

Fundamentals on Bayesian Statistics

From coin tossing to Bayesian regression

Jinkyoo Park · KAIST

Model → update → predict, one example at a time

— 01 • FOUNDATIONS

1 • Introduction and Bayes' rule

Start with the original coin question: what is its probability of heads? Compare statistical viewpoints, then learn the update rule.

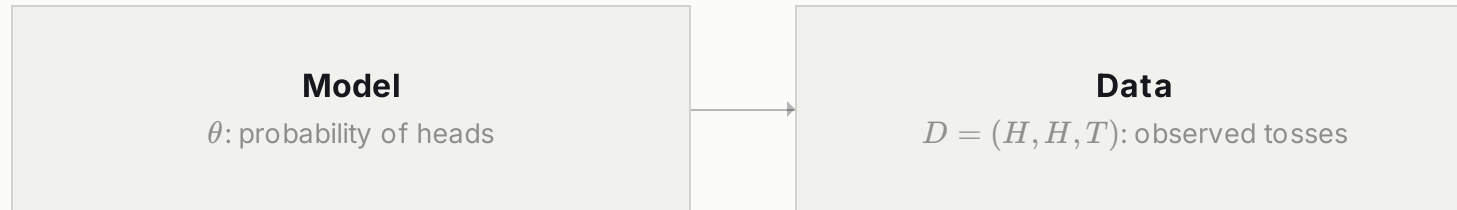
Where we are — the coefficients must be learned

PREVIOUS STEP	THIS CHAPTER'S QUESTION
Lecture 1 optimized a specified model.	Now represent an unknown parameter using a distribution.

Keep the likelihood, posterior and prediction target separate. The temperature example will distinguish an unknown mean from measurement noise.

Update a temperature mean and distinguish parameter uncertainty from a new noisy reading.

Statistics works backwards from data to a model



Given θ , the model tells us which data are plausible. Given data, inference asks which values of θ remain plausible. This alone is not a claim of causal identification.

OBSERVATION UNCERTAINTY

$$Y_i \mid \theta \sim \text{Bernoulli}(\theta).$$

Even if θ were known, the next toss would remain random.

PARAMETER UNCERTAINTY

We do not know θ . A Bayesian distribution over θ represents **our uncertainty about it**; it does not require the physical coin to change on every toss.

Two approaches, both able to quantify uncertainty

FREQUENTIST

θ is **fixed**; the *data* is what is random. So a probability describes the datasets you might have drawn — the long run of repeated, identical experiments.

Estimate by relative frequency, $\hat{\theta} = S/n$. Running the experiment with more and more tosses:

$\frac{1}{3}$	$\frac{5}{6}$	$\frac{5}{13}$	$\frac{129}{313}$	$\frac{61423}{123400}$
0.33	0.83	0.38	0.41	0.50

The difficulty. Small n swings wildly — that is high **variance**, not bias; S/n is centred on θ at every n . And "identical flip" is an idealisation no real experiment quite meets.

BAYESIAN

The **data** is fixed — you observed it. θ is what is uncertain, so a probability describes θ itself: a degree of belief.

Specify a prior $p(\theta)$, condition on the observed D , and read off the posterior:

$$p(\theta) \longrightarrow p(\theta \mid D)$$

The difficulty. The prior is a modelling choice, and choosing it honestly is real work. Check what it implies and compare reasonable alternatives.

The distinction is **what the probability statement describes**, not whether uncertainty matters.

The lecture route — keep each example together

Bring: Conditional probability, sums and a Gaussian distribution.

FIRST PASS	WHAT TO DO
Follow the idea	Specify → update → predict → check; repeat for coins, counts and regression
Work without the solution	Update a temperature mean and distinguish parameter uncertainty from a new noisy reading.
Return later	Complete-square derivations and additional conjugate calculations are in the appendix.

For the temperature thread: **predict** → **calculate** → **reveal and check** → **change one condition**. Complete the core calculation before reading the research extensions.

Bayes' rule — name each part before using it

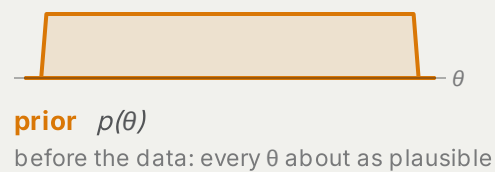
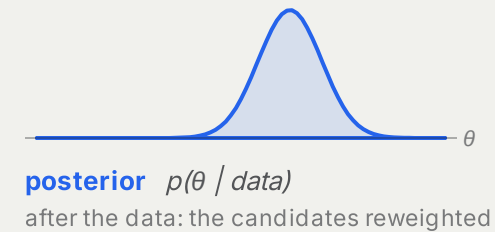
Q1 Why a distribution?

Q2 What does the coin teach us?

Q3 How does the model change?

Q4 How do we learn a regression?

$$\underbrace{p(\theta \mid \text{data})}_{\text{posterior}} = \frac{\overbrace{p(\text{data} \mid \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\text{data})}_{\text{evidence}}}$$

× likelihood $p(\text{data} \mid \theta)$
 ÷ evidence $p(\text{data})$
the data enter only here


The **evidence** $p(\text{data}) = \int p(\text{data} \mid \theta) p(\theta) d\theta$ is the probability of the data over every candidate at once. It does not depend on θ , so it only rescales the curve — which is why $\text{posterior} \propto \text{likelihood} \times \text{prior}$ is usually all you write.

Prior in, posterior out. **The likelihood tells us how to reweight the candidates.**

Why Bayes' rule works

The same joint probability can be factored in two orders:

$$p(\theta, D) = p(D \mid \theta)p(\theta) = p(\theta \mid D)p(D).$$

Divide by the evidence $p(D) > 0$:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}.$$

Multiply prior by likelihood, then **normalise** so the posterior sums or integrates to 1.

Two candidate coins — Bayes' rule with numbers

an illustrative calculation before the continuous case

A coin is either fair ($\theta = 0.5$) or head-biased ($\theta = 0.8$), with equal prior probabilities. We observe **two heads**, conditionally independently.

CANDIDATE	PRIOR	LIKELIHOOD OF HH	PRODUCT	POSTERIOR
$\theta = 0.5$	0.5	$0.5^2 = 0.25$	0.125	$0.125/0.445 = 0.281$
$\theta = 0.8$	0.5	$0.8^2 = 0.64$	0.320	$0.320/0.445 = 0.719$

The evidence is $p(HH) = 0.125 + 0.320 = 0.445$. Dividing by it makes the posterior probabilities add to 1.

Two heads favour the biased coin, but **do not prove which coin we have**.

One formula, two readings — probability and likelihood

the same expression, according to which argument you hold fixed

For the observed sequence $D = (H, H, T)$, conditional independence gives

$$p(D \mid \theta) = \theta \cdot \theta \cdot (1 - \theta) = \theta^2(1 - \theta).$$

That one expression has **two free arguments**, and which of them you hold fixed decides what kind of object you are looking at. Both readings below are of the formula above — they are not two different formulas.

FIX THE PARAMETER, VARY THE DATA — A PROBABILITY

Then it is $p(D \mid \theta)$, an ordinary probability over the $2^3 = 8$ sequences three tosses could produce. Those eight numbers **sum to 1**, for every θ .

FIX THE DATA, VARY THE PARAMETER — A LIKELIHOOD

Then it is $L(\theta; D)$, a *score* on each candidate θ . Nothing normalises it: $\int_0^1 \theta^2(1 - \theta) d\theta = \frac{1}{12}$, **not 1**.

Summing to 1 along one axis says nothing about the other. That **missing normalisation is exactly what $p(D)$ supplies** in Bayes' rule — which is why a likelihood is not yet a posterior.

For continuous θ , $p(\theta)$ is a **density**: interval areas are probabilities and $P(\theta = \theta_0) = 0$. A density may exceed 1. This differs from the two discrete candidates on the previous slide.

One toss, a sequence, and a count

Q1 Why a distribution?

Q2 What does the coin teach us?

Q3 How does the model change?

Q4 How do we learn a regression?

Assume $Y_i \mid \theta$ are independent Bernoulli variables. Let $S = \sum_{i=1}^n Y_i$.

WHAT IS OBSERVED?	PROBABILITY GIVEN θ	FOR TWO HEADS IN THREE TOSSES
One toss Y_i	$\theta^{Y_i}(1 - \theta)^{1-Y_i}$	Head: θ
One particular sequence	$\theta^S(1 - \theta)^{n-S}$	HHT: $\theta^2(1 - \theta)$
Only the total S	$\binom{n}{S}\theta^S(1 - \theta)^{n-S}$	HHT, HTH, THH: $3\theta^2(1 - \theta)$

The factor $\binom{n}{S}$ counts the sequences. It is constant in θ , so it changes the data probability but **not the MLE or posterior shape**.

The coin MLE — maximise the log likelihood

For $0 < S < n$, taking logs turns the product into a sum:

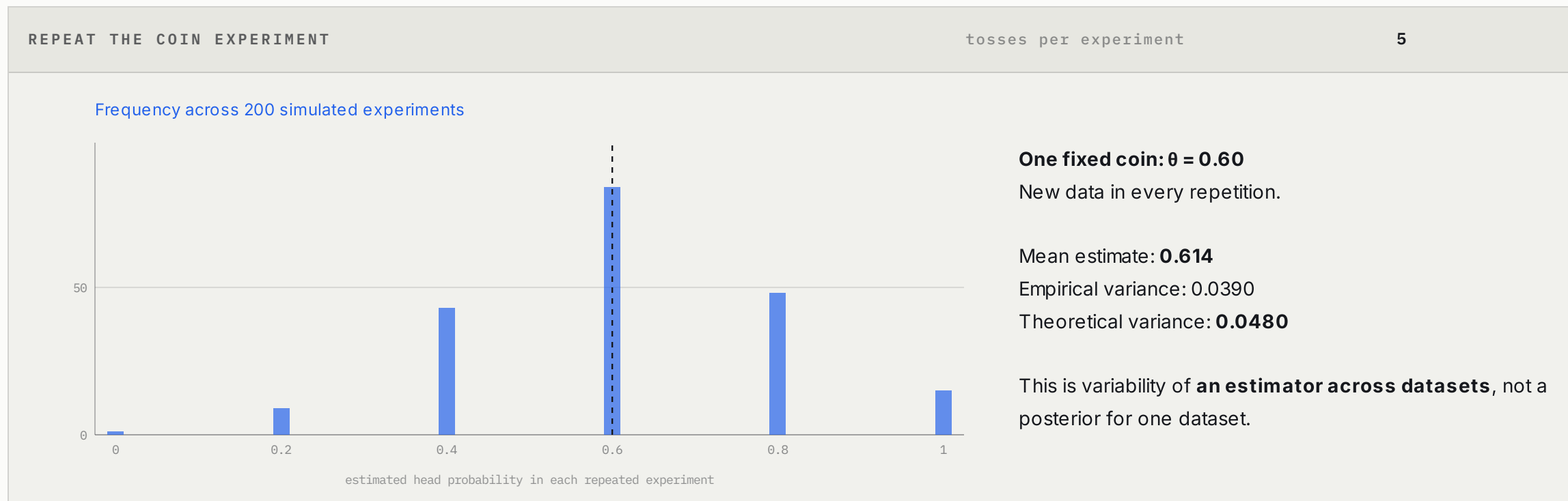
$$\ell(\theta) = S \log \theta + (n - S) \log(1 - \theta) + \text{constant}.$$

$$\ell'(\theta) = \frac{S}{\theta} - \frac{n - S}{1 - \theta} = 0 \quad \implies \quad S(1 - \theta) = (n - S)\theta \quad \implies \quad \hat{\theta}_{\text{ML}} = \frac{S}{n}.$$

The second derivative is negative on $(0, 1)$, so this is the maximum. For HHT, $\hat{\theta}_{\text{ML}} = 2/3$.

If $S = 0$ or $S = n$, the maximum is at the boundary, 0 or 1; there is no interior derivative-zero solution. Here S is sufficient for θ : the order adds no information under this model.

Experiment 1 — repeat the coin experiment



Predict: will 100 tosses vary less than 5? **Try:** change tosses per experiment, then repeat the 200 experiments. **Explain:** the coin stays fixed; the datasets and their estimates change.

The Bayesian move — carry a distribution

OBJECT	QUESTION IT ANSWERS	COIN EXAMPLE
Prior $p(\theta)$	What is plausible before these data?	How biased might the coin be?
Likelihood $p(D \mid \theta)$	How compatible is each candidate with the data?	How likely is HHT for this bias?
Posterior $p(\theta \mid D)$	What is plausible after these data?	Which biases still have support?
Predictive $p(\tilde{Y} \mid D)$	What might happen next?	How many heads in future tosses?

The parameter and the next observation are different unknowns. **A distribution over one is not a distribution over the other.**

Prior, posterior, predictive — keep the target clear

Q1 Why a distribution?

Q2 What does the coin teach us?

Q3 How does the model change?

Q4 How do we learn a regression?

DISTRIBUTION	RANDOM QUANTITY	INFORMATION USED
Prior $p(\theta)$	parameter	before observing D
Prior predictive $p(\tilde{y})$	possible data	prior + sampling model
Posterior $p(\theta \mid D)$	parameter	prior + observed data
Posterior predictive $p(\tilde{y} \mid D)$	future observation	posterior + sampling model

Posterior: what might the parameter be? Predictive: what might we observe?

