Week 3 — Conjugate Bayes

Friday, May 15, 2026

Prof. Joseph Austerweil

Agenda

Welcome back	0:00
Conjugacy as a pattern	0:10
Beta-Binomial	0:25
Break	0:55
Gaussian-Gaussian with N observations	1:05
Subjective randomness (G&T 2001)	1:40
Close	1:55

Conjugacy as a pattern

The move we saw in Week 2

Prior: $\mu \sim N(500, 20^2)$

Likelihood: observation $D = 510$, known $\sigma = 30$

Posterior: $\mu \mid D \sim N(503.3, 16.6^2)$

Notice: prior is Gaussian → posterior is Gaussian. Same family.

What “conjugate” means — in words

Conjugate = the posterior stays in the *same family* as the prior.

Gaussian prior + Gaussian likelihood → Gaussian posterior.

Different parameters, same shape of distribution.

We just saw it once. The next question: is this lucky, or a pattern?

What “conjugate” means — formally

A prior family $\mathcal{F}$ is **conjugate to a likelihood** $p(D \mid \theta)$ when the posterior stays in $\mathcal{F}$ (same functional form, updated params):

$$p(\theta) \in \mathcal{F} \quad \Longrightarrow \quad p(\theta \mid D) \in \mathcal{F}$$

The likelihood typically lives in a *different* family — e.g. **Beta** prior + **Binomial** likelihood → Beta posterior. Conjugacy is a property of the *pair* (prior family, likelihood).

Last week’s Gaussian-Gaussian was the special case where both happened to be in the same family — not the general rule.

Why conjugacy matters

Property	Why you care
Closed-form posterior	No integration, no sampling
Sequential updates	Today’s posterior = tomorrow’s prior
Interpretable hyperparameters	Prior knowledge as “pseudo-observations”
Fast, exact	Good pedagogy + fast enough to compute on the fly

Quick check — when does conjugacy *fail*?

Prior on μ: **bimodal** (mixture of two Gaussians). Likelihood: Gaussian. Posterior?

A. Gaussian (likelihood dominates) B. Bimodal Gaussian mixture C. Uniform (prior × likelihood cancels) D. Not closed-form — needs numerics

Quick check — answer

B. Bimodal Gaussian mixture.

Each component updates Gaussian-Gaussian independently; the mixture weights re-balance by how well each component explains the data.

Conjugacy is a property of the *prior family* × *likelihood pair*, not the prior alone.

Beta-Binomial

Back to the bentos — but now with counts

Chibany's prior belief about tonkatsu rate:

70% tonkatsu, 30% hamburger — but how confident?

This semester's data: **27 tonkatsu out of 40 bentos**.

What's Chibany's updated belief about p?

The Beta distribution

$$\mathrm{Beta}(\alpha, \beta): \quad p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$\theta \in [0, 1]$ is the rate we're inferring (e.g. tonkatsu rate). $\propto$ = "proportional to" — drops the normalizing constant.

α, β	**Shape interpretation**
$\alpha = \beta = 1$	Uniform on $[0, 1]$
$\alpha > \beta$	Concentrated above 0.5
Large $\alpha + \beta$	Narrow (confident)
Small $\alpha + \beta$	Wide (uncertain)
$\alpha = \beta < 1$	U-shaped at the edges

Poll — Tanaka’s attic of marbles

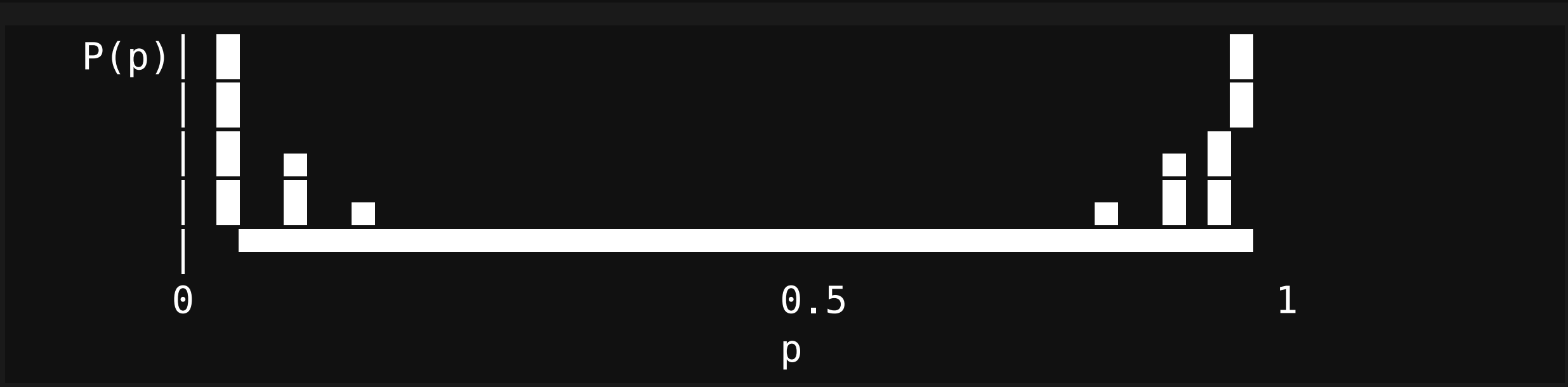
Tanaka finds bags of marbles in his parents’ attic. **Each bag is mostly one color** (white or black), but **overall the count is ~50/50**.

