Bayesian Inference

Priors, Posteriors, Shrinkage, Evidence, and Decisions

We Have Already Used a Prior Once

Deck 04 shrank $\bar{X}$ toward zero, $\hat{\mu}_2 = c\bar{X}$, and showed that a little bias can lower the MSE — the idea behind ridge regression. That estimator is exactly **maximum a posteriori** estimation under a Gaussian prior centred at zero, and the deck never said so.

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \underbrace{p(y \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

What that slide did not say

The MAP estimate is one number read off a **whole distribution**. That distribution — the posterior — carries the uncertainty as well, and it answers questions a point estimate cannot.

Regularization was the posterior's *mode*. This deck keeps the rest of it.

Bayes' Theorem for Parameters

Definition — Posterior Distribution

For data y and parameter θ,

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}, \qquad p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta$$

The denominator does not involve θ, so in practice

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta).$$

In words

Prior belief about θ, reweighted by how well each θ predicted what you actually saw.

The shift

θ is now a **random variable** — not because it moves, but because we are uncertain about it. Probability describes belief, not frequency.

Conjugate Updating

The Beta–Binomial Model

Estimating a proportion: a take-up rate, an unemployment share, a default probability.

Model

$$k \mid \theta \sim \text{Binomial}(n, \theta)$$

$$\theta \sim \text{Beta}(\alpha, \beta)$$

Posterior

$$\theta \mid k \sim \text{Beta}(\alpha + k,\ \beta + n - k)$$

The posterior is in the same family as the prior — the prior is **conjugate**.

The prior has an exchange rate

$\text{Beta}(\alpha, \beta)$ contributes exactly as much as α prior successes and β prior failures. A $\text{Beta}(2, 2)$ prior is worth four observations; $\text{Beta}(50, 50)$ is worth a hundred.

Conjugate Updating, Live

Prior

Uniform — Beta(1,1)

