`Project2`: Flipping a biased coin

Shahla Daneshmehr

March 2023

1 Abstract

The coin-flip problem is a classical problem in statistics. We toss a coin a number of times and record how many heads and tails we get. Based on this data we try to know how biased is the coin. This experiment measures ten different trials of coin flips and graphs the results. Python, by using alpha and beta parameters to compute the posterior distribution of the bias parameter using Bayes' theorem, will compute and plot the posterior.

2 Introduction

In this project, we will assume that we have already tossed a coin a number of times and we have recorded the number of observed heads and tails, so the data part has been done in project 1. In this project, we are going to generalize the concept of bias. For a coin with a bias of 1, it will always land heads, one with a bias of 0 will always land tails, and one with a bias of 0.5 will land half of the time heads and half of the time tails. To represent the bias, we use the parameter θ, and to represent the total number of heads for an N number of tosses, we use the variable y. According to Bayes' theorem [1], we have the following formula:

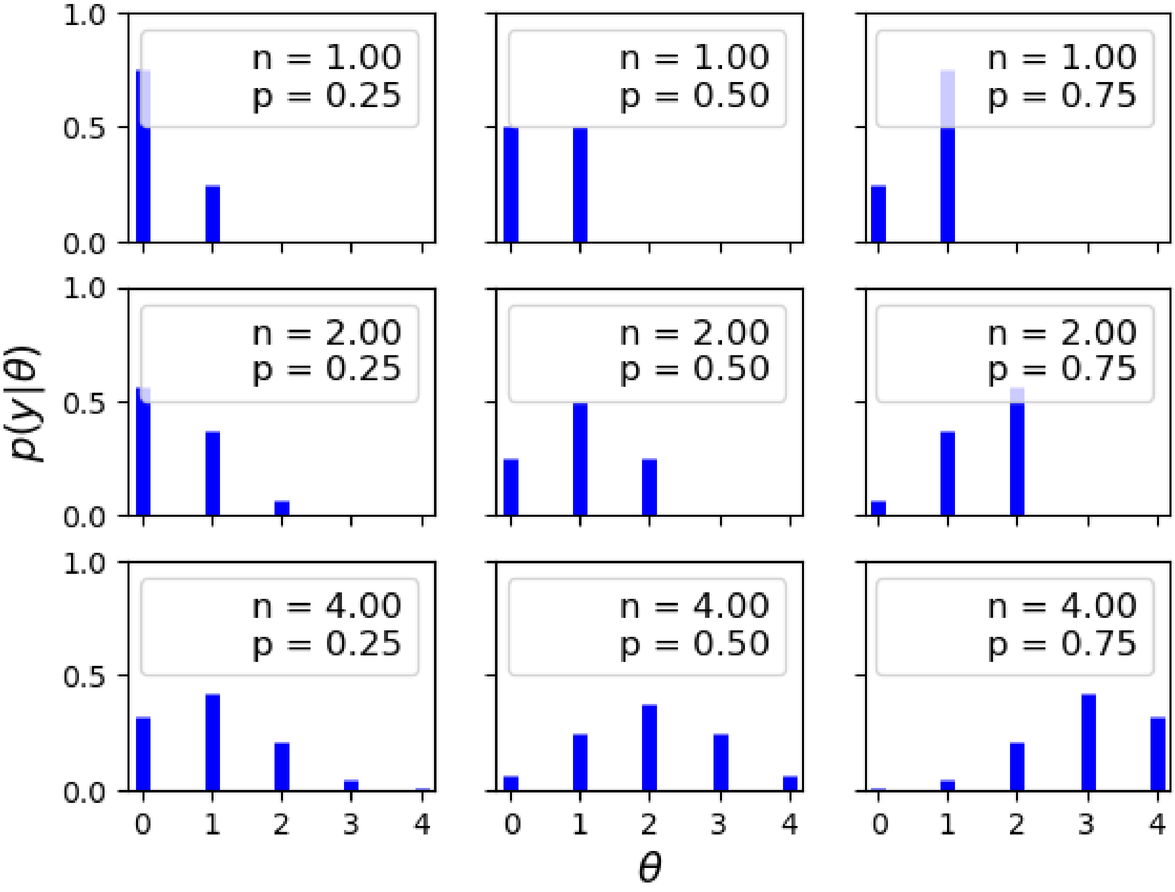
$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

We will specify prior $p(\theta)$ and likelihood $p(\ y \mid \theta)$.

