# A brief introduction to Bayesian statistics through coin-flipping

Berk A. Alpay[1]

**Abstract:** Statistics often seems opaque to beginners but its complexities flower from fundamental concepts that become intuitive when studied through example. Here I explain as much of Bayesian statistics as I can through one example: flipping a coin to infer its fairness. I include snippets of statistical computing and end with a glimpse of more extensive modeling.

**Keywords:** probability, inference, statistical computing, probabilistic programming

You flip a coin $n$ times and $y$ of them land heads: how fair is the coin? To be more precise, assume the coin has some intrinsic probability $\theta$ (between 0 and 1) to land heads and call this probability the coin's fairness. Given the data from this **random experiment**, what is $\theta$? It's a deep question and I hope my best attempt to answer it will give the reader an intuition for Bayesian statistics as it has for me since I first worked through it.

Although the beta-binomial model, which I'll spend this article describing, is often used as an example in Bayesian statistics courses, such courses are often not taught or required at the undergraduate level and often presuppose experience in probability and statistics [1]. I'll instead encapsulate core concepts of introductory probability and Bayesian statistics in an accessible example. I aim to be concise and recommend the curious reader to refer to the first few chapters of introductory textbooks, e.g. [2, 3], for more context.

## Randomness in a head count: the likelihood

Introductory probability typically begins by defining a **sample space**, the set of possible **outcomes** of an experiment. If we were to flip the coin a single time, the sample space would be $\{\mathrm{T}, \mathrm{H}\}$, T meaning the coin lands tails and H that it lands heads. Flipping it twice, the sample space would be $S = \{\mathrm{TT}, \mathrm{TH}, \mathrm{HT}, \mathrm{HH}\}$.

Even if the coin were fair, it's quite probable the coin wouldn't land evenly heads and tails. We can split the sample space of two flips into subsets of the sample space, or **events**. Two such events are $A = \{\mathrm{TH}, \mathrm{HT}\}$, the event that the flips fall evenly, and the event $B = \{\mathrm{TT}, \mathrm{HH}\}$ that they don't. Together these two events make up the whole sample space $S$ though other events can be defined as well. They may overlap in outcomes (such as the event that exactly one tails is flipped and the event that exactly one heads is flipped) or even fill up the whole space — the event that any combination of heads and tails are flipped. Let's do a "naïve" [2] calculation of the **probability** (technically a function of an event that returns a number between 0 and 1 and satisfies certain axioms) that both flips land

[1]Department of Systems Biology, Harvard Medical School (berk_alpay@g.harvard.edu)

evenly by comparing the sizes $|A|$ and $|B|$ of the events: $p(A) = |A|/|S| = \frac{|A|}{|A|+|B|} = 1/2$. The calculation is naïve because it assumes all outcomes are equally likely and therefore that the coin is fair. Note that even for a fair coin the probability that two flips fall evenly is only 1/2.

What's that probability if we flip the coin ten times? There being two possible outcomes of each of ten flips, the sample space would now have $2^{10}$ outcomes, a much bigger set than with two flips. It'd be tedious now to calculate the probability of the coin falling evenly by tallying the suitable outcomes one by one. It would pay to be more clever. There are $\binom{10}{5}$ outcomes in which the coin falls evenly: all the different ways to choose exactly five coins to fall heads from the ten available. This $\binom{10}{5}$ is called a binomial coefficient and is computed in general as $\binom{n}{k} = n!/(k!(n-k)!)$. There are $2^{10}$ possible outcomes, so the probability that the coin falls evenly over ten flips is $\binom{10}{5}/2^{10} \approx 1/4$, which is even less than with two flips.

So there's quite a large probability that even a fair coin doesn't fall evenly. But then what's the probability that it falls, say, almost evenly? Or that it falls entirely heads or tails? Here, $y$ is a real-number representation of the outcome of the experiment and is therefore a **random variable**, representing the number of flips that land heads. To understand the randomness in the number of heads $y$, it'd be ideal to know the relative probability of each possible number of heads, what's known as the **probability distribution** of $y$.

