Bayesians vs. Frequentists

~or~

The amazing discovery of pentaquarks, and the great un-discovery of the CMB cold spot

Elena Sellentin
Imperial Data Analysis Workshop (2018)

Sterrewacht
Universiteit Leiden, NL

Bayesians vs Frequentists

None of these two schools of thought is ‘better’ than the other.

But you can definitely frustrate yourself by tackling Bayesian problems with frequentist tools, and frequentist problems with Bayesian tools.

Differences in interpretation

- 'Frequentist' comes from frequency: rely on an actual or hypothetical repetition of the experiment.
- Bayesians interpret probability as a credibility.

Frequentists are friends of limit theorems $N \to \infty$ and typically forward model the distribution of their estimators.

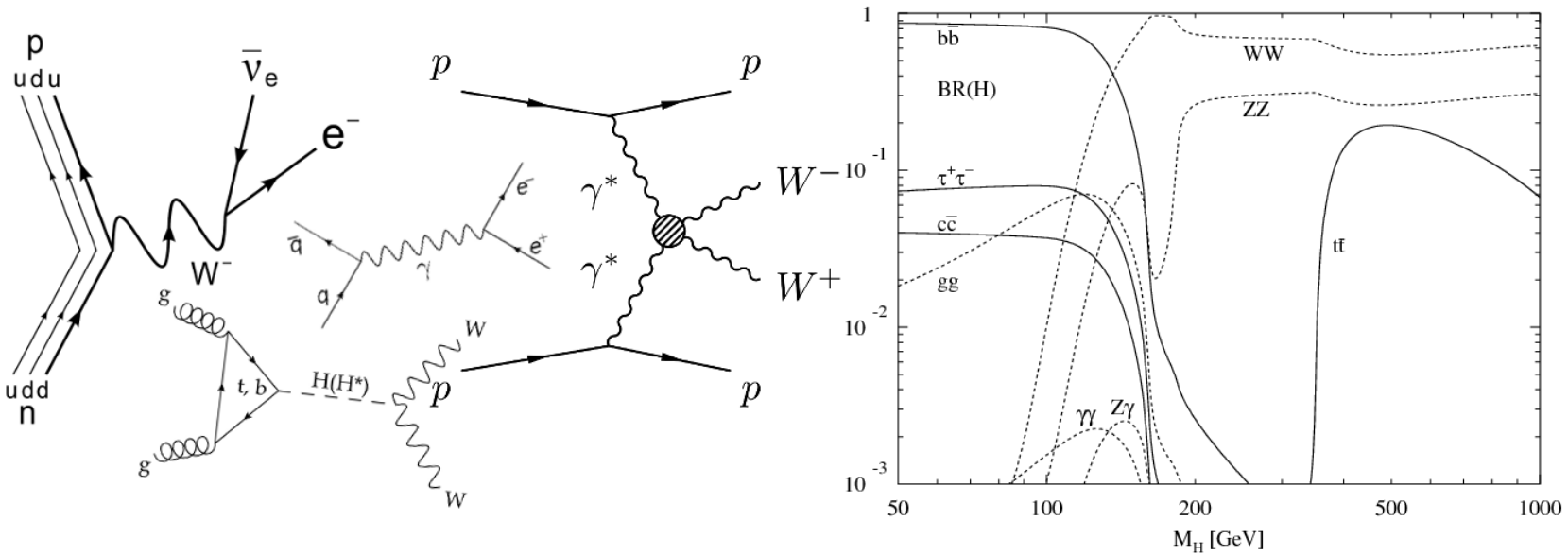
Mindset: 'I have measured the mass of the proton 1 million times. I always get $1.672621898(21) \cdot 10^{-27}$ kg.
I think if I measure once more, I'll again get $1.672621898(21) \cdot 10^{-27}$ kg.

