Inference about the bias of a coin

Giorgio Corani

Bayesian Data Analysis and Probabilistic Programming

References

- The Beta-Binomial model: Ch. 3 of `Bayes Rules! An Introduction to Applied Bayesian Modeling`
  - https://www.bayesrulesbook.com/chapter-3.html#chapter-3
  - Alicia A. Johnson, Miles Q. Ott, Mine Dogucu

Estimating the bias θ of a coin

- A coin falls tails with probability $\theta \in [0, 1]$
- θ is the *bias* of the coin
  - θ =0: it always lands tails
  - θ =1: it always lands heads
- We flip the coin n times and we measure the number of heads y.
- Let us start with a prior distribution $f(\theta)$ and update it with the data, to obtain a posterior distribution $p(\theta \mid y)$.

The coin problem

- The coin stands in for many real-world applications, such as estimating the proportion of supporters of a political party or the click-through rate of an online advertisement.

Assumptions

- Each observation takes a binary value (head or tail; also referred to as *success* and *insuccess*)
- The *success* usually refer to the rarer event among the two.
- The flips are independent: the probability of *heads* at the next flip does not depend on the outcome of the previous flips.
- θ is constant in all flips.

The coin problem

A single flip takes:

- *heads* with probability θ
- *tails* with probability $1-\theta$
- Assuming a constant θ and the independence of the flips, the sequence $H \quad T \quad T \quad H \quad H$ has probability $ □(1-□) (1-□) □□= □^2 (1-□)^3$
- The probability of any sequence containing y heads in n flips is: $\theta^y(1-\theta)^{n-y}$

Binomial likelihood

- We can get $\binom{n}{y} = \frac{n!}{k!(n-y)!}$ sequences containing y successes in n trials.
- The probability of observing y successes in n trials is:

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

- This is probability of the observing y tails within n flips, given a fixed value of θ.

Binomial likelihood

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

* This a *likelihood* function if interpreted in this way: * the probability is a function of θ. * the observation y are fixed * The likelihood function shows how the probability of the observed data varies with θ.

The Beta prior

- We need a continuous prior $f(\theta)$ for θ, i.e., a probability density function (pdf).
- It specifies all possible values of θ and the relative plausibility of each.
- It has to be limited on (0,1)
- It integrates to 1, much like a discrete pmf sums to 1.

The Beta distribution

- Beta(a, b), is a suitable prior for variables restricted to the $[0, 1]$ interval.
- Its parameters are $a > 0$ and $b > 0$. Parameters used in prior models are referred to as *hyperparameters*.
- The density function is:

$$f(\theta) = \frac{1}{\underbrace{B(a, b)}_{\text{normalizing constant}}} \theta^{a-1}(1-\theta)^{b-1} \propto \theta^{a-1}(1-\theta)^{b-1} \qquad a, b > 0$$

- In the following we ignore the normalizing constant.

Note that:

- θ is raised to the power of $a - 1$ (not a)
- $1 - \theta$ is raised to the power of $b - 1$ (not b)

Density function

- The density $f(\theta)$ evaluated for a specific value θ is not a probability.
- Probabilities are obtained by integrating the pdf over an interval i.e.:

$$P(c < \theta < d) = \int_c^d f(\theta)d\theta$$

- It is possible that $f(\theta) > 1$ even though the density integrates to 1.

The prior $p(\theta)$

- The mean captures the average value of θ:

$$E[\theta] = \int_x x \cdot f(x)dx$$

  - each possible θ value is weighted by its corresponding pdf value
- The mode is the value of θ at which the pdf f(θ) is maximized.
  - Mode($\theta = \arg_{max} f(\theta)$)

$$\begin{aligned} E(\theta) &= \frac{\alpha}{\alpha+\beta} \\ \text{Mode}(\theta) &= \frac{\alpha-1}{\alpha+\beta-2} \quad \text{when} \;\; \alpha, \beta > 1. \end{aligned} \tag{1}$$