Check — what the distribution is over

A Bayesian writes $p(\theta)$ for a physical constant with one true, fixed value. What does the distribution describe?

- ☒ A Our uncertainty about the fixed unknown
- ☐ B The constant changing on every measurement
- ☐ C The histogram of future observations
- ☐ D Nothing: a fixed quantity cannot be described probabilistically

The distribution represents **epistemic** uncertainty. Data can reduce it when informative. **Aleatoric** uncertainty is variation in future observations conditional on the model's parameters; learning the parameter alone does not remove that variation.

— 02 · BAYESIAN COIN

2 · Bayesian estimation for coin flipping

Keep the same coin: choose a Beta prior, observe heads and tails, quantify uncertainty, and predict the next tosses.

A Beta prior — location and strength

Use a density supported on the possible coin biases, $0 < \theta < 1$:

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad \alpha, \beta > 0.$$

WHERE IS THE BELIEF CENTRED?

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha + \beta}.$$

Beta(2, 2) centres at 1/2; Beta(2, 8) centres at 1/5.

HOW STRONG IS THAT BELIEF?

$$\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Write $\mu = \mathbb{E}[\theta]$ and $\nu = \alpha + \beta$: this is $\mu(1 - \mu)/(\nu + 1)$. At a fixed centre the numerator is fixed, so **spread falls like $1/\nu$** .

Both priors below centre at 1/2, so only ν separates them: Beta(2, 2) has variance 1/20, Beta(20, 20) has 1/164 — **8.2 times smaller**, and the standard deviation falls from 0.22 to 0.08. That is what “ $\alpha + \beta$ is the strength of the prior” means, and in the next slide it is why they cost ν pseudo-counts to overturn.

$B(\alpha, \beta)$ makes the density integrate to 1. The parameters act like pseudo-counts in the update; they need not represent actual past flips. Uniform Beta(1, 1) is still a choice of prior.

Derive the posterior by collecting powers

Start with a Beta prior and observe S heads in n tosses:

$$\begin{aligned} p(\theta \mid D) &\propto \underbrace{\theta^S (1 - \theta)^{n-S}}_{\text{likelihood}} \underbrace{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}_{\text{prior}} \\ &= \theta^{\alpha+S-1} (1 - \theta)^{\beta+n-S-1}. \end{aligned}$$

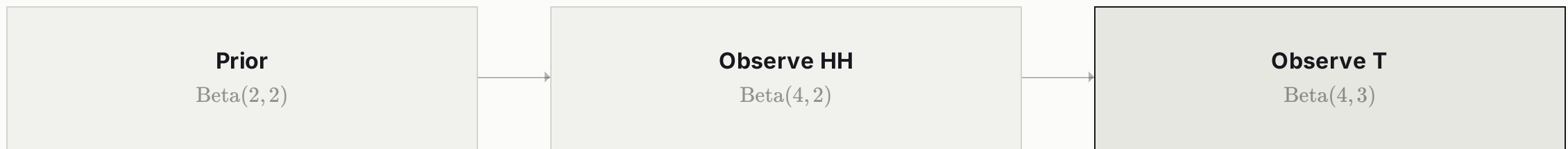
Recognise a Beta density and restore its normalising constant:

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

Add heads to the first parameter; add tails to the second. For a Beta(2, 2) prior and HHT, the posterior is Beta(4, 3).

Sequential updating — do not count old data twice

The posterior after one batch becomes the prior before the next batch:



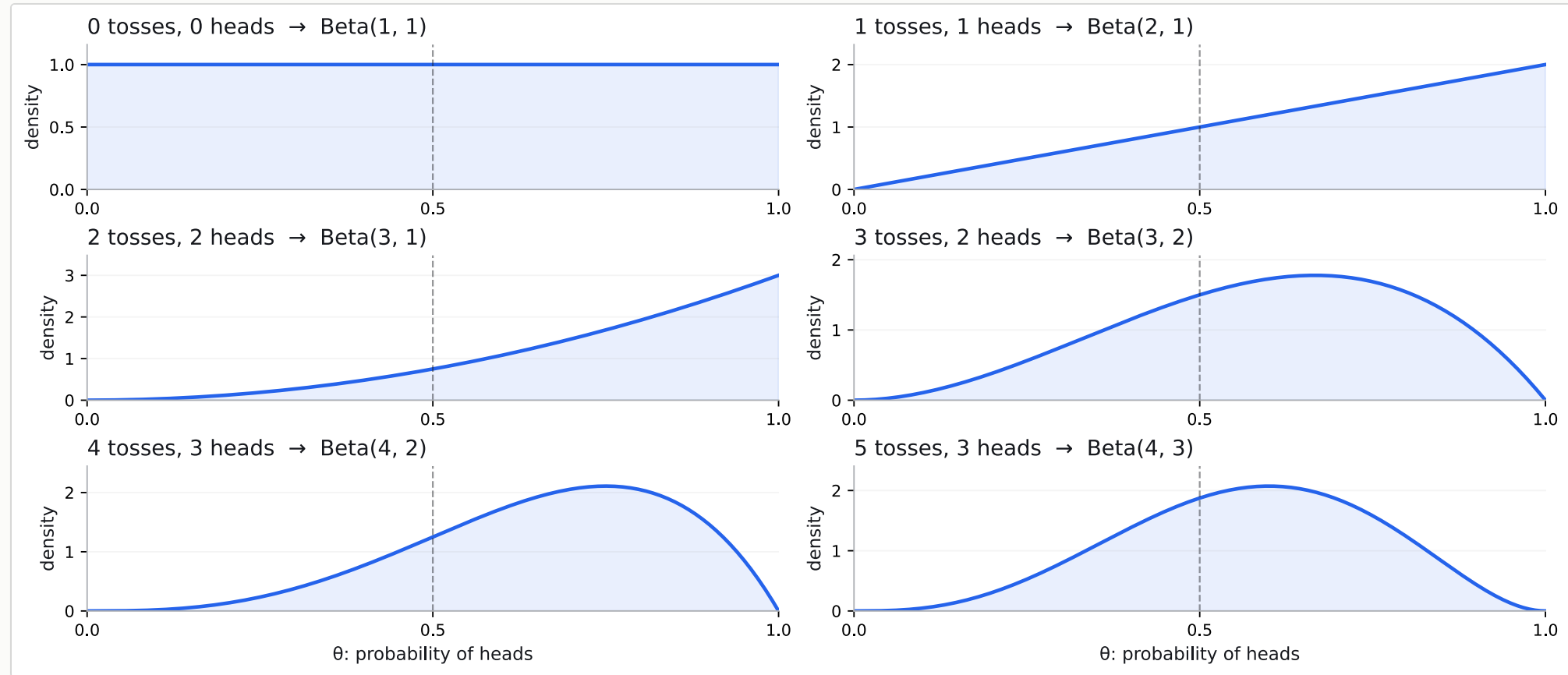
Updating once with HHT gives exactly the same answer. More generally,

$$p(\theta \mid D_1, D_2) \propto p(D_2 \mid \theta) p(\theta \mid D_1),$$

when the batches are conditionally independent given θ .

The old data is already in $p(\theta \mid D_1)$. **Multiply by the new likelihood once.**

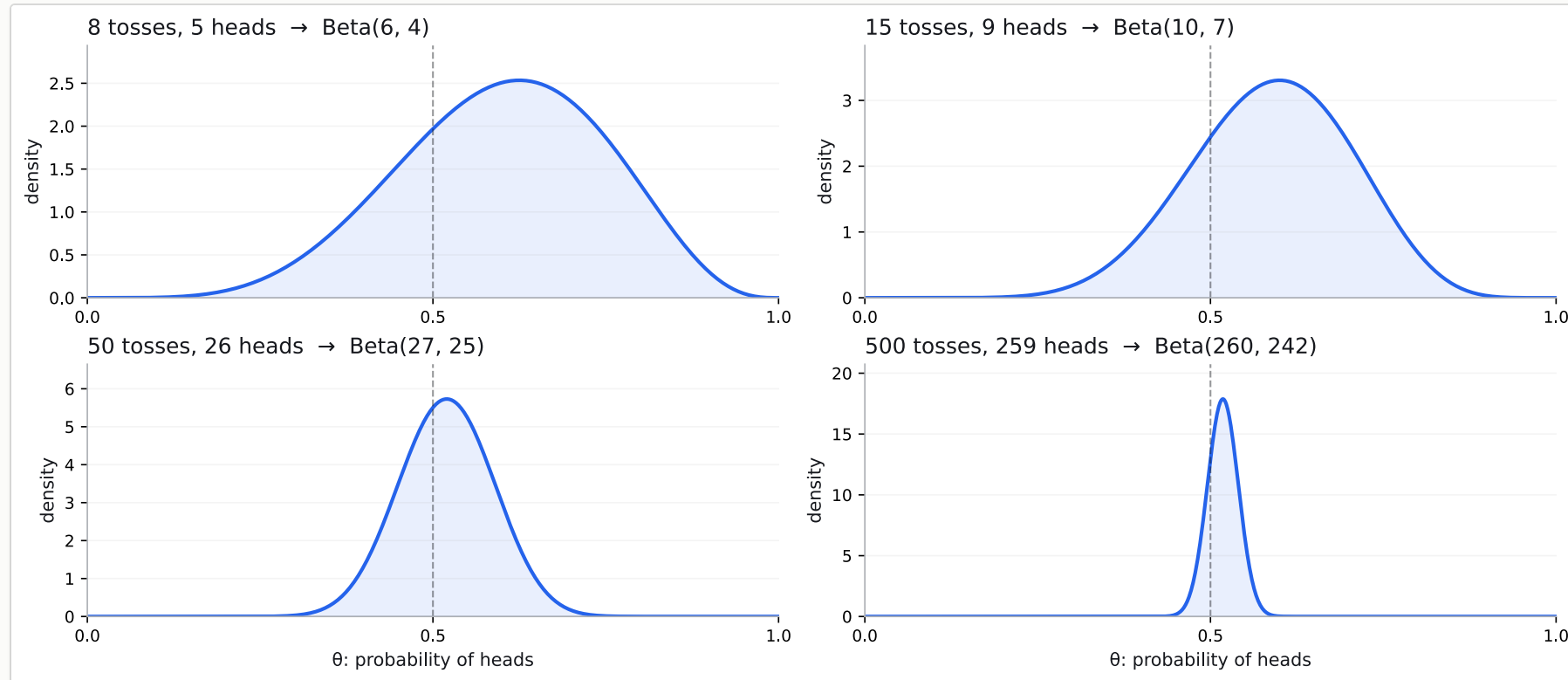
Bayesian coin updating — the first five tosses



All six early states. Start from Beta(1,1): each head adds 1 to α and each tail adds 1 to β .
The dashed line marks $\theta = 0.5$.

Read left to right, then down. **The first two heads pull the curve right; the next tail pulls it back.**

Bayesian coin updating — from 8 to 500 tosses



Heads out of tosses: 5/8, 9/15, 26/50 and 259/500. Every curve has area 1; the vertical density scales differ.

The final two observed proportions are similar. **The larger dataset supports a much more precise estimate.**

Beta(4,3) — the mean and the mode, by hand

HHT and a Beta(2,2) prior gave Beta(4,3); now read the numbers out of it

With $\alpha = 4$, $\beta = 3$ the normaliser is $1/B(4,3) = 60$, so the posterior is the polynomial $p(\theta \mid D) = 60 \theta^3 (1 - \theta)^2$.

THE MEAN — INTEGRATE θ AGAINST IT

$$\mathbb{E}[\theta \mid D] = \int_0^1 \theta \cdot 60 \theta^3 (1 - \theta)^2 d\theta = \frac{\alpha}{\alpha + \beta} = \frac{4}{7} = 0.571$$

The general Beta mean from earlier, with the numbers put in.

THE MODE — SET THE LOG-DERIVATIVE TO ZERO

$$\frac{d}{d\theta} \log p = \frac{3}{\theta} - \frac{2}{1 - \theta} = 0 \implies \theta^* = \frac{3}{5} = 0.600$$

In general $\frac{\alpha-1}{\alpha+\beta-2}$. Under a flat prior it would be the MLE $\frac{2}{3}$; the Beta(2, 2) prior pulled it toward $\frac{1}{2}$.

Mean 0.571 and mode 0.600 differ because the posterior is **skewed** — three tosses cannot make it symmetric.

Beta(4,3) — the interval and the next toss

A 95% CREDIBLE INTERVAL — INVERT THE CDF

Integrating the polynomial gives the CDF in closed form,

$$F(\theta) = \int_0^\theta 60 t^3 (1-t)^2 dt = 15\theta^4 - 24\theta^5 + 10\theta^6,$$

and the interval is the pair of roots of $F(a) = 0.025$ and $F(b) = 0.975$, solved numerically: [\[0.223, 0.882\]](#).

