System Modelling & Simulation (Assignment-II)

Manish.B.Shriyan
4AL16CS131
VIII-B

Write a note on following topic:

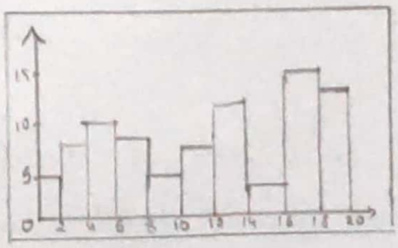
1) Data collection:

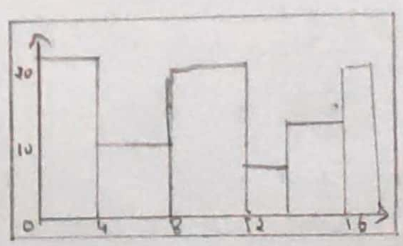
→ ① Data collection is one of the biggest task in solving real world problem
② It is the most important and difficult problem in simulation
③ Even if when data are available, they have rarely been recorded in form that is direltly useful for simulation input modelling.
④ "GIGO" or "Garbage-in, Garbage Out" is a basic component in computer Science & it applies equally in area of discrete system simulation.
⑤ Many are fooled by a pile of computer output or a sophesticated 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.
- Inappropriotely analysed.
- Not representative of the environment.

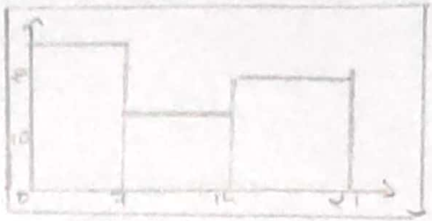
Suggestion: that enhance data collection : ① plan ahead: processing Session. ② Analyze data bring collected ③ Check for variable relationship ④ check for autocorrection.

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 with; however unequal width may be used if hight of frequencies are adjusted.
② Lable the horizontal axis to confirm to the intervals selected.
③ Determine the frequencies of occurences within each interval.
④ label the vertical axis so that occurences can be plotted for each interval.
⑤ plot the frequencies on the vertical axis. So that total occurences on the





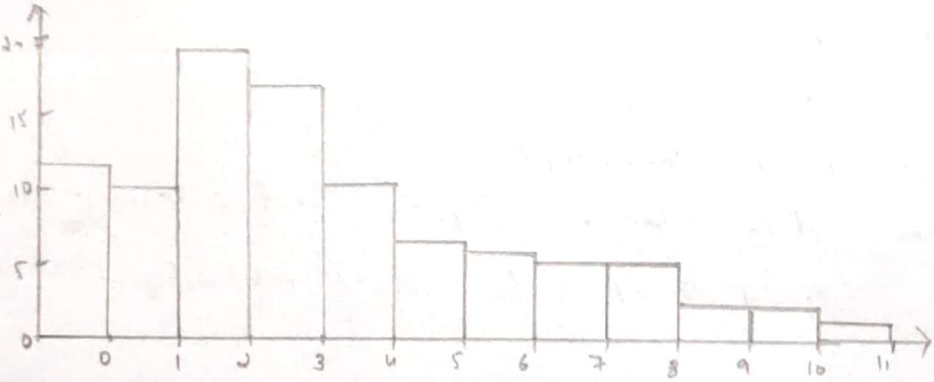
⑥ The number of class interval depends on
-> The no. of observation
-> The dispersion of data.
-> suggested No. of intervals.



⊙ For continuous data:
-> corresponds to probability density function of theoretical distribution.
⊙ For discrete data:
-> corresponds to probability mass function.
⊙ if few data points are available.
-> combine adjacent cell to eliminate the legal appearance of diagram.
- Figure above shows some data with different inter sizes.

Ex: The No. of vehicles arriving at the northeast corner of intersection in 5 min b/w 7 am & 7.05 am. was monitored for 5 day over 20 week period. Table shows resulting data. The 1st entry in table indicate 12: 5 min period during which 0 vehicles arrived & so on. The no. of automobiles in a discrete variable & there are sample data, so the histogram may have a cell for each possible value in range. The resulting histogram is shown below.

