The Normal Distribution

Introduction

For data scientists and machine learning professionals, the normal (also referred to as Gaussian) distribution stands out as one of the most commonly used distribution models. This lesson provides an introduction to the normal distribution, its characteristics, and its significance in data science.

Objectives

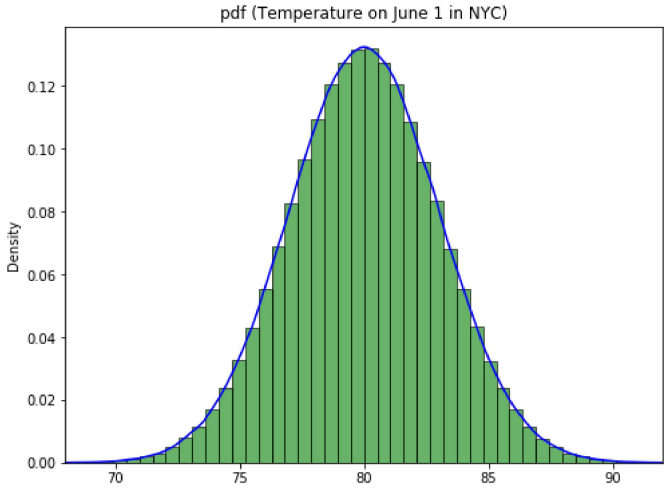
You will be able to:

- List the unique characteristics of a normal distribution
- Identify real world instances of things that follow a Gaussian distribution
- Use `numpy` to generate a random normal distribution

The Normal Distribution

The normal distribution is the most important and most widely used distribution in statistics and data science. It is also called the "bell curve," due to its bell shape, or the "Gaussian curve" after the German mathematician Karl Friedrich Gauss.

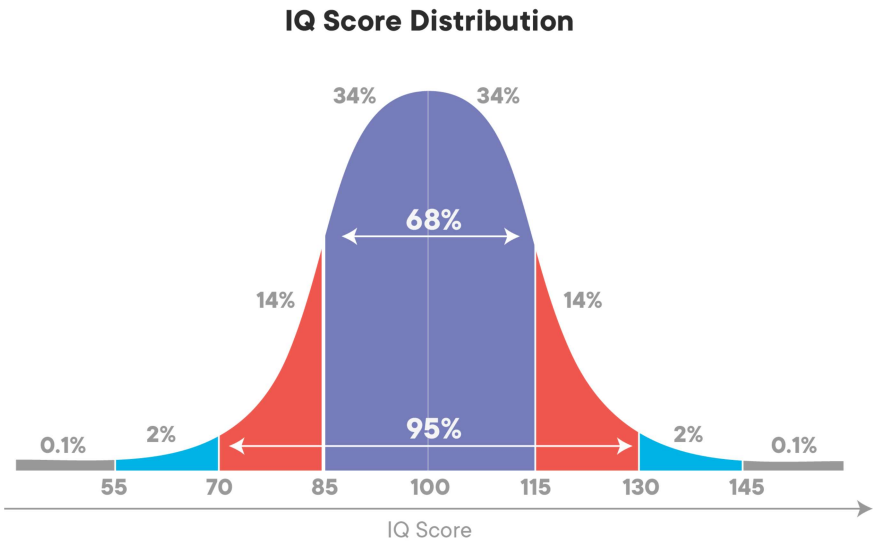
Recall our NYC weather distribution. This is a classic example of a normal distribution. The idea is that there is sort of an expectation around what the temperature will be on June 1 (80 degrees Fahrenheit) and that temperatures much lower or much higher are less likely the further they move away from this expected temperature. This type of behavior is present in many phenomena, as you'll see later.



The normal distribution is **a continuous distribution**. In practice though, you'll see many discrete distributions that follow a bell curve shape:

- The observed values are actually discrete. For example, human IQ follows a normal distribution, but IQ is only specified up to the unit digit level, e.g. an IQ of 90, 91, or 92.
- The values in our distribution are actually continuous (e.g. our temperature example) but recorded up to a certain constant because there is (obviously) no "exact" thermometer that measures temperature up to an infinite amount of digits.

