Bernoulli trials with Beta and Beta-mixture priors

Closed form, Monte Carlo integration and sampling importance resampling

Hedibert Freitas Lopes

Insper

October 2026

These slides were prepared with the assistance of an AI model (Claude, Anthropic), from the prompt reproduced on the next slides. All derivations and numbers were checked by running the accompanying code.

The prompt that generated these slides

```
Observed data: y=4 successes out of n=12 Bernoulli trials
Model: x1,...,xn | theta are iid Bernoulli(theta), for theta in (0,1)
Equivalent model representation:  y|theta ~ Binomial(n,theta)
Prior 1: theta ~ Beta(2,8)
Prior 2: theta ~ 0.7*Beta(2,8)+0.3*Beta(2,2)

Derive the following quantities for both mode/prior combinations:

a) Posterior density for theta
b) Posterior mean, median, standard deviation and 95 percent
   credibility and HPD intervals.
b) Prior predictive: Pr(y=k) for k=0,...,n
c) Posterior predictive:  Pr(x_{n+1}|y,n)
```

The prompt (cont.)

```
For prior 1, closed form derivation is available
For prior 2, use when necessary Monte Carlo integration based on
prior draws.

Finally, for model/prior 2, repeat the analysis replacing MC
integration with draws from the posterior obtained via sampling
importance resampling with proposal density a Beta(1.5,1.5).

Write the slides in latex/beamer as I have been doing lately so
I can share them with my students.
```

Lesson: the prompt states the data, the model, both priors, every quantity wanted, which tool to use for each prior, and the output format. Clear, specific prompts get clear, specific answers.

Roadmap

1. Model, data and the two priors
2. Prior 1: Beta$(2, 8)$ — everything in closed form
3. Prior 2: $0.7\,\mathrm{Beta}(2, 8) + 0.3\,\mathrm{Beta}(2, 2)$
   - exact answer (mixtures of conjugate priors are conjugate) — our benchmark
   - Monte Carlo integration with draws from the prior
   - Sampling importance resampling (SIR) with proposal $\mathrm{Beta}(1.5, 1.5)$
4. Comparison and take-aways
5. Appendix: code

Model and data

- $x_1, \ldots, x_n \mid \theta \overset{iid}{\sim} \mathrm{Bern}(\theta), \quad \theta \in (0,1)$.
- Sufficient statistic $y = \sum_i x_i$, so equivalently $y \mid \theta \sim \mathrm{Bin}(n, \theta)$:

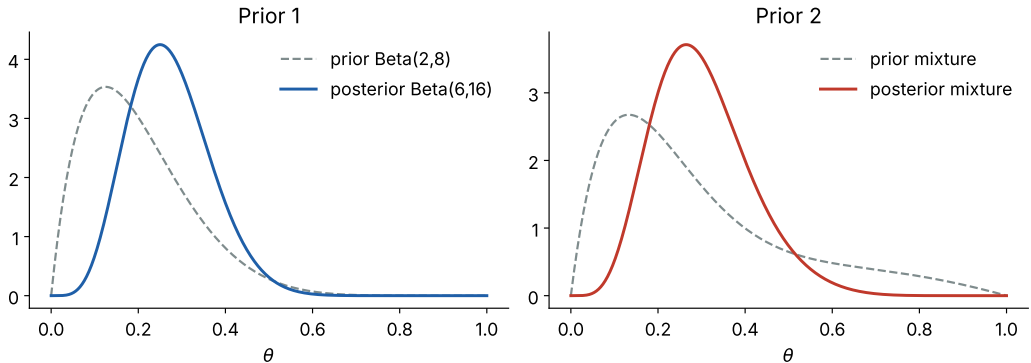
$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

- Data: $n = 12$, $y = 4$, MLE $\hat{\theta} = y/n = 1/3$.
- Priors

Prior 1: $\theta \sim \mathrm{Beta}(2, 8)$, Prior 2: $\theta \sim 0.7\,\mathrm{Beta}(2, 8) + 0.3\,\mathrm{Beta}(2, 2)$.

- Prior 1 has mean 0.2 and is fairly informative. Prior 2 keeps 70% of that belief and puts 30% on a vague component centred at 0.5: a heavier right tail.