We assume that a coin toss does not affect other tosses, which means coin tosses are independent of each other. We also assume that only two outcomes are possible, heads or tails. Given these assumptions, a good candidate for the likelihood is the binomial distribution:

$$P(y|\theta) = \frac{N!}{N!(N-y)!}\theta^{y}1-\theta^{(N-y)}$$

This is a discrete distribution returning the probability of getting y heads (or in general, success) out of N coin tosses (or in general, trials or experiments) given a fixed value of (θ) [2]. For example, we write a code that generates 9 binomial distributions:



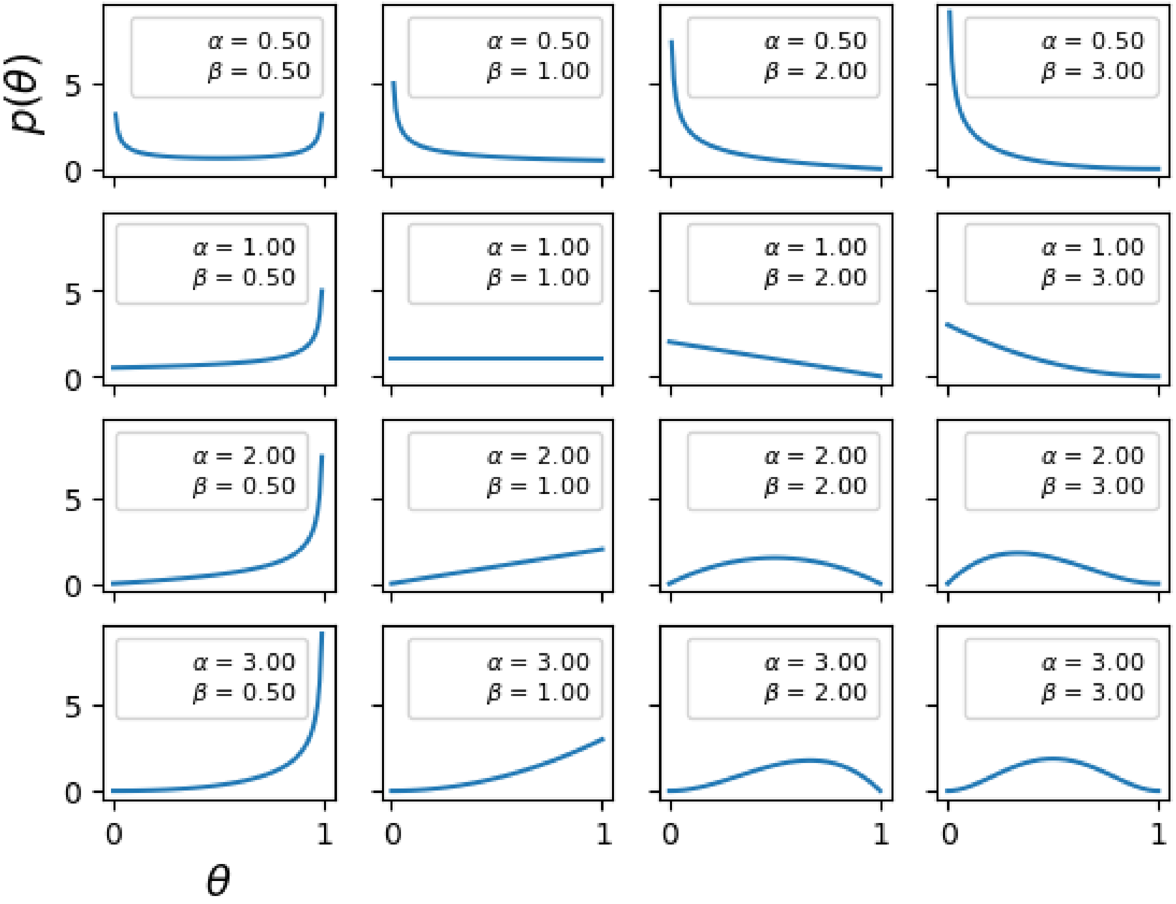
n = 1.00
p = 0.25
n = 1.00
p = 0.50
n = 1.00
p = 0.75
n = 2.00
p = 0.25
n = 2.00
p = 0.50
n = 2.00
p = 0.75
n = 4.00
p = 0.25
n = 4.00
p = 0.50
n = 4.00
p = 0.75
$p(y|\theta)$
θ
0.0
0.5
1.0
0
1
2
3
4

