Beta Distribution — Intuition, Examples, and Derivation

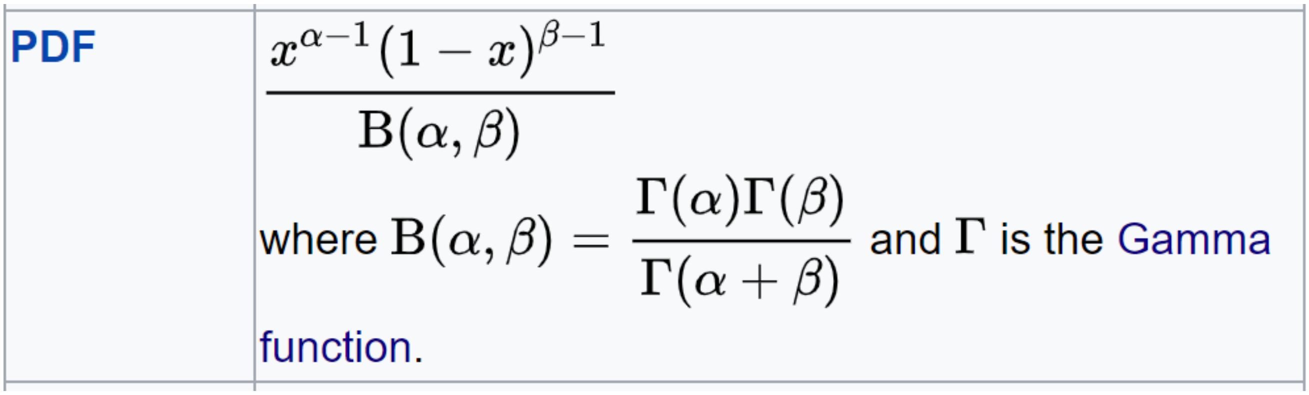
When to use Beta distribution

Aerin Kim Jan 8 · 7 min read ★

The Beta distribution is **a probability distribution** ***on probabilities***. For example, we can use it to model the probabilities: the Click-Through Rate of your advertisement, the conversion rate of customers actually purchasing on your website, how likely readers will clap for your blog, how likely it is that Trump will win a second term, the 5-year survival chance for women with breast cancer, and so on.

Because the Beta distribution models a probability, its domain is bounded between **0** and **1**.

1. Why does the PDF of Beta distribution look the way it does?



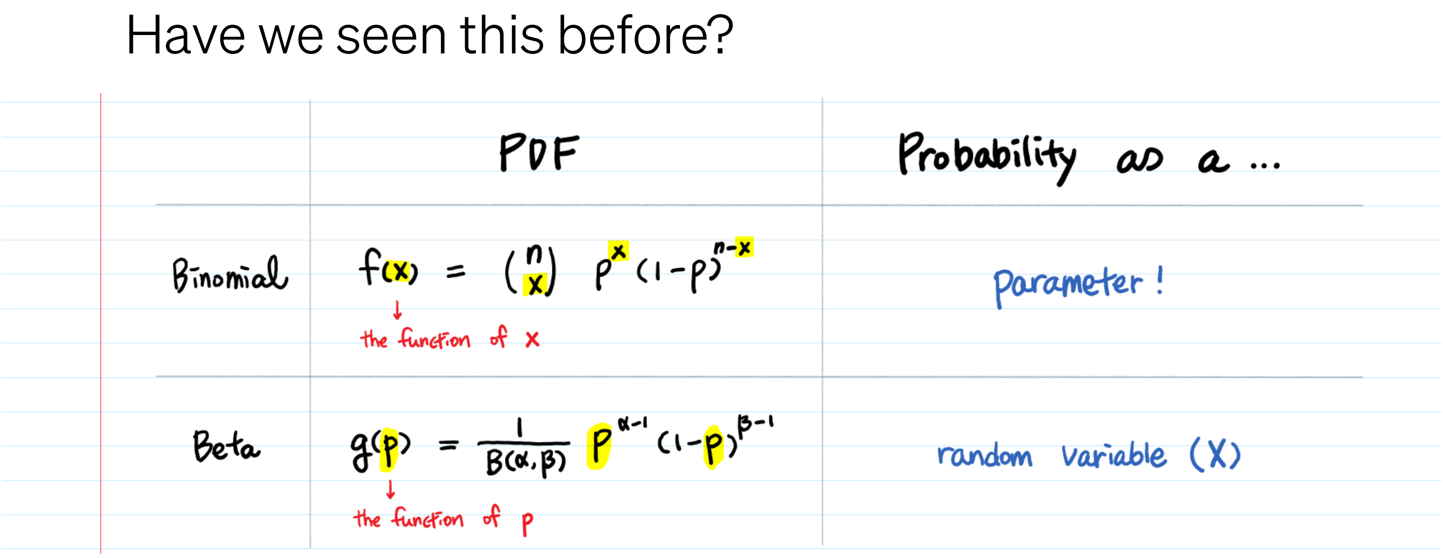
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.

An excerpt from Wikipedia

What’s the intuition?

Let’s ignore the coefficient 1/B(α,β) for a moment and only look at the numerator **x^(α-1) * (1-x)^(β-1),** because **1/B(α,β)** is just a normalizing constant to make the function integrate to 1.

Then, the terms in the numerator — **x to the power of something multiplied by 1-x to the power of something**— look familiar.



Have we seen this before?

	PDF	Probability as a ...
Binomial	$f(x) = \binom{n}{x} p^{x} (1-p)^{n-x}$ ↓ the function of x	Parameter !
Beta	$g(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1} (1-p)^{\beta-1}$ ↓ the function of p	random variable (X)

X ~ Binomial(n, p) vs. X ~ Beta(α, β)

The difference between the binomial and the beta is that **the former models the number of successes (x), while the latter models the probability (p) of success.**

In other words, the probability is a **parameter** in binomial; In the Beta, the probability is a **random variable**.

Interpretation of α, β

You can think of **α-1 as the number of successes** and **β-1 as the number of failures,** just like **n** & **n-x** terms in binomial.

You can choose the α and β parameters however you think they are supposed to be. If you think the probability of success is very high, let's say 90%, **set 90 for α** and **10 for β.** If you think otherwise, 90 for β and 10 for α.

As **α** becomes larger (more successful events), the bulk of the probability distribution will shift towards the right, whereas an increase in **β** moves the distribution towards the left (more failures).

