

Bayesian inference in particle physics

Lorenz Gärtner¹, Toni Sculac, Judith Katzy

¹*LMU Munich*

February 11, 2025

What is a probability?

Kolmogorov's 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, \dots ,

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

Conditional probability



Is **defined** as the probability of an event x if we know that an event y is true

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

The probability axioms and probability rules allow you to calculate new probabilities from old ones.

But axioms 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.

Assign a probability as relative frequency

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

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

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

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

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

The common ground

Frequentist inference is based on

- the model of the observable data

$$p(x|\theta)$$

Bayesian inference is based on

- the model of the observable data

$$p(x|\theta)$$

- your prior belief

$$p(\theta)$$

Physics does not care...



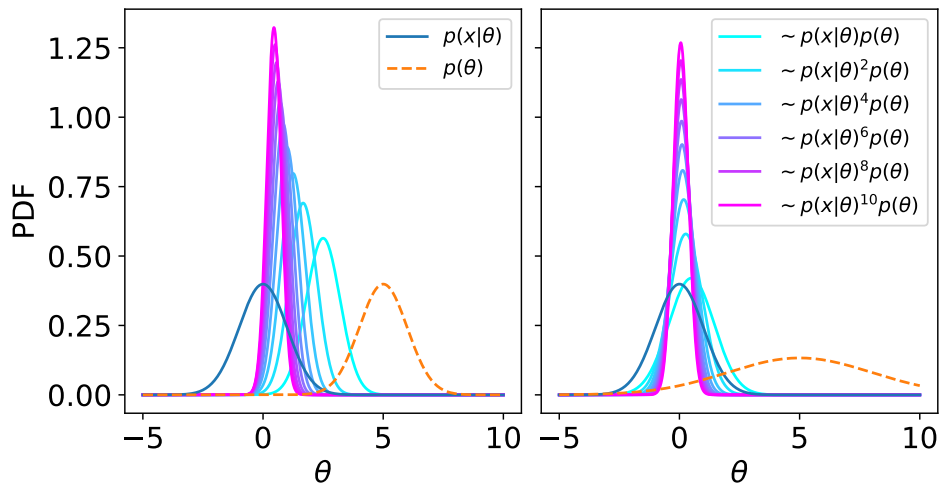
... about our interpretation

For many inference problems, the frequentist and Bayesian approaches give similar numerical values, even though they answer different questions and are based on fundamentally different interpretations of probability. Pour all energy into constructing the best possible $p(\mathbf{x}|\theta)$

Parameter values θ under which data x_1, x_2 is more probable on average get weighted up, other values get weighted down.

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

Bayesian updating



Nuisance parameters and priors

Models are not perfect → **systematic bias**

More general models include additional **nuisance** parameters, ν ,

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

Leads to increased statistical uncertainties on POIs (due to correlation)

Want to constrain 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 (masses, couplings, ...)
- missing higher-order corrections
- MC normalizations
- ...

Constrain nuisance parameters using "auxiliary data" a ,

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

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

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

If one has auxiliary data, the prior is

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

If $p_0(\nu)$ is chosen to have minimal impact, this overlaps with the frequentist treatment.

In the Bayesian case, other prior choices are also allowed, e.g.

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

Bayesian approach also requires priors for POIs.

Parameter inference

Point estimates

Identify the most probable parameter point.

Interval estimation

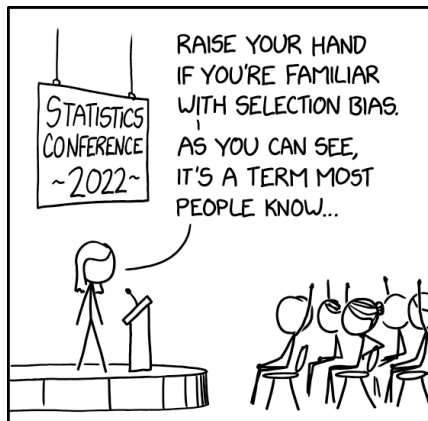
Identify extended regions in parameter space based on compatibility with the data.

