Probability Distributions

Galvanize

Objectives

By the end of this lesson, you will be able to:

- ❏ Describe the difference between a discrete and continuous random variable
- ❏ Describe the relationship between distribution functions:
  - ❏ Point Mass Function (PMF) or Point Density Function (PDF)
  - ❏ Cumulative Distribution Function (CDF)
- ❏ Describe and provide real-world examples of common distributions
- ❏ Calculate the probability of an event using the scipy.stats module

Random Variables

A **random variable**, usually written X, is a **variable** whose possible values are numerical outcomes of a **random** phenomenon.

Discrete Case:

X = Sum of two rolled dice

Continuous Case:

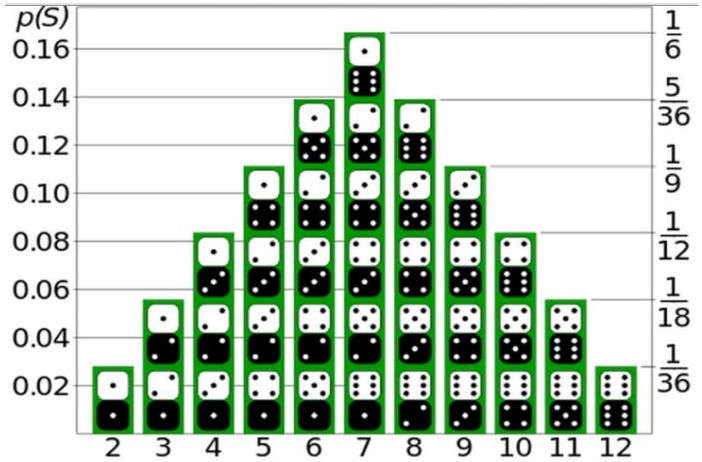
X = IQ Score of random individual

Distributions

The pattern of probabilities of a random variable is called its **distribution**.

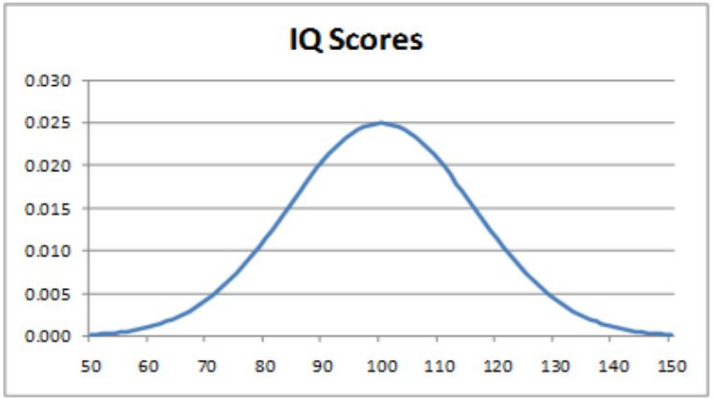
Discrete Case:

X = Sum of two rolled dice



Continuous Case:

X = IQ Score of random individual



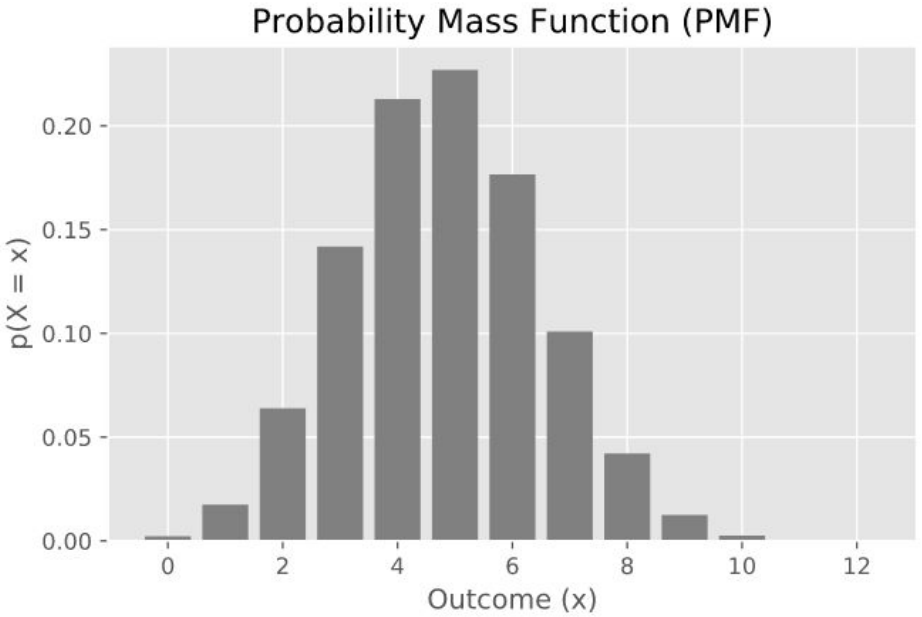
Discrete Distributions

The **probability mass function** (PMF) of a discrete distribution gives the probability of observing each possible outcome (x) of the random variable X.

Example:

The probability of observing a 4 is

$$p(X = 4) = 0.21$$



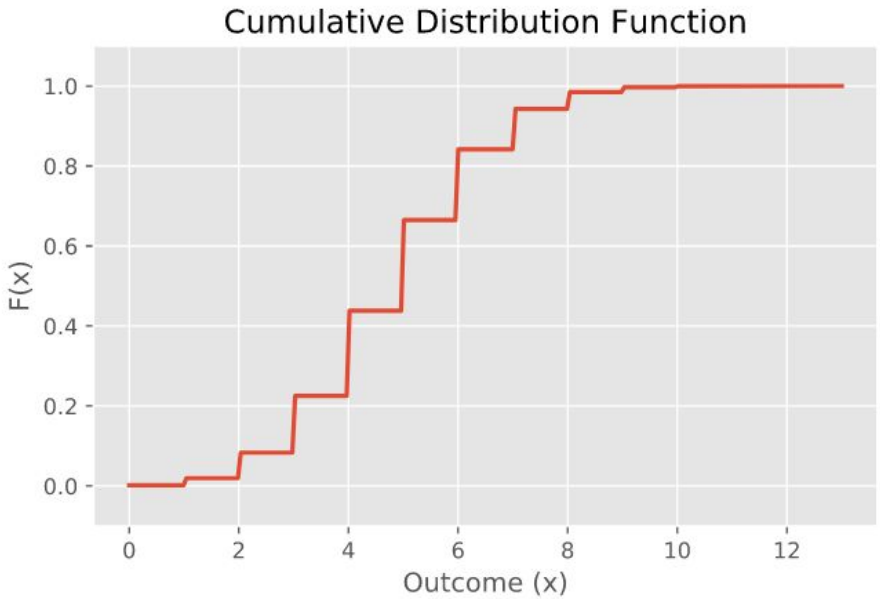
Discrete Distributions

The **cumulative distribution function** (CDF) or just **distribution function** is the probability of observing an outcome less than or equal to *x*.