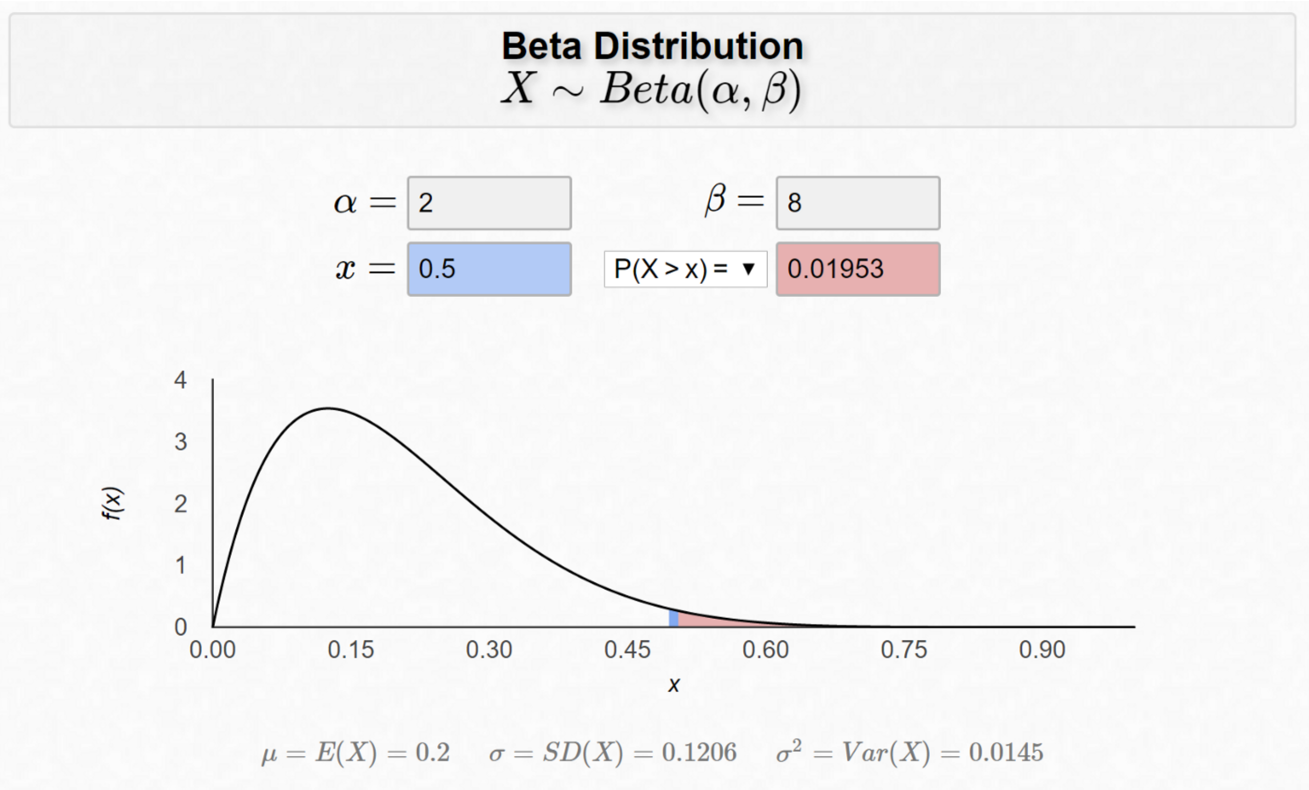
Also, the distribution will narrow if both **α** and **β** increase, for we are more certain.

2. Example: Probability of Probability

Let's say how likely someone would agree to go on a date with you follows a Beta distribution with **α** = 2 and **β** = 8. **What is the probability that your success rate will be greater than 50%?**

P(X>0.5) = **1- CDF(0.5)** = 0.01953

I'm sorry, it's very low. 😢



Beta Distribution

$X \sim Beta(\alpha, \beta)$

$\alpha =$ 2 $\beta =$ 8

$x =$ 0.5 P(X > x) = ▼ 0.01953

$\mu = E(X) = 0.2$ $\sigma = SD(X) = 0.1206$ $\sigma^2 = Var(X) = 0.0145$

Dr. Bognar at the University of Iowa built the calculator for Beta distribution, which I found useful and beautiful. You can experiment with different values of **α** and **β** and visualize how the shape changes.

3. Why do we use the Beta distribution?

If we just want the probability distribution to model the probability, any arbitrary distribution over (0,1) would work. And creating one should be easy. Just take any function that doesn't blow up anywhere between 0 and 1 and stays positive, then integrate it from 0 to 1, and simply divide the function with that result. You just got a probability distribution that can be used to model the probability. In that case, why do we insist on using the beta distribution over the arbitrary probability distribution?

What is so special about the Beta distribution?

If you don't know what the Conjugate Prior or Bayesian Inference is, read first

The Beta distribution is the **conjugate prior** for the Bernoulli, binomial, negative binomial and geometric distributions (seems like those are the distributions that involve success & failure) in Bayesian inference.

Bayesian Inference — Intuition and Implementation

The art of Bayesian Inference lies in how you implement it...

Computing a posterior using a conjugate prior is very convenient, because you can avoid expensive numerical computation involved in Bayesian Inference.

then

Conjugate Prior

There are two things that make the posterior calculation expensive. First, we are computing the posterior for every...

