

## Chapter-4

### Generation of Random Numbers

Random numbers are basic ingredient in the simulation discrete systems. A random number is a number generated by a Simulation program, whose outcome is unpredictable, and which cannot be subsequently reliably reproduced.

#### Properties of Random Numbers

There are two important statistical properties: uniformity and independence:

Each random number  $R_i$  is an independent sample drawn from a continuous uniform distribution between 0 and 1. The probability density function (pdf) is given by

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$



Figure 4.1: The pdf for random numbers

The expected value of each  $R_i$  is given by

$$E(R) = \int_0^1 x dx = \left. \frac{x^2}{2} \right|_0^1 = \frac{1}{2}$$

The variance is given by

$$\begin{aligned}
 V(R) &= \int_0^1 x^2 dx - [E(R)]^2 \\
 &= \left[ \frac{x^3}{3} \right]_0^1 - (1/2)^2 = 1/3 - 1/4 \\
 &= 1/12
 \end{aligned}$$

The consequences of uniformity and independence properties are:

If the interval (0, 1) is divided into n classes or subintervals of equal length, then the expected number of observations in each interval is N / n, where N is the total number of observations.

The probability of observing a value in a particular interval is independent of previous values drawn.

There are numerous methods that can be used to generate the values.

There are a number of important considerations for generating the random numbers.

The routine should be fast. Individual computations are inexpensive, but simulation could require many millions of random numbers. The total cost can be managed by selecting a computationally efficient method of generation.

The routine should be portable to different computer, simulation program will produce the same results wherever it is executed.

The routine should have a sufficiently long cycle. The cycle length, or period, represents the length of the random number sequence before previous numbers begin to repeat themselves in an earlier order.

The random numbers should be replicable. Given the starting point it should be possible to generate the same set of random numbers. completely independent of the system that is being simulated.

Most important, the generated random numbers should closely approximate the ideal statistical properties of uniformity and independence.

## **Generation of Pseudo-Random Numbers**

Pseudo" means false, so false random numbers are being generated. The pseudo random numbers behave like a random number but they are not random numbers. For example, 2, 6, 4, 9, 1 looks like a random but they are not random numbers. They can replicate easily and small in numbers.

Pseudo random numbers can be easily identified. Generated numbers might not be uniformly distributed, discrete-valued instead of continuous-valued, mean and variance are too high or too low.

### **Techniques for Generating Random Numbers**

The most widely used techniques for generating random numbers are:

Middle square Method

Linear Congruential Method (LCM)

Multiplicative Congruential Method (MCM)

Combined Linear Congruential Generators (CLCG)

#### **Middle square Method:**

In middle square method, the center number of square results is used to generate the next random number. For example,  $(73)^2 = 5329$ ,

$(32)^2 = 1024$ ,

$(02)^2 = 0004$ ,

$(00)^2 = 00$

OR  $(77)^2 = 5929$ ,

$(92)^2 = 8464$ ,

$(46)^2 = 2116$ ,

$(11)^2 = 0121$ ,

$(12)^2 = 0144$ ,

$(14)^2 = 0196$ ,

$(19)^2 = 0361$ ,

$$(36)^2=1296,$$

$$(29)^2= 0841,$$

$$(84)^2=7056,$$

$$(05)^2=0025,$$

$$(02)^2= 0004,$$

$$(00)^2=00.$$

### **Linear Congruential Method (LCM):**

To find the random number it uses linear mathematical equation. This method produces a sequence of integers,  $X_1, X_2 \dots$  between 0 and  $m-1$  by following a recursive relationship:

$$X_i=(X_{i-1}+C) \bmod M$$

The initial value  $X$  is called seed value.  $a, c, m$ , are constant, drastically affects the statistical properties and the cycle length.

If  $c \neq 0$  in the above equation, then it is called as Mixed Congruential method.

If  $c = 0$  the form is known as Multiplicative Congruential method.

The random numbers ( $R_i$ ) between 0 and 1 can be generated by

$$R_i = \frac{X_i}{m}, \quad i = 1, 2, \dots$$

Use linear congruential method to generate sequence of random numbers with  $X_0 = 27, a = 17, c = 43$ , and  $m = 100$ .

The random integers ( $X_i$ ) generated will be between the range 0 - 99

Equations  $\rightarrow X_{i+1} = (aX_i + c) \bmod m, R_i = X_i / m, i=1, 2, \dots$

$$X_1 = (17 * 27 + 43) \bmod 100$$

$$=(459+43) \bmod 100$$

$$=(502) \bmod 100$$

$$= 2, R_1 = 2 / 100 = 0.02$$

$$X_2 = (17 * 2 + 43) \bmod 100 = 77, R_2 = 77 / 100 = 0.77$$

$$X_3 = (17 * 77 + 43) \bmod 100 = 52, R_3 = 52 / 100 = 0.52$$

⋮

Hence the numbers are generated.