It's getting harder to count outcomes corresponding to each event, so let's be even cleverer now and find a general formula that provides the probability of each possible number of heads. Simultaneously, this formula will also handle coins that aren't fair. The actual fairness is uncertain but let's condition on the fairness being some given $\theta$, meaning we'll take for granted that the fairness is a certain known $\theta$. The way we express this **conditional probability** distribution of $y$ given $\theta$ is as $p(y|\theta)$, which is shorthand notation for the probability of each possible $p(y = k|\theta)$. The formula to compute those probabilities is

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y},$$

which can be derived by counting as follows. For exactly $y$ flips to land heads, the coin must land heads $y$ times and all other flips must land tails $n - y$ times. Since these flips are independent, the **joint probability** of a particular sequence of $y$ heads and $n - y$ tails is simply the product of the individual probabilities of each flip, giving $\theta^y(1-\theta)^{n-y}$. But there can be several ways in which exactly $y$ heads can land; the first $y$ flips might land heads and the rest tails, or vice-versa, to name two. The binomial coefficient $\binom{n}{y}$ counts all these ways.

This formula fully describes the distribution of a random variable and so it is called its **probability function**. In this example this distribution represents the relative probabilities of different numbers of heads. But more generally, we're counting the number of times something happens **independently** over a certain number of trials, with a given constant probability of it happening each time; flipping a heads doesn't change the probability that the next flip is a tails. You could use the same kind of distribution to describe, for example, the probability that a die rolls a five, with $\theta = 1/6$ for a fair die. This distribution is so generally useful that it has a name: the binomial distribution. We can say that $y|\theta$ is distributed as a binomial random variable with parameters $n$ and $\theta$, which is written in notation as $y|\theta \sim \text{Binomial}(n, \theta)$. Let's plot it for $n = 10$ and $\theta = 1/2$, the distribution of the number of times a fair coin tossed ten times lands heads (Figure 1A).

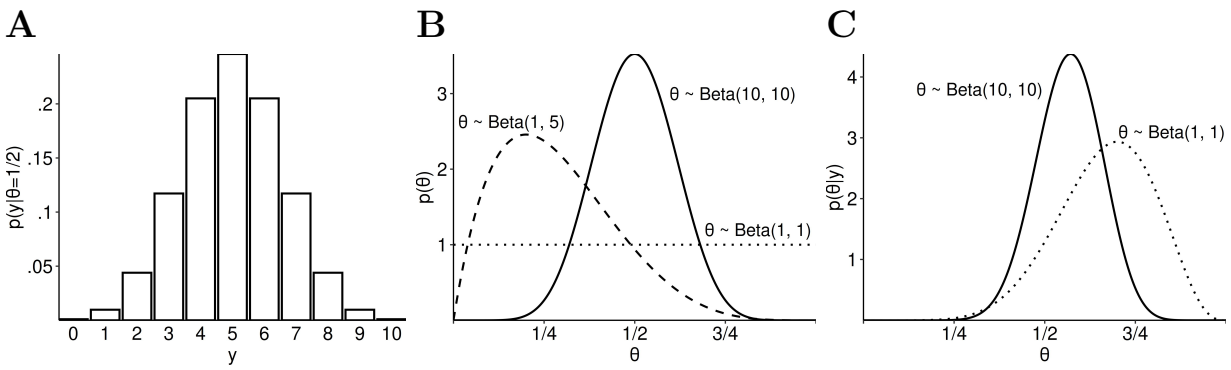


**Figure 1:** **(A)** The probability distribution of a Binomial$(10, 1/2)$ random variable. **(B)** The probability distributions of different beta random variables. **(C)** The posterior distribution of $\theta$ under two different priors given the data $y = 7$ with $n = 10$, where $y \sim \text{Binomial}(n, \theta)$.

We can do some calculations of useful properties of $y|\theta$. Its **expected value, or mean**, is the average of the values it can take, each value weighed by its probability, $\text{E}(y|\theta) = \sum_k kp(y = k|\theta)$. Its **variance** is the average squared distance from the mean. With a bit of work, the mean of a binomially distributed random variable can be shown to be $n\theta$ and its variance $n\theta(1 - \theta)$. It makes sense, for example, that the expected number of heads of a fair coin tossed ten times is five, and that the average distance from the mean increases as the scale of possible values increases.

We can simulate the distribution by simulating sets of ten random coin flips and counting how many experiments record $k$ heads for all possible $k$. Properties of the distribution can be estimated from these simulations, for example in R:

```
# Simulate a Binomial(10, 1/2) random variable 10^5 times
y <- rbinom(n=10^5, size=10, prob=1/2)

# Calculate sample mean and variance
mean(y)
var(y)
```

# First draft of fairness: the prior

We've been dealing with the distribution of the number of heads given a known fairness $\theta$. But really we don't know what $\theta$ is: that's why we're doing the experiment. Let's imagine what $\theta$ could be. If the coin were rigged, it might land heads almost all the time or tails almost all the time. But $\theta$ could be anywhere between 0 and 1, the valid range of a probability, and we don't yet have any experimental data of what it is.