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

Desired properties

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

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



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

$$\frac{\partial \ln L(x|\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0$$

With the property

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

This implies consistency, asymptotic unbiasedness and efficiency

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

Mode / MAP

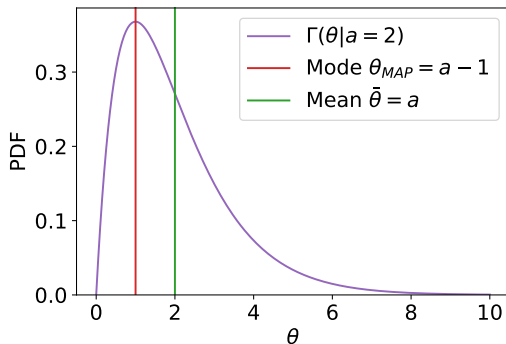
Value of θ with maximum a-posteriori probability

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

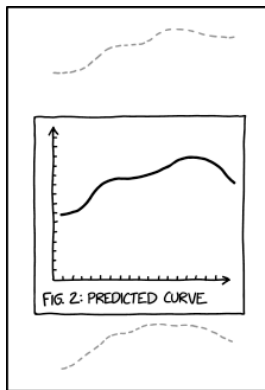
Mean

Expected value of θ under the posterior

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

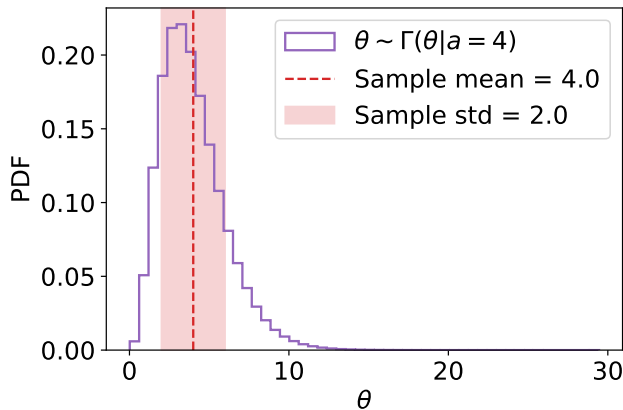


Intervals and limits



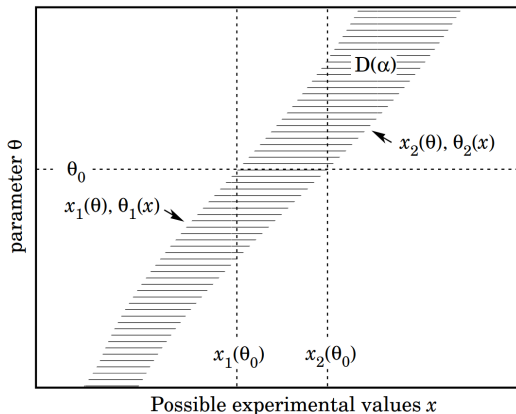
SCIENCE TIP: IF YOUR MODEL IS
BAD ENOUGH, THE CONFIDENCE
INTERVALS WILL FALL OUTSIDE
THE PRINTABLE AREA.

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



Neyman confidence belt

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



Credible intervals $[\theta_l, \theta_u]$ cover $1 - \alpha$ of the posterior

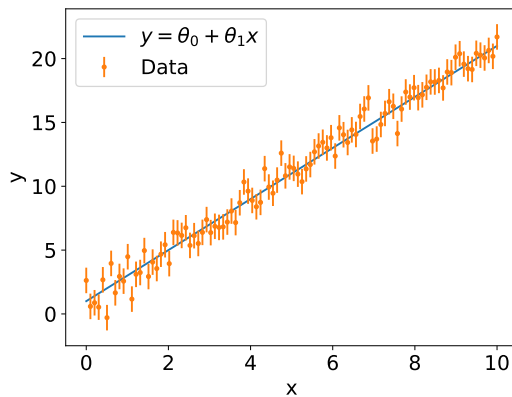
$$1 - \alpha = \int_{\theta_l}^{\theta_u} d\theta p(\theta, x)$$

