Model data with the normal distribution

Recently, you've been learning about continuous probability distributions, and how they help data professionals model their data. Recall that continuous probability distributions represent continuous random variables, which can take on all the possible values in a range of numbers. Typically, these are decimal values that can be measured, such as height, weight, time, or temperature. For example, you can keep on measuring time with more accuracy: 1.1 seconds, 1.12 seconds, 1.1257 seconds, and so on.

In this course, we focus on a single continuous probability distribution: the normal distribution. In this reading, you'll learn more about the main characteristics of the normal distribution, and how the distribution can help you model your data.

Continuous probability distributions

Before we get to the specific attributes of the normal distribution, let's discuss some general features of all continuous probability distributions.

Probability Density and Probability

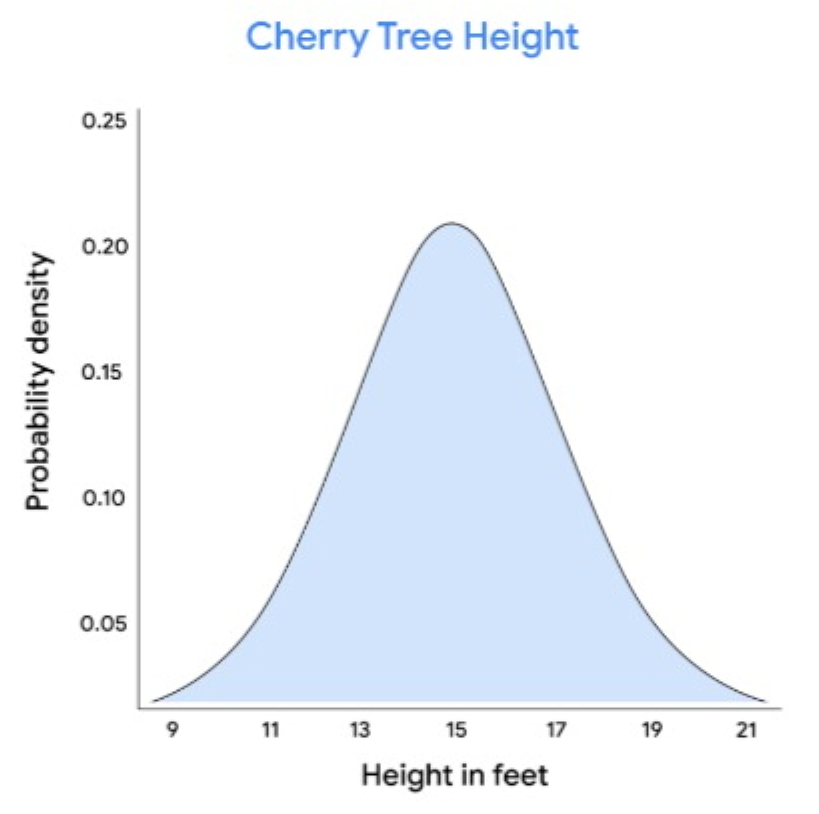
A probability function is a mathematical function that provides probabilities for the possible outcomes of a random variable.

There are two types of probability functions:

- Probability Mass Functions (PMFs) represent discrete random variables
- Probability Density Functions (PDFs) represent continuous random variables

A probability function can be represented as an equation or a graph. The math involved in probability functions is beyond the scope of this course. For now, it's important to know that the graph of a PDF appears as a curve. You've learned about the bell curve, which refers to the graph for a normal distribution.

As an example, imagine you have data on a random sample of cherry trees. Assume that the heights of the cherry trees are approximately normally distributed with a mean of 15 feet and a standard deviation of 2 feet.