Some random processes, at the mint, or at a trick coin factory, wear and tear, the style of flipping, contribute to this coin's fairness $\theta$, which we can treat as a random variable. What is the distribution of $\theta$? Since we haven't seen these prior processes, haven't measured anything about the coin, haven't even tossed it, it falls to us to make a first judgment about $p(\theta)$.

The beta class of distributions gives us a nice way of quantifying these judgments. A beta distribution is of the form

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\alpha > 0$ and $\beta > 0$ are parameters that tune its shape. Here B is a function of $\alpha$ and $\beta$ that we won't concern ourselves with since it serves purely as a normalization: it doesn't depend on $\theta$ and so doesn't affect the shape of the distribution, only its area. In notation, we write that $\theta \sim \mathrm{Beta}(\alpha, \beta)$.

The beta distribution for different settings of $\alpha$ and $\beta$ are shown in Figure 1B. Since $\theta$ is continuous, its mean and variance are calculated not by summation as in the case of $y|\theta$, where it was possible to enumerate all possible $y$, but by integration, so that its mean for example is $\mathrm{E}(\theta) = \int_\theta \theta p(\theta) d\theta$. This mean can be shown to be $\alpha/(\alpha+\beta)$, so setting $\beta > \alpha$, for example, pushes the mean toward the left. Its variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ is more complicated but for example setting $\alpha = \beta = 10$ centers the distribution at $1/2$ with seven times less variance than when $\alpha = \beta = 1$, which in fact represents a uniform distribution and the belief that all possible fairnesses are relatively speaking equally likely.

When we spoke of the distribution $p(y|\theta)$, the $p$ truly denoted a probability for each possible number of heads. But now we're dealing with fairness, a continuous quantity, which makes the interpretation of $p(\theta)$ subtler. Here $p$ isn't truly a probability. The probability that $\theta$ is any particular value is zero; there are an infinite number of possible values of $\theta$ between zero and one and each can't have its own chance. Instead, $p$ here represents density. One must integrate over an interval of $\theta$ to observe any mass, meaning any nonzero probability. Thus we use $p$ to denote probability for discrete distributions, and density for continuous distributions. This notation is convenient since probabilities and densities can often be treated similarly in mathematical derivations.

Note that by modeling the coin's fairness continuously, we assume the coin has no chance of being truly fair. There is however a nonzero probability that the coin is within, say, 0.05 points of being fair: $p(\theta \in (0.45, 0.55)) > 0$. For example, assuming a uniform distribution over $\theta$, this probability is 0.1, the area over that interval.

## Second draft of fairness: the posterior

We now have the structure of how the data was generated: each flip of the coin lands heads with probability $\theta$. If we knew $\theta$ we'd be able to say exactly how the number of heads is distributed. But we don't know $\theta$ and we must use the $y$ we measure to infer it.

In other words, we know how the data is generated except for the setting of a parameter, and we want to use our data to learn about the true value of that parameter. This kind of situation is at the heart of statistics: we have a model, we have data, and we want to use that data to improve the model. Bayes' Theorem, which can be easily derived from the axioms of probability, elegantly and exactly expresses how to make this update:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

In technical terms, $p(y|\theta)$ is called the **likelihood**, $p(\theta)$ the **prior**, and $p(y)$ has a number of names but let's call it the **normalization** (so called because it doesn't depend on $\theta$ and thus is practically a constant that simply rescales the numerator). What we infer, the parameter or parameters given the data, is the **posterior**, $p(\theta|y)$.

In the last two sections we settled on a reasonable likelihood and form of the prior. We'll therefore know the form of the numerator in our application of Bayes' theorem. Note that the normalization doesn't depend on $\theta$: by definition it's the distribution of the data not given $\theta$. Let's ignore it for now since we're interested in the shape and not the area of the posterior distribution of $\theta$. In fact, let's prune further by removing all the factors in the likelihood and prior that don't contain a $\theta$. Instead of computing $p(\theta|y)$ exactly, we compute what it's proportional to, its shape:

