Statistics for Data Science – Part II: Probability

This article is a second of a series and I will cover the parts of probability that are related to data science. Probability is very important for the data scientist and I will try to answer the following questions:

- Basic Concepts of Probability,
- Probability Distributions.

You may find the first article of this series here.

Types of Probability

Once again, let us start with the Wikipedia definition for probability: "**Probability** is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates impossibility of the event and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur."

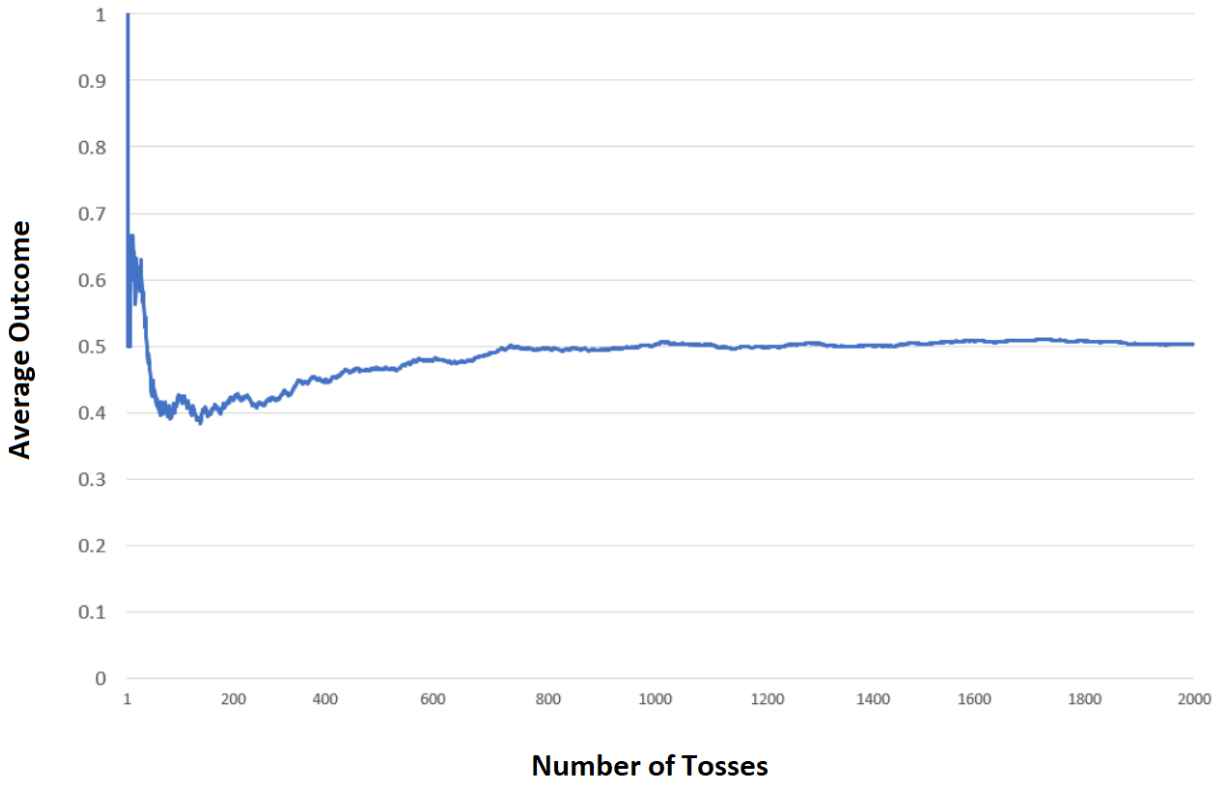
Classical (or **theoretical**) **probability** is used when each outcome in a sample space is equally likely to occur. The classical probability for an event *E* is given by:

$$P(E) = \frac{Number\ of\ outcomes\ in\ event\ E}{Total\ number\ of\ outcomes\ in\ sample\ space}$$

Empirical (or **statistical**) **probability** is based on observations obtained from probability experiments. The empirical probability of an event *E* is the relative frequency of event *E*.

$$P(E) = \frac{Frequency\ of\ event\ E}{Total\ frequency}$$

As the number of times a probability experiment is increased, the empirical probability of an event approaches the theoretical probability of the event. This is called **Law of Large Numbers**.



