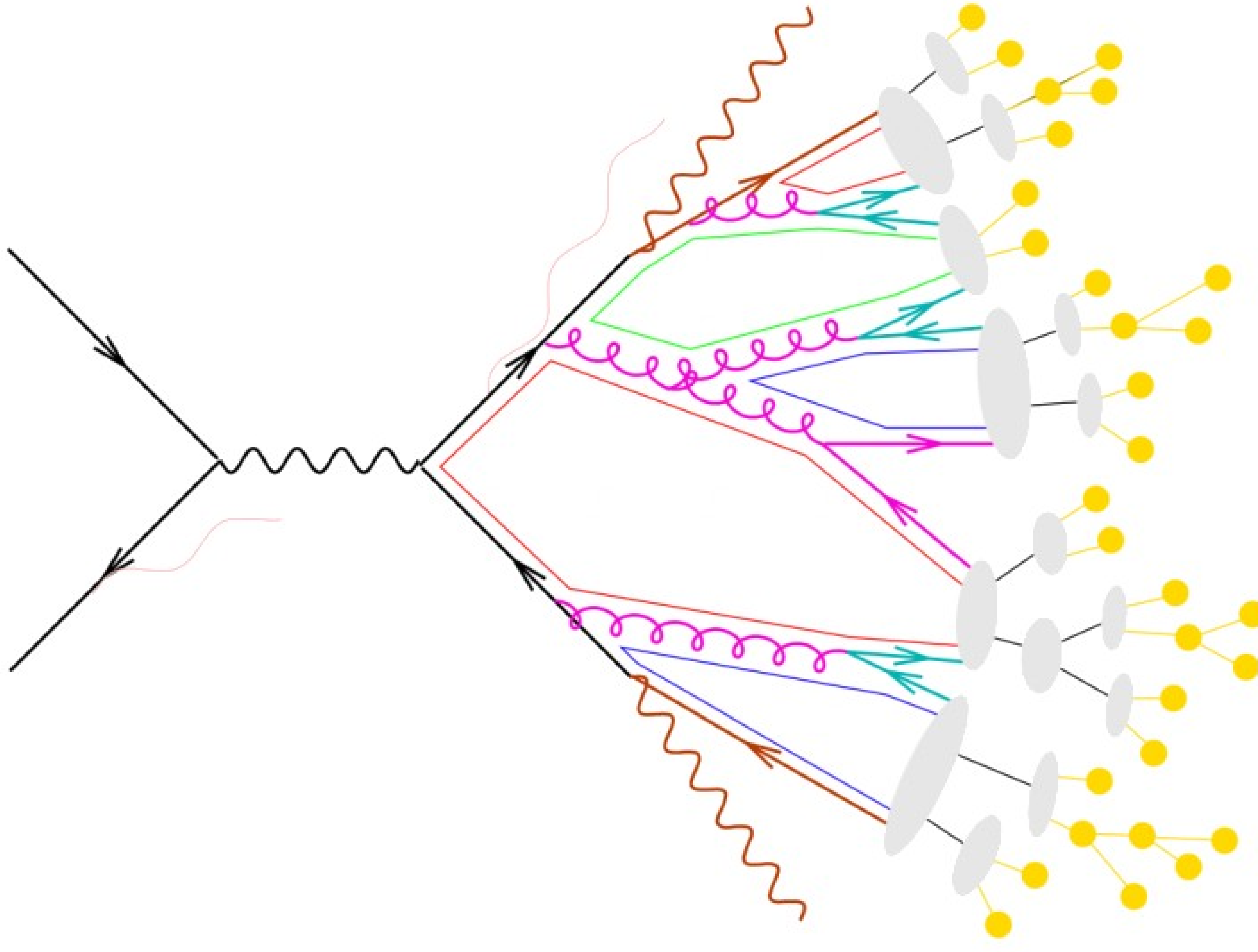




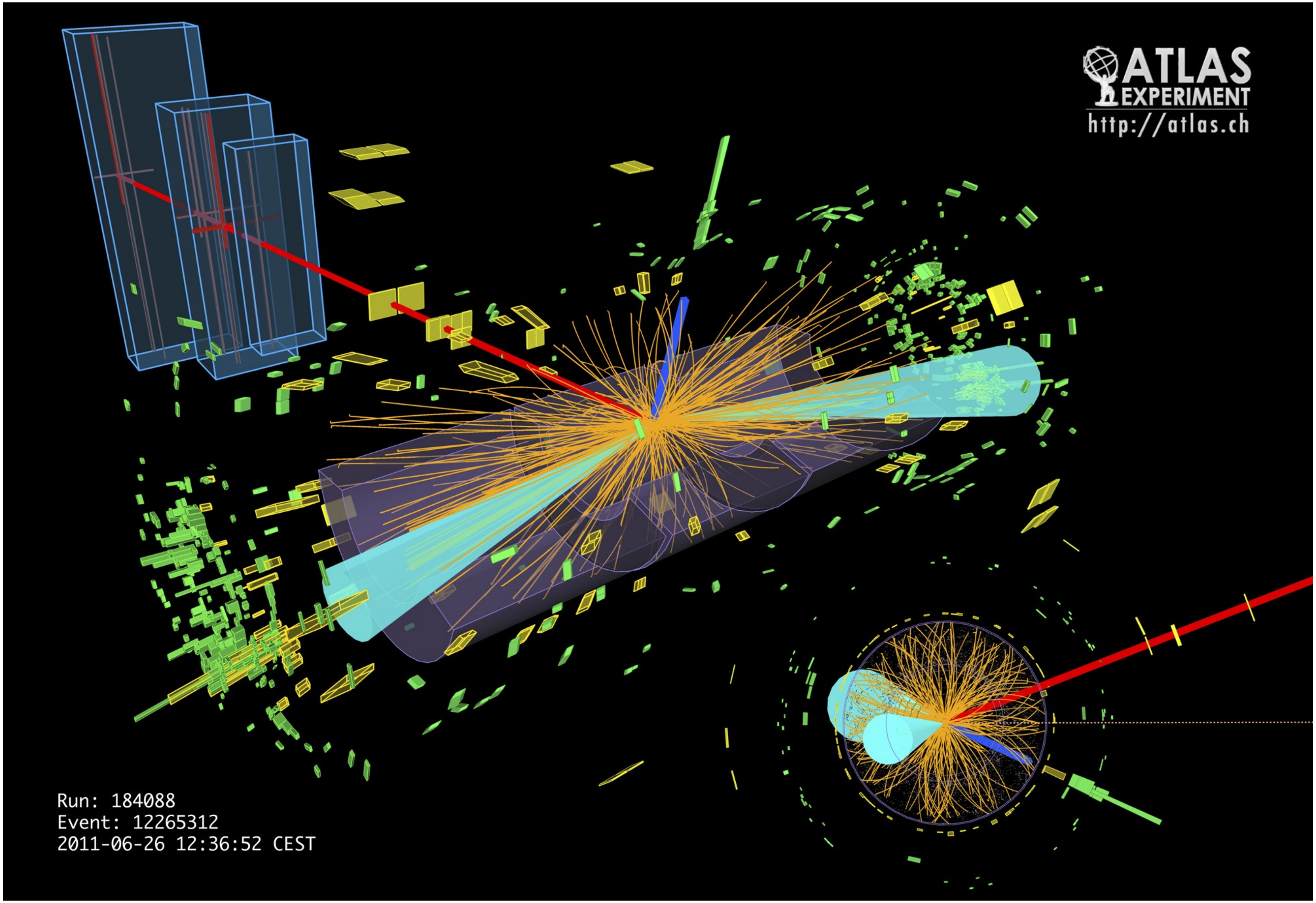


+ many many more...

Particle decay is again frequentist



Detection is (mainly) Frequentist...



…. with some Bayesian nightmares.

Sampling distributions for test statistics