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}. \end{aligned}$$

This expression looks a lot like the forms of both the beta and binomial distributions. But $\theta|y$ can't be binomially distributed since it's the posterior fairness, which is continuous, and so it must be beta-distributed. Indeed, the probability function of a $\text{Beta}(\alpha+y, \beta+n-y)$ random variable would be proportional to what we wrote for $p(\theta|y)$, and so we can conclude

$$\theta|y \sim \text{Beta}(\alpha+y, \beta+n-y).$$

It's called conjugacy when the distributional form of the prior is the same as that of the posterior. As we have just shown, the beta distribution is conjugate to the binomial distribution. Conjugate models are nice since the posterior can be computed so simply. Conjugacy also lends a beautiful interpretation to the model. Each time we toss the coin, we increment the first parameter of the posterior beta distribution if it lands heads, and otherwise increment the second parameter; the first parameter counts heads, the second tails. Thus, we can view $\alpha$ as the prior number of coin flips that landed heads and $\beta$ as that that landed tails. To set a uniform, $\text{Beta}(1,1)$ prior is equivalent to saying that we've seen two coin flips land evenly before beginning the experiment. Such a prior experiment need not have actually happened. It's a prior belief: it's as if we've seen such a prior experiment upon distilling our prior knowledge. (For another example of a beautiful conjugate model, I refer the reader to the Gamma-Poisson model, which can infer the underlying rate of an event from counts of the event over separate time intervals.)

# Inference and the influence of the prior

We now have our model and we know how to infer the parameter $\theta$ of that model from data. Before performing our experiment, let's reason about the prior we might actually want to set in light of the insight that the posterior distribution is combining a (potentially imaginary) prior experiment with our new one. We've been using the uniform prior as an example. It's useful to think about how the choice of prior influences the posterior before we compute it using real data. If we were to observe, say, just one flip whose result is heads, the

posterior would be distributed $\text{Beta}(2, 1)$ and the posterior expectation of the fairness would be $\text{E}(\theta|y) = 2/3$.

Would we really expect a random coin to be so unfair? The uniform prior over $\theta$, although uniform, is itself an assumption and may not be realistic. Knowing nothing else, for example, I acknowledge the random coin could be slightly unfair for some physical reasons I'm unaware of, but based on my prior experience with coins I'd be very skeptical of it being very unfair. I might therefore assume a $\text{Beta}(10, 10)$ prior, under which the posterior expectation given one heads in one flip would be $11/20 \approx 0.55$, which to me seems more reasonable if my goal were to give my best estimate of the fairness of the coin. Under this stronger prior, the posterior is less sensitive to the noise in the data.

In science we often have at least some prior belief, and there's no good way in general to posit an absence of belief. Weak priors can be bad priors, which is important if the data are small and/or noisy. As more data is collected, the influence of the prior wanes. If we were to flip the coin a million times, whether we set $\alpha = \beta = 1$ or $\alpha = \beta = 100$ won't matter much to the posterior.

Now we run the actual experiment by flipping the coin $n = 10$ times. We see it lands heads $y = 7$ times. To repeat the original question: how fair is the coin? We're now in a position to answer. Figure 1C shows the exact distribution of the fairness inferred under our model, with two different priors for comparison.

## Some extensions

I've given a detailed explanation of the beta-binomial model which is often useful by itself. But one of the wonderful aspects of Bayesian statistics is the adaptability of its models. I'll share two extensions of what we've done.

It's often useful to compare two random variables (for example, in clinical medicine, in inferring the difference in patient outcomes between control and treatment groups). To demonstrate the power of inferring posterior distributions, let's extend our inference to two coins and ask how different they are in fairness. One is the coin we just flipped and which landed heads seven of ten times, and the other is a new coin which in two flips lands evenly heads and tails. Let's call the fairness of the first coin $\theta_1$ and that of the second $\theta_2$. We're now interested in the distribution of $\theta_2 - \theta_1$, which is itself a random variable. Rather than deriving a new inference method, we can use our conjugate model twice, once for each coin, and simulate the difference:

```
# Draw 10^5 samples of each posterior theta assuming uniform priors
post1 <- rbeta(10**5, 1+7, 1+3)
post2 <- rbeta(10**5, 1+1, 1+1)

# Sample the difference in the posterior fairnesses
diff <- post2 - post1
quantile(diff, probs=c(0.025, 0.5, 0.975))
```

The last line of code estimates the median posterior difference to be $-0.17$, meaning the second coin seems to tend considerably more toward tails, but with a central 95% credible

interval of $(-0.65, 0.33)$. This interval contains 95% of the posterior probability mass of the difference. Given the width of this interval, we're quite uncertain of our estimate of the difference, which makes sense given the small number of observations.