Using the multiplicative congruential method, find the period of the generator for  $a = 13$ ,  $m = 2^6$  and  $X_0 = 1, 2, 3$ , and  $4$ .

$c=0$  (multiplicative congruential method),  $m = 2^6 = 64$

Equation  $\rightarrow X_{i+1} = (a X_i + c) \bmod m$

When  $X_0 = 1, i = 1, X_2 = (13 * 1 + 0) \bmod 64 = 13 \bmod 64 = 13$

When  $X_0 = 1, i = 2, X_3 = (13 * 13 + 0) \bmod 64 = 169 \bmod 64 = 41$

When  $X_0 = 1, i = 3, X_4 = (13 * 41 + 0) \bmod 64 = 533 \bmod 64 = 21$

When  $X_0 = 1, i = 16, X_{17} = (13 * 5 + 0) \bmod 64 = 65 \bmod 64 = 1$

⋮

When  $X_0 = 2, i = 1, X_2 = (13 * 2 + 0) \bmod 64 = 26 \bmod 64 = 26$

When  $X_0 = 2, i = 2, X_3 = (13 * 26 + 0) \bmod 64 = 338 \bmod 64 = 18$

⋮

When  $X_0 = 2, i = 8, X_9 = (13 * 10 + 0) \bmod 64 = 130 \bmod 64 = 2$

Similarly, for  $X_0 = 3$  and  $4$  are calculated. The values are tabulated below in the table 4.1

Therefore

For  $X_0=1, 3$  maximal period is 16

For  $X_0=2$ , maximal period is 8

For  $X_0=4$ , maximal period is 4

Table 4.1: Generation of random numbers for different seeds.

i	$X_i$ $X_0 = 1$	$X_i$ $X_0 = 2$	$X_i$ $X_0 = 3$	$X_i$ $X_0 = 4$	Seed
0	1	2	3	4	
1	13	26	39	52	
2	41	18	59	36	
3	21	42	63	20	
4	17	34	51	4	
5	29	58	23		
6	57	50	43		
7	37	10	47		
8	33	2	35		
9	45		7		
10	9		27		
11	53		31		
12	49		19		
13	61		55		
14	25		11		
15	5		15		
16	1		3		

## Combined Linear Congruential Generators

As the computing power increases, the complexity of the system to simulate also increases. So, a longer period generator with good statistical properties is needed. One successful approach is to combine two or more multiplicative congruential generators.

Theorem: If  $W_{i,1}, W_{i,2}, \dots, W_{i,k}$  are any independent, discrete-valued random variables and  $W_{i,1}$  is uniformly distributed on integers 0 to  $m_1 - 1$ , then

$$W_i = \left( \sum_{j=1}^k W_{i,j} \right) \bmod m_1 - 1$$

is uniformly distributed on the integers 0 to  $m_1 - 1$ .

To see how this result can be used to form combined generators,

Let  $X_{i,1}, X_{i,2}, \dots, X_{i,k}$  be  $i^{\text{th}}$  output from  $k$  different multiplicative congruential generators, where the  $j^{\text{th}}$  generator has prime modulus  $m_j$  and multiplier  $a_j$  is chosen so that the period is  $m_j - 1$ . Then the  $j^{\text{th}}$  generator is producing  $X_{i,j}$  that are approximately uniformly distributed on 1 to  $m_j - 1$  and  $W_{i,j} = X_{i,j} - 1$  is approximately uniformly distributed on 0 to  $m_j - 2$ .

Therefore, combined generator of the form,

$$X_i = \left[ \sum_{j=1}^k (-1)^{j-1} X_{i,j} \right] \bmod m_1 - 1 \quad \text{Hence, } R_i = \begin{cases} \frac{X_i}{m_1}, & X_i > 0 \\ \frac{m_1 - 1}{m_1}, & X_i = 0 \end{cases}$$

The maximum possible period for a generator is

$$P = \frac{(m_1 - 1)(m_2 - 1) \dots (m_k - 1)}{2^{k-1}}$$

Note:  $(-1)^{j-1}$  coefficient implicitly performs the subtraction  $X_{i,1} - 1$

For 32-bit computers, L'Ecuyer [1988] suggests combining  $k = 2$  generators with  $m_1 = 2,147,483,563$ ,  $a_1 = 40,014$ ,  $m_2 = 2,147,483,399$  and  $a_2 = 40,692$ . This leads to the following algorithm:

Step 1: Select seeds

$X_{0,1}$  in the range  $[1 - 2,147,483,562]$  for the 1st generator

$X_{0,2}$  in the range  $[1 - 2,147,483,398]$  for the 2nd generator

Set  $i=0$

Step 2: For each individual generator, evaluate

$X_{i+1,1} = 40,014 X_{i,1} \bmod 2,147,483,563$

$X_{i+1,2} = 40,692 X_{i,2} \bmod 2,147,483,399$

Step 3:  $X_{i+1} = (X_{i+1,1} - X_{i+1,2}) \bmod 2,147,483,562$

Step 4: Return

$$R_{i+1} = \begin{cases} \frac{X_{i+1}}{2,147,483,563} & , X_{i+1} > 0 \\ \frac{2,147,483,562}{2,147,483,563} & , X_{i+1} = 0 \end{cases}$$

Step 5: Set  $i = i+1$ , go back to step 2

The combined generator has period:  $(m_1-1)(m_2-1)/2 \approx 2 \times 10^{18}$

Tests for Random Numbers

The two main properties of random numbers are uniformity and independence.