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 $x_1, x_2, \dots x_n$, the sample mean & sample variance are: ⊙ The sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
⊙ The sample variance is $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$
⊙ If data are discrete in a frequency distribution, then we can re-write the equation as
$\bar{x} = \frac{\sum_{j=1}^{k} f_j x_j}{n}$ and $s^2 = \frac{\sum_{j=1}^{k} f_j x_j^2 - n\bar{x}^2}{n-1}$
where k is no. of distinct value of x and f_j is observed for value x_j of X.
⑤ If the data are contineous we "discretize" then estimate the mean $\bar{x} = \frac{\sum_{j=1} f_j m_j}{n}$
and variance $s^2 = \frac{\sum_{j=1} f_j m_j^2 - n\bar{x}^2}{n-1}$
where f_j is observed frequency of j^{th} case interval, m_j is point of j^{th} interval & c is no. of class intervals

(a) A parameter is an unknown constant, but an estimator in statistic.
Eg: Vehicle arrival example. Table in histogram of vehicles example can be analyzed to
$n_x = 100$, $f_1 = 12$, $x_1 = 0$, $f_2 = 10$, $x_2 = 1$. & $\sum_{j=1}^{h} f_j x_j = 364$ and $\sum_{j=1} f_j x_j^2 = 2080$

- The Sample mean & variance are

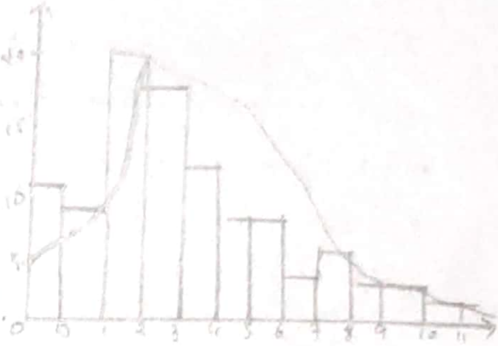
$$\bar{x} = \frac{364}{100} = 3.64$$

$$s^2 = \frac{2080 - 100 \times (3.64)^2}{99} = 7.63$$

- The histogram suggests x to have poisson distribution.
- However, mean sample is not equal to sample variance.

Therically: poision with parameter $\lambda \Rightarrow u = \sigma^2 = x$

Reason: Each estimator is random variable, its not perfect.



4) Goodness-of fit tests:

→ ⊙ previously helpful guidance for evaluating suitability of a input model exists.
⊙ There is no single correct distribution in usual application exists.
⊙ If very little data are available, it is unlikely to reject all candidate distribution.
⊙ Conduct hypothesis testing on input data distribution using:
→ Chi-square test & → kolmgrov-Smirnov Test.

(*) Chi-Square test:
one procedure for testing hypothesis random size n of random variable + follows specific distribution form in random variable x follow specific distribution form in specific distribution form in random variable x follow specific distribution form in specific class chi-square, the test procedure begins by arranging n-observation into sets of k class intervels or statistics is given by $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ where $O_i \to$ observed frequency, $E_i \to$ Expected frequency.
The expected frequency for each class internals is computed as $E_i = nP_i$
where $P_i \to$ is theoretical hypothesis probability associated with ith class interval.

The hypothesis are following:
H0: The random variable x, conforms the distributional assumption with parameters given by parameter estimate.
H1: the random variable x does not conform.

⊙ Each value of random variable should be class interval unless combining is necessary as $P = p(x_i) = p(x = x_i)$
⊙ for continous case with assuming pdf $f(x)$, or assumed cdf $F(x)$, p_i can be computed by,

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

where a_{i-1} $a_i \to$ endpoints of ith class internal
$f(x)$ assumed pdf.
$F(x)$ assumed cdf

Table below are made to aid in determing the no. of class interval for continous data.

Sample size n	Number of class interval k
20	Do not use chi-square
50	5 to 10
100	10 to 20
100	$\sqrt{n}$ to $n/5$

Table : Recomendation for no. of class-internal for contineous data.

