Assignment-2

Manoj. L
4AL16CS125

System Modelling & Simulation

Write a short note on the following topics

1. Data Collection

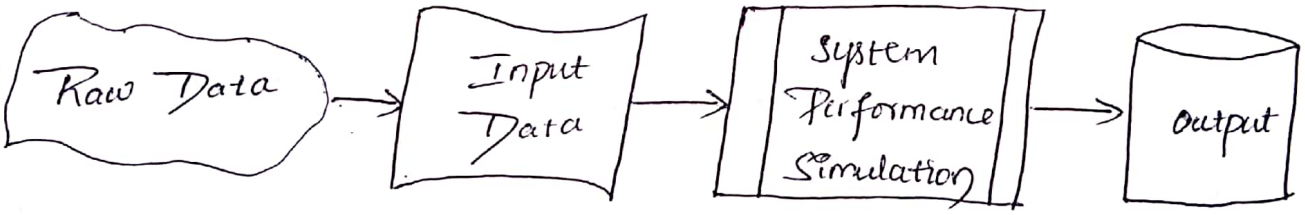
⇒① Data Collection is one of the biggest task in solving real problem.

⇒② It is One of the most Important and difficult Problems in Simulation.

⇒③ Even if when data are available. they have rarely been recorded in form that is directly useful for Simulation Input Modelling.

⇒④ "Garbage-in, Garbage-out" is a basic Concept in Computer Science and it applies equally in area of discrete system Simulation.

⇒⑤ Many are fooled by a pile of Computer Output or a Soplsticated animation, as if there where the absolute truth.



⇒⑥ Even when Model Structure is Valid Simulation results can be misleading, if the Input data is

- Inaccurately Collected
- Inappropriately analzed
- Not representative of the Environment.

* Suggestion that enhance data Collection:

(i) Plan ahead : Pre observing Session.

(ii) Analyze data being Collected.

(iii) check for Variable relationship.

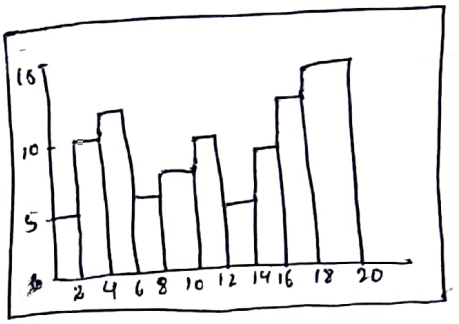
(iv) check for auto Correction

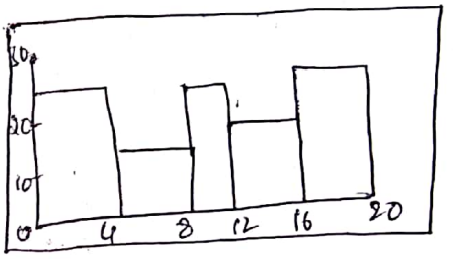
2] Identifying the Distribution with Data?

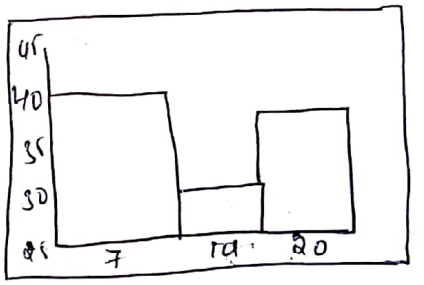
⇒ ⊙ A frequency distribution or histogram is useful in identifying the shape of a distribution.

⊙ A histogram is constructed as follows:

1. Divide the range of the data into intervals [usually of equal width; however, unequal width may be used if hight of frequencies are ajusted]
2. Label the horizontal axis to conform to the Intervals slected.
3. Determine the frequencies of occurance within each Interval.
4. Label the vertical axis so that total Occurance can be plotted for each Interval.
5. Plot the frequencies so that total Occurance on the Vertical axis.