Even though the IQ level is not actually recorded as a continuous variable, you'll see that the distribution is generally represented as a smooth curve!



The Probability Density Function

The probability density function equation for the normal distribution is given by the following expression:

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

Here,

- μ is the mean
- σ is the standard deviation
- $\pi \approx 3.14159$
- $e \approx 2.71828$

Don't worry if your head is spinning right now. Don't worry about the formula, what you really need to remember is that:

- A normal distribution has 2 key parameters, μ and σ, which define the mean and the spread of the distribution, respectively.
- If you apply our formulas of expected values and variance seen in the PDF lesson before, where $X \sim N(x)$:
  - $E(X) = \int_{-\infty}^{+\infty} p(x)x dx = \mu$
  - $E((X-\mu)^2) = \int_{-\infty}^{+\infty} p(x)(x-\mu)^2 dx = \sigma^2$

  where μ and σ are as specified in the formula of $N(x)$

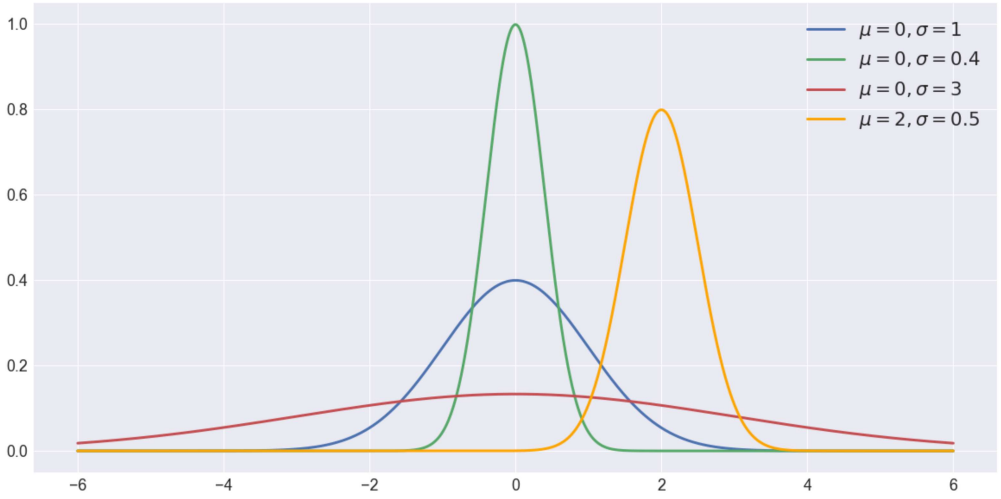
Mean and Standard Deviation

Here is a first simple definition for normal distributions:

> **The Normal Distribution is symmetrical and its mean, median and mode are equal.**

A normal distribution is **centered around its mean**, so the distribution is not skewed (you'll have a chance to learn more about skewness later). This doesn't mean that normal distributions cannot appear in different shapes and forms. How exactly the distribution behaves depends on the 2 key parameters, as specified before: the **mean** and the **standard deviation**.

The following figure shows four normal distributions:

