# Beta Estimators

Vasco Villas-Boas

September 5, 2022

## BETA DISTRIBUTION

The *Beta distribution* is a continuous distribution defined on an open or closed finite real interval. If $X$ has a beta distribution with shape parameters $\alpha, \beta > 0$, we say that $X$ has probability density function (pdf) and expectation:

$$f_X(x) := \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{1}$$

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{2}$$

where $B(\alpha,\beta) := \frac{\Gamma(\alpha)(\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(\cdot)$ is the gamma function ($B(\cdot,\cdot)$ serves as a normalizing constant so that the density $f_x$ integrates to 1)[1]. The Beta distribution is useful to model the probability of probabilities since it's support is $[0,1]$ and since it's the conjugate prior for the Bernoulli, Binomial, and Geometric distributions. That means if we have a beta prior on some parameter $p$ and observe an outcome of a Bernoulli, Binomial, or Geometric distributions that have probability parameter $p$, the posterior distribution of $p$ given the prior and a Bayesian update is also a beta distribution. Conjugate priors are useful because they save us expensive and potentially impossible computation in computing posteriors since we know the posterior follows the same distribution as the prior.

The Beta distribution has a straightforward and interpretable Bayesian update. Suppose that we observe $n$ $iid$ outcomes $X_1, ...X_n$ where each $X_i \sim$ Bernoulli$(p)$. For ease of notation, define $S_n := \sum_{i=1}^{n} X_i$. We have a prior on $p \sim$ Beta$(\alpha,\beta)$ for some shape parameters $\alpha, \beta > 0$. Our posterior distribution on $p$ follows $p|X_1,...,X_n \sim$ Beta$(\alpha + S_n, \beta + (n - S_n))$. The posterior distribution on $p$ is very simple to compute, we simply add the number of successes to $\alpha$ and the number of failures to $\beta$ and define a new beta distribution with those parameters.

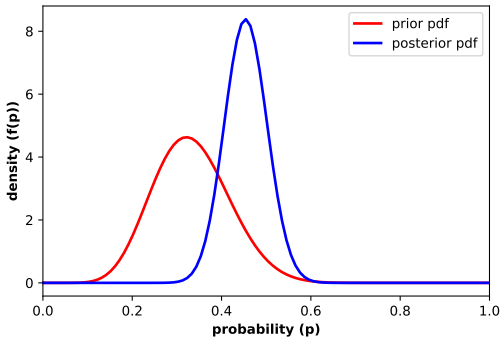


**Figure 1.** Demonstrating Bayesian Update for Beta Prior and Binomial Data

In figure 1, We can see a concrete example the prior and posterior densities when updating a beta distribution after observing Bernoulli outcomes. Suppose $X \sim$ Bernoulli$(p)$ and $p$ has prior $p \sim$ Beta$(\alpha = 10, \beta = 20)$. Suppose we

[1] Although not needed in this article, it's also possible to derive the beta distribution with four parameters: two shape parameters $\alpha, \beta > 0$ and two parameters $a, b$ that determine the interval of support of the beta distribution as $(a,b)$ or $[a,b]$. Suppose that $Y \sim$ Beta$(\alpha,\beta,a,b)$. We can define the linear transformation $Y := a + (c-a)X$ so that $X = \frac{Y-a}{c-a}$ to map the $[0,1]$ beta distribution to the $[a,b]$ beta distribution. Then $f_Y(y) = \frac{1}{c-a} f_X(\frac{y-a}{c-a}) = \frac{(\frac{y-a}{c-a})^{\alpha-1}(1-\frac{y-a}{c-a})^{\beta-1}}{(c-a)B(\alpha,\beta)}$.

observe 80 $iid$ draws $(X_1, ..., X_{80})$ of $X$ where 40 of them are successes and 40 are failures. The posterior distribution for $p$ is $p|X_1, .., X_{80} \sim \text{Beta}(\alpha = 50, \beta = 60)$. Graphically, there are a few nice points to observe in the update. The posterior distribution is tighter than the prior distribution since the $\alpha_{\text{posterior}} > \alpha_{\text{prior}}$ and $\beta_{\text{posterior}} > \beta_{\text{prior}}$- this reflects the fact that with more data, we're more confident about the precise area where the true $p$ lies; the more data points we observe, the tighter the posterior will be. Second, the expectation of the posterior will lie between the expectation of the prior $(1/3)$ and the sample frequency of the data $(1/2)$- we favor the expectation of the prior more if $\alpha_{\text{prior}}, \beta_{\text{prior}}$ are relatively large compared to the amount of data points, and we favor the sample frequency more if we have lots of data points. In other words, if we're confident in our prior we weight our prior heavily, if we receive lots of data, we weight our sample heavily in judging $p$.

### BETA REGRESSION DISTRIBUTION

In the typical Beta distribution $X$ has shape parameters $\alpha, \beta$ and support on [a,b]. In the beta regression distribution, $\alpha, \beta$ are no longer native parameters to $X$, instead $\alpha, \beta$ are functions of regressors so that the shape of the beta distribution can vary with the regressors (it's possible for only one of $\alpha, \beta$ to be a function of regressors too).