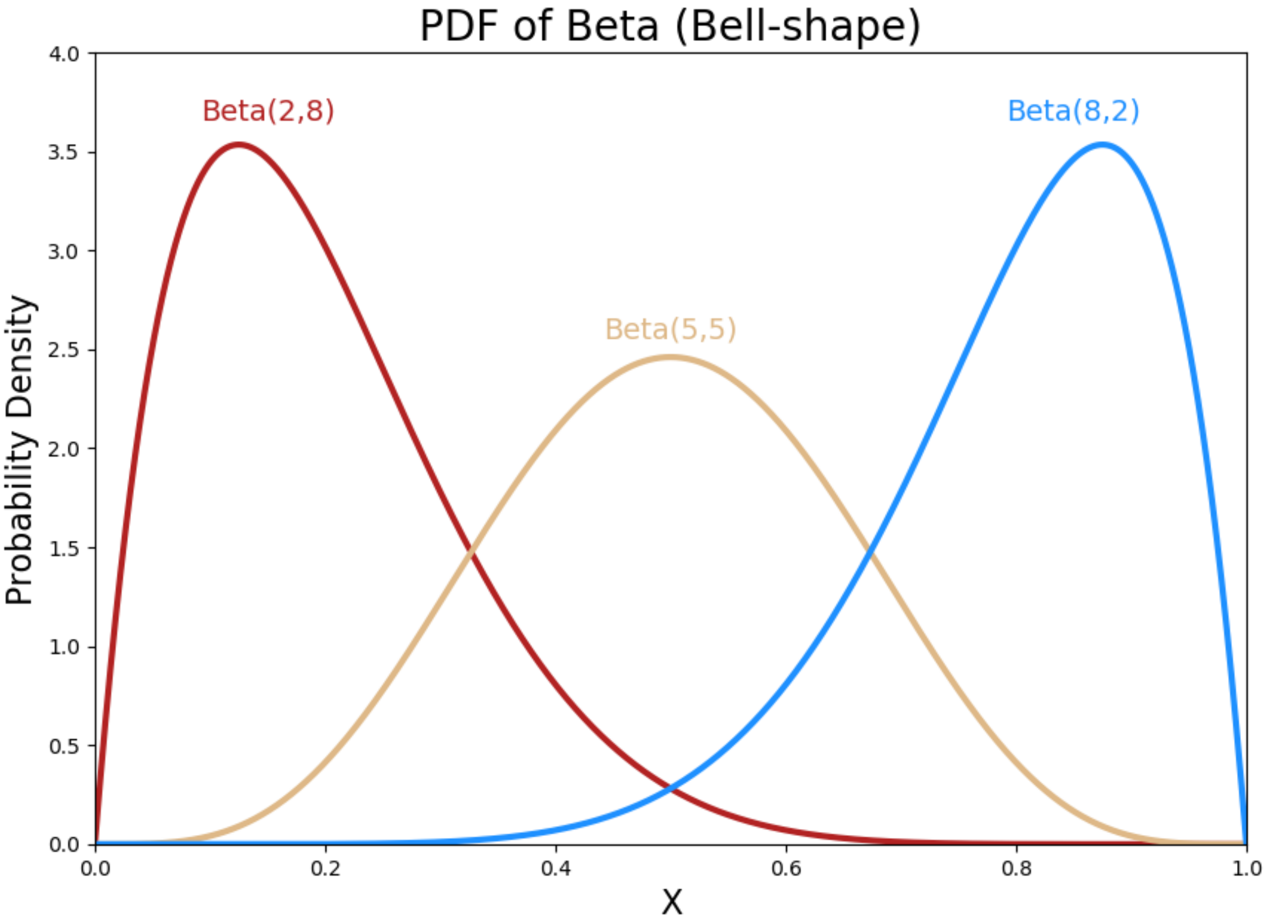
As a data/ML scientist, your model is never complete. You have to update your model as more data come in (and that's why we use Bayesian Inference).
The computation in Bayesian Inference can be very heavy or sometimes even intractable. But if we could use the closed-form formula with the conjugate prior, the computation becomes a piece of cake.

In our date acceptance/rejection example, the beta distribution is a conjugate prior to the binomial likelihood. **If we choose to use the beta distribution as a prior, during the modeling phase, we already know the posterior will also be a beta distribution.** Therefore, after carrying out more experiments (asking more people to go on a date with you), **you can compute the posterior simply by adding the number of acceptances and rejections to the existing parameters α, β respectively**, instead of multiplying the likelihood with the prior distribution**.**

4. Beta distribution is very flexible.

The PDF of Beta distribution can be U-shaped with asymptotic ends, bell-shaped, strictly increasing/decreasing or even straight lines. As you change **α** or **β**, the shape of the distribution changes.

