# Pramod Parajuli
# Simulation and Modeling, CS-331

## Chapter 10
### Analysis of Simulation Output

## Nature of the Problem

- Making a variable stochastic changes the properties of remaining variables to stochastic since the endogenous events make one variable depend upon another
- 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

## Central Limit Theorem

- The average of a sample of observations drawn from some population with any shape-distribution is approximately distributed as a normal distribution if certain conditions are met
- Most dominant conditions are; distribution of the parent population and the actual sampling procedure
- Simple application is;

If random sample size is ‘n’ ($n > 30$) from an infinite population

Finite standard deviation

Then, the standard sample mean converges to a standard normal distribution

(Look in the handout given for the details)

## Estimation Methods

- IID – independently and identically distributed, random numbers are mutually independent
- Central limit theorem

  It states that the sum of *n* IID variables, drawn from a population that has a mean of $m$ and a variance of $s^2$ , is approximately distributed as a normal variable with a mean of $nm$ and variance of $ns^2$ .

$$X_{norm} \equiv \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} m_i}{\sqrt{\sum_{i=1}^{N} s_i^2}}$$

where, $X_1, X_2, X_3, ..., X_N$ : set of independent random variates

each Xi has arbitrary probability distribution of $P(x_1, x_2, x_3, .., x_N)$ with mean $m_i$, *and variance* $s_i^2$

## Estimation Methods

- Any normal distribution can be transformed into a standard normal distribution (mean = 0, variance 1)
- Therefore, the normal variate will be

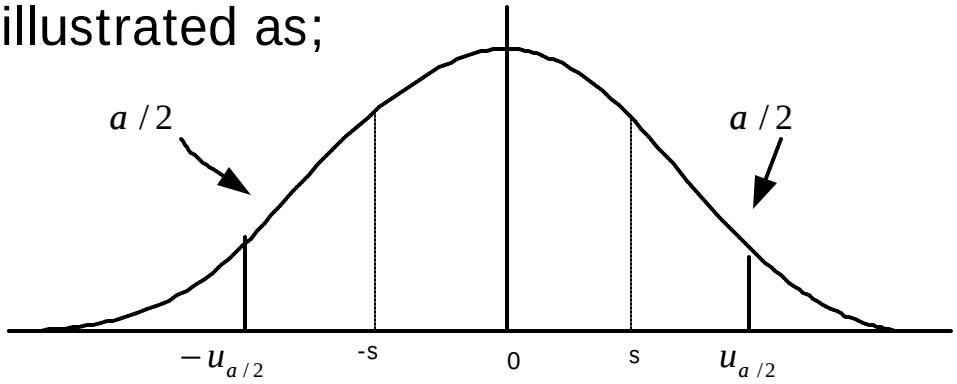
$$z = \frac{\sum_{i=1}^{n} x_i - nm}{\sqrt{n}.s}$$

- If, sample mean $\bar{x} = \frac{1}{n}.\sum_{i=1}^{n} x_i$

Then, $$z = \frac{\bar{x} - m}{s / \sqrt{n}}$$

## Estimation Methods

- The probability distribution of standard normal variate can be illustrated as;




- The integral from -8 to a value u is the probability that z is less than or equal to u
- This integral is generally denoted by F (u)
- Let for a chosen value of u is satisfying

$$\Phi(u) = 1 - \frac{a}{2}$$

$a$ =some constant less than 1

## Estimation Methods

- Let the value of u be $u_{a/2}$ i.e. the probability that z is greater than $u_{a/2}$ is $a/2$.
- Same holds for $-u_{a/2}$
- The probability that z lies between $u_{a/2}$ and $-u_{a/2}$ is

$$\Pr ob\{-u_{a/2} \le z \le u_{a/2}\} = 1-a$$

- In terms of sample mean, the probability can be written as;

$$\Pr ob\left\{\bar{x} - \frac{s}{\sqrt{n}} u_{a/2} \le m \le \bar{x} + \frac{s}{\sqrt{n}} u_{a/2}\right\} = 1-a$$

- The constant $1-a$ is confidence level and

$\bar{x} + \frac{s}{\sqrt{n}} . u_{a/2}$ is confidence interval

## Estimation Methods

- The size of he confidence interval depends upon the confidence level
- If confidence level is 90%, then the value of $u_{a/2}$ is 1.65
- That is, μ will be covered by the confidence interval $\bar{x} \pm \frac{1.65s}{\sqrt{n}}$ with probability 0.9. i.e. greater the value of 'n' (more no. of simulation runs), the confidence interval can be expected to cover the value μ on 90% of the repetitions
- In practice population variance $s^2$ is not known, in such case, it is replaced by an estimate calculated using;

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

## Estimation Methods

- $s^2$ has Student-t distribution with n-1 degrees of freedom i.e. n-1 number of parameters may be independently varied
- Useful if samples drawn are normally distributed

## Simulation Run Statistics

- For CLT, two more considerations exists;
  1. The observations are mutually independent
  2. The distribution from which they are drawn is stationary
- But in most of the cases, both of these considerations do not hold
- E.g. (mutually independent)

