Unit-7: Analysis of Simulation Output

Why analysis of simulation output?

- Many simulation includes some sort of randomness, which can arise in a variety of ways e.g. in a simulation of manufacturing system, the processing times required at a station may have random variations or the arrival times of new jobs may not be known in advance.
- In a bank, customers arrive at random times & amount of time spent at a teller is not known beforehand. Because of the randomness of the components driving a simulation, the o/p from simulation is also random. So, statistical techniques must be used to analyze the results.
- To test different ideas, to learn about the system behavior in new situation, to learn about simulation model and the corresponding simulation system.

Nature of the problem

Once a stochastic variable has been introduced into simulation model, almost all the system variables describing the system behavior also become stochastic. Hence it needs some statistical method to analyze the simulation output. A large body of statistical methods has been developed over the years to analyze results in science, engineering and other fields.

It seem natural to attempt applying these methods to analyze the simulation output but most of them pre-suppose that the results are mutually independent IID (Independent and Identically Distributed) and the simulation process almost never produce raw output that is IID.

Note: *A collection of random variables is independent and identically distributed (IID) if each random variable has the same probability distribution as the others and all are mutually independent.*

For example: Customer waiting times from queuing system are not IID. Thus it is difficult to apply classical statistical techniques to analysis of simulation model.

While characterizing the property of a variable, in every time unit the value changed is compared to predefined intervals known as confidence interval to find similarity between estimated and simulated value. Simulation results are not mutually independent.

Some definitions:

a) Independently and Identically Distributed (IID) Random Variables: Usually a random variable is drawn from an infinite population that has probability distribution with finite mean μ and finite variance σ^2. This mean that the population distribution is not affected by the number of sample already made or does it change with time. Further the value of sample is not affected in anyway by value of another sample. Random variables that meet all these conditions are said to be independently and identically distributed.

b) Central Limit Theorem: The central limit theorem, which is a statistical theory, states that when a large sample size has a finite variance σ^2, the samples will be normally distributed, and the mean μ of samples will be approximately equal to the mean of the whole population.

As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean. If the sample size is small, the actual distribution of the data may or may not be normal, but as the sample size gets bigger, it can be approximated by a normal distribution. This statistical theory is useful in simplifying analysis while dealing with stock indexes and much more.

1. Type of simulation on the basis of output

a) Termination simulation or finite simulation: The termination of a finite simulation takes place at a specified time or is caused by some specific events.

For example: Banks opens at 8:30am with no customers present and 8 of 11 teller working and closed at 4:30 pm. The terminating simulation runs for some specified duration of time TE, where E is a specified event that stops the simulation. It starts at time 0 under well specified initial condition and ends at the stooping time TE. For the banking system, the simulation analyst chooses to consider a terminating system because object of interest is one day's operations on the bank.

b) Non-Terminating Simulation (Steady state simulation): The main propose of steady state simulation is the study of long run behavior of system. Performance measure is called a steady state parameter if it is a characteristic of the equilibrium distribution of an output stochastic process.

For example: Continuously operating communication system such as telephone system, hospital emergency rooms, network routers, internet etc. where the objective of computation of mean delay of packet in the long run.

Here the non-terminating simulation has:

1. Runs continuously or at least over a very long period of time.
2. Initial conditions defined by analyst.
3. Runs for some analyst specified period of time T_E
4. Study the steady state (long run) properties of the system, properties that are not influenced by the initial condition of model.

*(**Note:** whether a simulation is considered to be terminating or non-terminating depends on both the objective of study and nature of the system)*

Estimation Methods

Estimation is a process in which we obtain the values of unknown population parameters with the help of sample data. In other words, it is a data analysis framework that combines effect sizes and confidence intervals to plan an experiment, analyze data, and interpret the results.

A random variable is drawn from an infinite population that has a stationary probability distribution with a finite mean μ and finite variance σ^2. Random numbers that meet all these conditions are said to be IID (Independently and Identically Distributed) variable for which the ***Central Limit Theorem*** can be applied.