- $if\ x_i\ \sim \mathcal{D}(x|\vec{\theta}),\ \ then\ T(x) \sim ?,\ hence\ T(x_{obs})...$
- Can often be derived analytically:

$$if\ x_i \sim \mathcal{N}(0,1),\ then\ \chi^2 = \sum_i^p x_i^2 \sim \chi_p^2$$

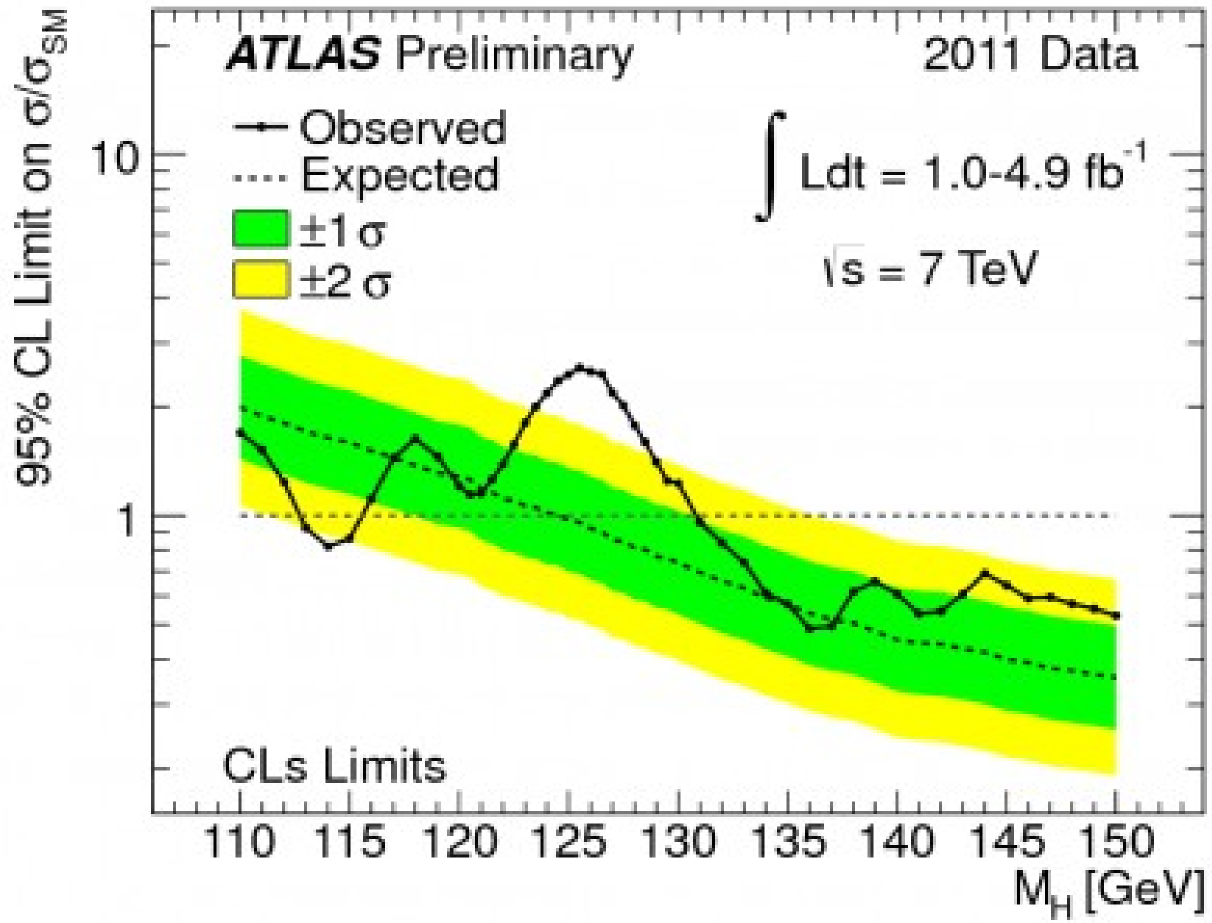
$$if\ u \sim \chi_p^2,\ and\ v \sim \chi_q^2,\ then\ \frac{u/p}{v/q} \sim \mathcal{F}$$

