

Bayesian Inference: Ten Coins and a Stamped Box

Weifeng Zeng

A Bayesian answer is a density, and the whole of it is one product divided by one integral. This note keeps to a box of coins whose label names the density the box draws its coins from, and asks six questions about what that label does to the arithmetic: which single number the ten posteriors share and what it does to the spread of the answers; what one coin can and cannot say about the box it came from; why the answer reduces to adding the label's own counts to each coin's tally, and what makes the order of a record invisible; where the label's answer beats a coin's own frequency and where it must lose; what the label prices for a coin nobody has flipped yet; and what this arithmetic cannot see when the box holds two kinds of coin, or when a wider label would have been the better bet. Every question is posed against a box whose arithmetic is checkable by hand, and every number quoted is printed by one script beside the note.

October 12, 2026

Contents

1. A label on a box, and the number every answer shares	3
2. What one coin can say about its box, and what ten can	5
3. Why the answer is “add the label’s counts”	7
4. What the label buys, and what it costs at the edges	8
5. The price of the next flip, for a coin and for a stranger	10
6. What this arithmetic does not see	11
7. Sources and further reading	13

1. A label on a box, and the number every answer shares

Question: The box's label says that each coin's bias is drawn from $\text{Beta}(3, 5)$. Ten coins have been flipped ten times each: 0, 1, 2, 3, 3, 4, 5, 5, 6, 8 heads. Table 1 lists the ten posteriors the label produces. Predict four things about them. First, one number appears in all ten posterior means and in none of the ten records: name it and say what it counts. Second, predict the ratio between the gap separating the 0-head and 8-head answers and the gap their two recorded frequencies already have, and then say why **any** other pair of coins in this box gives the same ratio. Third, the label is replaced by $\text{Beta}(12, 20)$ – the same mean, four times the strength: predict the ten means again and the new gap between the two extreme coins. Fourth, the coin that came up with no heads at all: say what the label lets you claim about it that its own frequency of zero does not.

The box. A coin's bias θ is the probability that it lands heads, and the label says that each coin's θ is drawn independently from

$$p(\theta) = \theta^{3-1}(1 - \theta)^{5-1} / B(3, 5), \quad (1)$$

in which the normalizing constant is the Beta function,

$$B(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}. \quad (2)$$

so the label names a mean of $3/8 = 0.375$ and a standard deviation of $\sqrt{3 \cdot 5 / (8^2 \cdot 9)} = 0.1614$. Nothing else about a coin is written down. Ten coins came out of the box and were flipped ten times each; the counts of heads were 0, 1, 2, 3, 3, 4, 5, 5, 6, 8, or 37 heads in 100 flips. Two answers to “how biased is coin seven?” can be read off the records alone: that coin's own frequency, $6/10 = 0.6$, and, if you are willing to pretend that all ten coins are the same coin, the pooled frequency $37/100 = 0.37$.

The ten answers. A coin flipped ten times with h heads has likelihood $p(h | \theta) = \binom{10}{h} \theta^h (1 - \theta)^{10-h}$, and Bayes's rule divides the label times that likelihood by its own integral,

$$p(\theta | h) = p(h | \theta)p(\theta) / \int_0^1 p(h | t)p(t) dt. \quad (3)$$

The division is the only hard step in this note, and Equation 1 is what makes it easy: a power of θ times a power of $1 - \theta$ is a Beta density carrying its own normalizer. Table 1 gives the result for each of the ten coins.

HEADS h	COINS	POSTERIOR	MEAN	STANDARD DEVIATION
0	1	Beta(3, 15)	0.1667	0.0855
1	1	Beta(4, 14)	0.2222	0.0954
2	1	Beta(5, 13)	0.2778	0.1028
3	2	Beta(6, 12)	0.3333	0.1081
4	1	Beta(7, 11)	0.3889	0.1118
5	2	Beta(8, 10)	0.4444	0.1140
6	1	Beta(9, 9)	0.5000	0.1147
8	1	Beta(11, 7)	0.6111	0.1118

Table 1: The ten posteriors the label produces, one per coin, in the order the coins were recorded. Every entry is an exact consequence of Equation 1 and that coin’s likelihood, reproduced by *scripts/pooling_checks.py*. Inspect the third column for the two numbers that move with the record, and the last column for what the label’s arithmetic does to a coin’s own uncertainty.

