# Analyzing Randomness

Simulation output can often be challenging to analyze due to the inherent randomness and complexity of the underlying processes. For example, when considering consecutive customer waiting times in a queue, the data might not be [normally distributed](https://en.wikipedia.org/wiki/Normal_distribution), i[dentically distributed, or independent](https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables).

* Not normally distributed: The waiting times are usually skewed, with many short waiting times and a few very long ones.
* Not identically distributed: The patterns of waiting times change throughout the day due to variations in customer arrival rates and service times.
* Not independent: Waiting times are often correlated, as a long waiting time for one customer may lead to longer waiting times for subsequent customers in the queue.

These characteristics of simulation data make it difficult to analyze using traditional statistical methods. Thus, simulation analysts must consider alternative approaches.

There are two general cases to consider when analyzing simulation output:

In terminating simulations, the focus is on short-term behavior. The simulation stops after a specific event or time, and the interest is in understanding system performance within that limited timeframe. Examples of terminating simulations include:

* Average customer waiting time in a bank over the course of a day
* Average number of infected victims during a pandemic
* Evaluating the effectiveness of a marketing campaign over a set period

In steady-state simulations, the focus is on long-term behavior. The simulation runs indefinitely (or for a long enough time that the system under analysis reaches a [steady state](https://en.wikipedia.org/wiki/Steady_state)), and the interest is in understanding the system performance as it stabilizes over time. This type of simulation is often used to analyze processes that have no clear start or end, such as manufacturing or service systems. Examples of steady-state simulations include:

* A long-running assembly line in a factory
* The performance of a transportation network under various conditions
* Obtain an estimate of the unemployment rate 30 years from now

When analyzing randomness in simulation outputs, it is crucial to consider the type of simulation (terminating or steady-state) and the characteristics of the data. This understanding will help guide the choice of appropriate analysis techniques to draw meaningful conclusions from the simulation results.

# Terminating Simulations

Terminating simulations are those that focus on short-term behavior and end after a specific event or time. Analyzing the output of these simulations typically involves a method known as [Independent Replications](https://en.wikipedia.org/wiki/Replication_(statistics)). This method consists of the following steps:

* Make independent runs (replications) of the simulation:

Perform multiple independent runs of the simulation, each under identical conditions. These replications allow you to capture the variability in the system's performance and obtain a better estimate of the true performance measures.

* Sample means from each replication:

After each replication, calculate the [sample mean](https://en.wikipedia.org/wiki/Sample_mean_and_covariance) of the performance measure of interest, such as average customer waiting time or average number of infected victims during a pandemic. These sample means are summary statistics that capture the performance of each replication.

* Assume approximate identically and independently distributed (i.i.d). normal distribution of sample means:

The [Central Limit Theorem](https://en.wikipedia.org/wiki/Central_limit_theorem) states that, given a sufficiently large number of replications, the distribution of the sample means will approach a normal distribution, regardless of the underlying distribution of the original data. The sample means are also assumed to be identically and independently distributed, meaning that they share the same distribution and are not correlated.

* Use classical statistical techniques on the i.i.d. sample means:

With the assumption of i.i.d. Normal distribution for the sample means, classical statistical techniques can be applied to analyze the simulation results. These techniques include calculating [confidence intervals](https://en.wikipedia.org/wiki/Confidence_interval), [hypothesis testing](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing), and other methods to draw [inferences](https://en.wikipedia.org/wiki/Statistical_inference) about the system's performance.

It is important to note that the classical statistical techniques are applied to the i.i.d. sample means, not the original observations from the simulation. This is because the original data may not meet the assumptions of normality, identical distribution, or independence, which are necessary for using traditional statistical methods.

# Steady-State Simulations

Steady-state simulations focus on long-term behavior, with the simulation running indefinitely (or for a long enough time that the system under analysis reaches a steady state) to study the system's performance as it stabilizes over time. Analyzing steady-state simulations involves addressing several challenges, such as [initialization bias](https://en.wikipedia.org/wiki/Transient_state), and applying appropriate methods to handle the steady-state data.

An excerpt from Bratley, P., Fox, B., & Schrage, L. (1987). A guide to simulation (2nd ed.). Springer Science+Business addresses the idea that you can’t simulate something forever, so the only hope is that “the system achieves some approximation to steady-state or long-run behavior during a moderate length (simulated) time and that extrapolation from this performance to predicted long-run performance is reliable.”

Initialization bias occurs when the simulation starts from an arbitrary initial state that does not accurately represent the steady-state conditions. To mitigate this bias, it is common to "warm up" the simulation by running it for a certain period before collecting data. Failure to address this bias can lead to inaccurate statistical analysis.

Various methods are available for analyzing steady-state data, including [Batch Means](https://rossetti.github.io/RossettiArenaBook/ch5-BatchMeansMethod.html), [Overlapping Batch Means](https://dl.acm.org/doi/abs/10.5555/2431518.2431562), [Standardized Time Series](https://en.wikipedia.org/wiki/Time_series), and Regeneration (identifying points where the system returns to a specific state and the process starts anew). Each method has its advantages and disadvantages, depending on the specific characteristics of the data and the objectives of the analysis.

The Method of Batch Means is a widely used approach for analyzing steady-state simulation data. The procedure involves the following steps:

* Make one long run: Instead of conducting multiple short replications as in terminating simulations, a single long run of the simulation is performed.
* Warm-up simulation before collecting data: To address initialization bias, the simulation is run for a "warm-up" period before starting to collect data for analysis.
* Chop remaining observations into contiguous batches: After the warm-up period, the remaining observations are divided into contiguous, non-overlapping batches of equal size.
* Sample means from each batch: Calculate the sample mean of the performance measure of interest for each batch. These batch means are summary statistics that capture the performance of the system during each batch.
* Assume approximate i.i.d. normal distribution of batch means: As with the independent replications method, the Central Limit Theorem suggests that, given a sufficiently large number of batches, the distribution of the batch means will approach a normal distribution, and the means are assumed to be identically and independently distributed.
* Use classical statistical techniques on the i.i.d. batch means: With the assumption of i.i.d. Normal distribution for the batch means, classical statistical techniques can be applied to analyze the simulation results.

The Method of Batch Means is particularly useful for analyzing steady-state simulations, as it allows analysts to handle the challenges posed by initialization bias and leverage classical statistical techniques to draw inferences from the data.