We can see that (θ) indicates how likely it is that we will obtain a head when tossing a coin, and we have observed that event y times. Following the same line of reasoning we get that $1-\theta$ is the chance of getting a tail, and that event has occurred N-y times.

By having (θ), the binomial distribution will find out the expected distribution of heads [3]. The only problem is that we do not know(θ). In Bayesian statistics, every time we do not know the value of a parameter, we put a prior on it. Therefore, we are going to choose the prior. As a prior we will use a beta distribution, which is a very common distribution in Bayesian statistics and looks like this:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}1-\theta^{\beta-1}$$

It is obvious that the beta distribution looks similar to the binomial except for the term with the Γ[4]. This is the Greek uppercase gamma letter and represents what is known as the gamma function. All that we care about at this point is that the first term is a normalization constant that ensures the distribution integrates to 1 and that the beta distribution has two parameters, α, and β, that control the distribution.



The posterior probability is a type of conditional probability that results from updating the prior probability with information summarized by the likelihood via an application of Bayes' rule [5]. From an epistemological perspective, the posterior probability contains everything there is to know about an uncertain proposition (such as a scientific hypothesis, or parameter values), given prior knowledge and a mathematical model describing the observations available at a particular time [6]. After the arrival of new information, the current posterior probability may serve as the prior in another round of Bayesian updating [7].

In the context of Bayesian statistics, the posterior probability distribution usually describes the epistemic uncertainty about statistical parameters conditional on a collection of observed data. From a given posterior distribution, various point and interval estimates can be derived, such as the maximum a posteriori (MAP) or the highest posterior density interval (HPDI) [8]. But while conceptually simple, the posterior distribution is generally not tractable and therefore needs to be either analytically or numerically approximated [9]. In Bayesian probability theory, if the posterior distribution $p(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x \mid \theta)$.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise, numerical integration may be necessary. Further, conjugate priors may give intuition by more transparently showing how a likelihood function updates a prior distribution. Bayes' theorem says that the posterior is proportional to the likelihood times the prior:

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

So for our problem, we have to multiply the binomial and the beta distributions:

$$P(\theta|y) \propto \frac{N!}{N!(N-y)!}\theta^y 1-\theta^{(N-y)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}\theta^{\beta-1}$$

Now we want to simplify this expression. We can drop all the terms that do not depend on and our results will still be valid. So we can write the following:

$$P(\theta|y) \propto \theta^y 1-\theta^{(N-y)}\theta^{\alpha-1}\theta^{\beta-1}$$

Reordering it, we get the following:

$$P(\theta|y) \propto \theta^{\alpha-1+y}1-\theta^{\beta-1+N-y}$$

If we pay attention, we will see that this expression has the same functional form of a beta distribution (except for the normalization) with

$$\alpha_{posterior} = \alpha_{prior} + y$$

and

$$\beta_{posterior} = \beta_{prior} + N - y$$

which means that the posterior for our problem is the beta distribution:

$$P(\theta|y) = Beta(\alpha_{prior} + y, \beta_{prior} + N - y)$$

3 Hypotheses to Explain coin flipping

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments. It takes different values in coin tosses which it has not simply fixed values. We know that in Bayesian statistics, Posterior = Likelihood × Prior ÷ Evidence. In this work, I use beta distribution as prior with alpha= 0.6 and beta parameters are (1, 1), (0.5, 0.5), (20, 20) for ten trials with theta= 0.35.