THE NEXT TOSS — AVERAGE θ OVER THE POSTERIOR

$$P(H \mid D) = \int_0^1 \underbrace{\theta}_{P(H|\theta)} p(\theta \mid D) d\theta = \mathbb{E}[\theta \mid D] = \frac{4}{7}$$

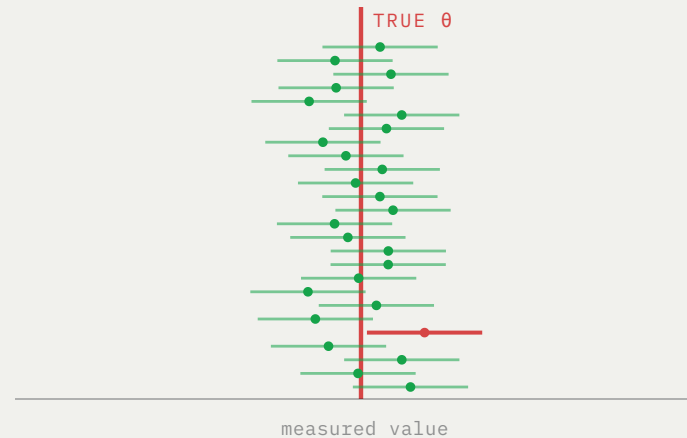
For a single toss the predictive integral collapses to the posterior mean. It is the $m = 1$ case of the Beta-Binomial to come.

Three tosses leave substantial uncertainty — the interval spans two thirds of $[0, 1]$. [The mean, mode and interval answer different questions](#), and none of them is the whole posterior.

Experiment 2 — repeat an interval procedure

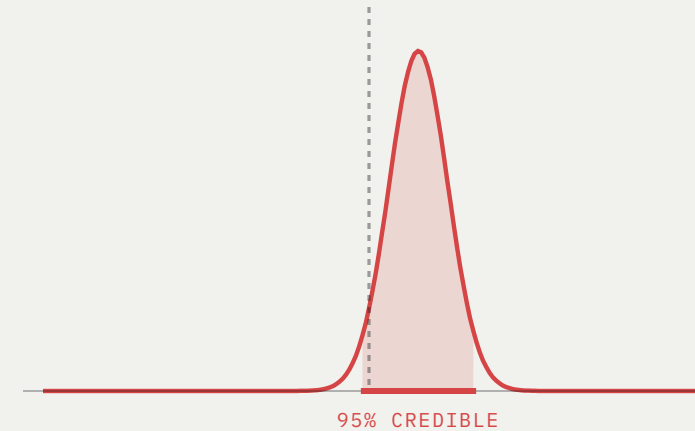
BOTH SAY 95% — AND THEY ARE DIFFERENT CLAIMS

FREQUENTIST — VARIABILITY IN THE DATA



25 of 26 intervals cover the true θ .
The 95% is a property of **the procedure**, not of any one interval.

BAYESIAN — VARIABILITY IN θ



From **this dataset**: $\Pr(\theta \in \text{CR} \mid \text{data}) \approx 0.95$.
Known Gaussian noise; flat prior on the mean.

Predict: must all 26 intervals cover the true value? **Try:** run the repeated experiments. **Explain:** compare the long-run coverage statement on the left with posterior probability for the observed dataset on the right.

Read the two 95% statements carefully

CONFIDENCE INTERVAL

Before drawing data, the random interval $C(D)$ covers a fixed parameter in 95% of repeated experiments:

$$P_{\theta}\{\theta \in C(D)\} = 0.95.$$

Once D is observed, the interval is fixed. Coverage alone does not assign 95% probability to that particular interval containing θ .

CREDIBLE INTERVAL

After observing D , an interval C contains 95% of the **posterior probability**:

$$P(\theta \in C \mid D) = 0.95.$$

This is conditional on the prior, likelihood and data. Its repeated-sampling coverage need not be 95%.

The previous picture uses Normal observations with known noise and a flat prior on the mean. The endpoints coincide in that special case; the interpretations still differ.

The posterior mean is a balance — in numbers

For the coin model,

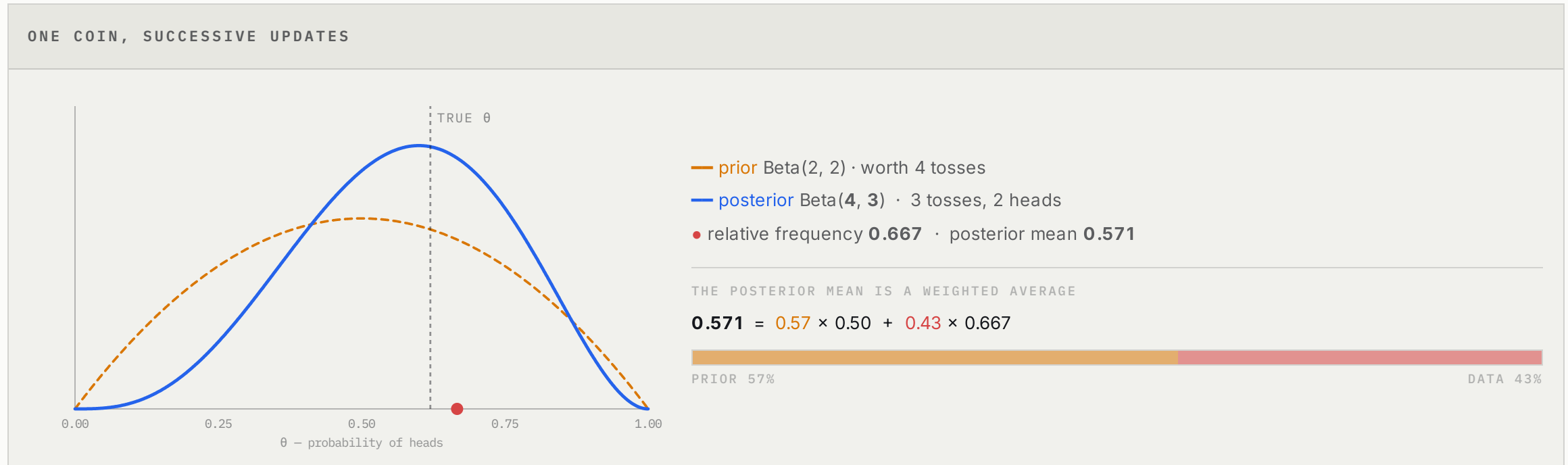
$$\mathbb{E}[\theta \mid D] = \frac{\alpha + S}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{S}{n}.$$

For Beta(2, 2) and HHT, this becomes

$$\underbrace{\frac{4}{7}}_{\text{posterior mean}} = \underbrace{\frac{4}{7}}_{\text{prior weight}} \underbrace{\frac{1}{2}}_{\text{prior mean}} + \underbrace{\frac{3}{7}}_{\text{data weight}} \underbrace{\frac{2}{3}}_{\text{sample proportion}}.$$

Compare **sample size** n with **prior strength** $\alpha + \beta$. Small n alone does not tell us which dominates.

Experiment 3 — change the prior, keep the data



Predict: which prior resists HHT most? **Try:** load HHT, switch the prior, then simulate more tosses. **Explain:** compare the displayed prior/data weights. H and T record hypothetical outcomes; simulated tosses use a fixed bias of 0.62.

Does more data always make a posterior narrower?

On average, yes, in the following precise sense. The average is over possible datasets under the joint model:

$$\text{Var}(\theta) = \mathbb{E}_D[\text{Var}(\theta \mid D)] + \text{Var}_D(\mathbb{E}[\theta \mid D]).$$

The last term is non-negative, so the expected posterior variance cannot exceed the prior variance. Also $\mathbb{E}_D[\mathbb{E}(\theta \mid D)] = \mathbb{E}(\theta)$.

ONE SURPRISING OBSERVATION CAN WIDEN IT

A $\text{Beta}(1, 20)$ prior has variance 0.00206. After one head, $\text{Beta}(2, 20)$ has variance 0.00359. The surprising head challenges a strong belief in tails.

From estimating the coin to predicting its next toss

The posterior answers **which coin biases are plausible**. A prediction asks **what might happen next**.

$$p(\tilde{y} \mid D) = \int p(\tilde{y} \mid \theta) p(\theta \mid D) d\theta.$$

For the Beta(4,3) posterior after HHT,

$$P(\tilde{Y} = 1 \mid D) = \int_0^1 \theta p(\theta \mid D) d\theta = \frac{4}{7}.$$

Average the next-toss prediction over the plausible biases. For one Bernoulli outcome, plugging in the posterior mean gives exactly the same distribution.

Predicting a batch — one uncertain coin is shared

Let K count heads in m **future tosses**. All of them are driven by the same unknown θ , so the honest prediction averages over everything θ might be:

$$p(K \mid D) = \int_0^1 \underbrace{p(K \mid \theta)}_{\text{Binomial}(m, \theta)} \underbrace{p(\theta \mid D)}_{\text{Beta}(\alpha', \beta')} d\theta.$$

This is Act 4's integral with a batch in place of a single y^* . Two things make it worth doing rather than plugging in $\hat{\theta}$: the answer is available in **closed form**, and it is a *different distribution*, not merely a wider one.

The next slide evaluates that integral. The answer has a name — **the Beta-Binomial**.

Where the Beta-Binomial comes from

the predictive integral, evaluated by recognising a Beta kernel

Write the binomial coefficient with gammas, $\binom{m}{y} = \frac{\Gamma(m+1)}{\Gamma(y+1)\Gamma(m-y+1)}$, and collect the powers of θ :

$$\begin{aligned} p(y) &= \int_0^1 \binom{m}{y} \theta^y (1-\theta)^{m-y} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= \frac{\Gamma(m+1)}{\Gamma(y+1)\Gamma(m-y+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \underbrace{\theta^{y+\alpha-1} (1-\theta)^{m-y+\beta-1}}_{\text{an unnormalised Beta}(y+\alpha, m-y+\beta)} d\theta. \end{aligned}$$

The integral is not a new problem — it is a Beta density missing its constant. Supply the constant and it integrates to **1**, leaving only gammas:

$$p(y) = \frac{\Gamma(m+1) \Gamma(\alpha+\beta) \Gamma(y+\alpha) \Gamma(m-y+\beta)}{\Gamma(y+1) \Gamma(m-y+1) \Gamma(\alpha) \Gamma(\beta) \Gamma(m+\alpha+\beta)} = \text{BetaBinomial}(y \mid m, \alpha, \beta).$$

Nothing here used which Beta it was. Feed it the prior and you get the *prior* predictive; feed it the posterior $\text{Beta}(\alpha + S, \beta + n - S)$ and the same formula returns the *posterior* predictive.

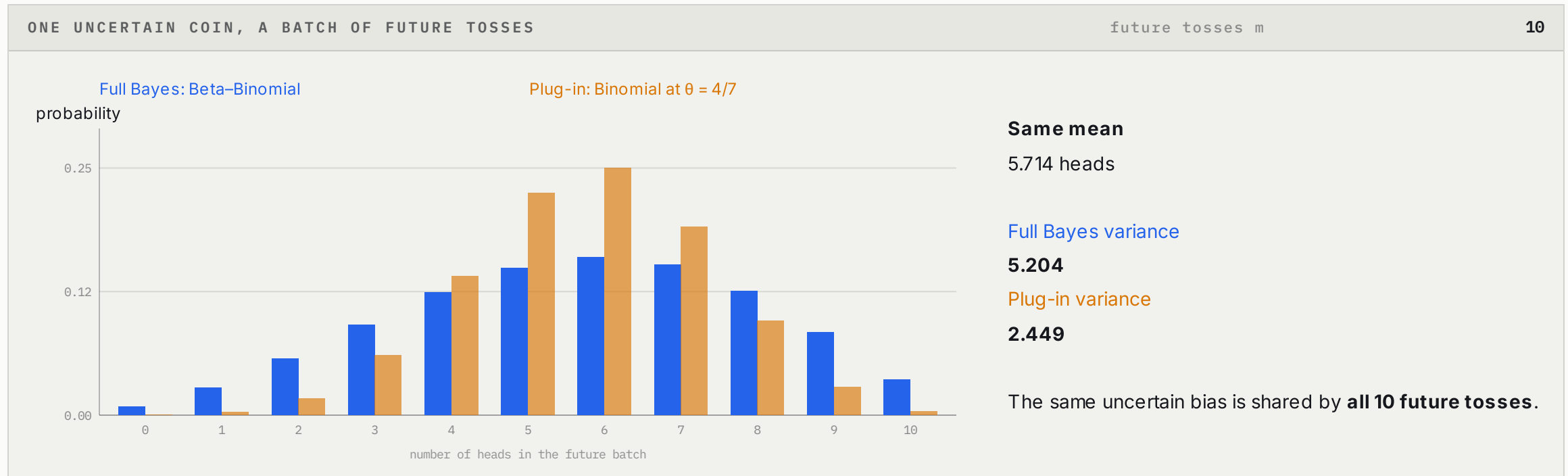
Which Beta you feed it is the whole difference

For our coin the posterior was $\text{Beta}(4, 3)$, so the batch of m future tosses is $\text{BetaBinomial}(m, 4, 3)$ — while plugging in the posterior mean $\hat{\theta} = 4/7$ would give $\text{Binomial}(m, 4/7)$.

