Understanding Random Variables and Distributions

Dmitri Nesteruk
QUANTITATIVE ANALYST

@dnesteruk http://activemesa.com

Goals:

Understand the notion of a *random variable* and the common distributions of random variables.

Overview

What is a Random Variable?

Discrete vs Continuous

Distributions and Probability Functions

Discrete distributions: uniform, binomial, geometric, hypergeometric

Continuous distributions: uniform, normal, gamma, beta

Random Variable

Formal: a real-value function on the sample space.

Informal: a variable that can take on a random value from a finite or infinite set of values.

Discrete vs Continuous

Discrete **random variables can take on values that are**

- Finite (e.g., die roll, coin toss)
- Countably infinite, i.e. can be put into 1-1 correspondence with natural numbers

Continuous **random variables can take on an infinite set of values**

- A person's exact height
- $P(\text{you are } \textit{exactly} \text{ 2 m. tall}) = 0$
- Can be turned into a discrete value by rounding

Notation

Random variables are typically denoted with a capital letter

$$X, Y, \text{etc.}$$

The probability of random variable X taking on a specific value (e.g., 3) is expressed as

$$P(X = 3) = \frac{1}{6}$$

The probability of random variable X taking on some value x is expressed as

$$P(X = x) = \frac{1}{x^2}$$

and this can be a function of x.

Discrete Random Variable

Random variable that takes on a finite (or countably infinite) set of values

Examples:

- Single coin toss (H or T)
- Number of heads in 10 coin tosses
- Die roll (6 possible values)
- Person’s ranking in a competition

Values don’t have to be equally likely

- E.g., a loaded die

Distribution

The *distribution* of random variable X is the collection of all probabilities $P(X \in S)$ for all sets of real numbers such that $\{X \in S\}$ is an event

Simple coin toss

$$P(X = H) = P(X = T) = 1/2$$

Number of heads in 10 coin tosses

- 2^{10} different outcomes, $P(X = x) = \frac{1}{2^{10}}$
- Need to count # of outcomes s such that $X(s) = x$
- Number of such outcomes = number of subsets of size x that can be chosen from 10 tosses, i.e., $\binom{10}{x}$
- $P(X = x) = \binom{10}{x} \frac{1}{2^{10}}$ for $x = 0, \ldots, 10$

Probability Function

Given X with a discrete distribution

The probability function (pf) of X is a function s.t. for every real number x

$$f(x) = P(X = x)$$

For example, for a fair die roll,

$$f(x) = \begin{cases} {}^{1}/_{6}, & x \in \{1,2,3,4,5,6\} \\ 0, & \text{otherwise} \end{cases}$$

Also known as *probability mass function*

Uniform Distribution of Integers

A lottery machine has balls corresponding to lottery numbers

Finite set 1..49

Each ball equally likely to be drawn

$P(X = 33) = 1/49$ (first draw)

A uniform distribution on k integers has probability $1/k$ for each integer

Given a random integer from a to b inclusive s.t. $a < b$, we have $b - a + 1$ possible values, so pf is

$$f(x) = \begin{cases} \dfrac{1}{b-a+1} & \text{for } x = a, \ldots, b \\ 0 & \text{otherwise} \end{cases}$$

Binomial Distribution

A manufactured item is defective with probability p

We want to find the probability of x items being defective in a production run of n items

We consider sequences of

$$\underbrace{FFF \dots FF}_{x}\,\underbrace{SSS \dots SS}_{n-x}$$

The probability of exactly x items being defective (and $n - x$ non-defective) is

$$p^x(1-p)^{n-x}$$

Binomial Distribution

The *number* of such sequences of success-failure pairs is $\binom{n}{x}$

It follows that

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

∴ the pf of X is

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0,1,\dots,n \\ 0, & \text{otherwise} \end{cases}$$

The distribution represented by this pf is the *discrete binomial distribution* with parameters n and p

