A Hidden Gem: Unlocking the Power of Bayesian Inference in Particle Physics

Lorenz Gärtner[1], Toni Sculac, Judith Katzy

[1]*LMU Munich*

February 14, 2025

About me

- BSc @ University of Manchester
  Physics with theoretical physics
  Little stats
- MSc @ LMU Munich
  Thesis on QFT in curved spacetime
  Almost no stats
- Currenly PhD @ LMU Munich
  A lot of stats

What is a probability?

Kolmogorov probability axioms

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

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

Conditional probability

Is **defined** as the probability of an event x if we know that an event y is true $p(x|y)$.

$$p(x \cup y) = p(x|y)p(y) \quad \rightarrow \quad p(x|y) = \frac{p(x \cup y)}{p(y)}$$

Probability interpretations

The probability axioms and rules

- allow you to calculate new probabilities from old ones.
- do not tell you how to assign probabilities to begin with. For this we 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

$$p(x) = \lim_{n\to\infty} \frac{n_x}{n}$$

- Can only be assigned to repeatable experiments
- Not everything is repeatable...
- No probabilities of single events

Bayesian interpretation

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

- Probability depends on the experimenters' knowledge.
- Inference results are subjective.
- Everything is a random variable.

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

Can you derive it?

The common ground

Frequentist inference is based on

- the model of the observable data

$$\boldsymbol{p(x|\theta)}$$

Bayesian inference is based on

- the model of the observable data

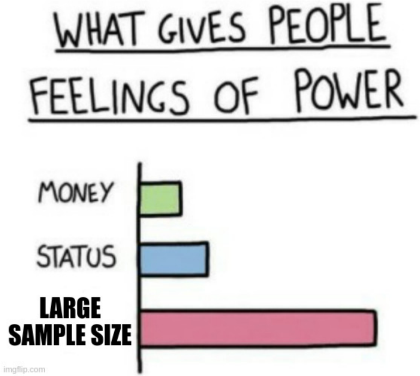
$$\boldsymbol{p(x|\theta)}$$

- your prior belief

$$p(\theta)$$

➜ You want the best possible $\boldsymbol{p(x|\theta)}$.

The best possible model...



Physics does not care...

... about our interpretation

For many inference problems, the frequentist and Bayesian approaches give similar numerical values, even though they answer different questions.

BUT if results are different, you should understand why.

Nuisance parameters and priors

Frequentist nuisance parameters

Models are not perfect ➜ **systematic bias**
Solution: Include additional **nuisance** parameters ν,

$$p(x|\theta, \nu).$$

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

… everything is data

"*The great advantage of the Bayesian approach is that it allows you to incorporate subjective beliefs, while the Frequentist approach pretends that you don´t have any.*"

– associated with Jim Berger by ChatGPT

... everything is data

Are we being honest here?

Prior knowledge in a typical frequentist analysis

- theory predictions
- model parameters
- missing higher-order corrections
- MC normalizations
- ...

Frequentist "priors"

Constrain nuisance parameters using "auxiliary data" a,

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

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

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

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, e.g.

$$p(\nu) = \mathrm{Gauss}(\nu|\nu_0, \sigma_\nu)$$

Bayesian updating

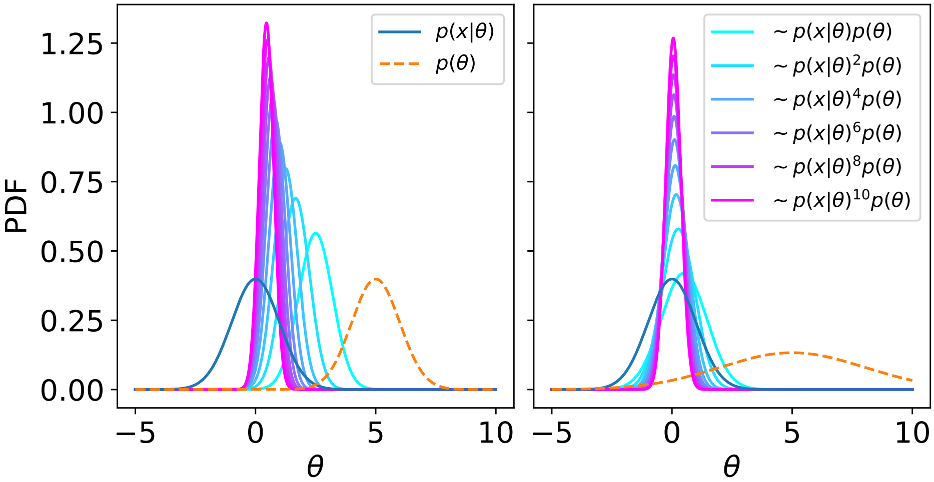
A posterior based on all LHC data x_{LHC}

