Probability Distributions - An Introduction

We'll cover the following

Introduction

We have learned that probability gives us the percent chance of an event occurring. Now, what if we want an understanding of the probabilities of all the possible values in our experiment? This is where probability distributions come into play.

A probability distribution is a function that represents the probabilities of all possible values. This is a very important concept in data science, by specifying the relative chance of all possible outcomes, probability distributions allow us to understand the underlying trends in our data. For example, if we have some missing values in our dataset, we can understand the distribution of our data using probability distributions and then replace missing values with the most likely values.

Random Variables

For the next couple of lessons, we are going to look at some of the most important probability distributions. But before we dive into probability distributions, we need to understand the different types of data we can encounter.

The set of possible values from a random experiment is called a **Random Variable**. Random Variables can be either discrete or continuous:

- **Discrete Data** (a.k.a. discrete variables) can only take specified values.

For example, when we roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.

- **Continuous Data** (a.k.a. continuous variables) can take any value within a range. This range can be finite or infinite. Continuous variables are measurements like height, weight, and temperature.

Types of Probability Distributions

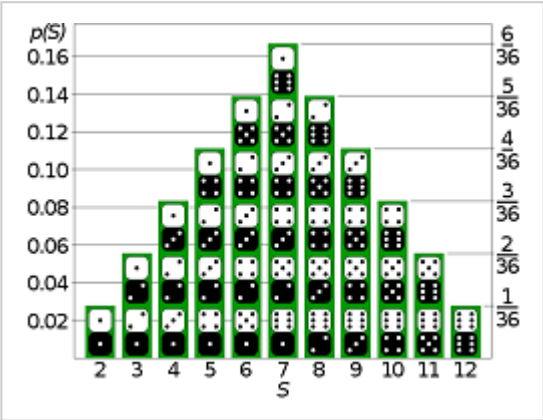
Since probability distributions describe the distribution of the values of a random variable, the kind of variable determines the type of probability distribution we are dealing with. This means that probability distributions can be divided into the following two types:

- Discrete probability distributions for discrete variables
- Probability density functions for continuous variables

Probability Functions

There is just one more concept we need to understand before jumping into the different distributions.

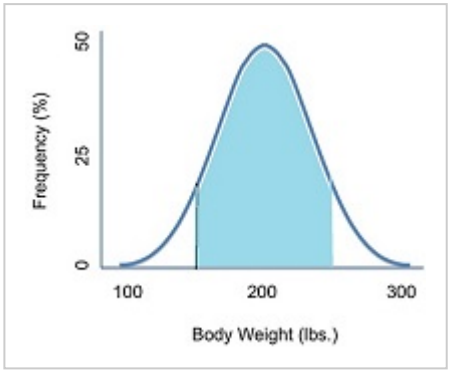
The probability function for a discrete random variable is often called **Probability Mass Function**, while for continuous variables we have the so-called **Probability Density Function** (a.k.a. Probability Distribution Function).



The probability mass function (pmf) p(S), or discrete probability distribution function for discrete variables, D, specifies the probability distribution for the sum S of counts from two dice. For example, the figure shows that p(11) = 2/36 = 1/18. The pmf allows the computation of probabilities of events such as P(S > 9) = 1/12 + 1/18 + 1/36 = 1/6, and all other probabilities in the distribution.

For example, we could have a continuous random variable Y that represents

possible weights of people in a group:



Probability Density Function

The probability density function shows all possible values for Y. For example, the random variable Y could be 100 lbs, 153.2 lbs or 201.9999 lbs.

Why are these functions important?

The probability density function can help us to answer things like: *What is the probability that a person will weigh between 170lbs and 200lbs?*

Now that we have done the ground work, in the next lessons we are going to cover the most important distributions for both discrete and continuous data types.