Beta distribution

- Expectation (i.e., mean):

$$E(\theta) = \frac{a}{a+b}$$

- Variance:

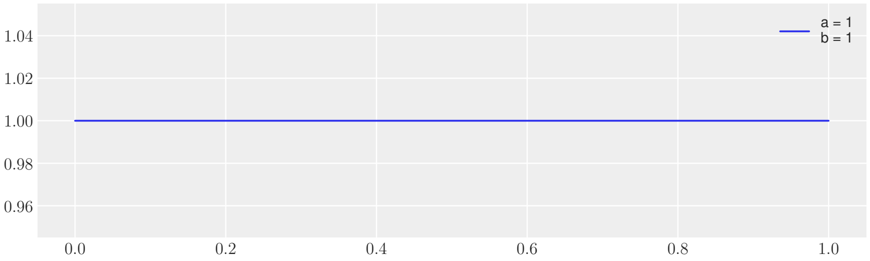
$$\mathrm{VAR}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$$

- Higher a: the expected value increases, the bulk of the distribution moves rightwards (vice versa for higher b)
- $a = b$ implies a symmetric distribution with expected value 0.5.
- Higher a and b: the distribution gets more concentrated around the mean.

Beta distribution with $a = b = 1$

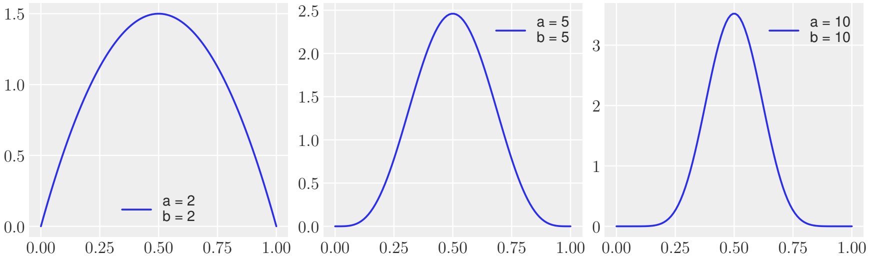
$$\begin{aligned} f(\theta) &\propto \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^0(1-\theta)^0 \\ &= 1 \end{aligned}$$

- This a *uniform* distribution: all values in $(0, 1)$ are equally probable.
- $E(\theta) = \frac{a}{a+b} = 0.5$.



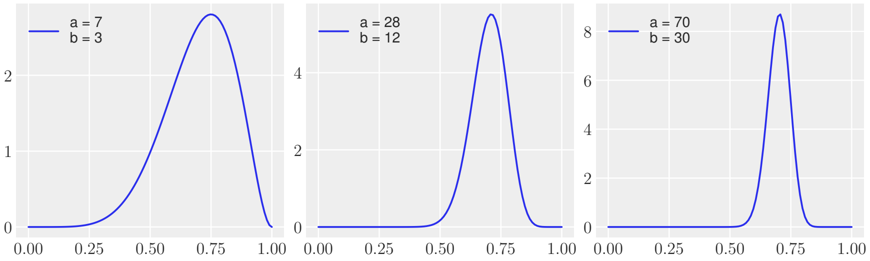
Increasing a and b the prior becomes more concentrated

- If we increase a and b together, the prior becomes more concentrated around the expected value $\theta = 0.5$
- This corresponds to be more confident that the coin is fair (θ=0.5).



If we think the coin is rigged towards *tails*

- If we suspect the coin to be 70% rigged towards heads, we set $a = \frac{7}{3}b$.
- We represent more confidence in this statement by:
  - increasing b
  - keeping $a = \frac{7}{3}b$.

