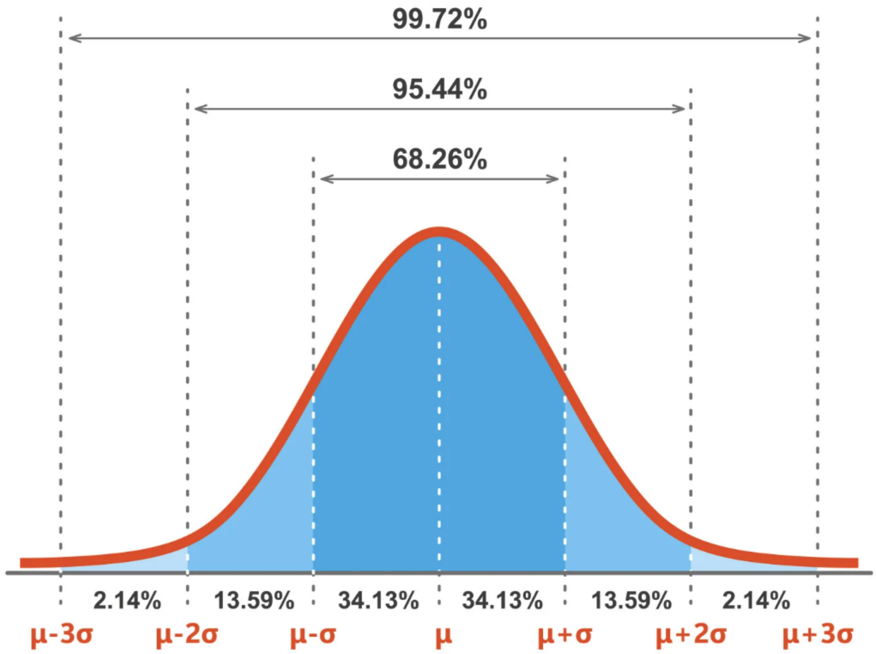
Gaussian Distribution (Normal Distribution)

The Gaussian distribution, also known as the **normal distribution**, is one of the most important and widely used probability distributions in statistics.

The Gaussian distribution states that when observing random phenomena, the observations are distributed in a particular way around a **mean**, forming a symmetric bell-shaped curve.

For this reason the normal distribution is also called **Gauss bell** characterized by its symmetrical bell shape around the mean, with most of the values concentrated near the mean itself and a gradual decrease in values as one moves away from it.



Normal distribution formula

The Gaussian distribution is defined by two parameters:

μ **(Mean)**:
represents the central value around which the data are distributed.

σ **(Standard deviation)**:
measure the dispersion or variability of data from the mean.
A smaller standard deviation indicates a more concentrated distribution around the mean, while a larger one indicates a more dispersed distribution.

The normal distribution is expressed by the following ***probability density function***, which is often referred to as the *Gaussian curve*:

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

where:

x: is the random variable

μ: is the mean of the distribution

σ^2: is the variance

A **probability density function** is a mathematical function that describes the probability distribution of a continuous random variable.

It describes the probability "density" at each point in the sample space.

Standardized normal distribution

The standardized normal distribution is a specific case of the normal distribution, which is obtained when the mean $\mu = 0$ and the standard deviation $\sigma = 1$

The standardized normal distribution is expressed by the following probability density function:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

where:

z: is the standardized random variable, measured in units of standard deviation from the mean.

This distribution is important because it allows you to standardize any normally distributed data with any mean and standard deviation to a standard form with mean zero and standard deviation one.

The standardized normal distribution is often used to calculate the probabilities associated with certain values in a normal distribution.
This is done by converting data from any normal distribution into corresponding values on the standardized normal distribution using the standardization formula:

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

where:

x: is the value of the random variable in the original normal distribution.

μ: is the mean of the original normal distribution.

σ: is the standard deviation of the original normal distribution.

z: is the corresponding value in the standardized normal distribution.

Normal Distribution Properties

The Gaussian distribution has several properties that make it an extremely important and versatile statistical model.

Symmetry:
The distribution is *symmetrical* about its mean. This means that half of the data is to the left of the mean and half to the right, with the same bell shape on both sides.

Data centrality:
Most of the data is concentrated around the mean of the distribution. The frequency of observations decreases as we move away from the mean, following a decreasing pattern conforming to a bell curve.

Probability of extreme data:
Extreme events are less likely than events close to the average.
The probability of occurrence of values very far from the mean decreases rapidly as we move away from it.

Mean, Median and Mode coincide:
In the normal distribution, the mean, median and mode (the most frequent value) coincide and are all in the center of the distribution, right at the mean.

Undefined extension:
The normal distribution curve extends infinitely in both directions, without ever touching the x-axis.