⊙ The number of class Interval depends on
→ The use of Observation.
→ The dispersion of Data.
→ Suggested ~~Data~~ No of Intervals.

⊙ for continous data:-
→ Compondes to Probabilety density function of theretical distribution.

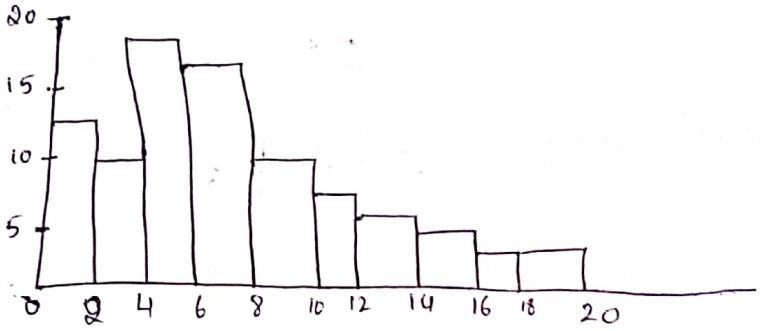
⊙ for discrete data:-
→ Coresponds to probabilrty on as function.

⊙ if few data Points are available.
→ Combins adjacent cell to elimenate the ragged apperence of histogram.

⊙ figure above shows same data with different Interval size.

Eg :- The No vehicle arriving at the nortest Corner of Interseaction in 5 min b/w 7AM & 7:05 AM was Monitored for 5 days over 90 Week Period. Table shows resulting data. The 1st entry in table indicate 12 of 5 Min period during which 0 vehicle arr -ived, 10 period during which 1 vehicle arrived & soon. The No of Automobils in a discrete Epoch. The No of Sample data, so the histogram may have a cell of each possible Value in range data.

Arrival Per period	Frequency
0	12
1	10
2	19
3	17
4	10
5	8
6	7
7	5
8	5
9	3
10	3
11	1



(3) Parameter Estimation :-

⇒ (1) After a family of distribution has been selected, the next step is to estimate Parameter of distribution.

(2) If Observation in a Sample of Size n are $x, x_2 \cdots x_n$

The Sample Mean & Sample Variance are :

- The Sample Mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- The Sample Variance is $S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2}{n-1}$
- If data are discrte in a frequence distribution, Then we can re-write the equation as :

$$\bar{x} = \frac{\sum_{j=1}^{k} f_i x_i}{n}$$

and $$S^2 = \frac{\sum_{j=1}^{k} f_j x_j^2 - n\bar{x}^2}{n-1}$$

where k is No of distint Value of x

f_j is obsered frequency of Value x_j of x.

③ If the data are continous we "discretize" Then estimate the Mean $\bar{x} \approx \frac{\sum_{j=1}^{n} f_j m_j}{n}$ and Varience $s^2 \approx \frac{\sum_{j=1}^{n} f_j m_j^2 - n\bar{x}^2}{n-1}$

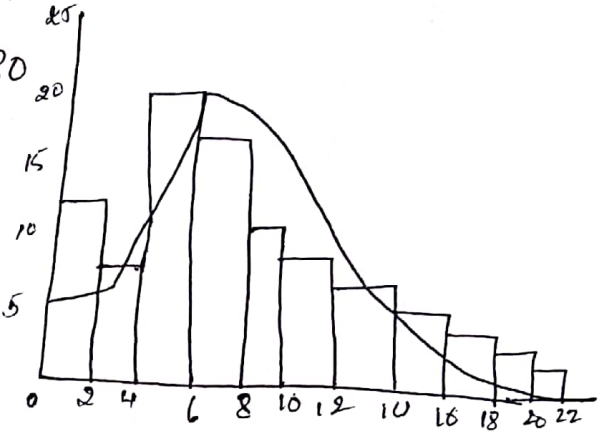
④ A parameter is an unknown Constant, but an estimator is statistic.

Eg: Vehicle Arrival Example.