Facts to reason with.

- The label is worth $\alpha + \beta = 8$ flips. Its mean, 0.375, sits 0.005 from the pooled frequency 0.37 – a tenth of the label’s own standard deviation, so the ten coins’ total cannot tell those two numbers apart.
- The ten records are independent given the box’s parameters: no coin’s likelihood mentions another coin, and the label mentions no data.
- A posterior is a density, not a number. Only its integrals over intervals are probabilities, and a single value of θ has none.
- The only integral this note works out by hand is Equation 2: a power of t times a power of $1 - t$ over $(0, 1)$. Everything after that is multiplication, division, or quadrature by script.
- The label’s standard deviation is 0.1614 while the ten frequencies run from 0 to 0.8, so most of the box’s coins fall outside the label’s own one-standard-deviation band.

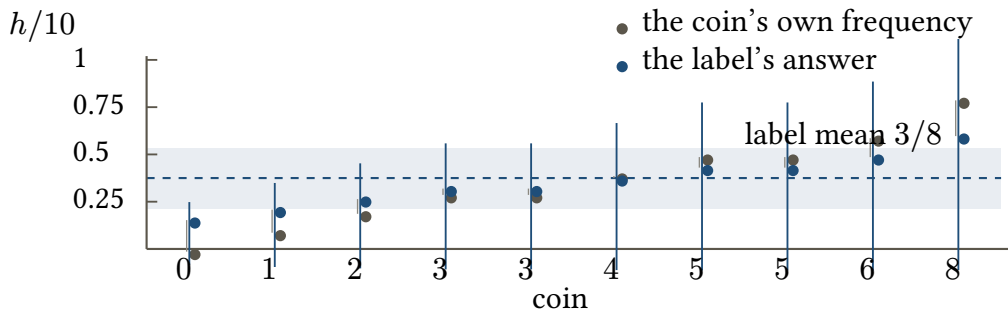


Figure 1: The ten coins, sorted by their count. Gray marks are each coin’s own frequency $h/10$ with a thin segment running to the label’s answer for that coin; navy marks are the ten posterior means of table 1, one standard deviation either side. The dashed rule is the label’s mean, $3/8$, and the band behind it is the label’s own one-standard-deviation interval, $(0.214, 0.536)$. Inspect which of the two zigzags the band contains, and how much of the taller one’s swing survives in the shorter one.

2. What one coin can say about its box, and what ten can

Question: The label is gone, and the same ten counts are all you have. You put a hyperprior on the box's two unknown parameters: α and β independently uniform on the log scale over $[0.1, 40]$, so that the density is proportional to $1/(\alpha\beta)$ inside that box and zero outside. Figure 2 shows what the counts say about $\alpha + \beta$, for one coin and for ten. Predict four things. First, whether a single coin narrows the box's width at all – and if not, what it does narrow. Second, which of the two things the ten coins pin down better, where the population sits or how wide it is, and by roughly what factor each 68 percent interval narrows against the hyperprior's own. Third, whether the label $\text{Beta}(3, 5)$ survives the ten coins' verdict, and how far the ten coins' own answer for a coin moves from table 1's answer for it. Fourth, whether dropping the label makes each coin's answer narrower or wider – and by how much at the two ends of the table.

The hyperprior, written out. With α and β confined to $[0.1, 40]$ and uniform on the log scale inside it, the density of the hyperprior is

$$\pi(\alpha, \beta) = 1/(\alpha\beta \cdot \log^2(400)) \quad \text{for } 0.1 \leq \alpha, \beta \leq 40, \quad (4)$$

and zero outside. The box's own prior, before any count is seen, gives $\alpha/(\alpha + \beta)$ a 68 percent interval of $[0.068, 0.932]$ and $\log(\alpha + \beta)$ one of $[0.39, 3.34]$.