He wants to encode this in a Beta prior over θ = probability of drawing white. Which $\mathrm{Beta}(\alpha, \beta)$?

A. $\mathrm{Beta}(1, 1)$ — uniform, no info B. $\mathrm{Beta}(2, 2)$ — gentle center at 0.5 C. $\mathrm{Beta}(10, 10)$ — strong center at 0.5 D. $\mathrm{Beta}(0.5, 0.5)$ — U-shaped at the edges

Poll — answer

D. $\mathrm{Beta}(0.5, 0.5)$.



U-shaped: mass piles up near 0 and 1 (bag-level extremity), symmetric overall. $\mathrm{Beta}(2, 2)$ and $\mathrm{Beta}(10, 10)$ are *unimodal at 0.5* — they encode “around half white” *within a bag*, which is the opposite of what Tanaka saw.

Beta-Binomial — set up the pieces

We’re inferring the tonkatsu rate. Call it p.

Prior: $p \sim \mathrm{Beta}(\alpha, \beta) \;\rightarrow\; p(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$

Likelihood: k tonkatsus in n bentos $\rightarrow$

$p(k \mid p) \propto p^{k}(1-p)^{n-k}$

Beta-Binomial — multiply and read off

Posterior $\propto$ Prior $\times$ Likelihood:

$$p(p \mid k) \;\propto\; p^{\alpha-1}(1-p)^{\beta-1} \cdot p^{k}(1-p)^{n-k}$$

$$= \; p^{(\alpha+k)-1}(1-p)^{(\beta+n-k)-1}$$

Recognize this: it's $\mathrm{Beta}(\alpha + k,\ \beta + n - k)$.

Beta-Binomial conjugate update

Prior: $p \sim \mathrm{Beta}(\alpha, \beta)$

Data: k successes in n trials

Posterior: $p \mid k \sim \mathrm{Beta}(\alpha + k,\ \beta + n - k)$

Just add the counts. Successes bump α, failures bump β.

Worked example — Chibany's semester

Prior: $p \sim \mathrm{Beta}(7, 3)$ ← encodes "70/30 with low confidence"

Data: 27 tonkatsu in 40 ($k = 27$, $n = 40$)

Posterior: $p \mid D \sim \mathrm{Beta}(7 + 27,\ 3 + 13) = \mathrm{Beta}(34, 16)$

Posterior mean: $34/50 = 0.68$. Prior was 0.70.

Posterior is tighter: the data shrank the spread.

Break

Resume — Gaussian-Gaussian with N observations

Agenda so far: Beta-Binomial ✓

Now: what if Chibany weighs N bentos, not one?

Gaussian-Gaussian — notation lock-in

Symbol	What it is
μ_0, σ_0^2	Prior mean and variance of μ
σ^2	Data noise (known, fixed)
$D_1, \ldots, D_N$	N iid observations
$\sum_i D_i$	Sum over the N observations: $D_1 + D_2 + \cdots + D_N$
μ_N, σ_N^2	Posterior mean and variance of μ after seeing N data

Gaussian-Gaussian — precision is additive

Posterior precision:

$$\underbrace{\frac{1}{\sigma_N^2}}_{\text{posterior}} = \underbrace{\frac{1}{\sigma_0^2}}_{\text{prior}} + \underbrace{\frac{N}{\sigma^2}}_{N \text{ data}}$$

Precision = 1/variance. Each observation adds $1/\sigma^2$ units. N observations add N/σ^2.

Sanity check: at $N = 1$, this matches Week 2's single-observation case.

