The Standard Normal Distribution

Introduction

In this lesson, we will introduce a special case of normal distributions: "The Standard Normal Distribution".

Objectives

You will be able to:

- Compare and contrast the normal and standard normal distribution
- Calculate and interpret the z-score (standard score) for an observation from normally distributed data
- Convert between a normal and a standard normal distribution

What is a Standard Normal Distribution?

Previously, you learned about the normal (or Gaussian) distribution, which is characterized by a bell shape curve. We also identified the mean and standard deviation as the defining parameters of this distribution. As mentioned before, normal distributions do not necessarily have the same means and standard deviations.

The standard normal distribution, however, is a **special case** of the normal distribution. The Standard Normal Distribution is a normal distribution with a mean of 0 and a standard deviation of 1.
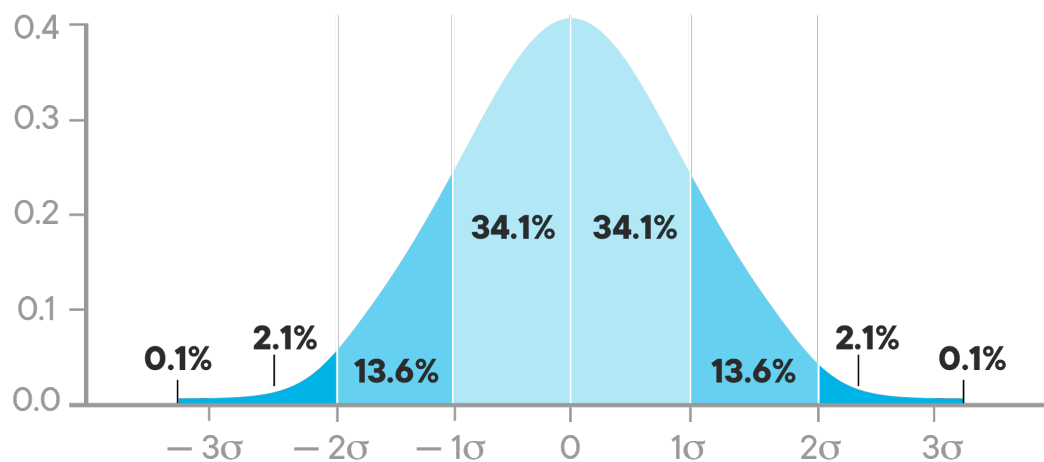
Plotting a continuous cumulative distribution function for the standard normal distribution, the CDF would look like this:

CDF for a normal distribution with $\mu = 0$ and $\sigma = 1$
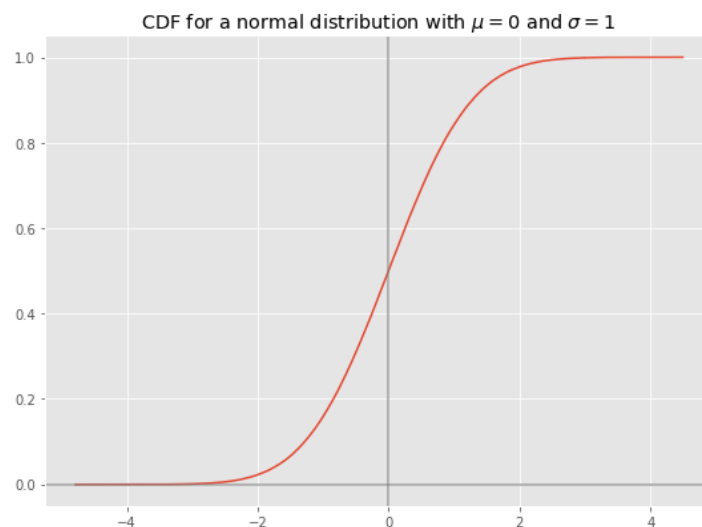
Thinking back to the standard deviation rule for normal distributions:

- 68% of the area lies in the interval of 1 standard deviation from the mean, or mathematically speaking, 68% is in the interval $[\mu - \sigma, \mu + \sigma]$
- 95% of the area lies in the interval of 2 standard deviations from the mean, or mathematically speaking, 95% is in the interval

- 99% of the area lies in the interval of 3 standard deviations from the mean, or mathematically speaking, 99% is in the interval $[(\mu - 3\sigma), (\mu + 3\sigma)]$

With a $\mu = 0$ and $\sigma = 1$, this means that for the standard normal distribution:

- 68% of the area lies between -1 and 1.
- 95% of the area lies between -2 and 2.
- 99% of the area lies between -3 and 3.

