Coin flipping

Suppose there is a coin that may be biased – this coin has unknown probability θ of giving a "heads." If we repeatedly flip this coin and observe the outcomes, how can we maintain our belief about θ?

Note that the coin-flipping problem can be seen as a simplification of the survey problem we discussed last time, where we assume that people always tell the truth, are sampled uniformly at random, and whose opinions are generated independently (by flipping a coin!).

Before we select a prior for θ, we write down the likelihood. For a particular problem, it is almost always easier to derive an appropriate likelihood than it is to identify an appropriate prior distribution.

Suppose we flip the coin n times and observe x "heads." Every statistician, regardless of philosophy, would agree that the probability of this observation, given the value of θ, comes from a binomial distribution:

$$\Pr(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Classical method

Before we continue with the Bayesian approach, we pause to discuss how a classical statistician would proceed with this problem. Recall that in the frequentist approach, the value θ can only be considered in terms of the frequency of success ("heads") seen during an infinite number of trials. It is not valid in this framework to represent a "belief" about θ in terms of probability.

Rather, the frequentist approach to reasoning about θ is to construct an *estimator* for θ, which in theory can be any function of the observed data: $\hat{\theta}(x, n)$. Estimators are then analyzed in terms of their behavior as the number of observations goes to infinity (for example, we might prove that $\hat{\theta} \to \theta$ as $n \to \infty$). The classical estimator in this case is the empirical frequency $\hat{\theta} = x/n$.

Bayesian method

An interesting thing to note about the frequentist approach is that it ignores all prior information, opting instead to only look at the observed data. To a Bayesian, every such problem is different and should be analyzed contextually given the known information.

With the likelihood decided, we must now choose a prior distribution $p(\theta)$. A convenient prior in this case is the *beta distribution,* which has two parameters α and β:

$$p(\theta \mid \alpha, \beta) = \mathcal{B}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}.$$

Here the normalizing constant $B(\alpha, \beta)$ is the *beta function:*

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, \mathrm{d}\theta.$$

The support of the beta distribution is $\theta \in (0, 1)$, and by selecting various values of α and β, we can control its shape to represent a variety of different prior beliefs.

Given our observations $\mathcal{D} = (x, n)$, we can now compute the posterior distribution of θ:

$$p(\theta \mid x, n, \alpha, \beta) = \frac{\Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)}{\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta) \, \mathrm{d}\theta}.$$

First we handle the normalization constant $\Pr(x \mid n, \alpha, \beta)$:

$$\begin{aligned}\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)\, \mathrm{d}\theta &= \binom{n}{x} \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta \\ &= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)}.\end{aligned}$$

Now we apply Bayes theorem:

$$\begin{aligned}p(\theta \mid x, n, \alpha, \beta) &= \frac{\Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)}{\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)\, \mathrm{d}\theta} \\ &= \left[\binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)}\right]^{-1} \left[\binom{n}{x} \theta^x (1-\theta)^{n-x}\right] \left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}\right] \\ &= \frac{1}{B(\alpha+x, \beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \\ &= \mathcal{B}(\alpha+x, \beta+n-x).\end{aligned}$$

The posterior is therefore another beta distribution with parameters $(\alpha + x, \beta + n - x)$; we have added the number of successes to the first parameter and the number of failures to the second.

The rather convenient fact that the posterior remains a beta distribution is because the beta distribution satisfies a property known as *conjugacy* with the binomial likelihood. This fact also leads to a common interpretation of the parameters α and β: they serve as "pseudocounts," or fake observations we pretend to have seen before seeing the data.

Figure 1 shows the relevant functions for the coin flipping example for $(\alpha, \beta) = (3, 5)$ and $(x, n) = (5, 6)$. Notice that the likelihood favors higher values of θ, whereas the prior had favored lower values of θ. The posterior, taking into account both sources of information, lies in between these extremes. Notice also that the posterior has support over a narrower range of plausible θ values than the prior; this is because we can draw more confident conclusions from having access to more information.

Hypothesis testing

We often wish to use our observed data to draw conclusions about the plausibility of various hypotheses. For example, we might wish to know whether the parameter θ is less than $1/2$. The Bayesian method allows us to compute this value directly from the posterior distribution:

$$\Pr(\theta < 1/2 \mid x, n, \alpha, \beta) = \int_0^{1/2} p(\theta \mid x, n, \alpha, \beta)\, \mathrm{d}\theta.$$

For the example in Figure 1, this probability is approximately 15%.

There is a sharp contrast between the simplicity of this approach and the frequentist method. The classical approach to hypothesis testing uses the likelihood as a way of generating fake datasets of the same size as the observations. The likelihood then serves as a so-called "null hypothesis" that allows us to generate hypothetical datasets under some condition.

From these, we compute *statistics,* which, like estimators, can be any function of the hypothesized data. We then identify some critical set C for this statistic which contains some large portion $(1 - \alpha)$ of the values corresponding to the datasets generated by our null hypothesis. If the

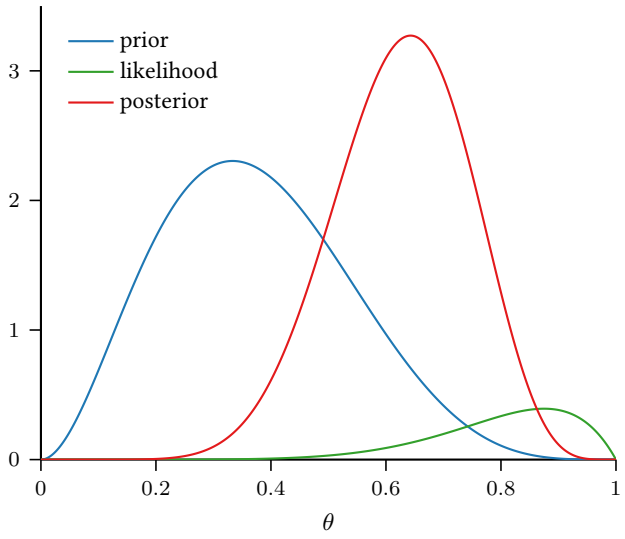

Figure 1: An example of Bayesian updating for coin flipping. Figure produced by `plot_beta_example.m`.

statistic computed from the observed data falls outside this set, we reject the null hypothesis with "confidence" α. Note that the "rejection" of the null hypothesis in classical hypothesis testing is purely a statement about the observed data (that it looks "unusual"), and not about the plausibility of alternative hypotheses!