Central Limit Theorem:

The Central Limit Theorem states that the sum of n IID variable drawn from a population that has mean μ and a variance of σ^2 is approximately distributed as a normal variable with mean $n\mu$ and variance $n\sigma^2$.

Any normal distribution can be transform into a standard distribution that has a mean of 0 and variance of 1.

Let X$_i$ (i=1,2,3,………n) be n IID random variables. Using central limit theorem, we have normal variate:

$$z = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}}$$

In terms of sample mean ($\bar{x}$), Dividing by n in both denominator and numerator, we get:

$$\bar{x} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The variable $\bar{x}$ is the sample mean and it can be shown to be a consistent estimator for mean of the population form which sample is drawn.

<u>Types of Estimation Methods:</u>

There are two types of estimation methods:

- Point estimation
- Confidence interval

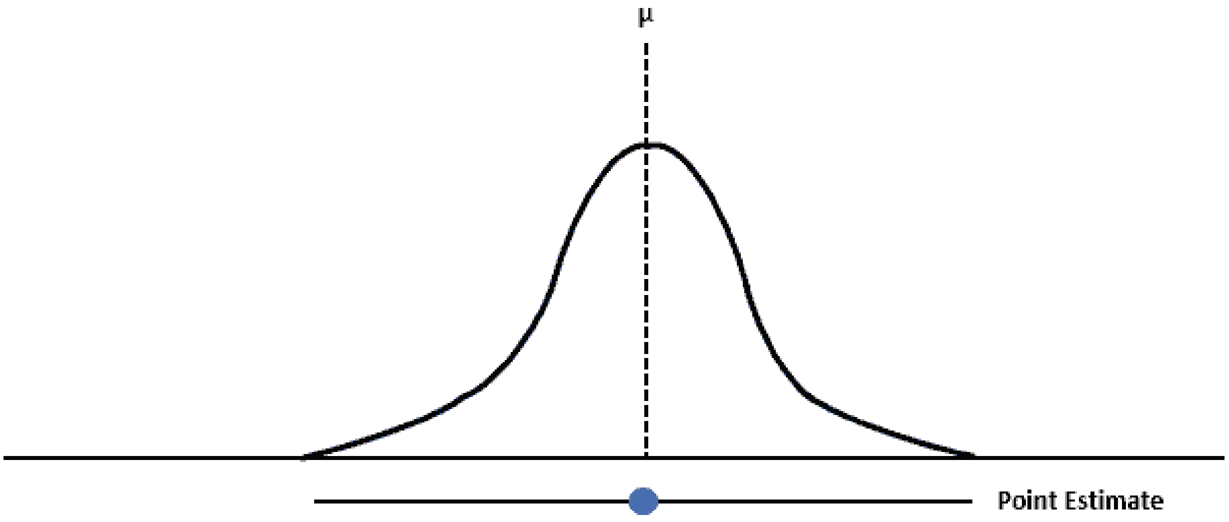
<u>Point estimation (For discrete data):</u>

A point estimate is a single value estimate of a parameter. For instance, a sample mean M is a point estimate of a population mean μ.

Let y1, y2 ...yn be discrete time data with ordinary mean θ, then point estimator

$$\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

This is unbiased if its expected value is θ i.e. **E($\hat{\theta}$) = θ** and is biased if **E($\hat{\theta}$) ≠ θ** and the difference **E($\hat{\theta}$) - θ** is called bias of $\hat{\theta}$.



For example: Assume you wanted to estimate the mean time of 12-year-olds athletes to run 100 yards. The mean running time of a random sample of 12-year-olds athlete would be an estimate of the mean running time for all 12-year-olds athletes. Thus, the sample mean M, would be a point estimate of the Population mean μ.

Example: Following are the random sample of height of people of the town. If the ordinary population mean is 6.1ft, find the bias of the point estimator.

