From Beta Distribution and its Intuitions/Applications to Conjugate Distributions

References:

1. Zhihu Question “如何通俗理解 beta 分布”, https://www.zhihu.com/question/30269898
2. Bilibili video “Beta and Dirichlet Distributions”, https://www.bilibili.com/video/BV1JD4y1m7TP

Pre-requisites:

1. Knowing the Bayes theorem;
2. It is recommended that readers have gone through elementary training about statistics.

Written by Tongyu Lu, March 14, 2021

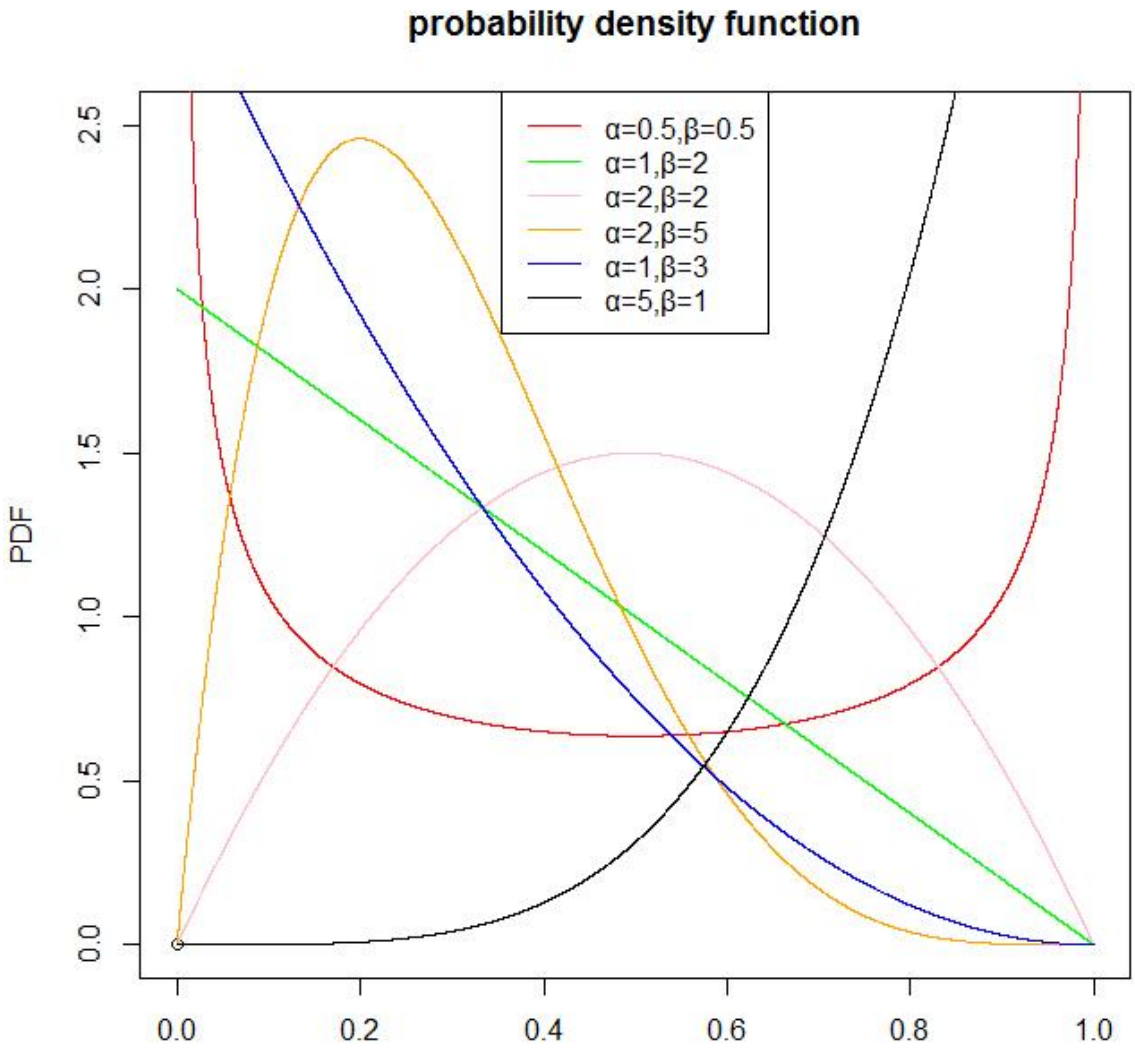
Contents:

Beta Distribution

Introduction

- Beta distribution is over real values on $[0, 1]$
- Therefore, Beta distribution is useful for modeling percentages and proportions
- The pdf is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$