- For upper/lower limits: set θ_l or θ_u manually
- Highest (posterior) density intervals (HDI): smallest possible interval

Example



- Our independent data : x_i, y_i, σ_i



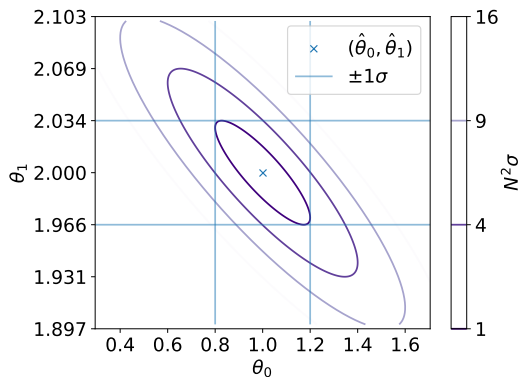
- Our independent data: x_i, y_i, σ_i
- Our model:

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

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

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

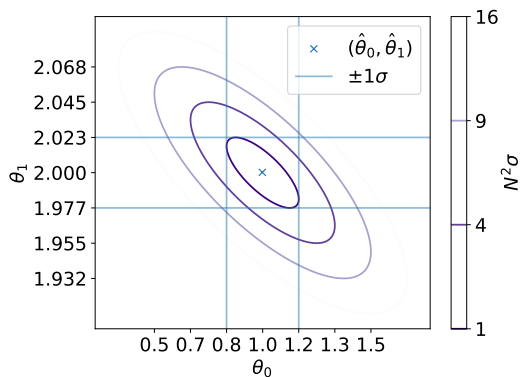
$$-2 \log p(\mathbf{x}|\theta_0, \theta_1) = \sum_{x_i \in \mathbf{x}} \frac{(x_i - \mu(x_i|\theta_0, \theta_1))^2}{\sigma_i^2}$$



Constrained likelihood



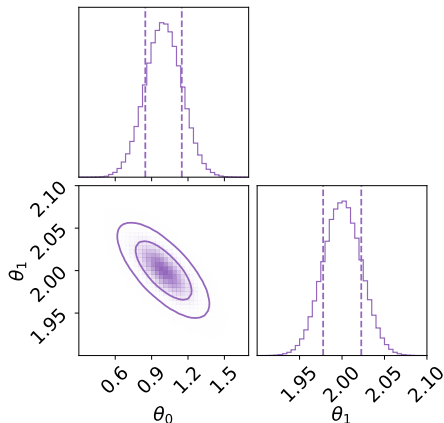
$$-2 \log p(\mathbf{x}|\theta_0, \theta_1) = \sum_{x_i \in \mathbf{x}} \frac{(x_i - \mu(x_i|\theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2}$$



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

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

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



Marginal posterior

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

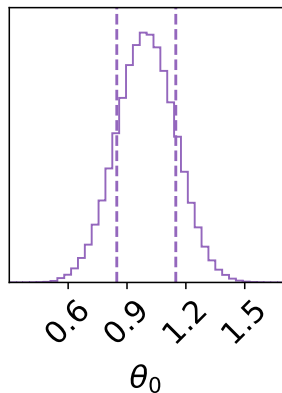
In this case, we get:

- Same $\hat{\theta}_0$ as for ML

$$\hat{\theta}_0 =$$

- Same σ_{θ_0} as for ML

$$\sigma_{\theta_0}^2 = \sum_{x_i, \sigma_i \in \mathbf{x}, \sigma} \sigma_{t_1} x_i^2 + \sigma_i^2$$