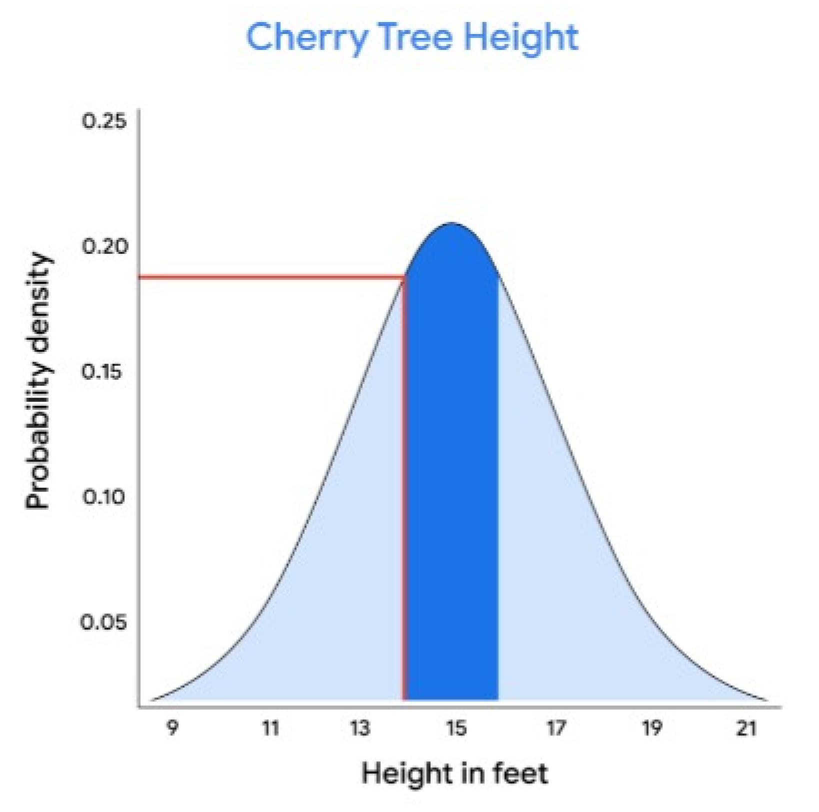
On a continuous distribution, the x-axis refers to the value of the variable you're measuring - in this case, cherry tree height. The y-axis refers to probability density. Note that probability density is not the same thing as probability.

The probability distribution for a continuous random variable can only tell you the probability that the variable takes on a range or interval of values. This is because a continuous random variable may have an infinite number of possible values. For instance, the height of a randomly chosen cherry tree could measure 15 feet, or 15.1 feet, or 15.175 feet, or 15.175245 feet, and so on.

Let's say you want to know the probability that the height of a randomly chosen cherry tree is exactly 15.1 feet. Because the height of the tree could be any decimal value in a given interval, the probability that the tree is exactly any single value is essentially zero.

So, for continuous distributions, it only makes sense to talk about the probability of intervals, such as the interval between 14.5 feet and 15.5 feet.

To find the probability of an interval, you calculate the area under the curve that corresponds to the interval. For example, the probability of a cherry tree having a height between 14.5 feet and 15.5 feet is equal to the area under the curve between the values 14.5 and 15.5 on the x-axis. This area appears as the shaded rectangle in the center of the graph.



In this case, the area of the rectangle is around 0.20. So, there is a 20% chance that the height of a randomly chosen cherry tree is between 14.5 feet and 15. 5 feet.

Note: data professionals typically use statistical software to calculate probabilities on a continuous distribution.