Some cases are clearly Bayesian

Some cases are Frequentist

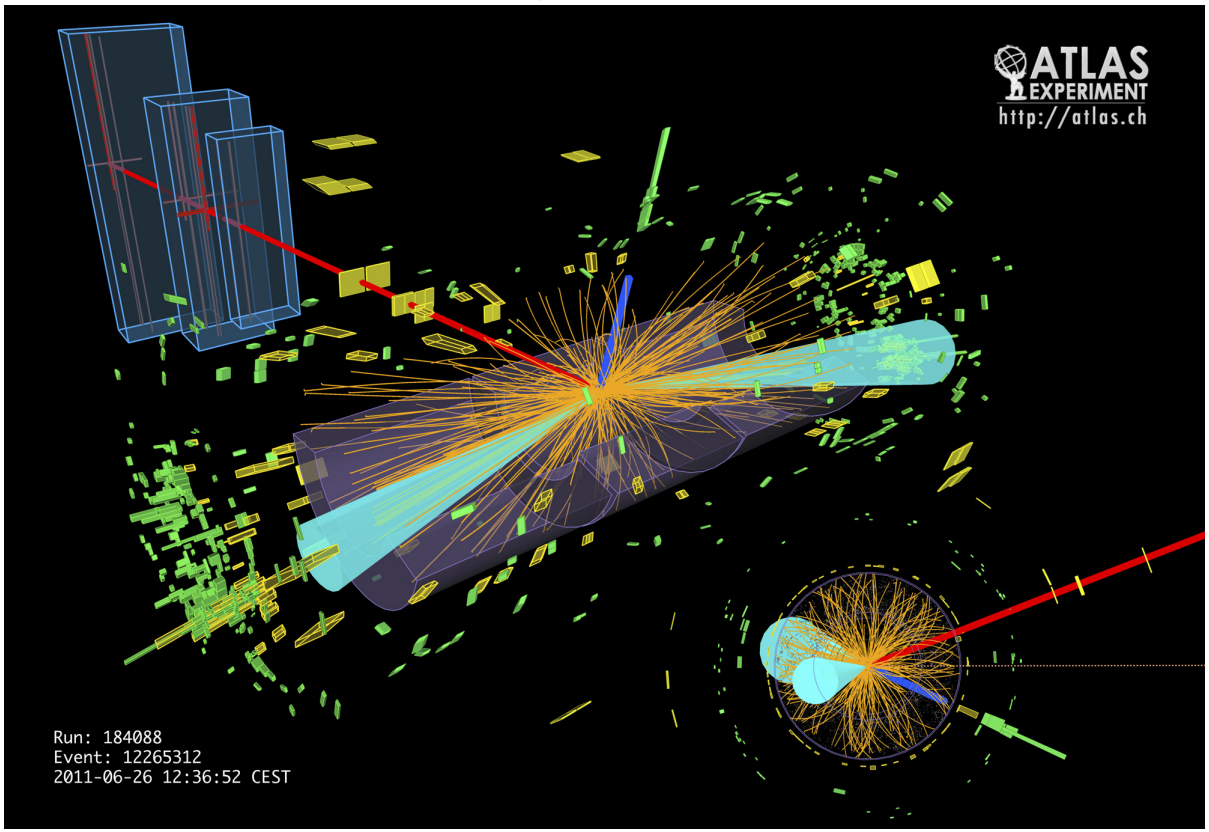
- Tests of medical tests: true positives, false positives, true negatives, false negatives. (Population studies.)
- Producing & detecting particles: particle creation **is** frequentist by nature.

  p + p $\rightarrow$ scattering matrix $|\langle f|M|i\rangle|^2$, transition **probabilities**.



Some cases are clearly Frequentist

Hadronization, detection and (mis)classification is again frequentist...



... with some Bayesian nightmares (trigger!)

Frequentist questions to Bayesians

1. How exactly do you get those priors?
2. Do you really just fit a model, without checking previously that your ‘signal’ isn’t just noise?
3. You do know that each time you fit, it is guaranteed that you get an answer? Even if it was just noise?
4. How do you get rid of a bad model? Without replacement?

Replacements for concepts

- Priors $\Rightarrow$ null-hypothesis + sampling distribution of test statistics T
- Model comparisons:
  - hypothesis rejection & p-values
  - likelihood-ratio tests, $\Delta\chi^2$
- Parameter estimation: quite similar! ML-estimators $\hat{\boldsymbol{\theta}} = \mathrm{argmax}[L(\boldsymbol{x}|\boldsymbol{\theta})]$, LS-estimators $\mathrm{minim}\chi^2$, sample estimators $\bar{x} = 1/N \sum_i x_i$...
- Inversion of the workflow: order of parameter estimation and model/hypothesis selection is interchanged

Workflows

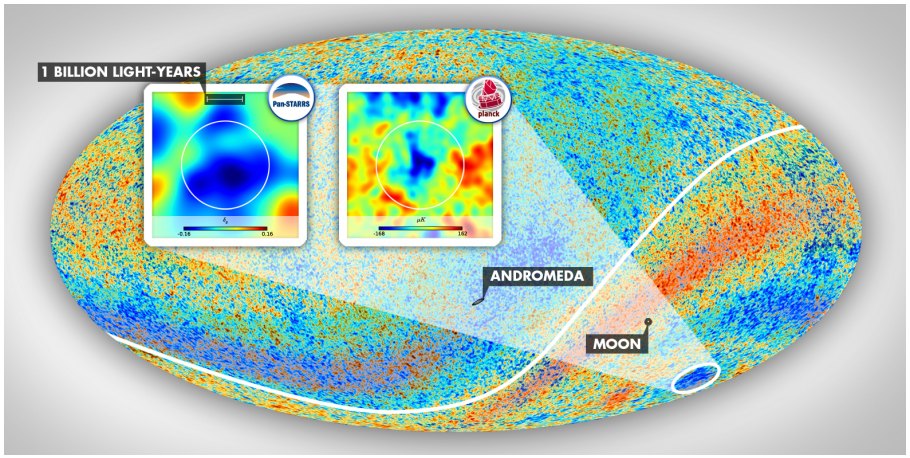
Astronomy:

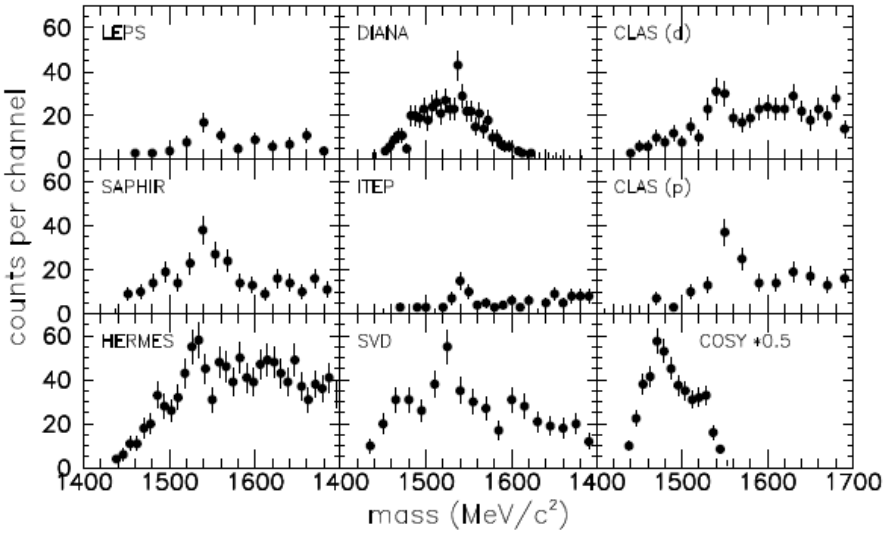
1. Get data = true signal + noise
2. select parametric model (decides which 'signal' is in the data)
3. estimate the model parameters
4. doubt model, compare it to a competitor model (evidences)

Particle physics:

1. Get data. Null-hypothesis H_0 as neutral as possible, no prejudices about potentially hidden signals.
2. non-parametric checks: is it maybe still noise? (p-values)
3. It's not noise!
4. select model and estimate its parameters.