Observations n

20

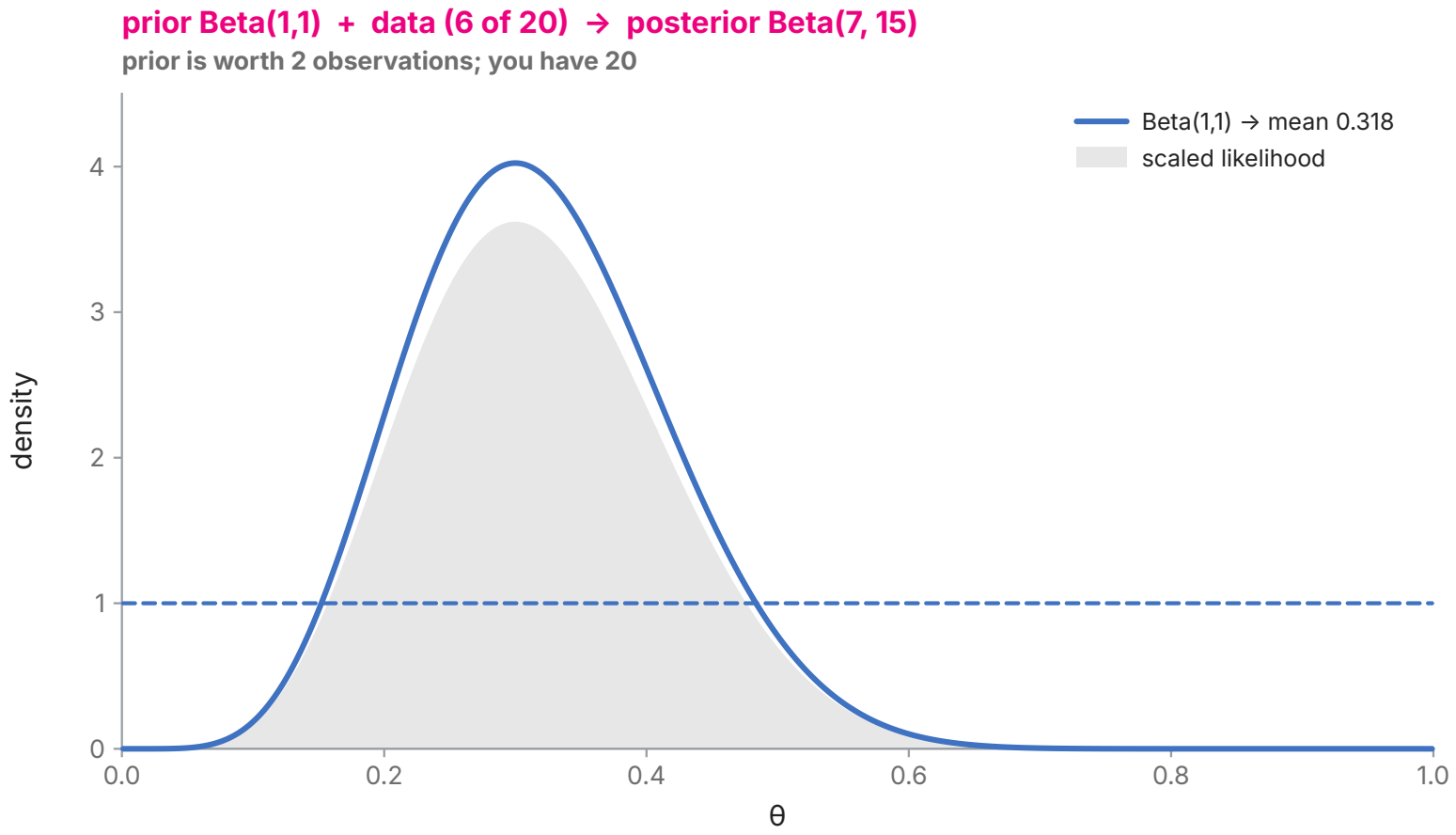
Observed share k/n

0.3

Overlay all four priors

At $n = 0$ the posterior **is** the prior. As n grows the likelihood takes over and the four posteriors converge — the data outvote the prior.

Turn the overlay on and slide n from 10 to 400 to watch it happen.



What the Update Does

- With $n = 0$ the posterior is exactly the prior — Bayes' rule degrades gracefully to "we knew nothing new"
- Each observation shifts α or β by one. The prior is **data you are pretending to have already seen**
- As n grows the four posteriors become indistinguishable: the likelihood dominates any fixed prior

Bernstein–von Mises, informally

Under regularity conditions the posterior concentrates on the true θ and becomes approximately normal with variance $1/(nI(\theta))$ — the same Cramér–Rao quantity deck 04 derived. Asymptotically, Bayesian and frequentist answers agree.

Where it fails is where it matters

Small samples, many parameters, and priors that put **zero** mass on the truth. Convergence is no comfort when n is 40.

Shrinkage

The Normal–Normal Model

Now a mean rather than a proportion: $x_1, \dots, x_n \sim N(\theta, \sigma^2)$ with prior $\theta \sim N(\mu_0, \tau^2)$.

Posterior

$$\theta \mid x \;\sim\; N\!\left(\frac{\tau^{-2}\mu_0 + n\sigma^{-2}\bar{x}}{\tau^{-2} + n\sigma^{-2}},\quad \frac{1}{\tau^{-2} + n\sigma^{-2}}\right)$$

Writing precision as the reciprocal of variance, the posterior mean is a **precision-weighted average** of the prior mean and the sample mean, and the posterior precision is the **sum** of the two precisions.

$$\mathrm{E}[\theta \mid x] \;=\; w\,\mu_0 + (1-w)\,\bar{x}, \qquad w = \frac{\tau^{-2}}{\tau^{-2} + n\sigma^{-2}}$$

One Formula, Four Names

In estimation

- **Ridge regression** — the Gaussian prior behind deck 04's $c\bar{X}$, shrinking coefficients toward zero
- **Empirical Bayes** — shrinking group estimates toward the grand mean

And elsewhere

- **The Kalman filter** — each update is exactly this, with the prior carried forward in time
- **Credibility theory** — how insurers price a small policyholder against the portfolio

Why economists should care

Regional unemployment from 40 respondents, school effects from one cohort, firm productivity from three years — all are noisy estimates of a quantity that resembles its neighbours. Shrinkage is the answer, and it is not a fudge: it **lowers mean squared error**.