This simplicity makes a standard normal distribution very desirable to work with. The exciting news is that you can very easily **transform** any normal distribution to a standard normal distribution!

Standard Score (z-score)

The standard score (more commonly referred to as a z-score) is a very useful statistic because it allows us to:

1. Calculate the probability of a certain score occurring within a given normal distribution and
2. Compare two scores that are from different normal distributions.

Any normal distribution can be converted to a standard normal distribution and vice versa using this equation:

$$z = \frac{x - \mu}{\sigma}$$

Here, x is an observation from the original normal distribution, μ is the mean and σ is the standard deviation of the original normal distribution.

The standard normal distribution is sometimes called the z-distribution. A z-score always reflects the number of standard deviations above or below the mean.

An example

Imagine some test results follow a normal distribution with a mean score of 50 and a standard deviation of 10. One of the students scored a 70 on the test. Using this information and z-scores makes it easy to tell how she performed in terms of standard deviations from the mean. Converting a test score of 70 to a z-score, an x of 70 would be, in this case:

$$z = \frac{70 - 50}{10} = 2$$

By transforming the test result of 70 to a z-score of 2, we now know that the student's original score was 2 standard deviations above the mean score. Note that the z distribution will only be a normal distribution if the original distribution of x was normal.

In summary, calculating the z-score gives us quick and easy access to understanding how **extreme** a certain result is. Looking at the original distribution ($\mu = 50$, $\sigma = 10$) and the standard normal distribution ($\mu = 0$, $\sigma = 1$) while highlighting $x = 70$ and $z = 2$ gives the following result:
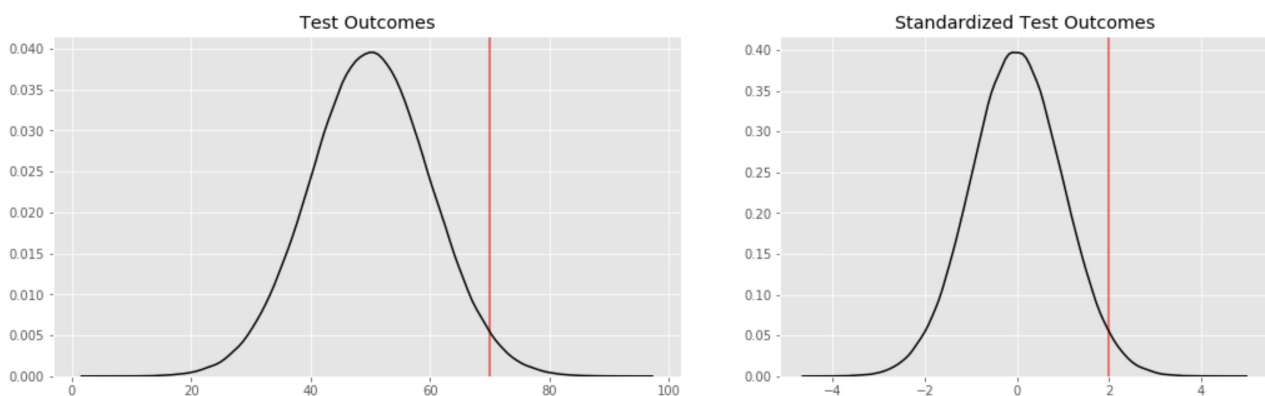
Visually, the idea is that the area under the curve, left and right from the vertical red line, are identical in the left plot and the right plot!

Thinking along these lines, you can also convert a z-score back to an original score x by using the same formula as:

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

For the above example, this would work out as:

$x = 50 + 2 * 10 = 70$

Data Standardization

Data standardization is a common data preprocessing skill, which is used to compare a number of observations belonging to different normal distributions which may have distinct means and standard deviations.