### Testing for Uniformity

The hypotheses are as follows

$H_0: R_i \sim U[0, 1]$

$H_1: R_i \not\sim U[0, 1]$

The null hypothesis  $H_0$ , reads that the numbers are distributed uniformly on the interval  $[0, 1]$ . Rejecting the null hypothesis means that the numbers are not uniformly distributed.

### Testing for Independence

The hypotheses are as follows

$H_0: R_i \sim \text{independently}$

$H_1: R_i \not\sim \text{independently}$

This null hypothesis,  $H_0$ , reads that the numbers are independent. Rejecting the null hypothesis means that the numbers are not independent. This does not imply that further testing of the generator for independence is unnecessary.

For each test, a level of significance  $\alpha$  must be stated.



$$\text{Level of significance } \alpha = \frac{\text{probability of rejecting the test}}{\text{probability of accepting the test}}$$

= P (reject H0 | H0 true)

Frequently,  $\alpha$  is set to 0.01 or 0.05.

There are five types of tests. The first is concerned for testing the uniformity whereas second through five with testing for independence.

**Frequency test** - Compares the distribution of set of numbers generated to a uniform distribution by using the Kolmogorov-Smirnov or the chi-square test.

**Autocorrelation test** - The correlation between numbers is tested and compares the sample correlation to the expected correlation of zero.

### Frequency Tests

The fundamental test performed to validate a new generator is the test for uniformity. The two different methods of testing are

Kolmogorov-Smirnov test

Chi-Square test

### Kolmogorov-Smirnov test

It compares the continuous cumulative distribution function (cdf) of the uniform distribution with the empirical cdf, of the N sample observations. The cdf of an empirical distribution is a step function with jumps at each observed value.

Notations used

$F(x) \rightarrow$  Continuous cdf

$SN(x) \rightarrow$  Empirical cdf

$N \rightarrow$  Total number of observations

$R_1, R_2 \dots R_N \rightarrow$  Samples from Random generator

$D \rightarrow$  Sample statistic

$D\alpha \rightarrow$  Critical value

By definition,

$$F(x) = x, \quad 0 \leq x \leq 1$$

$$S_N(x) = \frac{\text{number of } R_1, R_2 \dots R_n \text{ which are } \leq x}{N}$$

As  $N$  becomes larger,  $S_N(x) \approx F(x)$ .

Maximum deviation over the range of random variable is given by

$$D = \max |F(x) - S_N(x)|$$

The sampling distribution of  $D$  is known and is tabulated as a function of  $N$  in table A.8.

Procedure for testing uniformity using Kolmogorov-Smirnov test

Step 1– Rank the data from smallest to largest. Let  $R(i)$  denote the  $i^{\text{th}}$  smallest observation, so that

$$R(1) \leq R(2) \leq \dots \leq R(N)$$

Step 2 – Compute

$$D^+ = \max \{(i/N) - R(i)\} \quad 1 \leq i \leq N$$

$$D^- = \max \{R(i) - [(i-1)/N]\} \quad 1 \leq i \leq N$$

Step 3 – Compute  $D = \max(D^+, D^-)$

Step 4 – Determine the critical value  $D\alpha$ , from the table A.8 for the specified significance level  $\alpha$  and the given sample size  $N$ .

Step 5 If  $D > D\alpha$ , the null hypothesis that the data are sample from a uniform distribution is rejected.

If  $D \leq D\alpha$  then there is no difference detected between the true distribution of  $\{R_1, R_2 \dots R_N\}$  and the uniform distribution. So it is accepted.

Suppose 5 generated numbers are 0.44, 0.81, 0.14, 0.05, and 0.93. It is desired to perform a test for uniformity using Kolmogorov-Smirnov test with a level of significance  $\alpha = 0.05$ .  $D_{0.05, 5} = 0.565$ .

$N=5, i = 1, 2, 3, 4, 5$

**Step 1 -**

$R_i$	0.05	0.14	0.44	0.81	0.93
$i/N$	0.20	0.40	0.60	0.80	1.00
$i/N - R_i$	0.15	0.26	0.16	-	0.07
$R_i - [(i-1)/N]$	0.05	-	0.04	0.21	0.13

**Step 2 -**

Arrange  $R_i$  from smallest to largest

$D^+ = \max \{i/N - R_i\}$

$D^- = \max \{R_i - [(i-1)/N]\}$

**Step 3-**  $D = \max (D^+, D^-) = 0.26$

**Step 4-** For  $\alpha = 0.05, N = 5$

$$D\alpha = D_{0.05, 5} = 0.565$$

$$D < D\alpha \rightarrow 0.26 < 0.565$$

Therefore,  $H_0$  is not rejected, i.e. no difference between the distribution of generated numbers and the uniform distribution.

