

# Illuminating the dark side of statistics: Bayesian inference in particle physics

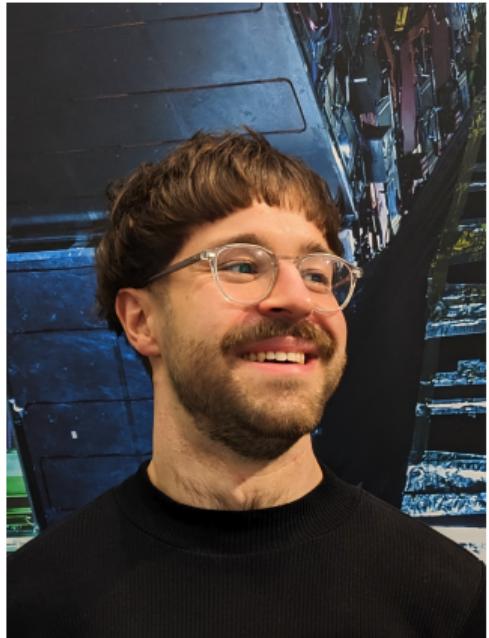
Lorenz Gärtner<sup>1</sup>

<sup>1</sup>*LMU Munich*

March 13, 2025

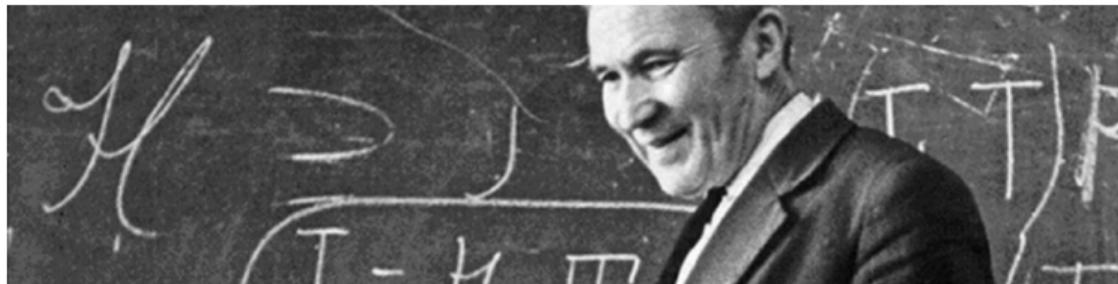
# About me

- BSc @ University of Manchester  
Physics with theoretical physics
  - MSc @ LMU Munich  
More theory ...  
— — — almost no stats — — —
  - Currently PhD @ LMU Munich  
A lot of stats
- Never too late to start



# What is a probability?

# Kolmogorov probability axioms



1.  $p(\Omega) = 1$ , where  $\Omega$  is the sample space.
2.  $p(x) \geq 0$  for any event  $x \subseteq \Omega$ .
3. For any sequence of disjoint events  $x_1, x_2, \dots$ ,

$$p\left(\bigcup_{i=1}^{\infty} x_i\right) = \sum_{i=1}^{\infty} p(x_i)$$

# Probability interpretations

Axioms tell you how to calculate with probabilities.

How do we assign probabilities in the first place?  
→ Need probability interpretations.

The interpretations **share the same mathematical framework**, but the meaning of  $p(x)$  is different.

# Frequentist interpretation

Assign a probability as relative frequency

$$p(x) = \lim_{N \rightarrow \infty} \frac{N_x}{N}$$

- **Data** is random.
- For repeatable experiments only.

# Bayesian interpretation

Assign a probability  $p(x)$  as *degree of belief*.

- **Parameters** are random.
- Inference results are subjective.

# Bayes' theorem

The **posterior** is

$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})p(\text{theory})}{p(\text{data})}$$

- **Likelihood**  $p(\text{data}|\text{theory})$

- **Prior**  $p(\text{theory})$

- **Marginal likelihood**

$$p(\text{data}) = \int p(\text{data}|\text{theory})p(\text{theory})$$

Can you derive it?

[...] nearly all physicists tend to misinterpret frequentist results as statements about the theory given with the data.