Shrinkage, Live

Prior mean μ_0

0

Prior sd τ

0.5

Sample mean $\bar{x}$

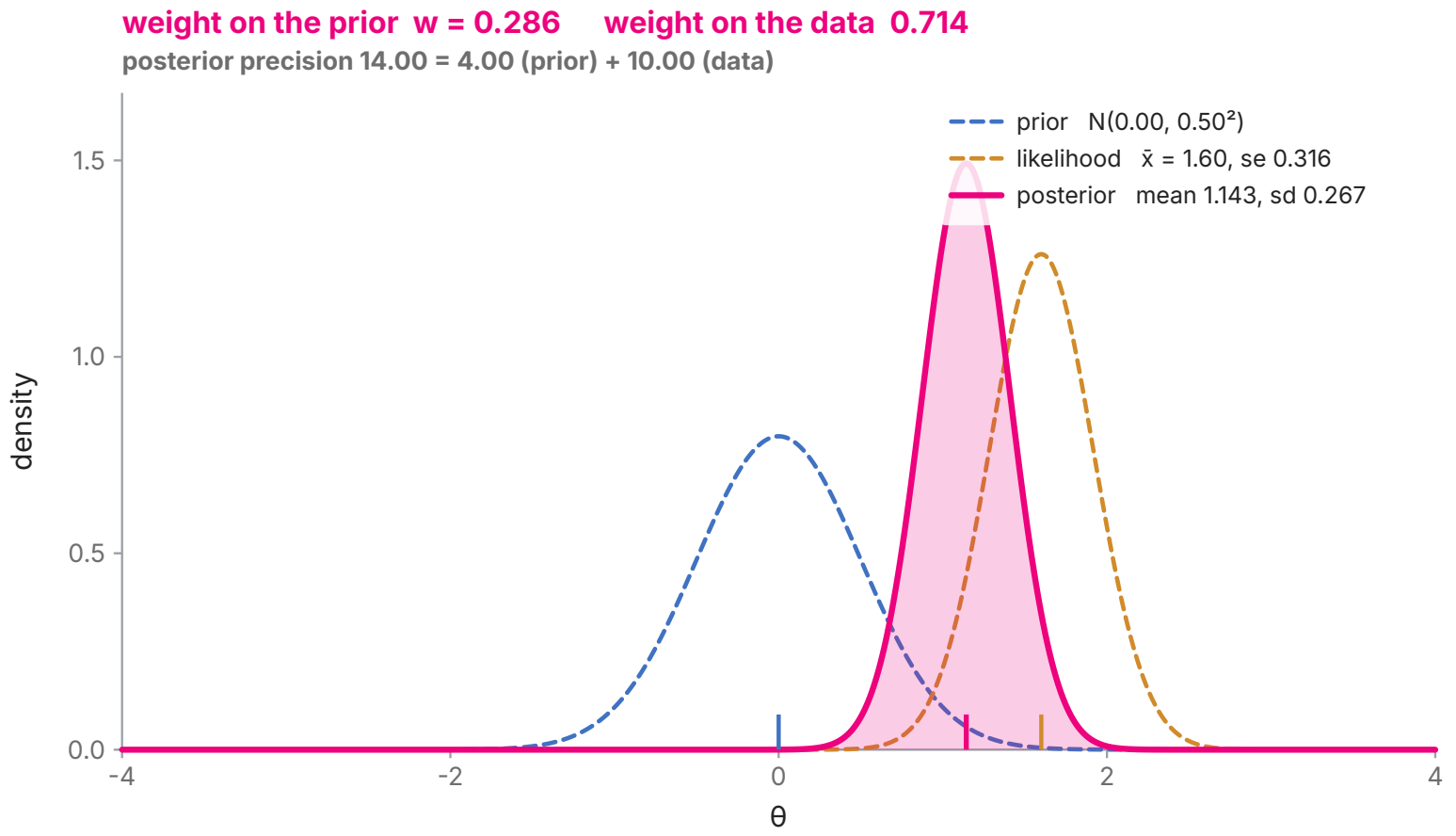
1.6

n

10

The posterior always sits **between** the prior mean and the sample mean, and it is always **narrower than both**.