Example:

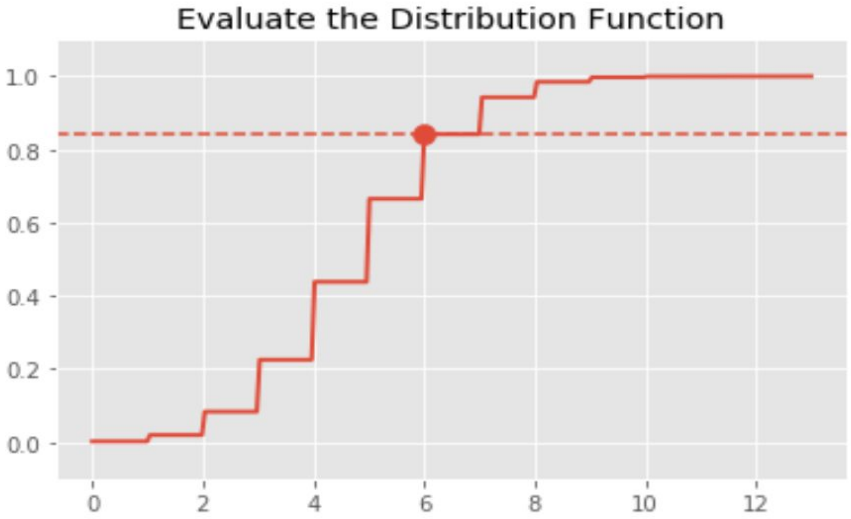
The probability of observing an outcome of 0, 1, 2, 3, or 4 is

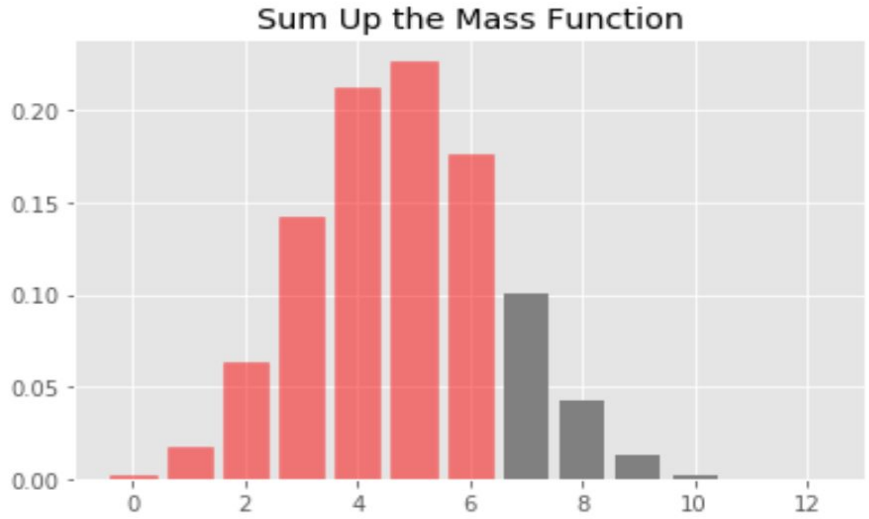
$$F_X(4) = 0.41$$



Discrete Distributions

The CDF of an outcome is a running sum of the PMF up to that outcome.





$F_X(6) = 0.82$

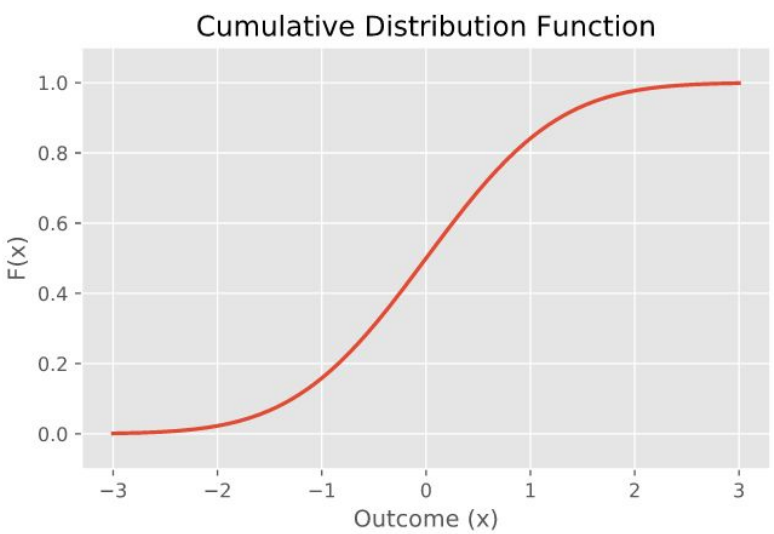
$p(X = 0) + p(X = 1) + \ldots + p(X = 6)$

0.82

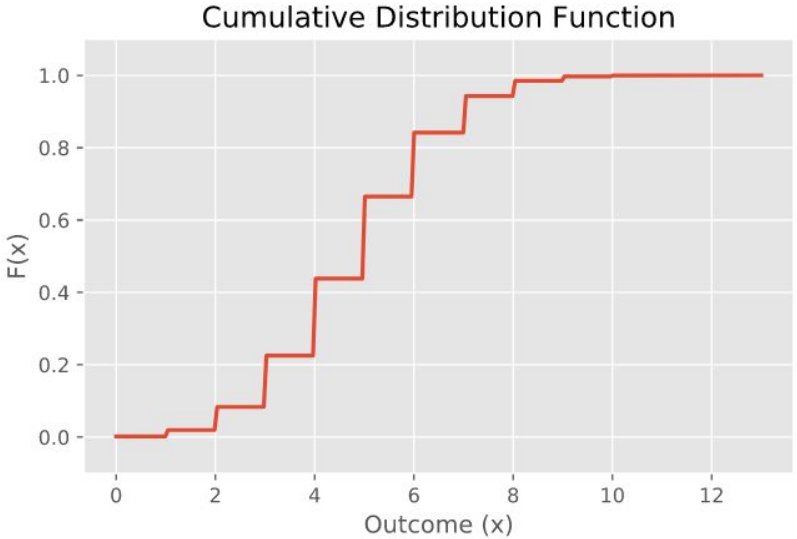
Continuous Distributions

Continuous distributions have a CDF that is smooth, not “jumpy”.

Continuous Distribution



Discrete Distribution



Continuous Distributions

Speaking about probability is a bit tricky for continuous random variables.

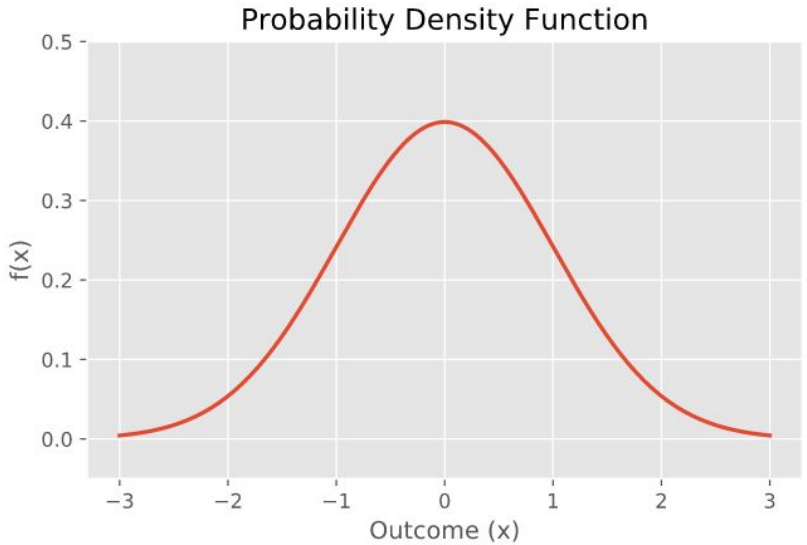
For example:

Suppose a species of bacteria typically lives 4 to 6 hours. What is the probability that a bacterium lives *exactly* 5 hours? The answer is 0%. A lot of bacteria live for *approximately* 5 hours, but there is no chance that any given bacterium dies at *exactly* 5.0000000000... hours.

Source: Wikipedia

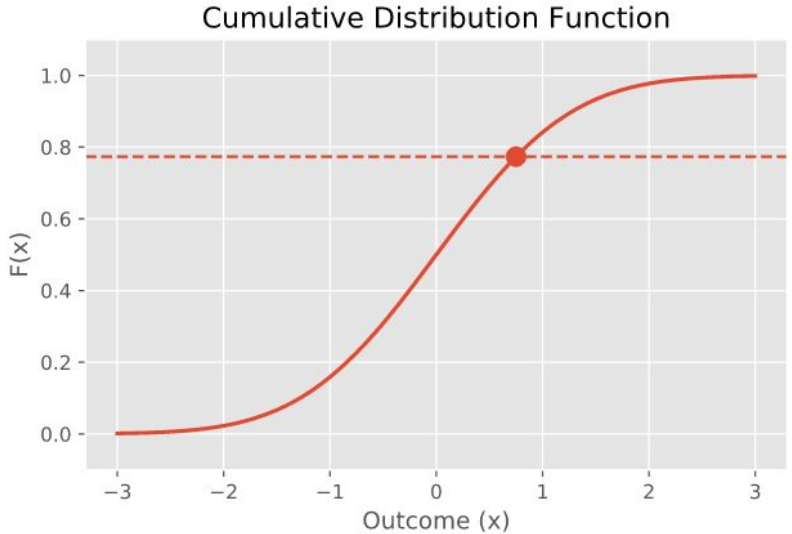
Continuous Distributions

However, we know that some outcomes are more likely than others. The probability of an outcome *x* relative to all other possible outcomes is described by a **probability density function (PDF)**.