It's actually easier to reason the construction of the beta regression model if we re-parametrize the beta distribution and for simplicity fix the interval of support to $[0, 1]$. Let $X \sim \text{Beta}(\alpha, \beta)$ with interval of support on $[0, 1]$ so that $X$ can be interpreted as a prior on some parameter $x$. Let's define the mean of $X$ as $\mu := \frac{\alpha}{\alpha+\beta}$ and the precision of $X$ as $\phi := \alpha + \beta$ like in (Graf 2021). Note that $X \sim \text{Beta}(\mu\phi, (1-\mu)\phi)$ so that $X$ can be equivalently characterized by mean and precision. I call this the beta regression model because I want to make $\mu$, the mean of the beta distribution a function of regressors. Let's define $\mu(x_j) = \sigma(\sum_{j=1}^{k} b_j x_j) = \frac{1}{1+e^{-\sum_{j=1}^{k} b_j x_j}}$. What's happening is that the beta distribution is now natively characterized by $\phi, b_1, ...b_k$ and for characteristics $[x_1, ..., x_k] \in \mathbb{B}^k$, we produce a Beta distribution. The function $\sigma : \mathbb{R}^k \mapsto [0, 1]$, often referred to as the sigmoid function, is called a link function because it "links" from regression parameters ($b$s) to the parameters of the probability distribution ($\mu$). The Beta distribution is a special case of the Beta regression distribution where the link function is just a constant function.

The Beta regression distribution is useful, for one, when we expect the mean of the prior to vary with some characteristics and we want to take that into account in our prior(s). For example, suppose we want to construct a prior distribution for the probability that a bunch of newborns will graduate college. A newborn child coming from Harvard educated upper class parents has very different likelihood of graduating college than a newborn coming from a lower-class family- with one prior distribution these two newborns have the same prior for graduating Harvard. With priors controlling for parental education and income, we can have a continuum of priors for populations of newborns depending on parental income and education.

How do we fit a Beta regression distribution to data using a maximum likelihood approach? Suppose we have a $N$ data points where data point $i$ takes the form: $z_i = (p_i, \vec{x}_i)$ where $p_i$ is a probability for point $i$ that has characteristics $\vec{x}_i = x_{i1}, ..., x_{ik}$. We assume that each $p_i$ is distributed $iid$ from a Beta distribution with density $f_p$ with mean $\mu$ and precision $\phi$ where $\mu(\vec{x}_i) = \sigma(\sum_{j=1}^{k} b_j x_{ij})$. Our objective is to solve the following maximization problem:

$$\begin{aligned}
\max_{b_1,\ldots,b_k,\phi} \mathcal{L}(\text{Data}|b_1, \ldots, b_k, \phi) &= \max_{b_1,\ldots,b_k,\phi} \Pi_{i=1}^{N} f_p(p_i; \mu(\vec{x}_i), \phi) \\
&= \min_{b_1,\ldots,b_k,\phi} -\sum_{i=1}^{N} \log[\frac{\Gamma(\phi) p_i^{\mu(\vec{x}_i)\phi-1}(1-p_i)^{(1-\mu(\vec{x}_i))\phi-1}}{\Gamma(\mu(\vec{x}_i)\phi)\Gamma((1-\mu(\vec{x}_i))\phi)}] \\
&= \min_{b_1,\ldots,b_k,\phi} -\sum_{i=1}^{N} \log\Gamma(\phi) - \log\Gamma(\mu(\vec{x}_i)\phi) - log\Gamma((1-\mu(\vec{x}_i))\phi) + (\mu(\vec{x}_i)\phi - 1)log(p_i) + ((1-\mu(\vec{x}_i))\phi - 1)log(1-p_i)
\end{aligned}$$

We can write the following program to approximate the optimal $b_1, ..., b_k, \phi$ by solving the last minimization problem:

```
import numpy as np
from scipy.special import loggamma
from scipy.optimize import minimize

def calculate_mu_i(X, b):
    return 1 / (1 + np.exp(-(np.dot(X, b))))

def calculate_neg_loglikelihood(params, p, Z):
    b = params[0:-1]
    phi = params[-1]
```

```
    mu_i = calculate_mu_i(Z, b)
    ll = loggamma(phi) - loggamma(mu_i*phi) - loggamma((1 - mu_i)*phi) + (mu_i*phi - 1)*np.log(p) +
        ((1-mu_i)*phi-1)*np.log(1-p)

    return -np.sum(ll)


def calculate_optimal_beta_and_phi():
    # initialize the data for optimization
    p = ... # N dimensional list of probabilities
    Z = ... # N x k matrix of characteristics indexed in the same way as p

    # initialize the parameters for optimization
    params = [1]*(k+1) # The first k dimensions are for b and the last dimension is for phi

    res = minimize(calculate_neg_loglikelihood, x0=params, args=(p,Z), bounds=[(None,None), ..., (
        None,None), (0,None)], options={'ftol': 1e-06, 'gtol': 1e-06}) # phi, precision of beta
        distribution, must be nonnegative

    if res.success:
        return res.x
    else:
        return [float("nan")]*(k+1)
```

## REFERENCES

Francisco Cribari-Neto and Achim Zeileis. Beta Regression in R. *Journal of Statistical Software*, 34(2):1–24, 2010. URL: https://www.jstatsoft.org/index.php/jss/article/view/v034i02.

Christian Graf. A Guide to the Regression of Rates and Proportions, February 18, 2021. URL: https://towardsdatascience.com/a-guide-to-the-regression-of-rates-and-proportions-bcfe1c35344f#:~:text=The%20general%20idea%20of%20the,by%20maximizing%20the%20corresponding%20likelihood.

Aerin Kim. Beta Distribution — Intuition, Examples, and Derivation, January 08, 2020. URL: https://towardsdatascience.com/beta-distribution-intuition-examples-and-derivation-cf00f4db57af.

David Robinson. Understanding the beta distribution (using baseball statistics), December 20, 2014. URL: http://varianceexplained.org/statistics/beta_distribution_and_baseball/.

David Robinson. Understanding beta binomial regression (using baseball statistics), May 31, 2016. URL: http://varianceexplained.org/r/beta_binomial_baseball/.