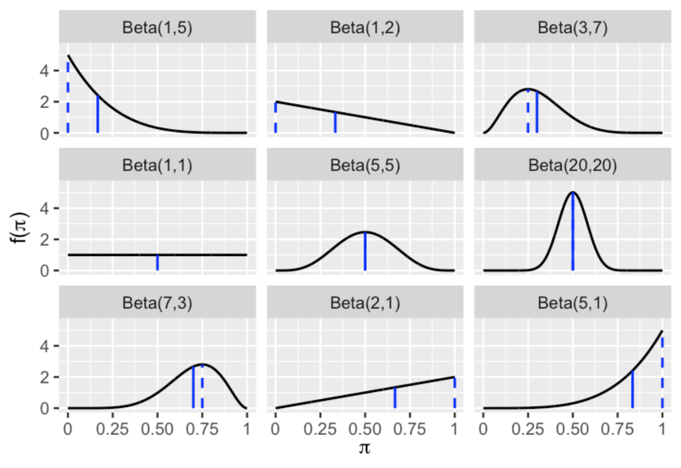


Tuning the beta

- mode is shown by the dashed line

- mode is shown by the dashed line

```r
knitr::include_graphics("beta-tuning-1.png")
```



Choice of a and b

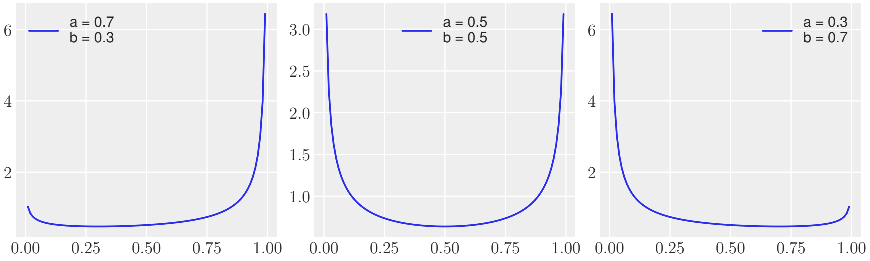
- The choice of a and b can be fine-tuned considering also the quantiles of your prior beliefs.
- You might expect a drug to heal some 70% of the patients, but with which uncertainty?
- This is how the 10-the and the 90-th percentile vary with a and b:

(a,b)	(7, 3)	(28, 12)	(70, 30)
5-th quantile	0.45	0.58	0.62
95-th quantile	0.90	0.81	0.77

```
#{python, echo=TRUE, fig.height=3, cache=TRUE,
fig.align="center"} #for (a, b) in  [(7, 3), (28, 12),
(70, 30)]: #    quantiles=[0.05, 0.95] #    q =
beta.ppf(quantiles, a=a, b=b) #    print("a=", a , " b=",
b, ", q10=", q[0], ", q90=", q[1]) #
```

We suspect the coin to be rigged but we do not in which direction

- $a < 1, b < 1$ yield a convex distribution



Discussion

The beta distribution can represent different types of prior beliefs about θ, such as:

- all values of θ are equally probable a priori
- we think the coin is likely to be fair, but we are not fully sure (bell centered in θ=0.5).
- we think the coin is likely to be rigged towards tails (asymmetric distribution centered in e.g. θ=0.7, less or more concentrated).
- we think the coin to be rigged, but she does not know in which way (convex distribution).

Posterior

Adopting a beta prior for θ and a binomial distribution as *likelihood*, we obtain a beta *posterior* distribution with updated parameters:

$$\begin{aligned} p(\theta) &\propto \theta^{a-1}(1-\theta)^{b} \\ p(y \mid \theta) &= \theta^{y}(1-\theta)^{n-y} \\ p(\theta \mid y) &\propto \theta^{y+a-1}(1-\theta)^{n-y+b-1} \end{aligned}$$

The beta prior is *conjugate* with the binomial likelihood, as we obtain a beta posterior.

Conjugacy

According to Bayes' theorem, the posterior is the product of the likelihood and the prior:

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

In our case:

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

which is a Beta distribution (without expressing the normalization constant).