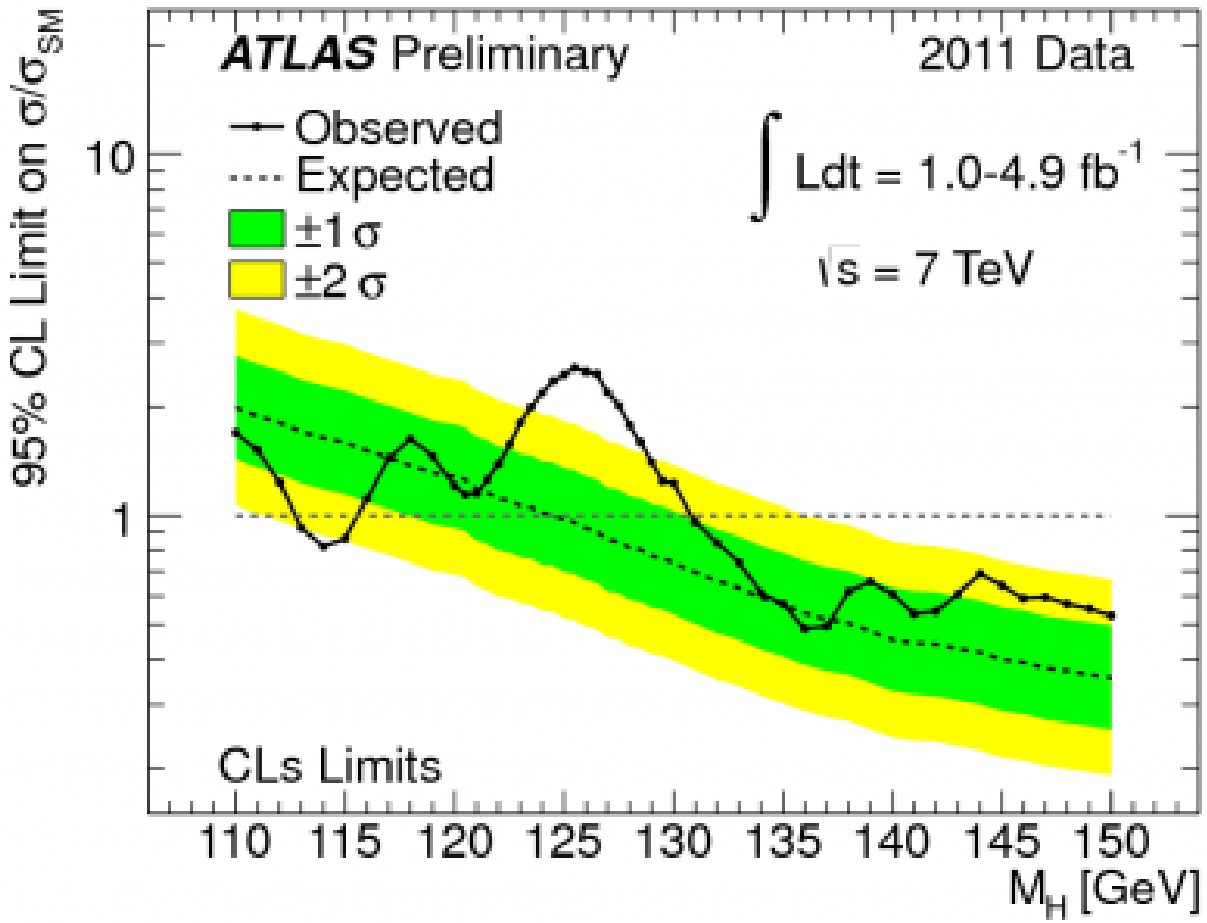
Inverted Workflows





Constructing confi.-intervals with H_0

events →



Monte Carlo Simulations ⇒ sampling distribution

P-VALUES...

... something that made sense, until someone had a bad idea.

p-values: tails of sampling distribs

Def: if $\boldsymbol{x}_{\rm obs} \sim \mathcal{D}(\boldsymbol{x}|H_0)$, and $T(\boldsymbol{x})$ a test statistic of $\boldsymbol{x}$...

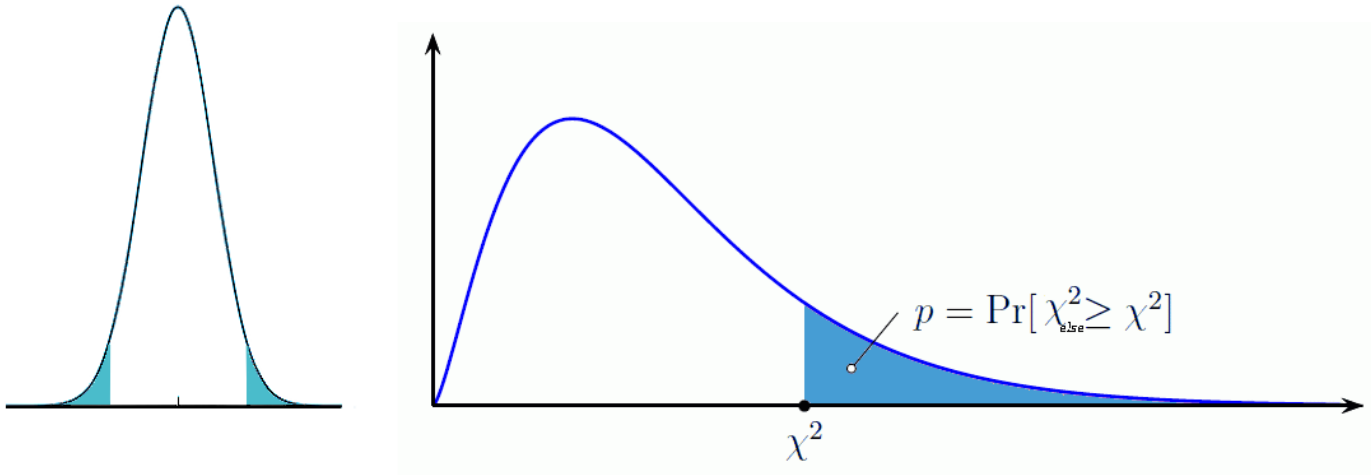
... then $p = \mathcal{P}[T(\boldsymbol{x})] \geq T(\boldsymbol{x}_{\rm obs})$

- Large values of T typically indicate bad agreement.
- p-value for large T is then *small*.
- The sampling distribution needs to be correct to have a reliable p-value.