$$p(\theta|x_{LHC}) = \frac{p(x_{LHC}|\theta)}{p(x_{LHC})}p(\theta)$$

can be updated with LHC-HL data x_{HL}, with $p(\theta|x_{LHC})$ as a prior

$$p(\theta|x_{LHC}, x_{HL}) = \frac{p(x_{HL}|\theta)}{p(x_{HL})}p(\theta|x_{LHC}) = \frac{p(x_{HL}|\theta)}{p(x_{HL})}\frac{p(x_{LHC}|\theta)}{p(x_{LHC})}p(\theta).$$

Bayesian updating



Parameter inference

Parameter inference

Point estimates

Identify the most probable parameter point.

Interval estimation

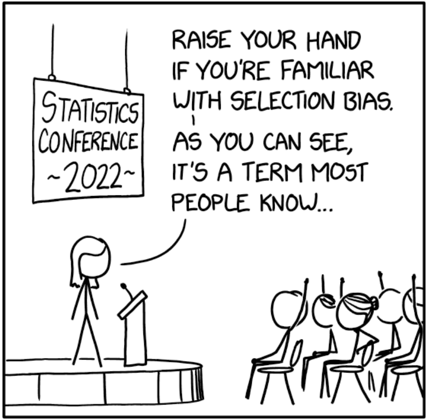
Identify extended regions in parameter space based on compatibility with the data.

Frequentist point estimates: estimators

Estimator is a *statistic* $\hat{\theta}(x)$, i.e. a well chosen function of the data.
Desired properties

- *consistency* $\lim_{N_x \to \infty} E(\hat{\theta}) = \theta_{true}$
- *unbiasedness* $b = E(\hat{\theta}) - \theta_{true}$
- *efficiency* (minimum variance)
- ...

A good data set is the key to success ...



Method of maximum likelihood

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

$$\left.\frac{\partial p(x|\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

With the property

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

This implies consistency, asymptotic unbiasedness and efficiency

$$\lim_{N\to\infty} V(\hat{\theta}) = I(\theta)^{-1} = E\left[\frac{\partial \ln p(x|\theta)}{\partial \theta}\right]^{-1}$$

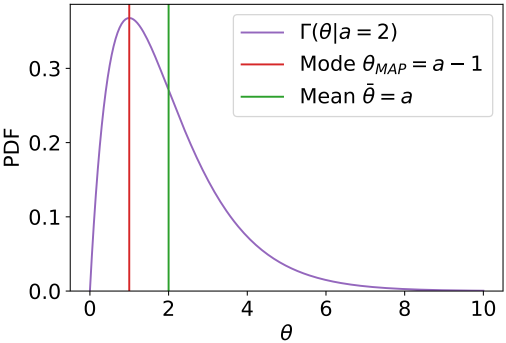
Bayesian point estimates

Mode
Value of θ with maximum a-posteriori probability

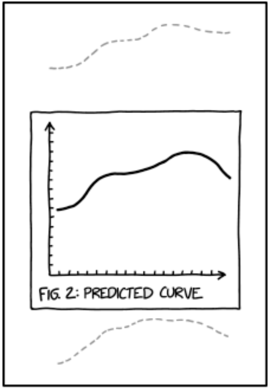
$$\theta^* = \text{argmax}_\theta \, p(\theta|x)$$