The normal distribution

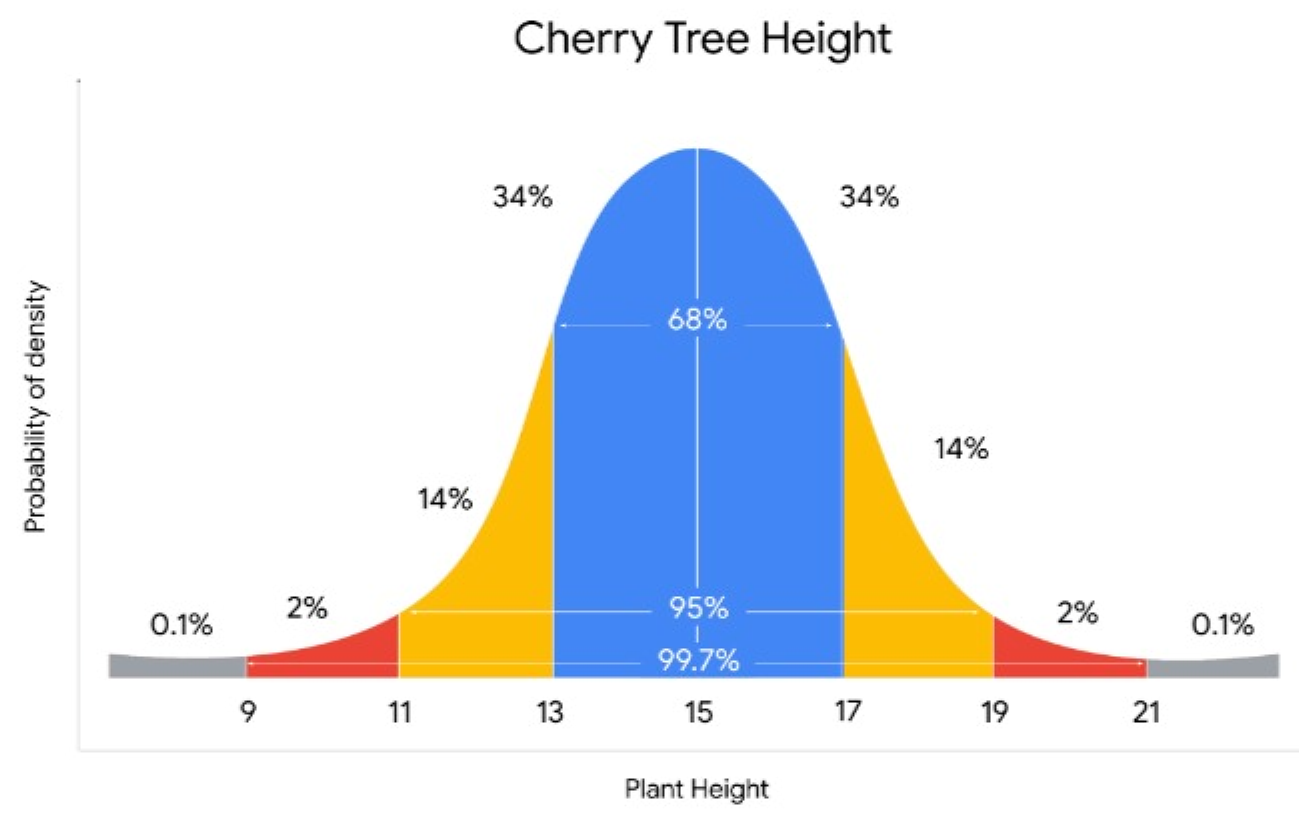
The normal distribution is a continuous probability distribution that is symmetric about the mean and bell-shaped. It is also known as the Gaussian distribution, after the German mathematician Carl Gauss, who first described its formula. The normal distribution is often called the bell curve because its graph has the shape of a bell, with a peak at the center and two downward sloping sides.

The normal distribution is the most common probability distribution in statistics because so many different kinds of datasets display a bell-shaped curve. For example, if you randomly sample 100 people, you will discover a normal distribution curve for continuous variables such as height, weight, blood pressure, shoe size, test scores, and more.

All normal distributions have the following features:

- The shape is a bell curve
- The mean is located at the center of the curve
- The curve is symmetrical on both sides of the mean
- The total area under the curve equals 1

Let's use our cherry tree example to clarify the features of the normal distribution. Recall that the mean height is 15 feet with a standard deviation of 2 feet.