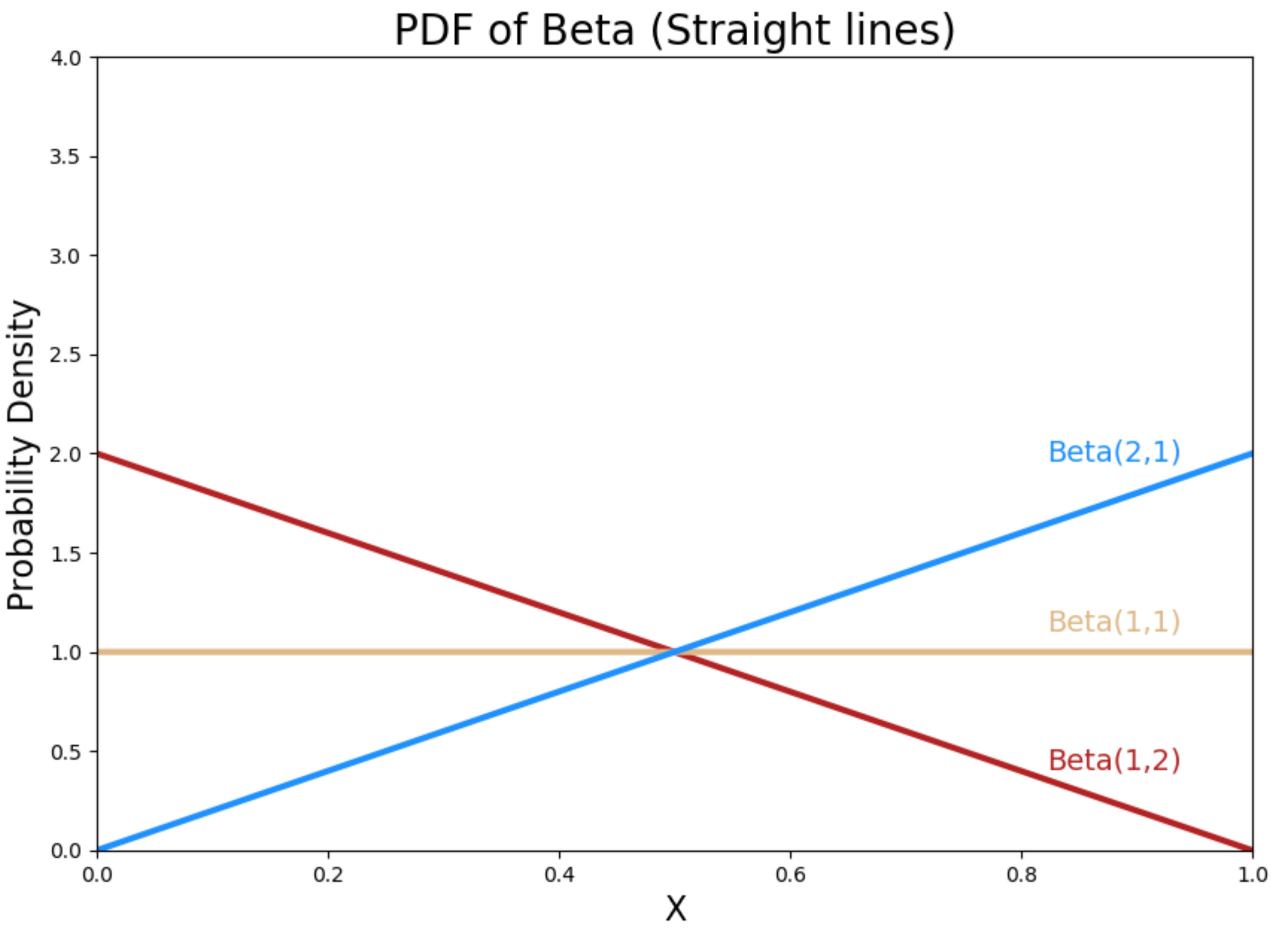
a. Bell-shape



Notice that the graph of PDF with **α** = 8 and **β** = 2 is in blue, not in read. The x-axis is the probability of success.

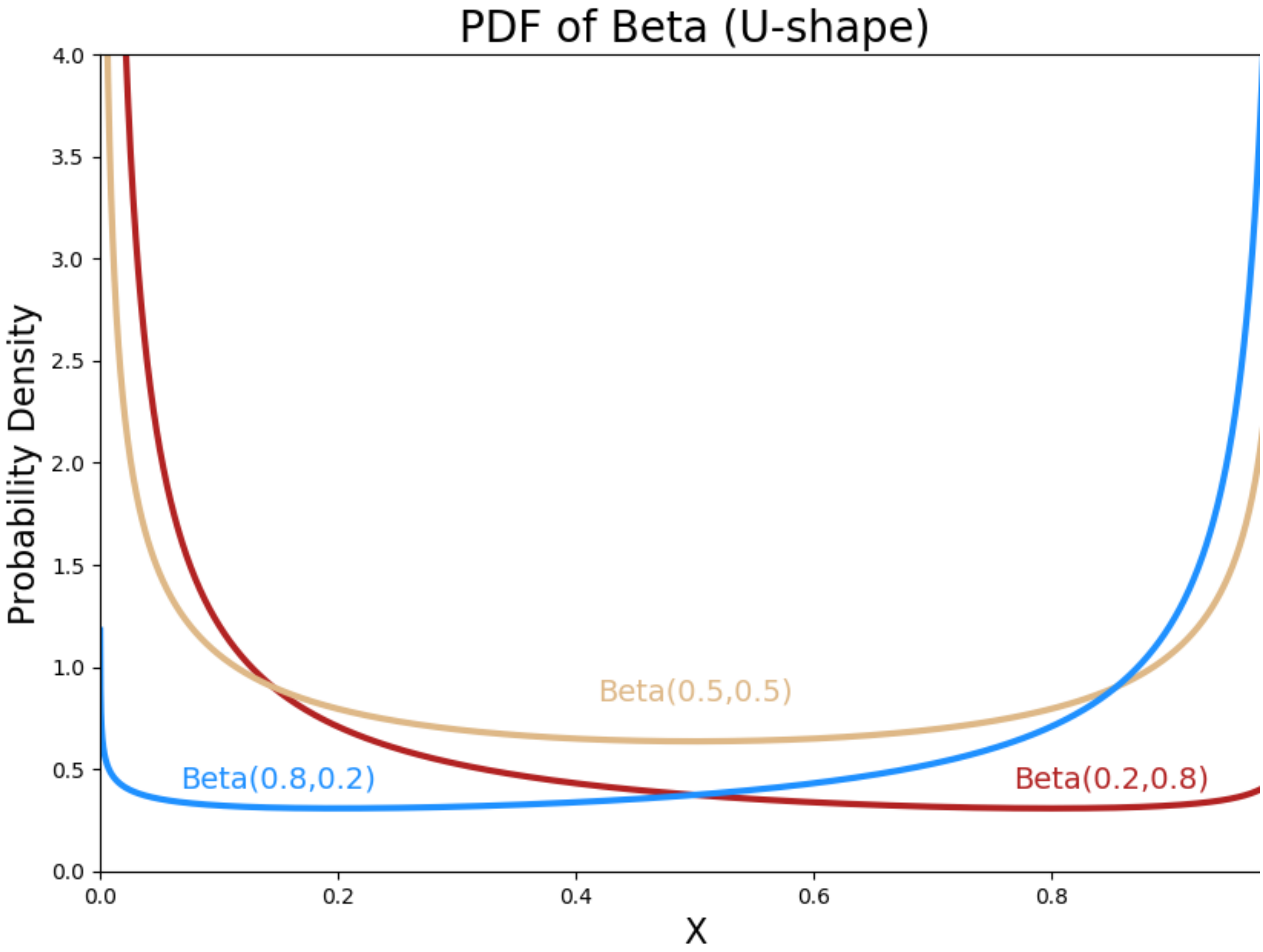
The PDF of a beta distribution is approximately normal if **α** +**β** is large enough and α & β are approximately equal.

b. Straight Lines



The beta PDF can be a straight line too!

c. U-shape



When **α <1, β<1,** the PDF of the Beta is U-shaped.