Geometric Probability Distribution

Similar to the Binomial experiment with success probability p

Measuring different thing

The random variable X corresponds to the trial on which the first success occurs

E_1: S, success on first trial

E_2: F, S, success on second trial

E_3: F, F, S, success on third trial

$\vdots$

E_n: $\underbrace{F, F, F, \dots, F}_{n-1}, S$, success on n^{th} trial

Geometric Probability Distribution

Random variable X is the number of trials up to and including the first success

Any event E_n does not include any prior outcome E_m where $m < n$

Because trials are independent, for

$$x = 1, 2, 3, \dots,$$

$$p(x) = P\left(\underbrace{FFF \dots FF}_{x-1} S\right) = \underbrace{qqq \dots qq}_{x-1} p = q^{x-1} p$$

Geometric Probability Distribution

A random variable X has a geometric probability distribution iff

$$p(x) = q^{x-1}p$$

where

$$x = 1, 2, 3, \ldots, \qquad 0 \leq p \leq 1$$

and $q = 1 - p$

Geometric Distribution Example

Suppose the probability of engine malfunction in a 1-hour period is $p = 0.03$

Find the probability that the engine will survive 2 hours

Let X denote number of 1-hour intervals until first malfunction

$P(\text{survive 2hrs}) = P(X \geq 3) = \sum_{y=3}^{\infty} p(x)$

Since $\sum_{x=1}^{\infty} p(x) = 1$,

$P(\text{survive 2hrs}) = 1 - \sum_{x=1}^{2} p(x) = 1 - p - qp$

$= 1 - 0.03 - 0.97 \cdot 0.03 = 0.9409$

Hypergeometric Probability Distribution

Consider a population of N elements that have a characteristic with 2 possible states

E.g., color of balls in a bag

Suppose r elements are red and $b = N - r$ are blue

A sample of n elements is selected

We are interested in X, the number of successful cases (e.g., red balls) selected

X follows a hypergeometric distribution

Hypergeometric Probability Distribution

A random variable X follows a hypergeometric distribution if its pf is

$$p(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

N – population size

r – number of success states in population

n – number of draws

x – number of observed successes

Hypergeometric Distribution Example

A factory has 10 machines, 4 are defective. If we pick 5 machines at random, what's the probability none of them are defective?

6 are non-defective, so

$$N = 10, r = 6, n = 5, x = 5$$

$$P(X = 5) = \frac{\binom{6}{5}\binom{10-6}{5-5}}{\binom{10}{5}} = \frac{1}{42} = 0.00238$$

Continuous Distributions

Continuous distributions assign probability 0 (zero!) to individual values

$$P(X = x) = 0 \text{ for each } x$$

This means a pf makes no sense

But we can talk about the probability that X falls between some values

$$P(a \leq X \leq b)$$

Given the parameter x, we define the cumulative distribution function (cdf) $F(x)$ as

$$F(x) = P(X \leq x)$$

Cumulative Distribution Function Example

Consider X that has a binomial distribution with $n = 2, p = 1/2$. Let's find $F(x)$…

The pf for X is

$$p(x) = \binom{2}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{2-x}, \qquad x = 0,1,2$$