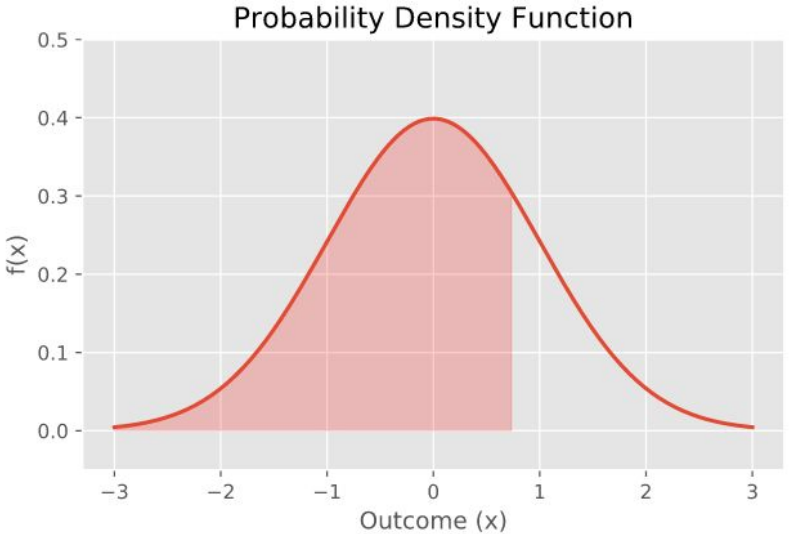


Continuous Distributions

We can't take a running sum of the PDF to get a CDF, but in calculus there is something similar called integration. Think of it as taking the area under the curve.



$F_X(0.8) = 0.78$



Area under the curve: 0.78

Check for Understanding

By the end of this lesson, you will be able to:

- ❏ Describe the difference between a discrete and continuous random variable
- ❏ Describe the relationship between distribution functions.

1. How is a random variable in statistics different than a variable in algebra?
2. What is the difference between a discrete random variable and a continuous random variable?
3. Describe the relationship between the probability mass function and the continuous distribution function of a distribution.
4. Is question 3 talking about a discrete or continuous distribution? How can you tell?

Types of Discrete Distributions

Humans have discovered and catalogued many, many distributions that are intended to describe various situations that arise in science and data analysis. It would be impossible (and useless) to describe them all, so we will stick to the ones that either:

1) Will be used in this class.
2) Will commonly arise in the work and research of an everyday data scientist.

Uniform Distribution

The Uniform Distribution is the most familiar discrete distribution. It describes a situation with a finite number of outcomes, where each outcome is as equally likely as any other. For example, a die roll is uniformly distributed, with 6, or 10, or 12, or 20 possible outcomes, depending on the number of sides of the die.

The probability mass function of the (discrete) uniform distribution is:

$$f(k) = \frac{1}{\#\text{ of outcomes}}$$

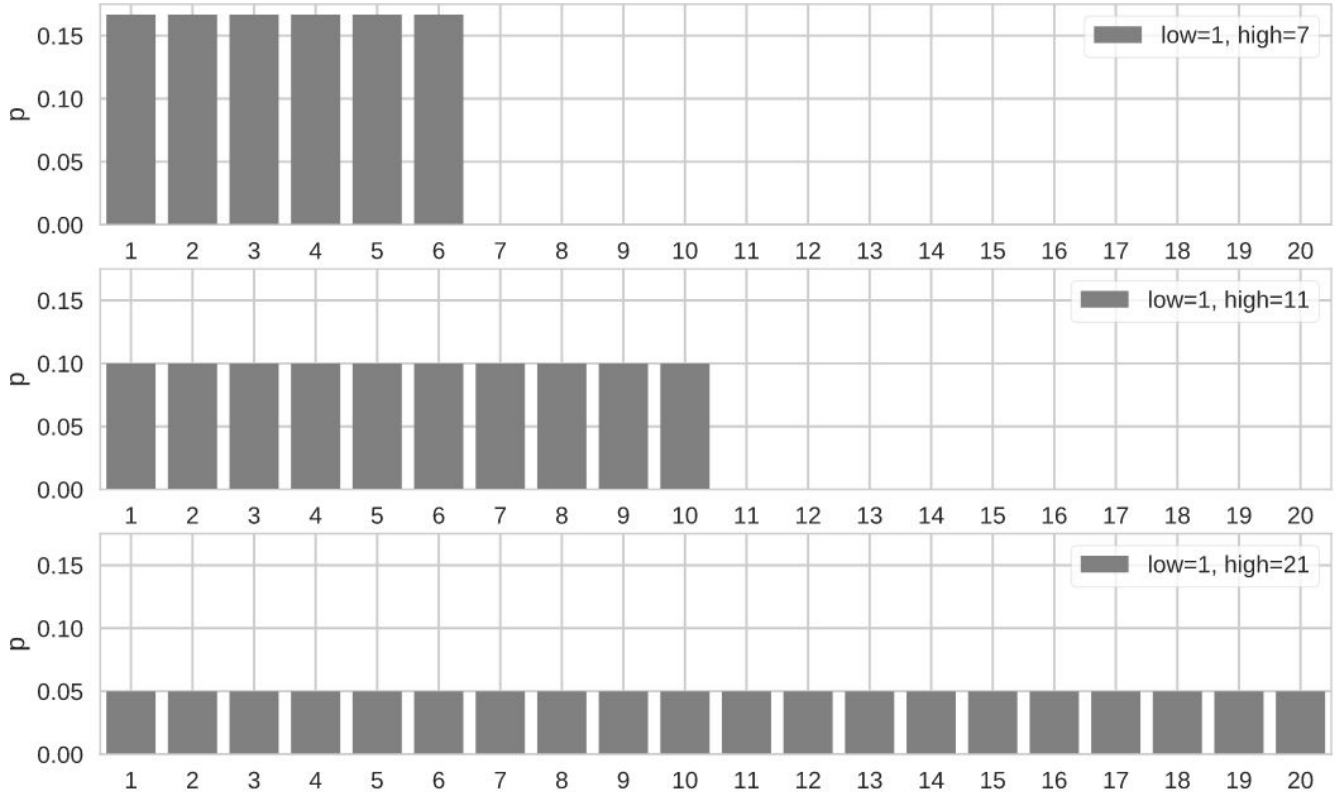
and the distribution function is:

$$f(k) = \frac{\#\text{ of outcomes} \leq k}{\#\text{ of outcomes}}$$

Uniform Distribution

$$f(k) = \frac{1}{\#\text{ of outcomes}}$$

Probability Mass Function



6-sided die

10-sided die

20-sided die

Bernoulli Distribution

The Bernoulli distribution is the simplest discrete distribution. It is a model of a single flip of a (possibly unfair) coin.

A random variable X has a Bernoulli distribution if:

- There are only two possible outputs for X, traditionally labeled 0 and 1.
- There is a probability of p that X outputs 1.