Below is the code to produce the beautiful graphs above.

```python
import numpy as np
from scipy.stats import beta
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = [10, 7]

# Bell shape
x = np.linspace(0, 1, 10000)
y1 = beta.pdf(x, 2, 8)
y2 = beta.pdf(x, 5, 5)
y3 = beta.pdf(x, 8, 2)

plt.title("PDF of Beta (Bell-shape)", fontsize=20)
plt.xlabel("X", fontsize=16)
plt.ylabel("Probability Density", fontsize=16)
plt.plot(x, y1, linewidth=3, color='firebrick')
plt.annotate("Beta(2,8)", xy=(0.15, 3.7), size = 14, ha='center', va='center', color='f
plt.plot(x, y2, linewidth=3, color='burlywood')
plt.annotate("Beta(5,5)", xy=(0.5, 2.6), size = 14, ha='center', va='center', color='bu
plt.plot(x, y3, linewidth=3, color='dodgerblue')
plt.annotate("Beta(8,2)", xy=(0.85, 3.7), size = 14, ha='center', va='center', color='d
plt.ylim([0, 4])
plt.xlim([0, 1])
plt.show()

# Straight lines
x = np.linspace(0, 1, 10000)
y1 = beta.pdf(x, 1, 2)
y2 = beta.pdf(x, 1, 1)
y3 = beta.pdf(x, 2, 1)

plt.title("PDF of Beta (Straight lines)", fontsize=20)
plt.xlabel("X", fontsize=16)
plt.ylabel("Probability Density", fontsize=16)
plt.plot(x, y1, linewidth=3, color='firebrick')
plt.annotate("Beta(1,2)", xy=(0.88, 0.45), size = 14, ha='center', va='center', color='
plt.plot(x, y2, linewidth=3, color='burlywood')
plt.annotate("Beta(1,1)", xy=(0.88, 1.15), size = 14, ha='center', va='center', color='
plt.plot(x, y3, linewidth=3, color='dodgerblue')
plt.annotate("Beta(2,1)", xy=(0.88, 2.0), size = 14, ha='center', va='center', color='d
plt.ylim([0, 4])
plt.xlim([0, 1])
plt.show()

# U shape
```

5. Classical Derivation: Order Statistic

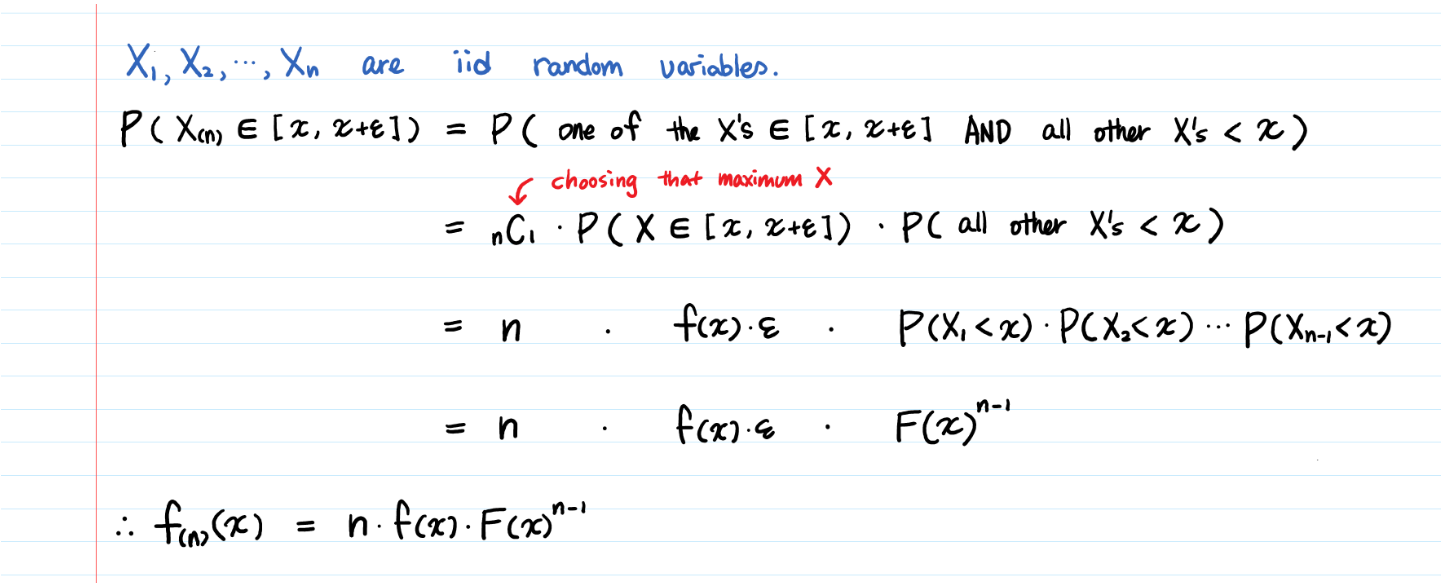
When I learned Beta distribution at school, I derived it from the order statistic. The order statistic isn't the most widely used application of the Beta distribution, but it helped me think about the distribution deeper and understand it better.

Let **X_1**, **X_2**, . . . , **X_n** be iid random variables with PDF **f** and CDF **F**.

We're re-arranging them in increasing order so that **X_k** is **the k-th smallest X**, called the **k-th order statistic**.

a. What's the density of the maximum X?

(Not familiar with the term "Density"? Read "PDF is NOT a probability")



$X_1, X_2, \cdots, X_n$ are iid random variables.

$$P(X_{(n)} \in [x, x+\varepsilon]) = P(\text{one of the } X\text{'s} \in [x, x+\varepsilon] \text{ AND all other } X\text{'s} < x)$$

choosing that maximum X

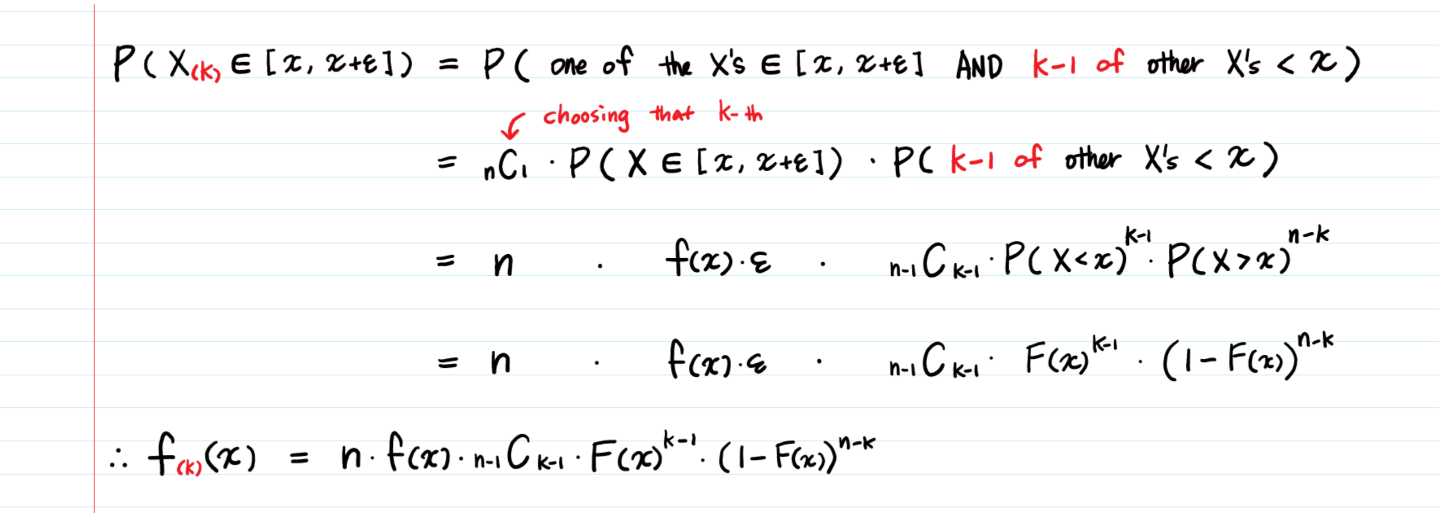
$$= {}_nC_1 \cdot P(X \in [x, x+\varepsilon]) \cdot P(\text{all other } X\text{'s} < x)$$