Mean
Expected value of θ under the posterior

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



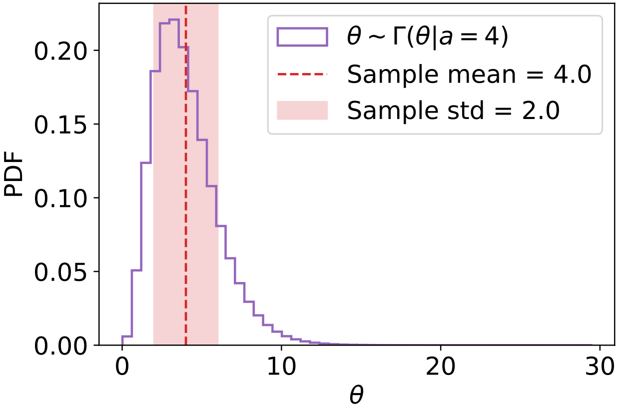
Intervals and limits



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

Intervals and limits

If an estimator PDF is not Normal, $\hat{\theta} \pm \sigma_\theta$ is meaningless.



Frequentist intervals

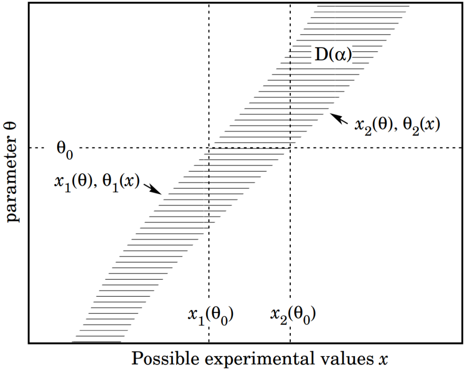
Neyman confidence belt

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

Not unique ➜ *central interval*

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

or *upper/lower interval*



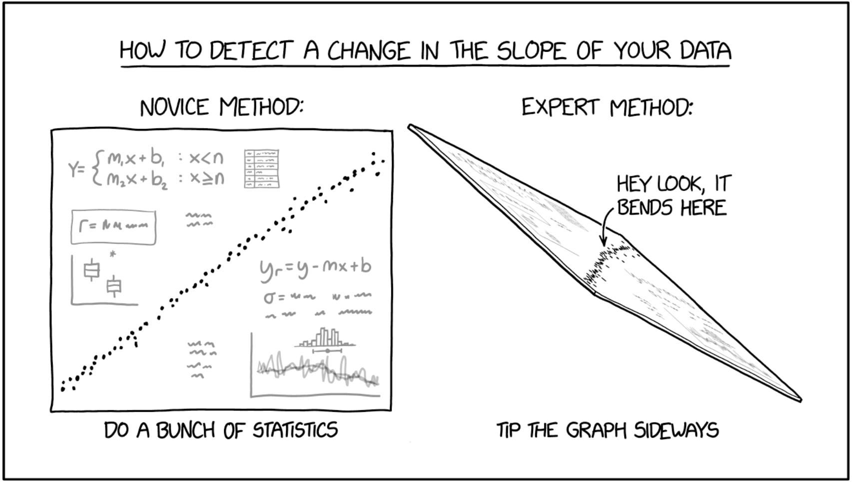
Bayesian intervals

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

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

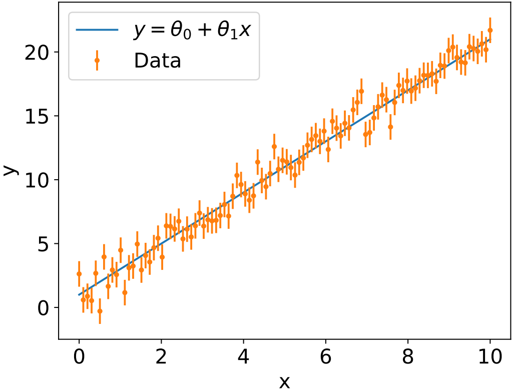
- For upper/lower limits: set θ_1 or θ_2 to boundary
- *Smallest possible interval*
  ➜ Highest (posterior) density intervals (HDI)