Shrink τ toward 0.1 for a confident prior; push n up to let the data win.



Two Kinds of Interval

Credible versus Confidence

Credible interval

A set C with

$$P(\theta \in C \mid y) = 1 - \alpha$$

"Given the data and the prior, there is a 95% probability that θ lies in here."

Confidence interval

A rule with

$$P_\theta\big(\theta \in C(Y)\big) = 1 - \alpha \;\; \forall \theta$$

"95% of intervals built this way contain θ."

The credible interval is what people think a confidence interval is

Deck 05 listed $P(H_0 \mid \text{data})$ as something a p-value is **not**, and deck 04 listed the probability statement as something a confidence interval is **not**. The Bayesian posterior is the object that actually licenses it — at the cost of requiring a prior.

Two Intervals, One Dataset

n

20

successes k

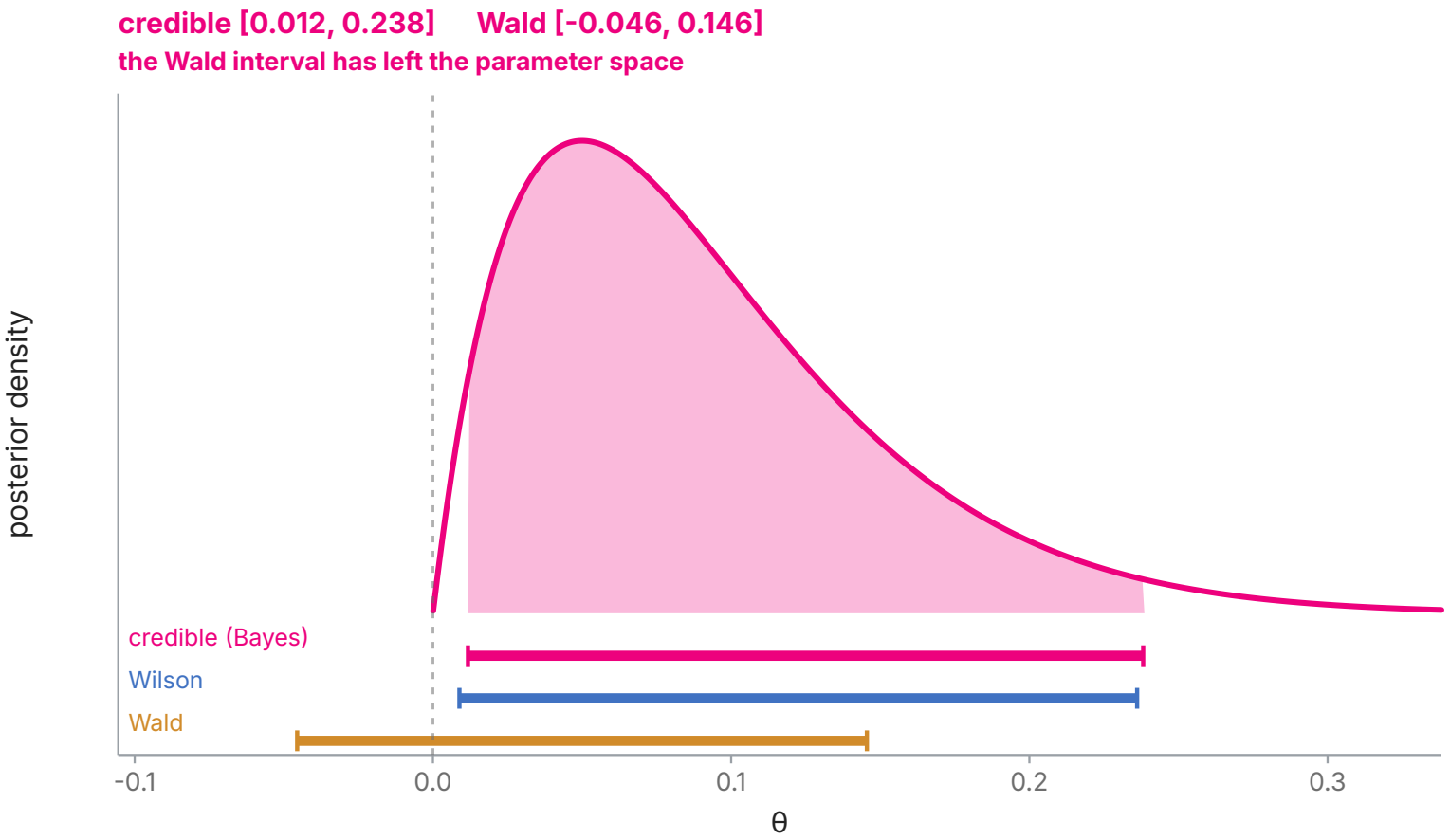
1

Prior

Uniform — Beta(1,1)

With a flat prior and a large n the two intervals nearly coincide — which is why the misinterpretation survives.

Push k toward 0 or n, or shrink n, and they separate. Only one of them stays inside $[0, 1]$ by construction.



The Two Sentences

- With a flat prior and plenty of data the intervals nearly coincide numerically — but they are answers to **different questions**
- A credible interval is a statement about θ given **these** data. A confidence interval is a statement about the **procedure** across hypothetical repetitions
- The posterior respects the parameter space: a probability cannot be negative, and the credible interval never is

The honest trade

The Bayesian statement is the one you wanted. It costs you a prior, and someone can always ask where the prior came from.

Evidence

Bayes Factors

Definition — Bayes Factor

$$\mathrm{BF}_{10} = \frac{p(y \mid H_1)}{p(y \mid H_0)} = \frac{\int p(y \mid \theta)\, p_1(\theta)\, d\theta}{p(y \mid \theta_0)}$$

It converts prior odds into posterior odds:

$$\frac{P(H_1 \mid y)}{P(H_0 \mid y)} = \mathrm{BF}_{10} \times \frac{P(H_1)}{P(H_0)}$$

What it does that a p-value cannot

It can favour H_0. A p-value of 0.6 is not evidence *for* the null; a Bayes factor of $\frac{1}{12}$ is.

What it costs

The integral needs a prior **under** H_1 — and unlike the posterior, the Bayes factor never stops depending on it.

Lindley's Paradox, Live

Hold the p-value fixed at exactly 0.05 and let the sample grow.