The probability mass function of the Bernoulli distribution is:

$$f(k) = \begin{cases} 1-p, & \text{if } k = 0 \\ p, & \text{if } k = 1 \end{cases}$$

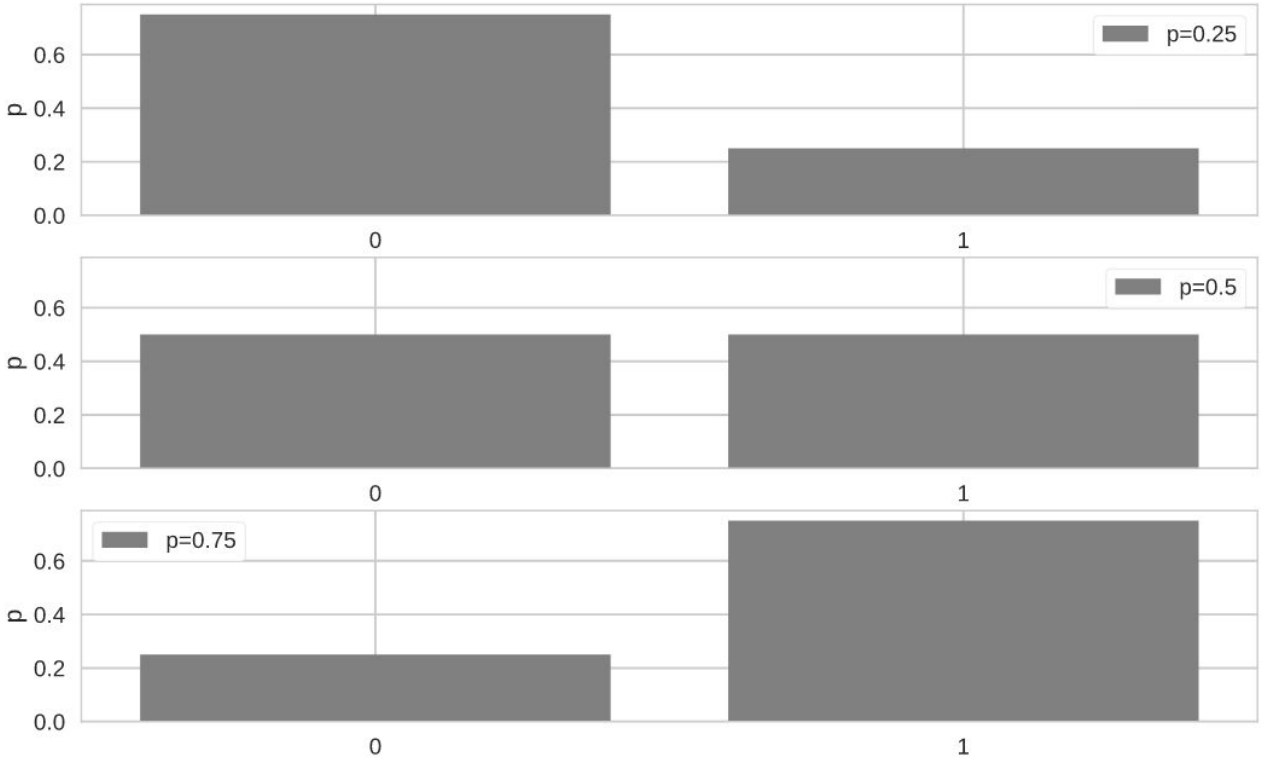
The distribution function of the Bernoulli distribution is:

$$F(k) = \begin{cases} 0, & \text{for } k < 0 \\ 1-p, & \text{for } 0 \leq k < 1 \\ 1, & \text{for } k \geq 1 \end{cases}$$

Bernoulli Distribution

Probability Mass Function

$$f(k) = \begin{cases} 1-p, & \text{if } k = 0 \\ p, & \text{if } k = 1 \end{cases}$$



tails-heavy coin

fair coin

heads-heavy coin

Binomial Distribution

The **Binomial distribution** is a counting distribution. It models flipping a (possibly unfair) coin some number of times, and counting how many times the coin lands heads.

The probability mass function is $$f(\text{k heads}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where:

- n is the number of times you flipped the coin.
- p is the probability a single flip results in heads.

Note: $\binom{n}{k}$ is pronounced “n choose k”: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial Distribution

An unfair coin has a 0.8 probability of landing on heads. Calculate the probability of having 7 heads in 10 flips of the coin.

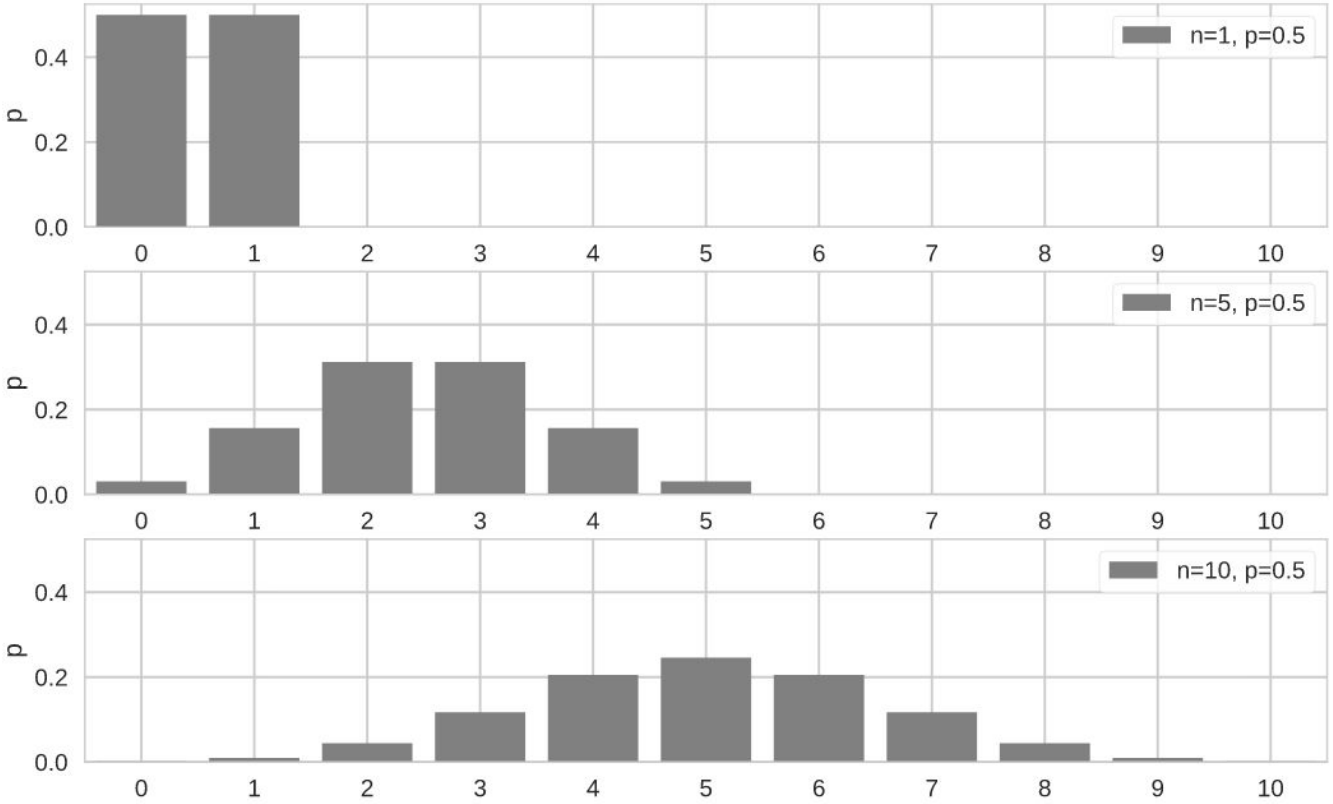
Binomial PMF: $f(\text{k heads}) = \binom{n}{k} p^k (1-p)^{n-k}$

$$\binom{10}{7} 0.8^7 0.2^3 = \frac{10!}{7!3!} 0.8^7 0.2^3 = 0.201$$

Binomial Distribution

$$f(\text{k heads}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability Mass Function



