Microsoft

Data Science Essentials

Simulations and Confidence Intervals

Simulations and confidence intervals are both concerned with estimating real-world data distributions from statistical suppositions or data samples.

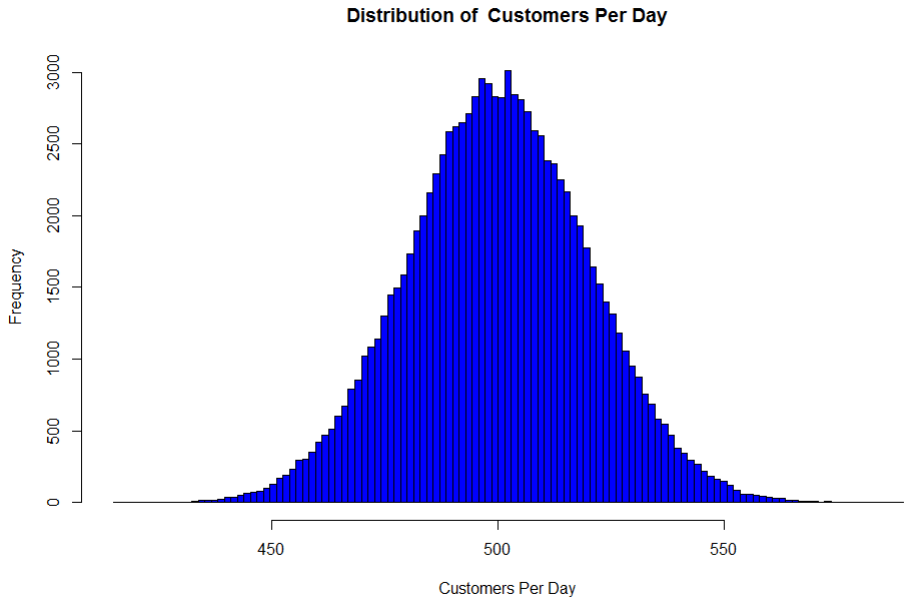
Simulations

Data scientists often need to use statistical methods to experiment with data and model real-world scenarios. When the data consists of a known number of independent random draws from a simple distribution, quantities of interest can often be estimated relatively simply. However, many real-world scenarios are more complex, and cannot be easily modeled as arising from a normal distribution or another simple distribution. In these cases, you can use a *simulation* to model the variables and gain an understanding of how the scenario is likely to work in reality.

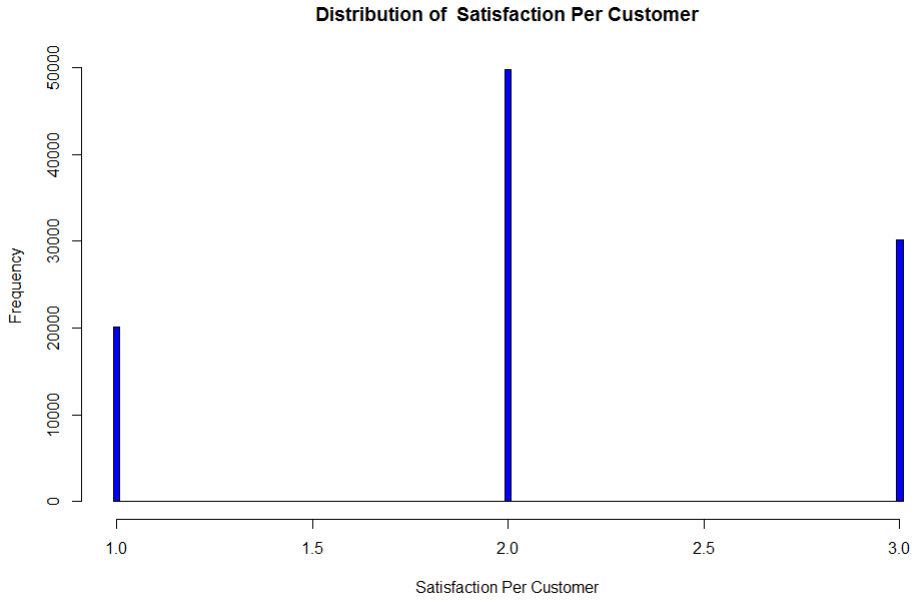
To run a simulation, you must identify the possible outcomes for each random variable in the scenario along with its probabilities, and the relationships between random variables. This defines the probability distribution you are working with. You then create a number of random draws and look at the outcomes; this is the simulation. For example, suppose you need to model customer satisfaction at a store where each customer can rate service as 1 for poor, 2 for acceptable, and 3 for excellent. The individual ratings are then totaled each day to give an overall satisfaction score. There are two random variables that need to be taken into consideration for the scenario: the number of customers and the ratings they give.