P-values describe necessary noise

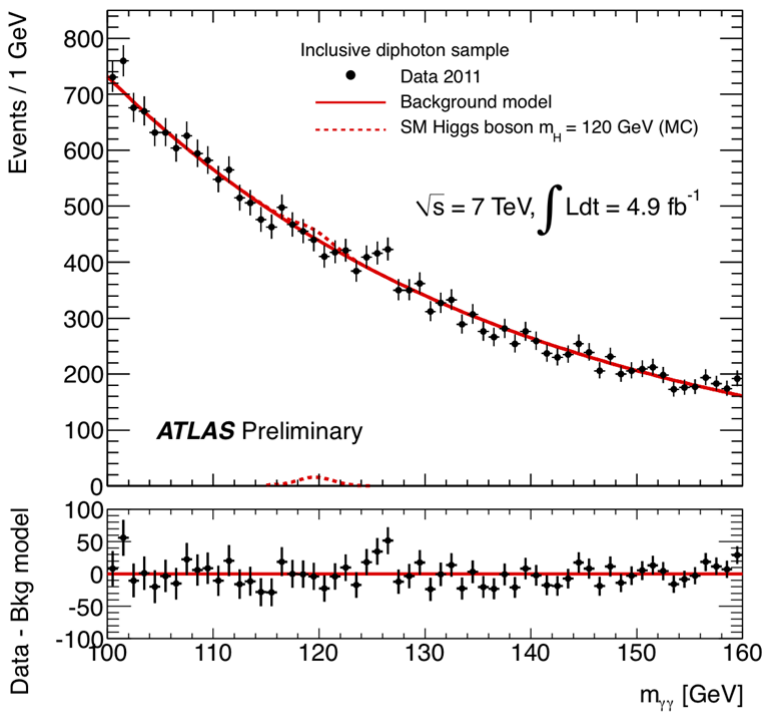
Example for p-values as tails of distributions:

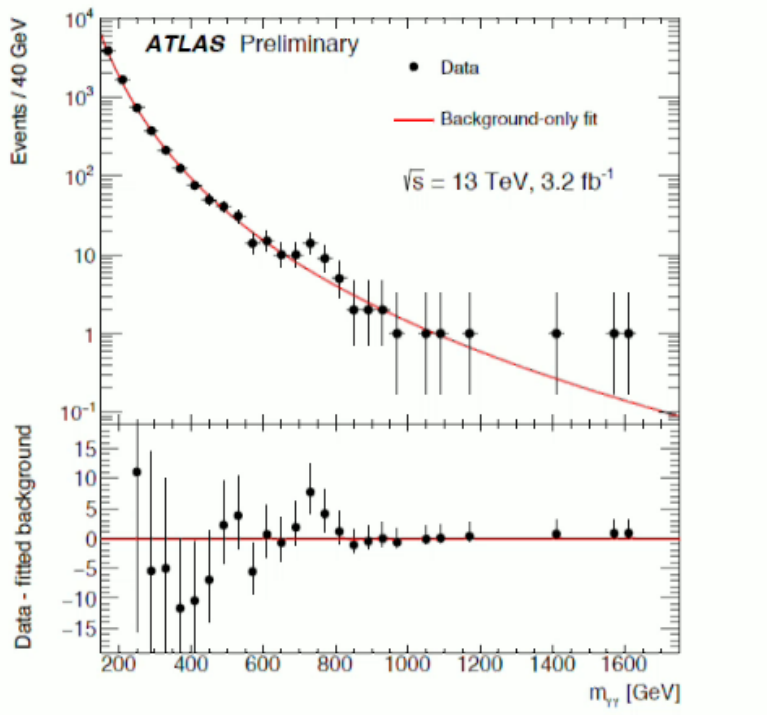
$H_0 : x_1, ..., x_n \sim \mathcal{G}(\mu_T, \sigma_T)$, choose $T = \chi^2 = \sum_i \left(\frac{x_i - \mu_T}{\sigma_T}\right)^2$



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

You can use p-values...





... to estimate how likely something is due to noise.

p-values and hypothesis rejection

— THE TEMPTATION —

The temptation

Start with the belief H_0 is true.

Conduct one measurement.

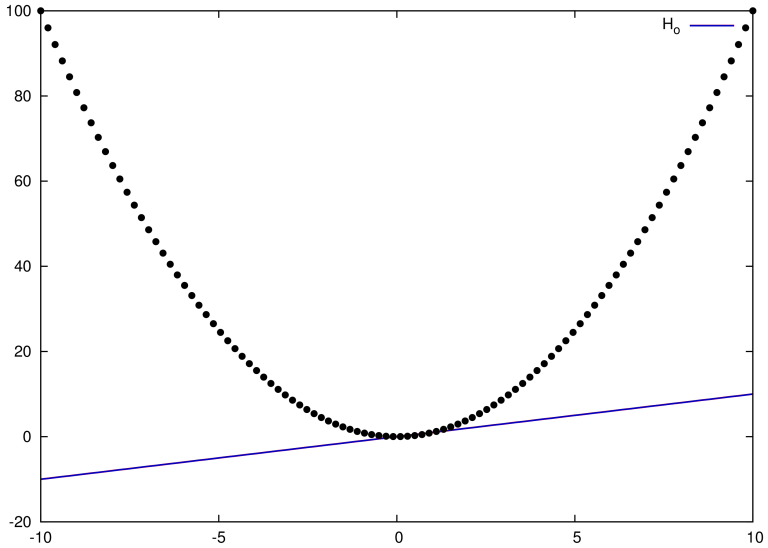
- p = 0.01: once out of 100 times, noise on top of H_0 is that weird.
- p = 0.001: once out of 1000 times, noise on top of H_0 is that weird.
- p = 10^{-9}: once in a billion, noise on top of H_0 is that weird.