5.5	6.1	5.7	6.6	5.2	6.0	5.6	6.3	5.9	5.8

Solution:

Here, given that

Ordinary population mean (θ) = 6.1

Number of sample (n) = 10

Then, the point estimator

$$\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

= (5.5 + 6.1 + 5.7 + 6.6 + 5.2 + 6.0 + 5.6 + 6.3 + 5.9 + 5.8)/10

$\hat{\theta}$ = 5.87

Now,

Bios of estimator = **$\hat{\theta}$ – θ**

= 5.87-6.1

= - 0.23

Hence the bios of estimator is – 0.23

For the continuous data {y(t): < t < T} with mean Φ, then the point estimator is given by

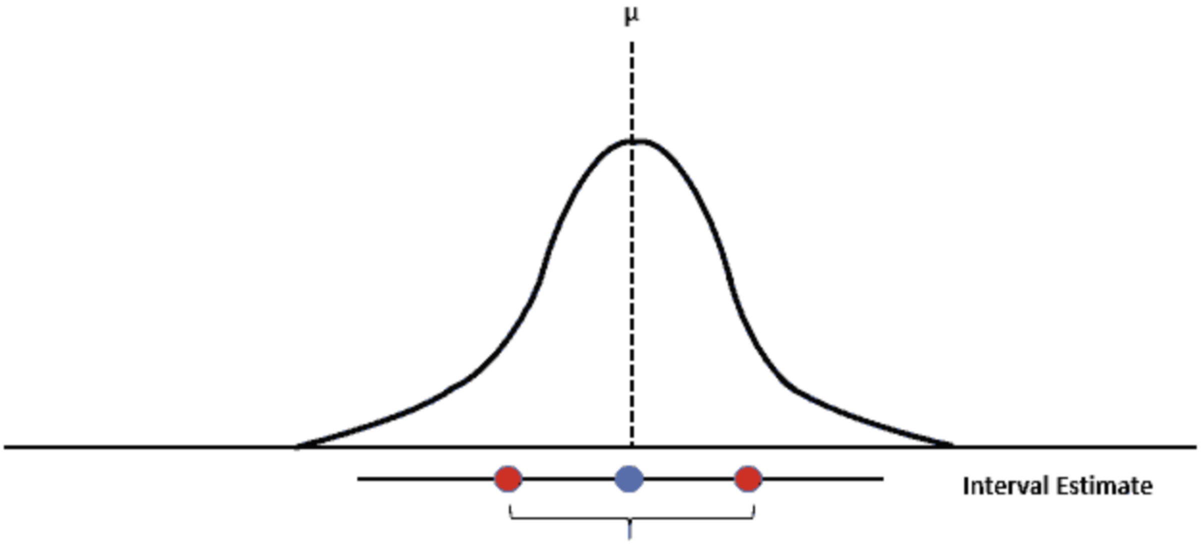
$$\hat{\phi} = \frac{1}{T_E}\int_{0}^{T_E} y(t)dt$$

Where, T_E is the specific duration of time where simulation runs.

<u>Confidence Interval Estimation (Interval Estimation):</u>

A confidence interval estimate gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

Confidence interval is a measure used to analyze the correctness of the point estimator. In the above example we got point estimation of height of the people of a city but we don't know whether to accept or reject it.



Confidence intervals are based on the premise that the data being produced by the simulation is represented well by a probability model. Suppose the model is the normally distributed with Mean ($\bar{x}$), Variance (σ^2) and we have a sample of **n** size then the confidence interval is given by:

$$\bar{x} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

In practice, the population variance is usually not known; in this case variance is replaced by the estimate calculated by the formula:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

This has a student t-distribution, with n–1 degree of freedom. In terms of estimated variance s^2, the confidence interval for $\bar{x}$ is defined by:

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}$$

Here, the quantity $t_{n-1,\alpha/2}$ is found on the student t-distribution table.

Example:

The daily production time of a product in a factory for 120 days is 5.8 hours and sample standard deviation (s) is 1.6. Calculate confidence interval for 95% confidence level.