Standardization applies a z-score calculation, as shown above, on each element of the distribution. The output of this process is a **z-distribution** or a **standard normal distribution**.
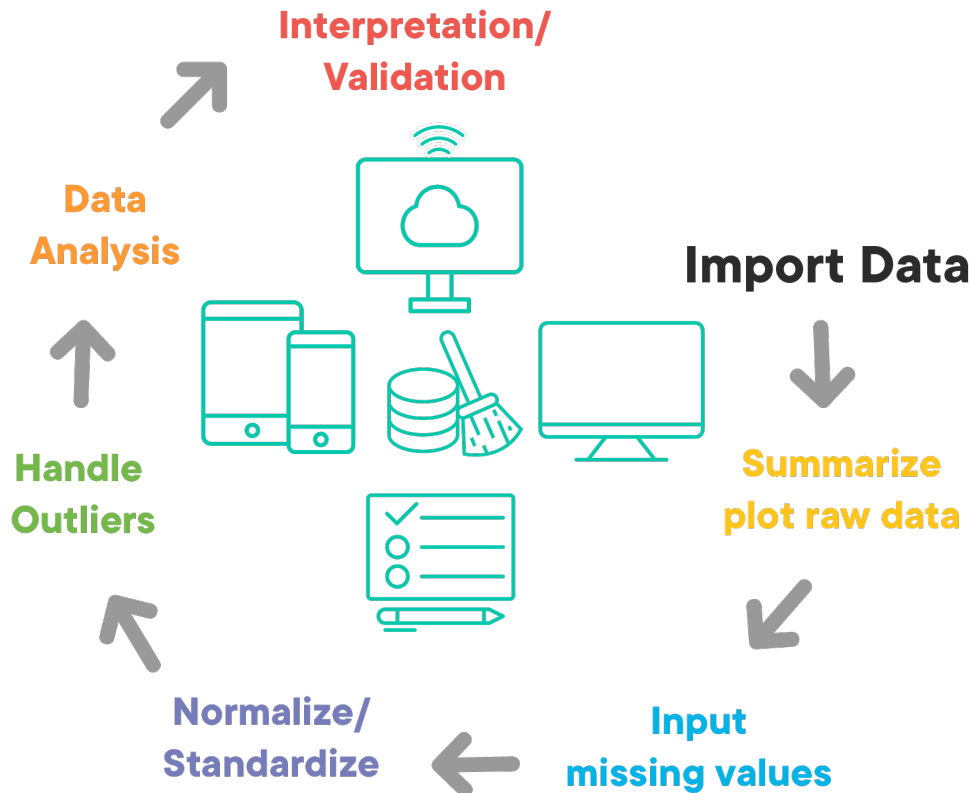
Let's look at a quick example. First, we'll randomly generate two normal distributions with different means and standard deviations. Let's generate 1000 observations for each. Next, we'll use `seaborn` to plot the results.

In [1]:

```python
import numpy as np
import seaborn as sns
mean1, sd1 = 5, 3 # dist 1
mean2, sd2 = 10, 2 # dist 2
d1 = np.random.normal(mean1, sd1, 1000)
d2 = np.random.normal(mean2, sd2, 1000)
sns.distplot(d1);
sns.distplot(d2);
```

You can see that these distributions differ from each other and are not directly comparable.

For a number of machine learning algorithms and data visualization techniques, it is important that the effect of the scale of the data is removed before you start thinking about building your model. Standardization allows for this by converting the distributions into a z-distribution, bringing them to a common scale (with $\mu = 0$, $\sigma = 1$). Let's standardize the above distributions and look at the effect.

In [2]:

```python
# Stardardizing and visualizing distributions

sns.distplot([(x - d1.mean())/d1.std() for x in d1]);
sns.distplot([(x - d2.mean())/d2.std() for x in d2]);
```
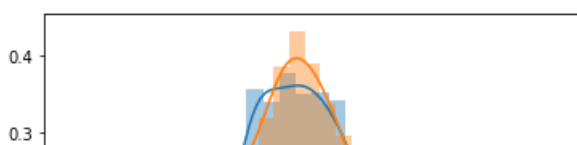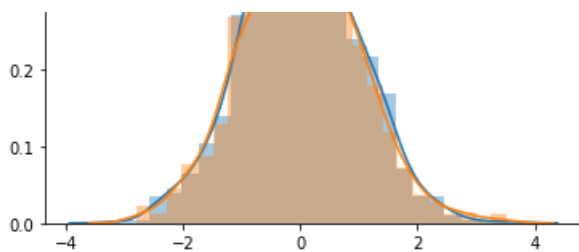
You see that both distributions are directly comparable on a common standard scale. As mentioned earlier, this trick will come in handy with analytics experiments while training machine learning algorithms.

Level up (Optional)

Convert standard distributions back to the original normal distributions using the formula given above. Visualize them to see your original distributions.

Summary

In this lesson, you learned about a special case of the normal distribution called the standard normal distribution. You also learned how to convert any normal distribution to a standard normal distribution using the z-score. You'll continue working on this in the following labs.