A. L. Read

# Bayesian beliefs

$p(SM|data)$

$p(Higgs|data)$

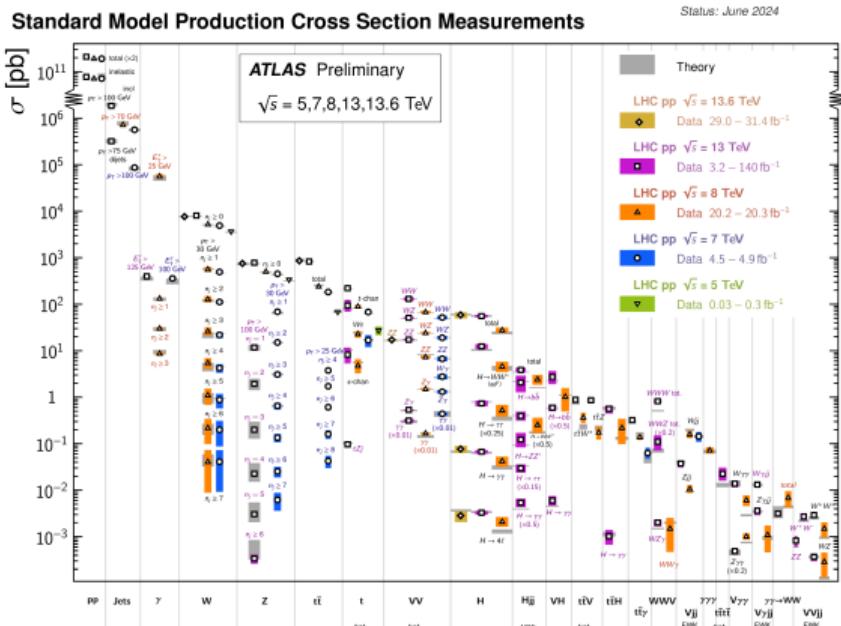
$p(SUSY|data)$

# Bayesian beliefs

$p(\text{SM}|\text{data})$

$p(\text{Higgs}|\text{data})$

$p(\text{SUSY}|\text{data})$

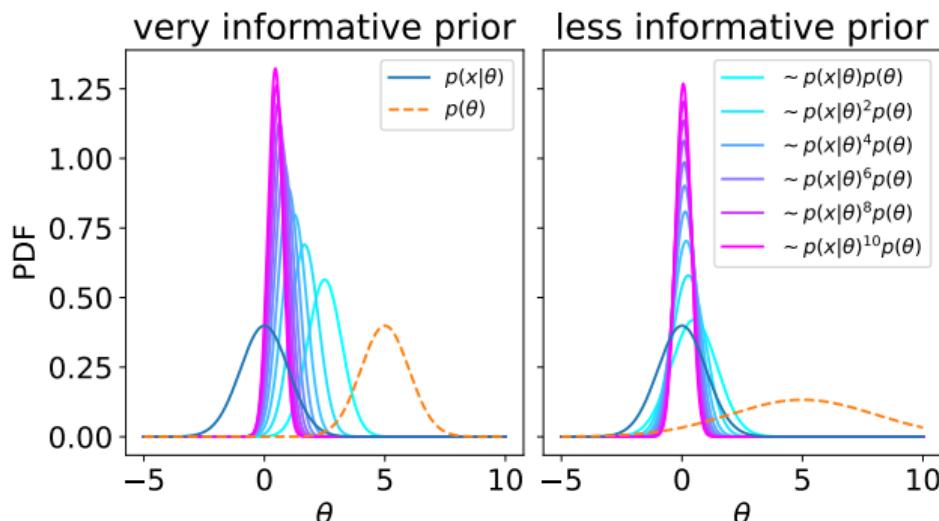


ATLAS

# Bayesian updating

Can generally use measurements to *update* our posterior

$$p(\theta|x_1, x_2) = \frac{p(x_2|\theta)}{p(x_2)} p(\theta|x_1) = \frac{p(x_2|\theta)}{p(x_2)} \frac{p(x_1|\theta)}{p(x_1)} p(\theta).$$



# The common ground

**Frequentist inference** is based on

$$p(x|\theta)$$

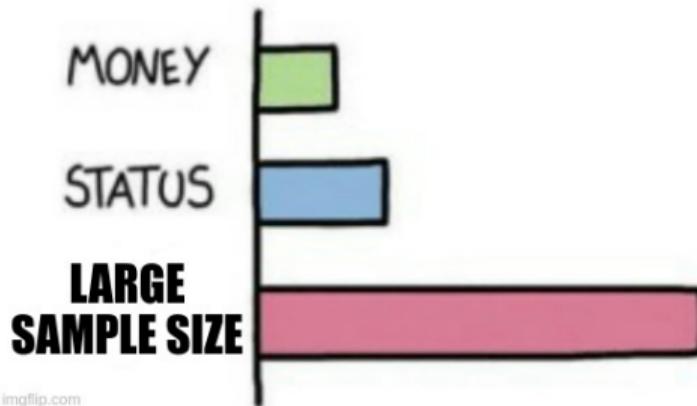
**Bayesian inference** is based on

$$p(\theta|x) = \frac{p(x|\theta) p(\theta)}{p(x)}$$

You want the best possible  $p(x|\theta)$  !

The best possible model...

## WHAT GIVES PEOPLE FEELINGS OF POWER



imgflip.com

# Parameter inference

## **Point estimates**

Identify the most probable parameter point.

## **Interval estimation**

Identify extended regions in parameter space based on compatibility with the data.

# Frequentist point estimates: estimators



Estimator is a *statistic*  $\hat{\theta}(x)$ , with desired properties

- **consistency**

$$\lim_{N_x \rightarrow \infty} E_x[\hat{\theta}] = \theta_{true}$$

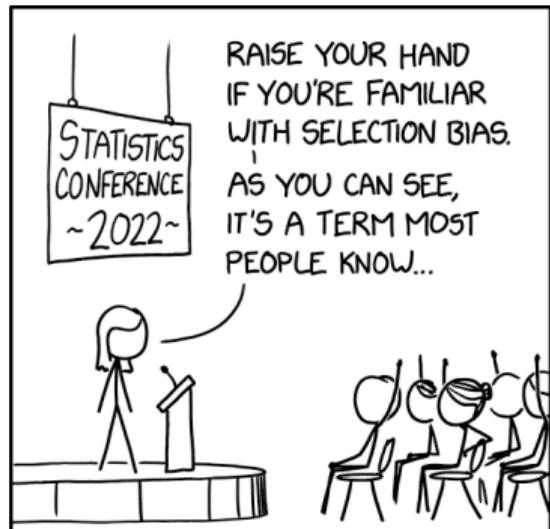
- **unbiasedness**

$$b = E_x[\hat{\theta}] - \theta_{true}$$

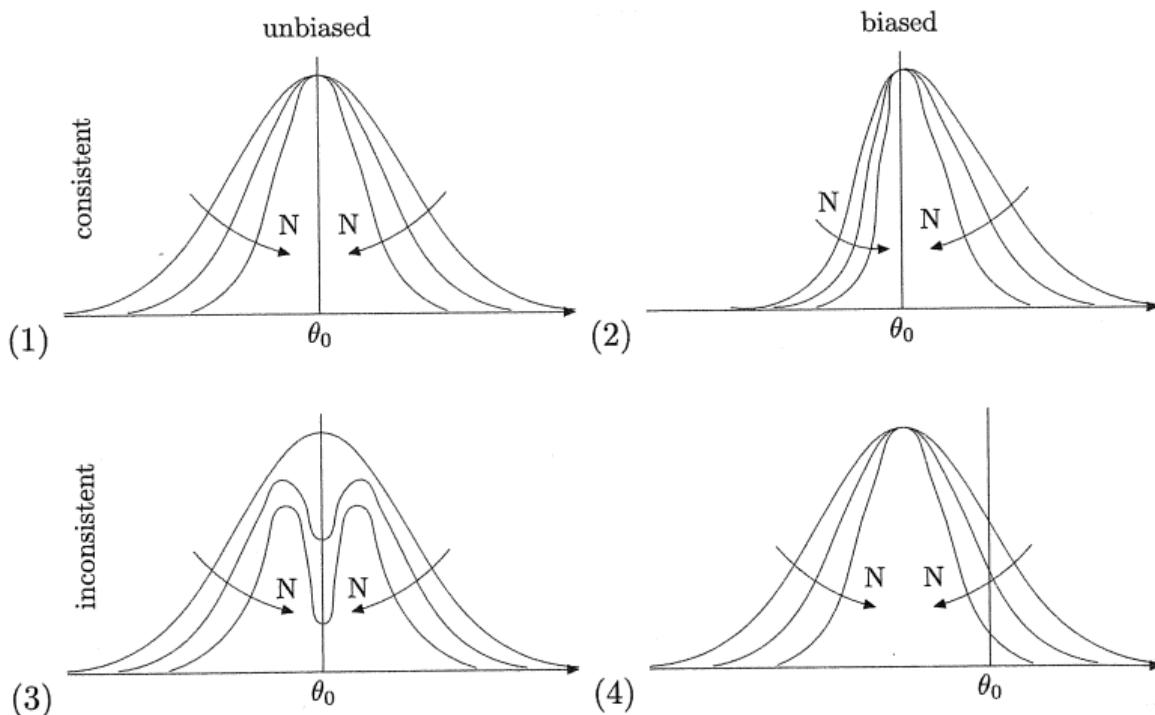
- **efficiency**

$$V(\hat{\theta}) = I(\theta)^{-1} = E_x \left[ \left( \frac{\partial \ln p(x|\theta)}{\partial \theta} \right)^2 \right]^{-1}$$

- ...