This gives us $p(0) = \frac{1}{4}, p(1) = \frac{1}{2}, p(2) = \frac{1}{4}$

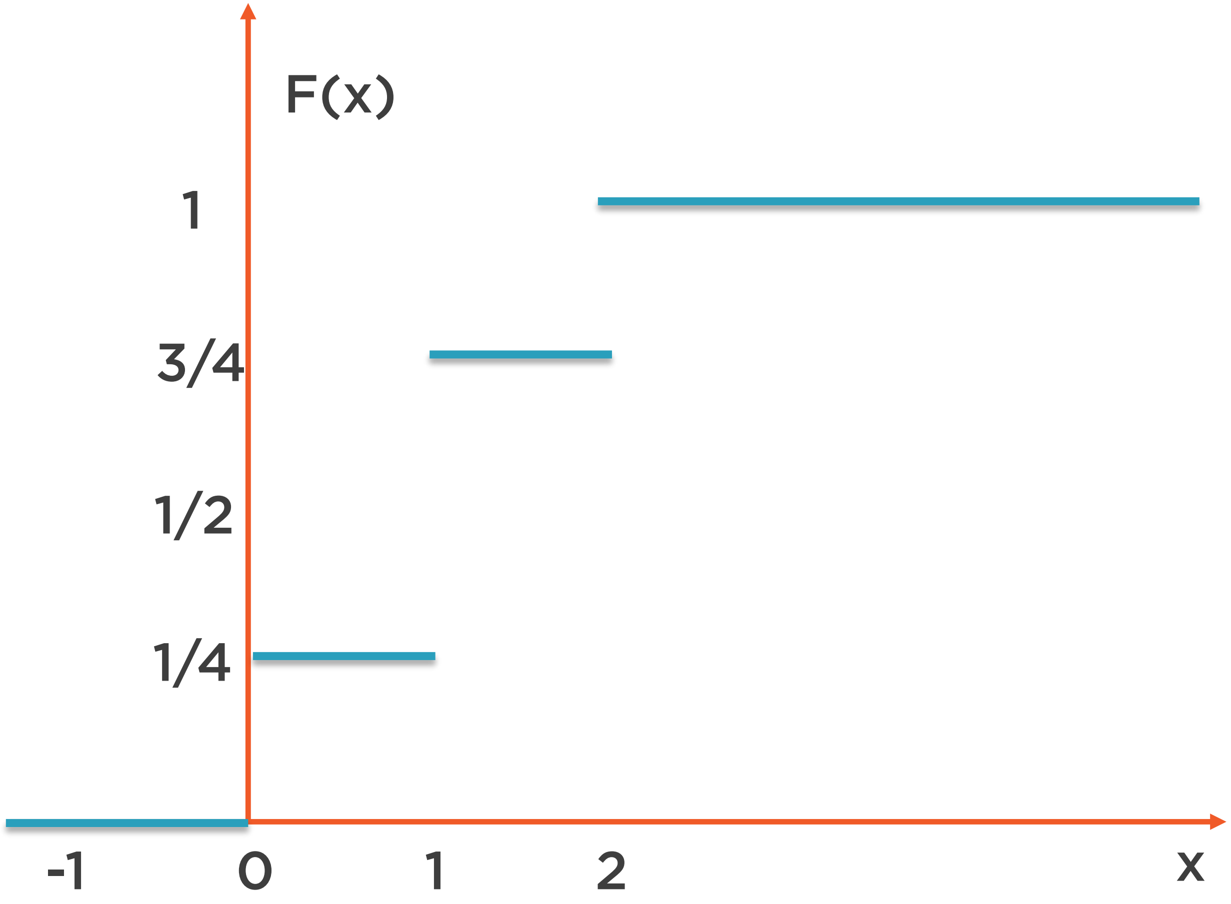
Now we plot the cdf

$F(x) = P(X \leq x)$

$$p(0) = \frac{1}{4}, p(1) = \frac{1}{2}, p(2) = \frac{1}{4}$$

So for each $F(x)$ we add up all the different probabilities $p(a)$ where $a \leq x$

$$F(x) = \begin{cases} 0, & x < 0 \\ {}^{1}/_{4}, & 0 \leq x < 1 \\ {}^{3}/_{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$



Properties of a Distribution Function

$F(-\infty) = 0$

$F(\infty) = 1$

$F(x)$ is a nondecreasing function

A random variable X is continuous if $F(x)$ is continuous for $-\infty < x < \infty$

Probability Density Function

If $F(x)$ is the distribution function for a continuous random variable X, we define $f(x)$ as

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

This is the *probability density function* (pdf) of the random variable X.

- $f(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Calculating Probability Values

Given a pdf $f(x)$,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Suppose you are given the pdf

$$f(x) = \begin{cases} \frac{1}{8}x, & 0 < x < 4 \\ 0, & \text{otherwise} \end{cases}$$

(Notice how $\int_0^4 x/8\ dx = \left.\frac{x^2}{16}\right|_0^4 = 1$**)**

$$P(1 \leq X \leq 2) = \int_1^2 \frac{1}{8}x\ dx = \frac{3}{16}$$

$$P(X > 2) = \int_2^4 \frac{1}{8}x\ dx = \frac{3}{4}$$

Uniform Probability Distribution

A train always arrives between 6:30 and 6:40

The probability it will arrive in any subinterval is proportional to the length of the subinterval

Let X denote amount of time a person has to wait for a train if they arrive at 6:30

X has a continuous uniform probability distribution

Uniform Probability Distribution

If $a < b$, a random variable X is said to have a continuous uniform probability distribution on the interval (a, b) iff the density function of X is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

The constants a and b are the parameters of the density function.