$$if\ x_i \sim \mathcal{G}(\mu,\sigma),\ then\ (\bar{x}-\mu)/(s/\sqrt{n}) \sim\ Student-\mathcal{T}$$

- Else: derive it from Monte Carlo Simulations: $\hat{\theta}_{ML} \sim ?$
- Aim: How typical is my measurement, compared to hypothetical others?

Neyman-construction with H_0

events →

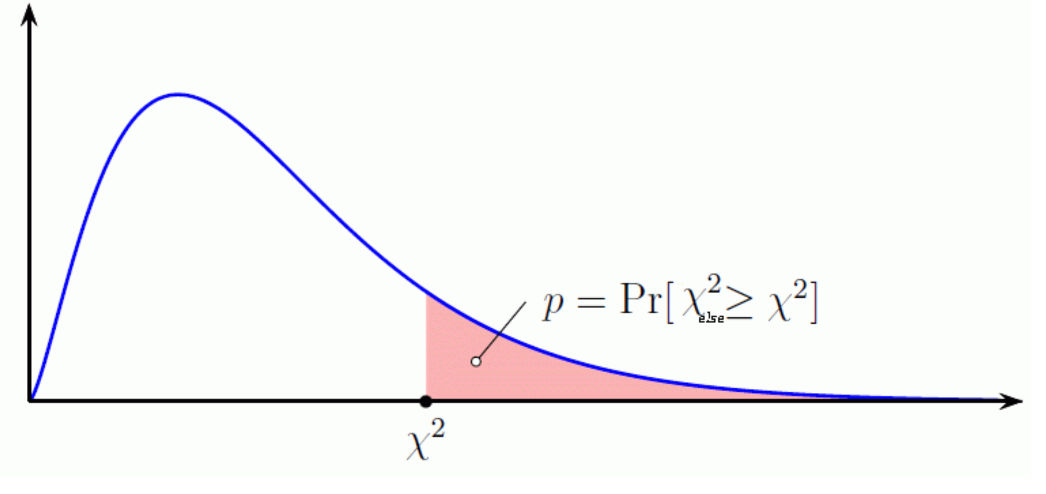


- Monte-Carlo simulations
- Target: sampling distribution

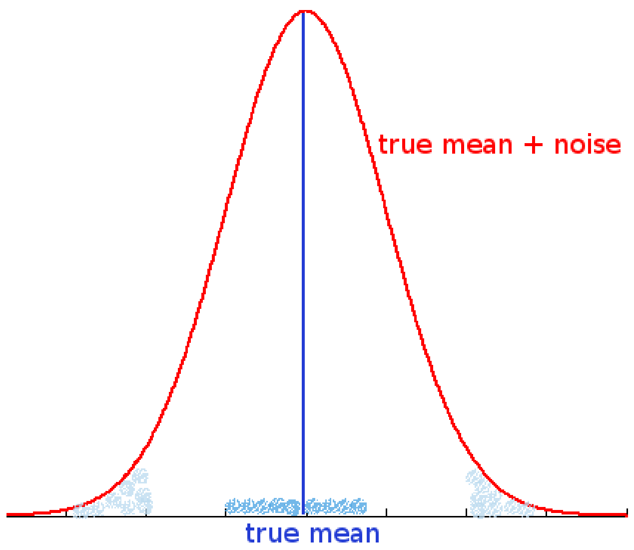
P-values: tails of sampling distributions

$$\vec{x}_{obs} \sim f(\vec{x}|H_o) \qquad then \qquad p = \mathcal{P}[T(\vec{x}) \geq T(\vec{x}_{obs})]$$

- Large values of T typically indicate bad agreement.
- P-value for a large T is then **small**.
- For continuous sampling distributions: p-values are upper-tail integrals.
  → the sampling distribution affects your p-value.
- Example: χ^2