⊛ Chi-square Test with equal Probabilities :

⊙ if contineous assumption is being tested class interval are equal in probability rather than equal in width of internal should be read.

⊙ unfortunately there is not method for determining the probability associated with each internal that maximize the power of that of given size.

$E_i = n \; p_i \geq 5$

Substituting for p_i yields. $n/k \geq 5$ & Solving for k yields $k \leq n/5$

⊛ kolmagorov - Smirnov Goodness - of - fit - tests :

This test is particularly useful when sample size are small & when no parameters have been estimated from data and Ex: Suppose 50 interarrival times are collected over the following 100 min interval.

0.44, 0.53 2.04, 2.74, 2.00, 0.30, 2.45, 0.52, 2.02, 1.89, 1.53, 0.21, 2.80, 0.04, 1.35, 8.32, 2.34, 1.95 0.10, 1.42, 0.46, 0.071, 1.09, 0.76, 5.55, 3.93, 1.07, 2.26, 2.88, 0.67, 1.21, 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, 2.90, 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 show d $T_1, T_2 ..$ } time exponential, the arrival times distributed on internal $(0, T)$. The arrival $T_1, T_1+T_2, T_1+T_2+T_3 + .. + T_1 + ... T_{50}$, are obtained by adding time on a $(0,1)$ internal the points will be $[T_1/T, (T_1+T_2)/T, \; .. \; (T+ .. + T_{50})/T]$

0.0044	0.0097	0.301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1055	0.1676	0.1956	0.1960	0.2095	0.2927	0.1161	0.3033	0.3366	0.3508
0.3535	0.3561	0.3670	0.3746	0.3746	0.4694	0.4796	0.5225	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6514	0.6845	0.7009	0.7154	0.7262	0.7408
0.7553	0.7636	0.7880	0.7982	0.7982	0.8417	0.8734	0.9022	0.9660	0.9944

Following the procedure d D^+ of 0.1054 d D^- of 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 $\alpha = 0.05$ & $n = 50$ & 0.005, so $= 1.36/\sqrt{n} = 0.1923$; but $D = 0.1054$, So the hypothesis that the interarrival times are exponentially observed cannot be rejected.

⊛ P-Values & Best fit :

p-value for test statistics - the significance level at which we would reject H0 for the given statistis.

A measure of fit, the larger the better

⊙ large - p-value : good fit

⊙ Small p-value : poor fit.

5) Fitting non stationery puission process
→ ① Fitting a NSPP data in difficult, possible appreaches.
② Fit a very flexible model with lots of paramates.
③ Approximate constant arrival over some basic interval time but vary it from time interval.
④ Suppose we need to model arrival over time $[0, T]$ over approach is most appropriate.
(i) observe the time period repeatedly.
(ii) Count arrival/record arrival time.
(iii) Divide the time period into k equal interval of length $\Delta t = T/k$
(iv) Over n period of observation let a_{ij} no. of arrivals during ith interval on jth period.

The estimated arrival rate during ith time period $(i-1)\Delta t \le i\,\Delta t$ is:

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

• n = no. of observation periods
• Δt = time $n\Delta t$ = interval length
• a_{ij} = no. of arrival during time interval of jth observation period.

Ex: Divide 10 hours buisness day [8am. 6pm] interval $k = 20$ whose length $\Delta = 1/2$. observe over $n = 3$ days.

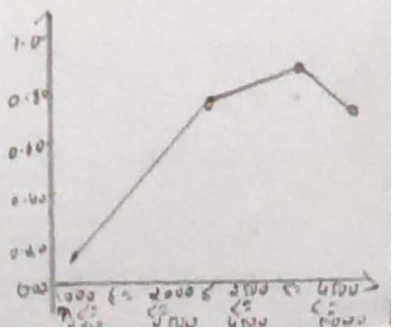
Time Period	No of Arrivel Day 1	Day 2	Day 3	Estimate Arrival Rate (arrivals/hr)
8:00 - 8:30	12	14	10	24
8:30 - 9:00	23	26	32	54
9:00 - 9:30	27	18	32	52
9:30 - 10:00	20	13	12	30

For instance, $1/3(0.5) * (0.3 + 26 + 32) =$ 54 arrival/hr.