Uniform Probability Distribution Example

Suppose trains arrive within a 30-minute period

What's the probability the train will arrive in the last 5 minutes of that interval?

We have a uniform distribution with $a = 0$ and $b = 30$

$$P(25 \leq X \leq 30) = \int_{25}^{30} \frac{1}{30}\, dx = \frac{30-25}{30} = 1/6$$

Normal Probability Distribution

A random variable X has a normal probability distribution iff, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The normal density function has two parameters, μ and σ. A distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard* normal distribution.

Normal Distribution

Consider the standard normal distribution ($\mu = 0, \sigma = 1$):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

To find out $P(a \leq X \leq b)$ we would need to evaluate

$$\int_a^b e^{-x^2/2} \, dx$$

No closed-form of this integral exists

Numeric integration techniques required

- `pnorm(x,μ,σ)` gives $P(X \leq x)$
- `qnorm(p,μ,σ)` gives the value x s.t. $P(X \leq p) = x$ (pth quartile)

Normal Distribution Example

Suppose we know that test scores are normally distributed with $\mu = 75$ and $\sigma = 10$

What fraction of scores lie between 80 and 90?

Calculate using tables

- We can transform this distribution into a standard one using

$$z = \frac{x - \mu}{\sigma}$$

- This gives us $z_1 = \frac{80-75}{10} = 0.5$ and $z_2 = \frac{90-75}{10} = 1.5$
- Look up the values and subtract

```
pnorm(90,75,10) - pnorm(80,75,10)
```

Answer: 0.24173

Uses of Normal Distribution

Used extensively in natural and social sciences

Brownian motion (physics, mathematical finance)

Gamma Probability Distribution

A random variable X has a gamma distribution with positive parameters α and β iff the density function of X is

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, & 0 \leq x < \infty \\ 0, & \text{otherwise} \end{cases}$$