1 flip, fair coin

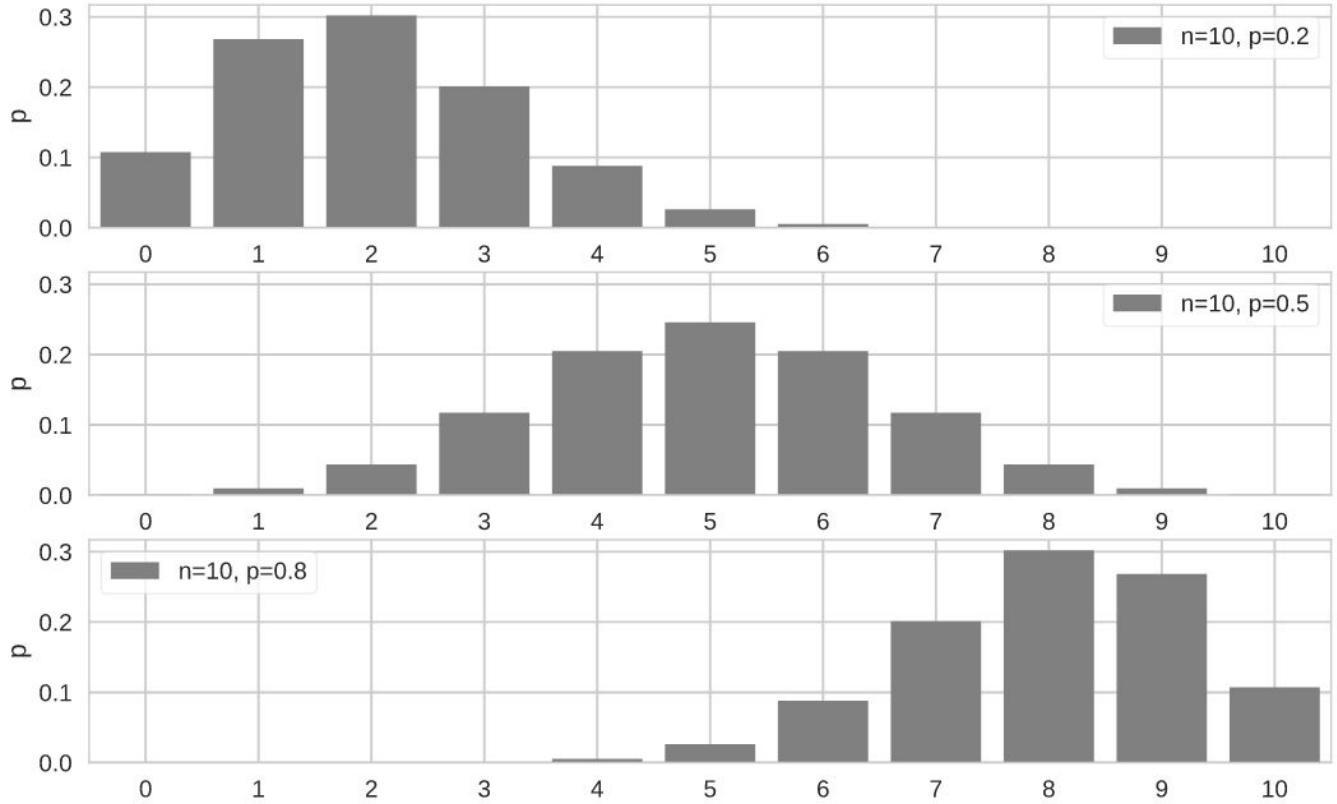
5 flips, fair coin

10 flips, fair coin

Binomial Distribution

$$f(\text{k heads}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Probability Mass Function



10 flips,
tails-heavy coin

10 flips, fair coin

10 flips,
heads-heavy coin

Hypergeometric Distribution

The Hypergeometric distribution is a another counting distribution. This one models a deck of cards of two types (say red cards and blue cards). If you shuffle the deck, draw some number of cards, and then count how many blue cards you have, this count is hyper geometrically distributed.

$$f(k \text{ blue cards}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Where:

- N is the total number of cards in the deck.
- K is the total number of blue cards in the deck.
- n is the size of the hand you drew.

Hypergeometric Distribution

A deck of cards contains 20 cards: 6 blue cards and 14 red cards. 5 cards are drawn randomly *without replacement*. What is the probability that exactly 4 blue cards are drawn?

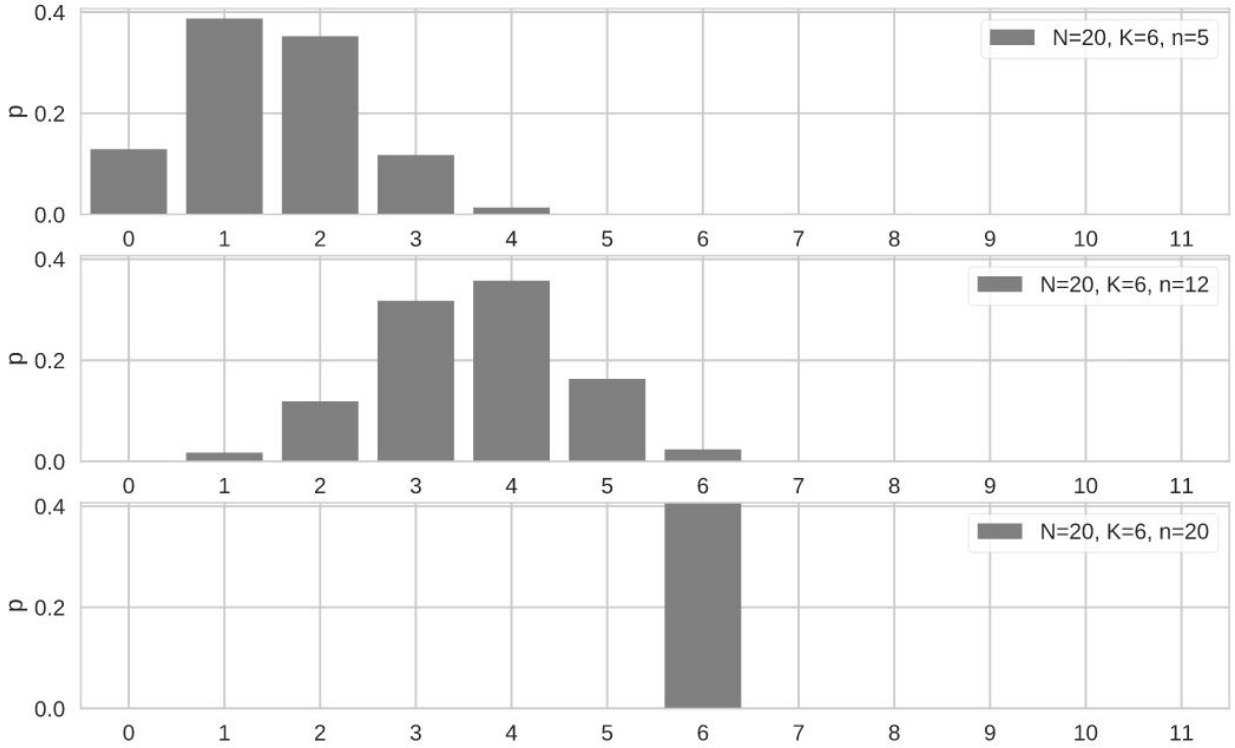
Hypergeometric PMF: $$f(k \text{ blue cards}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$\frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}} = \frac{\frac{6!}{4!2!}\frac{14!}{1!13!}}{\frac{20!}{5!15!}} = \frac{15 \times 14}{15504} = 0.0135$$