In the above calculation empirical cdf  $SN(x)$  is compared to uniform cdf  $F(x)$ . It is seen that  $D^+$  is the largest deviation of  $SN(x)$  above  $F(x)$  and  $D^-$  is the largest deviation of  $SN(x)$  below  $F(x)$ .

## Chi-square test

It uses the sample statistic

$$X_0^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Where  $O_i \rightarrow$  observed number in  $i^{\text{th}}$  class

$E_i \rightarrow$  expected number in  $i^{\text{th}}$  class

$n \rightarrow$  number of classes

For uniform distribution  $E_i$  is given by

$$E_i = \frac{N}{n}$$

Where  $N \rightarrow$  Total number of observations

It can be shown that the sampling distribution  $\chi^2_0$  is approximately the chi-square distribution with  $n-1$  degrees of freedom (i.e.,  $\chi^2_0 \leq \chi^2_{\alpha, n-1}$ )

Use Chi-square test with  $\alpha=0.05$  to test whether the data shown below are uniformly distributed.

0.34	0.90	0.25	0.89	0.87	0.44	0.12	0.21	0.46	0.67
0.83	0.76	0.79	0.64	0.70	0.81	0.94	0.74	0.22	0.74
0.96	0.99	0.77	0.67	0.56	0.41	0.52	0.73	0.99	0.02
0.47	0.30	0.17	0.82	0.56	0.05	0.45	0.31	0.78	0.05
0.79	0.71	0.23	0.19	0.82	0.93	0.65	0.37	0.39	0.42
0.99	0.17	0.99	0.46	0.05	0.66	0.10	0.42	0.18	0.49
0.37	0.51	0.54	0.01	0.81	0.28	0.69	0.34	0.75	0.49
0.72	0.43	0.56	0.97	0.30	0.94	0.96	0.58	0.73	0.05
0.06	0.39	0.84	0.24	0.40	0.64	0.40	0.19	0.79	0.62
0.18	0.26	0.97	0.88	0.64	0.47	0.60	0.11	0.29	0.78

Let  $n=10$ , the interval  $[0-1]$  divided in equal lengths,  $(0.01-0.10)$ ,  $(0.11-0.20)$ , ---,  $(0.91-1.0)$

$N = 100$

$E_i = N/n = 100/10 = 10$

The calculations are tabulated below in table 4.2

$$X_{0.05,9}^2 = 16.9 \text{ (check the table A.6 -using } \alpha, n-1 \text{)}$$

$$X_0^2 < X_{0.05,9}^2 = 3.4 < 16.9$$

Therefore, null hypothesis of uniform distribution is not rejected.

Table 4.2: Computation of chi-square test.

Interval	$O_i$	$E_i$	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	8	10	-2	4	0.4
2	8	10	-2	4	0.4
3	10	10	0	0	0.0
4	9	10	-1	1	0.1
5	12	10	2	4	0.4
6	8	10	-2	4	0.4
7	10	10	0	0	0.0
8	14	10	4	16	1.6
9	10	10	0	0	0.0
10	11	10	1	1	0.1
	100	100	0		3.4

Note:

In general, for any value choose 'n' such that  $E_i \geq 5$ .

Kolmogorov-Smirnov test is more powerful than chi-square test because it can be applied to small sample sizes, whereas chi square requires large sample, say  $N \geq 50$ .

## Tests for Autocorrelation

The tests for autocorrelation are concerned with dependence between numbers in a sequence. The equation for autocorrelation is

$$Z_0 = \frac{\hat{\rho}_{im}}{\sigma_{\hat{\rho}_{im}}}$$

$\hat{\rho}_{im}$  = Hypothesis of larger random number

$\sigma_{\hat{\rho}_{im}}$  = Standard deviation of  $\hat{\rho}_{im}$

Where,  $\hat{\rho}_{im}$  is given by

$$\hat{\rho}_{im} = \frac{1}{M+1} \left[ \sum_{k=0}^m R_{i+km} R_{i+(k+1)m} \right] - 0.25$$

and  $\sigma_{\hat{\rho}_{im}}$  is given by

$$\sigma_{\hat{\rho}_{im}} = \sqrt{\frac{13M+7}{12(M+1)}}$$

M → largest integer,  $i + (M + 1) m \leq N$

N → Total number of values in the sequence

i → starting position of observation

R<sub>i</sub> → Random number at the starting position of observation

m → lag, difference of observation.

If  $-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2}$ , then accept the null hypothesis of independence.