Table in Histogram of Vehicle Example can be Analysis to obtain

$n_k = 100$, $f_1 = 12$ $x_1 = 0$ $f_2 = 10$ $x_2 = 1$

$\sum_{j=1}^{k} f_j x_j = 364$ and $\sum_{j=1}^{k} f_j x_j' = 2080$



* The Sample Mean and Varance are

$\bar{x} = \frac{364}{100} = 3.64$

$s^2 = \frac{2080 - 100 \times (3.64)^2}{99} = 4.63$

⑤ The Histogram Suggests x to have poisson Distribution.

- However, most Sample is Not equal to Sample varance.

Theoritically: Poisson With Parameter $\lambda \Rightarrow \mu = \sigma^2 = \lambda$

Reason: Each estimator is random Variable, its Not Perfect.

⑥ Goodness of fit Tests:-

⇒ • Provids helpful guidana for evaluating Suitability of a Input modal

- There is no single Correct distribution in real application exists.
- If Very little data are available, it is unlikely to reject any candidate distribution.
- If a lot of data are available, it is likely to reject all candidate distribution.
- Conduct hypothesis treting on Input data distribution using:
  - Chi-square test
  - Kolmagorov-Smirnov Test.

* Chi-Square Test:

One procedure for testing hypotheses of random size n of random variable x follows specific distribution form in chi-square. The test Procedure begins by arranging n observation into set of k class Interval & chi-statis is given by $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Where $O_i \rightarrow$ Observaved frequency

$E_i \rightarrow$ Excepted frequency.

The Expected frequency for each class Interval is Computed as $E_i = np_i$

The Hypothesis are following:

H_0 : The random variable x, conforms the distributional assumption with parameter given by Parameter Estimate.

H_1 : The random varable x does not conform.

- Each value of random variable should be class Interval, unless Combining is necesary and $p = p(x_i) = p(x = x_i)$
- for Continous case with Assumed $f(x)$ or assumed cdf $f(x)$, P_i can be computed by,

$$P_i = \int_{a_{i-1}}^{a_i} f(x)\,dx = f(a_i) - f(a_{i-1})$$

Table Bellow are Made to aid in determining The No of class Interval for Continous data.

Sample size $\hat{n}$	Number of Class Interval k
20	Do Not use chi-square
50	5 to 10
100	10 to 20
>100	$\sqrt{n} + 0.5$

⊛ chi-Square Test with equal Probabilities.

- If Continous assumption is being tested, class Interval are equal in probability rather than equal in width of Interval should be used.
- unfortunately, there is yet no Method for determining the probability associated with each Interval that Maximize the Power of that of given size

$E_i = n\ p_i \geq 5$

⊛ Kolmogoro-Smirnov Goodness of file Tests

This test is particularly useful when sample size are small & when no parameters have been estimated from data.

Eg :- Suppose 50 Interval arival Times are Collected Over the following 100 min Interval.

0.44, 0.53, 2.04, 2.74, 2.00, 0.30, 2.54, 0.52, 2.02, 1.89, 1.53, 0.21, 2.80, 0.04, 1.35, 8.32, 2.34, 1.95, 0.1, 1.42, 0.46, 0.07, 1.09, 0.75, 5.55, 3.93, 1.07, 2.26, 2.88, 0.61, 1.12, 0.26, 4.57, 5.37, 0.12, 3.19, 1.63, 1.46, 1.08, 2.06, 0.85, 0.83, 2.44, 2.11, 3.15, 2.9, 6.58, 0.64

H0 : The Interarrival Times are Exponentially distributed.

H1 :- The Interarrival Times are not exponentially distributed.

The Data Were Collected Over the Interval 0 to T = 100 min can be shocon $\{T_1, T_2, \ldots\}$ time exponential, the arrival times distributed on Interval (0, T). The arrival times T_1, T_1+T_2, $T_1+T_2+T_3 \cdots + T_{50}$ are obtained by adding times on a (0,1) interval, The points will be $[T_1/T, (T_1+T_2)/T, \ldots (T_1+T_2+\ldots+T_{50})/T]$.