The third type of probability is **subjective probability**. Subjective probabilities result from intuition, educated guesses and estimates. As an example, given a patient's health and extent of injuries, a doctor may feel that the patient has a 90 % chance of a full recovery.

A probability value cannot be negative or greater than 1. So, the probability of an event E is between 0 and 1 - $0 \leq P(E) \leq 1$. If the probability is 1, the event is certain to occur. If the probability of an event is 0, the event is impossible.

Conditional Probability

Conditional probability is the probability of an event occurring, given that another event has already occurred. The conditional probability of event B occurring, given that event A has occurred, is denoted by $P(B \mid A)$ and is read as "probability of B, given A".

In some cases, one event does not affect the probability of another (e.g., rolling the dice and flip a coin). These two events are independent. Two events are **independent** if the occurrence of one of the events does not affect the probability of the occurrence of the other event.

Two events *A* and *B* are independent if:

$$P(B \mid A) = P(B) \quad \text{or if} \quad P(A \mid B) = P(A)$$

Events that are not independent are **dependent**, which means $P(B) \neq P(B \mid A)$.

The probability that two events occur A and B will occur in sequence is:

$$P(A \text{ and } B) = P(A) \bullet P(B \mid A)$$

If events *A* and *B* are independent, P(*B* | *A*) will be equal to P(*B*).

Permutations

A **permutation** is an **ordered** arrangement of objects. The number of different permutations of *n* distinct objects is *n*!.

In some cases, some of the objects in a group have to be taken and put in an order. Such an order is called a permutation of n objects taken r at a time.

$$_nP_r = \frac{n!}{(n-r)!}$$

where *n* is the total number of objects and *r* is the number of objects selected.

Combinations

The number of ways to choose *r* objects from a group of *n* objects without regard to order is called the number of **combinations** of *n* objects taken *r* at a time.

$$_nC_r = \frac{n!}{r!(n-r)!}$$

where *n* is the total number of objects and *r* is the number of objects selected.

Probability Distributions

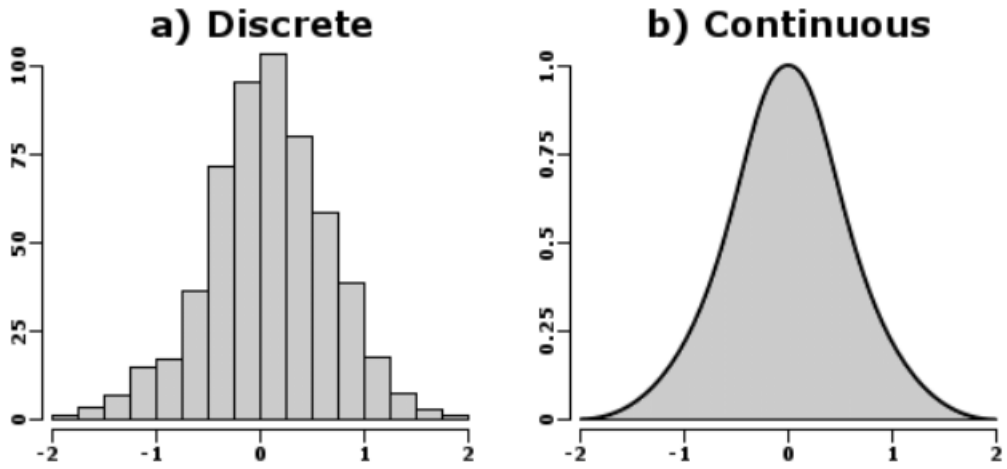
In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. Another definition is, a probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

A **random variable *x*** is a numerical value associated with each outcome of a probability experiment. There are two types of random variables: discrete and continuous.

A random variable is **discrete** if it has a finite or countable number of possible outcomes.

A random variable is **continuous** if it has an uncountable number of possible outcomes, represented by an interval on the number line.

An example to show the difference between the two types of variables may be the number of calls (**discrete**) a salesperson makes per day vs. time spent (**continuous**) by the salesperson on making phone calls.