6) Selecting input models without data.
→ * if data is not available, some possible source to obtain information about process are:
① Engineering Data: often product are process has performace ratings provided by manufactures, or company rules specific time or production standard.
② Expert option: perssimitic and most-likely times and they may know the variability.
③ Physical or conventional limitations: physical limit on performance, limit or bound, that narrow the range of input process.

Ex: Production planning simulation.
① Input of sales volume of various product is required, salesperson of product XYZ says that: → No fewer than 1000 units & no more than 5000 units, at 90% choose of selling more than 4500 units
- Translating these information into a cumulative probability or being less than or equal to those goals for simulation input:
① The nature of the process
② The uniform triangular & beta distribution are often used as input models

i	interval (Sales)	PDF	Cummulative frequency
1	$1000 \le x \le 2000$	0.1	0.10
2	$2000 \le x \le 2500$	0.65	0.75
3	$2500 \le x \le 4500$	0.24	0.99
4	$4500 < x \le 5000$	0.01	1.00



7) Multivariate & Time series input models.

- ⊙ The random variable discussed until now we considered to be independent of any other variables within the context of the problem.
  - However, variable may be related
  - If they appear as input, the relationship should be investigated & taken into consideration

* Multivariate input model:
  ⊙ Fixed, finite numbers of random variable $x_1, x_2, \cdots x_n$
  ⊙ Ex: lead time & annual demand for an inventory model.
  ⊙ An increase in demand results in lead time increase, hence variable are dependent

* Time Series input Model:
  ⊙ Infinite sequence of random variables, ex: $x_1, x_2, x_3 \ldots$
  ⊙ Ex: time between arrival of orders to buy & sell stocks.
  ⊙ Buy & sell orders tends to arrive in bursts; hence, time between arrival are dependent.

8) Types of simulation with respect to output analysis.

-> ⊙ Terminating versus non-terminating simulation.
  ⊙ Terminating Simulation:
  -> Runs for some duration of time T_E, where E is a specified unit that stops the simulation.
  -> Starts at time 0 under well-specified intial conditions.
  -> Ex: opens at 8:30 am (time 0) with no customers present at 8 of the or 11 teller working. (initial condition) & closes at 4:30 pm (Time $T_E = 80$)
  -> The simulation analyst chooses to consider it a terminating system because the object of interest is one days operation.

9) Stochastic nature of output data:

-> ⊙ Model output consists of one or more random variables (r.v) because the model is an input-output model transformation & input variables are r.v.s.
  ⊙ M/G/1 queuing eg:
  -> poisson arrival rate = 0.1 per minute. Service time ~ N(m = 9.5, s = 1.75)
  -> System performance: long-term mean queue length $L_Q(t)$.
  -> Suppose we seen a single simulation for a total of 5000 minutes.
  -> Divide the time interval [0, 5000) into 5 equal subinterval of 1000
  -> Average no. of customers in queue from time $(j-1)1000$ to $j(1000)$ is Y_j
  • M/G/1 queuing eg (cont:)
  -> Batched Average queue length for 3 independent replication

Batching interval (minutes)	Batch j	REPLICATIONS 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 $Y_{11} \dots Y_{15}$ are not.

(10) Absolute measure of performance and their estimation.
→ ⊙ Consider the estimation of a performance parameter, θ (or ϕ) of simulated system.
 → Discrete time data: $[Y_1, Y_2 \dots Y_n]$ with ordinary mean: θ
 → Continuous time data: $\{Y(t),\ 0 \le t \le T_E\}$ with time weighted mean: ϕ.
⊙ Point estimation of discrete time data.
 → The point estimator: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, is unbiased if its expected value is θ.
 i.e. if: $E(\hat{\theta}) = \theta$, is biased if:

* point estimator:
 • point estimator for continuous time data
 → The point estimator: $\hat{\phi} = \frac{1}{T_E}\int_0^{T_E} Y(t)\,dt$
 → Is biased in general when: An unbiased or low-bias estimator is desired.
• Usually, system performance measures can be put into the common framework of θ or ϕ:
eg: The performance of days on which sales are lost through out of stock situation let:

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

⊙ Performance measure that doesnot fit: quantile or percentile:
→ estimating quantile: the increase of problem of estimating a proportion or probability $Pr\{Y \le \theta\} = P$.
→ Consider a histogram of observed value Y: Find such that 100p% of the histogram is to the left of :

* Confidence-Interval Estimation:
⊙ To understand this fully, it is important to distinguish b/w measures of error, i.e. eg: confidence interval versus prediction interval.
⊙ Suppose the model is normal distribution with mean θ, variance σ^2 (both unknown).
→ let Y_i be avg cycle time for parts produced on ith replication of simulation.
→ Avg cycle time will vary from day to day, but over the long-term the avg of the averages will be close to θ.

→ Sample variance across R replication : $s^2 = \frac{1}{R-1}\sum_{i=1}^{R}(Y_i - \bar{Y}_{..})^2$

* Confidence - Interval estimation

→ Confidence interval (CI):

- A measure of error.
- when Y_i are normal distributed: $\bar{Y}_{..} \pm t_{\alpha/2, R-1}\frac{S}{\sqrt{R}}$
- We cannot know for certain how far is $\bar{Y}$ but C_I attempts to bound that error.
- A CI, 95% tell us how we can trust interval actually bound the error.
- The more replication we make, less error occurs.

→ prediction interval (PI):

- A measure of risk.
- A good guess for average cycle time on a particularly is over estimator but it is unlikely to be exactly right
- PI is designed to be wide enough to contain the actual average cycle time on any particular day with high productivity.
- Normal - theory prediction interval:

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

- The length of P_I will not go to 0 as R increases because we can never simulate away risk.
- PI's limit is : $\theta \pm Z_{\alpha/2}\sigma$

11) Output analysis for terminating

→ 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 output $Y(t)$, $0 \le t \le T_E$

In general, independent replication are used, each run using a different random no. stream & independently chosen initial conditions.

* Statistical Background.

- Important to distinguish within replication data from across - replication data.
- For Eg: simulation of manufacturing System.

(i) 2 performance measures of system: cycle time for part and work in (WP).

(ii) Across replication data are formed by summarizing within replication data.

(a) Overall samples avg and interval replication sample avg are always unbiased estimators of expected daily avg cycle time or daily avg WP.

(b) Across replication data are independent and identically distributed, but within - replication data donot 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.

$\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i$, for discrete measure.

$\phi = \lim_{T_E\to\infty} \frac{1}{T} \int_0^{T_E} Y(t)\,dt$, 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 artificial & unrealistic.

→ Intelligent initialization

- Initialize the simulation in a state that is more representative of long run conditions.

⊙ If the system exists, collect data on it and use these data to specify more nearly typical initial conditions.

→ Divide simulation into 2 phases:

- An initialization phase from 0 to time T_0.

(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 system with steady state, produce an output process that is approximately stationery.

(iii) Replication Method: ⊙ Approach make R replication initializing & deleting from each one the same way.

⊙ Important to do through j obj of investigation condition bias:

- Basic row output data $\{Y_{rj}, r=1, R, j=1, \dots n\}$ is derived

- Individual observation from with 'r'.

- Protch mean from within replication 'r' of some of no. of discrete time observation.

---

13) Model building, verification & validation.

– ⊙ The first step in model building consists of observing real system & interaction among its various components & collecting data on its behaviour.

⊙ Operators, techicians, repairs & maintainance personnel, engineer & managers under certain aspects of system which may be unfamilier to other.

⊙ As model development proceeds, new question may arrives & model developers will returns to this step of learning true system structure & behaviors.

⊙ 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 parameters, illustrated by following figure.

⊙ The third step is translation of the operational model into computer recognizable form - the computerized model.