0.0044	0.0097	0.301	0.0515	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3151	0.3556	0.3866	0.3878
0.3553	0.3561	0.3610	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.7553	0.7636	0.7880	0.7982	0.8206	0.8417	0.873	0.9022	0.968	0.9944

following the procedure in Kolmogorous-Smirnov test studied earlier, Procedure D^+ 0.1054 & D^- 0.0080. Therefore This statistic is

$D = \max(D^+, D^-) = \max(0.1054, 0.0080) = 0.1054$

The Critical Value of D for level of significance of $\alpha = 0.05$ & $n = 50$ is $D_{0.005, 50} = 1.36/\sqrt{n} = 0.1923$.

But $D = 0.1054$, so the hypotheses that the Interarrival Times are Exponentially distributed cannot be rejected.

(4) P-values & Best-fit :-

P value for testing statistics. The significationce level at which one would reject H0 for the given statistics.

A Measure of fit, The large the better,

- Large-p-value : good fit.
- Small p-value : poor fit.

(5) Fitting non stationary poisson process.

(*) fitting a NSPP data is difficult, possible approaches :

(i) fit a very flexible model with lots of parameters.

(ii) Approximate constant arrival rate over some basic Interval time, but vary it from time Interval.

(*) Suppose we need to model arrival rate time $[0, T]$, Our approach is most appropriate when we can :

(i) Observe the time Period repeatedly.

(ii) count arrival/round arrival times.

(iii) Divide the Time Period into k equal Interval of length $\Delta t = T/k$

(iv) Over n Period of observation let C_{ij} be no of arrivals during the Interval on j^{th} Period.

The estimated arrival rate during i^{th} time period $(i-1)\Delta t \le i\ \Delta t$ is:

$$\lambda(t) = \frac{1}{n\Delta t} \sum_{j=1}^{n} c_{ij}$$

Eg:- Divide 10 hours business day [8am, 6PM] into Interval k=20 whose length $\Delta t = 1/2$, & observe over n = 3 days.

Time Period	No. of Arrival			Estimate Arrival Rate (arrival/hr)
	Day1	Day2	Day3	
8:00 - 8:30	12	14	10	24
8:30 - 9:00	23	26	32	54
9:00 - 9:30	27	18	32	52
9:30 - 10:00	20	13	12	30

for Instances,
1/3 (0.5) * (23 + 26 + 32) = 54 Arrival/hour

(6) Selecting Input Models Without Data.

⇒ (*) If data is not available, some possible source to obtain Information about process are:

- Engineering data: Often product or process has performance rating Provided by Manufactures or Company rules specific time or Production Standard.
- Expert Option: Pessimistic and Most-likely times, and they may know the variability as well.
- Physical or Conventional limitations: physical limit on performance, limit or bounds that narrow the range of Input process.

(*) Eg:- Production Planning Simulation

Input of Sales Volume of Varbales product is required, salesperson of Product xyz says That:

q	Interval [sales]	PDF	cummulative frequency
1	1000 <= x <= 2000	0.1	0.10
2	2000 <= x <= 2500	0.65	0.75
3	2500 <= x <= 4500	0.24	0.99
4	4500 <= x <= 5000	0.01	1.00

- No fewer than 1000 units and no more than 5000 units will be sold.
- Given Experiences, says There is 90% chance of selling More than 2000 units, a 25% chance of selling more than 2500 unit & 1% of chance of selling More than 4500 units.
- Translating these information into a Amulative probability or being less than or Equal to Those goals simulation Input.

(7) Multivariate and Time Series Input Models.

⇒ The random Variable disussed until now were Considered to be Independent of any other Variables within The Context of The problem.

- However, Variables may be related.
- If they appear as I/p, The relationship should be investigated and taken into consideration.