Gaussian-Gaussian — posterior mean

$$\mu_N = \sigma_N^2 \left(\underbrace{\frac{\mu_0}{\sigma_0^2}}_{\text{prior precision} \times \text{prior mean}} + \underbrace{\frac{\sum_i D_i}{\sigma^2}}_{\text{data precision} \times \text{data sum}} \right)$$

μ_N is a **precision-weighted average** of the prior mean and the data sum. Whoever has more precision wins.

Poll — Jamal's shortcut

While deriving a posterior over μ, Jamal notices the non-constant terms (w.r.t. μ) **have the form of a Gaussian**. He drops everything else and concludes the posterior is Gaussian with parameters read off the surviving form. Is he correct?

A. Yes — the dropped terms are absorbed into the normalization constant B. Yes — you can drop any term, even those involving μ C. No — he dropped some terms that involve μ D. Only if he later multiplies his answer by the dropped terms

Poll — answer

A. Yes — the dropped terms are part of the normalization constant.

A posterior is a probability density in μ. Anything not depending on μ is a multiplicative constant — absorbed into $Z = \int p(\mu \mid D)\, d\mu$.

Recognize the functional form → read off parameters → normalization handles itself.

Sequential updating — same rule, no new math

Observations arrive one at a time. Posterior after k observations becomes prior for observation $k+1$.

$$\mathrm{Beta}(34, 16) \xrightarrow[+1\ \text{hamb}]{\text{see 1 more}} \mathrm{Beta}(34, 17)$$

$$N(503.3, 16.6^2) \xrightarrow[D=498]{\text{see 1 more}} N(502.2, 14.5^2)$$

This is why conjugacy is useful in practice: online updates, no re-fit.

Three conjugate pairs, one pattern

Prior	Likelihood	Posterior
$\mathrm{Beta}(\alpha, \beta)$	Binomial(n, p)	$\mathrm{Beta}(\alpha + k, \beta + n - k)$
$\mathrm{Dirichlet}(\vec{\alpha})$	Multinomial$(n, \vec{p})$	$\mathrm{Dirichlet}(\vec{\alpha} + \vec{k})$
$N(\mu_0, \sigma_0^2)$	$N(\mu, \sigma^2)$	$N(\mu_N, \sigma_N^2)$

Row 2 is the multi-category generalization: $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\vec{k} = (k_1, \ldots, k_K)$. Same “add the counts” rule.

Subjective randomness — a Bayesian story

Why does HHTHTTTH look “more random” than HHHHHHHH?

Both sequences have the same probability under a fair coin: $(1/2)^8 = 1/256$.

So why do people consistently say the first is “more random”?

Griffiths & Tenenbaum (2001): people aren’t computing $P(x \mid \text{random})$. They’re computing a Bayes factor.

The reframe — likelihood ratio, not likelihood

Wrong question: “How likely is x under randomness?” $\rightarrow P(x \mid \text{random})$

Right question: “How much *more* does x favor randomness over a regular pattern?”

$$\text{subjective randomness}(x) \;=\; \log \frac{P(x \mid \text{random})}{P(x \mid \text{regular})}$$

A **likelihood ratio** — the same Bayes-factor object that appears in every conjugate update.

Defining “regular”

What competes with “random” depends on the stimulus.

Stimulus	“Regular” hypothesis
Binary sequence	Local representativeness: each subsequence has ~50/50 split
Numbers (pick 1–10)	Properties like “is prime,” “is even,” “ends in 7”
Spatial points	Clusters, lines, symmetry

The “regular” model is whatever your mind treats as the salient alternative — the implicit prior over what patterns exist.

Zenith radio data — the binary sequence experiment

1937 publicity stunt: Zenith broadcast 5 H/T sequences over the radio, asked listeners to “transmit” their guesses via ESP.

Result: people responded with sequences biased toward switching (away from runs). $P(\text{switch}) \approx 0.6$ instead of 0.5.

G&T fit this with their likelihood-ratio model: switches *raise* the posterior odds of “random.” Predicted bias falls out of the math, not added as a free parameter.

The takeaway for today's class

- The "mistake" in subjective randomness is **not** that people are bad at probability.
- They're computing $\log P(x \mid \text{random}) / P(x \mid \text{regular})$ — a perfectly Bayesian quantity.
- Today's conjugate-update mechanics ("drop the normalizer, recognize the form") are *exactly* the operation behind this model.

Open question: is human cognition Bayesian-by-default, or just Bayesian-when-tractable?

Close

Next week — Week 4 preview

Ira leads. **Hierarchical Bayes.**

Chibany's bento rate isn't the same across every semester, but semesters aren't totally independent either. How do we share information without collapsing?

Read T3 Ch 5 before class.