The posterior is a compromise of prior and likelihood

- Given the prior Beta(a,b), the prior mean of θ is:

$$\frac{a}{a+b}$$

- Having observed y tails in n flips, the posterior distribution of θ is Beta($y+a$,$n-y+b$). The posterior mean is:

$$E_{\text{post}}[\theta] = \frac{a+y}{a+y+b+n-y} = \frac{a+y}{a+b+n}$$

- Rearranging:

$$\underbrace{\frac{a+y}{a+b+n}}_{\text{posterior}} = \underbrace{\frac{y}{n}}_{\text{observed proportion}} \underbrace{\frac{n}{n+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior mean of }\theta} \underbrace{\frac{a+b}{n+a+b}}_{\text{weight of the prior}}$$

- The posterior mean is a weighted average of the prior mean and the observed proportion.
- The weight of the observed proportion increases with n; the weight of the prior mean increases with a and b.

Conjugacy

- If for a certain *likelihood* the functional form of the *a priori* and that of the *a posteriori* coincide, it is said that the *a priori* is conjugated with the *likelihood*.
- Historically, problems in Bayesian statistics were restricted to the use of conjugate priors, because of mathematical tractability.
- However modern computational techniques allow obtaining posteriors even when conjugacy does not hold, allowing the resurgence of Bayesian statistics in recent years.

Exercise: grid sampling (DISCUTERE CON MARCO)

- Choose a value of θ and simulate 1000 Bernoulli trials from it; collect the value of n and y
- Define your prior by setting a and b (do not abuse the fact that you know θ)
- Discretize the values of $\theta \in (0, 1)$ by 0.01
- Compute for each value of θ the unnormalized posterior $p(⍰| mid y) ⍰Beta(⍰;a,b) p(y ⍰⍰) $
- Check the following:
  - given the large sample size, your posterior remains practically the same if you change the prior
  - given to the large sample size, your posterior is concentrated around the true values of θ
  - your posterior is practically equivalent to $Beta(\theta; a + y, b + n - y)$ also for small sample size.

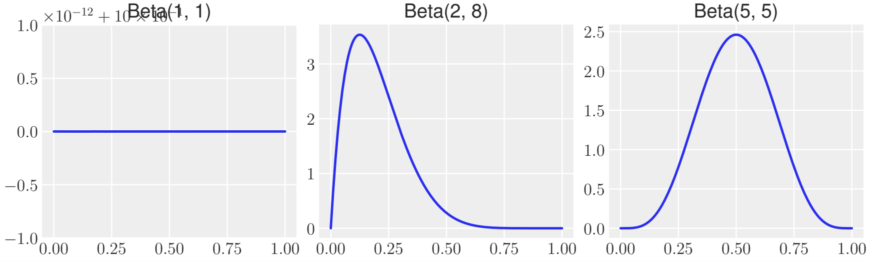
CODE TO BE IMPLEMENTED

Computation

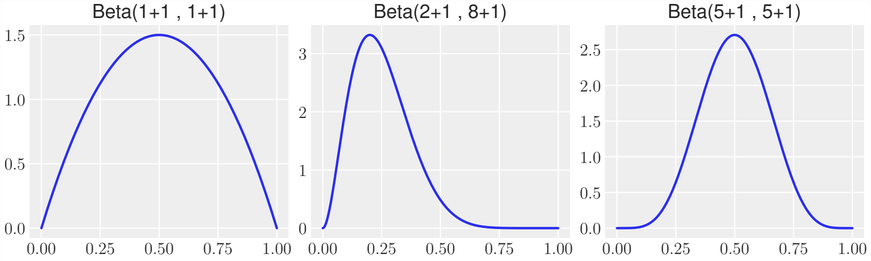
- In the course we will see how to use computational methods to compute the posteriori even with non-conjugate priors.
- In the following we exploit conjugacy in order to explor the sensitivity of the posterior on the prior.

