Probability Theory & Statistics

Beta-Binomial Conjugacy

Prerequisites

- Random Variables and their Distributions

Recap

Random variables are *functions* mapping the sample space, S, to the real number line, $\mathbb{R}$. For example, consider a coin-tossing problem. The structure of the problem is a sequence of trials with two possible outcomes for each trial. The outcomes are either heads (H) or tails (T), or equivalently "success" or "failure", or "1" or "0".

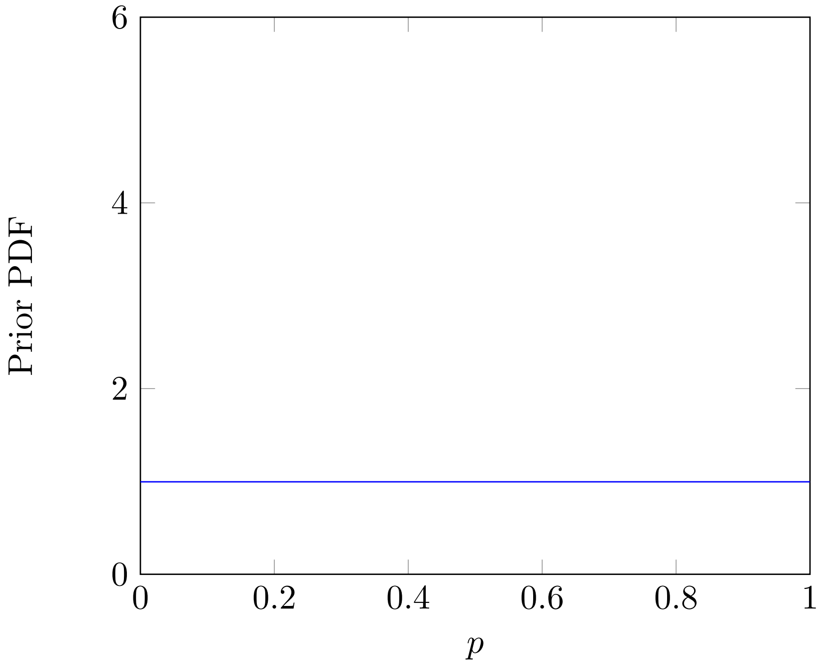

Figure 1: The function X, mapping s_j to the real number line.

For an experiment where the coin is flipped twice, the sample space consists of four possible outcomes

$$S = \{HH, HT, TH, TT\} \tag{1}$$

that can also be mapped to a set of numbers. Figure **??** is one possible mapping, $X(s_j)$, from the sample space to the real number line, i.e.,

$$X(HH) = 2,\ X(HT) = X(TH) = 1,\ X(TT) = 0 \tag{2}$$

where the mapping corresponds to the total number of heads in the experiment. Notice how the mapping is somewhat arbitrary. We could have also defined a mapping on the same sample space counting the number of tails.

Distributions

The *distribution* of a random variable X, is a full description of the probabilities of the events associated with X, e.g., $\{X = 2\}$ and $\{0 \le X \le 2\}$. The distribution of a *discrete* random variable can be defined either by its *Probability Mass Function* (PMF) or its *Cumulative Distribution Function* (CDF). The PMF of a discrete random variable, X, is the function

$$P(A) = P(\{X = x\}) = P(X = x) \tag{3}$$

for $x \in \mathbb{R}$.

For example, consider a 2-flip coin experiment,

$$P(X = 2) = P(\{X(s) = 2\}) = P(\{X(s_1)\}) = P(\{s_1\}) = 1/4 \tag{4}$$

The CDF of X is the function

$$P(B) = P(\{X \le x\}) = P(X \le x). \tag{5}$$

For a 2-flip coin experiment we get,

$$P(X \le 2) = P(\{X(s) \le 2\}) = P(\{X(s_1), X(s_2), X(s_3), X(s_4)\}) = P(\{S\}) = 1. \tag{6}$$

A PMF is *valid* if it is nonnegative and sums to 1. A CDF is valid if it is right-continuous and increasing, and if it converges to 0 as x tends to $-\infty$, and converges to 1 as x tends to ∞.