# Estimator properties



# Method of maximum likelihood

We find maximum likelihood estimators  $\hat{\theta}$  by solving

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(x|\theta)$$

It has the property

$$\lim_{N \rightarrow \infty} p\left(\sqrt{N}(\hat{\theta} - \theta_{true})\right) = \mathcal{N}\left(0, I^{-1}(\theta)\right),$$

implying consistency, asymptotic unbiasedness and efficiency.

# Bayesian point estimates

## Mode

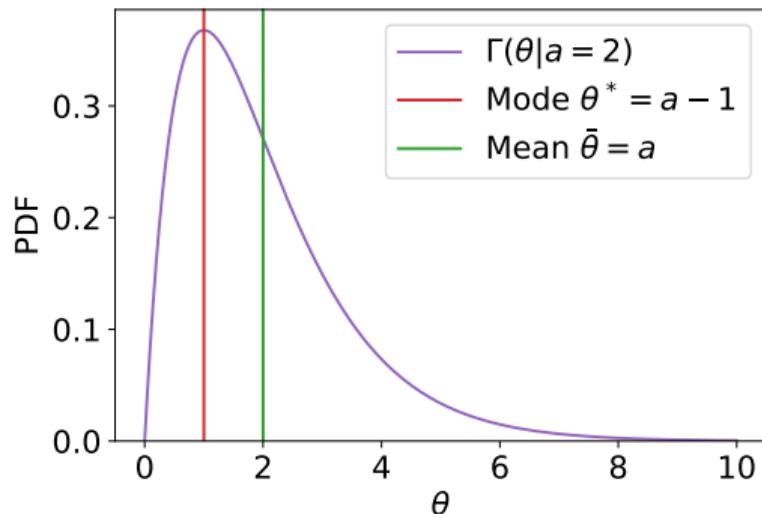
Value of  $\theta$  with maximum a-posteriori probability

$$\theta^* = \operatorname{argmax}_\theta p(\theta|x)$$

## Mean

Expected value of  $\theta$  under the posterior

$$\bar{\theta} = E_{p(\theta|x)}[\theta]$$



# Bayesian point estimates

## Mode

Value of  $\theta$  with maximum a-posteriori probability

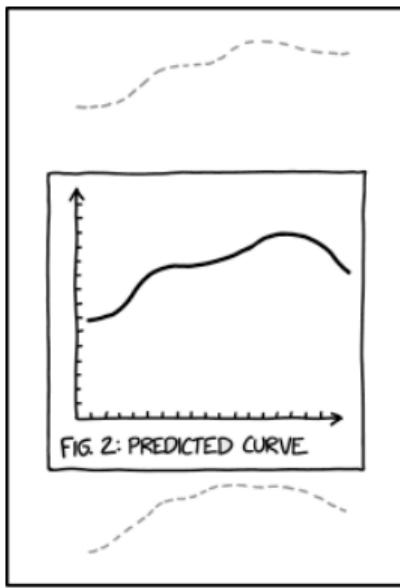
$$\theta^* = \operatorname{argmax}_\theta p(\theta|x)$$

Stationary point of the posterior at  $\theta^*$ :

$$0 = \frac{\partial p(\theta|x)}{\partial \theta} \Big|_{\theta=\theta^*} \propto \left( \frac{\partial p(x|\theta)}{\partial \theta} p(\theta) + p(x|\theta) \frac{\partial p(\theta)}{\partial \theta} \right) \Big|_{\theta=\theta^*}$$
$$\implies \theta^* = \hat{\theta} \quad \text{if} \quad \frac{\partial p(\theta)}{\partial \theta} \Big|_{\theta=\theta^*} = 0$$

Posterior mode and ML estimate agree for *flat* priors.

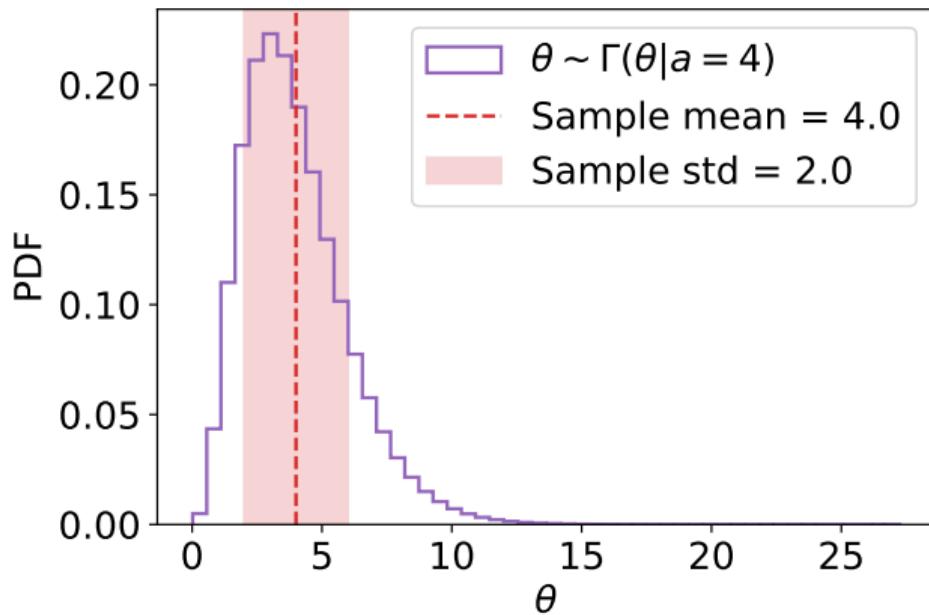
# Intervals and limits



SCIENCE TIP: IF YOUR MODEL IS  
BAD ENOUGH, THE CONFIDENCE  
INTERVALS WILL FALL OUTSIDE  
THE PRINTABLE AREA.

# Non-Normal PDFs

For non-normal estimator PDFs,  $\hat{\theta} \pm \sigma_{\hat{\theta}}$  can be misleading.