Hypergeometric Distribution

$$f(k \text{ blue cards}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Probability Mass Function



draw 5 cards

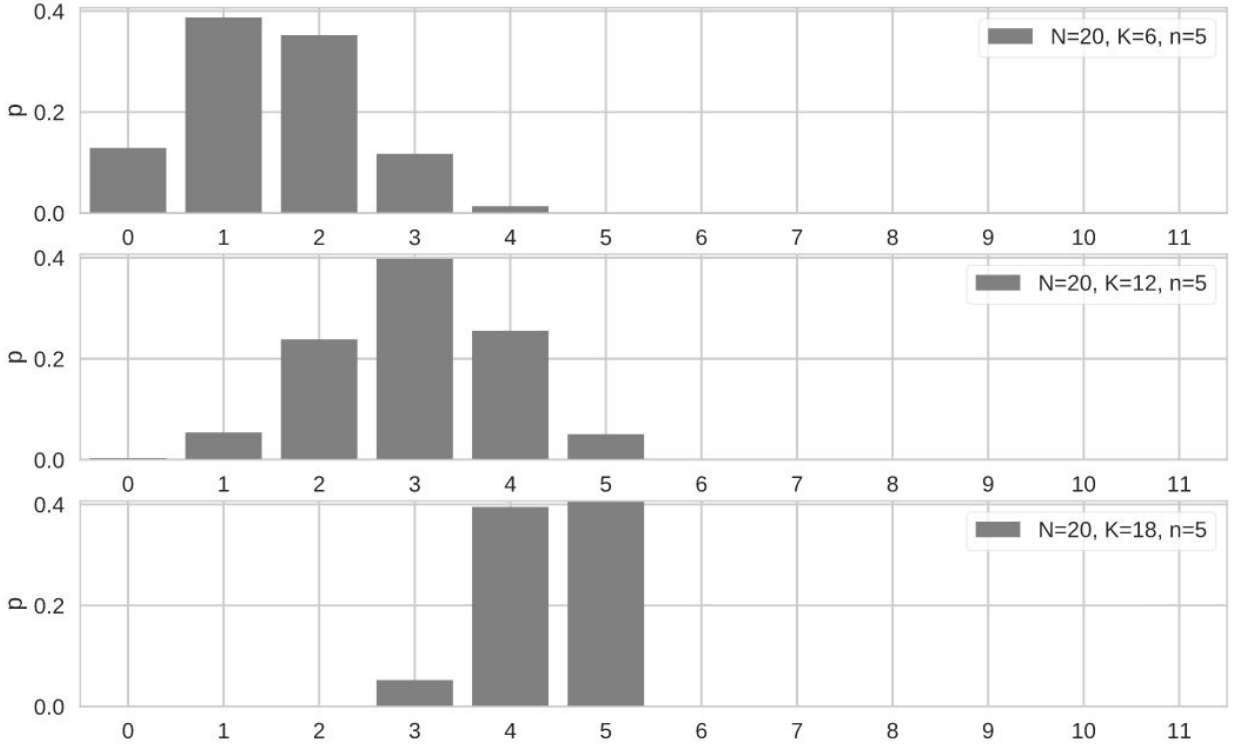
draw 12 cards

draw 20 cards

Hypergeometric Distribution

$$f(k \text{ blue cards}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Probability Mass Function



6 blue cards in deck

12 blue cards in deck

18 blue cards in deck

Poisson Distribution

The **Poisson distribution** is yet another counting distribution. The Poisson distribution models a process where events happen at a fixed **rate or frequency**, and you're watching it for a fixed amount of time.

This distribution is applicable only if the events are **independent** and **identically distributed.**

- **Independent** - knowledge of event A tells you nothing about the probability of event B
- **Identically distributed** - the probability of event A has the same distribution as that of event B

Poisson Distribution

The probability mass function of the Poisson distribution is:

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ is the rate at which the events occur (for example, 2 per-hour, 6 per-day, 122 per-second, etc...).

Poisson Distribution

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes are sold tomorrow?

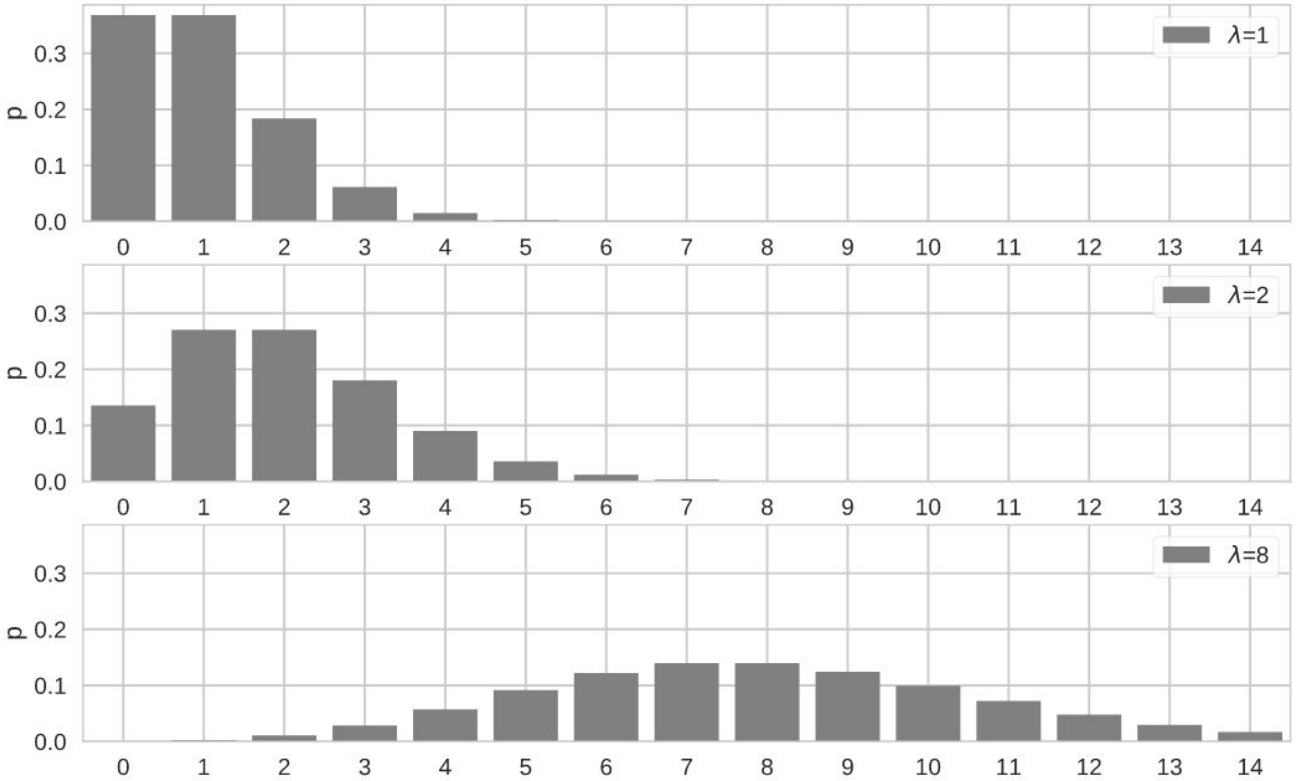
Poisson PMF: $$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\frac{2^3 e^{-2}}{3!} = 0.180$$

Poisson Distribution

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Probability Mass Function



average 1 per day

average 2 per day

average 8 per day

Continuous Distributions

We just studied the following discrete distributions:

- Uniform
- Bernoulli
- Binomial
- Hypergeometric
- Poisson

Now let’s learn about some common continuous distributions...

Uniform Distribution

There is also a continuous version of the Uniform Distribution. It also describes a set of outcomes that are all equally likely, but this time any number in an interval is a possible output of the random variable. For example, the position a raindrop falls on a line segment (in a very large rainstorm) is uniformly distributed.