$\log_{10}$ n

2

Prior sd under H_1, τ

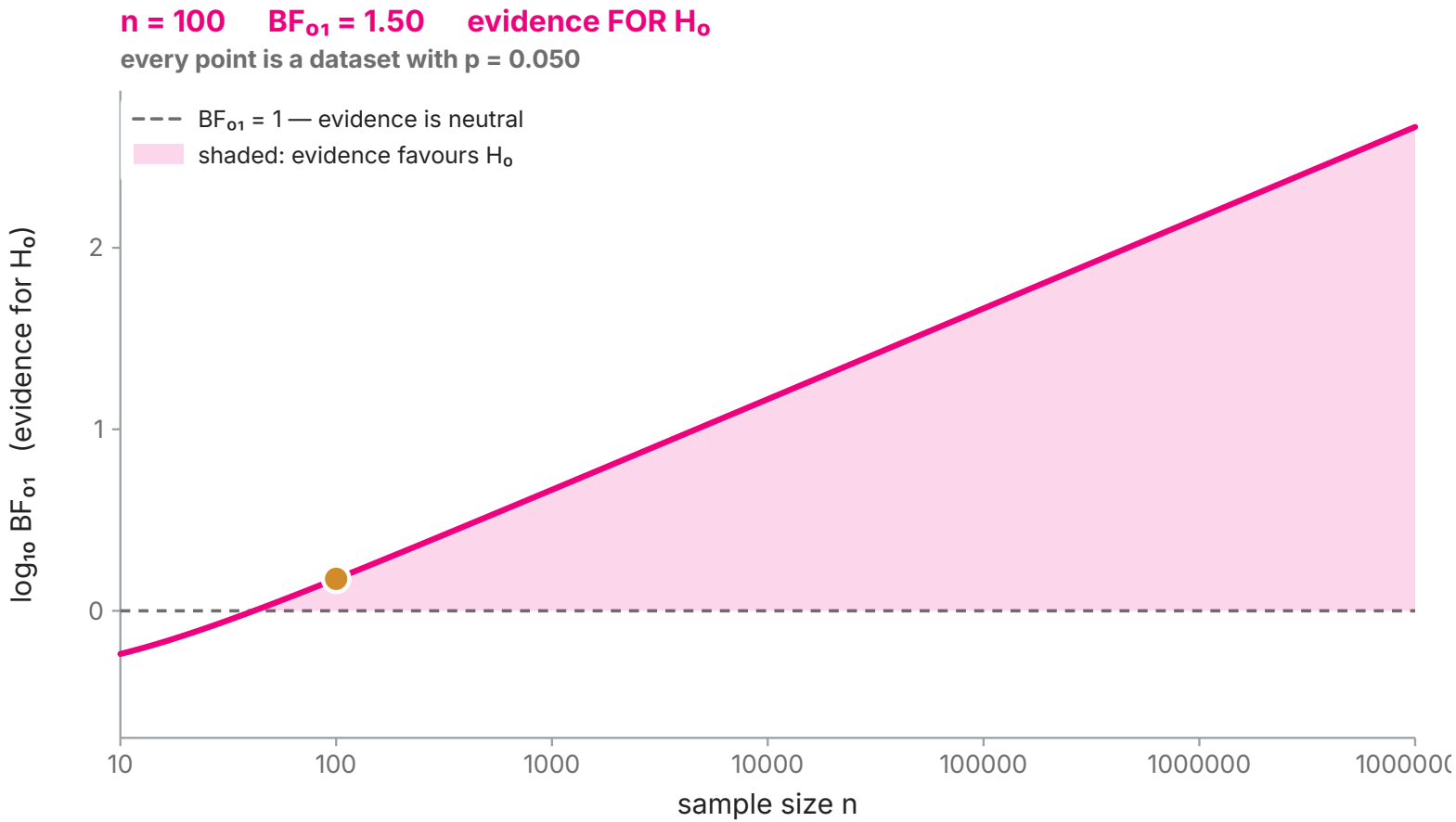
1

Held-fixed p-value

p = 0.05

Every point on this curve is a dataset that a frequentist would reject at the same fixed level.

Above the line, the evidence favours H_0 — the hypothesis just rejected.



What the Paradox Says

- At small n, "$p = 0.05$" is genuine evidence against H_0
- At large n the **same** p-value is evidence *for* H_0, because a fixed significance threshold detects ever smaller and less interesting deviations
- The crossing point depends on τ — the prior under H_1 — and never stops depending on it

Why economists meet this constantly

Administrative datasets have millions of rows. At $n = 10^7$ everything is significant and nothing is large. Deck 05's closing reminder — statistical significance is not practical significance — stops being a caution and becomes the operating condition.

How Often Is a Significant Result True?