Solution:

Here, given:

Total sample size (n) = 120

Production mean ($\bar{x}$) = 5.8

Standard deviation (s) = 1.6

95% confidence interval indicates that (α) = 100-95 = 5% = 0.05

We know that,

Confidence Interval = $\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}$

= $5.8 \pm \frac{1.6}{\sqrt{120}} t_{120-1,0.05/2}$

= $5.8 \pm \frac{1.6}{10.95} t_{119,0.025}$

= $5.8 \pm (0.15 * 1.960)$ $\quad (t_{119,0.025} = 1.960)$

Confidence Interval = 5.8 ± 0.294

Hence, the estimates between 5.8± 0.294 can be accepted for 95% confidence interval.

Assignment:

1. Following are the random sample of marks of student of the prime college. If the population mean is 51.1, find the bias of the point estimator.

51	52	56	50	48	49	59.8	45.9	58	47

2. The daily present time in a simulation class for 120 days is 50 minutes and sample standard deviation (s) is 1.6. Calculate confidence interval for 95% confidence level. $(t_{119,0.025} = 1.960)$
3. The daily present time in a futsal for 120 days is 60 minutes and sample standard deviation (s) is 1.5. Calculate confidence interval for 95% confidence level. $(t_{119,0.025} = 1.960)$

Simulation Run Statistics:

In the estimation method, it is assumed that the observations are mutually independent and the distribution from which they are draws is stationary. Unfortunately many statistics of interest in simulation do not meet these conditions.

For example: Consider a single server system in which the arrival occurs with poison distribution and service time has an exponential and queue discipline is FIFO.

Where, Poisson distribution means a discrete frequency distribution which gives the probability of a number of independent events occurring in a fixed time.

Suppose the study objective is to measure the mean waiting time. In simulation run, the simplest approach to estimate the mean waiting time by accumulating the waiting time of **n** successive entities and dividing by number of entities **n**. This is the sample mean, denoted by $\bar{x}$(n). Waiting time measured in this way is not independent.

If $x_i(i = 1,2, \ldots \ldots , n)$ are the individual waiting time then,

$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Whenever a waiting line forms, the waiting time of each entity on the line clearly depends upon the waiting time of its predecessors. Such data are called auto-correlated.

Another problem that must be faced is that distribution is not stationary. The early arrivals get the service quickly, so a sample mean that include early arrivals is biased.

The following figure is based on theoretical results which shows how the expected value of sample mean depends upon the sample length for $M/M/1$ system, starting from an initial empty state with a server utilization of 0.9.