A simple linear model



A simple linear model

- Our independent data : x_i, y_i, σ_i



A simple linear model

- Our independent data: x_i, y_i, σ_i
- Our model:

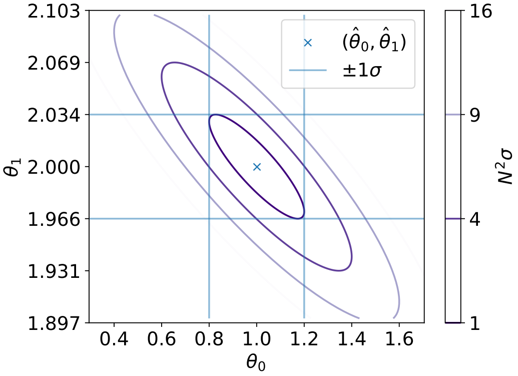
$$p(\mathbf{x}, \mathbf{y}|\theta_0, \theta_1) = \prod_{x_i, y_i \in \mathbf{x}, \mathbf{y}} \mathrm{Gauss}(y_i|\mu(x_i|\theta_0, \theta_1), \sigma_i)$$

$$\mu(x_i|\theta_0, \theta_1) = \theta_0 + \theta_1 x_i$$

➜ We want to know about θ_0, do not care about θ_1.

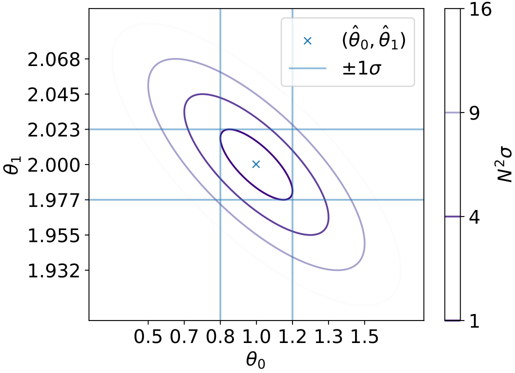
Unconstrained likelihood

$$-2\log p(\mathbf{x},\mathbf{y}|\theta_0,\theta_1) = \sum_{x_i,y_i\in\mathbf{x},\mathbf{y}} \frac{(y_i-\mu(x_i|\theta_0,\theta_1))^2}{\sigma_i^2}$$



Including a measurement of θ_1: (t_1, σ_{t_1})

$$-2\log p(\mathbf{x},\mathbf{y}|\theta_0,\theta_1) = \sum_{x_i,y_i\in\mathbf{x},\mathbf{y}} \frac{(y_i-\mu(x_i|\theta_0,\theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2}$$



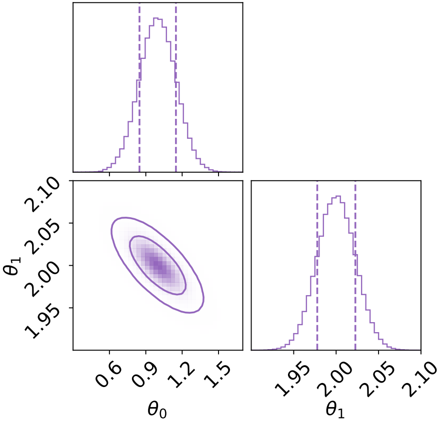
Posterior

$$p(\theta_0, \theta_1 | \mathbf{x}, \mathbf{y}) \propto p(\mathbf{x}, \mathbf{y} | \theta_0, \theta_1) p(\theta_0) p(\theta_1)$$

$$p(\theta_0) = \text{Uniform}(0, 2)$$

$$p(\theta_1) = \text{Gauss}(\theta_1 | t_1, \sigma_{t_1})$$

Corner plots are great for visualization.