TWO FUTURE TOSSES	$P(K = 0)$	$P(K = 1)$	$P(K = 2)$	MEAN	VARIANCE
Plug in $\theta = 4/7$	$9/49 = 0.184$	$24/49 = 0.490$	$16/49 = 0.327$	$8/7$	$24/49$
Average over θ	$3/14 = 0.214$	$6/14 = 0.429$	$5/14 = 0.357$	$8/7$	$27/49$

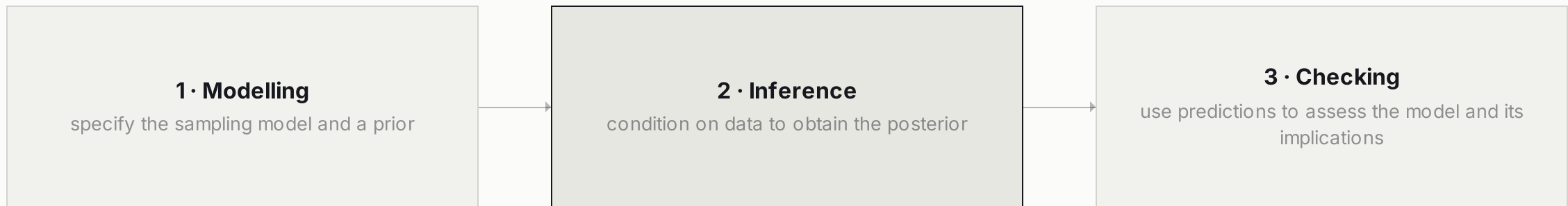
Identical means, and mass moved out of the middle onto both ends. Averaging did not shift the prediction — it widened it, by exactly the $3/49$ the plug-in threw away when it pretended to know θ .

Experiment 4 — predict one toss, then a batch



Predict: do the two predictions agree for one toss? **Try:** set future tosses to 1, then 10, then 30. **Explain:** the posterior stays Beta(4,3); more future tosses share the same uncertain bias.

Three steps in Bayesian approaches



For the coin: Bernoulli/Binomial + Beta $\rightarrow$ updated Beta $\rightarrow$ predictions of future heads.

We have completed one full example. [Keep these steps; change the data model next.](#)

Check — who wins as the data piles up?

With a fixed $\text{Beta}(\alpha, \beta)$ prior and independent Bernoulli observations, what happens to the prior weight in the posterior mean as n grows?

- ☐ A It stays fixed because the prior was chosen first
- ☒ B It tends to zero: $(\alpha + \beta) / (\alpha + \beta + n) \rightarrow 0$
- ☐ C It grows because the posterior becomes more concentrated
- ☐ D It becomes exactly zero after $n = \alpha + \beta$

This is exact for the specified Beta–Binomial model. At $n = \alpha + \beta$, prior and data weights are equal. At finite n the prior still has positive weight. General claims about posterior concentration require additional assumptions.

3 · Conjugate pairs

Repeat the coin workflow in the original order: counts with Poisson–Gamma, measurements with Normal–Normal, and categories with Multinomial–Dirichlet.

Conjugacy — recognise the same family after updating

A prior is **conjugate** to a likelihood when the posterior stays in the prior's family. We can recognise and normalise the result analytically.

DATA MODEL AND UNKNOWN QUANTITY	PRIOR	POSTERIOR
Binomial success probability	Beta	Beta
Poisson rate	Gamma (shape, rate)	Gamma
Normal mean, variance known	Normal	Normal
Normal variance, mean known	Inverse Gamma	Inverse Gamma
Normal mean and precision both unknown	Normal–Gamma	Normal–Gamma
Multinomial category probabilities	Dirichlet	Dirichlet

Conjugacy is algebraic convenience. It does not establish model fit or make every posterior mean a simple average of prior and sample means. Unknown Normal mean and variance require an appropriate joint prior.

Binomial–Beta recap — our template for the next models

STEP	COIN CALCULATION
Sampling model	$S \mid \theta \sim \text{Binomial}(n, \theta)$
Prior	$\theta \sim \text{Beta}(\alpha, \beta)$
Posterior	$\theta \mid D \sim \text{Beta}(\alpha + S, \beta + n - S)$
Prediction	One toss: posterior mean; a batch: Beta–Binomial

Choose a model for the observations, update its parameters, then predict observations again. The next model counts events rather than successes out of a fixed number of trials.

Counting events — why Poisson and Gamma?

Let Y_i count events in equal observation windows, with a common unknown rate λ :

POISSON LIKELIHOOD

$$p(Y_i = y_i \mid \lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}.$$

Y_i is a non-negative integer. Its conditional mean and variance both equal λ .

GAMMA PRIOR

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}.$$

$\lambda > 0$. Here a is **shape** and b is **rate**, with mean a/b and variance a/b^2 .

We use a, b for Gamma to distinguish them from the coin's Beta parameters. A Gamma **scale** would be $1/b$, not b .

Poisson–Gamma updating — counts and exposure add

For conditionally independent, equal-exposure counts,

$$\begin{aligned} p(\lambda \mid D) &\propto \lambda^{\sum_i y_i} e^{-n\lambda} \cdot \lambda^{a-1} e^{-b\lambda} \\ &= \lambda^{a+\sum_i y_i-1} e^{-(b+n)\lambda}. \end{aligned}$$

$$\lambda \mid D \sim \text{Gamma}(a + \sum_i y_i, b + n), \quad \mathbb{E}[\lambda \mid D] = \frac{b}{b+n} \frac{a}{b} + \frac{n}{b+n} \bar{y}.$$

Add events to the shape; add exposure to the rate. If $Y_i \mid \lambda \sim \text{Poisson}(t_i \lambda)$, replace n by total exposure $\sum_i t_i$.

The original example — Pokémon counts in 20 districts



The twenty districts, with their counts.

THE OBSERVATIONS

14, 13, 7, 10, 15, 15, 2, 13, 13, 11, 10, 13, 5, 13, 9, 12, 9, 12, 8, 7.

Twenty districts: $n = 20$, total $\sum_i y_i = 211$, average $\bar{y} = 10.55$.

The source specifies prior mean 20 and standard deviation 10. In shape–rate notation,

$$\frac{a}{b} = 20, \quad \frac{a}{b^2} = 100 \implies b = 0.2, \quad a = 4.$$

The model treats districts as comparable units with one common rate. Unequal areas, search effort or spatial dependence require a richer model or exposure adjustment.

The count estimate — calculate

every number on this slide comes from a formula two slides back

Prior $\text{Gamma}(a, b) = \text{Gamma}(4, 0.2)$; data $n = 20$ districts with $\sum_i y_i = 211$, so $\bar{y} = 10.55$.

QUANTITY	FORMULA	PUT THE NUMBERS IN	VALUE
Posterior	$\text{Gamma}(a + \sum_i y_i, b + n)$	$\text{Gamma}(4 + 211, 0.2 + 20)$	$\text{Gamma}(215, 20.2)$
Posterior mean	$\frac{a + \sum_i y_i}{b + n}$	$\frac{215}{20.2}$	10.644
Posterior variance	$\frac{a + \sum_i y_i}{(b + n)^2}$	$\frac{215}{20.2^2} = \frac{215}{408.04}$	0.527
Standard deviation	$\sqrt{\text{Var}}$	$\sqrt{0.527}$	0.726
95% credible interval	2.5% and 97.5% quantiles of $\text{Gamma}(215, 20.2)$	by software, or the normal approximation $10.644 \pm 1.96 \times 0.726$	[9.27, 12.11] exact; [9.22, 12.07] approx.

The interval needs a quantile function — `qgamma` in R, `scipy.stats.gamma.ppf` in Python. The normal approximation lands within 0.05 of it here because the shape 215 is large; a Gamma is right-skewed, so the exact interval sits slightly to the right.

The count estimate — interpret

the same numbers, read as statements about the rate

WHY THE MEAN IS 10.644

Two slides back the posterior mean was written as a weighted average of the prior mean and the sample mean:

$$\mathbb{E}[\lambda \mid D] = \underbrace{\frac{0.2}{20.2}}_{0.99\%} \times 20 + \underbrace{\frac{20}{20.2}}_{99.01\%} \times 10.55 = 0.198 + 10.446.$$

The prior rate $b = 0.2$ is worth **0.2 of a district**. Twenty real districts outweigh it a hundred to one, so the estimate lands next to $\bar{y}$.

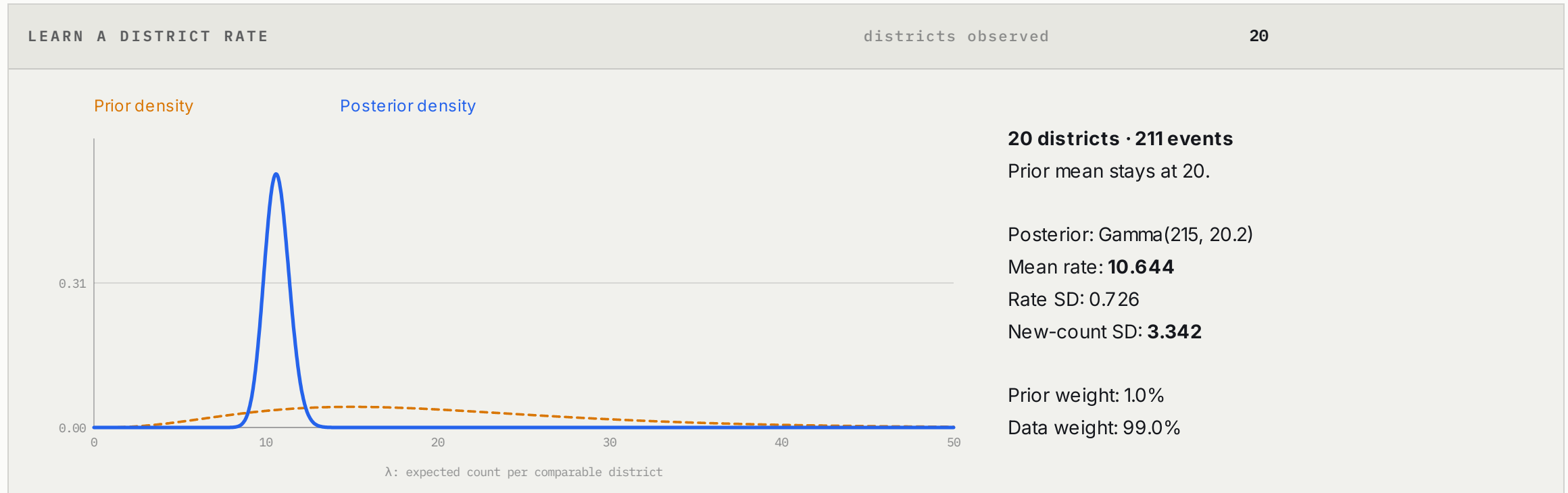
WHAT THE INTERVAL SAYS

With 95% probability the **common rate** λ lies in $[9.27, 12.11]$: about nine to twelve sightings per district, on average.

It is a statement about λ , a parameter. It is **not** a range for the count in one new district — that count also carries Poisson noise, and its interval is wider. The next slides make that distinction exact.

This interval describes the **common rate** λ . It is not a prediction interval for the count in one new district.

Experiment 5 — observe the districts one by one



Predict: will 20 districts outweigh the original prior? **Try:** move from 0 to 1 to 20 districts; then increase prior exposure while keeping its mean fixed. **Explain:** distinguish uncertainty in the rate from variability in a new district.

Predict one new district — rate uncertainty is not count noise

Return to $\lambda \mid D \sim \text{Gamma}(215, 20.2)$. For one comparable new district, $\tilde{Y} \mid \lambda \sim \text{Poisson}(\lambda)$.

The law of total variance gives

$$\begin{aligned}\mathbb{E}[\tilde{Y} \mid D] &= \mathbb{E}[\lambda \mid D] = 10.644, \\ \text{Var}(\tilde{Y} \mid D) &= \underbrace{\mathbb{E}[\lambda \mid D]}_{\text{count noise}} + \underbrace{\text{Var}(\lambda \mid D)}_{\text{rate uncertainty}} \\ &\approx 10.644 + 0.527 \approx 11.17.\end{aligned}$$

The rate's posterior standard deviation is **0.726**; a new count's predictive standard deviation is **3.342**. **Knowing the rate well does not make every district identical.**

The predictive distribution is Negative Binomial, a Poisson–Gamma mixture. Its probability mass function and parameter convention are in the appendix.

Prior $\times$ likelihood — write down what is being multiplied

Let $Y_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$ independently with σ^2 known, and $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$.

LIKELIHOOD

$$p(y \mid \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \theta)^2}{2\sigma^2} \right\}$$

PRIOR

$$p(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp \left\{ -\frac{(\theta - \mu_0)^2}{2\tau_0^2} \right\}$$

Only the shape in θ matters, so every factor without a θ in it goes into the constant:

$$p(\theta \mid y) \propto p(y \mid \theta) p(\theta) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 - \frac{1}{2\tau_0^2} (\theta - \mu_0)^2 \right\}$$

Multiply two Gaussians and the exponent is **a quadratic in θ** — so the posterior has no choice but to be Gaussian.