Priors and posteriors at a glance



Prior 1 (a): posterior density

Beta prior $\times$ binomial likelihood:

$$p(\theta \mid y) \;\propto\; \theta^{a-1}(1-\theta)^{b-1}\,\theta^{y}(1-\theta)^{n-y} = \theta^{a+y-1}(1-\theta)^{b+n-y-1},$$

so

$$\theta \mid y \sim \text{Beta}(a+y,\; b+n-y) = \text{Beta}(2+4,\; 8+8) = \text{Beta}(6, 16).$$

$$p(\theta \mid y) = \frac{\theta^5(1-\theta)^{15}}{B(6,16)}, \qquad 0 < \theta < 1.$$

Posterior mean as a weighted average:

$$E(\theta \mid y) = \frac{a+b}{a+b+n}\,\frac{a}{a+b} + \frac{n}{a+b+n}\,\frac{y}{n} = \frac{10}{22}(0.2) + \frac{12}{22}\left(\tfrac{1}{3}\right) = \frac{6}{22}.$$

Prior 1 (b): posterior summaries

For $\theta \mid y \sim \mathrm{Beta}(A, B)$ with $A = 6,\ B = 16$:

$$E(\theta \mid y) = \frac{A}{A+B} = \frac{6}{22} = 0.2727$$

$$SD(\theta \mid y) = \sqrt{\frac{AB}{(A+B)^2(A+B+1)}} = \sqrt{\frac{96}{22^2 \cdot 23}} = 0.0929$$

$$\text{Median} = F^{-1}_{\mathrm{Beta}(6,16)}(0.5) = 0.2657 \quad \left(\approx \frac{A - 1/3}{A+B-2/3} = 0.2656\right)$$

$$\text{Mode} = \frac{A-1}{A+B-2} = 0.25$$

- 95% equal-tail interval: $\left(F^{-1}(0.025), F^{-1}(0.975)\right) = (0.1128,\ 0.4717)$.
- 95% HPD: $\{\theta : p(\theta \mid y) \geq c\}$ with mass 0.95; for a unimodal density solve $p(L \mid y) = p(U \mid y)$, $F(U) - F(L) = 0.95$: $(0.1012,\ 0.4557)$.
- HPD is shorter (0.354 vs 0.359) and shifted left: the posterior is right-skewed.

Prior 1 (c): prior predictive

$$\Pr(y=k) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \, \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \, d\theta = \binom{n}{k} \frac{B(a+k,\ b+n-k)}{B(a,b)},$$

the Beta-binomial(n, a, b) distribution, $k = 0, \ldots, 12$, with $a = 2$, $b = 8$.

- $E(y) = n \frac{a}{a+b} = 2.4$; $\quad V(y) = n \frac{ab}{(a+b)^2} \frac{a+b+n}{a+b+1} = 3.84$ (vs. 1.92 for $\text{Bin}(12, 0.2)$).
- The observed value: $\Pr(y = 4) = 0.1095$. This is the marginal likelihood (evidence) of model/prior 1.
- Table with all k on slide 18.

Prior 1 (d): posterior predictive

$$\Pr(x_{n+1} = 1 \mid y) = \int_0^1 \theta\, p(\theta \mid y)\, d\theta = E(\theta \mid y) = \frac{a+y}{a+b+n} = \frac{6}{22} = \frac{3}{11} = 0.2727,$$

$$\Pr(x_{n+1} \mid y) = \left(\tfrac{3}{11}\right)^{x_{n+1}} \left(\tfrac{8}{11}\right)^{1-x_{n+1}}, \qquad x_{n+1} \in \{0, 1\}.$$

More generally, for m new trials, $\tilde{y} \mid y \sim$ Beta-binomial$(m,\, 6,\, 16)$.

Prior 2: the exact answer (benchmark)

With $\pi(\theta) = \sum_{i=1}^{2} \pi_i \,\mathrm{Beta}(\theta \mid a_i, b_i)$, $\pi_1 = 0.7$, $\pi_2 = 0.3$:

$$p(\theta \mid y) = \sum_{i=1}^{2} w_i \,\mathrm{Beta}(\theta \mid a_i + y,\ b_i + n - y), \qquad w_i = \frac{\pi_i \, m_i(y)}{\sum_j \pi_j \, m_j(y)},$$

where $m_i(y) = \binom{n}{y} B(a_i + y, b_i + n - y)/B(a_i, b_i)$ is the Beta-binomial evidence of component i.

