Chapter 7. Statistical Estimation

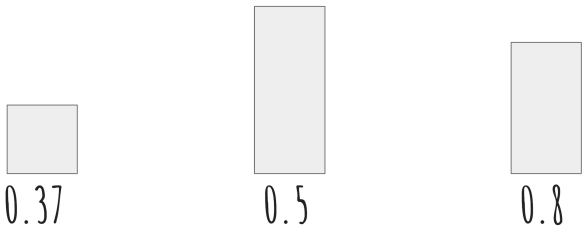
7.4: The Beta and Dirichlet Distributions

Slides (Google Drive) Alex Tsun Video (YouTube)

We'll take a quick break after learning two ways (MLE and MoM) to estimate unknown parameters! In the next section, we'll learn yet another approach. But that approach requires us to learn at least one other distribution, the Beta distribution, which will be the focus of this section.

7.4.1 The Beta Random Variable

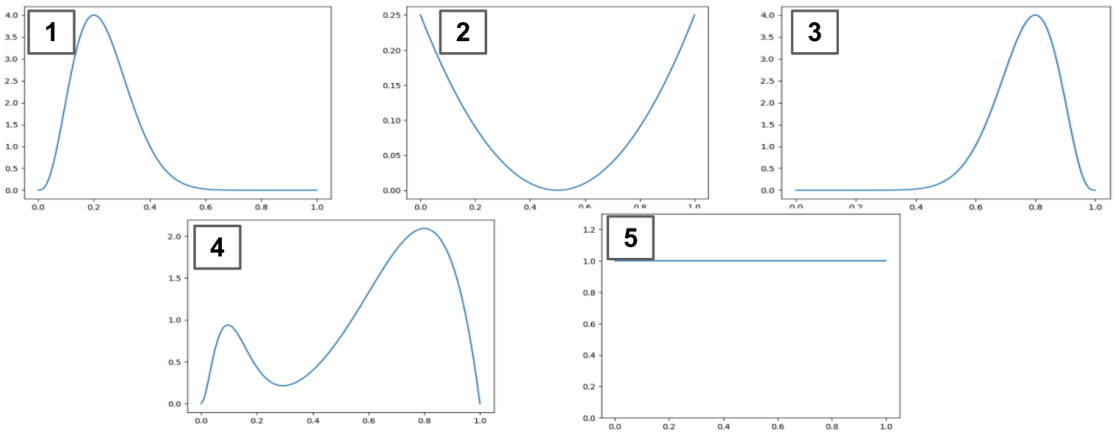
Suppose you want to model your belief on the unknown probability X of heads. You could assign, for example, a probability distribution as follows:



This figure below shows that you **believe** that $X = \mathbb{P}(\text{head})$ is most likely to be 0.5, somewhat likely to be 0.8, and least likely to be 0.37. That is, X is a *discrete* random variable with range $\Omega_X = \{0.37, 0.5, 0.8\}$ and $p_X(0.37) + p_X(0.5) + p_X(0.8) = 1$. This is a probability distribution on a probability of heads!

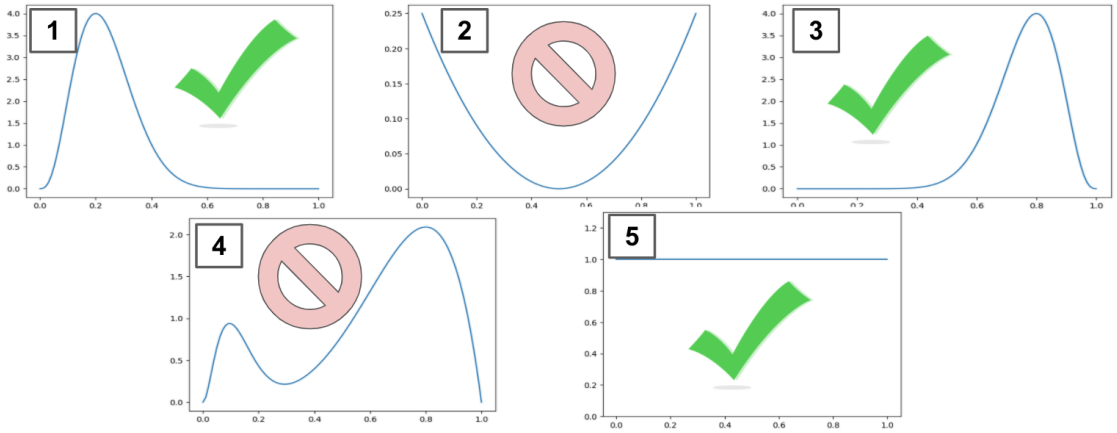
Now what if we want $\mathbb{P}(\text{head})$ to be open to any value in $[0, 1]$ (which we should want; having it be just one of three values is arbitrary and unrealistic)? The answer is that we need a **continuous** random variable (with range $[0, 1]$ because probabilities can be any number within this range)! Let's try to see how we might define a new distribution which might do a good job modelling this belief! Let's see which of the following shapes might be appropriate (or not).