Here $\boldsymbol{M/M/1}$ indicates,

1st M = Inter-arrival time is distributed exponentially

2nd M = service time is distributed exponentially

1 = one server

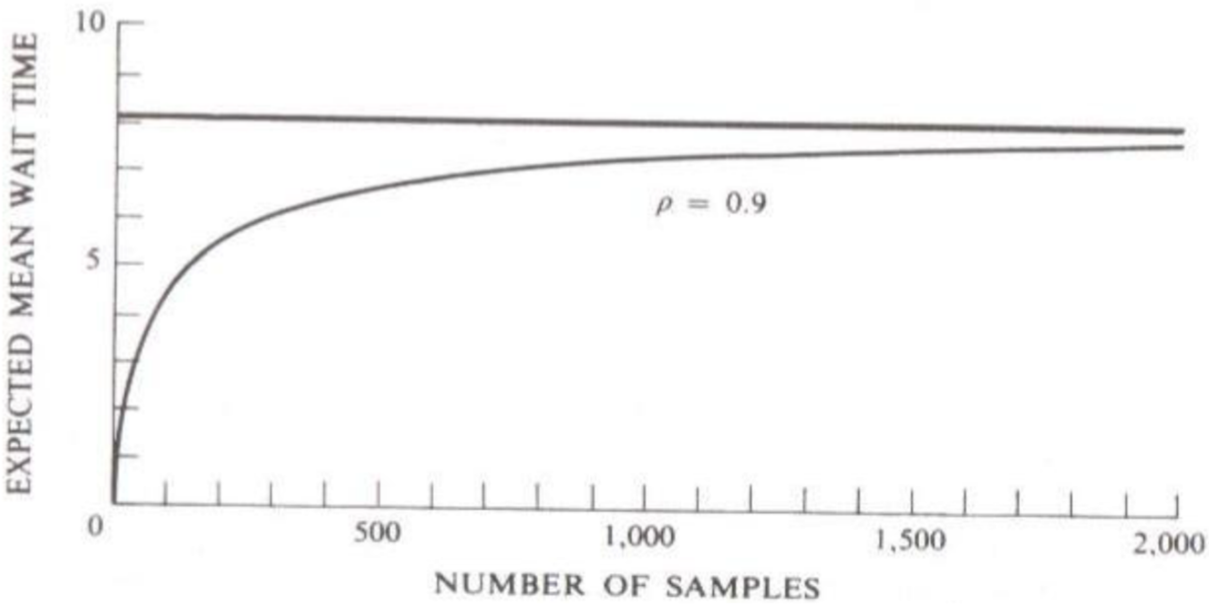


Fig: Mean wait time in M/M/1 system for different sample size

A simulation run is started with the system in some initial state, frequently the idle state in which no service is being given and none entities are waiting. The early arrivals that have a more than normal probability of obtaining service quickly so, a sample mean that includes the early arrivals will be biased.

For a given sample size starting from a given initial condition the sample mean distribution is stationery, but if the distribution would be compared for different sample sizes the distribution will be slightly different.

Replications (Repetition) of Runs:

One way of obtaining independent result is to repeat simulation. Repeating the experiment with different random numbers for the sample size **n** gives a set of independent determination of sample mean **($\bar{x}$).**

Suppose the experiment is repeated **p** times with independent random values of **n** sample sizes.

Let **x_{ij}** be the **ith** observation in **jth** run and let the sample mean and the variance for the **jth** run is denoted by **$\bar{x}_j$(n)** and **s_j^2(n)** respectively. Then for **jth** run, the estimates are:

$$\bar{x}_j(n) = \frac{1}{n}\sum_{i}^{n} x_{ij}$$

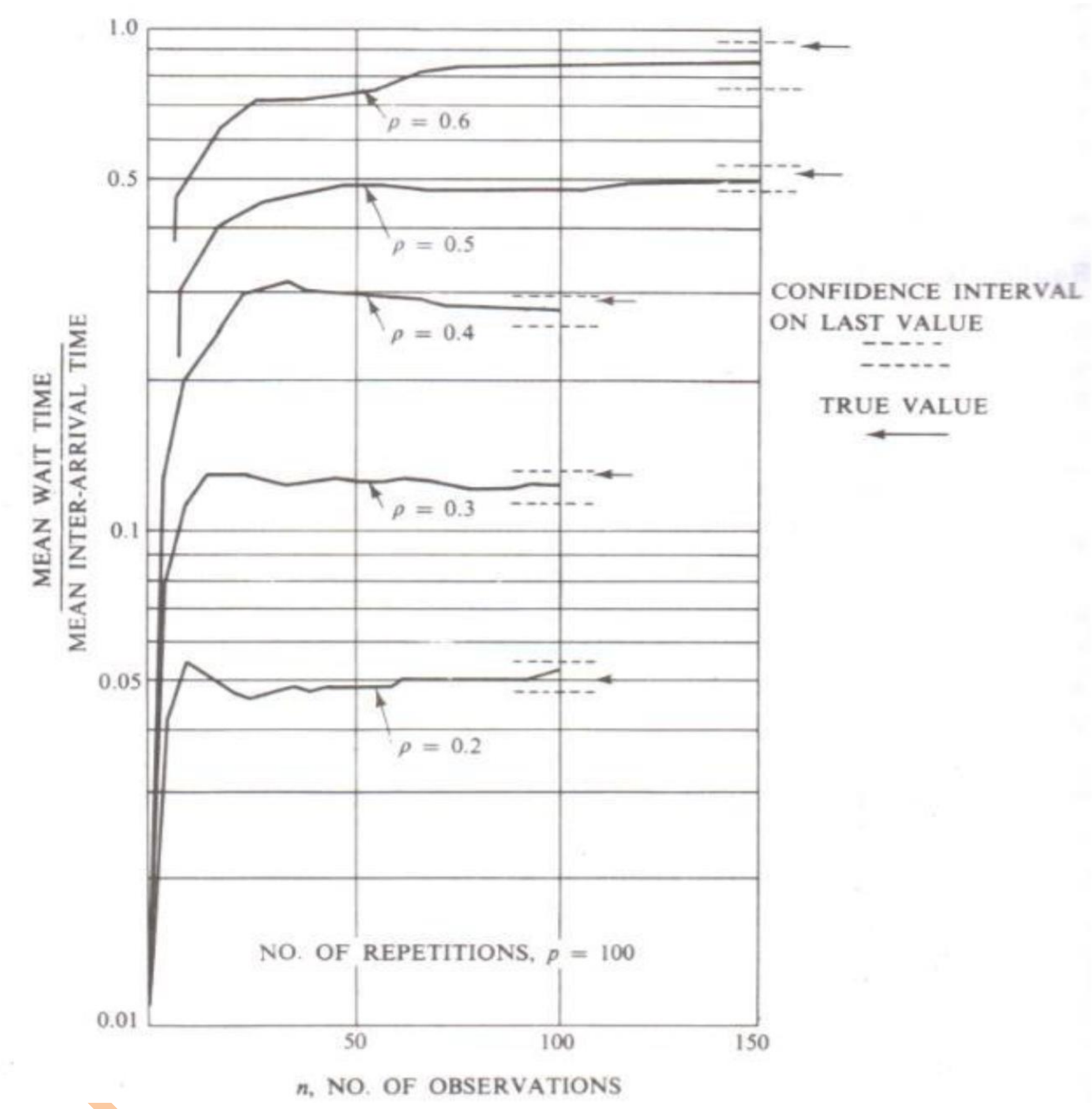
$$s_j^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_{ij} - \bar{x}_j(n)\right]^2$$