- Observe that Bernoulli distribution pmf: $P(X = x|p) = p^x(1-p)^{1-x}$, $x = 0, 1$; therefore, Beta distribution could be regarded as “a continuous version of Bernoulli”.
- Beta distribution pdf examples:



Definition of Beta Distribution

Beta distribution pdf is

$$p(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

The expectation is $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$

Intuitions of this pdf:

- When $\beta = 1$, then the distribution is exponential (and actually $\alpha = 5$)
- When $\beta = \alpha$, then the distribution is symmetric around $x = 0.5$

Just skip the Gamma function, because it just serves as normalization term.
Note that $\alpha, \beta > 0$ and $x \in [0, 1]$

Dirichlet Distribution

Introduction

- Dirichlet distribution is a generalization of Beta distribution for multiple random variables (RVs).
- Actually, Dirichlet distribution is tailored for vectors whose components are all in $[0, 1]$ and the sum of its components is 1.

The Dirichlet distribution pdf is defined as such: $p(x|\alpha) = \frac{\Pi_{k=1}^{K}\Gamma(\alpha[k])}{\Gamma(\sum_{k=1}^{K}\alpha[k])}\Pi_{k=1}^{K}x[k]^{\alpha[k]-1}$, where $\alpha \in \mathbf{R}_{+}^{K}$

- Example: if $\alpha = [1, 1, 1]$, then the Dirichlet distribution is uniform within triangular space $x_1 + x_2 + x_3 = 1, x_i \geq 0$
- Dirichlet distribution could be regarded as a continuous version of multinomial distribution.

Intuitions of Beta Distribution

To understand the user cases of Beta distribution, the key might come from that the RV X is restricted betwee 0 and 1.
Think: what kind of values situate between 0 and 1?

- Probabilities
- Proportions x/y when $x \geq 0, y > 0$ and $x \leq y$
- Percentages etc.
  Therefore, way could say that Beta distribution may come into use when we want to model an RV which represents probabilities or proportions.

Beta Distribution could Model Rates

For example, if we want to model the distribution of winning rate (maybe we are interested in a game player), we just need to estimate α, β for it. A handy method is that: assume that player has played for 300 games, where 81 of them wins. Then, we could say that the distribution for the winning rate x is $Beta(81 + 1, 219 + 1)$. When that player played another 300 games, where 100 of them wins, then we say that the distribution for the winning rate x becomes $Beta(81 + 100 + 1, 219 + 200 + 1)$.

Why saying that α is the number of winning plus one? I will reveal this in the next part.

Beta Distribution could Model Probabilities and Parameters

Suppose we are doing a coin flipping game. The probability of "head top" is θ. But we do not know θ exactly (maybe the coin is fake, we cannot say that $\theta = 0.5$). How to estimate it?

First, we model this problem in statistic language: RVs X_k are subject to a Bernoulli distribution $B(\theta)$. We want to estimate the distribution parameter θ.

An easy method is: we do $A + B$ games, where we have A head tops. Then we may say that $\theta = A/(A + B)$.
Well, of course that is OK. Because that is indeed an unbiased estimation for θ (which is a result of maximum-likelihood estimation). However, are we 100% certain about that? What if $\theta = A/(A + B) - 0.0001$?

Obviously, we cannot say that θ is exactly $A/(A + B)$. We can only say that $A/(A + B)$ is a likely value of θ. But how confident are we on that estimation? Here is where beta distribution is of use.

Return to our experiments. If $Y_{A+B} = \sum_{k=1}^{A+B} X_k$ denotes the number of "head top" cases after we have done $A + B$ experiments, we could say that $P(Y_{A+B} = A|\theta) = C_{A+B}^{A}\theta^A(1 - \theta)^B$ (which is the likelihood). This likelihood is a function of θ.
Define $l_{A,B}(\theta) = P(Y_{A+B} = A|\theta) = C_{A+B}^{A}\theta^A(1 - \theta)^B$.
We want the distribution of θ. Therefore, what we want is actually $p(\theta|Y_{A+B} = A)$.