A random variable has a *continuous distribution* if its CDF is *differentiable*.[1] A *continuous random variable* is a random variable with a continuous distribution. It is often much more convenient to work with the derivative of a continuous CDF, a function called the *Probability Density Function* (PDF)

$$f(x) = F'(x) = \frac{d}{dx} P(X \le x) \tag{7}$$

where $F(x)$ is the CDF of a continuous random variable X.

Bernoulli and Binomial Discrete Distributions

A random variable X has a *Bernoulli distribution* with parameter p if $P(X = 1) = p$, and $P(X = 0) = 1 - p$, where $0 < p < 1$.

We write this as

$$X \sim Bern(p) \tag{8}$$

[1] Excluding endpoints.

where the symbol $\sim$ is read as "is distributed as". An experiment that can result in either a "success" or a "failure" but not *both* is a called a *Bernoulli trial.* A Bernoulli random variable is the indicator of a success or failure.

Suppose that n independent Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the *Binomial distribution* with parameters n and p. We write

$$X \sim Bin(n, p) \tag{9}$$

when a random variable X has a *Binomial distribution.* If $X \sim Bin(n, p)$, the PMF of X is

$$P(X = k) = \binom{n}{k} p^k (1-p)^k = \binom{n}{k} p^k (1-p)^{n-k}. \tag{10}$$

Beta Continuous Distribution

A random variable X has a *Beta distribution* with parameters a and b, where $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\beta(a, b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1, \tag{11}$$

where the constant $\beta(a, b)$ is chosen to make the PDF integrate to 1. We write this as

$$X \sim Beta(a, b). \tag{12}$$

Bayesian Inference

Suppose we have a coin that lands heads with probability p, but we do not know the value of p. Our goal is to *infer* the value of p after observing n coin tosses. Ideally, the more observations we make, the better our estimate.

Bayesian inference treats all unknown quantities as random variables. Therefore, in the above described problem, we treat p as a *random variable* and give it a *distribution.* The distribution we give to p reflects our uncertainty about the true value of p before we observe any coin tosses. We call this distribution, the *prior distribution.* After observing an experiment, and the data from the experiment have been collected, the prior distribution is updated using Bayes' rule. This yields the *posterior* distribution that now reflects our new beliefs about p.

Beta-Gamma Conjugacy

Suppose our *prior belief* about the true value of p, the probability a coin lands heads, is described by the *beta distribution*

$$p \sim Beta(a, b) \tag{13}$$

where a and b are *known constants.*

Next, let X be the number of heads in n coin tosses. Knowing the true value of p means that each realization of X is an independent Bernoulli trial with a p probability of success

$$X|p \sim Bin(n, p) \tag{14}$$

Note, X is not *marginally* Binomial, it is *conditionally* Binomial. Its marginal distribution is called the *Beta-Binomial distribution.*

We can update our belief, after observing data, using Bayes' rule (in hybrid form) in exactly the same way we did in the *biased coin problem.* Letting $f(p)$ be the prior distribution,[2] and $f(p|X = k)$ be the *posterior distribution* after observing k heads

$$f\left(p|X = k\right) = \frac{P\left(X = k|p\right) \cdot f(p)}{P(X = k)} \tag{15}$$

Using the definitions of the Beta and Binomial distributions from the previous sections,

$$f\left(p|X = k\right) = \frac{\binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{1}{\beta(a,b)} x^{a-1} (1-x)^{b-1}}{P(X = k)} \tag{16}$$

The denominator, the *marginal distribution* of X, is obtained by integrating over the support of the *conditional distribution* (equivalently the joint distribution by the definition of conditional probability)

$$P(X = k) = \int_0^1 P\left(X = k|p\right) f(p) dp = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p) dp. \tag{17}$$

For $a = b = 1$, and

$$p \sim Unif(0, 1) \tag{18}$$

so that $f(p) = \frac{x}{b-a}$, it can be shown that $P(X = k)$ has a *Discrete Uniform distribution.* For general a and b the problem seems difficult.

The posterior distribution $f(p|X = k)$ is a function of p, i.e.,

$$f\left(p|X = k\right) = g(p) \tag{19}$$

which means that everything that does not depend on p is a constant. We can therefore drop all of the constant terms in the expression and determine the normalizing constant to be whatever is needed to make the PDF integrate to 1. This gives

$$f\left(p|X = k\right) \propto p^{a+k+1} \cdot (1-p)^{b+n-k-1}, \tag{20}$$

[2]Note how we have switched "probability" for "distribution" compared to the previous session's definition. It is not trivial to prove, but it can be shown that a general rule applies.

which is the $Beta(a+k, b+n-k)$ PDF, up to a multiplicative constant!