where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx$

Gamma Function

$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$

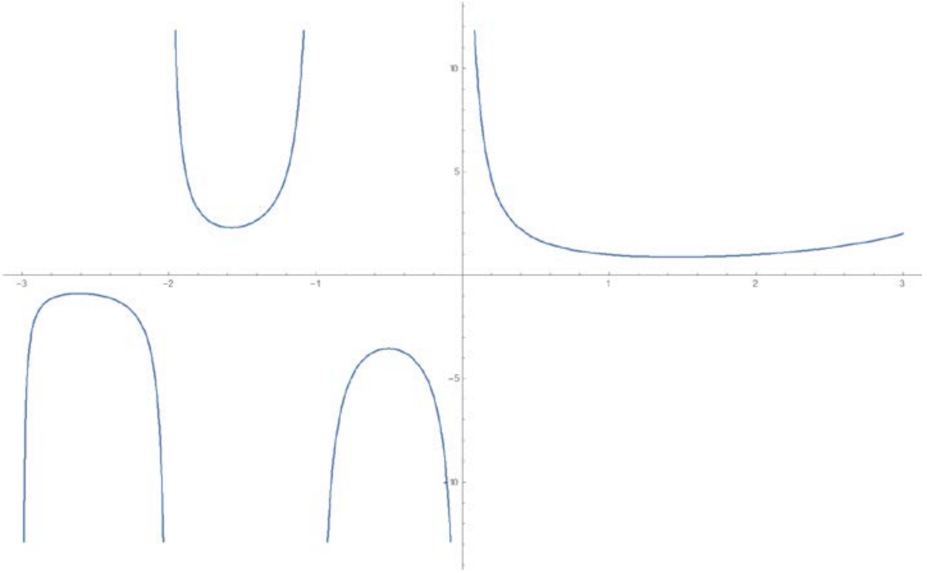
is called the *gamma function*

$\Gamma(1) = \int_0^\infty e^{-x}\, dx = 1$

Integration by parts gives the relation

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

Thus, for $n \in \mathbb{N}$**,** $\Gamma(n) = (n-1)!$

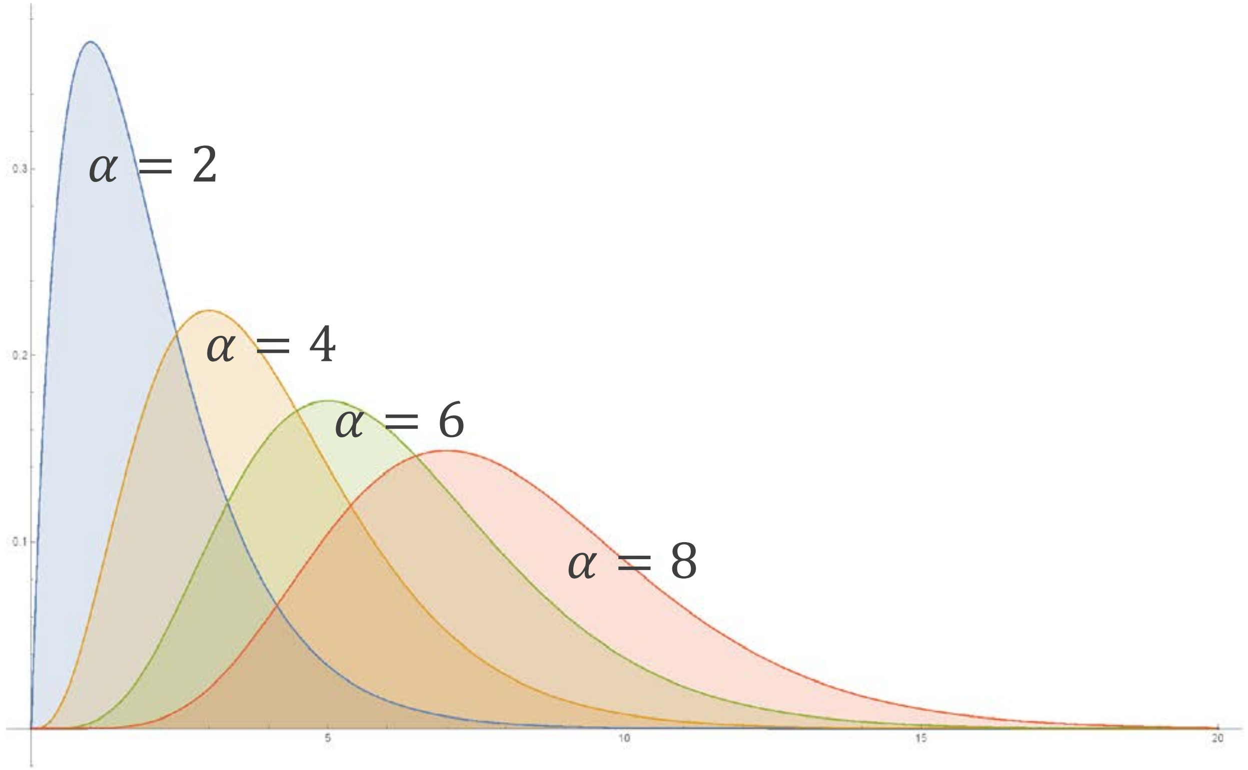


Let's plot the gamma pdf

$$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$$

Assign α (shape parameter) values of $2, 4, 6$ and 8

Fix $\beta = 1$ (scale parameter)



Uses of Gamma Distribution

Insurance claims

Rainfall

Wireless communication (multi-path fading of signal power)

Neuroscience (distribution of inter-spike intervals)

Multi-level Poisson regression models

Beta Probability Distribution

A random variable X is said to have a beta probability distribution with parameters $\alpha > 0$ and $\beta > 0$ iff the density function of X is

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

where

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

Beta Distribution

The cdf for the beta random variable is called the *incomplete beta function*

$$F(x) = \int_0^x \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}\,dt = I_x(\alpha,\beta)$$

When α and β are both positive integers, integration by parts gives us

$$F(x) = \int_0^x \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}\,dt = \sum_{i=\alpha}^{n} \binom{n}{i} x^i (a-x)^{n-i}$$