Combining the result of **p** independent measurement gives the following estimate for the mean **($\bar{x}$)** and variance **s^2** of the populations as:

$$\bar{x} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_j$$

$$s^2 = \frac{1}{p}\sum_{j=1}^{p} s_j^2$$

The following figure shows the result of applying the procedure to experiment results for the $M/M/1$ system.



This variance can further be used in established confidence interval for $(p-1)$ degree of freedom.

The length of run of replication is so selected that all compiled combined comes to sample size N i.e. $p.n = N$, Where,

p = No. of sample

n = No. of observation per sample

By measuring the number of replications and shortening their length of run, confidence interval can be narrower but due to shortening of length of replication the effect of starting condition will increase.

The result obtained won't be accurate especially when the initialization of the runs is not proper.

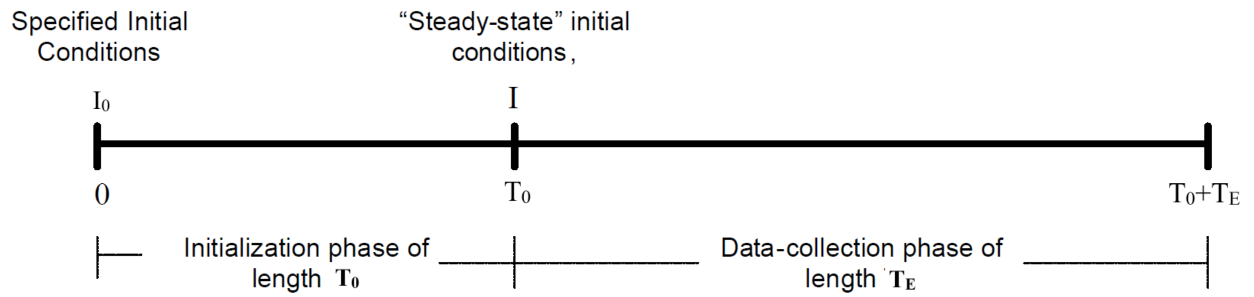
There is not established procedure of dividing the sample size N into replications. Dividing 'N' observations into number of ranges is quite difficult. However it is suggested that the number of replications should not be very large and that the sample means should approximate in normal distribution.