$$= n \cdot f(x)\cdot\varepsilon \cdot P(X_1 < x)\cdot P(X_2 < x) \cdots P(X_{n-1} < x)$$

$$= n \cdot f(x)\cdot\varepsilon \cdot F(x)^{n-1}$$

$$\therefore f_{(n)}(x) = n \cdot f(x) \cdot F(x)^{n-1}$$

b. What's the density of the k-th order statistic?



$$P(X_{(k)} \in [x, x+\varepsilon]) = P(\text{ one of the } X\text{'s} \in [x, x+\varepsilon] \text{ AND } k-1 \text{ of other } X\text{'s} < x)$$

$$= \overset{\swarrow \text{ choosing that } k\text{-th}}{{}_nC_1} \cdot P(X \in [x, x+\varepsilon]) \cdot P(k-1 \text{ of other } X\text{'s} < x)$$

$$= n \cdot f(x)\cdot\varepsilon \cdot {}_{n-1}C_{k-1} \cdot P(X<x)^{k-1} \cdot P(X>x)^{n-k}$$

$$= n \cdot f(x)\cdot\varepsilon \cdot {}_{n-1}C_{k-1} \cdot F(x)^{k-1} \cdot (1-F(x))^{n-k}$$

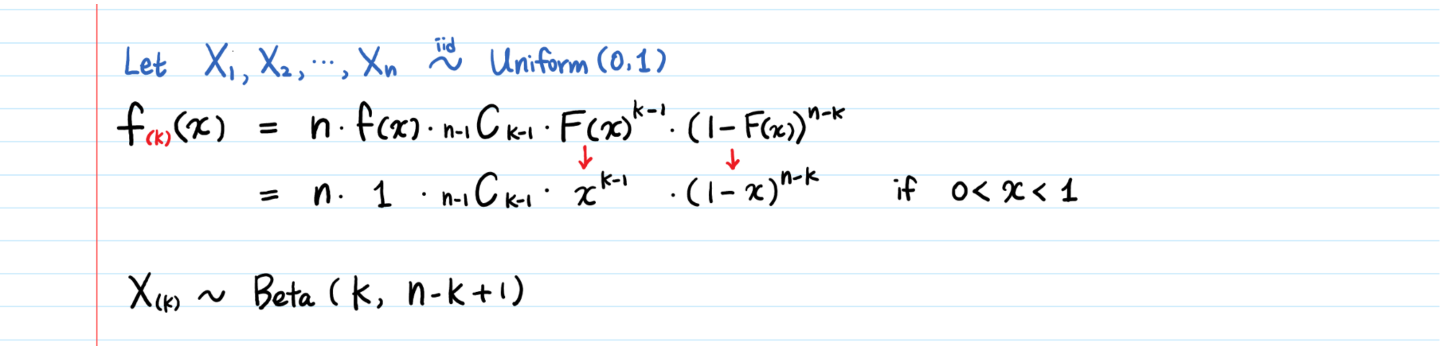
$$\therefore f_{(k)}(x) = n \cdot f(x) \cdot {}_{n-1}C_{k-1} \cdot F(x)^{k-1} \cdot (1-F(x))^{n-k}$$

Notice we don’t need to choose nor permute **X**s bigger than **x**.

c. How can we derive the Beta distribution using the k-th order statistic?

What happens if we set **X_1**, **X_2**, . . . , **X_n** as iid Uniform(0,1) random variables?

Why Uniform(0,1)? Because the domain of the Beta is [0,1].



Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \text{Uniform}(0,1)$

$$f_{(k)}(x) = n \cdot f(x) \cdot {}_{n-1}C_{k-1} \cdot \underset{\downarrow}{F(x)}^{k-1} \cdot (1-\underset{\downarrow}{F(x)})^{n-k}$$

$$= n \cdot 1 \cdot {}_{n-1}C_{k-1} \cdot x^{k-1} \cdot (1-x)^{n-k} \quad \text{if } 0 < x < 1$$

$$X_{(k)} \sim \text{Beta}(k, n-k+1)$$

The CDF of uniform distribution (0,1) is **x**.

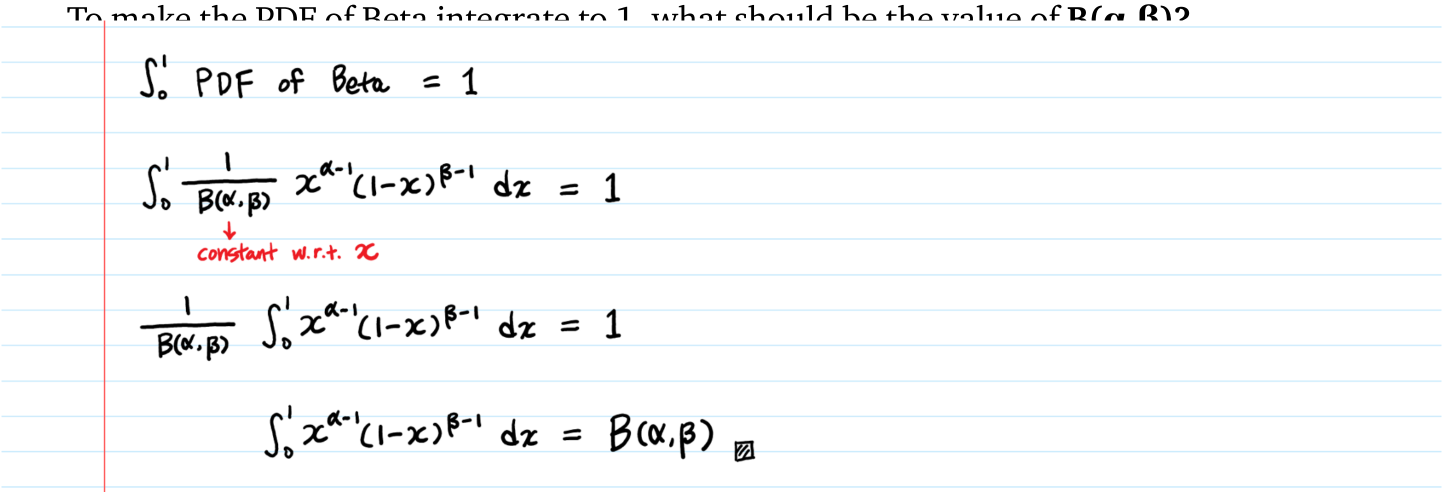
Here, we have the Beta!