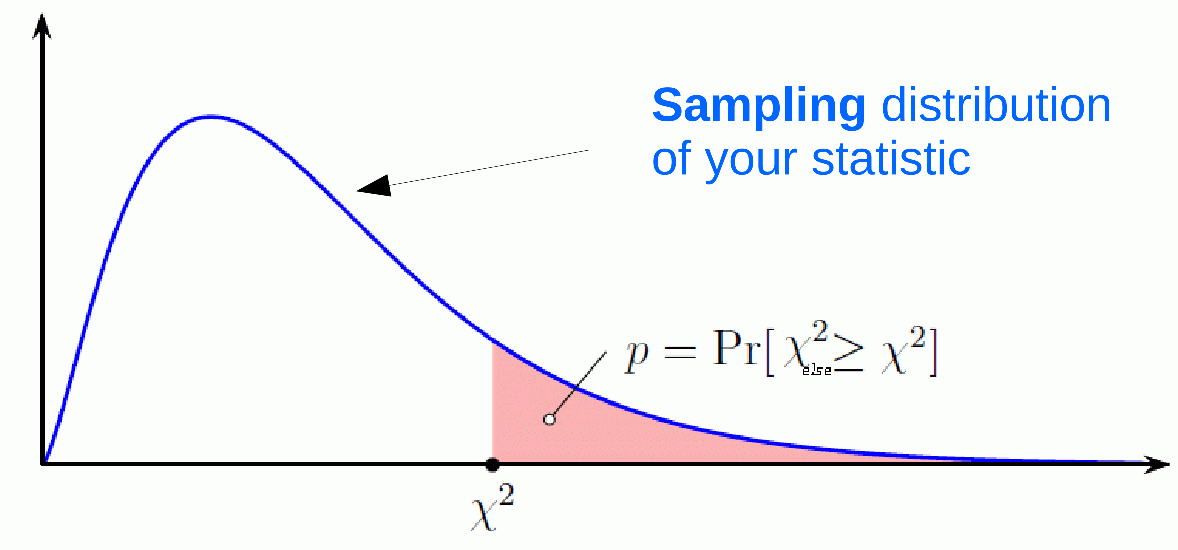


P-values describe necessary noise



$$H_o : x_1, ...x_n \sim \mathcal{G}(\mu_T, \sigma_T)$$

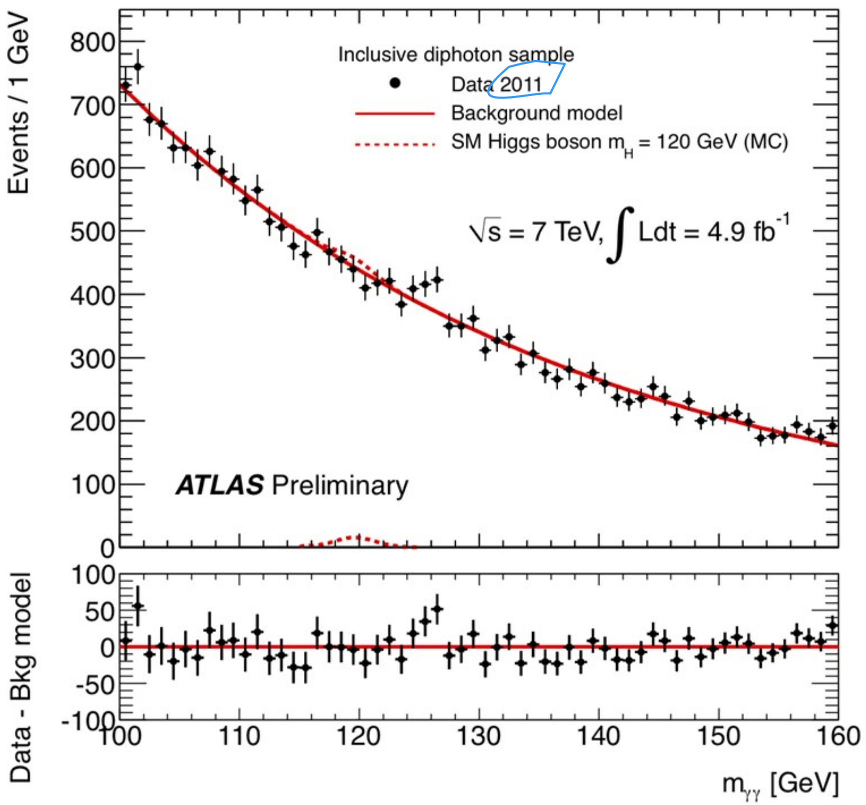
$$choose\ T = \chi^2 = \sum_i \left(\frac{x_i - \mu_T}{\sigma_T}\right)^2$$



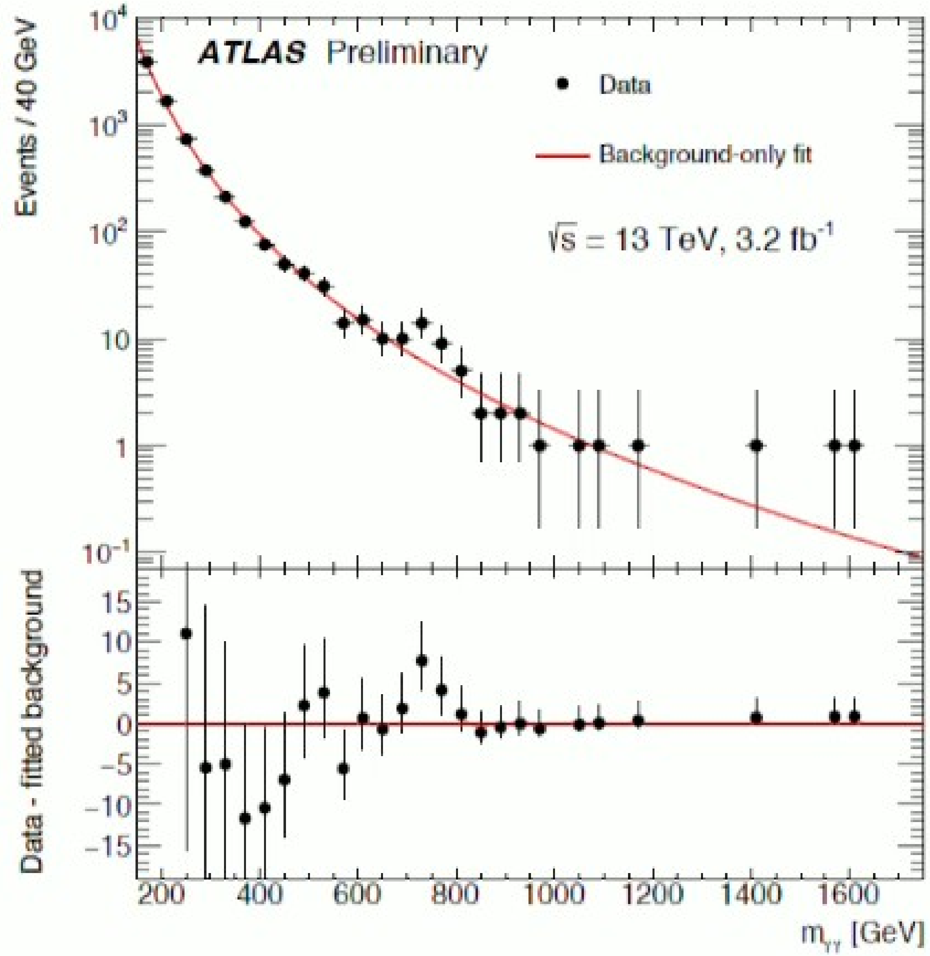
→ P-values describe how typical your noise is, for a certain hypothesis H_0: once out of x times, you **will** get such noise. And there is nothing you can do about it.