(*) Multivariate Input Models:

- fixed, finite No of random Variables $x_1, x_2 \ldots x_n$
- for Eg. lead time and Annul demand for an Inventory Model.
- An Increase in demand results in lead time Increase, Heence Variable are dependent.

(*) Time Series Input Models:

- Infinit Sequence of random Variables Eg $x_1, x_2 x_3 \ldots$
- for Eg: Time b/w arrival of orders to buy & sell stocks.
- Buy and sell Orders tends to arrive in bursts hence, Times b/w arrivals are dependent.

(8) Types of Simulation with respect to output Analysis.

⇒ • Terminating Verses Non-Terminating Simulation.

• Terminating Simulation :

→ Runs for some duration of Time T_E, where E is a specified event that stops the simulation.

→ Starts at time 0 under well-specified initial Condition.

→ Prank eg: Opens at 8:30 AM (time 0) with no Customer Present at 8 of the 11 teller Working & closes at 4:30 PM (Time T_E = 480)

⇒ The Simulation Analyst choosen to Consider it a terminating system because the Object of Interest is One day's Operation.

(9) Stochastic Nature of output data.

⇒ Model output Consist of One or More random Variable (r.v) Because the Model is an I/o transformation and Input Variables are r.v's.

. M/G/1 queing eg:

→ poisson Arrival rate = 0.1 per MIN

Serice Time ~ N (m=9.5, s=1.75)

→ System Performance : long-run Mean queue length $L_Q(t)$.

→ Suppose we run a Single Simulation for a total of 5000 MIN.

. Divide the Time Interval [0, 5000] into 5 equal subintval of 1000.

Average No of Customer in queue from Time (j-1)1000 to j(1000)

• M/G/1 queuing

→ Batched Average queue length for 3 Independent replication

Batching Interval [Men]	Batch, j	Replication		
		1, y_{1j}	2, y_{2j}	3, y_{3j}
0, 1000	1	3.61	2.91	7.67
1000, 2000	2	3.21	9.00	19.53
2000, 3000	3	2.18	16.15	20.36
3000, 4000	4	6.92	24.53	8.11
4000, 5000	5	2.82	25.19	12.62
0, 5000		3.75	15.56	13.66

• Inherent Variability in Stochastic Simulation both within a single replication and across different Application

• The Average across 3 replications, can be regarded as Independent Observations, but Average within a replication $y_1 \cdots y_{15}$ are Not.

(10) Absolute Measure of Performance and there estimation.

⇒ • Consider the estimation of a performance parameter, $q(or\ f)$ of Simulated system.

→ Discrete Time data : $[y_1\ y_2 \ldots y_n]$ with Ordinary Mean : q

→ Continous-time data : $\{y(t), 0 \le t \le T_E\}$ with time-weighted Mean : f

• Point estimation of discrete time data

→ The Point Estimator : $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i$

is unbaised if it's Expected value is θ, i.e if $E(\hat{\theta}) = \theta$.

Is biased if:

* Point Estimator

- Point estimator for continous-Time data.
  - → The point estimator: $\hat{\phi} = \frac{1}{T_E} \int_0^{T_E} y(t)\,dt$
  - → Is biased in General Where,
    - An unbiased or loco-bairs estimator is desired.

- usually, System performance Measure can be put into Common framework of q or f:

Eg: The performance of days on which sales are lost through an out of stock situation let:

$$y(t) = \begin{cases} 1, & \text{if out of stock on day } t \\ 0, & \text{otherwise.} \end{cases}$$

- Performance Measure that does not fit: quantile or percentile.

→ Estimating quantiles: The Inverse of problem of estimating a Proportion or probability. $Pr\{y \le 0\} = p$.

→ Consider a Histogram of observed value y: find such that 100 % of The histogram is to the left of.

(*) Confidence-Interval Estimation:

- To understand this fully, it is ~~histogram~~ important to distinguish b/w Measures of errors, risk, eg Confidence Interval versus Prediction Interval.
- Suppose the Model is Normal distribution with Mean & Variance s^2 (both unknown)

→ Let y_i be avg cycle time for parts produced on i^{th} replication of the Simulation.