Component	prior weight	$m_i(4)$	posterior weight
$\mathrm{Beta}(2,8) \to \mathrm{Beta}(6,16)$	0.7	0.1095	0.7209
$\mathrm{Beta}(2,2) \to \mathrm{Beta}(6,10)$	0.3	0.0989	0.2791

- $p(y = 4) = 0.7(0.1095) + 0.3(0.0989) = 0.1063$.
- $E(\theta \mid y) = 0.7209 \cdot \frac{6}{22} + 0.2791 \cdot \frac{6}{16} = 0.3013$; median, quantiles and HPD by numerically inverting the mixture cdf.
- The data barely discriminate between components ($m_1 \approx m_2$), so weights hardly move.

Prior 2: Monte Carlo integration with prior draws

Draw $\theta^{(1)}, \ldots, \theta^{(M)} \sim \pi(\theta)$ (pick component with prob. $0.7/0.3$, then draw the Beta), $M = 100{,}000$.

- **Evidence:** $p(y) = E_\pi[p(y \mid \theta)] \approx \hat{p}(y) = \frac{1}{M} \sum_j p(y \mid \theta^{(j)})$.
- **Posterior density:** $\hat{p}(\theta \mid y) = \pi(\theta)\, p(y \mid \theta) / \hat{p}(y)$ (evaluate on a grid).
- **Posterior expectations:** with $w_j \propto p(y \mid \theta^{(j)})$, $\sum_j w_j = 1$,

$$E[g(\theta) \mid y] = \frac{E_\pi[g(\theta) p(y \mid \theta)]}{E_\pi[p(y \mid \theta)]} \approx \sum_j w_j\, g(\theta^{(j)}).$$

  Mean: $g(\theta) = \theta$; SD: also $g(\theta) = \theta^2$; median/quantiles: weighted quantiles.
- **HPD:** highest-density region of $\hat{p}(\theta \mid y)$ on the grid.
- **Prior predictive:** $\Pr(y = k) \approx \frac{1}{M} \sum_j \binom{n}{k} (\theta^{(j)})^k (1 - \theta^{(j)})^{n-k}$.
- **Posterior predictive:** $\Pr(x_{n+1} = 1 \mid y) \approx \sum_j w_j\, \theta^{(j)}$.

Prior 2: MC results

- $\hat{p}(y = 4) = 0.1063$ (exact 0.1063); $\int_0^1 \hat{p}(\theta \mid y)\, d\theta = 0.9996$.
- Effective sample size $\mathrm{ESS} = 1/\sum_j w_j^2 \approx 60,779$ out of $100{,}000$: the prior is a good proposal here because y/n sits in the bulk of the prior.

	exact	MC (prior draws)
Mean	0.3013	0.3019
Median	0.2896	0.2903
SD	0.1103	0.1102
95% equal-tail	(0.1197, 0.5479)	(0.1195, 0.5473)
95% HPD	(0.1028, 0.5219)	(0.1026, 0.5225)
$\Pr(x_{n+1} = 1 \mid y)$	0.3013	0.3019

Prior 2: sampling importance resampling

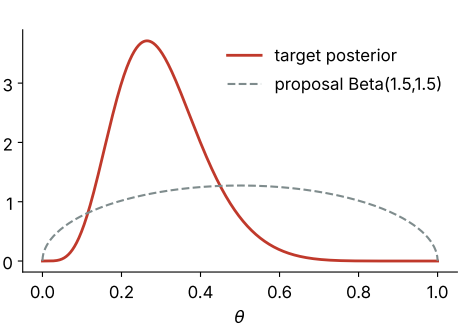
Proposal $q(\theta) = \mathrm{Beta}(1.5, 1.5)$; target $p(\theta \mid y) \propto \pi(\theta) p(y \mid \theta)$.