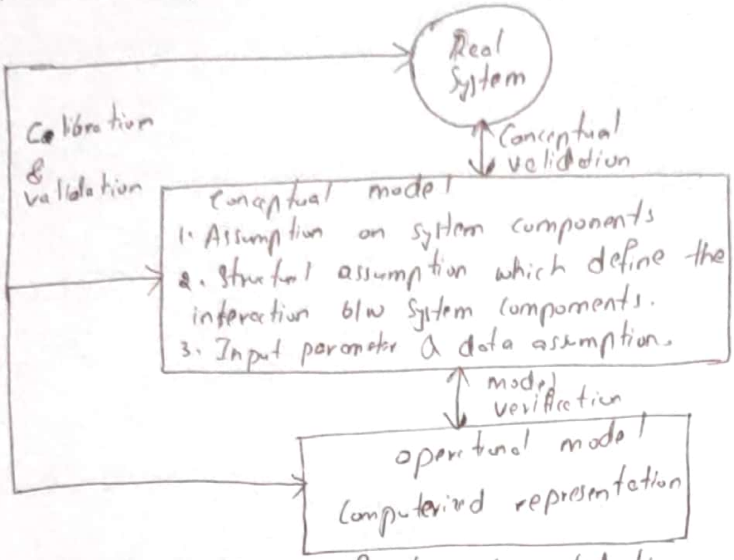


Fig: Model Building, verification & validation.

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 indicates each logically possible action a system can take during an event.

⊙ Closely examine the model output for reasonableness under a variety of input parameter.

⊙ Make the operational model as self documentary as possible.

⊙ The interactive run controller (IRC) assist in debugging in the following ways:

ⓐ Simulation can be focused on a particular line of logic.

ⓑ Attention can be focused on a particular line of logic.

ⓒ Values of selected model components can be observed.

ⓓ The simulation can be temporarily suspended or paused, not only to view information but also to reassign value or redirect entities.

⊙ Graphical interface are recommended for accomplishing verification & validation.

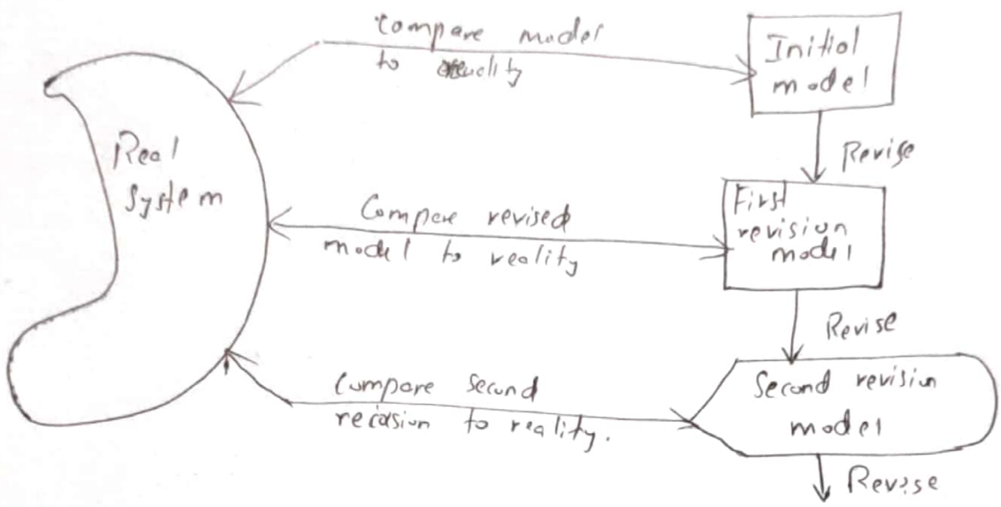
15) Caliberation & validation for models, optimization via simulation.

→ ⊙ Verification & validation although are conceptually distinct, usually are conducted simultaneously by the models.

⊙ Validation is the overall process of comparing the model to real system & its behaviour.

⊙ Caliberation is the interactive process of comparing the model to real system, making adjustments to the model, comparing again & so on.

- Comparison of model to real system.
- Subjective tests - Requires data on real systems behaviours & the output of the model.
- ~~Subj~~ Validation is not an proposition - no model is ever totally representative of system under study.
- In addition, each revision of model, as in fig. involves some cost, time and efforts.



- As an aid in validation process, Naylor & Tinger 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 input-output transformation for the real system.