

Computational Cognitive Science

Lecture 5: Bayesian Parameter Estimation 1

Frank Keller

5 October 2026

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Slide credits: Chris Lucas, Benjamin Peters

From Likelihood to Bayesian Inference

Information in the Prior

Analytic Posterior: Biased Coin

Interpreting the Prior

Reading: Chapter 6 of [Farrell and Lewandowsky \(2018\)](#).

From Likelihood to Bayesian Inference

Why likelihood is not enough

Last time, we used likelihood to estimate parameters:

$$L(\theta|\mathbf{y}) \propto P(\mathbf{y}|\theta)$$

- Likelihoods allow **relative comparisons** between parameter values.
- But likelihoods are not probabilities over parameters.
- A point estimate, such as the maximum likelihood estimate (MLE), does not tell us how uncertain we are.

For example: if an experiment gives a working-memory estimate of 3.2 items, we want to know the **likely range** of the true value, not just the best-fitting value.

The problem Bayes' theorem solves

In modeling, likelihood asks:

$P(\mathbf{y}|\boldsymbol{\theta})$ How probable are the data if the parameter is fixed?

Bayesian inference asks:

$P(\boldsymbol{\theta}|\mathbf{y})$ How probable are parameter values after seeing the data?

Bayes' theorem lets us reverse the direction of conditionality.

Bayes' theorem applied to parameters

$$\underbrace{P(\theta|\mathbf{y})}_{\text{posterior}} = \frac{\overbrace{P(\mathbf{y}|\theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(\mathbf{y})}_{\text{evidence}}}$$

Prior: knowledge or assumptions about parameters before observing data.

Likelihood: probability of the observed data under a parameter value.

Evidence: overall probability of the data, used for normalization.

Posterior: updated probability distribution over parameters.

If we maximize the posterior, we get the maximum a posteriori (MAP) estimate.

The evidence is a marginal likelihood

The denominator is the probability of the data across all possible parameter values.

For discrete parameter values:

$$P(\mathbf{y}) = \sum_{\theta} P(\mathbf{y}, \theta) = \sum_{\theta} P(\mathbf{y}|\theta)P(\theta)$$

For continuous parameters:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta) d\theta$$

This integral is often the difficult part of Bayesian inference.

Information in the Prior

How much information is in the prior?

To build an intuition, let's start with a baseball example. We have batting data from Johnnie and Jane:

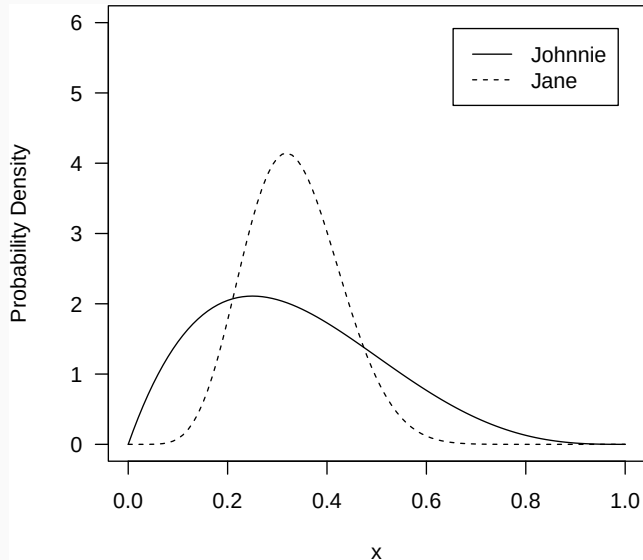
- Johnnie: 2 hits in 6 attempts.
- Jane: 8 hits in 24 attempts.

Both batting averages are .33, but we should be more confident about Jane, as we have more data for her.

To represent this, we can use the **Beta distribution**: the same mean can be paired with a broad or a narrow distribution.

More observations mean a more concentrated distribution.

How much information is in the prior?



Analytic Posterior: Biased Coin

The textbook example: a biased coin

Now let's look this motivating example in detail: **Bernoulli likelihood and Beta prior**.

Suppose a coin-like process has probability θ of producing heads.

For a single toss $k \in \{0, 1\}$:

$$f(k|\theta) = \theta^k (1 - \theta)^{1-k}$$

For n independent tosses with h heads:

$$f(h|\theta) = \theta^h (1 - \theta)^{n-h}$$

If the data are fixed and θ varies, this is the likelihood function.

A prior over probabilities

We need a probability distribution over possible values of θ .

For a Bernoulli process, the natural prior is the **Beta distribution**:

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

- Its domain is $[0, 1]$, so it can describe probabilities.
- α and β can be interpreted as prior successes and failures.
- It can capture both central tendency and amount of prior information.
- $B(\alpha, \beta)$ is a normalizing constant (ensures probabilities sum to 1).

A prior for the biased coin

For the coin example, suppose we expect a slight bias around $\theta = .5$.

For a Beta-distributed variable:

$$E(X) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

If we want the prior centered on .5, choose $\alpha = \beta$.

The textbook uses:

$$P(\theta) = \text{Beta}(\theta|12, 12)$$

Time for a short quiz on WooClap!