A count, as a function of the box. A count of h heads in ten flips has, under the box's parameters, the probability

$$p(h \mid \alpha, \beta) = \int_0^1 p(h \mid t)p(t) dt, \quad (5)$$

a Beta-binomial probability with no other ingredients: the coin's own likelihood and the label's own density, integrated against each other. The ten coins' hyperposterior is Equation 5 multiplied over the ten counts and then multiplied by Equation 4.

Facts to reason with.

- The ten coins' hyperposterior has its mode at $(\alpha, \beta) = (2.21, 3.76)$, a population $\text{Beta}(2.21, 3.76)$ with mean 0.3709; its 68 percent set is α between 0.75 and 10.39 and β between 1.26 and 17.99, and inside that set $\alpha + \beta$ runs from 2.08 to 28.38.
- The coin with three heads, taken alone, puts the mode of (α, β) on the corner of the hyperprior's box, at $(17.70, 40.00)$: a population whose width has shrunk to the smallest the box allows. Its 68 percent set touches all four edges of the box.
- The probability of that coin's own record – three heads in ten – is 0.2461 at the corner population, 0.1697 under the label, and 0.0909 under a uniform population.
- The posterior mass with $\alpha + \beta$ above 20 is 0.114 for the ten coins and 0.278 for the single coin.
- The hyperprior's own 68 percent interval for $\alpha/(\alpha + \beta)$ has width 0.863; the ten coins' has width 0.142. For $\log(\alpha + \beta)$ the same two widths are 2.955 and 1.671.
- The box is a choice, and the ten counts do not settle it: the mode of (α, β) moves only to $(2.19, 3.74)$ as the box becomes $[0.2, 20]$ or $[0.05, 80]$, while the posterior mean of $\alpha + \beta$

moves from 7.96 to 9.85 to 12.25 and the same ten counts read as evidence for the three boxes differ by 0.86 nats.

DATA	MODE OF (α, β)	68 PERCENT SET	$\alpha/(\alpha + \beta)$ WIDTH	$\log(\alpha + \beta)$ WIDTH
hyperprior only	—	all of the box	0.863	2.955
the 3-head coin	(17.70, 40.00), a box corner	α and β to all four edges	0.414	2.889
all ten coins	(2.21, 3.76)	$\alpha + \beta$ in [2.08, 28.38]	0.142	1.671

Table 2: The same data, three times: the hyperprior alone, one coin’s count, and all ten. Widths are 68 percent intervals of the two functionals named. For the one-coin row the 68 percent set reaches the hyperprior’s boundary, so its $\alpha + \beta$ entry is a range and not an interval around a mode. Inspect the two width columns against the same two columns in the first row.

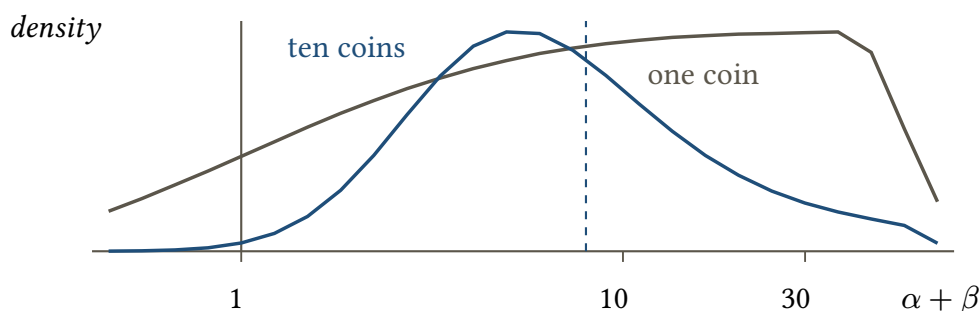


Figure 2: What the counts say about the label’s strength, $\alpha + \beta$, on a log axis; each curve is scaled to its own peak. Navy: all ten counts. Gray: the single coin that came up heads three times. The dashed rule marks $\alpha + \beta = 8$, the label’s own strength. Inspect which curve has an interior peak and which one keeps rising until the hyperprior’s box stops it.