If a fraction π of tested hypotheses are true, a test with level α and power $1-\beta$ gives

$$\text{PPV} = P(H_1 \text{ true} \mid \text{reject } H_0) = \frac{\pi(1-\beta)}{\pi(1-\beta) + (1-\pi)\alpha}$$

A well-run field

$\pi = 0.5$, power 0.8, $\alpha = 0.05$

$$\text{PPV} = \frac{0.40}{0.40 + 0.025} = 0.94$$

Exploratory economics

$\pi = 0.10$, power 0.4, $\alpha = 0.05$

$$\text{PPV} = \frac{0.04}{0.04 + 0.045} = 0.47$$

A significant result is then **more likely false than true** — and no individual researcher did anything wrong.

Decisions

Loss, Risk, and What "Best" Means

Deck 04 asked which parameter value fits the data *best*, and never defined "best". A loss function is the definition.

Definition — Bayes Estimator

Given a loss $L(\theta, d)$ for reporting d when the truth is θ, the **Bayes estimator** minimises posterior expected loss:

$$\hat{\theta} \;=\; \arg\min_{d} \; \mathrm{E}\left[L(\theta, d) \mid y\right] \;=\; \arg\min_{d} \int L(\theta, d)\, p(\theta \mid y)\, d\theta$$

Loss function	Bayes estimator
Squared error $(\theta - d)^2$	posterior **mean**
Absolute error $\lvert\theta - d\rvert$	posterior **median**
0–1 loss	posterior **mode** (the MAP)
Asymmetric linear, costs a under / b over	posterior **quantile** at $a/(a+b)$

Asymmetric Loss, Live

Under-forecasting a deficit and over-forecasting it do not cost the same.