→ Avg cycle time will vary from day to day, but Over the long-run the average of the averages will be close to q.

→ Sample Variance accross R replication : $S^2 = \frac{1}{R-1} \sum_{r=1}^{R} (y_r - \bar{y})^2$

* Confidence - Interval Estimation

→ Confidence Interval (CI):

- A Measure of Error.
- Where y_r are normal distributed : $\bar{y}.. \pm t_{\alpha/2, R-1} \frac{S}{\sqrt{R}}$
- We cannot know for certain how far is q but C, attempts to bound That Errors.
- A CI, 95% tell us how we can trust Interval Actually bound That Error.
- The More replication We Make, Less Error Occurs.

→ Prediction Interval (PI):

- A Measure of Error.
- ~~Where~~ A good guess for Average cycle time on a particular day is Our estimator but it is unlikely to be exactly right.
- PI is designed to be Wide enough to Contain the Actual Avrage cycle time On any particular day with high probability.
- Normal - Theory Prediction Interval:

$$\bar{y} = \pm t_{\alpha/2, R-1} S\sqrt{1+1/R}$$

- The length of PI will not go to 0 as R Increase because We can never Simulate away risk.
- PI's limit is : $\theta \pm z_{\alpha/2}\sigma$.

(11) Output Analysis for terminating Simulation.

⇒ A Terminating Simulation : runs over simulated time interval $[0, T_E]$

A common goal is estimate :

a) $\theta = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]$, for discrete output.

b) $\phi = E\left[\frac{1}{T_E}\int_0^{T_E} y(t)dt\right]$ for continous o/p $y(t)$. $0 \le t \le T_E$

In General, ~~Input~~ Independent replication are used, each run using a different random no stream & Independently choosen initial conditions.

* Statistical Background.

- Important to distinguish within replication data from across-replication data.
- For eg, Simulation of Manufacturing System.
  - (i) a performance Measure of system : cycle time for parts & work in process (wip).
  - (ii) Let Y_{ij} be cycle time for j^{th} part production in i^{th} replication
  - (iii) Across Replication data are formed by summarizing within Replication data.
- Overall Sample avg and interval replication Sample Avg are always unbaised estimators of Expected daily avg cycle time or daily avg wip
- Across replication data are Independent and identically distributed, but within-replication data do not have these properties.

12) Output Analysis for steady-time Simulations.

⇒ Consider a single run of Simulation Model to estimate a steady state long-run characteristics of system

- The Performance Measure

$$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} y_i, \text{ for discrete Measure.}$$

$$\phi = \lim_{T_E\to\infty} \frac{1}{T} \int_0^{T_E} y(t)\,dt, \text{ Continous Measure.}$$

- The Sample size is design choice, With several Consideration in Mind.

(i) Initialization Bias

Method to reduce the Point-estimator bias caused by using artifical & unrealistic

→ Intelligent Initialization

- Initialize the Simulation in a state that is More representation of long run Conditions.
- If the system exists, Collect data on it and use these data to specify more nearly typical initial Conditions.

(ii) Error Estimation

- If $\{y_1 \dots y_n\}$ are not statistically Independent then S^2/n is biased estimator of true Variance
- Suppose point estimator $\hat{\theta}$ is Sample Mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- Variance of $\bar{y}$ is very hard to estimate.
- for sly with steady state, Produce an o/p process That is Approximately covariance stationary