Critical value  $Z_{0.05/2} = Z_{0.025} = 1.96$

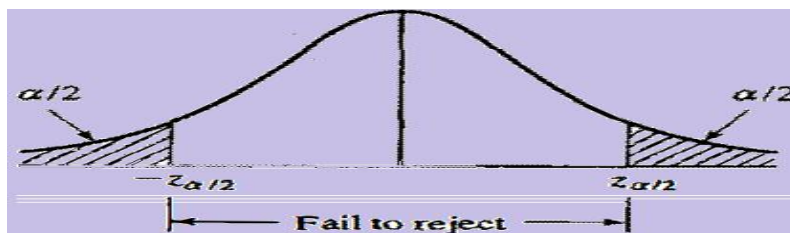


Figure 4.2: Null hypothesis of independence.

Test whether the 3<sup>rd</sup>, 8<sup>th</sup>, 13<sup>th</sup> and so on, numbers in the sequence are autocorrelated. The level of significance is 0.05.

0.12 0.01 0.23 0.28 0.89 0.31 0.64 0.28 0.83 0.93 0.99  
 0.15 0.33 0.35 0.91 0.41 0.60 0.27 0.75 0.88 0.68 0.49  
 0.05 0.43 0.95 0.58 0.19 0.36 0.69 0.87

i=1	i=2	R <sub>i</sub>							
0.12	0.01	0.23	0.28	0.89	0.31	0.64	0.28	0.83	0.93
0.99	0.15	0.33	0.35	0.91	0.41	0.60	0.27	0.75	0.88
0.68	0.49	0.05	0.43	0.95	0.58	0.19	0.36	0.69	0.87

N=30, m=5,

$i + (M + 1) m \leq N$

$3 + (M + 1)5 \leq 30$

$$(M + 1)5 = 30 - 3 = 27$$

$$M + 1 = 27/5 = 5.4 = 5$$

$$M = 4$$

$$\hat{p}_{35} = \frac{1}{4+1} [(0.23)(0.28) + (0.28)(0.33) + (0.33)(0.27) + (0.27)(0.05) + (0.05)(0.36)] - 0.25$$

$$= -0.1945$$

$$\sigma_{\hat{p}_{35}} = \frac{\sqrt{13(4) + 7}}{12(4 + 1)} = 0.1280$$

The test statistic,

$$Z_0 = -0.1945 / 0.1280 = -1.519$$

$$\text{Critical value } Z_{0.05/2} = Z_{0.025} = 1.96$$

$$-Z_{0.025} \leq Z_0 \leq Z_{0.025} = -1.96 \leq -1.519 \leq 1.96,$$

Therefore, null hypothesis of independence is not rejected.

### Exercises:

- Explain the properties of random numbers.
- Use the linear congruential method to generate a sequence of random numbers with  $X_0 = 23$ ,  $a = 13$ ,  $c = 56$ , and  $m = 65$ .
- Using the multiplicative congruential method, find the period of the generator for  $a = 15$ ,  $m = 2^5$ , and  $X_0 = 1, 2, 3$ , and  $4$ .
- Derive Kolmogorov-Smirnov test to validate a new generator for testing uniformity.
- Suppose 6 generated numbers are 0.34, 0.51, 0.16, 0.27, 0.45, and 0.53. It is desired to perform a test for uniformity using Kolmogorov-Smirnov test with a level of significance  $\alpha = 0.05$ .
- Test whether the 2<sup>nd</sup>, 8<sup>th</sup>, 14<sup>th</sup> and so on, numbers in the sequence are auto-correlated. The level of significance is 0.05.

0.14   0.17   0.23   0.28   0.89   0.31   0.64   0.28   0.83  
0.93

0.99   0.15   0.33   0.35   0.91   0.31   0.40   0.77   0.74  
0.38

0.65   0.39   0.55   0.46   0.45   0.68   0.19   0.36   0.69  
0.87

- Derive auto-correlation for the random number given below  
0.13   0.71   0.63   0.28   0.85   0.81   0.60   0.88   0.63  
0.73



0.95 0.75 0.83 0.65 0.91 0.41 0.60 0.27 0.75 0.85  
0.68  
0.47 0.55 0.33 0.96 0.38 0.69 0.76 0.49 0.87

## **Chapter-5**

### **Random Variate Generation**

#### **Variates:**

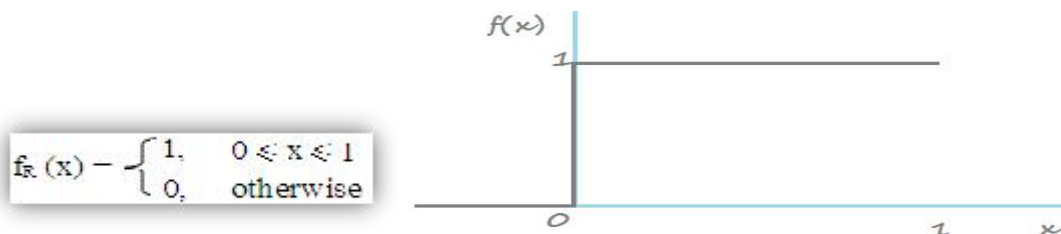
We know that a system can be studied through its simulation modeling. One of the fundamental concepts in simulation modeling is random variate generation and it is necessary to make the inference about the stochastic behavior of a random variable. There are different types of random variate generation techniques available.



