Probability Distributions

Galvanize

Objectives

- Define the concept of a random variable (informally)
- Define the concepts of PDF, PMF, and CDF.
- Relate the concepts of PDF and PMF to the CDF
- Use the CDF of a distribution to compute probabilities of events described by the given distribution.
- Define the following distribution, and give example of situations them

-Bernoulli	-Binomial
-Poisson	=Uniform
-Normal	=Exponential

Random Variables

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

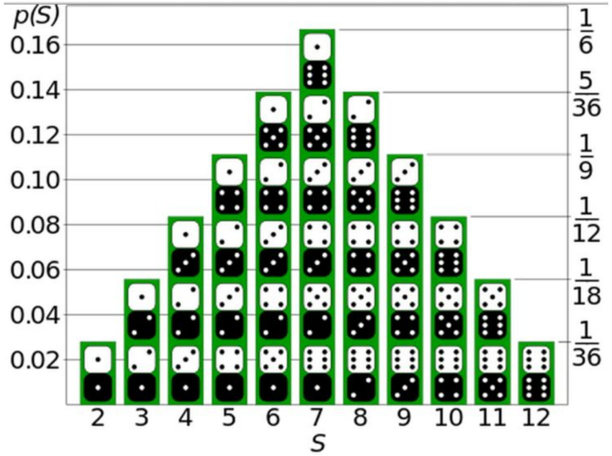
Random Variables

Discrete Case: P(X = i)

X = Sum of two rolled dice

P(X) => Probability Distribution of RV

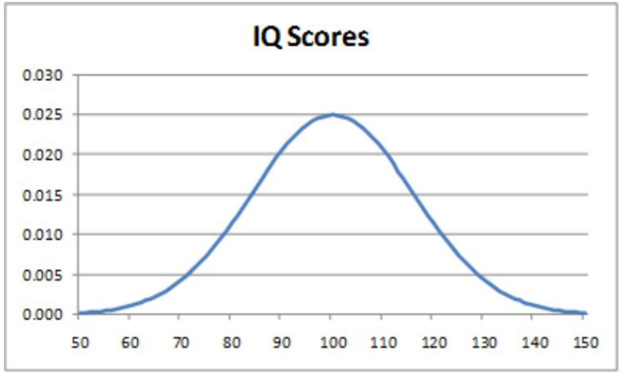
P(X = 12) = Probability sum of rolls is 12



Continuous Case: f(x) or p(x)

X = IQ Score of random individual

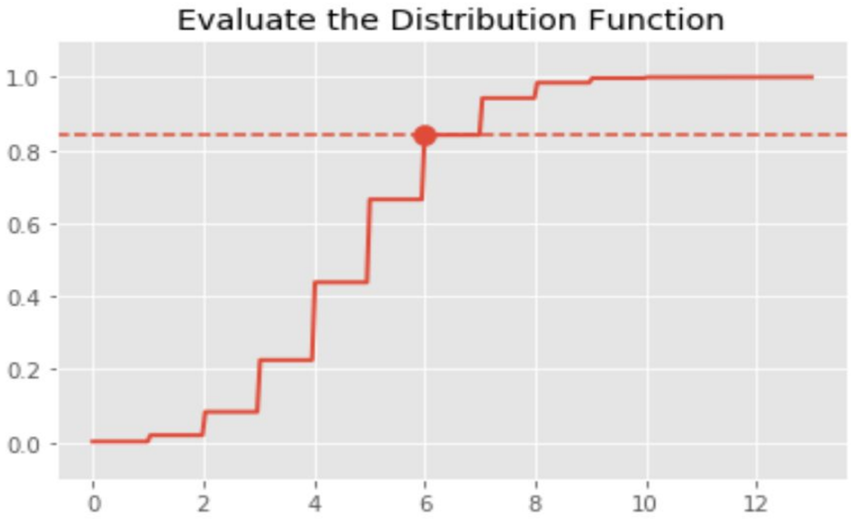
P(X < 120) = 0.909

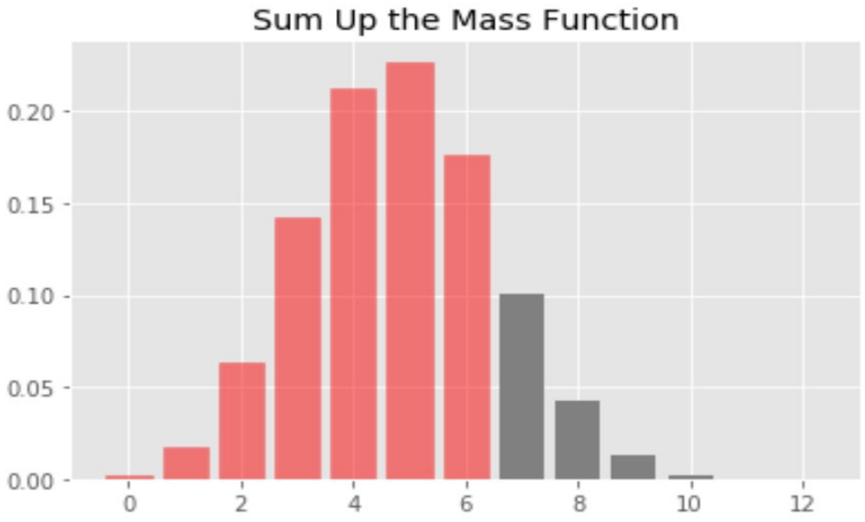


Distributions

The pattern of probabilities of a random variable is called its distribution, and there are a few mathematical do-dads used to describe them. Which do-dad is appropriate depends on whether the random variable generates discrete or continuous values.