Multiply out, collect, complete the square

Expand both squares, writing $\bar{y} = \frac{1}{n} \sum_i y_i$ and keeping only the terms that carry a θ :

$$\begin{aligned}\sum_i (y_i - \theta)^2 &= n\theta^2 - 2\theta n\bar{y} + \sum_i y_i^2 \\ (\theta - \mu_0)^2 &= \theta^2 - 2\theta\mu_0 + \mu_0^2\end{aligned}$$

Divide each by its variance and gather the two powers of θ :

$$-\frac{1}{2} \left[\underbrace{\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right)}_A \theta^2 - 2 \underbrace{\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right)}_B \theta \right]$$

One completed square finishes it — and the leftover B^2/A has no θ , so it joins the constant:

$$A\theta^2 - 2B\theta = A\left(\theta - \frac{B}{A}\right)^2 - \frac{B^2}{A} \quad \implies \quad p(\theta | y) \propto \exp \left\{ -\frac{A}{2} \left(\theta - \frac{B}{A}\right)^2 \right\}$$

That is a Normal density with **precision A and mean B/A** . Nothing was assumed; it fell out.

Normal observations — combine information by precision

Reading A and B/A back gives the posterior in the form worth remembering.

PRECISIONS ADD

$$\frac{1}{\tau_1^2} = \underbrace{\frac{1}{\tau_0^2}}_{\text{prior}} + \underbrace{\frac{n}{\sigma^2}}_{\text{data}}$$

$\bar{Y} \mid \theta \sim \mathcal{N}(\theta, \sigma^2/n)$, so the data carry precision n/σ^2 — **information adds, variances do not.**

THE MEAN IS A WEIGHTED AVERAGE

$$\mu_1 = \frac{(1/\tau_0^2)\mu_0 + (n/\sigma^2)\bar{y}}{1/\tau_0^2 + n/\sigma^2}$$

Each side is weighted by exactly the precision it brings, and $\theta \mid D \sim \mathcal{N}(\mu_1, \tau_1^2)$.

SHARPER PRIOR

$\tau_0^2 \downarrow$ — the prior mean μ_0 is more trusted and pulls μ_1 harder.

QUIETER SENSOR

$\sigma^2 \downarrow$ — each reading is more precise, so $\bar{y}$ pulls harder.

MORE DATA

$n \uparrow$ — $\bar{y}$ pulls harder still, and eventually the prior stops mattering.

Four temperature readings — add the precisions

the previous slide's two formulas, with numbers put in

Prior $\theta \sim \mathcal{N}(20, 2^2)$; measurement noise has standard deviation 3; four readings average 23.

Step 1 — state each source as a precision, since that is the currency that adds:

$$\underbrace{\frac{1}{\tau_0^2} = \frac{1}{2^2} = \frac{1}{4}}_{\text{the prior}} \quad \text{and} \quad \underbrace{\frac{n}{\sigma^2} = \frac{4}{3^2} = \frac{4}{9}}_{\text{four readings, } \bar{y}=23}.$$

Step 2 — add them, then invert to come back to a variance:

$$\frac{1}{\tau_1^2} = \frac{1}{4} + \frac{4}{9} = \frac{25}{36}, \quad \tau_1^2 = \frac{36}{25} = 1.44, \quad \tau_1 = 1.2.$$

Four temperature readings — weight the means

36% prior, 64% data — and the estimate lands at 21.92

Step 3 — weight each mean by the precision that carries it. Dividing through by the total precision $25/36$ turns the formula into an ordinary weighted average:

$$\mu_1 = \frac{\frac{1}{4}(20) + \frac{4}{9}(23)}{\frac{25}{36}} = \underbrace{\frac{9}{25}}_{0.36} (20) + \underbrace{\frac{16}{25}}_{0.64} (23) = 21.92.$$

The four readings took **64% of the weight and left the prior 36%**, so the estimate moved from 20 to 21.92 — most of the way to the data, but not all of it.

Precision decides the split, and four readings at $\sigma = 3$ are worth just under twice a prior at $\tau_0 = 2$.

Temperature thread — update an uncertain temperature

shared teaching example · predict before revealing the calculation

The true temperature θ has prior $\mathcal{N}(20, 4)$. One sensor reading is $y = 22$, with $y \mid \theta \sim \mathcal{N}(\theta, 1)$. The second parameter here is variance.

Predict: Will the posterior mean be closer to 20 or to 22? Will its variance equal the variance of a new reading?

Calculate and check. The gain is $K = 4/(4 + 1) = 0.8$. The posterior mean is $20 + 0.8(22 - 20) = \mathbf{21.6}$ and variance is $(1 - 0.8)4 = \mathbf{0.8}$. A new reading with independent sensor noise has predictive variance $0.8 + 1 = \mathbf{1.8}$.

Learning the unknown temperature and predicting another noisy measurement are different questions.

Try it — the sensor variance rises to 4

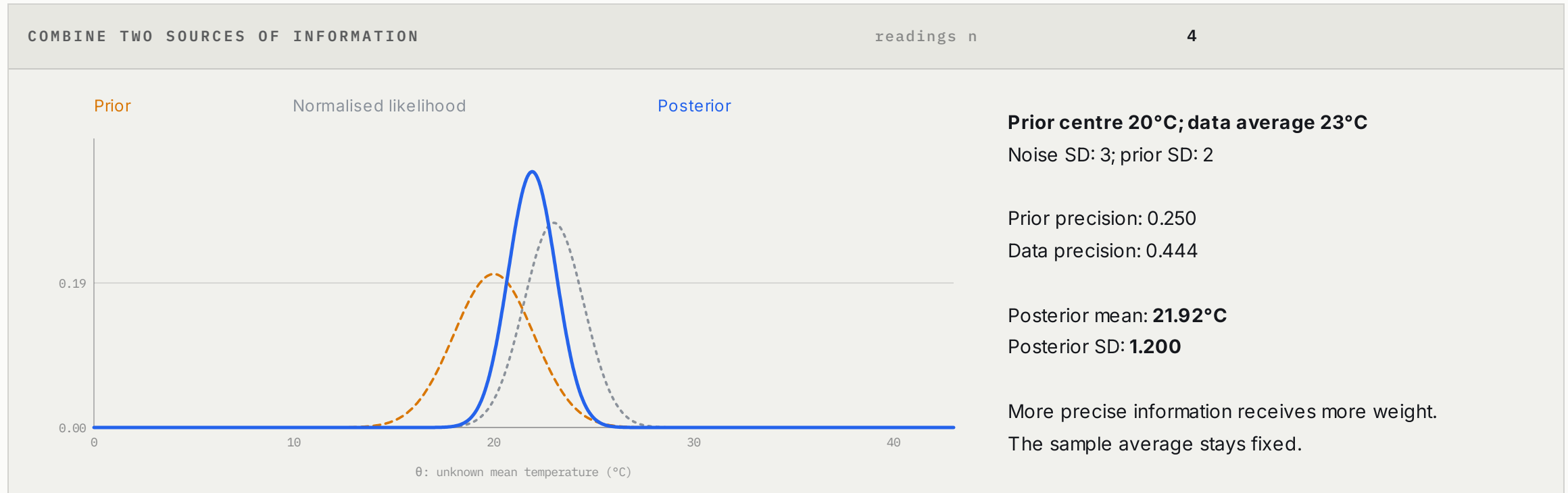
work independently · reveal only after writing an answer

Keep the prior and the observed value 22. Change the sensor variance from 1 to 4, for both this and the next reading. Find the posterior mean, posterior variance and predictive variance.

Check your answer. $K = 4/(4 + 4) = 0.5$, posterior mean **21**, posterior variance **2**, predictive variance **6**. The less precise reading pulls the posterior less strongly.

State which distribution each variance belongs to before adding or comparing numbers.

Experiment 6 — vary precision with the same average



Predict: does noisier data move the estimate closer to 20 or 23? **Try:** change noise SD, prior SD, and the number of readings. **Explain:** the average stays at 23; precision determines its weight. The grey likelihood is normalised for plotting.

Normal prediction — restore the observation noise

The updated mean parameter is $\theta \mid D \sim \mathcal{N}(\mu_1, \tau_1^2)$. A new reading is $\tilde{Y} = \theta + \epsilon$, with independent $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

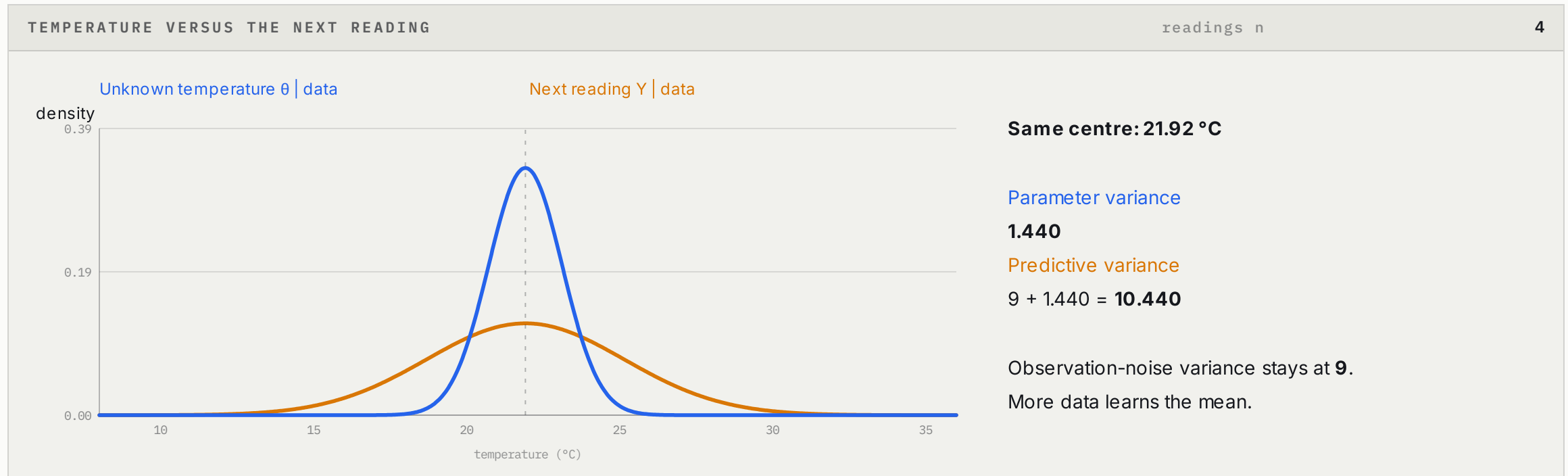
$$\tilde{Y} \mid D \sim \mathcal{N}(\mu_1, \tau_1^2 + \sigma^2).$$

For the previous four readings:

$$\theta \mid D \sim \mathcal{N}(21.92, 1.44), \quad \tilde{Y} \mid D \sim \mathcal{N}(21.92, 10.44).$$

Same centre, different uncertainty. **Learning the temperature does not remove noise from the thermometer.**

Experiment 7 — compare the mean and the next reading



Predict: which curve keeps a noise floor? **Try:** increase readings from 1 to 100. **Explain:** posterior parameter variance shrinks, while predictive variance remains 9 plus that variance. Here noise SD is fixed at 3 and prior SD at 2.

More than two outcomes — the Dirichlet update

the categorical model

Suppose each service request is **delivery**, **pickup** or **return**. The probabilities form a vector θ with non-negative entries adding to 1.

$$\theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3), \quad (Y_1, Y_2, Y_3) \mid \theta \sim \text{Multinomial}(n, \theta).$$

The likelihood adds one count to the corresponding exponent for each observation:

$$\theta \mid D \sim \text{Dirichlet}(\alpha_1 + Y_1, \alpha_2 + Y_2, \alpha_3 + Y_3).$$

The Beta update with a longer list of categories. With two categories, the Dirichlet reduces to a Beta distribution for the first probability.

An unseen category need not get zero probability

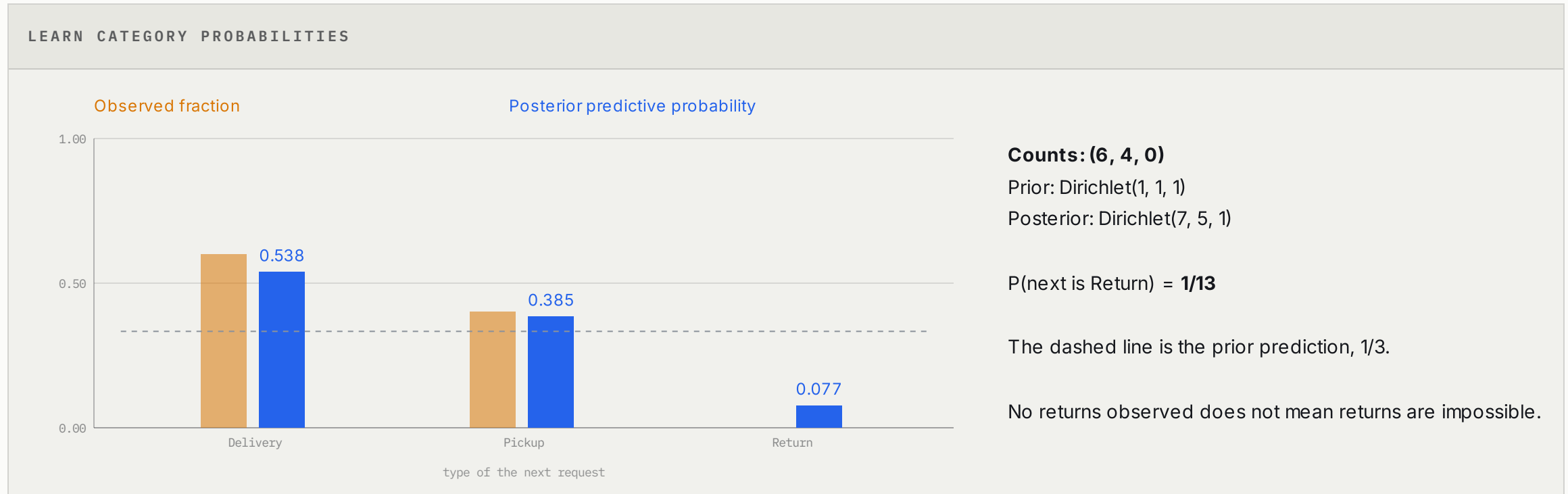
Start from $\text{Dirichlet}(1, 1, 1)$ and observe ten requests: **6 deliveries, 4 pickups, 0 returns**.