# Frequentist confidence intervals

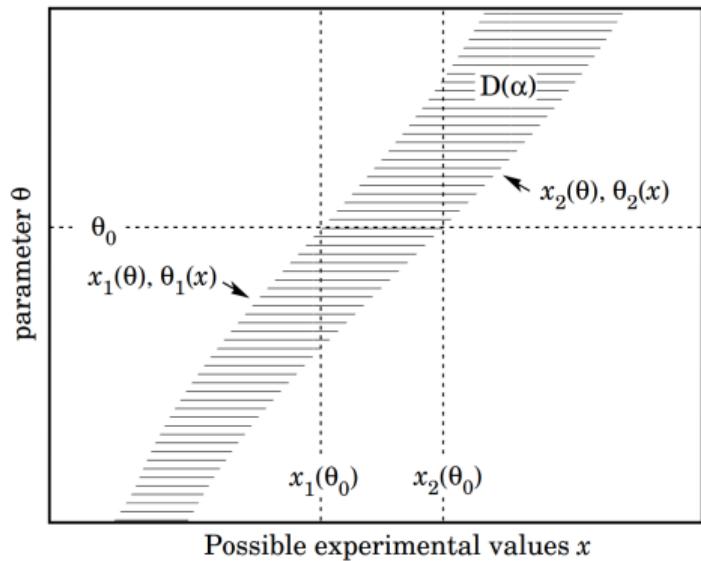
## Neyman confidence belt

$$\int_{x_1}^{x_2} dx \ p(x|\theta) = 1 - \alpha$$

Not unique  $\rightarrow$  central interval

$$\int_{-\infty}^{x_1} dx \ p(x|\theta) = \int_{x_2}^{\infty} dx \ p(x|\theta) = \alpha/2$$

or upper/lower interval

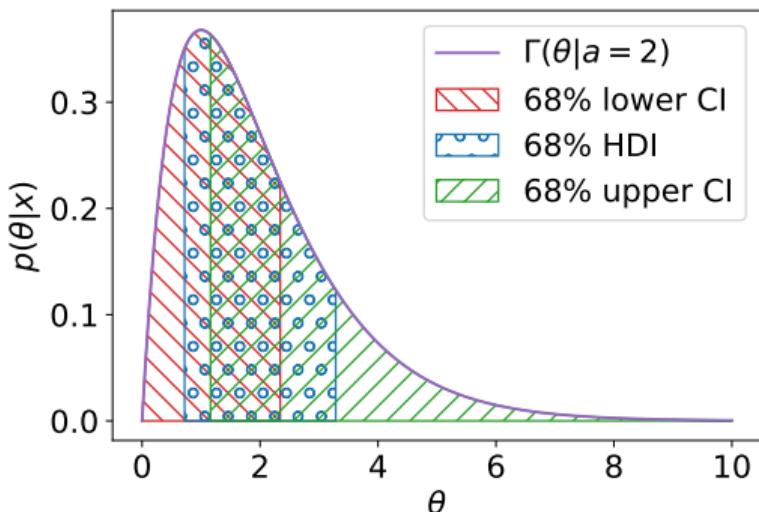


# Bayesian credible intervals

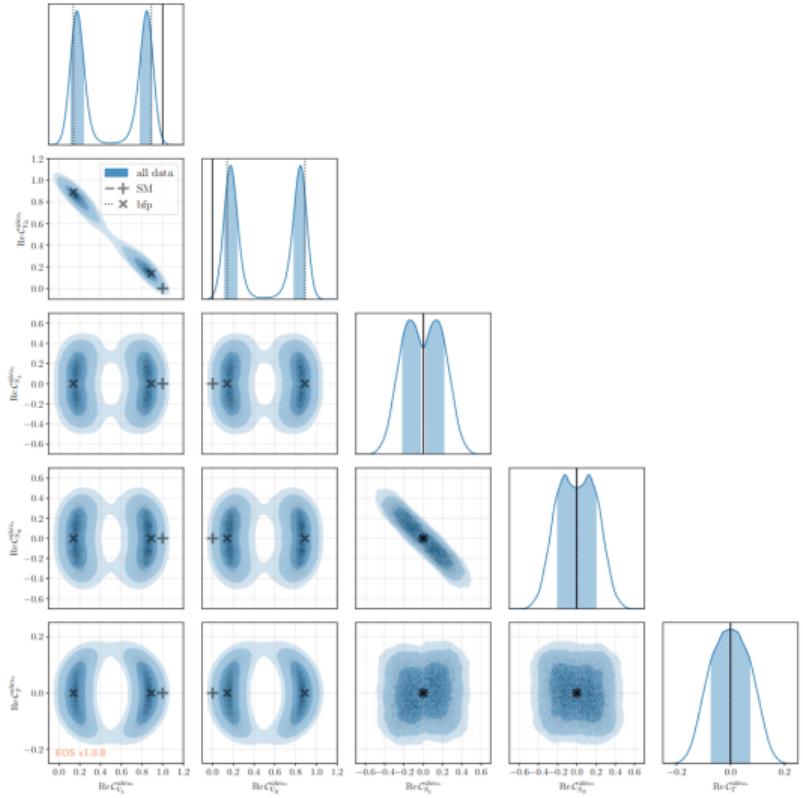
Credible intervals (CI)  $[\theta_1, \theta_2]$  cover  $1 - \alpha$  of the posterior

$$\int_{\theta_1}^{\theta_2} d\theta p(\theta|x) = 1 - \alpha$$

- Upper/lower CI
- Highest (posterior) density intervals (HDI)



# $b \rightarrow u l^- \bar{\nu}$ in the Weak Effective Theory



- Corner plots are great for visualization.
- Posterior for Wilson coefficients
- Identify modes, credible intervals

[arXiv:2302.05268v2 \[hep-ph\]](https://arxiv.org/abs/2302.05268v2)

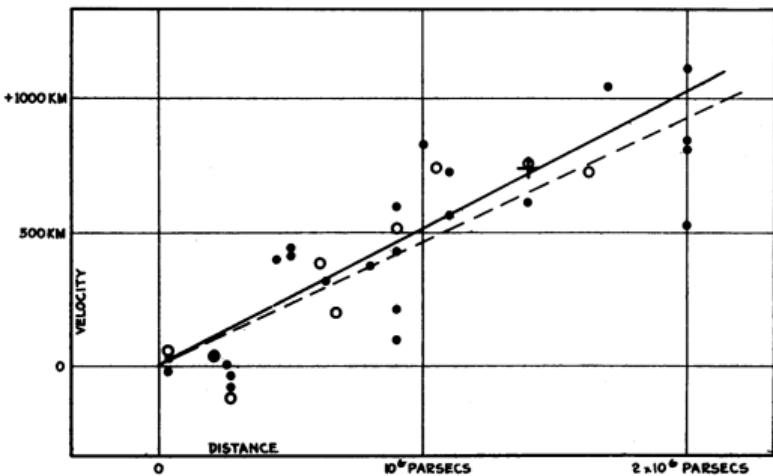
# Nuisance parameters and priors

# Nuisance parameters

Models are not perfect  
→ **systematic bias**

*Solution:*  
**Nuisance** parameters  $\nu$ ,

$$p(x|\psi, \nu)$$



Hubble 1929

Usually, we want to constrain nuisance parameters, but in the frequentist language **everything is data**.