So you **can** use p-values.

→ To estimate how likely something is due to noise.



Di-photon excess



P-values and hypothesis rejection

The tempation

Start with the believe: H_0 is true.

Conduct one measurement.

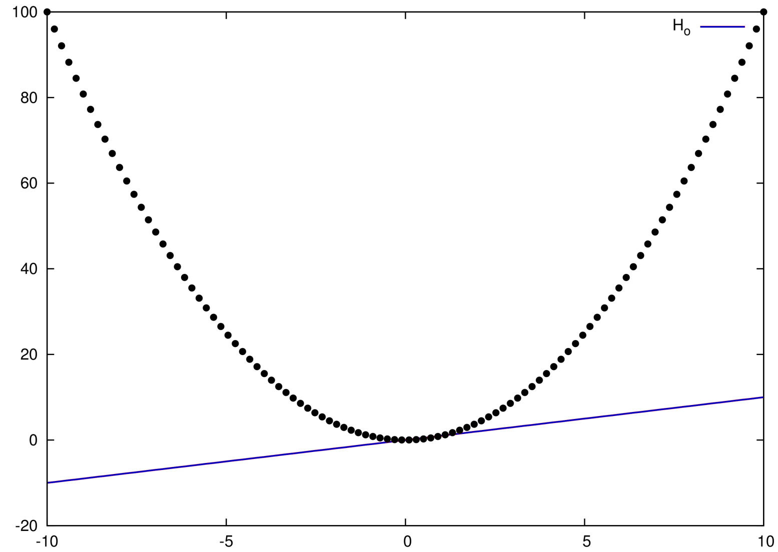
- P = 0.01: once out of 100 times, noise on top of H_0 is that weird.
- P = 0.001 once out of 1000 times, noise on top of H_0 is that weird.
- P = 1e-9: once in a billion, noise on top of H_0 is that weird.

Wait! I have measured only once! Why should my one measurement be that rare one in a billion case?

Low p-values make you doubt H_0

→ Wish: reject H_0 for low p.

It looks like a good idea:



But it is essentially impossible to control:

$$\vec{x}_{obs} \sim f(\vec{x}|\underline{H_o = true}) \qquad then \qquad p = \mathcal{P}[\chi^2 \geq \chi^2_{obs}]$$

→ But if H_0 is wrong, the p-value calculation is completely hypothetical.

Now what if H_0 actually is wrong?

$$\vec{x}_{obs} \sim f(\vec{x}|\underline{H_o = true}) \qquad then \qquad p = \mathcal{P}[\chi^2 \geq \chi^2_{obs}]$$

- Famous paper on the dangers of p-values: T. Selke, M. J. Bayarri, J. O. Berger, American Statistician (2001).
- Details on many possibilities of misinterpretation.
- Outreach-'friendly' versions of it exist.
  - Easy setup: count how often a true hypothesis is rejected.
  - Use of a 'precise' hypothesis (a yes/no answer), to avoid issues due to complexity.
  - Result: p-value of 0.05 → should reject H_0 5% of all times, but was measured to reject H_0 at least 23% of the times! For p = 0.01 H_0 is rejected at least 7% of the times.
  - Exact numbers depend on setup.

