What is a probability distribution?

A probability distribution is an idealized frequency distribution.

A **frequency distribution** describes a specific sample or dataset. It's the number of times each possible value of a variable occurs in the dataset.

The number of times a value occurs in a sample is determined by its **probability** of occurrence. Probability is a number between 0 and 1 that says how likely something is to occur:

- 0 means it's impossible.
- 1 means it's certain.

The higher the probability of a value, the higher its frequency in a sample.

More specifically, the probability of a value is its relative frequency in an infinitely large sample.

Infinitely large samples are impossible in real life, so probability distributions are theoretical. They're idealized versions of frequency distributions that aim to describe the population the sample was drawn from.

Probability distributions are used to describe the populations of real-life variables, like coin tosses or the weight of chicken eggs. They're also used in hypothesis testing to determine *p* values.

Discrete probability distributions

A **discrete probability distribution** is a probability distribution of a categorical or discrete variable.

Discrete probability distributions only include the probabilities of values that are possible. In other words, a discrete probability distribution doesn't include any values with a probability of zero. For example, a probability distribution of dice rolls doesn't include 2.5 since it's not a possible outcome of dice rolls.

The probability of all possible values in a discrete probability distribution add up to one. It's certain (i.e., a probability of one) that an observation will have one of the possible values.

Probability tables

A **probability table** represents the discrete probability distribution of a categorical variable. Probability tables can also represent a discrete variable with only a few possible values or a continuous variable that's been grouped into class intervals.

A probability table is composed of two columns:

- The values or class intervals

- Their probabilities

Continuous probability distributions

A **continuous probability distribution** is the probability distribution of a continuous variable.

A continuous variable can have any value between its lowest and highest values. Therefore, continuous probability distributions include every number in the variable's range.

The probability that a continuous variable will have any specific value is so infinitesimally small that it's considered to have a probability of zero. However, the probability that a value will fall within a certain interval of values within its range is greater than zero.

Probability density functions

A **probability density function** (PDF) is a mathematical function that describes a continuous probability distribution. It provides the **probability density** of each value of a variable, which can be greater than one.

A probability density function can be represented as an equation or as a graph.

In graph form, a probability density function is a curve. You can determine the probability that a value will fall within a certain interval by calculating the area under the curve within that interval. You can use reference tables or software to calculate the area.

The area under the whole curve is always exactly one because it's certain (i.e., a probability of one) that an observation will fall somewhere in the variable's range.

A **cumulative distribution function** is another type of function that describes a continuous probability distribution.