# Frequentist "priors"

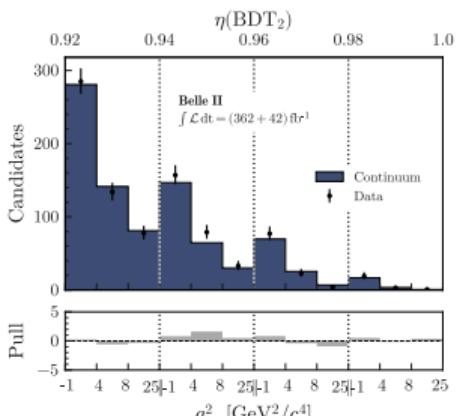
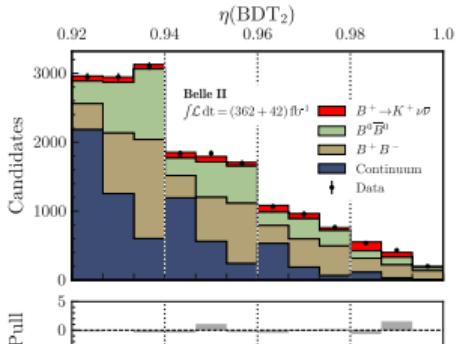
Constrain nuisance parameters using auxiliary data  $a$ ,

$$p(x|\psi, \nu)p(a|\nu)$$

Often auxiliary data is created to match our desired constraint term.

$p(a|\nu)$  represents degree of belief in  $\nu$ .

Belle II 2024



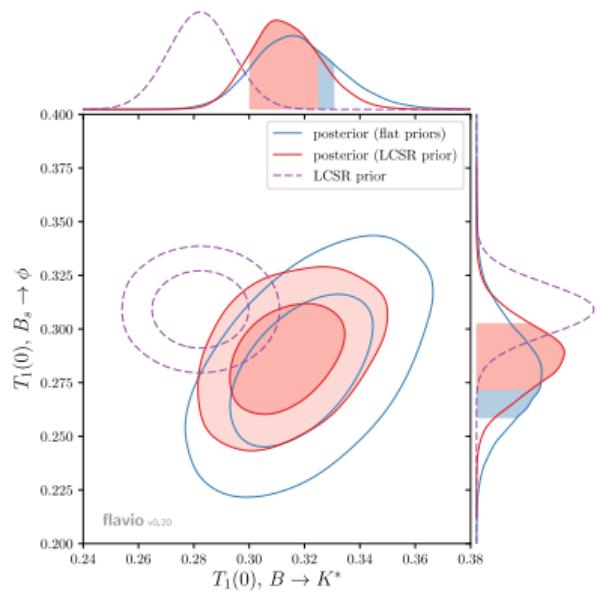
# Bayesian nuisance parameters

Priors can be defined using auxiliary data

$$p(\nu|a) \propto p(a|\nu)p_0(\nu)$$

Only in the Bayesian case, other prior choices are also allowed

$$p(\nu) = \mathcal{N}(\nu|\nu_0, \sigma_\nu^2)$$

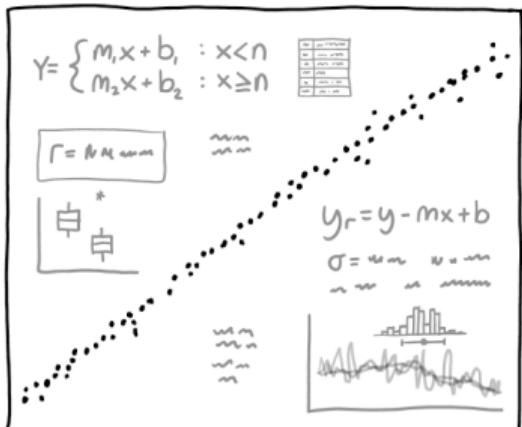


Paul 2017

# A simple linear model

## HOW TO DETECT A CHANGE IN THE SLOPE OF YOUR DATA

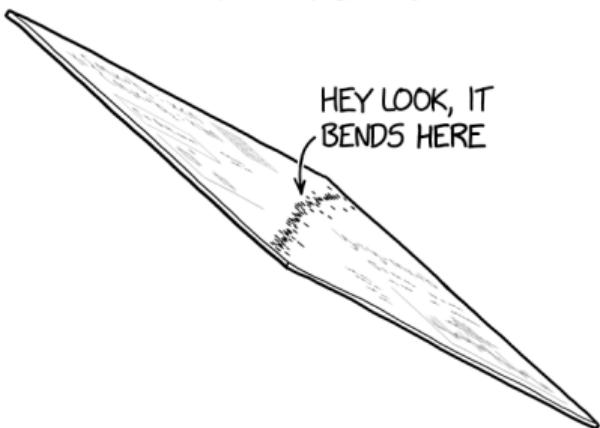
### NOVICE METHOD:



### EXPERT METHOD:

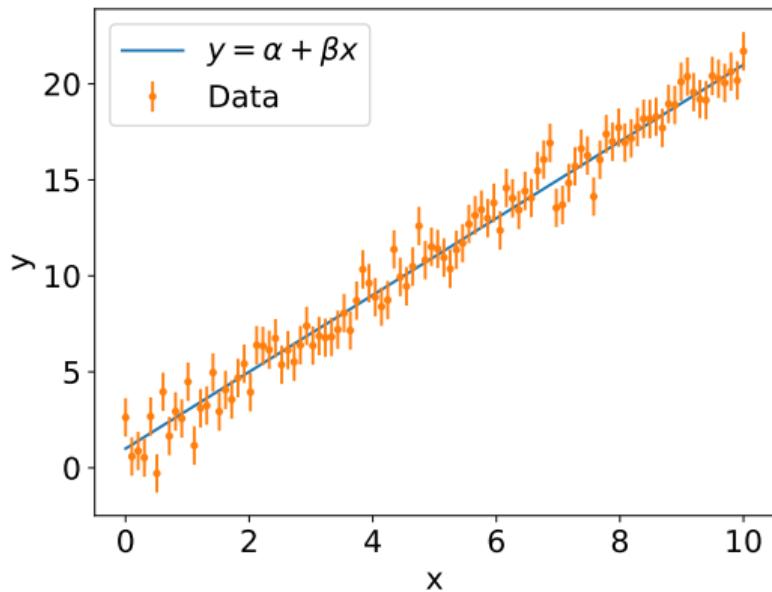
HEY LOOK, IT  
BENDS HERE

TIP THE GRAPH SIDEWAYS



# A simple linear model

- Independent data :  $\mathbf{X} = (x_i, y_i, \sigma_i)$



# A simple linear model

- Independent data:  $\mathbf{X} = (x_i, y_i, \sigma_i)$
- Model = product of normal distributions:

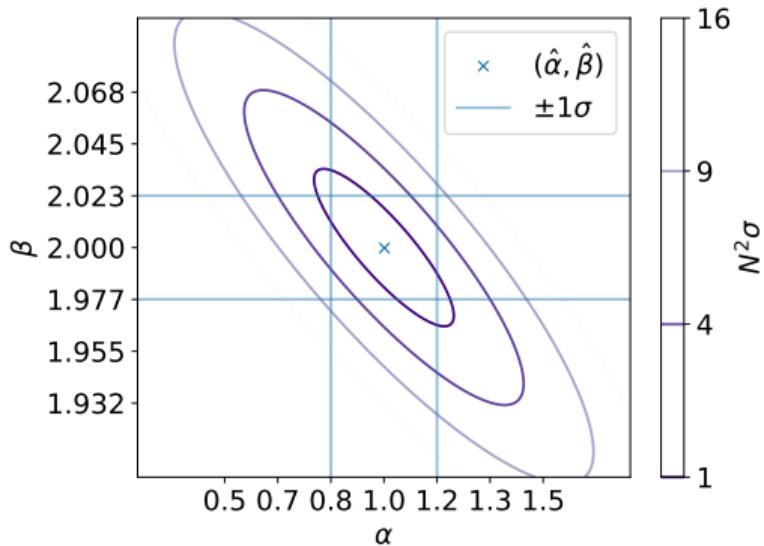
$$p(\mathbf{X}|\alpha, \beta) = \prod_{x_i, y_i, \sigma_i \in \mathbf{X}} \mathcal{N}(y_i | \mu(x_i | \alpha, \beta), \sigma_i^2)$$

$$\mu(x_i | \alpha, \beta) = \alpha + \beta x_i$$

- We want to know about  $\alpha$ , do not care about  $\beta$ .

# Frequentist analysis

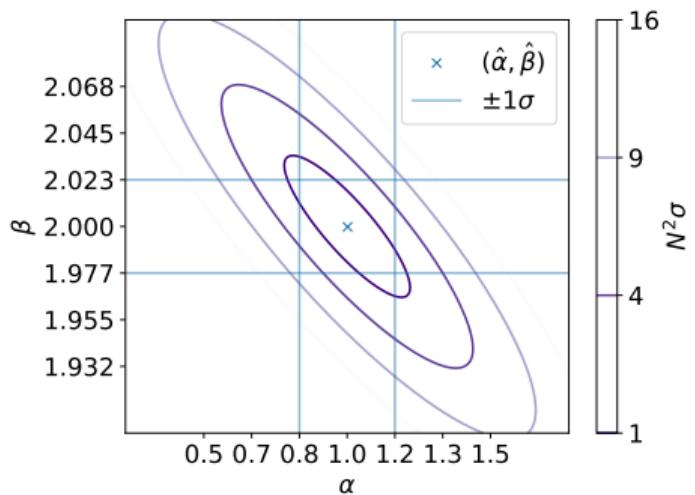
$$-2 \log p(\mathbf{X} | \alpha, \beta) = \sum_{x_i, y_i, \sigma_i \in \mathbf{X}} \frac{(y_i - \mu(x_i | \alpha, \beta))^2}{\sigma_i^2}$$



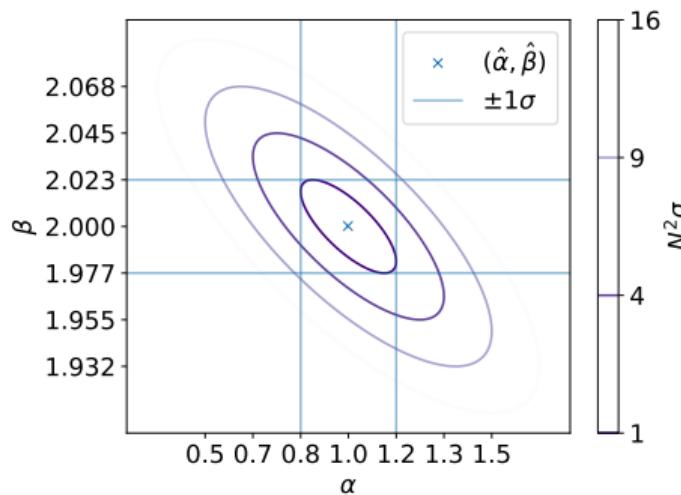
# Including a measurement of $\beta$ : $t, \sigma_t$

$$-2 \log p(\mathbf{X} | \alpha, \beta) = \sum_{x_i, y_i, \sigma_i \in \mathbf{X}} \frac{(y_i - \mu(x_i | \alpha, \beta))^2}{\sigma_i^2} + \frac{(\beta - t)^2}{\sigma_t^2}$$

unconstrained



constrained

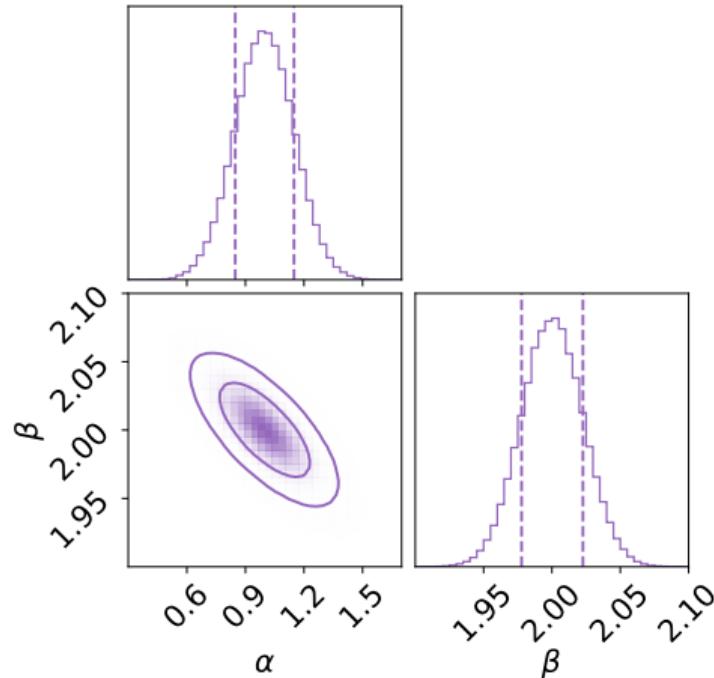


# Posterior

$$p(\alpha, \beta | \mathbf{X}) \propto p(\mathbf{X} | \alpha, \beta) p(\alpha) p(\beta)$$

$$p(\alpha) = \text{Uniform}(0, 2)$$

$$p(\beta) = \mathcal{N}(\beta | t_1, \sigma_{t_1}^2)$$

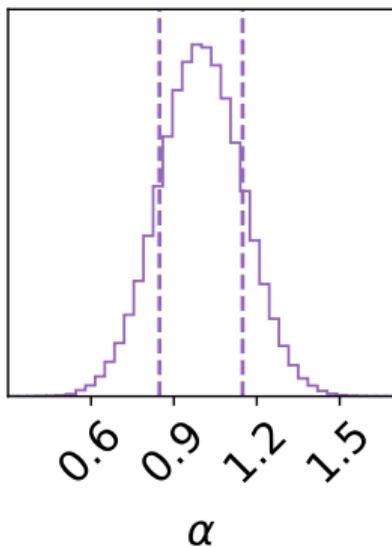


# Marginal posterior

$$p(\alpha|\mathbf{X}) = \int d\beta p(\alpha, \beta|\mathbf{X}) = \mathcal{N}(\alpha|\alpha^*, \sigma_\alpha)$$