  Let's consider a system with 'n' successive entities. Let $x_i$ represents wait time for $i^{th}$ entity. To represent mean wait time;

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

  As wait time of $i^{th}$ entity affect wait time of remaining entities, they are not independent. Therefore, they are auto-correlated

## Simulation Run Statistics

- In most of simulation run, the sample mean of auto-correlated data is approximated to a normal distribution as the sample size increases
- In such case, the equation in slide-10 holds
- However, variance of auto-correlated data will not exhibit same property as variance of independent data
- Therefore, some adjustment is made for auto-correlated data. Failure to adjust will result into under-estimation or over-estimation

- E.g. (stationary)
  - Early arrivals get service quickly, later arrivals have to wait
  - Large number of simulation runs reduce the bias

## Simulation Run Statistics

- If distribution for particular sample size is taken, it might differ from another sample size
- This leads to dynamic distribution
- E.g. mean wait time

- **fig( 14.2) Gordon**

- Even after 2,000 samples, the mean is not in steady state. For low server utilization, steady state is reached very fast but our concern is for higher server utilization

## Replication of Runs

- Repeated no. of simulations generate independent results
- For each simulation run, a different random numbers are used for same sample size ‘n’ and the simulation gives a set of independent determinations of sample mean

$$\overline{x}(n)$$

- Even though the sample mean depends on degree of auto-correlation, the independent determinations can be used to estimate the variance of the distribution
- Let the simulation is run for ‘p’ times with independent random number series
- Let $x_{ij}$ be the $i^{th}$ observation in the $j^{th}$ run and let the sample mean and variance for the $j^{th}$ run be;

$$\overline{x}_j(n) \qquad\qquad s_j^2(n)$$

## Replication of Runs

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

- Since there are 'p' number of runs,

$$x = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_j(n) \qquad s^2 = \frac{1}{p}\sum_{j=1}^{p} s_j^2(n)$$

- Look into figure 14-3 in Gordon, p = 100

  For sample size, n = 5, and for ? = 0.2, 0.3, 0.4

  For n = 10, ? = 0.5, 0.6

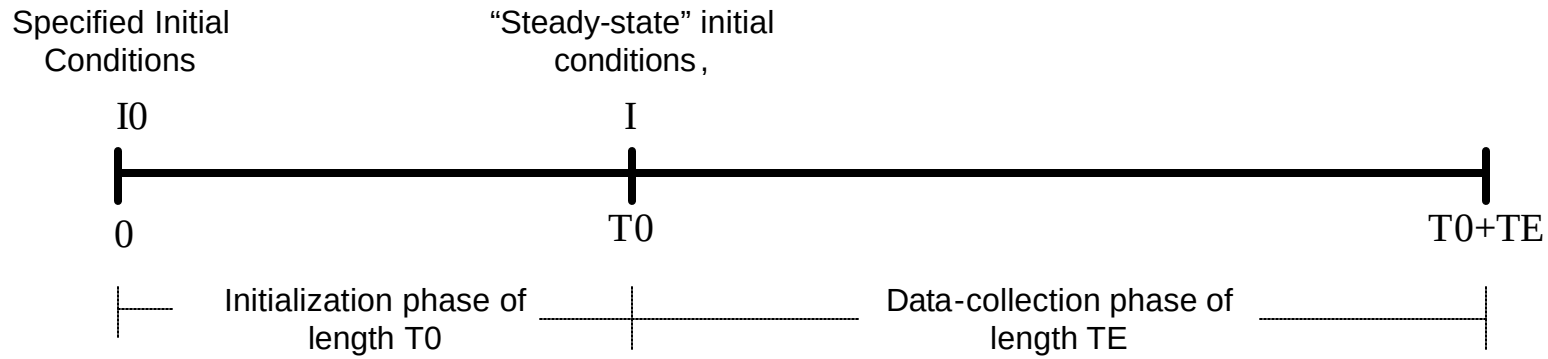
- 'p' repetitions of 'n' observatiosn involve a total of N = p.n observations.
- Dividing 'N' observations into number of ranges is quite difficult
- It is suggested that number of repetitions be as low as possible → reduction in the approximation of normal distribution with sample means

## Initialization Bias

- 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;
     1. Initialization phase from time 0 to $T_0$
     2. Data collection phase from time $T_0$ to stopping time $T_0+T_E$

# Initialization Bias




- 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

Two approaches to eliminate initial bias;

1. System can be started in a more representative state than the empty state
2. Or, first part of simulation run can be ignored i.e. take care of those data that come only after steady-state

Ideal situation to remove initial bias is to select initial conditions from steady state distribution.

More common approach to remove initial bias is to eliminate an initial section of the run. The run is started from an idle state and stopped after certain period of time. The run is restarted with statistics being gathered from the point of restart

# Elimination of Initial Bias

Pilot runs – set of runs which are conducted to judge how long the initial bias remains

Disadvantages

- More number of pilot runs required
- Eliminating the first part of a simulation introduces variance estimate, confidence limit etc. be based on less information