Impact of the prior on the posterior

- We can start from different priors depending on subjective beliefs (priors might be used to encode domain expertise, and different experts would provide you with reasonable but different assessment)
- Let us consider different priors



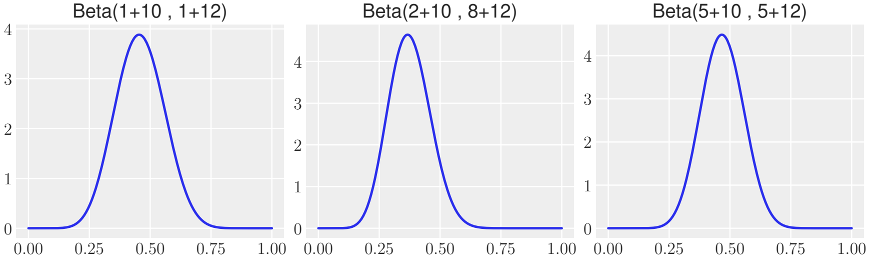
The posterior depends on the priors when observations are few (y=1 tails, n=2, heads=1)



- The Beta(2,8) represents the following beliefs:
  - expected value = $\frac{2}{2+8} = 0.2$
  -

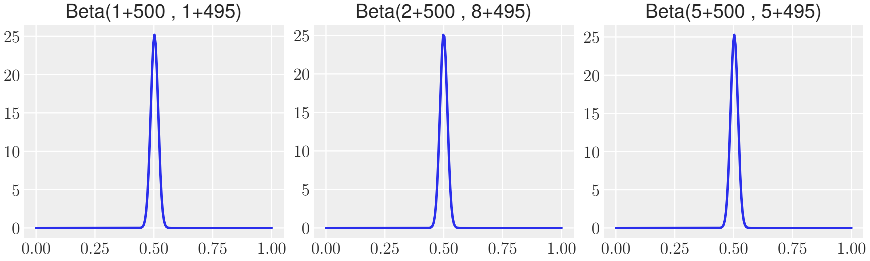
```
from scipy.stats import beta
quantiles=[0.05, 0.25, 0.5, 0.75, 0.95]
q = beta.ppf(quantiles, a=2, b=8)
print(q)
```

The posterior becomes similar as we observe more data (10 tails, 12 heads)



When the number of observations is large, the posterior is the same whatever the prior

- The likelihood overwhelms the prior f



The posterior means $E_{\text{post}}[\theta]$ obtained using the different priors are:

- $\frac{500+1}{500+1+495+1} = \frac{501}{997} = 0.502$
- $\frac{500+2}{500+2+495+8} = \frac{502}{1005} = 0.499$
- $\frac{500+5}{500+5+495+5} = \frac{505}{1005} = 0.502$
- Also the posterior variances and quantiles are practically identical in the different cases.

The posterior mean is just part of the information

- The result of the Bayesian analysis is the posterior distribution of θ, **not** a single value.
- The dispersion of the posterior distribution (posterior variance) is a measure of our uncertainty.
- The uncertainty decreases when the number of experiments is greater.
- Given a large amount of data, the posterior is practically the same with any prior, but how much data is needed varies with the problem.
- If we only have few data, the posterior can differ depending on the adopted prior; it makes sense to repeat the analysis with different priors (*sensitivity*).
- This is sensible: the prior encodes our previous knowledge and different experts could have different priors.

Discussion

- Priors and likelihood are assumptions which are part of the model.
- Uninformative priors (flat) provide the least possible amount of information and therefore have the least possible impact on the analysis.
- *Slightly informative* priors are recommended, which at least provide the order of magnitude of the parameter.
- In many cases we known that the parameter can only be positive, or that they are restricted to sum to 1 or the approximate range, etc.
- For instance even a Beta(1,1) prior is flat but limits the possible values of θ between 0 and 1.

Conclusions

- We have seen how Bayesian inference works when Bayes’ rule can be solved analytically, i.e., when the likelihood has a conjugate prior distribution.
- In this case the posterior distribution has the same mathematical form as the prior.
- Only simple likelihood functions have conjugate priors. In realistic applications the complex models have no conjugate priors. We will abandon exact mathematical solutions to use instead use numerical Markov chain Monte Carlo (MCMC) methods.