15.4.3 Dive Into Distributions

Now that the team feels comfortable with different data types, Jeremy wants to do a quick refresher on data distributions. In other words, Jeremy wants to understand the shape of his data before performing any analysis.

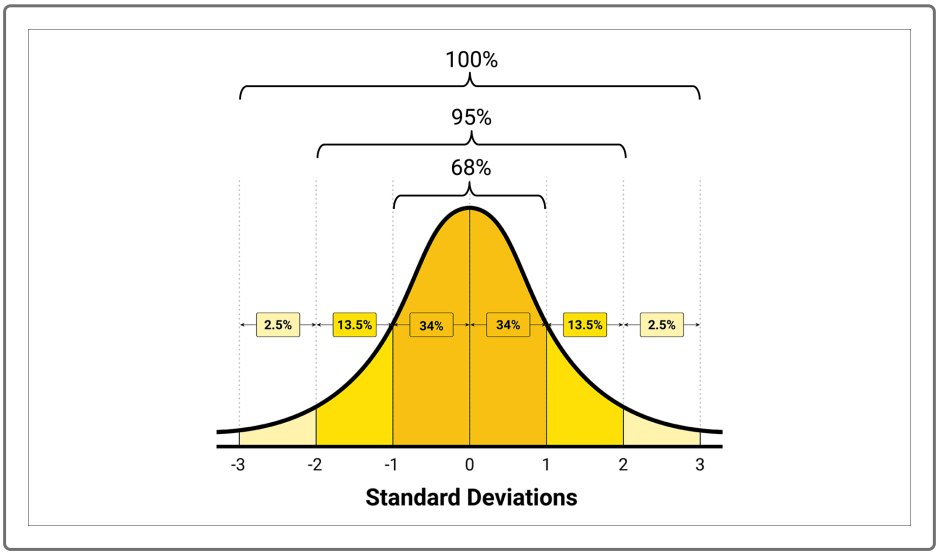
When it comes to data analysis, characterizing the distribution of numerical data is as important as characterizing the different data types. If we make incorrect assumptions about the distribution of our data, our statistical results could be meaningless. In general, most basic statistical tests assume that each numerical metric follows an approximate **normal distribution**.

In other words, we must confirm our data is normal before we can use a statistical test. If the data does not follow an approximate normal distribution, we would need to implement a more generalized (and oftentimes more complicated), non-normal statistical function.

Fortunately, there are qualitative and quantitative tests we can use to test our data for normality to avoid using these more generalized functions.

What Is Normal Distribution?

Normal distribution, or normality, is commonly referred to as "the bell curve," and describes a dataset where values farther from its mean occur less frequently than values closer to its mean.



When numerical data is considered to be normally distributed, the probability of any data point follows the **68-95-99.7** rule, stating that 68.27%, 95.45%, and 99.73% (effectively 100%) of the values lie within one, two, and three standard deviations of the mean, respectively.

In statistics, the **central limit theorem** is a key concept that states if you take sufficiently large samples of data from a dataset with mean μ (mu) and standard deviation σ (sigma), then the distribution will approximate normal distribution. Therefore, if we are using relatively large sample sizes, we should expect data to become more normally distributed.