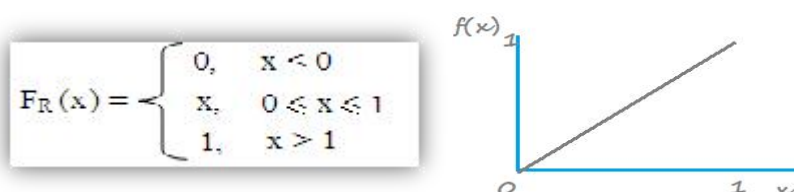
Figure 5.1: Impurities in random numbers.

All the techniques for generating random variates assumes that they are uniform (0, 1) random numbers  $R_1, R_2 \dots$  is readily available, where each  $R_i$  has

probability density function (pdf)



and cumulative distribution function (cdf)



## Inverse Transform Technique

This technique will be explained in detail for exponential distribution. Given pdf/pmf of a distribution, find the cdf  $F(X)$ , where  $X$  is a random variate. Set  $F(X) = R$ , where  $R$  is a random number. Then  $X = F^{-1}(R)$ ,  $F^{-1}$  is the solution of equation  $R = F(X)$  in terms of  $R$ , not  $1/f$ .

## Exponential Distribution

The exponential distribution has pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

cdf

$$F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Where  $\lambda \rightarrow$  Mean number of occurrences per unit time

Example

If interarrival times  $X_1, X_2, X_3 \dots$  had an exponential distribution with rate  $\lambda$ , then  $\lambda$  can be interpreted as mean number of arrivals per unit time. For any  $i$ ,

$$E(X_i) = \frac{1}{\lambda}$$

Where  $1/\lambda \rightarrow$  Mean interarrival time

Therefore, a procedure can be developed for generating values  $X_1, X_2, X_3 \dots$  having an exponential distribution.

Procedure for Inverse Transform Technique, by exponential distribution

Step1 -Compute the cdf of desired random variable  $X$ . For exponential distribution, cdf is

$$F(x) = 1 - e^{-\lambda x}, x \geq 0$$

Step2 -Set  $F(X) = R$  on range of  $X$ . For exponential distribution,  $1 - e^{-\lambda x} = R, x \geq 0$ . Since

X is a random variable;  $1 - e^{-\lambda x}$  is also a random variable, uniformly distributed over the interval (0, 1).

Step3 -Solve the equation  $F(X) = R$ , X in terms of R.

For exponential distribution, the solution is as follows

$$1 - e^{-\lambda x} = R$$

$$e^{-\lambda x} = 1 - R$$

$$-\lambda X = \ln (1 - R)$$

$$X = -\frac{1}{\lambda} \ln (1 - R) \quad (5.1)$$

Equation (5.1) is called a random variate generator for exponential distribution and can be written as  $X = F^{-1}(R)$ .

Step4 -Generate uniform random numbers  $R_1, R_2, R_3 \dots$  and compute desired random variate by

$$X_i = F^{-1}(R_i)$$

For exponential distribution, using (5.1)

$$F^{-1}(R) = -\frac{1}{\lambda} \ln (1 - R)$$

$$X_i = -\frac{1}{\lambda} \ln (1 - R_i)$$

for  $i = 1, 2 \dots$

$$X_i = -\frac{1}{\lambda} \ln R_i \quad [\text{Replace } 1 - R_i \text{ by } R_i] \quad (5.2)$$

Note: It is justified that both  $1 - R_i$  and  $R_i$  are uniformly distributed on  $(0, 1)$ .

## Uniform Distribution

Let  $X \rightarrow$  Random variable, uniformly distributed on interval  $[a, b]$ .

$R \rightarrow$  Random number

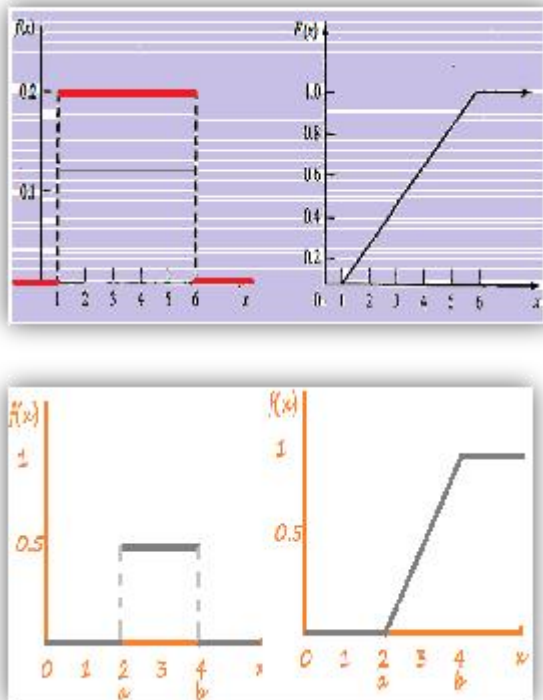


Figure 5.2: Uniform distribution.