Distribution and Mass Functions - Discrete Case





The cdf, or **cumulative distribution** function, often denoted as F(x), is the probability of that random variable **X** having a value <=x.

Given a possible output value from X, the **probability mass function** spits out the probability of that value occuring.

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 list them all here, 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:

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

The 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:

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

Binomial Distribution

The **Binomial distribution** is another common discrete distribution. It 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.

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.

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.

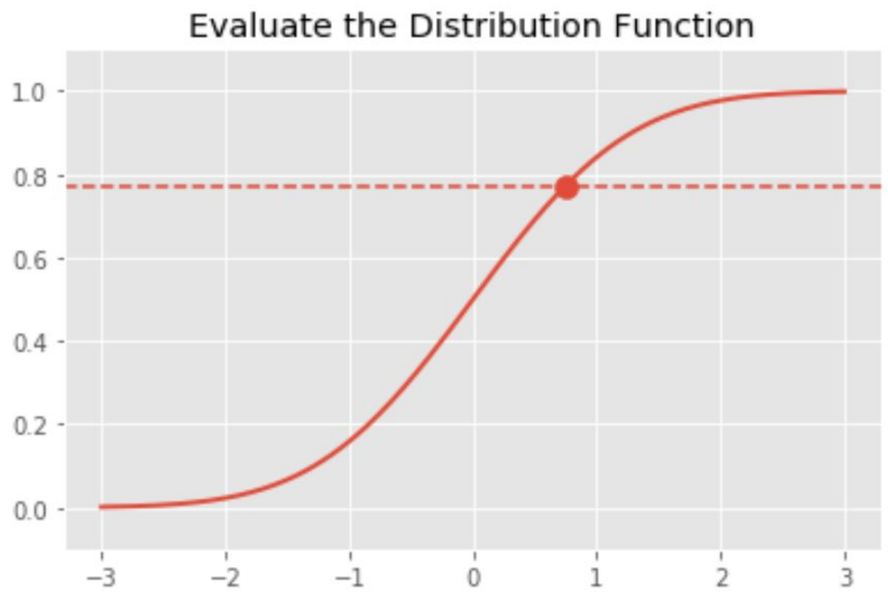
EX: the number of busses that arrive at a stop in an hour, or the number of times a person checks their phone in a day is (approximately) Poisson distributed

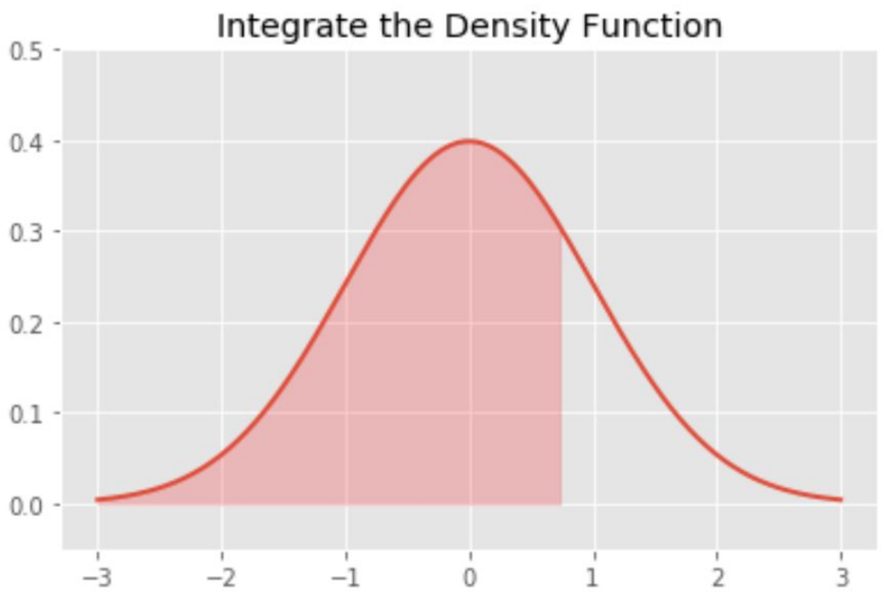
The probability mass function of the Poisson distribution is:

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

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

Distribution and Mass Functions - Continuous Case





Instead of a probability mass function, we now have a **probability density function**. While before, we went between the two by "adding up probability", now we need to use a concept from [calculus](), and [integrate]()

Distribution and Mass Functions - Continuous Case

In the continuous case, our random variable can output any value (sometimes within some range), so the concept of "adding up all the probabilities of possibilities" doesn't make sense.

Fortunately, the definition of the distribution function is exactly the same as in the discrete case:

$$F_X(t) = P(X \leq t)$$

The density function does not tell us the probability that our random variable will assume any specific value, but it does tell us the probability that the output of the random variable will fall into any given range. Again, this connection requires integration:

$$P(a < X \leq b) = \int_a^b f_X(t)dt$$

Summary

When we want to compute probabilities involving some random quantity, we can either:

- Evaluate the distribution function.
- Integrate the density function.

If niether the distribution or density/mass functions are available, then we cannot compute probabilities about the random variable.

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:

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

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 density function of the normal distribution is:

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

Exponential Distribution

The **Exponential Distribution** is a continuous distribution related to the Poisson distribution. Where the poisson distribution describes how many events you will observe if the events happen at a specific rate and you watch for a specific amount of time, the exponential distribution models the *amount of time* you will have to watch until you observe the first event.

For example, the amount of time you have to wait at a bus stop until a bus arives, and the amount of space you have to search before you find a dropped object tends to be exponentially distributed.

Continue...

See probability-distributions.ipynb