For this example, we'll assume that the number of customers can be represented by a normal distribution with a mean of 500 and a standard deviation of 20; and that 50% of the time these customers tend to give a rating of 2, 20% of the time they give a rating of 1, and 30% of the time they give a rating of 3.

Using these suppositions, you can run the simulation a large number of times, generating random values for the two variables based on their probability distributions, and use the results to model the likely distribution of total satisfaction scores. In this case, the distribution of customers for 100,000 runs (or *realizations*) of the simulation looks like this:

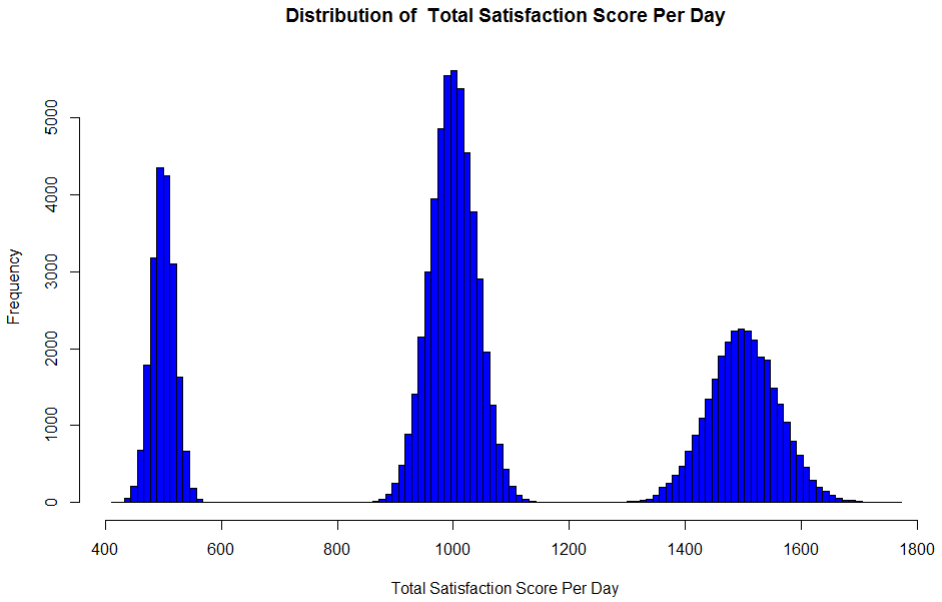


The mean number of customers per day is 500, and this was achieved around 3,000 days out of the total 100,000 simulated. The ratings given by those customers look like this:



Out of 100,000 realizations, around half of them (50,000) produce a rating of 2. There are 2,000 instances of a rating of 1, and 3,000 instances of a rating of 3. This corresponds with the probability we assumed for customer ratings.

When we combine the results of the simulations for both variables, we can see the likely distribution of total satisfactions scores below:



This distribution indicates that around 4,500 days out of 100,000, the total score will be around 500, around 5,000 days will achieve a total score of 1,000, and around 2,000 days will achieve a total score of 2,000. These peaks (*modes*) in the distribution are the result of combining the number of customers and customer scores generated in 100,000 realizations of the simulation based on the expected distribution of those individual variables.

Confidence Intervals

Without having access to the total population of the data, you need to be able to determine whether the sample mean ($\bar{x}$) is likely to approximate the population mean (μ). A *confidence interval* is a way to express the probability that a true population parameter falls within an interval of a statistic of the data. The confidence interval for the mean μ would be centered at $\bar{x}$, and the size of the confidence interval would depend on the sample standard deviation of points and the total number of sample points.