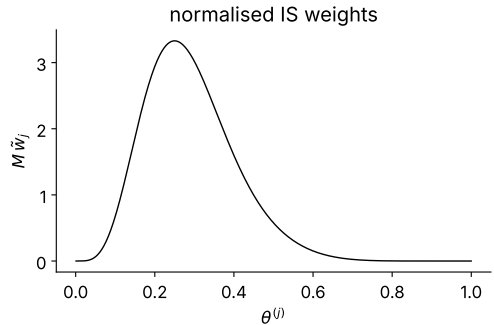
1. Draw $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(M)} \sim q, \quad M = 100{,}000.$
2. Weights $\omega_j = \dfrac{\pi(\tilde{\theta}^{(j)})\, p(y \mid \tilde{\theta}^{(j)})}{q(\tilde{\theta}^{(j)})}, \quad \tilde{w}_j = \omega_j / \sum_l \omega_l.$
3. Resample $\theta^{(1)}, \ldots, \theta^{(N)}$ from $\{\tilde{\theta}^{(j)}\}$ with probabilities $\{\tilde{w}_j\}$, $N = 20{,}000$.

Then $\{\theta^{(i)}\}$ is (approximately) a sample from $p(\theta \mid y)$:

- mean, SD, median, quantiles: sample versions; HPD: shortest interval containing 95% of the draws;
- $\Pr(x_{n+1} = 1 \mid y) \approx \frac{1}{N} \sum_i \theta^{(i)}$;
- prior predictive (needs no resampling): $\Pr(y = k) \approx \frac{1}{M} \sum_j \pi(\tilde{\theta}^{(j)}) \binom{n}{k} (\tilde{\theta}^{(j)})^k (1 - \tilde{\theta}^{(j)})^{n-k} / q(\tilde{\theta}^{(j)})$;
- $\hat{p}(y = 4) = \frac{1}{M} \sum_j \omega_j = 0.1065$; ESS $\approx 42{,}672$.

SIR: proposal versus target





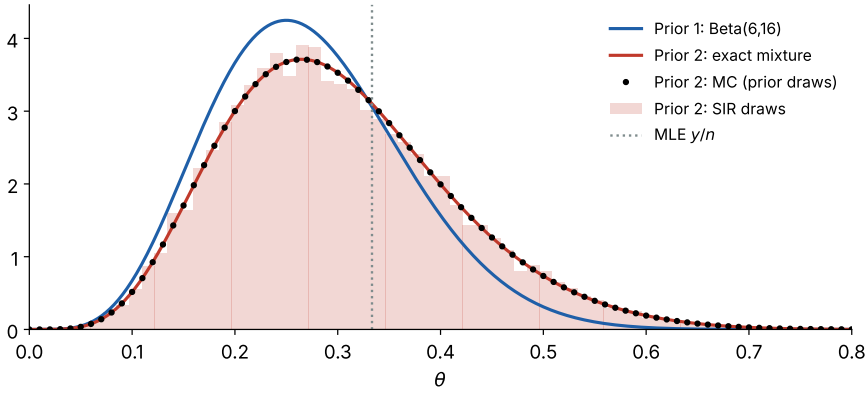
- q is flat-ish and has heavier tails than the posterior: weights are bounded, no explosion.
- The price: many draws land where the posterior is negligible, so ESS ($\approx$42,672) is lower than with the prior as proposal.

All results side by side

	Prior 1 exact	Prior 2 exact	Prior 2 MC (prior)	Prior 2 SIR
Mean	0.2727	0.3013	0.3019	0.3006
Median	0.2657	0.2896	0.2903	0.2870
Std. dev.	0.0929	0.1103	0.1102	0.1104
95% equal-tail	(0.1128, 0.4717)	(0.1197, 0.5479)	(0.1195, 0.5473)	(0.1192, 0.5463)
95% HPD	(0.1012, 0.4557)	(0.1028, 0.5219)	(0.1026, 0.5225)	(0.1080, 0.5267)
$\Pr(x_{n+1} = 1 \mid y)$	0.2727	0.3013	0.3019	0.3006
$p(y = 4)$	0.1095	0.1063	0.1063	0.1065

- Prior 2 shifts the posterior towards the MLE $1/3$ and widens it, mostly to the right.
- MC and SIR agree with the exact mixture to 2–3 decimals. The SIR HPD lower limit (0.108 vs 0.103) shows that the shortest-interval estimator is noisier than the grid-based one.
- Prior 1 has slightly larger evidence (0.1095 vs 0.1063): Bayes factor ≈ 1.03, no real preference.