CATEGORY	COUNT	MLE	POSTERIOR PARAMETER	NEXT-REQUEST PROBABILITY
Delivery	6	$6/10$	$1 + 6 = 7$	$7/13$
Pickup	4	$4/10$	$1 + 4 = 5$	$5/13$
Return	0	$0/10$	$1 + 0 = 1$	$1/13$

The posterior is $\text{Dirichlet}(7, 5, 1)$ and predictive probabilities add to 1. The prior gives an unseen but possible category some support.

This update can estimate the rows of a transition matrix. **Ten observations without a return do not prove returns are impossible.**

Experiment 8 — observe a previously unseen category



Predict: is a return impossible after zero observed returns? **Try:** add Return, change the prior strength, then clear the data. **Explain:** posterior parameters are prior pseudo-counts plus observed counts; divide by their sum to predict one request.

Check — what uncertainty belongs in a new measurement?

With known noise variance σ^2 and posterior mean variance τ_1^2 , what is the variance of the next Normal observation?

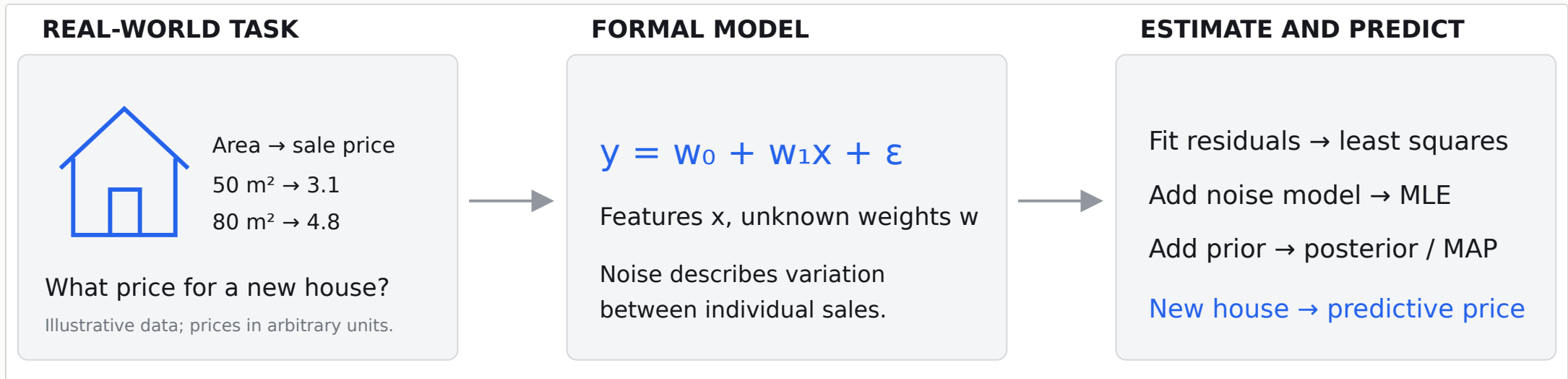
- ☐ A Only τ_1^2 , because the posterior contains everything
- ☐ B Only σ^2 , because the parameter has been estimated
- ☐ C Zero once enough observations have arrived
- ☒ D $\tau_1^2 + \sigma^2$, because both sources contribute

The unknown mean and independent noise both vary in the predictive calculation. More data learns the mean but does not eliminate the assumed observation noise. We will use exactly this idea for regression.

4 · Bayesian regression and regularisation

Follow the original regression route: formulate the task, solve least squares, interpret MLE, keep a Bayesian posterior, then motivate and interpret regularisation.

Problem solving — estimate a housing price



Task → model → algorithm. The two house-price observations are illustrative, not an empirical housing dataset.

We now learn a vector of weights. **The Bayesian logic stays the same as for a coin, a rate, or a mean.**

Regression roadmap — the same model from several views

We keep $y_i = x_i^\top w + \epsilon_i$ throughout.



One regression model, several ways to describe what learning means. We will connect the objectives instead of introducing unrelated methods.

Put a line into matrix form

Include the intercept by adding a constant feature 1. For the illustrative data $(x, y) = (0, 1), (1, 2), (2, 2)$,

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \quad Xw = \begin{pmatrix} w_0 \\ w_0 + w_1 \\ w_0 + 2w_1 \end{pmatrix}.$$

Each row is one observation. Each column is one feature. **The residual vector is $y - Xw$.**

In general $X \in \mathbb{R}^{n \times d}$, $w \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$. Here d includes the intercept. Keep this orientation throughout the derivation.

Least squares — use Lecture 1's optimality condition

Choose weights to minimise squared residuals:

$$J(w) = \frac{1}{2} \|y - Xw\|^2, \quad \nabla J(w) = X^\top (Xw - y).$$

The Hessian is $X^\top X \succeq 0$, so J is convex. Therefore

$$\hat{w} \text{ minimises } J \iff X^\top X \hat{w} = X^\top y.$$

This is the **normal equation**. The zero-gradient condition is exact here because the problem is convex, differentiable and unconstrained.

Only when X has full column rank is the minimiser unique and $(X^\top X)^{-1} X^\top y$ valid. Otherwise minimisers are not unique; use QR or SVD methods for the linear algebra instead of explicitly forming an inverse.

Solve the three-point regression by hand

For the previous three observations,

$$X^{\top} X = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}, \quad X^{\top} y = \begin{pmatrix} 5 \\ 6 \end{pmatrix}.$$

The normal equation gives $3w_0 + 3w_1 = 5$ and $3w_0 + 5w_1 = 6$. Subtract to get $w_1 = 1/2$, then $w_0 = 7/6$.

$$\hat{y} = \begin{pmatrix} 7/6 \\ 5/3 \\ 13/6 \end{pmatrix}, \quad y - \hat{y} = \begin{pmatrix} -1/6 \\ 1/3 \\ -1/6 \end{pmatrix}, \quad \|y - \hat{y}\|^2 = 1/6.$$

A best-fitting line need not pass through every point. **It balances the residuals to minimise their squared total.**

Why Gaussian noise turns MLE into least squares

Assume independent noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with known $\sigma^2 > 0$:

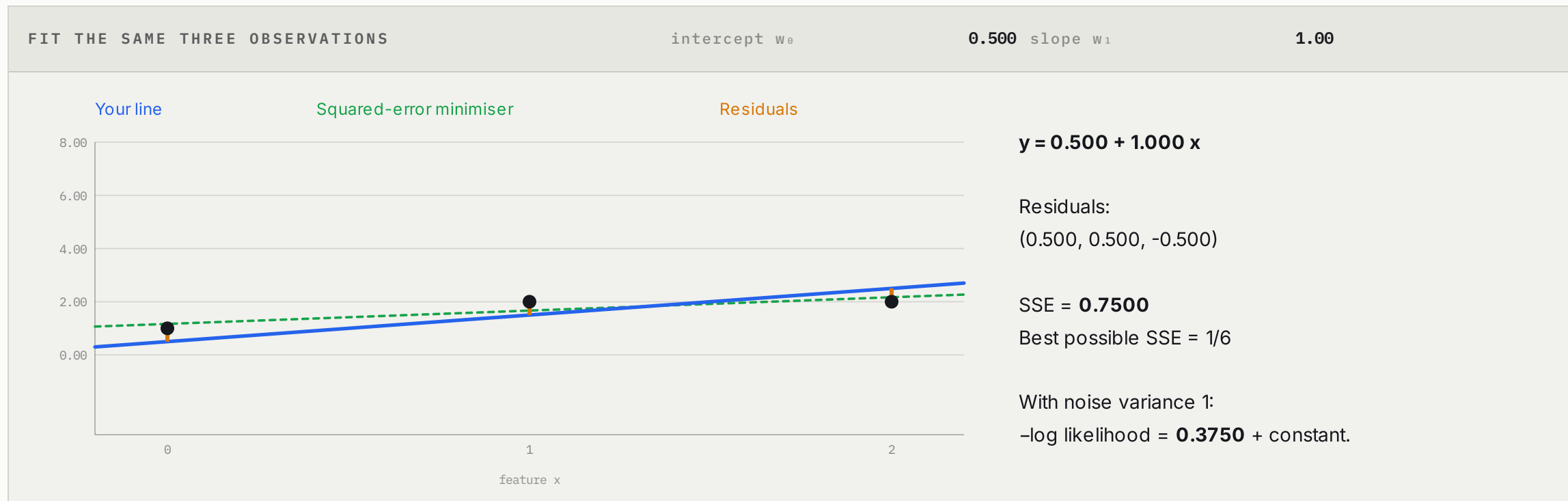
$$p(y \mid X, w) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - x_i^\top w)^2}{2\sigma^2}\right].$$

Take the negative log:

$$-\log p(y \mid X, w) = \underbrace{\frac{n}{2} \log(2\pi\sigma^2)}_{\text{constant in } w} + \frac{1}{2\sigma^2} \|y - Xw\|^2.$$

For fixed σ^2 , **maximising this likelihood is exactly minimising squared error**. Least squares itself does not require Gaussian noise; this probabilistic interpretation does.

Experiment 9 — move the line and watch its residuals



Predict: does making one residual zero minimise the total error? **Try:** adjust intercept and slope, then select Fit least squares. **Explain:** the same minimiser gives SSE = 1/6 and maximises the displayed Gaussian likelihood.

From MLE to Bayesian regression

MLE: CHOOSE ONE WEIGHT VECTOR

$$\hat{w}_{\text{ML}} = \arg \max_w p(y \mid X, w).$$

We retain the weight vector giving the largest likelihood.

BAYESIAN: UPDATE A DISTRIBUTION

$$p(w \mid X, y) \propto p(y \mid X, w)p(w).$$

MAP is its most dense point; full Bayes also keeps the spread and correlations.

The prior acts on weights, just as the Beta prior acted on the coin bias. A Gaussian prior makes this update analytically tractable.

Complete the square again — now with a matrix

Same two factors as before, with w in place of θ . Keep only what carries a w :

$$\log p(w \mid X, y) = -\frac{1}{2\sigma^2} (y - Xw)^\top (y - Xw) - \frac{1}{2\tau^2} w^\top w + C.$$

Expand $(y - Xw)^\top (y - Xw) = y^\top y - 2w^\top X^\top y + w^\top X^\top Xw$ and gather the two powers of w :

$$-\frac{1}{2} \left[w^\top \underbrace{\left(\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2} \right)}_A w - 2w^\top \underbrace{\frac{X^\top y}{\sigma^2}}_b \right]$$

The scalar identity survives transposition — A is symmetric, so

$$w^\top Aw - 2w^\top b = (w - A^{-1}b)^\top A (w - A^{-1}b) - b^\top A^{-1}b,$$

and the leftover carries no w . What is left is a Gaussian with **precision matrix A and mean $A^{-1}b$** .

Line for line the scalar derivation, with a matrix where the number was: A was $\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}$ and is now $\frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2}$; the mean was B/A and is now $A^{-1}b$. **Precision still adds; the mean is still precision-weighted.**

Full Bayes keeps the distribution over weights

With independent noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and prior $w \sim \mathcal{N}(0, \tau^2 I)$, the posterior is also Gaussian:

$$w \mid X, y \sim \mathcal{N}(\mu_w, \Sigma_w).$$

THE CENTRE

$$\mu_w = \Sigma_w \frac{X^\top y}{\sigma^2}.$$

Here the mean and mode coincide, so μ_w is also the MAP estimate. Later we will connect it to ridge regression.