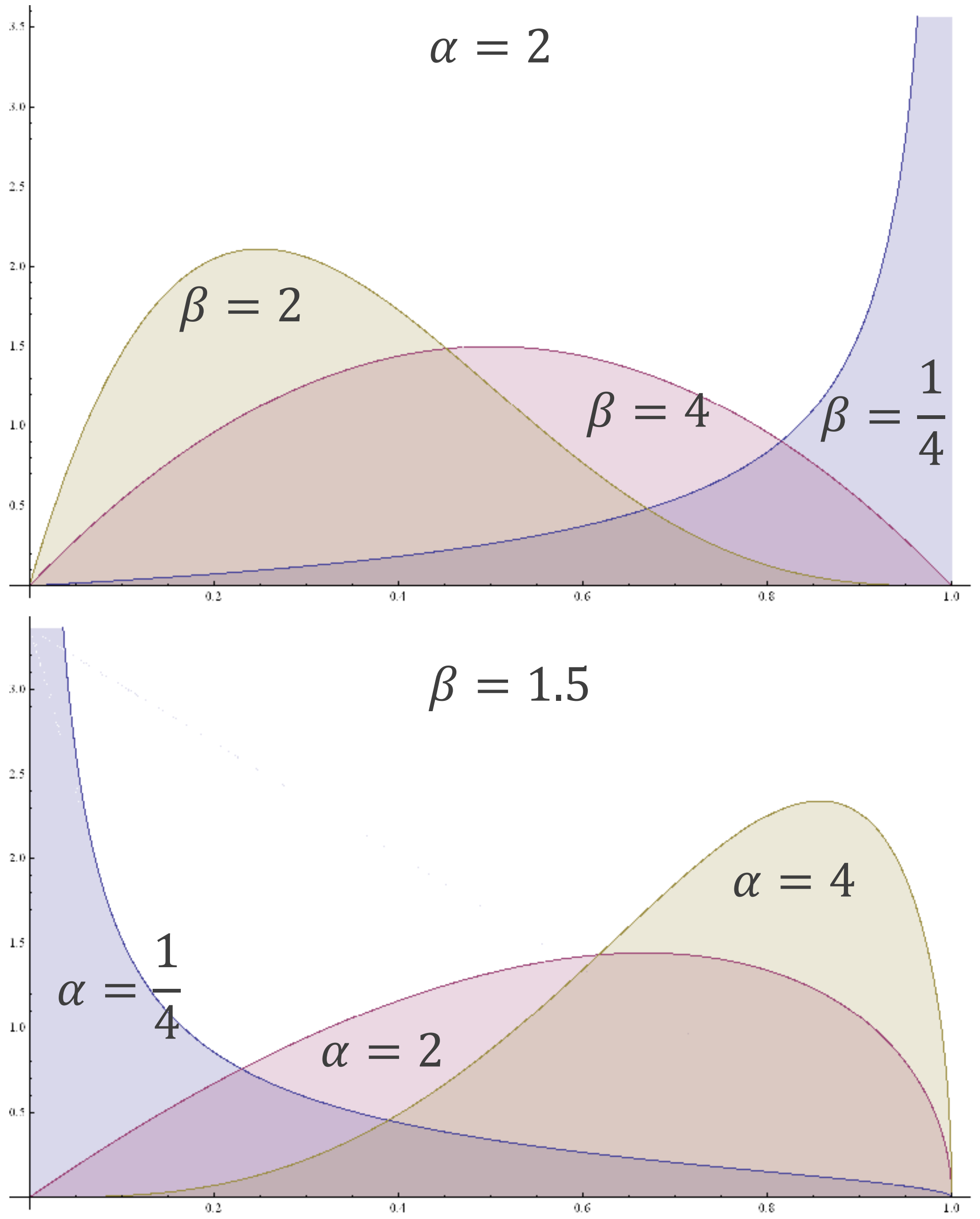
where $n = \alpha + \beta - 1$

This is a sum of probabilities associated with a binomial random variable with $n = \alpha + \beta - 1$ and $p = x$

Plot of the density function by fixing either α or β

$$\alpha = 2, \beta = \left\{\frac{1}{4}, 2, 4\right\}$$

$$\beta = 1.5, \alpha = \left\{\frac{1}{4}, 2, 4\right\}$$



Beta Distribution in R

`pbeta(x,α,1/β)`
yields $P(X \leq x)$

`qbeta(p,α,1/β)`
yields x s.t. $P(X \leq x) = p$

Summary

Discrete distributions are characterized by a probability function

Discrete distributions: uniform, binomial, geometric, hypergeometric

Continuous distributions are characterized by a probability density function (derivative of the cumulative distribution function)

Continuous distributions: uniform, normal, gamma, beta