In this example, we get

- $\alpha^* = \hat{\alpha}$
- 68% HDI =  $\hat{\alpha} \pm \sigma_\alpha$



# How do we marginalize?

- We only want the posterior for  $\psi$  alone.
  - Remove nuisance parameters by integrating over  $\nu$ .
- The *marginal posterior* is

$$p(\psi|x) = \int d\nu p(\psi, \nu|x)$$

- Commonly a high dimensional integral  
→ compute with Monte Carlo methods.

# Markov Chain Monte Carlo (MCMC)

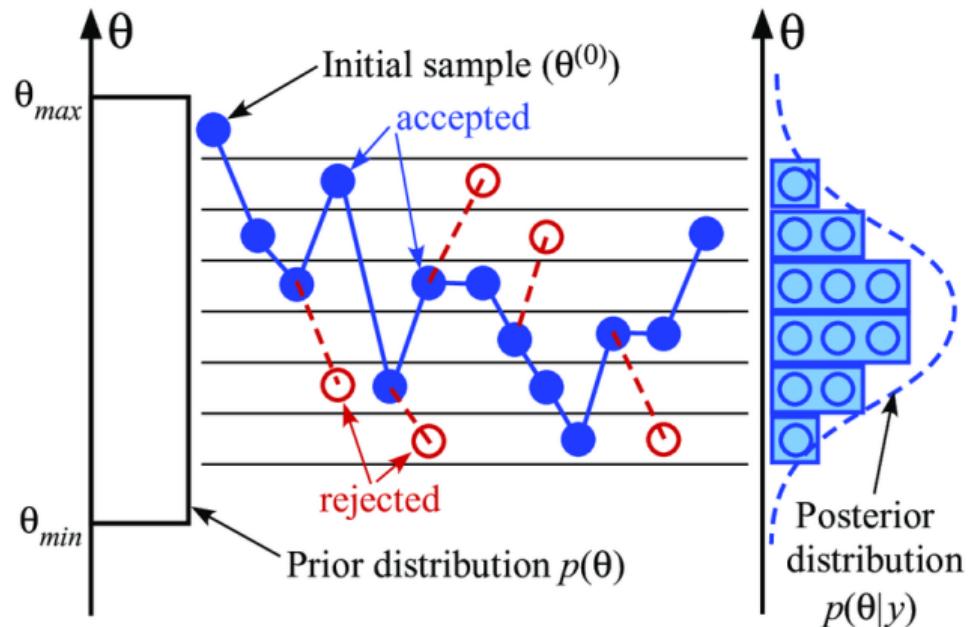
# Markov chain

A sequence of events, where probability of the next state depends solely on the current state

$$\dots \rightarrow \theta_i \sim g(\theta_i | \theta_{i-1}) \rightarrow \theta_{i+1} \sim g(\theta_{i+1} | \theta_i) \rightarrow \dots$$

for some *proposal distribution*  $g$ .

# MCMC integration



# Metropolis-Hastings

We loop

1. Generate  $\theta \sim g(\theta|\theta_i)$
2. Update

$$\theta_{i+1} = \begin{cases} \theta & u \leq \min\left(1, \frac{p(\theta)g(\theta|\theta_i)}{p(\theta_i)g(\theta_i|\theta)}\right) \\ \theta_i & \text{otherwise} \end{cases}$$

where  $u \sim \text{Uniform}(0, 1)$

Note: for example  $g(\theta|\theta_0) = \mathcal{N}(\theta|\theta_0, \sigma)$ .

# Chains

In MCMC we generate a sequence

$$\theta_0 \rightarrow \theta_1 \rightarrow \theta_2 \rightarrow \dots$$

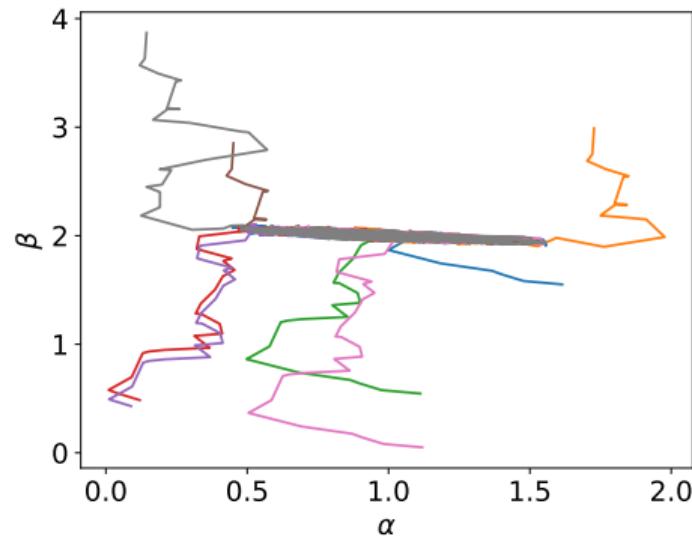
Only one start can land you  
in local minima.