This was an easy extension since the underlying model was unchanged. Consider however that maybe at night when the room is dark, coins that land heads are sometimes misread as tails with probability $\epsilon$. We can write this model:

$$\begin{aligned}
\theta &\sim \text{Beta}(1, 1) \\
\epsilon &\sim \text{Beta}(1, 5) \\
y|\theta &\sim \text{Binomial}(n_1, \theta) \\
\gamma|\theta &\sim \text{Binomial}(n_2, \theta(1-\epsilon)).
\end{aligned}$$

We observe $n_1$ flips $y$ of which land heads perfectly seen in daylight, but of $n_2$ flips at night we observe an error-prone count of $\gamma$ heads.

There are now two parameters $\theta$ and $\epsilon$ in this model and we must adjust our inference technique. Although there are simpler ways in this case, for generality's sake, supposing we might want to further modify the model later on, let's do inference using a probabilistic programming language. Probabilistic programming languages generally use a technique called Markov chain Monte Carlo to sample the posterior in a way that's flexible to the specification of the model, allowing the user to make adjustments to the model without having to reformulate an inference method each time. In Stan [4], a probabilistic programming language, our model can be written in a file named "model.stan" as:

```
data {
  int<lower=0> n1;
  int<lower=0> n2;
  int<lower=0> y;
  int<lower=0> gamma;
}
parameters {
  real<lower=0, upper=1> theta;
  real<lower=0, upper=1> epsilon;
}
model {
  theta ~ beta(1, 1);
  epsilon ~ beta(1, 5);
  y ~ binomial(n1, theta);
  gamma ~ binomial(n2, theta * (1-epsilon));
}
```

Saying we observe $y = 7$ heads of $n_1 = 10$ daytime flips and $\gamma = 2$ heads of $n_2 = 5$ nighttime flips, the posterior distributions of $\theta$ and $\epsilon$ can be inferred in R as:

```
library(rstan)
stan("model.stan", data=list(n1=10, n2=5, y=7, gamma=2))
```

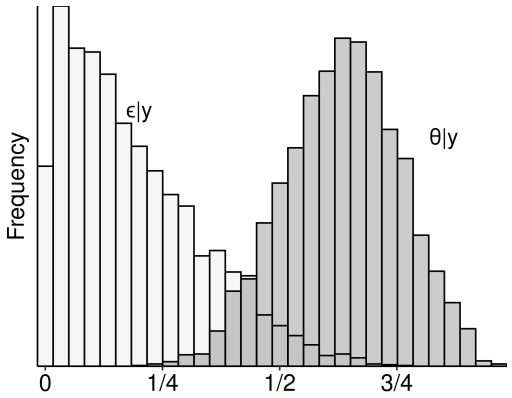


**Figure 2:** The posterior distributions of $\theta$ and $\epsilon$ where $\theta$ is still the underlying fairness of the coin but $\epsilon$ is the probability that each heads is misread as tails at night.

Stan then samples the posterior distribution and these samples can plotted as in Figure 2.

Another extension might be to say that the fairness of the coin depends on who's flipping it: maybe each person has a different technique that affects the fairness. Then we should take into account who's flipping the coin and the variation in fairness among them, as authors of a study of coin flips recently did [5].

# Conclusion

I often refer to the coin-flipping example in my research and teaching because it accessibly conveys important principles of applied statistics: to model carefully, think distributionally, and simulate. It's informative for the beginner and often useful when analyzing count data. It's also a conversation piece for experienced statisticians. What does it mean for a fairness to be random? Is a coin ever truly fair? How should you set priors? How extensive should your model be? Statistics is a beautiful but precarious discipline. Rich examples steady us as we climb and I hope my explanation of the beta-binomial model will serve in this respect.

# References

1. Dogucu, M. & Hu, J. The current state of undergraduate Bayesian education and recommendations for the future. *The American Statistician* 76, 405–413 (2022).
2. Blitzstein, J. K. & Hwang, J. *Introduction to Probability* (Chapman and Hall/CRC, 2019).
3. Gelman, A. *et al. Bayesian Data Analysis* (CRC Press, 2013).
4. Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* 76 (2017).
5. Bartoš, F. *et al.* Fair coins tend to land on the same side they started: Evidence from 350,757 flips. *arXiv preprint arXiv:2310.04153* (2023).