Wait! I have just measured once! Why should my one measurement be that rare once 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.



IF

$$\boldsymbol{x}_{\text{obs}} \sim f(\boldsymbol{x}|H_0 = \text{true})$$

THEN

$$p = \mathcal{P}[T(\boldsymbol{x}) \geq T(\boldsymbol{x}_{\text{obs}})].$$

But if H_0 is wrong, the whole p-value calculation is entirely hypothetical.

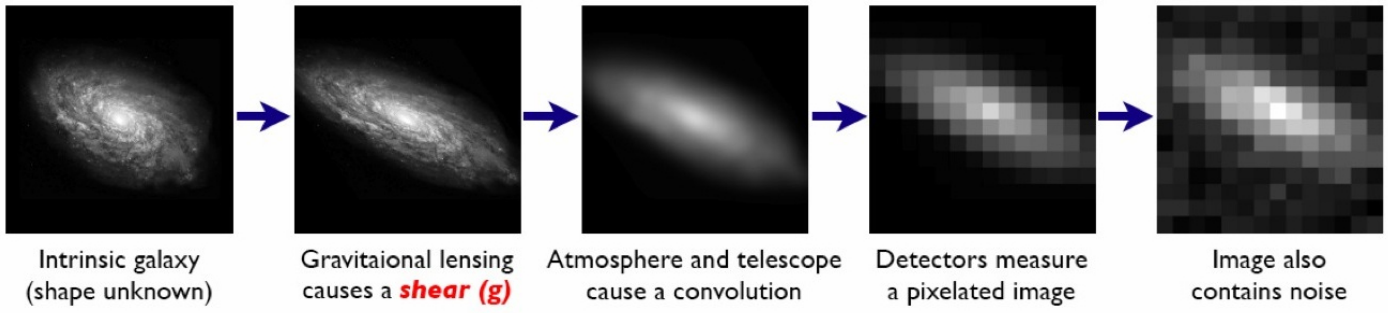
Complex analyses

- Let's assume H_0 is indeed true, but we don't know that.
- How reliable are p-values then for testing the truth of H_0?

1. Need the correct sampling distribution (Don't know it. Numerically tough!)
2. So let's opt out and use the χ^2- distribution as sampling distribution.

Cosmological example:
$H_0 : \hat{C}(\ell)^{\mathrm{WL}}_{\kappa} \sim C(\ell)^{\mathrm{WL}}_{\kappa}$ of ΛCDM with Planck best-fit parameters.

Complex analyses



- PSF, pixelization, noise
- Shape measurements with sophisticated algorithms
- Source misclassification, deblending
- Malmquist-bias
- Flat-sky approx., Limber approx.
- Intrinsic alignments: nuisance params & model uncertainties
- photometric redshift errors, outliers
- non-linear CDM power spectrum: N-body sims? Field theories? 5% accurate solutions? Halofit?
- Covariance matrix estimation: estimated on the right cosmology?
- And how good is the Gaussian approximation?

End result

$$\chi^2_{\hat{C}-C_{\text{true}}} + \chi^2_{\text{shape}} + \chi^2_{\text{IA}} + \chi^2_{\text{photo-z}} + \chi^2_{\text{Cov}} + \ldots = \chi^2_{\text{measured}}$$