The pdf of  $X$  is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

Procedure for deriving the equation for  $X$

Step1 – The cdf is

$$\text{Continuous probability} = \int_{-\infty}^x f(x) dx$$

It can re-written as 
$$= \int_{-\infty}^a f(x) dx + \int_a^x f(x) dx$$

In uniform distribution 
$$\int_{-\infty}^a f(x) dx = 0$$

Therefore, continuous density probability 
$$= \int_a^x f(x) dx$$
  

$$= \int_a^x \frac{1}{b-a} dx$$

$$= \frac{1}{b-a} \left( x \right)_a^x$$

$$= \frac{x-a}{b-a}$$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

Step2 – Set

$$F(X) = \frac{X-a}{b-a} = R$$

Step3 – Solve for X in terms of R, Then

$$X - a = R (b - a)$$

$$\text{Therefore } X = a + (b - a) R$$

## Triangular distribution

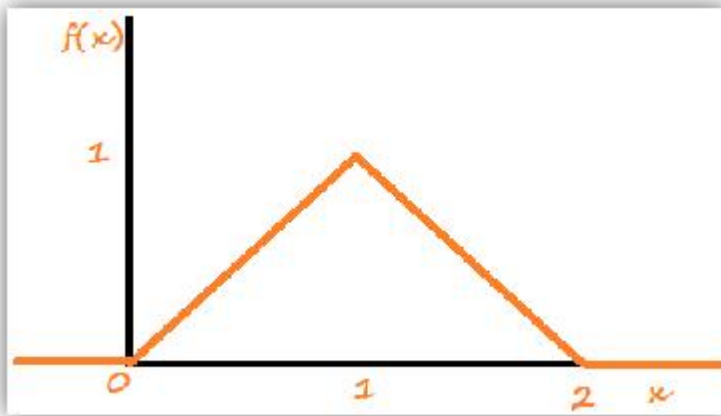


Figure 5.3: Triangular distribution.

In triangle  $x = 0$  to  $1$

$$f(x) = x$$

and  $1$  to  $2$

if  $x=1$ ,  $f(x) = 2 - 1 = 1$  (from figure 5.3),

$$x = 1.5, f(x) = 2 - 1.5 = 0.5$$

$$x = 2, f(x) = 2 - 2 = 0$$

hence, we can write mathematical equation for the  $f(x)$  as

$$f(x) = 2 - x$$

The pdf,

$$f(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ 2 - x, & 1 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

This is depicted in figure 5.1 with endpoints  $(0, 2)$  and mode at  $1$ .

The cdf,

$$\text{Continuous probability} = \int_{-\infty}^X f(x) dx$$

$$\text{it can re-written as} = \int_{-\infty}^0 f(x) dx + \int_0^1 f(x) dx + \int_1^2 f(x) dx + \int_2^X f(x) dx$$

$$= 0 + \int_0^1 x dx + \int_1^2 (2-x) dx + 1$$

The triangle is between 1 to 2

$$\int_0^1 x dx + \int_1^2 (2-x) dx$$

$$= \frac{x^2}{2} + 1 - \frac{(2-x)^2}{2}$$

For  $0 \leq X \leq 1$ ,

$$R = \frac{X^2}{2} \quad (5.4)$$

and for  $1 \leq X \leq 2$ ,



$$R = 1 - \frac{(2-X)^2}{2} \quad (5.5)$$

By equation (5.4),  $0 \leq X \leq 1$  implies  $0 \leq R \leq 1/2$  then

$$X = \sqrt{2R}$$

By equation (5.5),  $1 \leq X \leq 2$  implies  $1/2 \leq R \leq 1$  then

$$X = 2 - \sqrt{2(1-R)}$$

Thus, X is generated by

$$X = \begin{cases} \sqrt{2R}, & 0 \leq R \leq \frac{1}{2} \\ 2 - \sqrt{2(1-R)}, & \frac{1}{2} < R \leq 1 \end{cases}$$

## Weibull Distribution

Weibull distribution is an example of continuous probability distribution. It is more preferred in simulation modeling because of its flexibility property, means it can simulate different types of distributions such as: exponential and normal distributions.