P-values in complex analyses

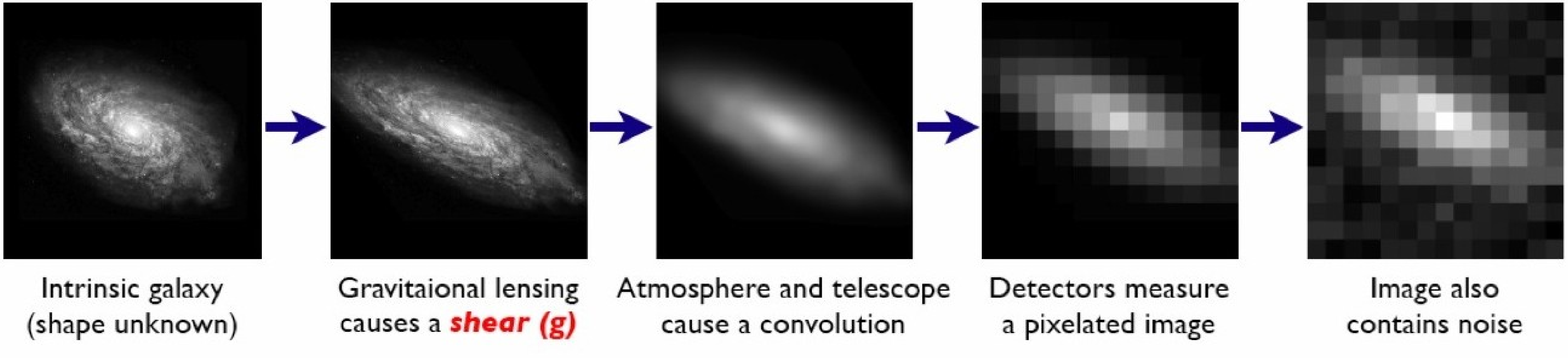
- Let's assume our H_0 is indeed true, but we don't know that.

How reliable are p-values in that case?

- Sampling distribution is not always χ_p^2
- But usually, that is what people use. (1st problem.)
- Illustrative example:

$$H_0 : \hat{C}_\kappa^{WL}(\ell) \sim C_\kappa^{WL}(\ell) \ of \ \Lambda CDM_{Planck \ BF}$$

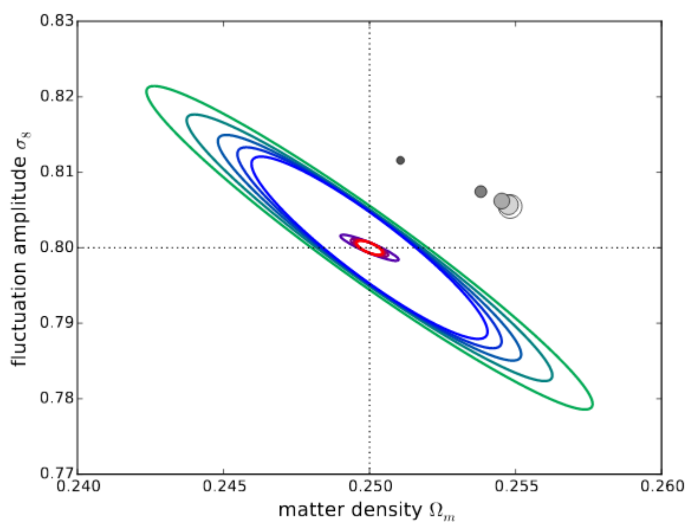
P-values in complex analyses



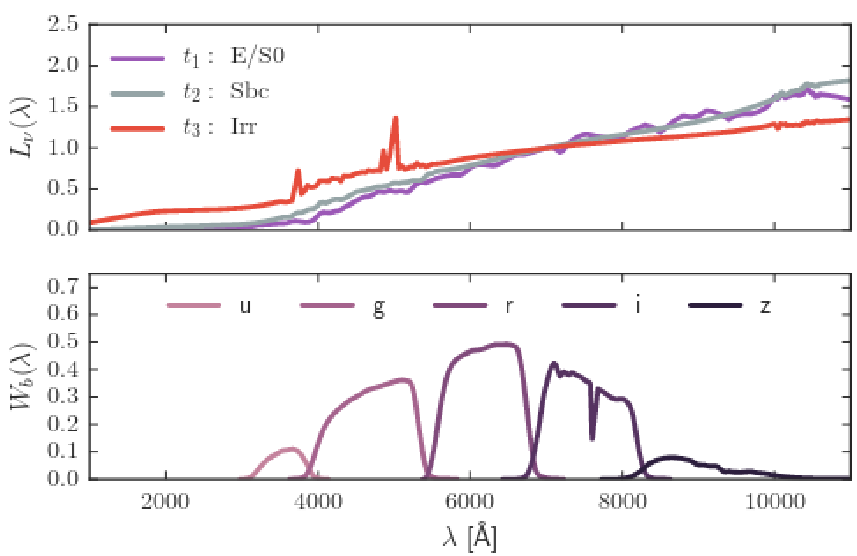
- Point spread function & blurring
- Pixelization
- Noisy images
- Shape measurements with sophisticated algorithms
- Source misclassification

P-values in complex analyses

- Intrinsic alignments: nuisance parameters & multiple models.
- Photometric redshift estimations: Galaxy position in z influences WL signal due to geometry.
- Non-linear CDM power-spectrum: N-body simulations? Field theories? “5 % accurate” solutions. Halo Fit?
- **Approximations?**