and

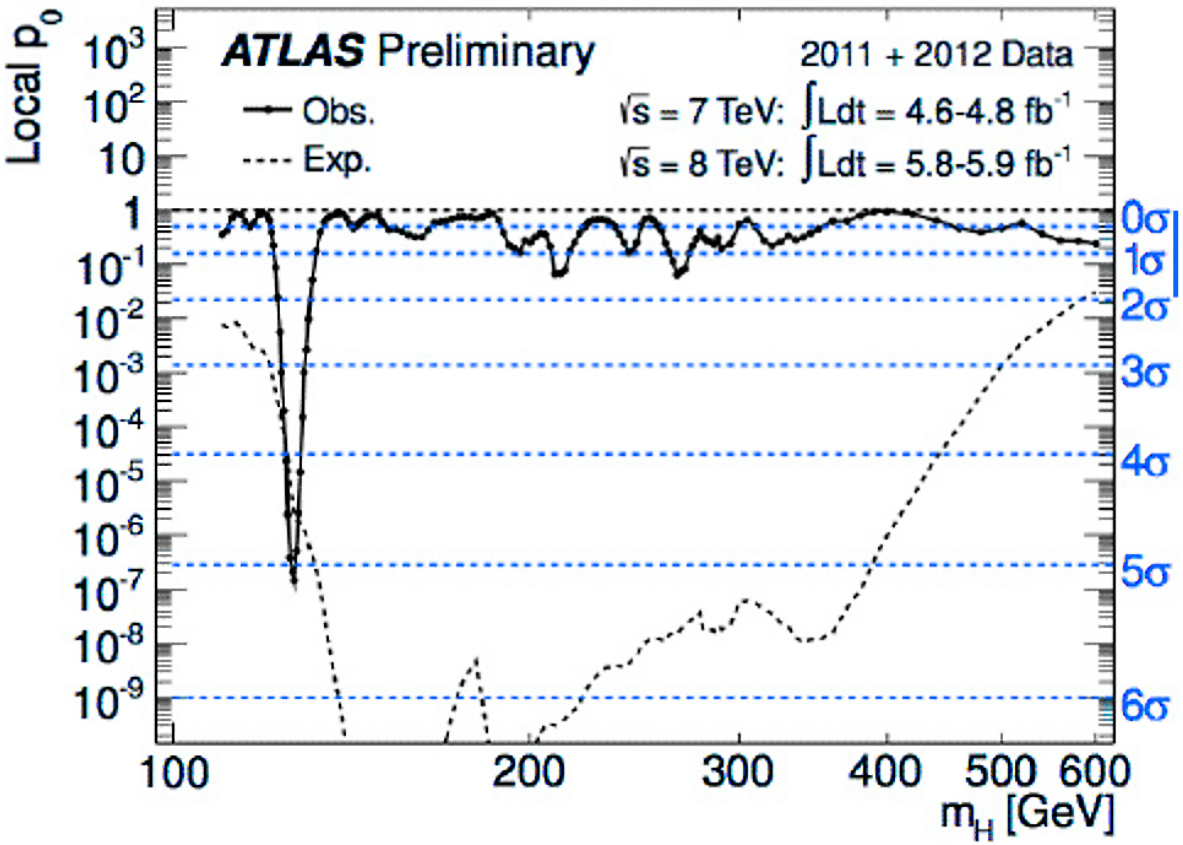
$$\chi^2_{\text{measured}} > \chi^2_{\hat{C}-C_{\text{true}}}$$

- H_0 was true, but we rejected it, because our data reduction was too bad/complex.
- p-values accumulate systematics. The aren't made for quick solutions to complex problems. That is why HEP's prepare for them over years!

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

The frequentists way

First check on noise.



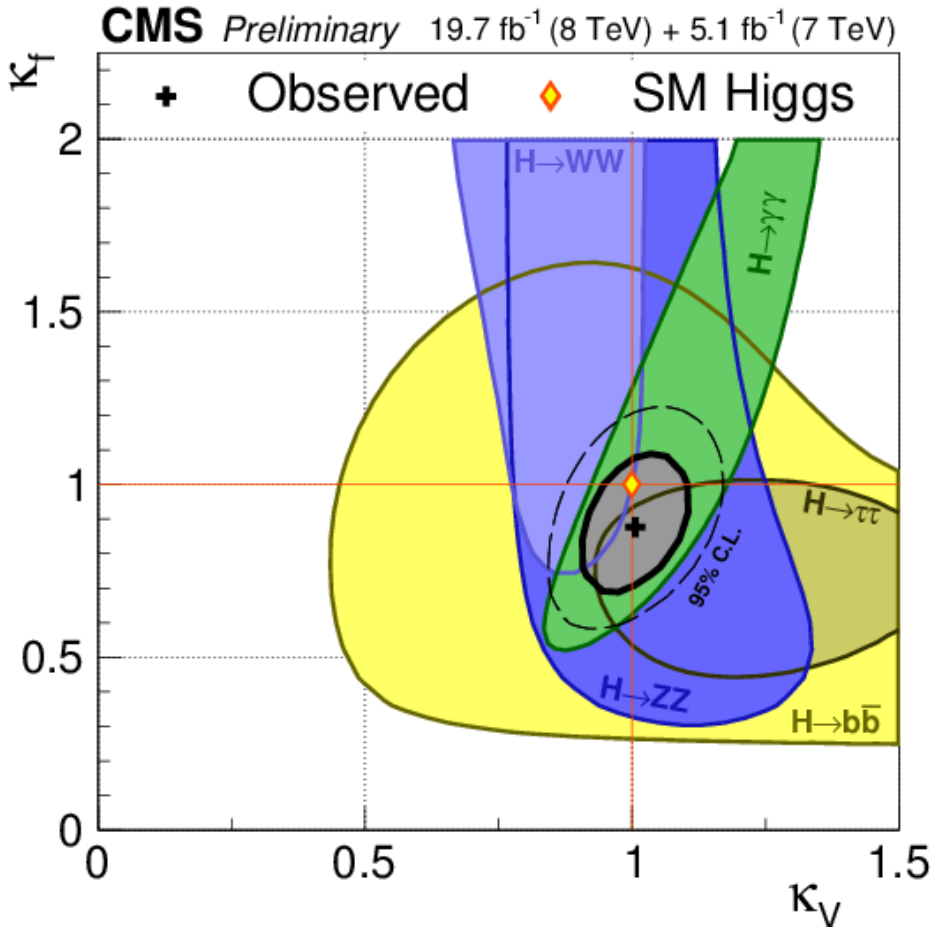
Usual noise.