This special relationship between the Beta and Binomial distributions is known as *conjugacy*. Beta is the conjugate prior of the Binomial, which means that if we have a Beta prior distribution on p and the data are conditionally Binomial given p, when going from prior to posterior, we do not leave the family of Beta distributions.

Example

Suppose our prior belief about the true value of p, the probability a coin lands heads, is described by the prior $Beta(1,1)$, which is equivalent to a $Unif(0,1)$ distribution. This assigns an equal probability to any value of $0 < p < 1$. In words, we don't assume anything about p since any $p \in [0,1]$ is equally likely.[3] Note, this doesn't mean the prior is *uninformative*, knowing the probability is equally likely means we still know something.

After $n = 5$ coin tosses we observe $k = 5$ Heads. The posterior then is $Beta(6,1)$ plotted in Figure **??**

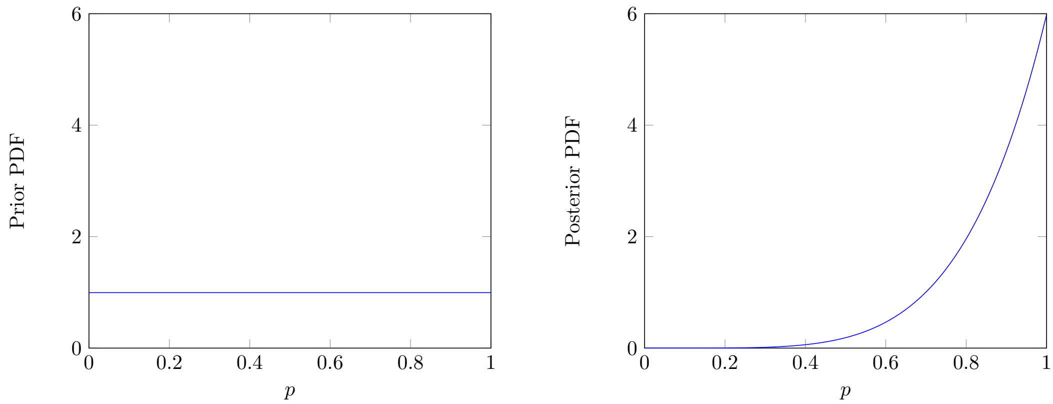

Figure 2: Uniform prior (left) and posterior $Beta(6,1)$ of bias in coin after 5 heads in 5 tosses are observed.

This model is the continuous analogue of the *biased coin problem* from the conditional probability session. In the biased coin problem, we also had a coin with probability of heads p that was unknown, but our prior information led us to believe that p could only take on one of two values: $1/2$ or $3/4$. The prior distribution, although we did not call it that, was discrete

$$P\left(p = \frac{1}{2}\right) = 1/2 \tag{21}$$

[3]Technically, the probability p falls in an equal length disjoint partition of the interval where the probability is proportional to the length of the subset.

$$P\left(p = \frac{3}{4}\right) = 1/2 \tag{22}$$

After observing 3 heads, we obtained the posterior PMF that assigned 0.23 to $p = 1/2$ and 0.77 to $p = 3/4$. The same logic applies here, only p can take on any value between 0 and 1.

Practice

Biased Coins

1) Draw the prior and posterior PMFs for the biased coin problem in its discrete form.

Clinical Trials

A new treatment has just been developed for a disease. A clinical trial is about to be conducted to study how effective the treatment is. The treatment will be applied to n patients who have the disease. Given p, the patients' outcomes are independent, with each patient having probability p of being cured by the treatment. But p is unknown. To quantify our uncertainty about p, we model p as a random variable, with prior distribution $p \sim Unif(0, 1)$.

1) Find the probability that exactly k out of the n patients will be cured by the treatment (unconditionally, *not* given p).

2) Suppose the treatment is extremely effective in the clinical trial: all n patients are cured! Given this information, find the probability that p exceeds 1/2.