MCMC

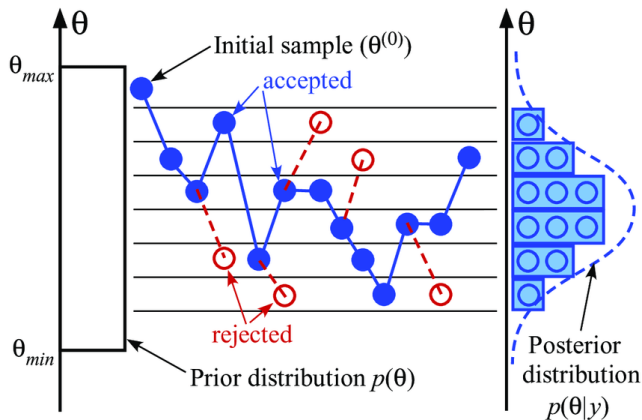
The hard part ...

- Usually, we are not interested in the nuisance parameters ν .
- One can obtain the posterior p.d.f. for θ alone by integrating over ν .
- The *marginal posterior* is

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

- High dimensional integral \rightarrow compute with Monte Carlo methods.

Markov Chain Monte Carlo (MCMC)



1. Generate $\theta \sim g(\theta|\theta_0)$
2. Form Hastings test ratio $\alpha = \min \left(\frac{p(\theta)g(\theta|\theta_0)}{p(\theta_0)g(\theta_0|\theta)} \right)$
3. Generate $u \sim \text{Uniform}(0, 1)$
4. If $u \leq \alpha$, let $\theta_1 = \theta$. Otherwise, let $\theta_1 = \theta_0$
5. Set $\theta_0 = \theta_1$ and return to step 1.

In MCMC we generate a sequence

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

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

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

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

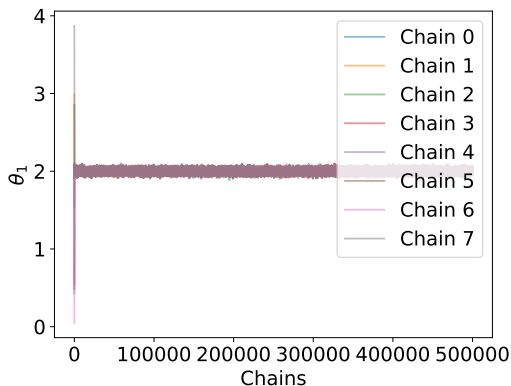
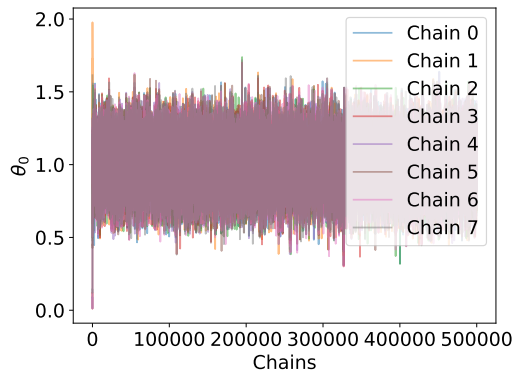
$$\theta_0^2 \rightarrow \theta_1^2 \rightarrow \theta_2^2 \rightarrow \dots$$

...

Convergence



Trace plots are a useful convergence diagnostic



... but one can become more fancy.

Tools to try

arviz, corner, pymc

a simple MCMC algorithm

Exercises: s+b problem



You do a binned analysis. You observe $[155, 121, 13]$ in each bin respectively. From simulation, you expect to see $[90, 30, 0]$ signal events (where the last bin acts as a control measurement) and $[50, 70, 10]$ background events.

Hint: It is better to work with log-probabilities

Exercises: s+b problem



- Construct a likelihood function (chi2)
- What is the best fit point?
- What are the frequentist limits?
- What would a sensible choice of priors be for this example?
- Use an implementation of the Metropolis Hastings algorithm to sample from the posterior
- Make a corner plot of the posterior
- How do the mode of the posterior and the credible intervals compare to the best fit point and frequentist limits?

Solutions to exercises