We've made at least one mistake.

Okay, it is impossible we made that many mistakes, that is a signal!

The frequentists way

Then measure parameters.



A historical incident involving love, tragedy and low p-values

The actors[1]

- Pierre-Simon Laplace (1749-1827)
- Thomas Bayes (1701-1761)
- Many people who needed money.
- Many people who wanted to be healthy.
- Samuel Hahnemann (1755-1843)

Feel of the time: Marie Antoinette (1755-1793) was alive, bloodletting was fashionable.

[1] We apologize to all humans whose story has not been handed down correctly throughout the years.

Ergo: There was no such thing as approved clinical trials.

People had difficulties learning from data:

- Data $\boldsymbol{x} \sim \mathcal{D}(\boldsymbol{x})$
- Hypothesis or theory H
- $\mathcal{L}(\boldsymbol{x}|H)$ ('Claiming the hypothesis is true, this is how often it generates data $\boldsymbol{x}$')
- $\mathcal{P}(H|\boldsymbol{x})$ ('Look, these are my data $\boldsymbol{x}$, how likely is my hypothesis?')

Bayes' Theorem

$$\mathcal{P}(H|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|H)\pi(H)}{\pi(\boldsymbol{x})} \tag{1}$$

Today: $\pi(H)$ m-a-t-t-e-r-s. (But let's forget that; we're in 1790.)

So we come up with a hypothesis... namely[2]...

[2] Don't even try this on your worst enemy!

H = 'Arsenic[3] is good for your health.'

No, it isn't! Arsenic is a deadly poison!
Don't eat it AT ALL!!

[3] Dutch & German: Arsenic is 'Arseen'. It's a terrible poison!

But we're in 1790...

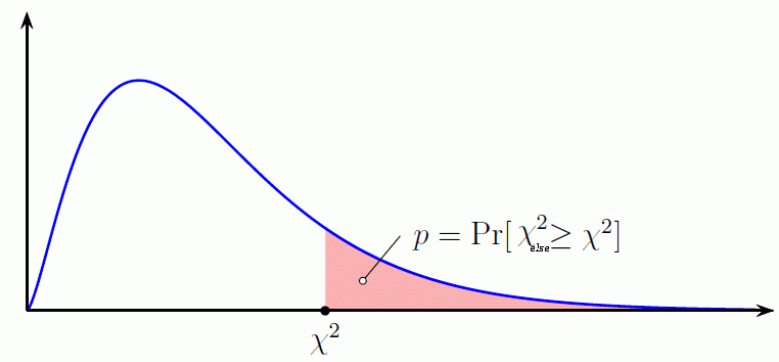
... so someone conducted a 'study'.

(And applied what we today call p-value reasoning.)

Outcome of the arsenic study

$$\mathcal{P}(H|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|H)\pi(H)}{\pi(\boldsymbol{x})}$$

- Person 1: dies.
  Analysis from back then: $\pi(H)$ and $\pi(x)$ don't matter, let's put them to 1. Then $\mathcal{P}(H|\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}|H)$. Since $H =$ '$Arsenic\ good$' probably right, x must be from tail.
  Conclusion: well, this happens.
- Person 2: dies.
  Conclusion: well, as above.
- Person 3.... dies.
- Person X.... dies.



Correct conclusion would have been:
Arsenic is a poison and kills people.

But the conclusion drawn was...

Homeopathy

Samuel Hahnemann (1755-1843)

- PhD completed at Erlangen University (Bavaria)
- Founder of homeopathy
- Results published in ‘Organon der Heilkunst’ (1810)

→ Take-home messages:

Frequentists: Don’t confuse what is hypothetical and what is real.

Bayesians: Always check on your priors.

Both: Your data *must* dominate your priors/hypotheses.