<u>Initialization of Bias:</u>

Initial system state representation is more representative of long-run conditions. This step is also known as intelligent-initialization

Two ways to specify initial conditions:

1. If the system exists, collect data on it and use these data to specify more typical initial conditions.
    - It requires a large data collection effort
    - If system being modeled does not exist, this method is impossible to implement
    - Collecting system data from a second model which is a simplified
    - model is suggested for complex systems
2. Reduce the impact of initial conditions, by dividing each simulation run into two phases;
    a. Initialization phase from time 0 to T_0
    b. Data collection phase from time T_0 to stopping time T_0+T_E


Specified Initial Conditions
"Steady-state" initial conditions,
I_0
I
0
T_0
T_0+T_E
Initialization phase of length T_0
Data-collection phase of length T_E

In such case, simulation begins at time 0 under specified initial conditions I_0, and runs for specified period of time T_0.

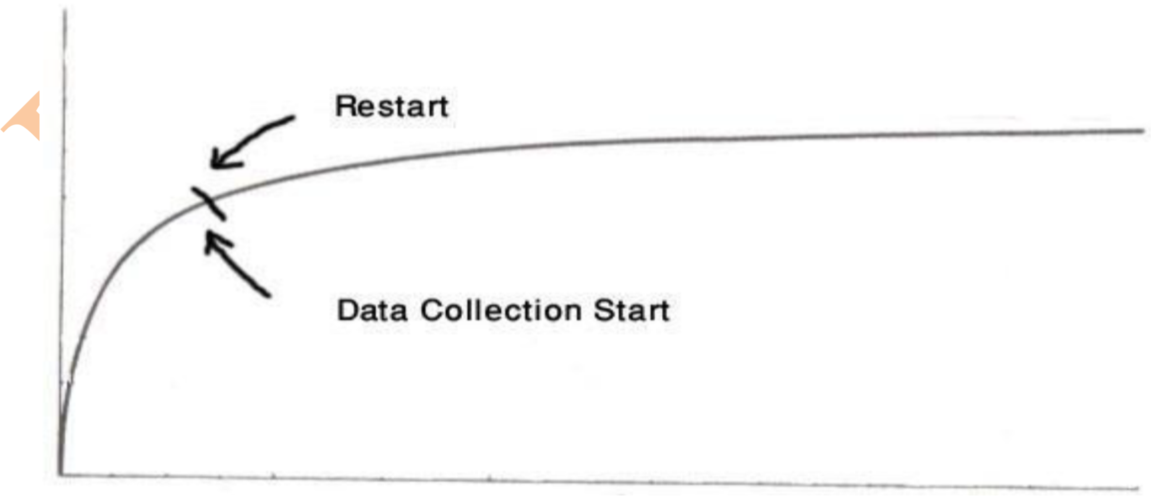
- Data collection on the response variables of interest starts from T_0 and until time T_0+T_E.
- The length T_E of the data collection phase should be long enough to guarantee sufficiently precise estimates of steady-state behavior

Elimination of initial bias:

The ideal situation for initial bias is first know the steady state distribution for the system and select the initial condition from the steady state distribution.

The more common approach to know the initial bias is to eliminate on initial selection of runs. The run is started from an idle state and stop after the certain period of time. The run is then restarted with statistics being gathered form the point of restart.

The initial bias as shown in figure below needs to be removed.



Two general approaches can be taken into remove the bias.

1. The system can be started in a more representative than an empty state.
2. The first part of the simulation can be removed.

It is usual to program the simulation so that statistics are gathered from the beginning and simply wipe out the statistics gathered up to the point of restart. No sample rules can be given beside how long an interval should be eliminated.

The disadvantage of eliminating the first part of the simulation run is that the estimate of the variance needed to estimate a confidence limit, mostly based on less information.

End of Unit-7