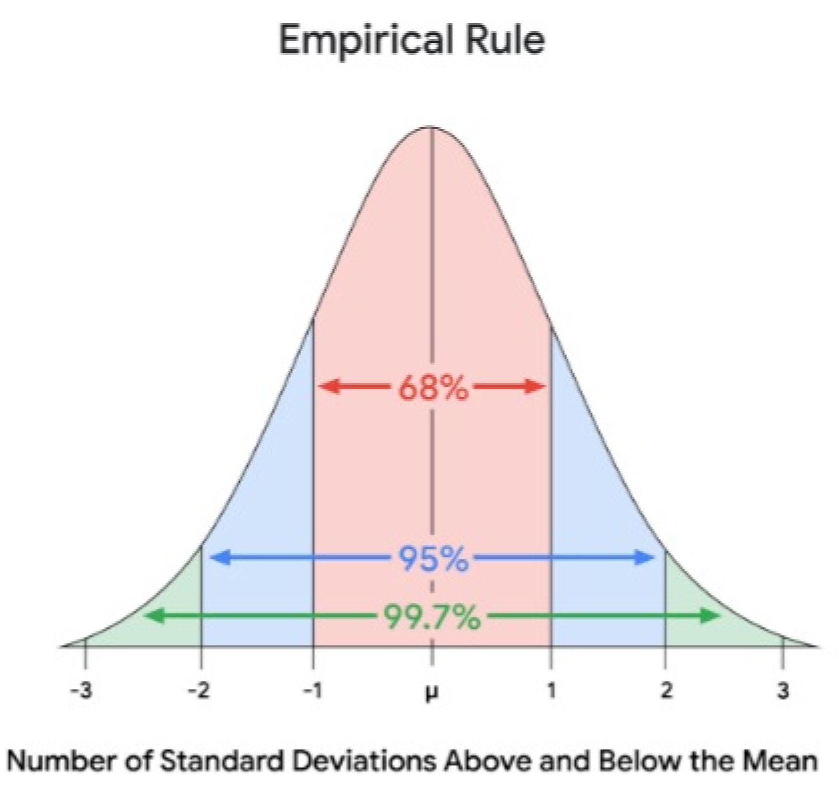
You may notice the following features of the normal curve:

- The mean is located at the center of the curve, and is also the peak of the curve. The mean height of 15 feet represents the most probable outcome in the dataset
- The curve is symmetrical about the mean. 50% of the data is above the mean, and 50% of the data is below the mean
- The farther a point is from the mean, the lower the probability of those outcomes. The points farthest from the mean represent the least probable outcomes in the dataset. These are trees that have more extreme heights, either short or tall
- The area under the curve is equal to 1. This means that the area under the curve accounts for 100% of the possible outcomes in the distribution

The empirical rule

You may also notice that the values on a normal curve are distributed in a regular pattern, based on their distance from the mean. This is known as **the empirical rule.** The rule states that for a given dataset with a normal distribution:

- 68% of values fall within 1 standard deviation of the mean
- 95% of values fall within 2 standard deviations of the mean
- 99.7% of values fall within 3 standard deviations of the mean



If you apply the empirical rule to our cherry tree example, you learn the following:

- Most trees, or 68%, will fall within 1 standard deviation of the mean height of 15 feet. This means that 68% of trees will measure between 13 feet and 17 feet, or 2 feet below the mean and 2 feet above the mean
- 95% of trees will measure between 11 feet and 19 feet, or within 2 standard deviations from the mean
- Almost all trees, or 99.7%, will measure between 9 feet and 21 feet, or within 3 standard deviations of the mean

The empirical rule can give you a quick estimate of how the values in a large dataset are distributed. This saves time and helps you better understand your data.

Knowing the location of your values on a normal distribution is also useful for detecting outliers. Recall that an outlier is a value that differs significantly from the rest of the data. Typically, data professionals consider values that lie more than 3 standard deviations below or above the mean to be outliers. It's important to identify outliers because some extreme values may be due to errors in data collection or data processing, and these false values may skew your results.

Key takeaways

As a data professional, you'll likely use the normal distribution to identify significant patterns in a wide variety of datasets. Understanding the normal distribution is also important for more advanced statistical methods, such as hypothesis testing and regression analysis, which you'll learn about later on.

Resources for more information

To learn more about continuous probability distributions and the normal distribution, check out the following resources:

- This article from Duke University provides a useful summary of the main characteristics of the normal distribution

Mark as completed

Like Dislike Report an issue