Solution:

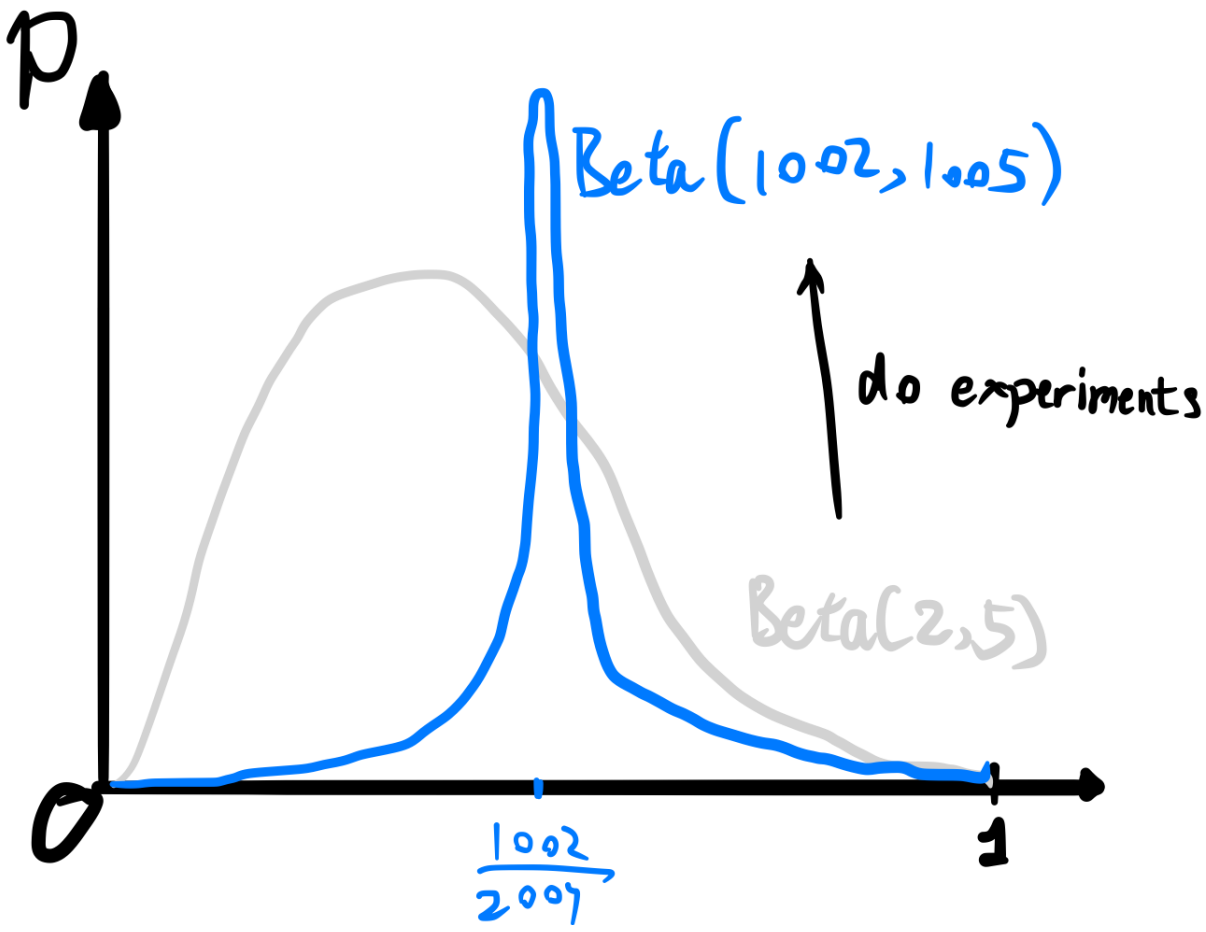
1. Plug in Bayes's rule: $p(\theta|Y_{A+B} = A) = P(Y_{A+B} = A|\theta)p(\theta)/P(Y_{A+B} = A)$.
2. The term $p(\theta)$ is our prior on θ. If we have no idea about the coin, we may say that $p(\theta) = 1, \theta \in [0, 1]$. Generally, we could assume that the prior distribution of θ is a Beta distribution $Beta(\alpha_0, \beta_0)$. In the "I do not know anything about the coin" case, $\alpha_0 = \beta_0 = 1$
3. The term $P(Y_{A+B} = A)$ is a constant. Actually, we could do integral over θ to get this constant: $P(Y_{A+B} = A) = \int_0^1 l_{A,B}(\theta)d\theta = \int_0^1 C_{A+B}^{A}\theta^A(1 - \theta)^B d\theta$
4. Summing up the discussions, we have:

$$p(\theta|Y_{A+B} = A) = \frac{C_{A+B}^{A}\theta^A(1 - \theta)^B\theta^{\alpha_0-1}(1 - \theta)^{\beta_0-1}}{B(\alpha_0, \beta_0)\int_0^1 C_{A+B}^{A}\theta^A(1 - \theta)^B d\theta} = \frac{\theta^{A+\alpha_0-1}(1 - \theta)^{B+\beta_0-1}}{B(\alpha_0, \beta_0)\int_0^1 \theta^A(1 - \theta)^B d\theta}$$

5. Skipping the "Beta\Gamma calculus", we conclude that $p(\theta|Y_{A+B} = A) = \frac{1}{B(A+\alpha_0, B+\beta_0)}\theta^{A+\alpha_0-1}(1 - \theta)^{B+\beta_0-1}$

Oh! The magical thing happens: after doing experiments, our belief of parameter θ changes from $Beta(\alpha_0, \beta_0)$ to $Beta(A + \alpha_0, B + \beta_0)$.