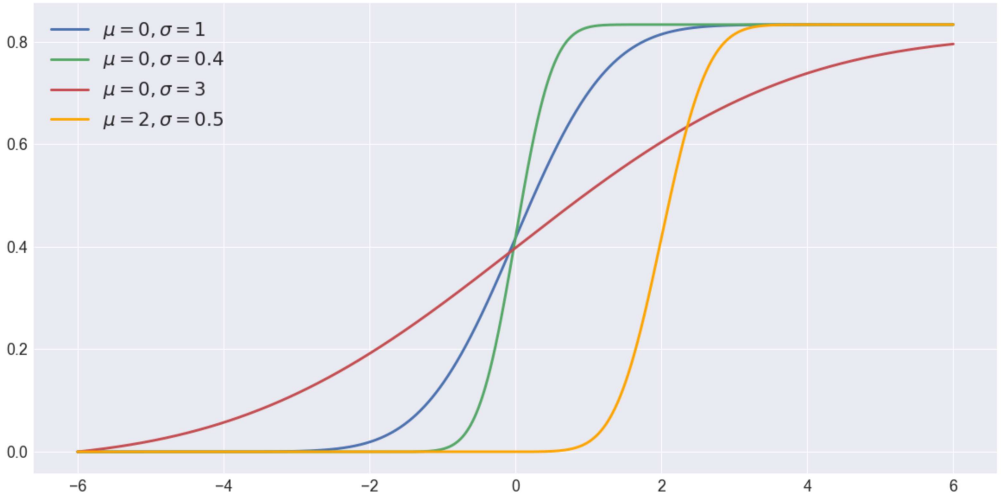


- The green distribution has a mean of 0 and a standard deviation of 0.4
- The distribution in blue has a mean of 0 and a standard deviation of 1.
- The distribution in red has a mean of 0 and a high spread with standard deviation of 3.
- The orange distribution has a mean of 2 and a standard deviation of 0.5.

All of these distributions have the following properties in common:

- They are symmetric around the mean,
- They have relatively higher densities of values at the center of the distribution and relatively lower density in the tails

The CDFs of these distributions are shown below:

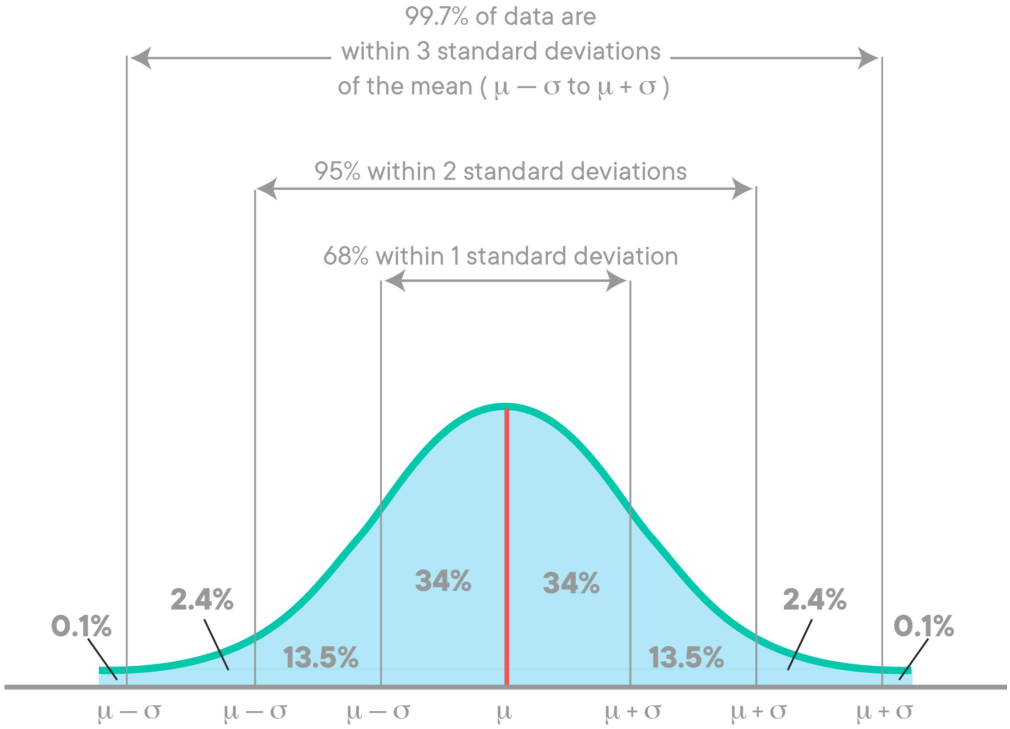


Some More Characteristics of the Normal Distribution

Let's summarize the key characteristics of the normal distribution below:

- Normal distributions are symmetric around their mean
- The mean, median, and mode of a normal distribution are equal
- The area under the bell curve is equal to 1.0
- Normal distributions are denser in the center and less dense in the tails
- Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
- Around 68% of the area of a normal distribution is within *one standard deviation* of the mean ($(\mu - \sigma)$ to $(\mu + \sigma)$)
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean ($(\mu - 2\sigma)$ to $(\mu + 2\sigma)$).