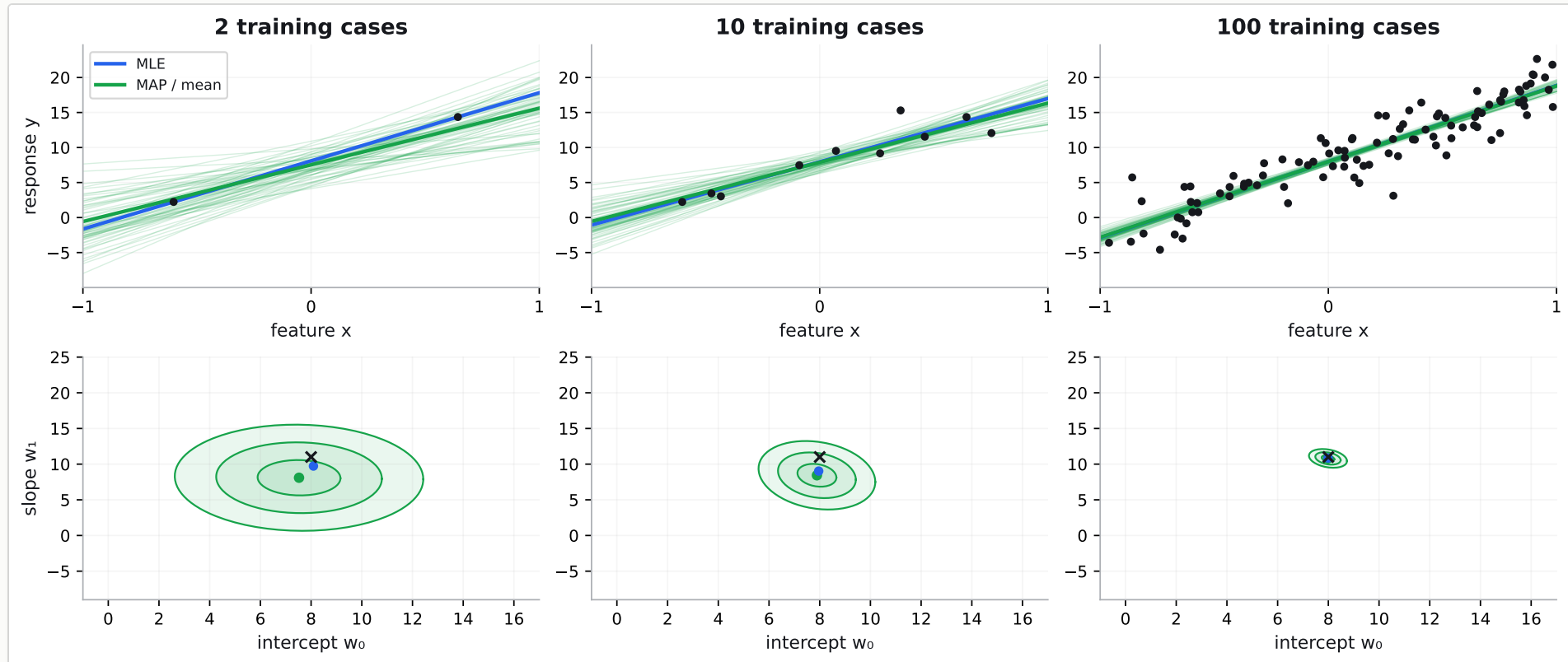
THE UNCERTAINTY

$$\Sigma_w^{-1} = \frac{X^\top X}{\sigma^2} + \frac{I}{\tau^2}.$$

Data precision plus prior precision. Large posterior variance marks directions in the weights that remain uncertain.

Read straight off the previous slide: $\Sigma_w = A^{-1}$ and $\mu_w = A^{-1}b$. Known positive σ^2 and τ^2 are assumed, which is what makes A positive definite — and therefore invertible — even when X has fewer rows than columns.

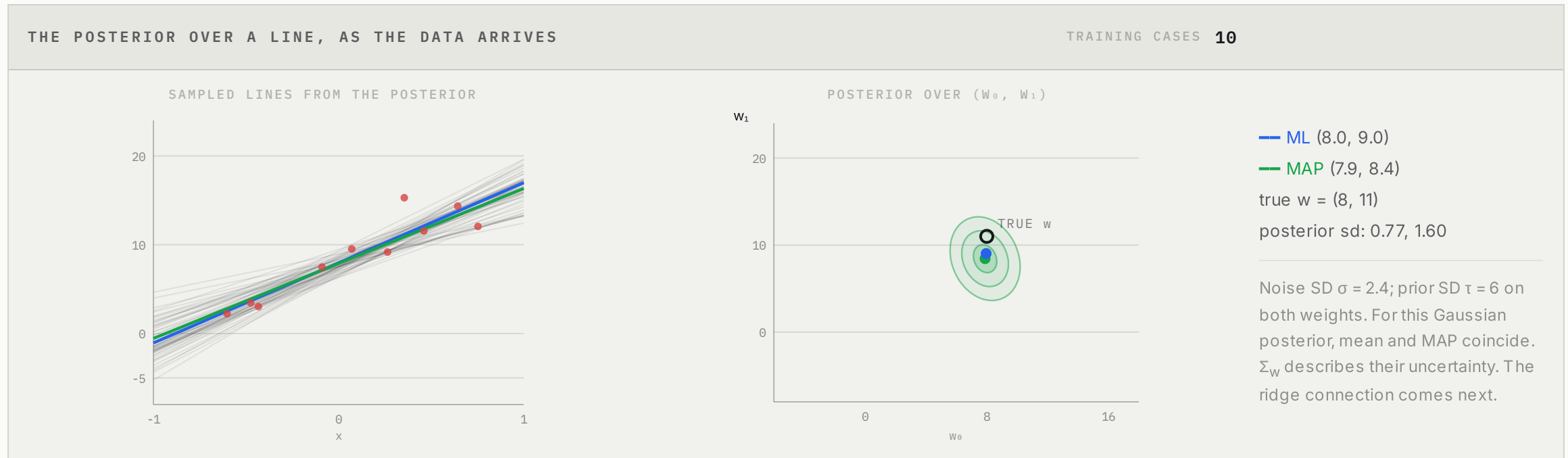
Bayesian regression — one posterior in two spaces



Two rows, three columns, on nested simulated datasets ($n = 2, 10, 100$). Blue: MLE; green: posterior mean/MAP and sampled lines; x: true weights. Corresponding axes use common scales.

Top: plausible mean functions. Bottom: plausible weights. [More informative data contracts both views of the same posterior.](#)

Experiment 10 — change data, watch the posterior over a line



Predict: which fit is most unstable with two cases? **Try:** compare 2, 10, and 100 nested cases, then draw new data. **Explain:** the left panel shows sampled mean functions; the right shows the same uncertainty in coefficient space.

Regression prediction — a line and a noisy observation

At a new feature vector x_* , the fitted mean is $f_* = x_*^\top w$. Since $w \mid D \sim \mathcal{N}(\mu_w, \Sigma_w)$,

$$f_* \mid D \sim \mathcal{N}(x_*^\top \mu_w, x_*^\top \Sigma_w x_*).$$

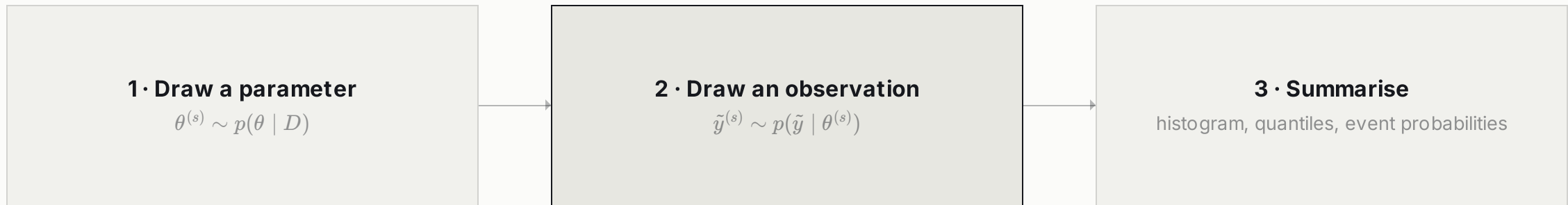
An individual response also includes noise, $\tilde{Y}_* = f_* + \epsilon_*$:

$$\tilde{Y}_* \mid D \sim \mathcal{N}\left(x_*^\top \mu_w, \underbrace{x_*^\top \Sigma_w x_*}_{\text{parameter uncertainty}} + \underbrace{\sigma^2}_{\text{observation noise}}\right).$$

Drawing posterior lines shows uncertainty about the mean function. **Add noise to predict actual observations.** Both depend on the linear model being appropriate.

When the integral is hard — simulate the two stages

For $s = 1, \dots, M$:



For regression, draw $w^{(s)}$ from its Gaussian posterior, then draw noise and set $\tilde{y}_*^{(s)} = x_*^\top w^{(s)} + \epsilon_*^{(s)}$.

Conjugate examples allow direct sampling. Other models may require MCMC or an approximation, whose accuracy must be checked. See the [Stan posterior prediction guide](#) for the same two-stage construction.

Why regularise a regression?

The original goal is to predict **new observations**, not only to fit the training cases.

A FLEXIBLE FIT

Extra features can lower training error. With limited or correlated data, fitted weights may change substantially when the sample changes.

A CONTROLLED FIT

Penalising large weights can reduce this sensitivity. We accept some fitting bias in exchange for potentially better predictions.

$$\hat{w}_{\text{ridge}} = \arg \min_w \{ \|y - Xw\|^2 + \lambda_2 \|w\|^2 \}.$$

Check the tradeoff on independent validation data. Neither more flexibility nor more shrinkage is automatically better.

Why shrinkage helps — and what it changes

Ridge changes the normal equation to

$$(X^{\top}X + \lambda_2 I)\hat{w} = X^{\top}y.$$

SMALL OR CORRELATED DATASETS

Several weight vectors may fit almost equally well. Small data perturbations can then produce large changes in the OLS weights.

Adding $\lambda_2 I$ makes the system positive definite for $\lambda_2 > 0$ and stabilises these directions.

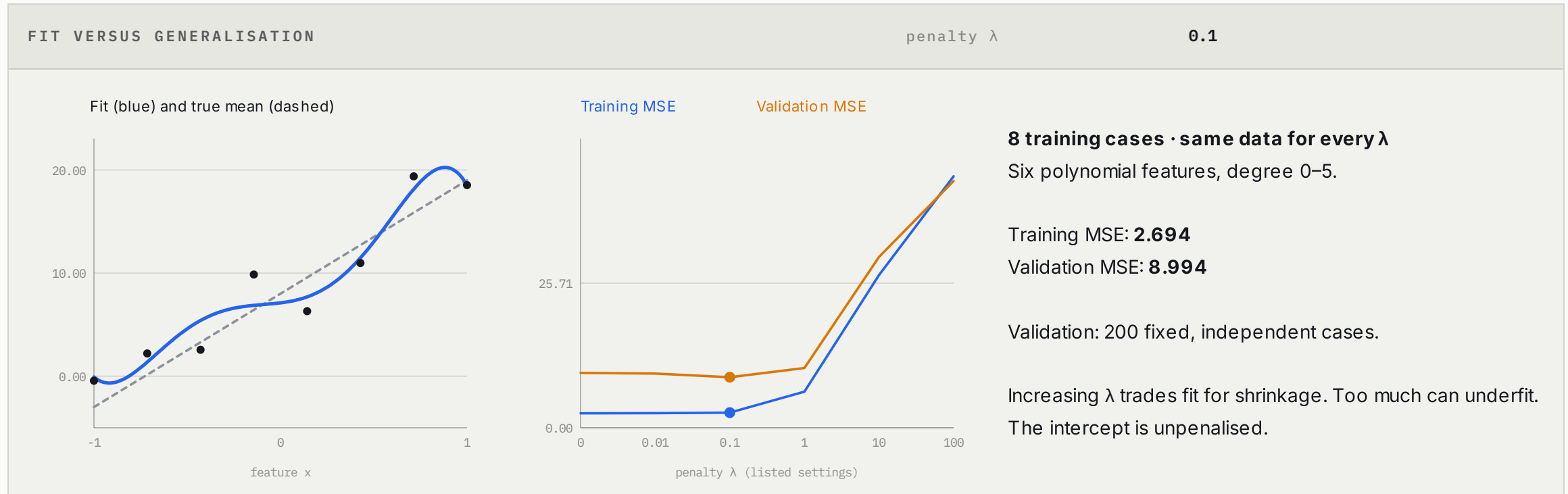
THE TRADEOFF

Shrinkage introduces bias toward the prior mean for potentially lower variance. It can help prediction but does not guarantee better test performance.

Choose strength with prior knowledge and appropriate validation; inspect feature scaling.

The Gaussian-posterior example gives all weights a prior. The next experiment leaves the intercept unpenalised; applications often treat it separately.

Experiment 11 — compare training fit with validation error



Predict: must the best training fit predict best? **Try:** sweep λ on one fixed dataset, then change its size or draw new training data. **Explain:** polynomial features keep the model linear in its weights; validation data stays separate and fixed.

Add a Gaussian prior — derive ridge, not just name it

Let $w \sim \mathcal{N}(0, \tau^2 I)$ with known $\tau^2 > 0$. The negative log posterior is

$$-\log p(w \mid X, y) = \frac{1}{2\sigma^2} \|y - Xw\|^2 + \frac{1}{2\tau^2} \|w\|^2 + \text{constant}.$$

Multiply by $2\sigma^2$, which does not change the minimiser:

$$\hat{w}_{\text{MAP}} = \arg \min_w \{ \|y - Xw\|^2 + \lambda_2 \|w\|^2 \}, \quad \lambda_2 = \sigma^2 / \tau^2.$$

A smaller prior variance penalises large weights more strongly. The regulariser is the **negative log prior**, with the noise scale accounted for.