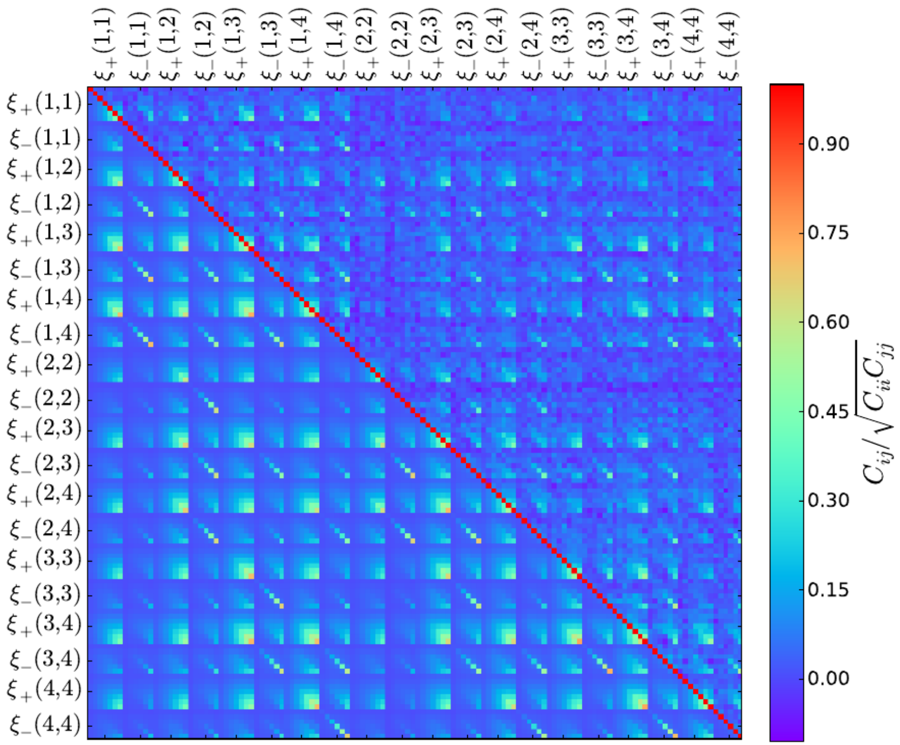


Merkel & Schäfer (2015)



B. Leistedt, DM, H. Peiris (2016)

P-values in complex analyses



From KiDS; Hildebrandt et al (2016)

- Estimated covariance matrix with N-body problems: grid resolution, boundary effects, super-survey modes.
- Analytically estimated covariance matrix with approximations.
- Cosmology of covariance is probably another than the best-fit cosmology.

End result

$$\underbrace{\chi^2_{\hat{C}-C_{true}}}_{\text{compatible}} + \chi^2_{shape} + \chi^2_{IA} + \chi^2_{photo-z} + \chi^2_{cov} + \ldots = \underbrace{\chi^2_{meas}}_{\text{too large}}$$

H_0 was true, but we rejected it, because our data reduction was too bad/complex.

P-values accumulate systematics. They aren't made for quick solutions to complex problems. And that's why Bayesian Hierarchical Models (BHM) are currently on the rise in astronomy (→ ask AH).

Conclusion:

Before you doubt a hypothesis due to p-values, doubt your analysis.

However... p-values can't be complete nonesense either

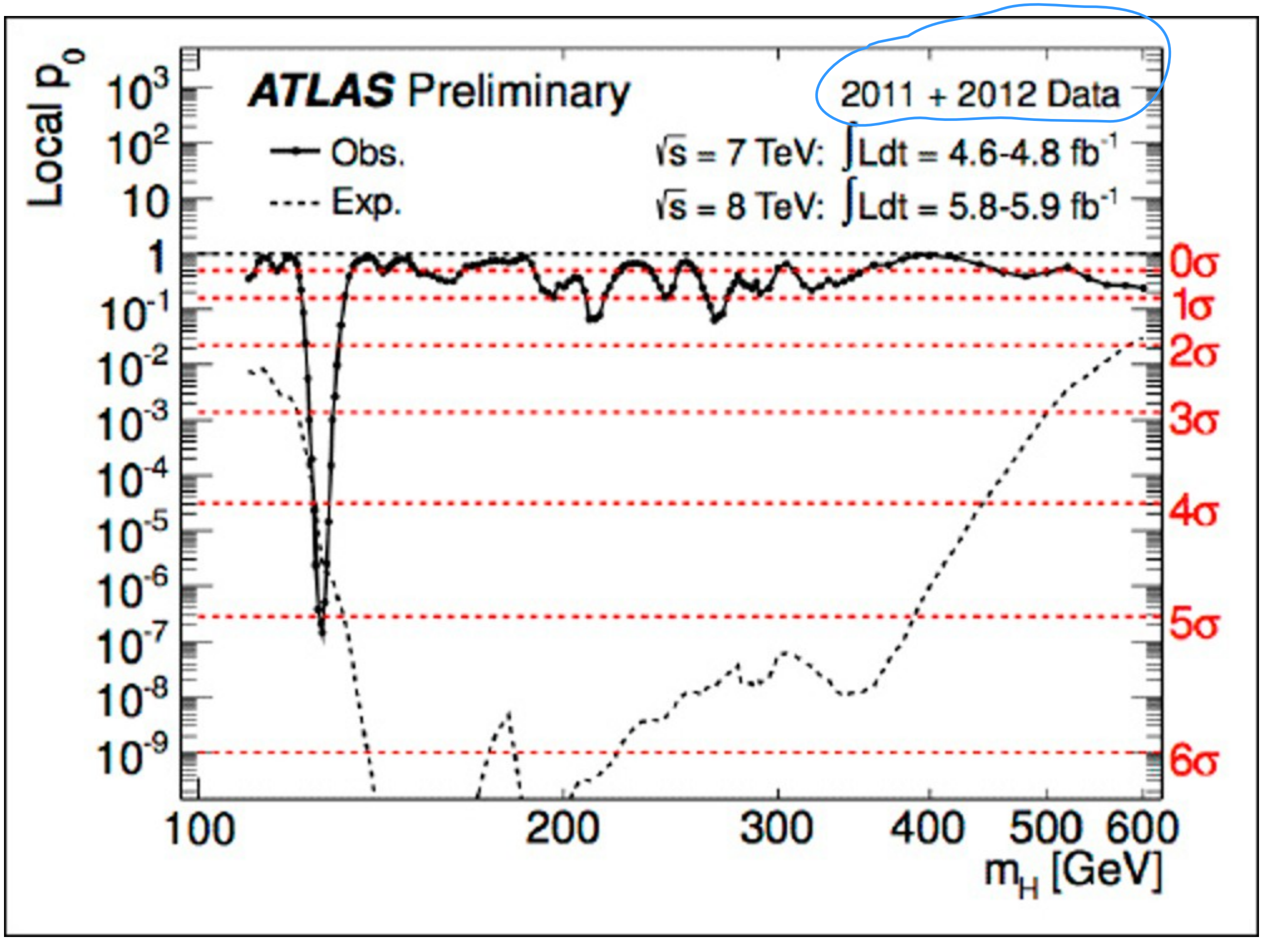
- H_0: “Arsenic is good for your health.”
- Conduct study[1]. → **extremely** low p-value.