The probability density function of the (continuous) uniform distribution is:

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

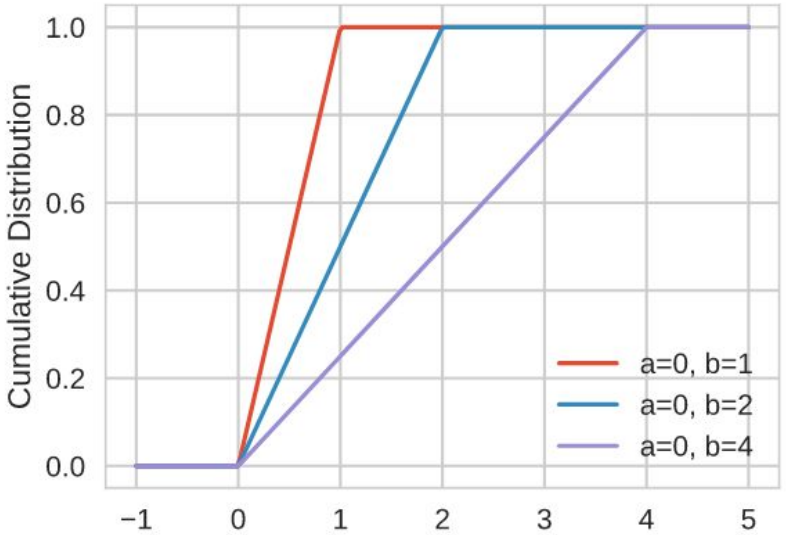
and the distribution function is:

$$F(t) = \begin{cases} 0 & t < a \\ \dfrac{1}{b-a}(t-a) & a < t \leq b \\ 1 & t \geq b \end{cases}$$

Uniform Distribution

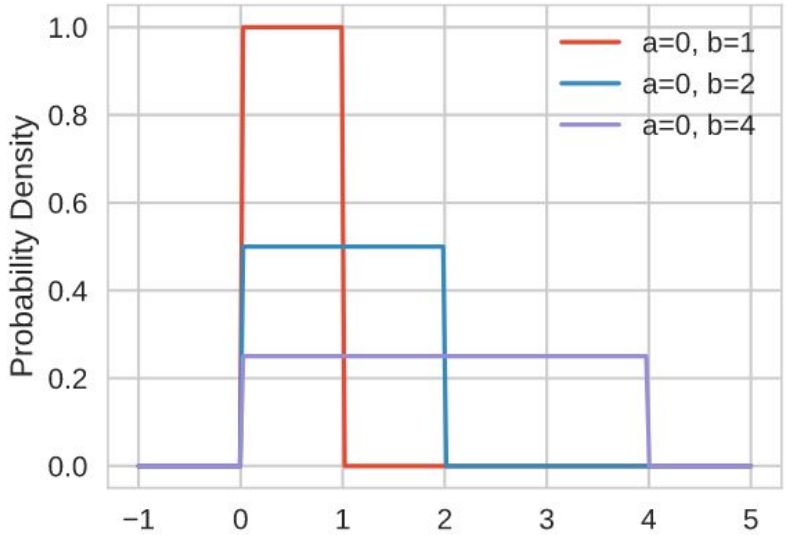
Cumulative Distribution Function

$$F(t) = \begin{cases} 0 & t < a \\ \dfrac{1}{b-a}(t-a) & a < t \leq b \\ 1 & t \geq b \end{cases}$$



Probability Density Function

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



Normal Distribution (Gaussian)

The **Normal Distribution** is of primary importance in probability and statistical theory due to the **Central Limit Theorem** (which we will discuss later in the course).

The probability density function of the normal distribution is:

$$f_Z(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma}\right]$$

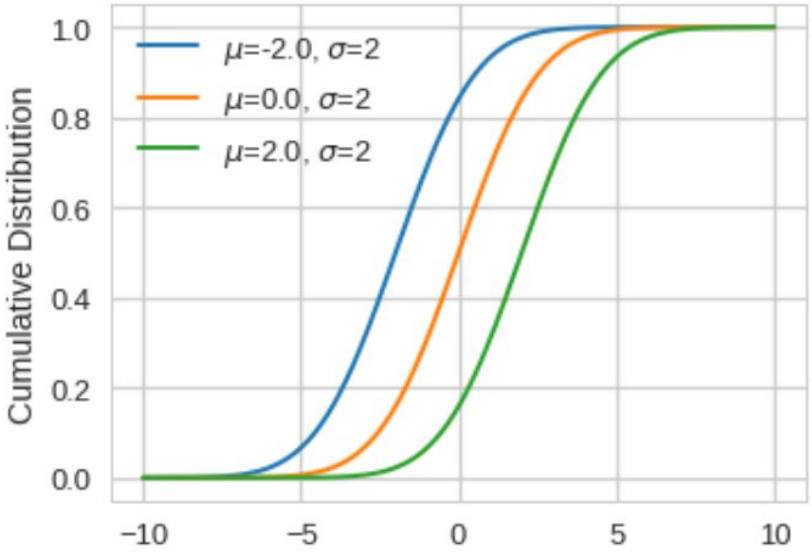
where

μ = mean value of the random variable

σ = standard deviation of the random variable

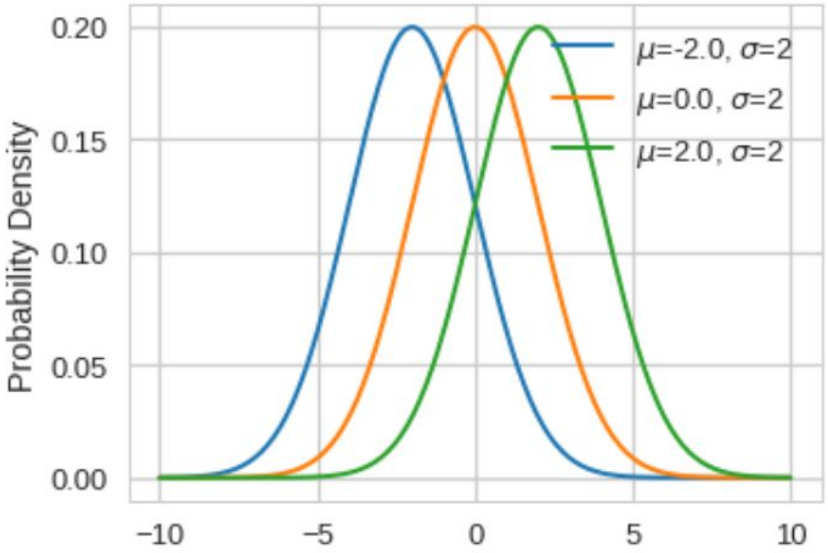
Normal Distribution (Gaussian)

Cumulative Distribution Function



Probability Density Function

$$f_Z(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma}\right]$$

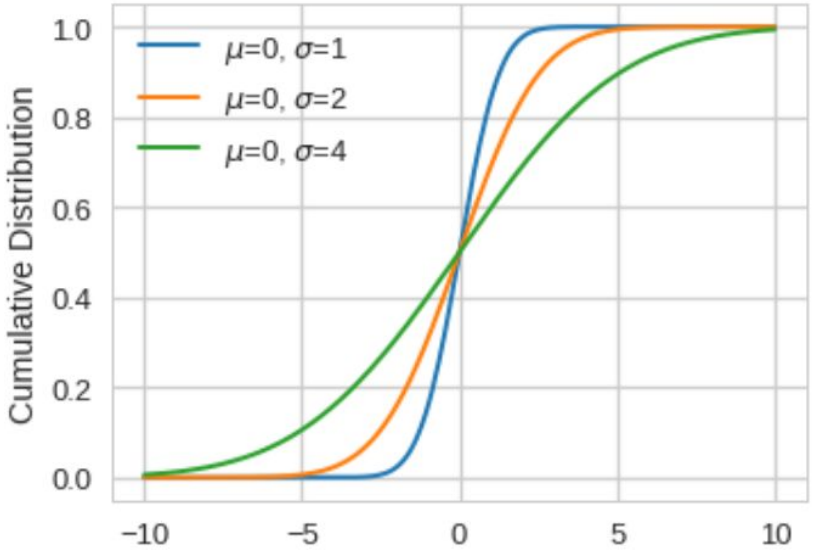


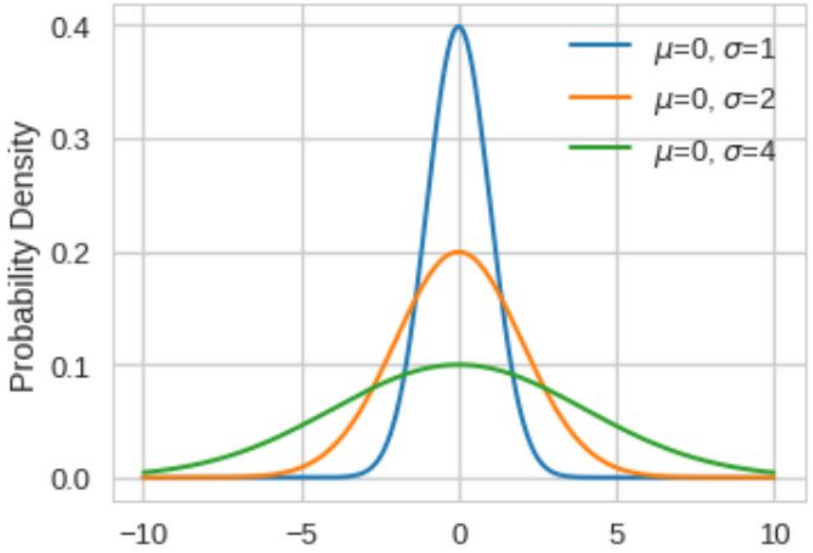
Normal Distribution (Gaussian)

Cumulative Distribution Function

Probability Density Function

$$f_Z(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma}\right]$$





Exponential Distribution

The **Exponential Distribution** is a continuous distribution related to the Poisson distribution.