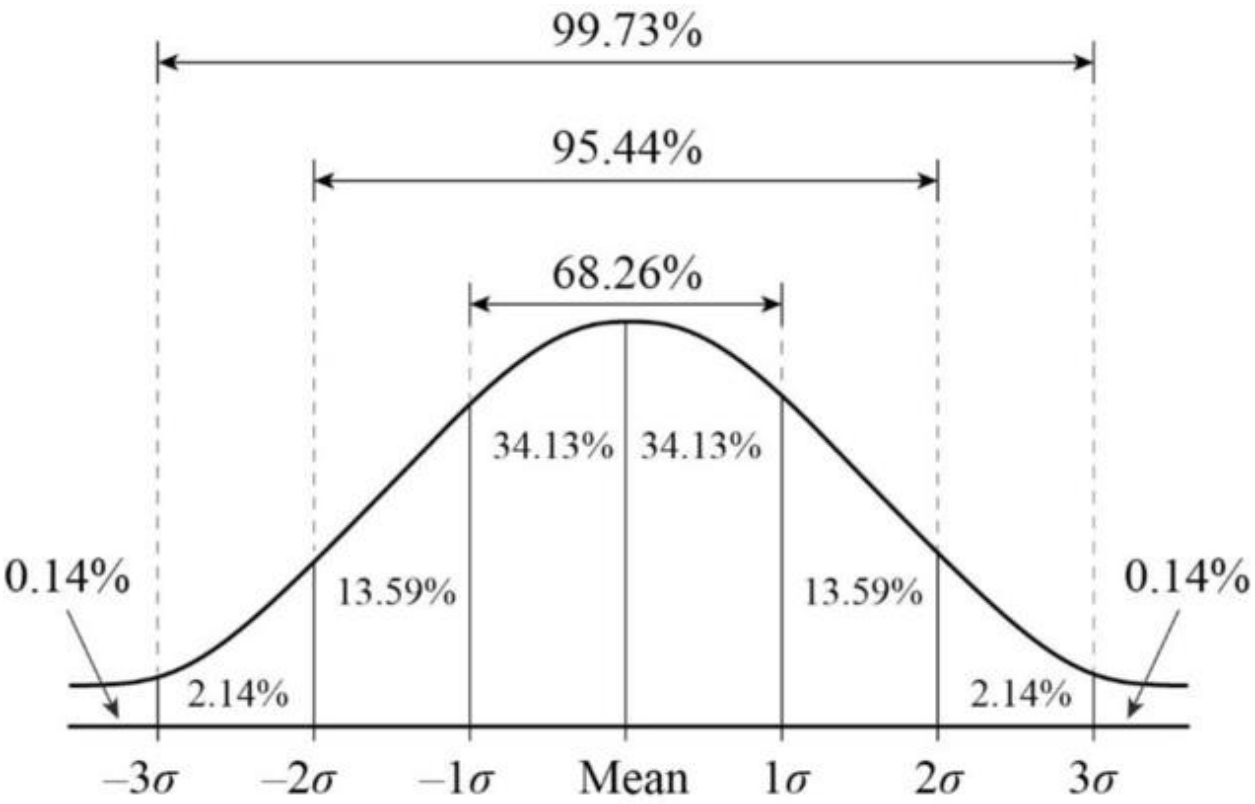
Discrete Vs Continuous

Normal Distribution

There are many different classifications of probability distributions. Some of them include the normal distribution, chi square distribution, binomial distribution, and Poisson distribution. The most commonly used distribution is the normal distribution, which is used frequently in finance, investing, science, and engineering.

Considering the definition of a continuous random variable, it has an infinite number of possible values that can be represented by an interval and its probability distribution is called a **continuous probability distribution**. The most important continuous probability distribution in statistics is the normal distribution. Normal distributions can be used to model many sets of measurements in nature, industry and business.

A **normal distribution** is a continuous probability distribution for a random variable. The graph of a normal distribution is called the normal curve.