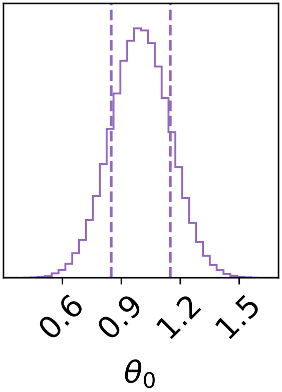


Marginal posterior

$$p(\theta_0|\mathbf{x},\mathbf{y}) = \int d\theta_1\, p(\theta_0,\theta_1|\mathbf{x},\mathbf{y}) = \mathrm{Gauss}(\theta_0|\hat{\theta}_0,\sigma_{\theta_0})$$

In this example, we get

- $\theta_0^* = \hat{\theta}_0$
- 68% CI = $\hat{\theta}_0 \pm \sigma_{\theta_0}$



MCMC

The hard part ...

- We only want the posterior for θ alone.
- Remove nuisance parameters by integrating over ν.

➜ The *marginal posterior* is

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

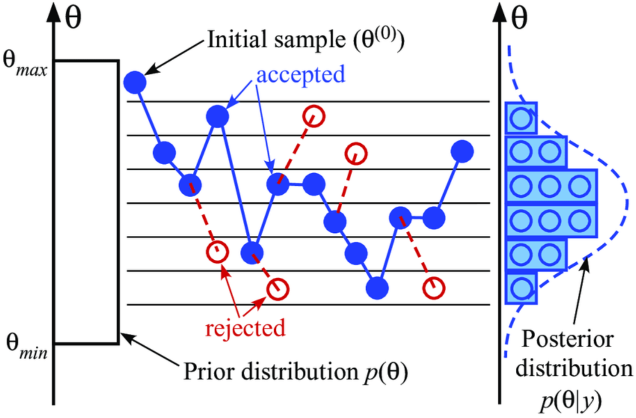
- Commonly a high dimensional integral
  ➜ compute with Monte Carlo methods.

Markov Process

The next element in the sequence (the Markov Chain) is proposed as a random variate θ_{i+1} of a PDF g that is conditioned on a location parameter, set to the current location θ_i,

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

Markov Chain Monte Carlo (MCMC)



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: We need to define a proposal distribution $g(\theta|\theta_0)$.

Chains

In MCMC we generate a sequence

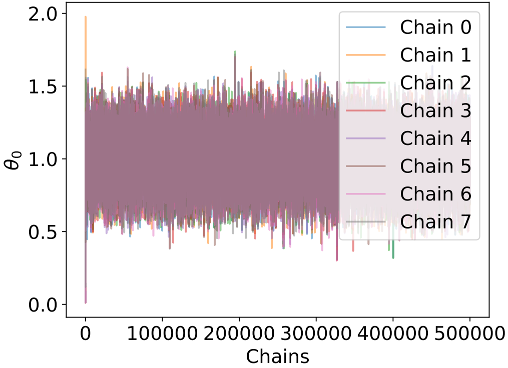
$$\theta_0 \to \theta_1 \to \theta_2 \to \dots$$

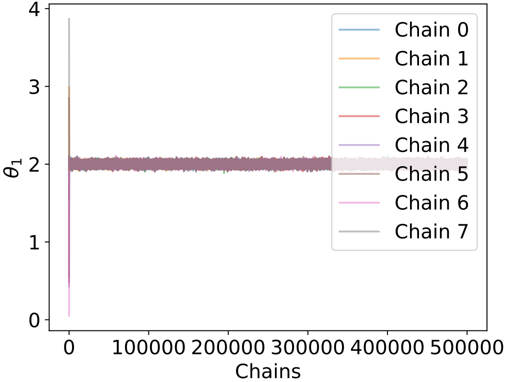
Only starting at one point can land you in local minima, hence often we sample

$$\begin{aligned}
&\theta_0^0 \to \theta_1^0 \to \theta_2^0 \to \dots \\
&\theta_0^1 \to \theta_1^1 \to \theta_2^1 \to \dots \\
&\theta_0^2 \to \theta_1^2 \to \theta_2^2 \to \dots \\
&\dots
\end{aligned}$$

Convergence

Trace plots are a useful convergence diagnostic





... but one can become more fancy.

Tools to try

PyMC

BAT
BAYESIAN ANALYSIS TOOLKIT

ArviZ