6. Beta Function as a normalizing constant

I proposed earlier:

> *Let's ignore the coefficient* ***1/B(α,β)*** *… because* ***1/B(α,β)*** *is just a normalizing constant to make the function integrate to 1.*

To make the PDF of Beta integrate to 1, what should be the value of **B(α,β)**?



$$\int_0^1 \text{PDF of Beta} = 1$$

$$\int_0^1 \underset{\substack{\downarrow \\ \text{constant w.r.t. } x}}{\frac{1}{B(\alpha,\beta)}} x^{\alpha-1}(1-x)^{\beta-1}\,dx = 1$$

$$\frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = 1$$

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = B(\alpha,\beta)$$ ▨

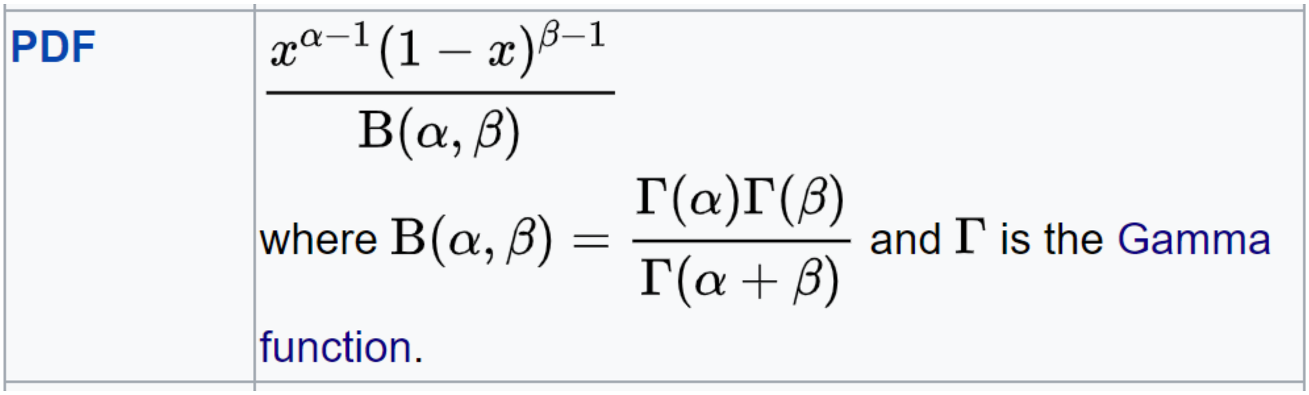
What should be the value of **B(α,β)**?

B(α,β) is **the area under the graph of the Beta PDF** from 0 to 1.

7. Simplify the Beta function with the Gamma Function!

This section is for the proof addict like me.

You might have seen the PDF of Beta written in terms of the Gamma function. The Beta function is the ratio of the product of the Gamma function of each parameter divided by the Gamma function of the sum of the parameters.

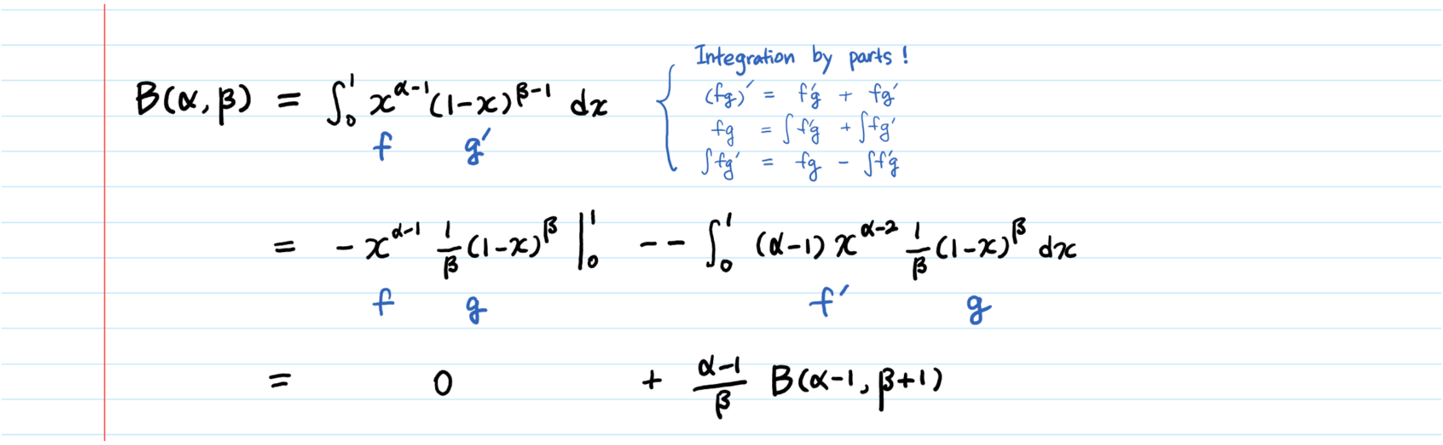


PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.