Example(s)



Suppose you flipped the coin n times and observed k heads. Which of the above density functions have a "shape" which would be *reasonable* to model your belief?

Solution Here is the answer:



It's important to note that Distributions 2 and 4 are **invalid**, because there is no possible sequence of flips that could result in the belief that is "bi-modal" (have two peaks in the graph of the distribution). Your belief should have a single peak at your highest belief, and go down on both sides from there.

For instance, if you believe that the probability of (getting heads) is most likely around 0.25, we have Distribution 1 in the figure above. Similarly, if you think that it's most likely around 0.85, we have Distribution 3. Or, more interestingly, if you have NO idea what the probability might be and you want to make every probability equally likely, you could use a **Uniform distribution** like in Distribution 5.

□

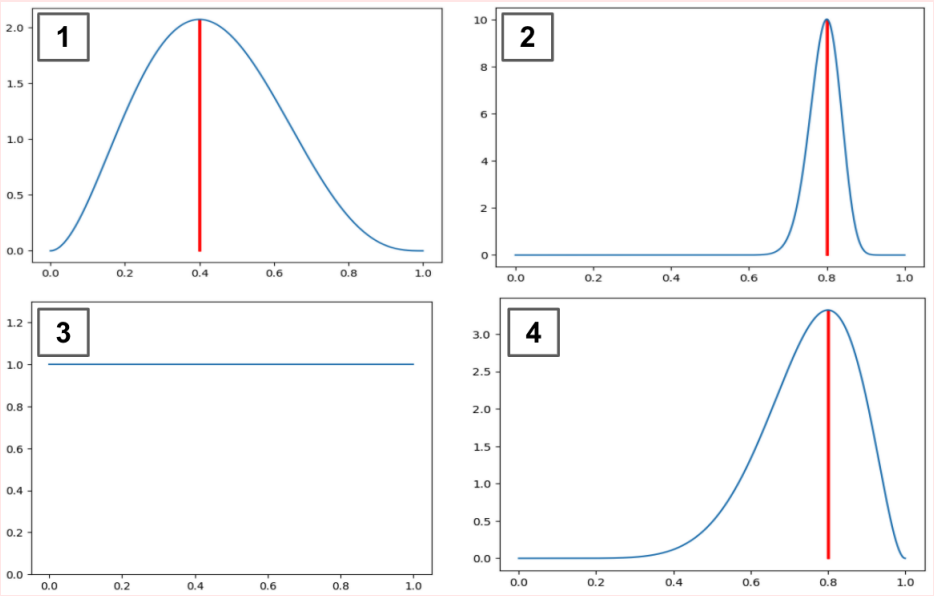
Let's have some practice with concrete numbers now.

Example(s)

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

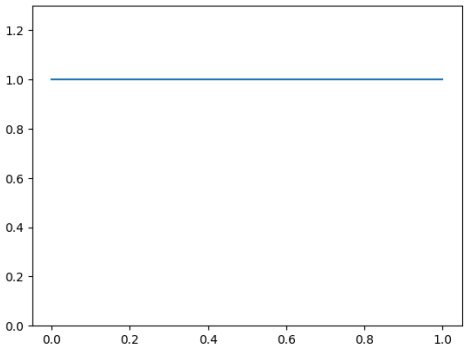
- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?

Match the four distributions below to the four scenarios above. Note the vertical bars in each distribution represents where the **mode** (the point with highest density) is, as that's probably what we want to estimate as our probability of heads!



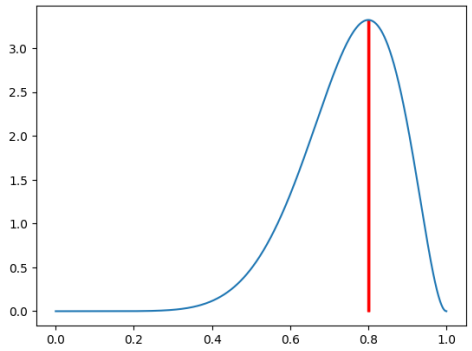
Solution

- You didn't observe anything? **Answer:** Distribution 3.



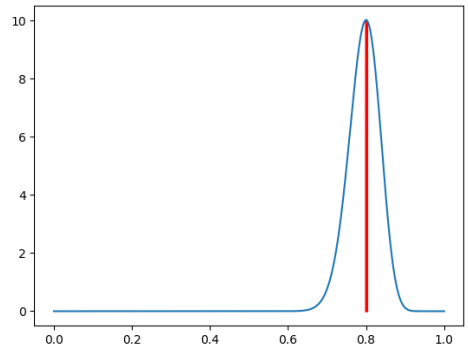
Explanation: Since we haven't observed anything yet, we shouldn't have preference over any particular value. This is encoded as a continuous Unif$(0, 1)$ distribution.

- You observed 8 heads and 2 tails? **Answer:** Distribution 4.