What the ten coins do not settle. Equation 4 is a choice, not a fact. The ten coins’ own agreement about the population’s width never stops improving as the population gets tighter inside the box, and only the box’s own edge decides how far that goes – which is why the mean of $\alpha + \beta$, not the mode of (α, β) , is the entry that moves when the box does.

3. Why the answer is “add the label’s counts”

Question: Look at the third column of table 1: the posterior of a coin with h heads is $\text{Beta}(3 + h, 15 - h)$, and the pair $(3 + h, 15 - h)$ is the label’s own pair $(3, 5)$ with the coin’s $(h, 10 - h)$ added to it. Predict three consequences. First, the ten flips of a coin can equally well arrive as one batch of ten, as two batches of five, or one at a time: say which of these can give a different posterior, and name the two numbers of the record that a posterior can see. Second, a coin that shows three heads can have its record written in $\binom{10}{3} = 120$ orders: say how many distinct posteriors those orders produce, and what that forces about what a likelihood can mean. Third, the label’s density carries its own normalizer $1/B(3, 5)$ and the posterior carries $1/B(3 + h, 15 - h)$: say which of the two decides the **shape** of a belief and which decides the **probability of a record**, and then give that probability for the coin that showed no heads.

The product. Bayes’s rule Equation 3 is a product and a division. The product is

$$p(h \mid \theta)p(\theta) \propto \theta^h (1 - \theta)^{10-h} \cdot \theta^{3-1} (1 - \theta)^{5-1} = \theta^{3+h-1} (1 - \theta)^{5+10-h-1}, \quad (6)$$

in which the exponents of θ add and the exponents of $1 - \theta$ add, because both factors are powers of the same two bases. Multiplying two powers is the whole mechanism, and the numbers it adds are the label’s 3 and 5.

Two routes to one posterior. Because the product is again a Beta kernel, one route is to name it: the posterior is the Beta density whose two parameters are the sums, with normalizer $B(3 + h, 15 - h)$. The other route is to feed the flips in one at a time: take the label as the prior for flip one, feed that flip in, take the posterior as the prior for flip two, and so on ten times. Both routes are one multiplication with its factors in a different order, which is what makes the batching question answerable without any computation.

Facts to reason with.

- The pair $(3 + h, 15 - h)$ sums to 18 for every coin in the box, and 18 is the label’s 8 plus the coin’s 10.
- A Beta density with both parameters above one has one interior peak and no atoms: the posterior gives every interval inside $(0, 1)$ positive probability, including intervals near the two ends.
- The label places no position in a record above another: it draws each coin’s θ once and then makes the ten flips independent with that one θ , so no flip is special to the model.
- The probability of a record is Equation 5 with the label substituted for the box’s parameters; the normalizer of the product Equation 6 is what makes it a number.
- The record “no heads in ten flips” is a single record, not two: $h = 0$ leaves the heads count no room to vary.

4. What the label buys, and what it costs at the edges

Question: For a coin with h heads the label's answer is an affine function of the coin's own frequency: $(3 + h)/18 = (5/9)(h/10) + (4/9)(3/8)$. Two risk curves follow, one for the frequency $h/10$ and one for the label's answer, both defined through Equation 7 and drawn in figure 3. Predict four things. First, where each curve reaches its maximum, and which of the two maxima is larger. Second, which of the two estimators has the smaller **average** risk when the coin's true bias really is drawn from the label, and by what factor. Third, which of the two has the smaller **worst** case over biases – and say what the label's answer costs you at $\theta = 1$. Fourth, which of your answers so far is a property of this particular label and which is a property of the label's method, and what that distinction says about a label nobody has checked.