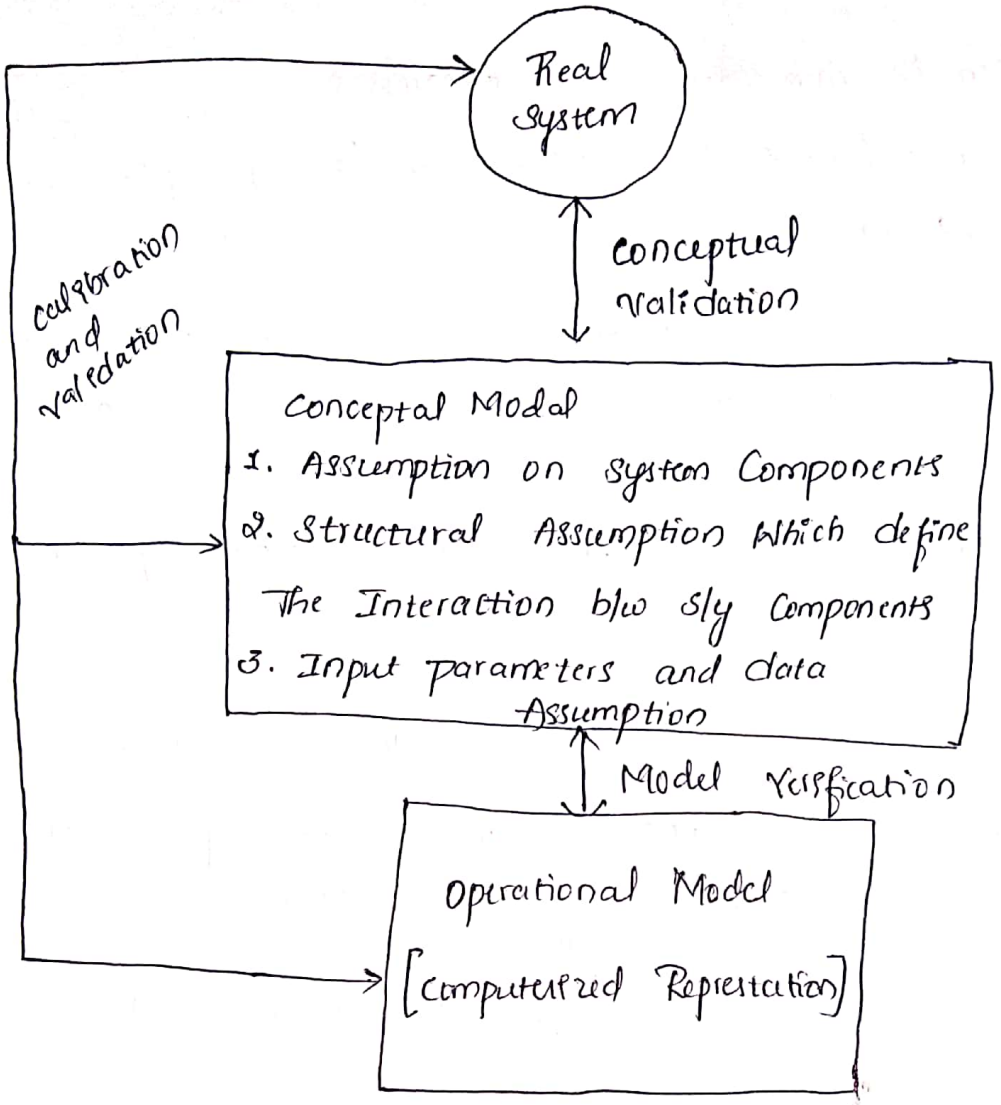
(iii) Replication Method

- Approach make R replication initializing & deleting from each one That same way.
- Important to do through job of Investigation Condition bias:
- Basic Rao o/p data $\{y_{ij}, r=1, R, j=1, \ldots n\}$ is desired
  - Individual Observation from within 'r'
  - Patch Mean from within replication 'r' of some of no. of discrete-time Observation).

(13) Model building, Verification and Validation.

- The first step in Model building consists of Observing real system and Interactions among its various Components & Collecting data on its behavior.
- Operators, technicians, repairs and Maintance personnel, Engineer and Managers under certain aspect of system which may be unfamiliar to Other.
- As Model development proceeds, new question may arise & Model development will return to this step of learning true system structure & behaviour.
- The second step in Model building is the Construction of a Conceptual Model → a Collection of Assumption on Components & structure of system, plan Hypothesis on The Value of Model into Parameter, illustrate by following figure.