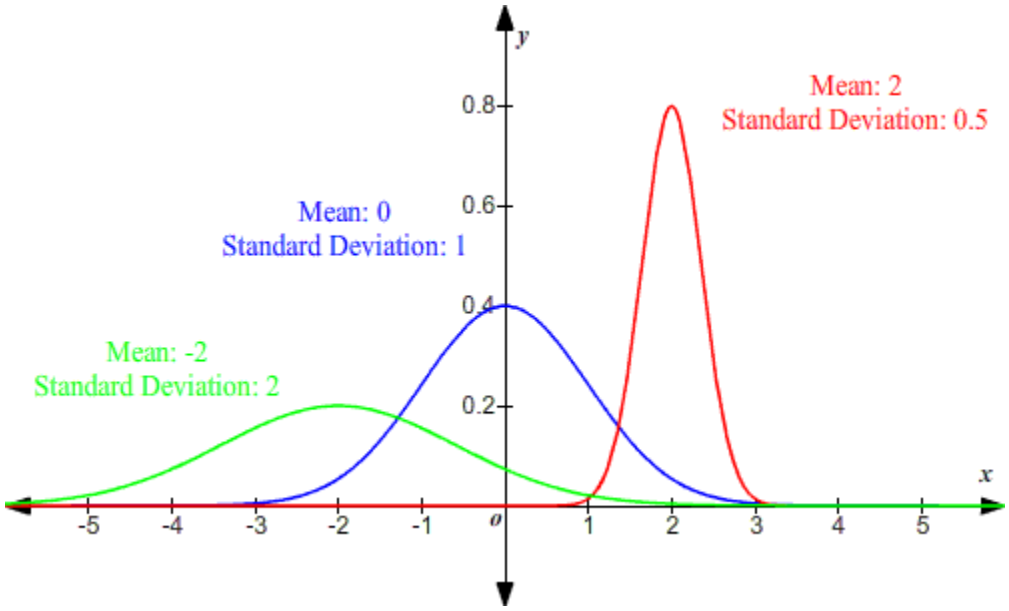
A normal distribution has the following properties:

1. The mean, median and mode are equal.
2. The normal curve is bell-shaped and is symmetric about the mean.
3. The total area under the normal curve is equal to one.
4. The normal curve approaches, but never touches the x-axis, as it extends farther and farther away from the mean.

A **probability density function** (**pdf**) defines the continuous probability distribution. A normal curve with mean μ and standard deviation σ can be graphed using the normal probability density function.

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

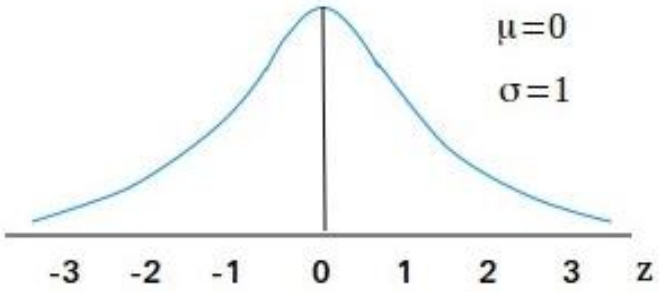
A normal distribution can have any mean and any positive standard deviation. These two parameters completely determine the shape of the normal curve. The mean gives the location of the line of symmetry and the standard deviation describes how much the data are spread out.



There are infinitely many normal distributions, each with its own mean and standard deviation. The normal distribution with a mean of 0 and a standard deviation of 1 is called the **standard normal distribution**.

The horizontal axis of the standard normal distribution graph corresponds to z-scores. A z-score is a measure of position that indicates the number of standard deviations a value lies from the mean. The basic formula for the z-score is:

$$z = \frac{Value - Mean}{Standard\ Deviation} = \frac{x - \mu}{\sigma}$$



If each value of a normally distributed random variable *x* is transformed into a z-score, the result will be the standard normal distribution.

Standard normal distribution has the following properties:

1. The cumulative area is close to 0 for z-scores close to z = -3.49.
2. The cumulative area increases as the z-scores increase.

3. The cumulative area for z = 0 is 0.5000.
4. The cumulative area is close to 1 for z-scores close to z = 3.49.

Considering the formula above, in order to transform a standard z-score to a data value x in a given population, the formula below is used:

$$x = \mu + z\sigma$$

The next article will be about **Statistical Inference**.

You can access to this article and similar ones here.

Photo by Edge2Edge Media on Unsplash