Cost of under ÷ over

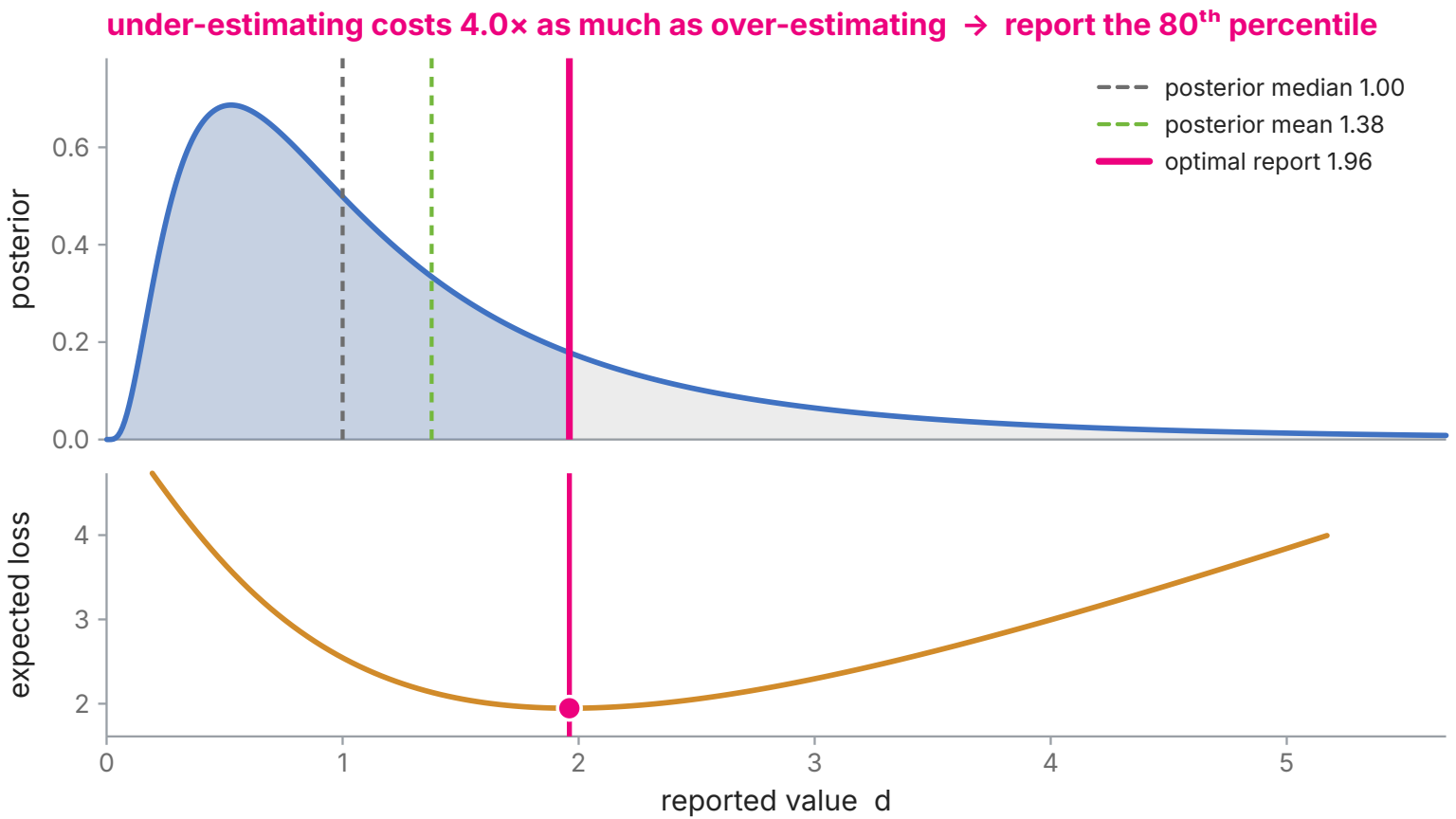
4

Posterior skew

0.8

With symmetric loss the optimum is the median. Make under-estimating four times as costly and the optimal report moves into the upper tail — deliberately, not as a bias.

The mean, median and mode of the posterior have not moved at all.



Why Economists Should Recognise This

- It is **expected-utility maximisation**, with the unknown parameter as the state of the world — the framework is already familiar from micro
- The estimator is not a property of the data alone. It depends on what a mistake costs
- Inventory, capacity, deficit forecasts, capital buffers and reserve prices are all asymmetric-loss problems, and reporting the posterior mean is simply the wrong answer for them

The unifying claim

"Best" is never a purely statistical question. Squared-error loss is a **choice**, and it is the one that happens to be conventional.

Computation and Practice

The Computation Ladder

- **Conjugate** — closed form, as in this deck. Fast, exact, and restricted to a handful of model families
- **Grid approximation** — evaluate the posterior on a mesh and normalise. Any model, up to about three parameters, then the mesh explodes
- **MCMC** — simulate a Markov chain whose stationary distribution is the posterior. Everything else, at the cost of convergence diagnostics

What changed in practice

None of this was usable for real models until cheap computation arrived. That is why the Bayesian–frequentist argument was philosophical for a century and is now largely a question of which tool fits the problem.