An excerpt from Wikipedia

How can we prove **B(α,β) = Γ(α) * Γ(β) / Γ(α+β) ?**

Let’s take the special case where **α** and **β** are integers and start with what we’ve derived above.



$$B(\alpha,\beta) = \int_0^1 \underbrace{x^{\alpha-1}}_{f}\underbrace{(1-x)^{\beta-1}}_{g'}\,dx$$

Integration by parts !

$$(fg)' = f'g + fg'$$
$$fg = \int f'g + \int fg'$$
$$\int fg' = fg - \int f'g$$

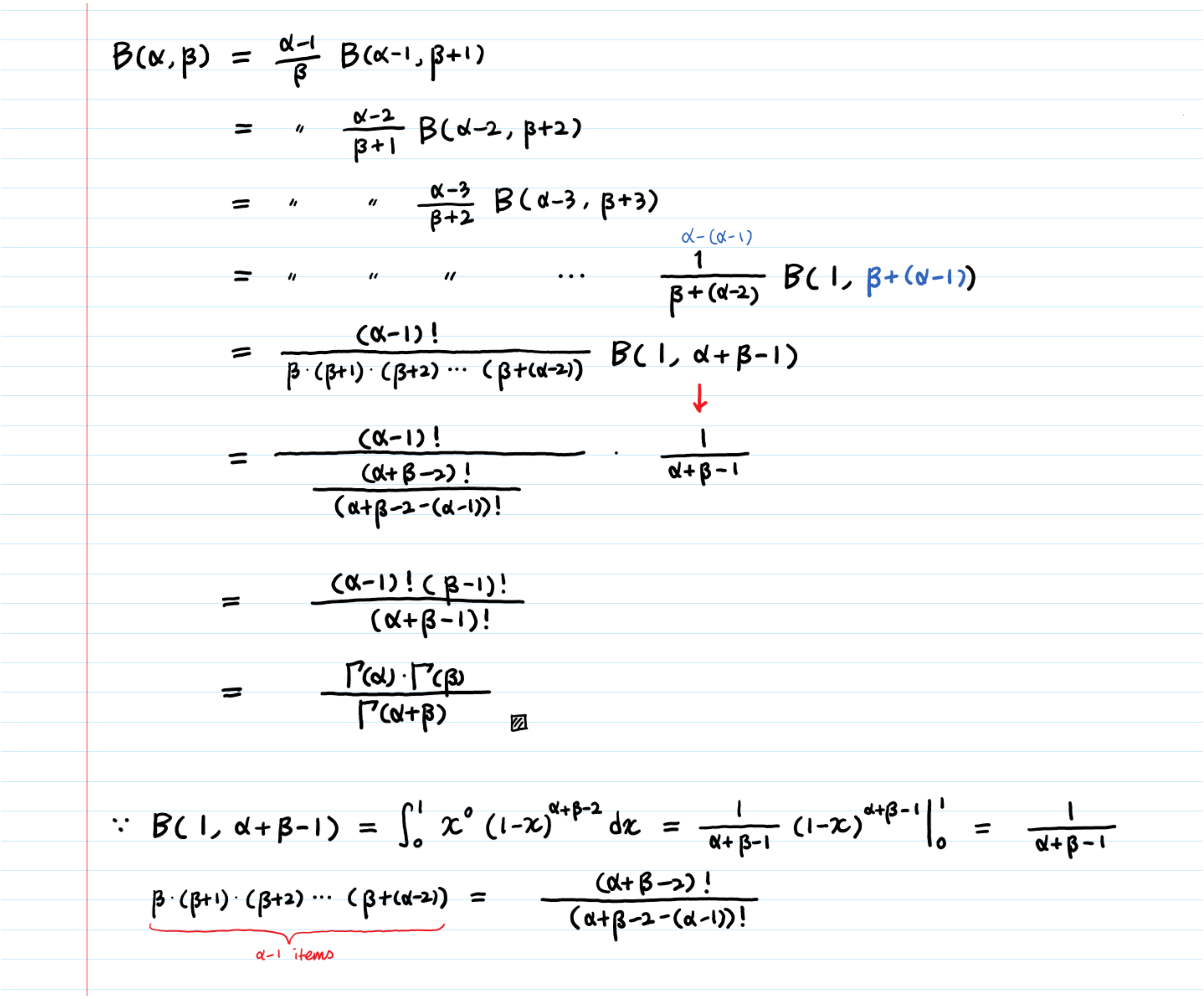
$$= \underbrace{-x^{\alpha-1}}_{f}\underbrace{\frac{1}{\beta}(1-x)^{\beta}}_{g}\Big|_0^1 \;-\,-\; \int_0^1 \underbrace{(\alpha-1)\,x^{\alpha-2}}_{f'}\underbrace{\frac{1}{\beta}(1-x)^{\beta}}_{g}\,dx$$

$$= 0 + \frac{\alpha-1}{\beta}B(\alpha-1,\beta+1)$$

We got a recursive relationship **B(α,β) = (α-1) * B(α-1,β+1) / β.**

How should we exploit this relationship?

We can try to get to the base case B(1, *).



$$B(\alpha,\beta) = \frac{\alpha-1}{\beta} B(\alpha-1,\beta+1)$$

$$= 〃 \frac{\alpha-2}{\beta+1} B(\alpha-2,\beta+2)$$

$$= 〃 \quad 〃 \frac{\alpha-3}{\beta+2} B(\alpha-3,\beta+3)$$

$$= 〃 \quad 〃 \quad 〃 \cdots \frac{\overset{\alpha-(\alpha-1)}{1}}{\beta+(\alpha-2)} B(1, \beta+(\alpha-1))$$

$$= \frac{(\alpha-1)!}{\beta\cdot(\beta+1)\cdot(\beta+2)\cdots(\beta+(\alpha-2))} \underset{\downarrow}{B(1, \alpha+\beta-1)}$$

$$= \frac{(\alpha-1)!}{\frac{(\alpha+\beta-2)!}{(\alpha+\beta-2-(\alpha-1))!}} \cdot \frac{1}{\alpha+\beta-1}$$

$$= \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

$$= \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad ▨$$

$$\because B(1, \alpha+\beta-1) = \int_0^1 x^0 (1-x)^{\alpha+\beta-2} dx = \frac{1}{\alpha+\beta-1}(1-x)^{\alpha+\beta-1}\Big|_0^1 = \frac{1}{\alpha+\beta-1}$$

$$\underbrace{\beta\cdot(\beta+1)\cdot(\beta+2)\cdots(\beta+(\alpha-2))}_{\alpha-1 \text{ items}} = \frac{(\alpha+\beta-2)!}{(\alpha+\beta-2-(\alpha-1))!}$$

Beautifully proved!

Data Science Machine Learning Probability Statistics