Lasso — a different prior changes the penalty

Independent Laplace priors with scale $b_L > 0$ have

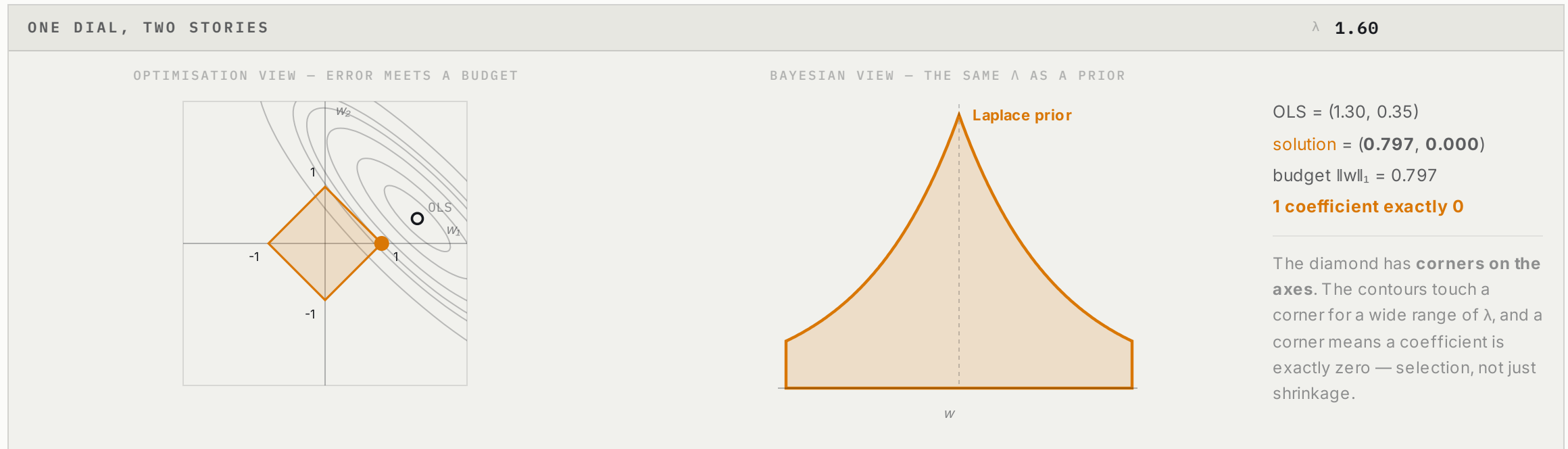
$$p(w) = \prod_j \frac{1}{2b_L} e^{-|w_j|/b_L}.$$

With the same Gaussian likelihood,

$$\hat{w}_{\text{MAP}} = \arg \min_w \left\{ \|y - Xw\|^2 + \lambda_1 \|w\|_1 \right\}, \quad \lambda_1 = \frac{2\sigma^2}{b_L}.$$

The absolute-value penalty has a corner at zero, so some **MAP coefficients** can be exactly zero. A continuous Laplace prior has **no probability mass at the single point zero**.

Experiment 12 — see why ridge shrinks and lasso can select



Predict: which constraint has corners on the axes? **Try:** change λ and switch between ridge and lasso. **Explain:** the contact point is the fitted weight vector. Priors are scaled to peak height 1 to compare shapes; this illustration uses noise variance 1.

Check — the prior wearing a disguise

Under Gaussian noise, what prior gives MAP the ridge objective $\|y - Xw\|^2 + \lambda_2 \|w\|^2$?

- ☐ A A Laplace prior, because ridge selects exact zeros
- ☐ B A uniform prior over all of $\mathbb{R}^d$
- ☒ C A zero-mean Gaussian prior with variance $\tau^2 = \sigma^2 / \lambda_2$
- ☐ D No prior can produce a regularisation term

Taking the negative log turns the Gaussian likelihood into squared residuals and the Gaussian prior into a squared-weight penalty. For this unaveraged objective, $\lambda_2 = \sigma^2 / \tau^2$. Averaging the data loss by n changes the coefficient convention.

What we can now calculate

TASK	CALCULATION	EXAMPLE
Update a belief	$\text{posterior} \propto \text{prior} \times \text{likelihood}$	Beta(2,2) + HHT $\rightarrow$ Beta(4,3)
Choose a point	MLE, MAP or a loss-based posterior summary	Gaussian prior + Gaussian noise $\rightarrow$ ridge
Predict data	average the sampling model over the posterior	uncertain Poisson rate $\rightarrow$ Negative Binomial
Check the model	simulate, compare and revise	replicated counts or regression residuals

Lecture 3 asks how to represent many interacting variables. [Conditional independence makes a large probabilistic model manageable.](#)

Questions?

Can you explain which uncertainty is over a parameter, which is over an observation, and where each appears in the calculation?

— APPENDIX

Appendix · extra derivations and checks

Optional details for questions and self-study; the main lecture has completed the original route.

Reading guide — one Bayesian calculation, several observation models

one main idea to explain, one comparison, one application

ROLE	READ OR REVISIT	QUESTION TO ANSWER
Core	Murphy, <i>Probabilistic Machine Learning: An Introduction</i> (2022): probability, Bayesian statistics and linear regression sections	What distribution is specified, updated, and used for prediction?
Compare	MLE, MAP and full Bayes on the same coin or regression data	Which answer is a point estimate, and which carries uncertainty?
Apply	The original Pokémon district counts and housing-price example	How does the observation model affect the estimate and prediction?

Use selected sections as a reference, not the entire book as an assignment. For each example write the likelihood, prior, posterior and predictive target before doing algebra.

Backup — integrate the Beta–Binomial prediction

Write $\alpha' = \alpha + S$, $\beta' = \beta + n - S$. For K heads in m **future tosses**,

$$\begin{aligned} P(K = k \mid D) &= \int_0^1 \binom{m}{k} \theta^k (1 - \theta)^{m-k} \frac{\theta^{\alpha'-1} (1 - \theta)^{\beta'-1}}{B(\alpha', \beta')} d\theta \\ &= \binom{m}{k} \frac{B(\alpha' + k, \beta' + m - k)}{B(\alpha', \beta')}. \end{aligned}$$

For $m = 2$ and $\text{Beta}(4, 3)$, probabilities of $K = 0, 1, 2$ are $3/14, 6/14, 5/14$. They sum to 1.

Before observing data, use α, β to obtain the **prior predictive**. With a $\text{Beta}(2, 2)$ prior and two tosses, its probabilities are 0.3, 0.4, 0.3. Future batch size m need not equal training size n .

Backup — MLE, MAP and posterior mean

ESTIMATE	DEFINITION	BETA(2,2) PRIOR + HHT
MLE	$\arg \max_{\theta} p(D \mid \theta)$	2/3
MAP	$\arg \max_{\theta} p(\theta \mid D) = \arg \max_{\theta} p(D \mid \theta)p(\theta)$	3/5
Posterior mean	$\mathbb{E}[\theta \mid D]$	4/7

MAP picks the posterior's mode. The posterior mean minimises posterior expected squared-error loss; the posterior median minimises expected absolute-error loss.

The posterior mode depends on the parameterisation. With abundant informative data and suitable regularity, MLE and MAP may become close; they need not be identical for a finite sample.

Backup — when does prior influence fade?

In the Beta–Binomial model, for a fixed proper Beta prior,

$$\text{prior weight} = \frac{\alpha + \beta}{\alpha + \beta + n} \longrightarrow 0.$$

This does not mean any prior can be overcome in any model:

- An identifiable model needs informative observations about the parameter.
- A prior excluding a whole region cannot acquire posterior mass there by multiplication.
- Model misspecification can produce confident but misleading conclusions.

Check prior sensitivity when data is limited. **More data is not a substitute for a plausible model.**

Backup — checking a fitted probability model

BEFORE FITTING: PRIOR PREDICTIVE CHECK

Draw parameters from the prior, then simulate data. Do plausible parameters produce plausible counts, prices or temperatures?

AFTER FITTING: POSTERIOR PREDICTIVE CHECK

Draw parameters from the posterior and simulate replicated datasets of the same design. Compare their spread, extremes and patterns with the observations.

For the Pokémon example: compare the observed mean 10.55 and sample variance 11.94 with replicated 20-district datasets. Also examine exposure and spatial patterns. Mean–variance similarity alone does not establish the Poisson assumptions.

Use a mismatch to improve the model. Evaluate future prediction on held-out data too; checking training data is not a test of generalisation.

Workflow reference: [Stan, posterior and prior predictive checks](#).

Backup — Beta–Binomial normalisation and moments

Let S heads be observed in n tosses and let $\alpha' = \alpha + S$, $\beta' = \beta + n - S$.

$$p(\theta \mid D) = \frac{\theta^{\alpha'-1}(1-\theta)^{\beta'-1}}{B(\alpha', \beta')}, \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

$$\mathbb{E}[\theta \mid D] = \frac{\alpha'}{\alpha' + \beta'}, \quad \text{Var}(\theta \mid D) = \frac{\alpha' \beta'}{(\alpha' + \beta')^2(\alpha' + \beta' + 1)}.$$

For K heads in m future tosses and $\mu = \alpha' / (\alpha' + \beta')$,

$$P(K = k \mid D) = \binom{m}{k} \frac{B(\alpha' + k, \beta' + m - k)}{B(\alpha', \beta')}, \quad \text{Var}(K \mid D) = m\mu(1-\mu) \frac{\alpha' + \beta' + m}{\alpha' + \beta' + 1}.$$

$\mathbb{E}[K \mid D] = m\mu$. The multiplier of Binomial variance is 1 at $m = 1$ and greater than 1 at $m > 1$. Future tosses share one latent bias; independently resampling a bias for each toss would be a different model.

Backup — the Poisson–Gamma predictive distribution

With posterior $\lambda \mid D \sim \text{Gamma}(a', b')$ in **shape–rate** form, a new unit-exposure count has

$$\begin{aligned} P(\tilde{Y} = k \mid D) &= \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} \frac{b'^{a'}}{\Gamma(a')} \lambda^{a'-1} e^{-b'\lambda} d\lambda \\ &= \frac{\Gamma(k + a')}{\Gamma(a')k!} \left(\frac{b'}{b' + 1} \right)^{a'} \left(\frac{1}{b' + 1} \right)^k, \quad k = 0, 1, \dots \end{aligned}$$

$$\mathbb{E}[\tilde{Y} \mid D] = \frac{a'}{b'}, \quad \text{Var}(\tilde{Y} \mid D) = \frac{a'}{b'} + \frac{a'}{b'^2}.$$

This is Negative Binomial with shape a' and probability $b'/(b' + 1)$ under the displayed convention; a' need not be an integer. For future exposure t , the mean is ta'/b' and variance $ta'/b' + t^2a'/b'^2$.

Backup — the Dirichlet predictive distribution

For category counts y_j , let $\alpha'_j = \alpha_j + y_j$ and $A' = \sum_j \alpha'_j$. Then

$$p(\theta \mid D) \propto \prod_j \theta_j^{\alpha'_j - 1}, \quad P(\tilde{Y} = j \mid D) = \frac{\alpha'_j}{A'}.$$

For a future batch of size m with counts k_j adding to m ,

$$P(K_1 = k_1, \dots, K_c = k_c \mid D) = \frac{m!}{\prod_j k_j!} \frac{\Gamma(A')}{\Gamma(A' + m)} \prod_j \frac{\Gamma(\alpha'_j + k_j)}{\Gamma(\alpha'_j)}.$$

This is the **Dirichlet–Multinomial** distribution. Updated α'_j give the posterior predictive; original α_j give the prior predictive.

Backup — complete the square for the Normal mean

Independent $Y_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$, known σ^2 , prior $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$:

$$\begin{aligned}\log p(\theta \mid D) &= -\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2 - \frac{1}{2\tau_0^2} (\theta - \mu_0)^2 + C \\ &= -\frac{1}{2} \left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right) \theta^2 - 2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right) \theta \right] + C' \\ &= -\frac{(\theta - \mu_1)^2}{2\tau_1^2} + C''.\end{aligned}$$

$$\tau_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2} \right)^{-1}, \quad \mu_1 = \tau_1^2 \left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2} \right).$$

C, C', C'' do not depend on θ . A new observation is θ plus independent Gaussian noise, so its predictive distribution is $\mathcal{N}(\mu_1, \tau_1^2 + \sigma^2)$.

Backup — three remarks on the regression posterior

The completion of the square now sits in the body, at *Complete the square again — now with a matrix*. Three things worth saying only if asked:

ALWAYS INVERTIBLE

$X^\top X$ alone is singular whenever X has fewer rows than columns. Adding I/τ^2 makes A positive definite, so a Bayesian fit is defined where least squares is not — the prior supplies the missing rank.

MEAN EQUALS MAP

A Gaussian's mean, mode and median coincide, so μ_w is simultaneously the posterior mean and the MAP estimate. That is special to the Gaussian and does not survive a skewed posterior.

WHAT THE POINT ESTIMATE DROPS

Σ_w is precisely what collapsing to μ_w throws away — and its large directions are the ones a prediction should hedge in.

Backup — why Lasso can select an exact zero

For one coordinate, write the objective, up to a constant, as

$$q(w_j) = a_j w_j^2 - 2r_j w_j + \lambda_1 |w_j|, \quad a_j > 0.$$

At zero, $\partial|w_j| = [-1, 1]$. The optimality condition is

$$0 \in -2r_j + \lambda_1 [-1, 1] \quad \Longleftrightarrow \quad |r_j| \leq \lambda_1/2.$$

Otherwise the solution is the soft-thresholded value

$$w_j = \frac{\text{sign}(r_j)}{a_j} \max(|r_j| - \lambda_1/2, 0).$$

This is a property of MAP optimisation. A continuous Laplace prior and the resulting continuous posterior have no atom at zero; full posterior samples are not sparse in this exact-zero sense.