Risk. Score an estimator $\hat{\theta}$ of the bias by its expected squared error at the true bias,

$$\text{risk}(\theta) = \mathbb{E}_{h \sim \text{Binomial}(10, \theta)} (\hat{\theta}(h) - \theta)^2, \quad (7)$$

so that a smaller risk means a better answer in the true bias's own units. The frequency $h/10$ has risk $\theta(1 - \theta)/10$, maximized at $\theta = 1/2$ with value $0.025 = 1/40$, and falling to zero at both ends: a coin that never shows heads is reported as having bias zero, and by this score that report is perfect if the coin really is that coin. The label's answer has

$$\text{risk}_{\text{label}(\theta)} = (25/81) \cdot \theta(1 - \theta)/10 + (16/81) \cdot (\theta - 3/8)^2, \quad (8)$$

in which the first term is the frequency's own sampling variance shrunk by the square of the weight $5/9$, and the second is the square of the pull toward the label's mean. The two curves are equal exactly at the roots of $432\theta^2 - 352\theta + 45 = 0$, at $\theta = 0.1588$ and $\theta = 0.6560$.

Facts to reason with.

- The label's risk at its own mean is 0.0072338 ; at $\theta = 0$ it is $1/36 = 0.02778$; at $\theta = 1$ it is $25/324 = 0.07716$.
- Averaged over $\theta \sim \text{Beta}(3, 5)$, the frequency's risk is $1/48 = 0.020833$ and the label's is $5/432 = 0.011574$.
- The label's answer is the function of h that minimizes posterior expected squared error under the label, for all ten coins of the box at once.
- The label's mean is $3/8 = 0.375$; the pooled frequency of the ten coins is 0.37 ; the two differ by less than a tenth of the label's own standard deviation.
- The frequency's risk vanishes where the label's does not: the frequency reports a coin that never shows heads as having bias zero, and reports it with no uncertainty at all.

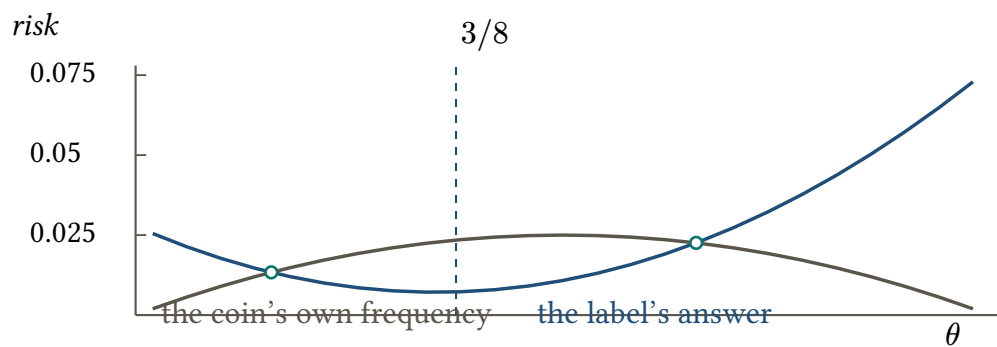


Figure 3: The two risk curves of Equation 7 and Equation 8, as functions of the coin's true bias. Gray: the frequency $h/10$. Navy: the label's answer $(3 + h)/18$. The dashed vertical rule is the label's mean, and the two open marks are the biases at which the curves are equal. Inspect the middle of the interval against its two ends.

What the two curves say. The label's answer is lower in the middle and higher at the ends: it is the best answer available **on average over the label**, and at a single bias it can be several times the frequency's error. Nothing in the arithmetic of section 3 chose the label's mean $3/8$; that choice is the centre of the second term of Equation 8, and the risk curve is its price list. Section 6 runs the same comparison where the label is not merely unchecked but wrong.