4 Code and Experimental Simulation

Python 3 computes analytical expression for the posterior in our experiment (coin tosses) and plots the results.

Let's look at a simulated code:

```
#! /usr/bin/env python
# imports of external packages to use in our code
import sys
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

theta_real = 0.35
trials = [0, 1, 2, 3, 4, 8, 16, 32, 50, 150]
data = [0, 1, 1, 1, 1, 4, 6, 9, 13, 48]

beta_params = [(1, 1), (0.5, 0.5), (20, 20)]
dist = stats.beta
x = np.linspace(0, 1, 100)

for idx, N in enumerate(trials):
    if idx == 0:
        plt.subplot(4,3, 2)
    else:
        plt.subplot(4,3, idx+3)
    y = data[idx]
    for (a_prior, b_prior), c in zip(beta_params, ('b', 'r', 'g')):
        p_theta_given_y = dist.pdf(x, a_prior + y, b_prior + N - y)
        plt.plot(x, p_theta_given_y, c)
        plt.fill_between(x, 0, p_theta_given_y, color=c, alpha=0.6)

    plt.axvline(theta_real, ymax=0.3, color='k')
    plt.plot(0, 0, label="{:d} experiments\n{:d} heads".format(N, y), alpha=0)
    plt.xlim(0,1)
    plt.ylim(0,12)
    plt.xlabel(r'$\theta$')
    plt.legend()
    plt.gca().axes.get_yaxis().set_visible(False)
plt.tight_layout()
```

5 Analysis

As a matter of fact, sometimes visual summaries of data can tell us more than text summaries. The plots show us ten different trials for having the head in biased coin flips.