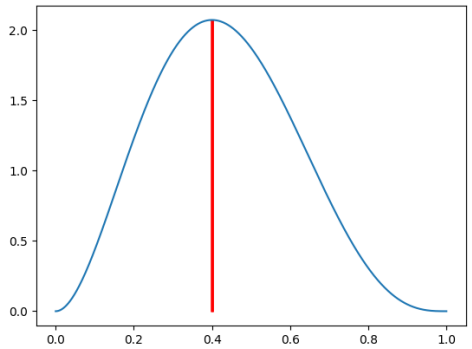
Explanation: We expect $\mathbb{P}$(head) to be around $\frac{8}{8+2} = 0.8$, so either distribution 2 or 4 is reasonable. BUT we have much **uncertainty** (since we only flipped it 10 times) so we have a wider distribution. Note that 0.8 is the MODE, not the mean.

- You observed 80 heads and 20 tails? **Answer:** Distribution 2.



Explanation: We expect $\mathbb{P}$(head) to be around $\frac{80}{80+20} = 0.8$ again, but now since we have way more flips, we can be more certain that the probability is more likely to be 0.8 (thus the "spread" is smaller than the previous).

- You observed 2 heads and 3 tails? **Answer:** Distribution 1.



Explanation: We expect $\mathbb{P}(\text{head})$ to be around $\frac{2}{2+3} = 0.4$, but since 5 flips are rather limited, we have much uncertainty in the actual distribution, therefore the "spread" is quite large!

□

There is a continuous distribution/rv with range $[0, 1]$ that parametrizes probability distributions over a probability just like this, based on two parameters α and β, which allow you to account for how many heads and tails you've seen!

Definition 7.4.1: Beta RV

$X \sim \text{Beta}(\alpha, \beta)$, if and only if X has the following density function (and range $\Omega_X = [0, 1]$):

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we've seen $\alpha-1$ successes and $\beta-1$ failures. Hence the mode (most likely value of the probability/point with highest density) $\arg\max_{x\in[0,1]} f_X(x)$, is

$$\text{mode}[X] = \frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

Also note the following:

- The first term in the pdf, $\frac{1}{B(\alpha,\beta)}$ is just a normalizing constant (ensures the pdf to integrates to 1). It is called the Beta function, and so our random variable is called a Beta random variable.
- There is an annoying "off-by-1" issue: ($\alpha - 1$ heads and $\beta - 1$ tails), so when choosing these parameters, be careful (examples below)!
- x is the probability of success, and $(1 - x)$ is the probability of failure.

7.4.2 Beta Random Variable Examples

Example(s)

If you flip a coin with unknown probability of heads X, identify the parameters of the most appropriate Beta distribution to model your belief:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?

Solution

- You didn't observe anything? $\text{Beta}(0+1, 0+1) \equiv \text{Beta}(1,1) \equiv \text{Unif}(0,1) \rightarrow$ NO mode (because it follows the Uniform distribution; every point has same density).
- You observed 8 heads and 2 tails? $\text{Beta}(8+1, 2+1) \equiv \text{Beta}(9,3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$
- You observed 80 heads and 20 tails? $\text{Beta}(80+1, 20+1) \equiv \text{Beta}(81,21) \rightarrow \text{mode} = \frac{(81-1)}{(81-1)+(21-1)} = \frac{80}{100}$
- You observed 2 heads and 3 tails? $\text{Beta}(2+1, 3+1) \equiv \text{Beta}(3,4) \rightarrow \text{mode} = \frac{(3-1)}{(3-1)+(4-1)} = \frac{2}{5}$

Note all the off-by-1's in the parameters! □

7.4.3 The Dirichlet Random Vector

The Dirichlet random vector generalizes the Beta random variable to having a belief distribution over $p_1, p_2, \ldots, p_r$ (like in the multinomial distribution so $\sum p_i = 1$), and has r parameters $\alpha_1, \alpha_2, \ldots, \alpha_r$. It has the similar interpretation of pretending you've seen $\alpha_i - 1$ outcomes of type i.

Definition 7.4.2: Dirichlet RV

$X \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_r)$, if and only if X has the following density function:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{r} x_i^{a_i - 1}, & x_i \in (0,1) \text{ and } \sum_{i=1}^{r} x_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

This is a generalization of the Beta random variable from 2 outcomes to r. The random vector X is typically the belief distribution about some unknown probabilities of the different outcomes, where we pretend we saw $a_1 - 1$ outcomes of type 1, $a_2 - 1$ outcomes of type 2, ..., and $a_r - 1$ outcomes of type r . Hence, the mode of the distribution is the vector, $\arg \max_{x \in [0,1]^d \text{ and } \sum x_i = 1} f_{\mathbf{X}}(\mathbf{x})$, is

$$\text{mode}[\mathbf{X}] = \left(\frac{\alpha_1 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \frac{\alpha_2 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \ldots, \frac{\alpha_r - 1}{\sum_{i=1}^{r}(a_i - 1)} \right)$$

Also note the following:

- Similar to the Beta RV, the first term in the pdf, $\frac{1}{B(\alpha)}$ is just a normalizing constant (ensures the pdf integrates to 1), where $\alpha = (\alpha_1, \ldots, \alpha_r)$.
- Notice that this is the probability distribution over the random vector x_i's, which is the vector of probabilities, so they must sum to 1 ($\sum_{i=1}^{r} x_i = 1$).