Poisson: How many events will you observe in a given time?
Exponential: How much time will it take to observe the first event?

probability density function

$$f(t) = \frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right)$$

cumulative distribution function

$$F(x) = 1 - \exp\left(-\frac{t}{\theta}\right)$$

where θ is the average time between events.

Exponential Distribution

It's also common to parameterize with the average rate λ at which the event occurs:

probability density function

$$f(t) = \lambda \exp\Big(-\lambda t\Big)$$

cumulative distribution function

$$F(x) = 1 - \exp\Big(-\lambda t\Big)$$

Exponential Distribution

Students arrive at a local bar and restaurant at a mean rate of 1 student every 10 minutes. What is the probability that the next student will enter the bar within the next 5 minutes?

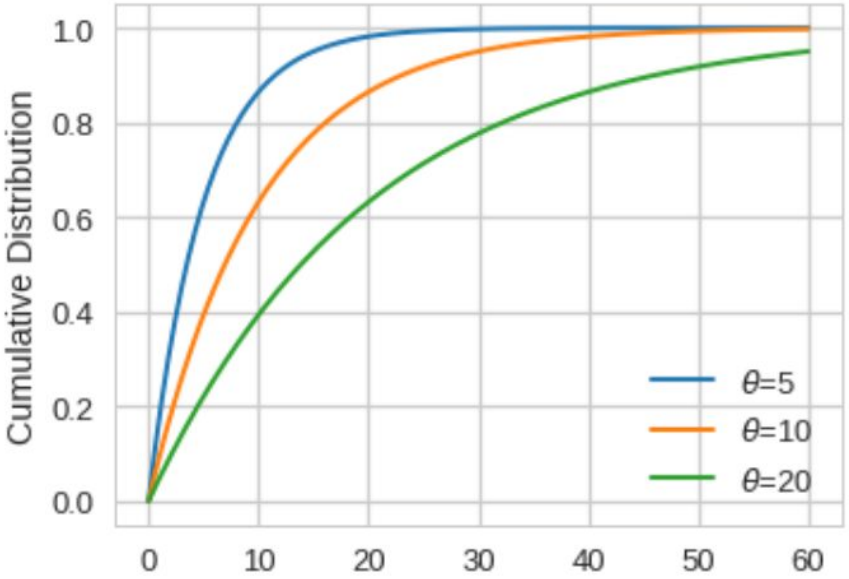
Exponential Cumulative Distribution: $F(x) = 1 - \exp\left(-\frac{t}{\theta}\right)$

$$1 - exp\left(-\frac{5}{10}\right) = 1 - e^{-0.5} = 0.393$$

Exponential Distribution

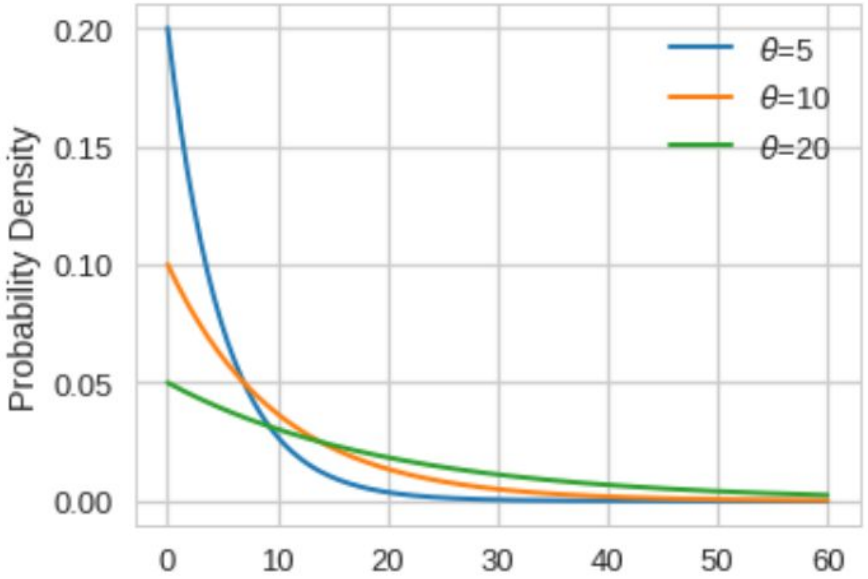
Cumulative Distribution Function

$$F(x) = 1 - \exp\left(-\frac{t}{\theta}\right)$$



Probability Density Function

$$f(t) = \frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right)$$



Check for Understanding

By the end of this lesson, you will be able to:

- ❏ Describe and provide real-world examples of common distributions

Which distribution (and function) would you use to calculate the probability of each of the following events?

1. Observing at most 2 buses at the corner in the next 10 minutes.
2. Observing the next bus arrive at the corner in the next 10 minutes.
3. Rolling a 4 on a 10-sided die.
4. Getting 3 black cards when you draw 5 cards from a deck *without replacement.*
5. Getting 3 black cards when you draw 5 cards from a deck *with replacement.*

Scipy Stats Module

The major tool for working with distributions in python is **scipy.stats**.

Pros:

- Has a LOT of distributions
- Relatively good documentation

Cons:

- The interface design is… not great
- Non-standard parameter names

Scipy Stats Module

Creating a Distribution Object

```
import scipy.stats as stats

# Discrete Distributions
uniform_disc = stats.randint(low=0, high=10) # k = 0, 1, ..., 9
bernoulli = stats.bernoulli(p=0.4)
binomial = stats.binom(n=50, p=0.4)
hypergeom = stats.hypergeom(M=20, n=7, N=12) # non-standard parameters!
poisson = stats.poisson(mu=5) # mu is the same as lambda


# Continuous Distributions
uniform_cont = stats.uniform(loc=0, scale=10) # non-standard parameters!
normal = stats.norm(loc=0.0, scale=1.0) # non-standard parameters!
exponental = stats.expon(loc=2.0) # non-standard parameters!
```

Scipy Stats Module

Calculating a CDF

```
print("P(Binomial(n=50, p=0.4) <= 20) = ", binomial.cdf(20))
print("P(Normal(mu=0.0, sigma=1.0) <= 1.0 = ", normal.cdf(1.0))
```

```
P(Binomial(n=50, p=0.4) <= 20) =  0.5610349320400658
P(Normal(mu=0.0, sigma=1.0) <= 1.0 =  0.8413447460685429
```

There are also methods that calculate a PMF or PDF. All three methods also accept lists or numpy arrays of values...very useful for plotting. :)

Scipy Stats Module

The .rvs (Random ValueS) method samples from a distribution object:

```
print("Ten random draws from a Binomial(n=50, p=0.4): ", binomial.rvs(10))
print("Ten random draws from a Normal(mu=0.0, sigma=1.0): ", normal.rvs(10))
```

```
Ten random draws from a Binomial(n=50, p=0.4):  [16 17 20 22 19 25 19 20 16 17]
Ten random draws from a Normal(mu=0.0, sigma=1.0):  [ 1.7594658   1.46029478 -0.06747751
-1.39694945 -0.20049015 -0.05654447
  0.29456879  1.26254489  2.30070426  0.51292104]
```

These are just a few of the methods available to a distribution object. Browse the scipy.stats documentation for more methods and distributions!