5. The price of the next flip, for a coin and for a stranger

Question: Three prices are wanted from the same box: the next flip of the coin that has shown 8 heads in its ten flips; the next flip of the coin that has shown none; and the first flip of a fresh coin nobody has flipped. Two of them follow from table 1 if a price is a posterior mean, so predict the third and say why it is the label's own mean, untouched by the ten records. Then say what the ten coins' own verdict from section 2 does to all three. Next, check the price rule: do the ten successive prices of the 0-head coin multiply out to the probability of that record from section 3? Last, table 3 gives a fresh coin's first ten flips under three routes – the label, the ten coins' mixture over labels, and a plug-in at the single most likely label. Say which of the three is the answer, and then predict the same three routes for a box that has produced one counted coin, three heads in ten, saying which row's tail moves most and why.

The price of one flip. A price that is to be coherent must equal the probability of the event given the record. Appending a head to a record with h heads turns that record's probability Equation 5 into the probability of the longer record, and the ratio of the two is what a head should cost; the same ratio computed from Equation 3 is the posterior mean of θ . For a fresh coin the same argument runs on the label alone:

$$P(H \mid \text{record}) = \int_0^1 \theta p(\theta \mid \text{record}) d\theta, \quad P(H \mid \text{fresh}) = \int_0^1 \theta p(\theta) d\theta. \quad (9)$$

Facts to reason with.

- The ten coins' hyperposterior puts 0.3709 of expectation on $\alpha/(\alpha + \beta)$, and the label's own mean is 0.375: a stranger is priced by the population, not by any recorded coin.
- A fresh coin's bias under that hyperposterior has variance 0.035464 from the population's own spread plus 0.005438 from the population's unknown mean; a recorded coin's answer carries no such second term, because its flips are in hand.
- The one-coin box's hyperposterior has no interior mode to plug in: its 68 percent set reaches every edge of the box.

ROUTE	MEAN	STD. DEV.	$P(h \leq 1)$	$P(h \geq 7)$	$P(h = 5)$
the label, Beta(3, 5)	3.7500	2.1651	0.1618	0.1170	0.1361
ten coins, mixture	3.7088	2.4525	0.2097	0.1437	0.1194
ten coins, plug-in	3.6992	2.3105	0.1933	0.1317	0.1246

Table 3: A fresh coin's first ten flips under three routes from the same ten counts. The plug-in uses the modal population Beta(2.21, 3.76); the mixture averages over the whole hyperposterior of section 2, and is what an unknown label forces a mixture of Beta-binomials to be.

Inspect the two ends of the support against the middle.

What an unknown label does to a stranger. The second integral of Equation 9 is replaced by its own average over the hyperposterior of section 2, and the fresh coin's count becomes a mixture rather than a single Beta-binomial. A mixture of Beta-binomials is not a Beta-binomial: it is wider, and table 3 is what that width buys.

6. What this arithmetic does not see

Question: A second box was stuffed by hand to hold two kinds of coin: five biased at 0.15 and five at 0.85, each flipped ten times, and its counts came out 0, 1, 1, 2, 2, 8, 8, 9, 9, 10. The same one-shape population now fits $\text{Beta}(0.593, 0.593)$. Predict three things. First, where the fit's answer for a fresh coin puts mass that the truth puts almost nowhere, and the ratio of the two at five heads. Second, table 4 gives five models of the same counts on both boxes: predict which model wins each column, and predict which two rows of the table describe **exactly** the same distribution of the ten counts even though they are written differently. Third, the population with a hyperprior loses to a model with neither a hyperparameter nor a prior over it on the first box: say why, and say what the second box shows that the hyperprior cannot buy. Last, say what none of this arithmetic could have told you about the second box.

The two boxes. The first box is section 1's: ten coins drawn from the label, counts 0, 1, 2, 3, 3, 4, 5, 5, 6, 8. The second was stuffed by hand and has no density to draw its coins from – it has two atoms, at 0.15 and 0.85 – so a one-shape population is now a description of the wrong kind of object, and the question is what such a description does when it is fitted anyway.

Evidence. The probability of a whole set of counts under a model is that model's own average of Equation 5 over whatever it leaves unknown, and two models are compared by the ratio of those averages on the same data. The numbers below are nats, so one nat is a factor of e .