1) Do NOT conduct this study!! Arsenic is extremely poisonous.

- “P-values just parameterize noise and are dominated by mistakes in complex analyses. H_0 is true.”

Is it?

First check on noise

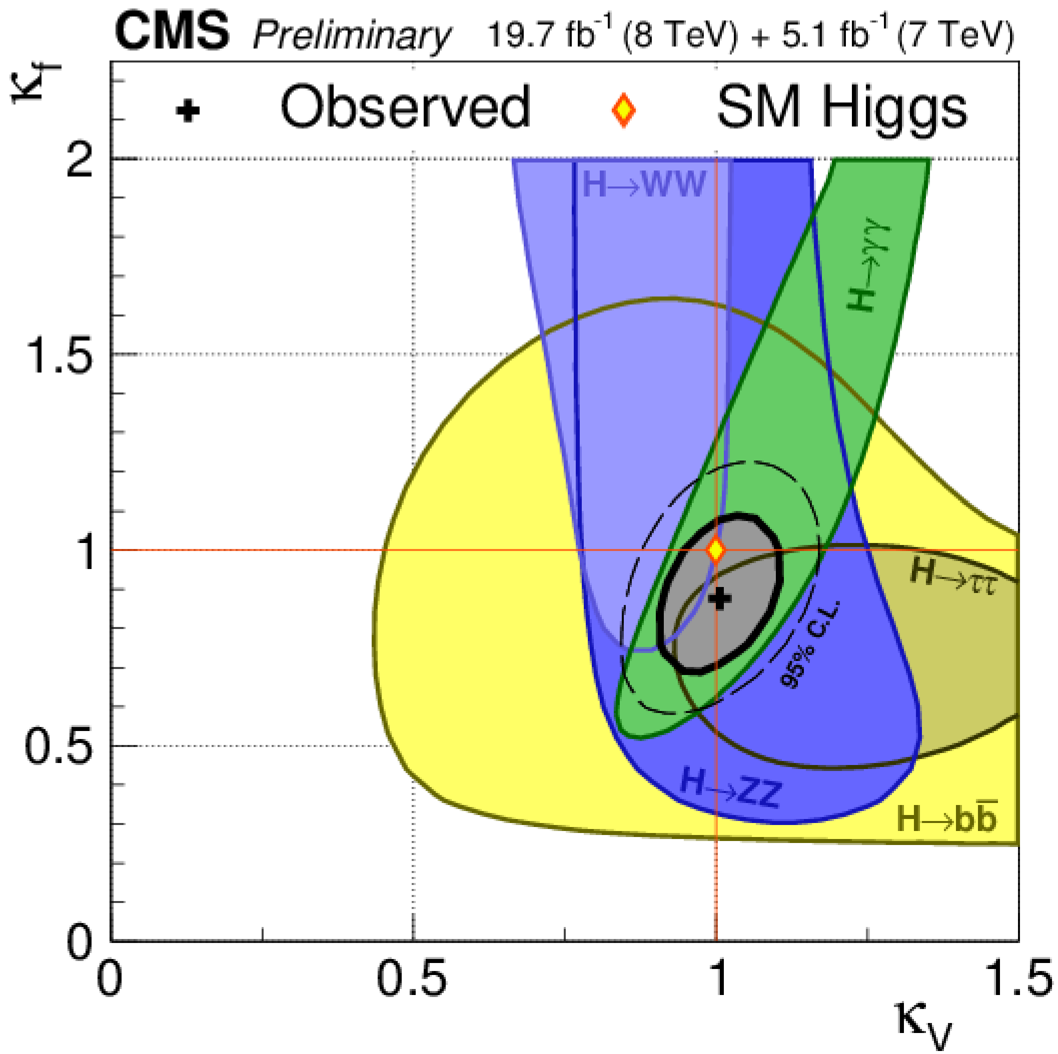


Usual noise

We have made at least one mistake.

Okay, it's impossible we made that many mistakes, that's a signal!
(With a significance of 5sigma, based on a calculation that were correct if the signal did not exist.)

Then measure parameters



Summary

- P-values: estimate the 'weirdness' of noise (that's fine).
- Noise is part of the game; p-values teach you to accept it.
- P-values: Hypothesis rejection/Model selection (take care!)
  - Prepare for being confused and don't ignore your confusion.
  - Low p-value (0.05 -- 1e-4): doubt your analysis before you doubt your hypothesis! Do you have a sampling distribution?
  - Extremely low p-value (<1e-5): probably physics, **if** all other cross-checks on your data turn out fine
  - 1e-5 is a convention from HEP
- Bayesian Hierarchical Models are designed to treat complex situations and force you to think about assumptions and specifications which p-values would clandestinely 'sweep up'.