Where Economists Meet This

Macro and finance

- **DSGE estimation** — short samples, many parameters, priors doing real work
- **Bayesian VARs** — shrinkage is what makes them forecast
- **Model averaging** — weight models by posterior probability instead of picking one

Micro and policy

- **Hierarchical models** for regional and firm panels
- **Small-area estimation** — official statistics for thin geographies
- **Sequential experiments** — yesterday's posterior is today's prior

The recurring theme

Every one has more parameters than the data can comfortably identify. A prior is not a philosophical stance there — it is what makes the problem tractable.

Summary

Key Takeaways

(1) The posterior is the whole answer; MAP, the posterior mean and the credible interval are all **readings** of it

(2) A conjugate prior is data you are pretending to have seen — and the likelihood outvotes it as n grows

(3) The posterior mean is a **precision-weighted average**, which is ridge, Kalman and empirical Bayes at once

(4) A credible interval licenses the probability statement people wrongly make about confidence intervals

(5) A **loss function** is what turns a posterior into a decision — and defines "best"

Bayesian and frequentist

- **Agree** asymptotically, under regularity conditions
- **Differ** where it matters: small n, many parameters, constrained spaces, and genuine prior information
- **Bayes factors** can support H_0; p-values never can

The cost, stated plainly

Every Bayesian answer is conditional on a prior. Report it, and show what happens when it changes.