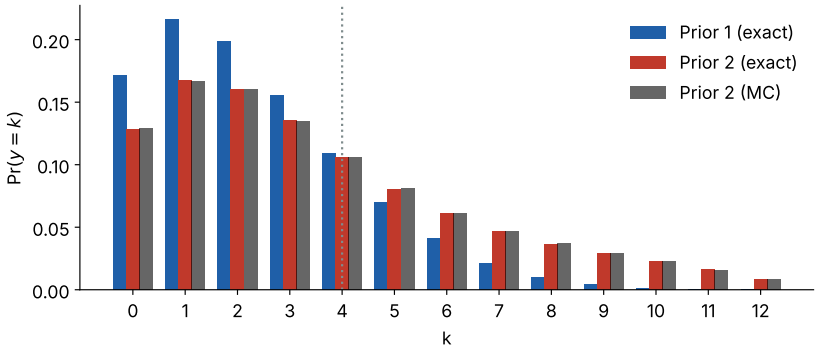
Posterior densities



Prior predictive $\Pr(y = k)$: table

k	Prior 1 exact	Prior 2 exact	Prior 2 MC (prior)	Prior 2 IS, $q = \mathrm{Be}(1.5, 1.5)$
0	0.1714	0.1286	0.1290	0.1298
1	0.2165	0.1674	0.1671	0.1691
2	0.1985	0.1607	0.1601	0.1620
3	0.1557	0.1354	0.1350	0.1360
4	0.1095	0.1063	0.1063	0.1065
5	0.0701	0.0807	0.0809	0.0807
6	0.0409	0.0609	0.0611	0.0610
7	0.0216	0.0467	0.0468	0.0468
8	0.0101	0.0367	0.0368	0.0368
9	0.0041	0.0292	0.0294	0.0292
10	0.0013	0.0227	0.0229	0.0225
11	0.0003	0.0161	0.0160	0.0159
12	0.0000	0.0086	0.0083	0.0085
Sum	1.0000	1.0000	1.0000	1.0048

Prior predictive $\Pr(y = k)$: picture



- Prior 2 puts far more mass on large k: $\Pr(y \geq 8) = 0.113$ vs 0.016 under prior 1.
- The IS column (SIR proposal) does not sum exactly to one: each k is estimated separately.

Take-aways

1. Conjugacy: Beta prior + binomial likelihood gives a Beta posterior, a Beta-binomial prior predictive and a Bernoulli posterior predictive with $p = E(\theta \mid y)$.
2. A finite mixture of conjugate priors is still conjugate: the posterior is a mixture of the updated components with weights $\propto$ prior weight $\times$ component evidence.
3. Monte Carlo with prior draws is importance sampling with the prior as proposal; weights are the likelihood. It works well when the likelihood is not too concentrated relative to the prior.
4. SIR turns weighted draws into an unweighted posterior sample; the ESS measures how well the proposal matches the target.
5. Always compare a numerical method with an exact answer when one exists.

Appendix: Monte Carlo with prior draws (Python)

```
import numpy as np; from scipy import stats
n, y, M = 12, 4, 100_000
rng = np.random.default_rng(20261029)
comp = rng.random(M) < 0.7 # mixture component
th = np.where(comp, rng.beta(2, 8, M), rng.beta(2, 2, M)) # prior draws
L = stats.binom.pmf(y, n, th) # likelihood
py = L.mean() # p(y)
w = L / L.sum() # weights
mean = np.sum(w*th); sd = np.sqrt(np.sum(w*th**2) - mean**2)
prior_pred = [stats.binom.pmf(k, n, th).mean() for k in range(n+1)]
post_pred = mean # Pr(x_{n+1}=1|y)
```

Appendix: SIR with $q = \mathrm{Beta}(1.5, 1.5)$ (Python)

```
def prior(t): return 0.7*stats.beta.pdf(t,2,8) + 0.3*stats.beta.pdf(t,2,2)
tq = rng.beta(1.5, 1.5, M)
om = prior(tq) * stats.binom.pmf(y, n, tq) / stats.beta.pdf(tq, 1.5, 1.5)
wt = om / om.sum(); ess = 1/np.sum(wt**2); py = om.mean()
ths = tq[rng.choice(M, size=20_000, replace=True, p=wt)] # posterior draws

def hpd(x, p=0.95): # shortest interval
    x = np.sort(x); m = int(p*len(x))
    j = np.argmin(x[m:] - x[:len(x)-m]); return x[j], x[j+m]

summ = ths.mean(), np.median(ths), ths.std(), np.quantile(ths,[.025,.975]), hpd(ths)
```

Full script (figures and tables included): `bernoulli_beta.py`.