Facts to reason with.

- The second box's fit is $\text{Beta}(0.593, 0.593)$: one hump, mean 0.5, standard deviation 0.338, and a 68 percent set with α and β between 0.26 and 1.28. The hump's mass is pushed toward the two ends of $(0, 1)$, which is the only way a single Beta can imitate two clumps.
- A fresh coin from that box really has bias 0.15 or 0.85 with equal probability, which puts 0.0085 of its first ten flips on exactly five heads and 0.0984 on no heads at all.
- The two ends of the second box are answered very differently by the two models: the coin with no heads gets 0.0553 from the fit and 0.1667 from a label; the coin with ten heads gets 0.9447 from the fit and 0.7222 from a label.
- The free-coin model scores -23.9790 nats with a uniform prior on each coin, -22.0635 with a $\text{Beta}(3, 5)$ prior, and -22.6060 with a $\text{Beta}(6, 10)$ prior: what such a model bets on is most of what it scores.
- The population model's evidence is an average over its own box, so the box's width is part of the score: on the first box it reads -23.9767 for $[0.9, 1.1]$, -23.9456 for $[0.5, 2]$, -24.6300 for $[0.2, 20]$, -25.1024 for $[0.1, 40]$ and -25.4883 for $[0.05, 80]$. The uniform shape on its own, with no box to average over, is -23.9790 .

MODEL OF THE TEN COUNTS	TEN COINS	TWO KINDS
every coin free, uniform prior	−23.9790	−23.9790
every coin free, Beta(3, 5) prior	− 22.0635	−32.0592
one shared bias for the box	−26.7706	−46.9620
population Beta(α, β), unknown	−25.1024	−26.6619
population of two unknown points	−24.6753	− 22.8268

Table 4: Log probabilities of the same two sets of ten counts under five models, in nats; larger is better and the best model in each column is marked. “Free” means every coin’s bias has its own distribution and no two coins share anything. Inspect which model wins each column, and which two rows are related in a way their descriptions do not suggest.

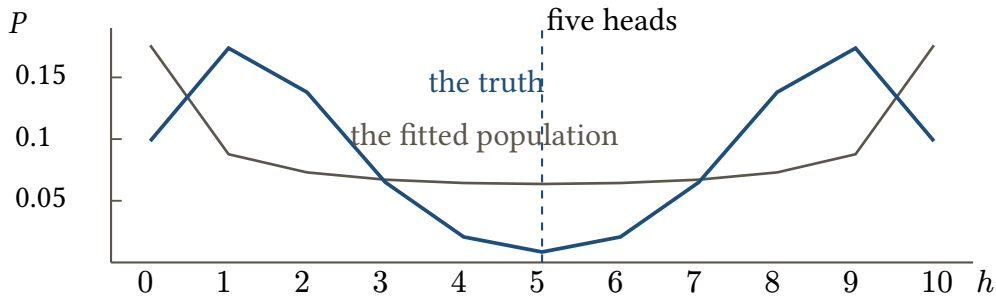


Figure 4: The first ten flips of a fresh coin from the hand-stuffed box. Navy: the truth, an even mixture of 0.15 and 0.85. Gray: the fitted one-shape population Beta(0.593, 0.593), whose shape has bent toward the two ends because that is the only way one hump can imitate two. Inspect the middle of the support against the two humps.

7. Sources and further reading

No external source has been consulted for this note, and it makes no claim about prior work, attribution, or the history of any method. Every number above is either an arithmetic consequence of the label in Equation 1 and the ten counts, or a quadrature of a density those two define; *scripts/pooling_checks.py* prints all of them, and recomputes each identity by a second route that shares no code with the first. The statements here that carry names elsewhere – that a posterior mean minimizes posterior expected squared error, that a labelled population and a prior on each coin are the same statement, that one-step prices are posterior means and therefore a martingale – are proved in this note only in the finite, checkable form the two boxes allow. The expected answers to the six questions are kept in *summary.md* beside this file, not in it.