Go to wooclap.com

Enter the event code in the top banner

Event code

LURCYCH



Combining likelihood and prior

Bayes' theorem for the biased coin:

$$P(\theta|h, n) = \frac{f(h|\theta)f(\theta)}{P(h, n)} = \frac{\theta^h(1 - \theta)^{n-h} \text{Beta}(\theta|\alpha, \beta)}{P(h, n)}$$

Substituting the Beta prior and dropping the denominator:

$$P(\theta|h, n) \propto \theta^h(1 - \theta)^{n-h}\theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

Therefore:

$$P(\theta|h, n) \propto \theta^{\alpha+h-1}(1 - \theta)^{\beta+n-h-1}$$

Conjugacy

The posterior has the same form as the prior:

$$P(\theta|h, n) = \text{Beta}(\theta|\alpha + h, \beta + n - h)$$

When the prior and posterior belong to the same distribution family, the prior is **conjugate** to the likelihood.

For the Bernoulli likelihood:

Beta prior + Bernoulli data $\Rightarrow$ Beta posterior

Interpreting the Prior

Estimating the bias of a coin

Our prior belief is $P(\theta) = \text{Beta}(\theta|12, 12)$, which is centered on $\theta = 0.5$.

We observe three datasets:

$$\{14, 26\}, \quad \{113, 213\}, \quad \{1130, 2130\}$$

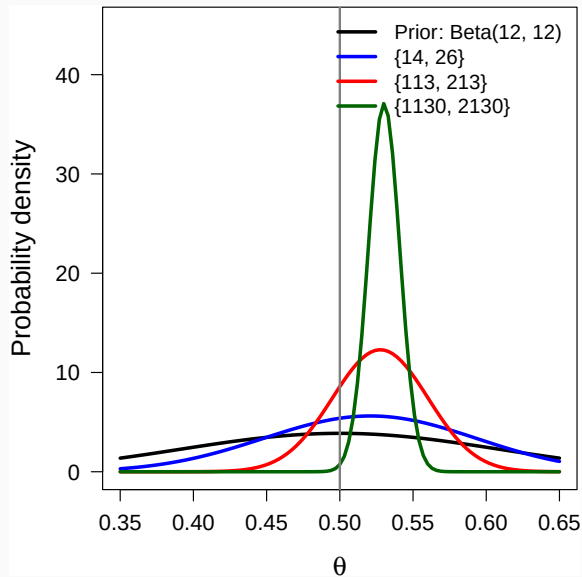
Their observed proportions of heads are almost identical:

$$0.538, \quad 0.531, \quad 0.531$$

Because the Beta prior is conjugate to the Bernoulli likelihood:

$$P(\theta|h, n) = \text{Beta}(\theta|12 + h, 12 + n - h)$$

How much information is in the prior?



More data means greater precision

Although the observed proportion of heads is almost the same in all three datasets:

- with little data, the posterior is relatively broad;
- with more data, the posterior becomes increasingly concentrated;
- the influence of the prior decreases as the amount of data increases.

The posterior means are

0.520, 0.527, 0.530

approaching the observed proportion as n increases.

Interpreting the posterior

For the largest dataset, the posterior is tightly concentrated above $\theta = 0.5$.

A 95% credible interval is approximately

$$0.51 < \theta < 0.55$$

This can be interpreted directly:

There is 95% posterior probability that θ lies between 0.51 and 0.55.

Bayesian estimation gives both a parameter estimate and its uncertainty.

This is the main difference from MLE: the posterior is interpretable as uncertainty about θ , not merely as relative fit.

Bayesian summaries

Once we have the posterior, we can summarize it in several ways.

- Posterior mean:

$$E[\theta|\mathbf{y}]$$

- Posterior mode, i.e., MAP estimate:

$$\operatorname{argmax}_{\theta} p(\theta|\mathbf{y})$$

- Credible interval:

$$P(a < \theta < b|\mathbf{y}) = .95$$

- Posterior predictive probability:

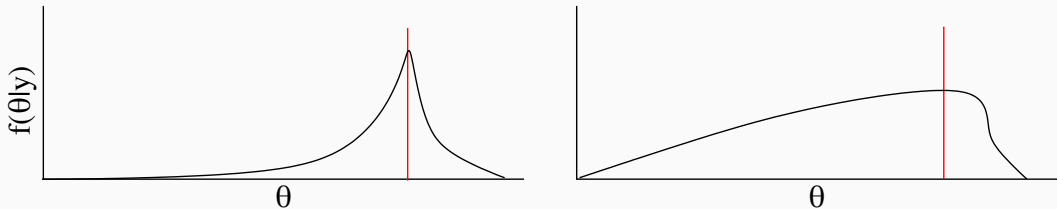
$$P(x_{N+1}|\mathbf{y})$$

Point estimate vs posterior distribution

The MAP estimate does not account for uncertainty or the shape of $p(\theta|\mathbf{y})$.

If our prior is uniform, MAP = MLE.

Compare two densities, where θ is the bias of a coin toward heads:



- The mode of the posterior is the same for both.
- But the uncertainty and posterior mean can differ.

- Bayesian estimation turns likelihood plus prior into a posterior distribution.
- The evidence normalizes the posterior by marginalizing over parameter values.
- For Bernoulli data, the Beta prior gives an analytic Beta posterior.
- MAP is only one summary of the posterior; Bayesian inference uses the whole distribution.
- Prior choice matters and should be explicit.

Farrell, Simon and Stephan Lewandowsky. 2018. *Computational Modeling of Cognition and Behavior*. Cambridge University Press, Cambridge.