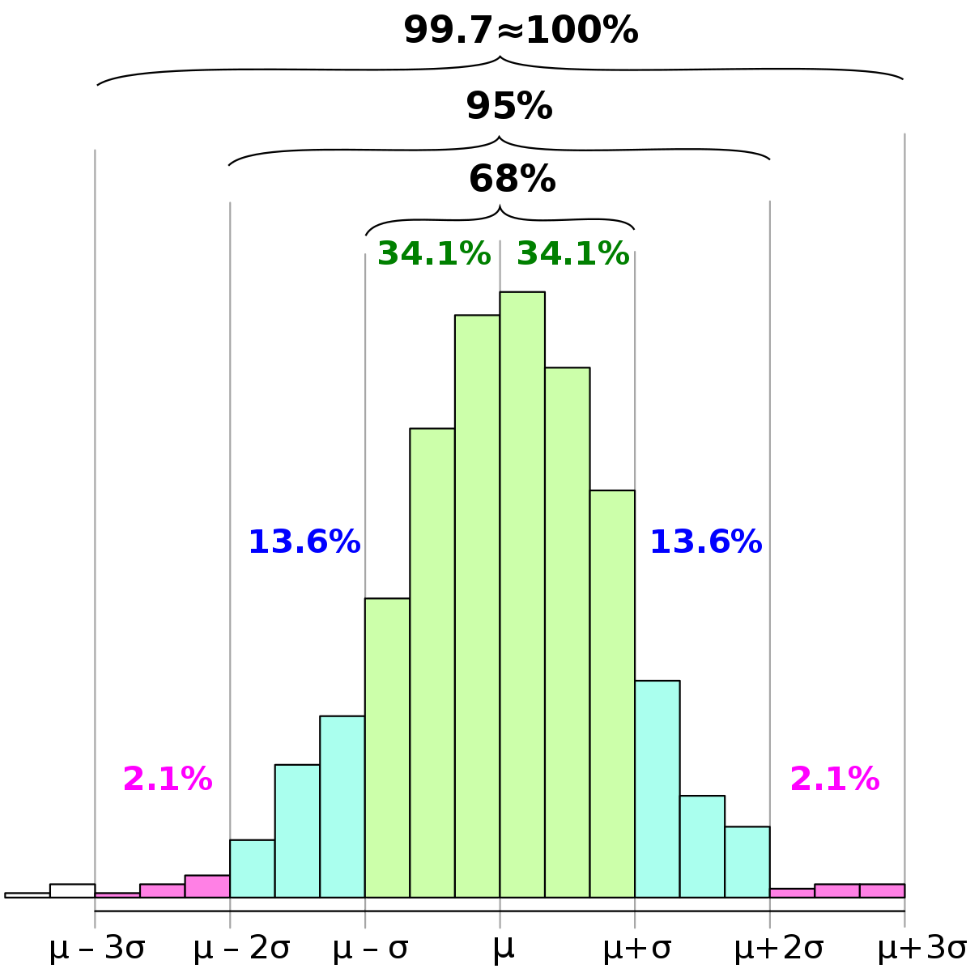
Sum and mean of normal random variables:
The sum of normal random variables is also a normal random variable, as is the average of a sample of normal random variables

Empirical rule (68-95-99.7 rule):

*The **68 - 95 - 99.7** rule, also known as the empirical rule, is an abbreviation used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations.*

more precisely:

- About 68% of the data is within one standard deviation of the mean
- About 95% of the data is within two standard deviations of the mean
- About 99.7% of the data is within three standard deviations of the mean



Relation with the Central Limit Theorem

The Central Limit Theorem (CLT) is one of the fundamental concepts of statistics and probability theory which concerns the analysis of the behavior of a sequence of random variables when the number of observations approaches infinity.

It states that, given the right conditions, the sum of a large number of independent random variables, each with any probability distribution, follows an approximately normal distribution, **regardless of the shape of the original distribution of the random variables.**

In other words, CLT states that even if individual random variables do not follow a normal distribution, the sum of these variables tends towards a normal distribution when the number of summed random variables becomes large.

This makes the Gaussian distribution useful in many contexts where you are working with the sum of many random variables.

The relationship between the Central Limit Theorem and the normal distribution can be represented by these two points:

1. Sum of independent random variables:
If we take a large number of independent random variables and add them, the result of this sum tends to be normally distributed.
These random variables can come from any distribution with a variety of shapes, but when added together in large numbers, their behavior always follows a normal distribution.

2. Application of normal distribution:
CLT is widely used in statistics because it allows us to make hypotheses about the distribution of a sample mean or the sum of random variables.

This allows us to use the properties of the normal distribution to make inferences about sampled data.

Let's now look at an example to make the link between the normal distribution and the central limit theorem clearer.

Suppose we have an experiment in which a random variable is represented by the result of rolling a six-sided fair die.

The distribution of this random variable is not normal, **but is uniform** (each face has the same probability of being obtained).

For example, if we roll the die and record the result (1 to 6), we will have a uniform distribution across six possible outcomes.

Now, consider rolling this die n times and adding the results of the rolls.

According to CLT, the more throws are made (increasing n), the greater the similarity of the distribution of the sum of the results to a normal distribution.

For example, if we roll the die 10 times and add the results, the distribution of the sum will be less like a normal distribution.

But if we increase the number of rolls to 1000, 10000 or more, the distribution of the sum of the results of the dice rolls will begin to look more and more like a normal distribution.

This example demonstrates how, although the initial data (dice roll results) are uniformly distributed and do not follow a normal distribution, CLT allows us to predict that the sum of a large number of these random variables (dice rolls) will converge to a normal distribution.

References:

Normal distribution - *Wikipedia*
https://en.wikipedia.org/wiki/Normal_distribution

The standard normal distribution - *Scribbr, Wikipedia*
https://www.scribbr.com/statistics/standard-normal-distribution/#:~:text=The%20standard%20normal%20distribution%2C%20also,the%20mean%20each%20value%20lies.
https://en.wikipedia.org/wiki/Normal_distribution

68-95-99.7 rule - *Wikipedia*
https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

Central Limit Theorem & its relationship with normal distribution - *Investopedia*
https://www.investopedia.com/terms/c/central_limit_theorem.asp#:~:text=Key%20Takeaways-,The%20central%20limit%20theorem%20(CLT)%20states%20that%20the%20distribution%20of,for%20the%20CLT%20to%20hold.