Let's look at the image below to get a better sense of the two last statements. In this image, the spread is differentiated between levels of deviation.



This forms a 68-95-99.7 rule, i.e., 68% values of a normal distribution are within 1 standard deviation of the mean, 95% within 2 standard deviations and 99.7 % within 3 standard deviations. This rule is also called the empirical rule. Normally distributed data is considered ideal for analysis due to this simplicity of description. Values in the extreme of tails (more than 3 standard deviations) can be considered "interesting events" as their probability of occurrence is very low (1 occurrence in about ~300!). In other cases, you'll consider them as outliers due to noise or error of measurement. It all depends on your analysis question.

Keeping this in mind, have another look at the IQ distribution and identify "extreme events" in terms of IQ!

Why So Popular?

In this section, you'll learn about some reasons why normal distributions are so popular among data scientists:

Ubiquitous in Natural Phenomena

An amazingly vast number of natural processes naturally follow the normal distribution. A simple normal distribution gives the best model approximation for natural processes like weight, height, blood pressure, etc. Errors committed during some measurements are also found to be normally distributed so they can be modeled and isolated with ease. The income, expenditure and other social attributes of masses are often normally distributed as well.

Central Limit Theorem

The Central Limit Theorem states:

> When you add **a large number** of independent random variables, irrespective of the original distribution of these variables, **their sum tends towards a normal distribution**.

The theorem provides a reason why many natural phenomena follow a normal distribution.

The key takeaway from the central limit theorem is that it allows different distributions to be processed as a normal distribution, even when they do not fulfill the normality requirements shown above. We'll discuss this further when we talk about hypothesis testing.

Simplified Computation

When undergoing transformations, a number of distributions tend to change their nature and may result in a totally new distribution. With normal distributions, we can add random variables, take their product or apply any other advanced transformations (like Fourier transformations or Convolutions) - the resulting distribution will always be normal.

For every normal model approximation, there may exist a complex multi-parameter distribution that gives a better approximation than the normal distribution. Even then, a normal distribution is often the preferred distribution to use because it makes the math a lot simpler!

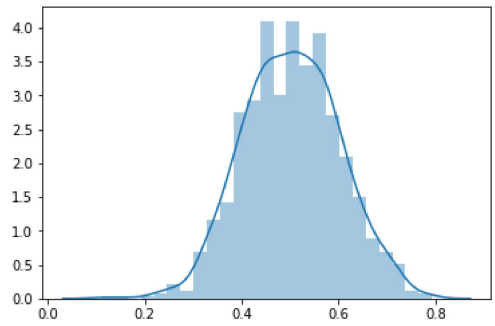
Normal Distributions in Python

In Python, the NumPy module provides a ton of methods to generate and inspect random variables.

You can generate a random normal distribution by providing its parameters μ and σ (mean and sd) to `np.random.norm()`, along with n (number of values to be generated for the normal distribution).

In [2]:
```
import numpy as np
import seaborn as sns

mu, sigma = 0.5, 0.1
n = 1000
s = np.random.normal(mu, sigma, n)
sns.distplot(s);
```



The density function of a normal distribution can also be plotted using a matplotlib *line plot* and using the formula given above. You'll try to do this in the next lab.

Summary

This lesson provides an introduction to normal distributions, the most common distribution family in the field of statistics and data analysis. You learned about the key characteristics of normal distributions, their density function based on mean and standard deviations, and briefly discussed the reasons behind their ubiquitous nature.