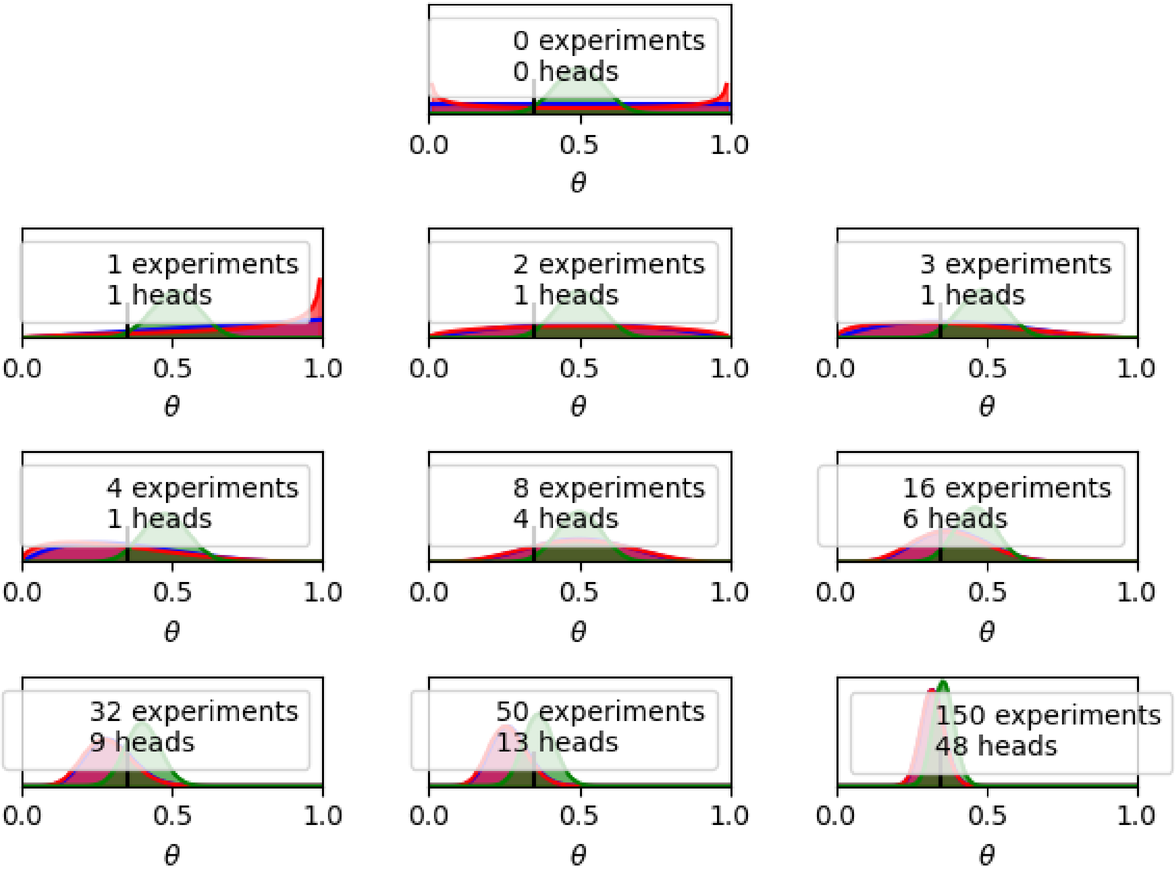


Figure 1: Simulations of biased coin flips probabilities

On the first plot, we have 0 experiments done, hence these curves are just our priors. We have three curves, one per prior: The blue one is a uniform prior. This is equivalent to saying that all the possible values for the bias are equally probable a priori. The red one is similar to the uniform. This means we are a bit more confident that the bias is either 0 or 1 than the rest of the values. The green and last one is centered and concentrated around 0.5, so this prior is compatible with information indicating that the coin has more or less about the same chance of landing heads or tails. We could also say this prior is compatible with the belief that most coins are fair(a word commonly used in Bayesian discussions). The rest of the subplots show posteriors $p(\theta \mid y)$ for successive experiments. We consider posteriors as updated priors given the data. The number of experiments (or coin tosses) and the number of heads are indicated in each subplot's legend. There is also a black vertical line at 0.35 representing the true value for θ.

6 Conclusion

The result of a Bayesian analysis is the posterior distribution, not a single value but a distribution of plausible values given the data and our model. The most probable value is given by the mode of the posterior (the peak of the distribution). Given a sufficiently large amount of data, two or more Bayesian models with different priors will tend to converge to the same result. How fast posteriors converge to the same distribution depends on the data and the model. In the plots, we see that the blue and red posteriors look almost indistinguishable after only 8 experiments, while the red curve continues to be separated from the other two even after 150 experiments. Something not obvious from the figure is that we will get the same result if we update the posterior sequentially than if we do it all at once. We can compute the posterior 150 times, each time adding one more observation and using the obtained posterior as the new prior, or we can just compute one posterior for the 150 tosses at once. The result will be exactly the same. This feature not only makes perfect sense, also leads to a natural way of updating our estimations when we get new data, a situation common in many data analysis problems.

7 References

1. Puga, J., Krzywinski, M. Altman, N. Bayes' theorem. Nat Methods 12, 277–278 (2015).

2. Li, Li. "Discrete distributions." (1972).

3. Edwards, A. W. F. "The meaning of binomial distribution." Nature 186, 1074-1074 (1960).

4. Howard, Ronald A. "Decision analysis: Perspectives on inference, decision, and experimentation." Proceedings of the IEEE 58.5, 632-643 (1970).

5. Lambert, Ben. "The posterior – the goal of Bayesian inference". A Student's Guide to Bayesian Statistics. Sage. pp. 121–140 (2018).

6. Grossman, Jason. Inferences from observations to simple statistical hypotheses (PhD thesis) (2005).

7. Etz, Alex. "Understanding Bayes: Updating priors via the likelihood". (2015-07-25) The Etz-Files. Retrieved 2022-08-18.

8. Gill, Jeff. "Summarizing Posterior Distributions with Intervals". Bayesian Methods: A Social and Behavioral Sciences Approach (Third ed.). Chapman Hall. pp. 42–48 (2014).

9. Press, S. James. "Approximations, Numerical Methods, and Computer Programs". Bayesian Statistics: Principles, Models, and Applications. New York: John Wiley Sons. pp. 69–102 (1989).