$$\theta_0^0 \rightarrow \theta_1^0 \rightarrow \theta_2^0 \rightarrow \dots$$

$$\theta_0^1 \rightarrow \theta_1^1 \rightarrow \theta_2^1 \rightarrow \dots$$

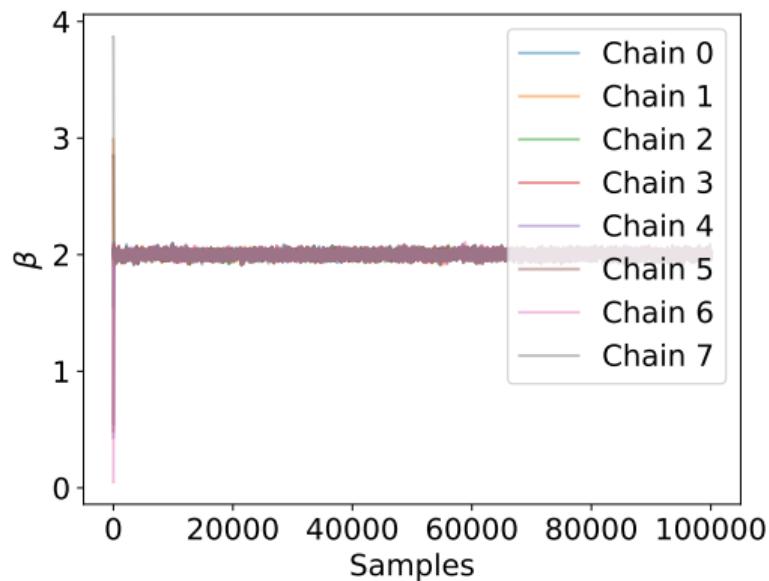
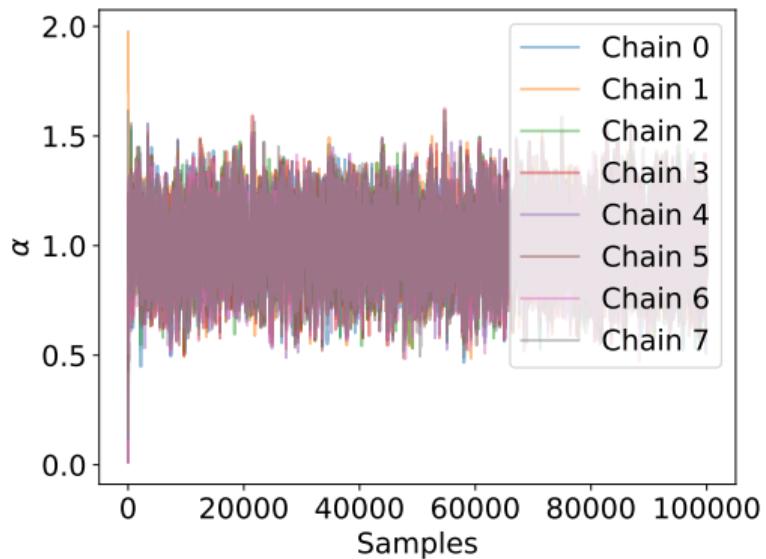
$$\theta_0^2 \rightarrow \theta_1^2 \rightarrow \theta_2^2 \rightarrow \dots$$

...



# Convergence

Trace plots are a useful convergence diagnostic

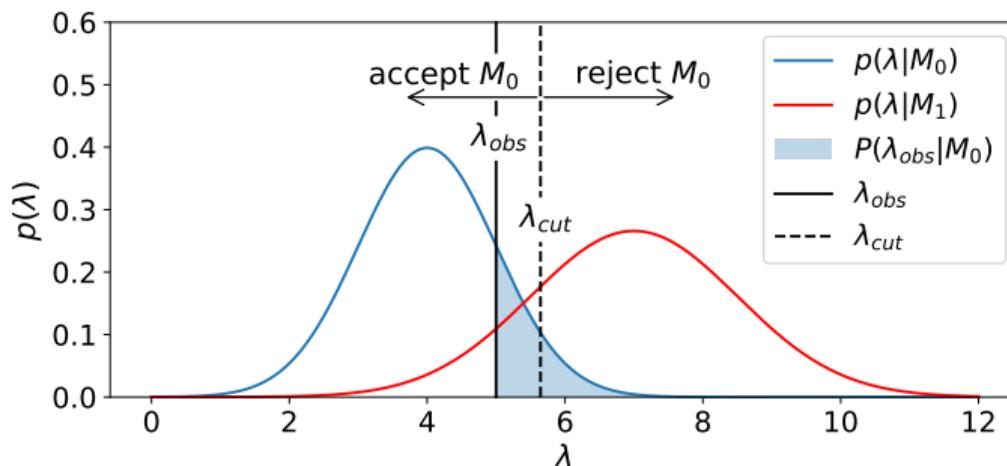


... but one can become more fancy.

# Model comparison

# Frequentist: P-values

$$P(\lambda_{obs}|M_0) = \int_{\lambda_{obs}}^{\infty} d\lambda p(\lambda|M_0), \quad \lambda = -2 \ln \frac{p(x|\hat{\theta}_0, M_0)}{p(x|\hat{\theta}_1, M_1)} \dagger$$



$\dagger$  Likelihood ratio = optimal test statistic  $\rightarrow$  Newman-Pearson lemma

# Averaged: Bayes factor

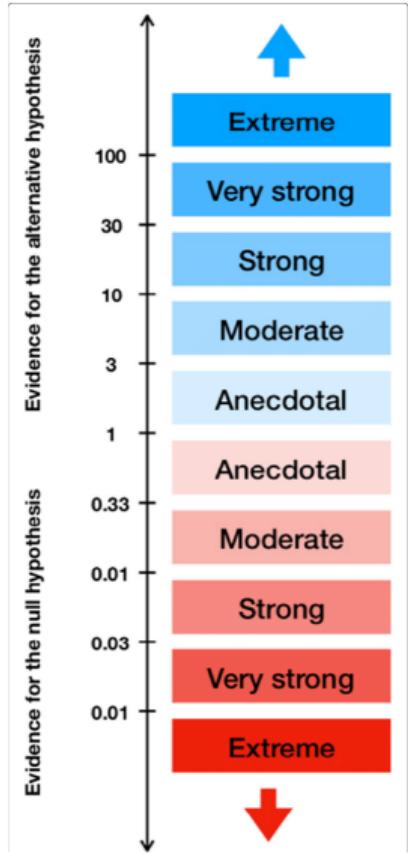
Compare the probabilities of the observed data being produced by a given model.

$$p(\theta|x, M) = \frac{p(x|\theta, M) p(\theta|M)}{p(x|M)}$$

$$p(x|M) = \int d^n\theta p(x|\theta, M) p(\theta|M)$$

$$B = \frac{p(x|M_1)}{p(x|M_0)}$$

*Do you see a potential hazard?*



# $b \rightarrow u l^- \bar{\nu}$ in the Weak Effective Theory

fit model $M$	goodness of fit			
	$\chi^2$	d.o.f.	$p$ value [%]	$\ln Z(M)$
SM	44.18	48	63.03	$372.5 \pm 0.4$
CKM	43.75	47	60.78	$372.4 \pm 0.4$
WET	36.13	43	76.17	$376.5 \pm 0.4$

**Table 1.** Goodness-of-fit values for the three main fits conducted as part of this analysis. We provide  $\chi^2 = -2 \ln P(\text{data} | \vec{x}^*)$  at the best-fit point  $\vec{x}^*$  next to the  $p$  value and the natural logarithm of the evidence  $\ln Z$ . We find that the  $p$  values associated with each individual likelihood are larger than 42%.

$$B = \exp(\ln Z(WET) - \ln Z(SM)) = 54.6$$

arXiv:2302.05268v2 [hep-ph]

# Tools to try