For example, if we firmly believe that the coin is biased before experiments (e.g., $\alpha_0 = 2, \beta_0 = 5$), then after getting results $A = 1000, B = 1000$, we could say that we now believe the coin is nearly unbiased because its posterior distribution is $Beta(1002, 1005)$.



Intuitions behind Conjugate Distribution

In the coin flipping game, we have two main RVs: 1. the flipping results X_k; 2. the parameter θ.
Their relationship: X_k is subject to $B(\theta)$, θ is subject to prior $Beta(\alpha_0, \beta_0)$.
After experiments, we could calculate the posterior of θ, which is magically $Beta(A+\alpha_0, B+\beta_0)$.
This magical Beta prior distribution (which remains Beta as posterior) of a Bernoulli RV is called "the conjugate distribution of Bernoulli distribution".

Generally, we have a RV whose distribution $p_X(x|\Theta)$ is controlled by parameter Θ whose prior distribution is $p_\Theta(\theta|\alpha)$, where α controls the prior distribution.
After observing a bunch of data $X_{1:K} = [X_k, k = 1, \dots, K]$, we may have a posterior distribution $p_{\Theta|X_{1:K}}(\theta) = p_\Theta(\theta|\alpha)p_X(X_{1:K}|\theta)/p(X_{1:K})$. Here comes the definition of conjugate distribution:

if there exists $\hat{\alpha}$, such that posterior distribution $p_{\Theta|X_{1:K}}(\theta) = p_\Theta(\theta|\hat{\alpha})$, then we say that the density function p_Θ defines a conjugate distribution of p_X.

In other words, "$p_\Theta(\theta|\alpha)$ and $p_X(x|\Theta)$ are conjugate" if and only if "given any prior parameter α and observation $X_{1:K}$, there exists parameter $\hat{\alpha}$ and constant c, such that:
$p_\Theta(\theta|\hat{\alpha}) = cp_\Theta(\theta|\alpha)p_X(X_{1:K}|\theta), \forall\theta$".

In the Bernoulli-Beta case, given prior parameter α_0, β_0, there exists parameter $\hat{\alpha} = A + \alpha_0, \hat{\beta} = B + \beta_0$ and constant $c = (\int_0^1 C_{A+B}^A \theta^A (1-\theta)^B d\theta)^{-1}$ such that
$p_{\Theta|X_{1:K}}(\theta) = p_\Theta(\theta|\hat{\alpha}) = cp_\Theta(\theta|\alpha)p_X(X_{1:K}|\theta), \forall\theta$

Not every distribution has conjugate distribution. But we could say that all “exponential family distributions” have conjugate distribution. What on earth are “exponential family distributions”? We leave this topic to further discussions.

Comments on the pdf Notations

Normally, if I do not add footnotes on pdf (e.g. $p(X|\Theta)$, $p(\theta|Y_{A+B}=A)$), I treat the operation $p(\cdot|\cdot)$ as a functional which operates on RVs.
For example,

- $p(X|\Theta)$ is a functional on RVs X and Θ, where RV Θ serves as prior for RV X;
- $p(\theta|Y_{A+B}=A)$ is a functional on RVs θ, where RV Y_{A+B} is given as A and serves as prior for RV θ.
- Every “given(determined) RV” is marked in the “$X=5$” form; “given RV” is a constant actually.

If I add footnotes on pdf (e.g. $p_X(x|\Theta)$, $p_\Theta(\theta|\alpha)$, $p_{\Theta|X_{1:K}}(\theta)$), I treat the operation $p_\_(\cdot|\cdot)$ as a function, whose form is defined by notation $p_\_$.
For example,

- $p_X(x|\Theta)$ is a density function p_X which operates on variable x and this function is a random function controlled by RV Θ;
- $p_\Theta(\theta|\alpha)$ is a density function p_Θ which operates on variable θ and is controlled by a determined parameter α;
- $p_{\Theta|X_{1:K}}(\theta)$ is a density function $p_{\Theta|X_{1:K}}$ which operates on variable θ.

Sorry that in the conjugate function section I denote Θ as RV while denote θ as variable, but in the Beta distribution section I denoted θ as RV. Hope that you could understand.