⊙ The Third step is Translation of the Operational Model into Computer recognizable form - The Computerized Model.



(14) Verification of Simulation Model

⇒ The purpose of Model Verification is to Assume That Conceptual Model is reflected accurately in Computerized representation.

- Make someOne else check The Model.
- Make a flow diagram That includes each logically Possible Action a s/y can take during an event.
- closly Examine The model o/p for reasonable under a Variety of Input Parameter setting.
- Make the Operational Model as self documentory as possible.
- The Intuactive run Controller (IRC) assist in debugging in the following ways:

(a) Simulation can be monitared as it progresses.

(b) Attention can be fowused On a Particular line of logic.

(c) Values of Selected Model Component can be Observed

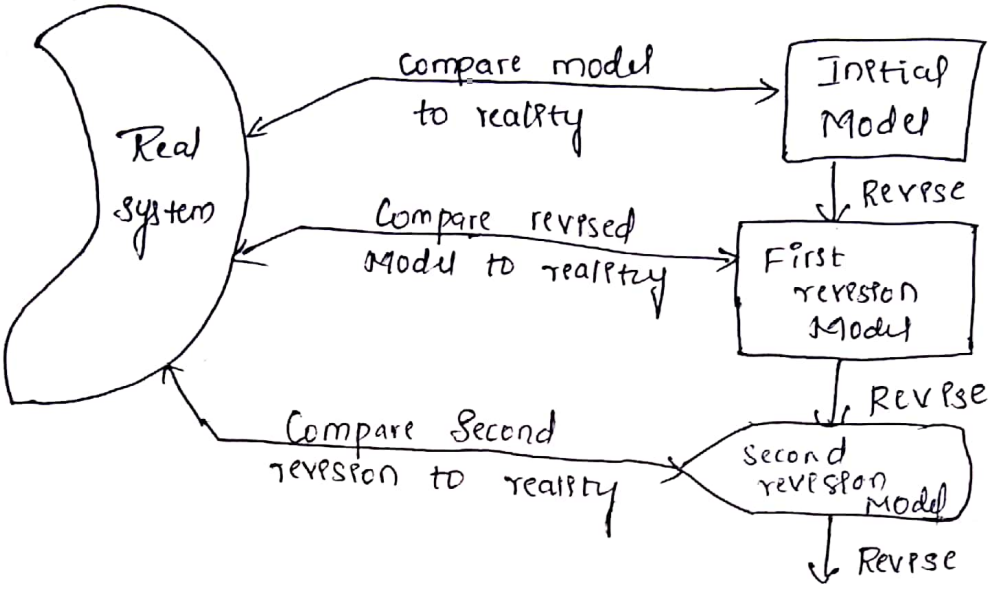
(d) The Simulation can be temporarily suspended or pased, not only to view Information but also to reassign Values or redirect entities.

- Graphical Interfaces are recommended for accomplishing Verification and Validation.

(15) Calibration and Validation for Models, Optimization via Simulation.

⇒ Verification and validation Although are Conceptually distinct, usually are Conducted Simultaneously by the Models.

- Validation is the Overal procus of Comparing the Model to real system and its behavsour.
- Caliberation is the Interactive procus of Comparing the Model to real System, Making adjustment to the modal, Comparing again & so on.
- Comparison of Model to real System.
- Subjective Tests. People who are knowledgable about the System.
- Objective Tests - Requires Data On eval System's behaviour & the output of the Model.
- Validation is not an proposition - no model is ever totaly reprentative of sly under Study.
- In Addition, each revesion of Model, as to fig. Involves some cost, tims and efforts.



- As an aid in Validation Process, Naylor and finger formulated a 3 step approach which been followed:

1) Build a Model that has high face validity.

2) Validate Model Assumption.

3) Compare the Model Input-output transformation to Corresponding I/o transformation for the real sly.