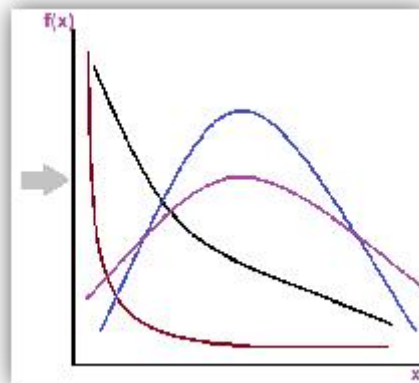


Figure 5.4: Weibull Distribution

The pdf of contamination is shown below

$$f(x) = \begin{cases} \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right) & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The cdf of exponential is given as below

$$F(x) = 1 - \exp\left(-\frac{x}{\alpha}\right)$$

Hence the cdf for equation

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)$$

is given as

$$R = 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)$$

$$\exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) = 1-R$$

$$-\left(\frac{x}{\alpha}\right)^\beta = \ln(1-R)$$

$$\left(\frac{x}{\alpha}\right)^\beta = -\ln(1-R)$$

$$\left(\frac{x}{\alpha}\right) = \left([-\ln(1-R)]\right)^{1/\beta}$$

$$x = \alpha \left([-\ln(1-R)]\right)^{1/\beta}$$

If the modeler fails find a perfect match with the theoretical distribution that provides a good model for the input data, then it may be necessary to use the empirical distribution of the data. One possibility is to simply resample the observed data itself. This is known as using the empirical distribution, and it gives very good result after experimenting with large data. On the other hand, if the data are drawn from what is believed to be a continuous-valued input process, then it makes sense to interpolate between the observed data points to fill in the gaps.

Five observations of fire-crew response times (in minutes) to incoming alarms have been collected to be used in a simulation investigating possible alternative staffing and crew-scheduling policies. The data are

2.76 1.83 0.80 1.45 1.24

Before collecting more data, it is desired to develop a preliminary simulation model that uses a response-time distribution based on these five observations. Thus, a method for generating random variates from the response-time distribution is needed. Initially, it will be assumed that response times  $X$  have a range  $0 \leq x \leq c$

where  $c$  is unknown, but will be estimated by  $c = \max \{ X_i : i = 1, \dots, n \} = 2.76$ , where  $\{X, i = 1, \dots, n\}$  are the raw data and  $n = 5$  is the number of observations. Arrange the data from smallest to largest and let  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  denote these sorted values. The smallest possible value is believed to be 0. Assign the probability  $1/n = 1/5$  to each interval.

$i$	Interval $x_{(i-1)} < x \leq x_{(i)}$	Probability $1/n$	Cumulative Probability, $i/n$	Slope $a_i$
1	$0.0 < x \leq 0.80$	0.2	0.2	4.00
2	$0.80 < x \leq 1.24$	0.2	0.4	2.20
3	$1.24 < x \leq 1.45$	0.2	0.6	1.05
4	$1.45 < x \leq 1.83$	0.2	0.8	1.90
5	$1.83 < x \leq 2.76$	0.2	1.0	4.65

If the graph is drawn for the interval, the slope at the  $i^{\text{th}}$  interval is given by

For example, if a random number  $R1 = 0.51$  is generated, then  $R1$  is seen to lie in the 3<sup>rd</sup> interval (between  $2/5 = 0.40$  and  $3/5 = 0.60$ );

$$a_i = \frac{x(i) - x(i-1)}{i/n - (i-1)/n}$$

$$X = a_i (R - (i-1)/n) + x(i-1)$$

$$X = 1.24 + 1.05 (0.51 - 0.4)$$

## Discrete Distributions

The discrete distributions can be generated by using inverse transform technique either numerically (table-lookup procedure) or algebraically (formula). It includes empirical distribution and two standard discrete distributions – (discrete) uniform and geometric distributions.

### Example 5.2 (Empirical Discrete distribution)

At the end of the day, number of shipments on loading dock of ABC Company is 0, 1 or 2 with relative frequency of occurrence of 0.50, 0.30 and 0.20 respectively. The internal consultants were asked to develop a model to improve efficiency of loading and hauling operations, as a part they are required to generate values  $X$ , to represent number of shipments on loading dock at end of each day. The discrete random variable with distribution is given in the table 5.1 and 5.2.

Table 5.1: Distributions of number of shipments  $X$

$x$	$p(x)$	$F(x)$
0	0.50	0.50
1	0.30	0.80
2	0.20	1.00

Table 5.2: Generating the discrete variate  $X$

$i$	Input ( $r_i$ )	Output ( $x_i$ )
1	0.50	0
2	0.80	1
3	1.00	2

pmf is given by

$$p(0) = P(X=0) = 0.50$$

$$p(1) = P(X=1) = 0.30$$

$$p(2) = P(X=2) = 0.20$$

The cdf of discrete random variable always consists of horizontal line segments with jumps of size  $p(x)$  at points  $x$ , which the random variable can assume. There is a jump of size  $p(0) = 0.5$  at  $x = 0$ ,  $p(1) = 0.3$  at  $x = 1$  and  $p(2) = 0.2$  at  $x = 2$ .

cdf,

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.5, & 0 \leq x < 1 \\ 0.8, & 1 \leq x < 2 \\ 1, & 2 \leq x \end{cases}$$

Suppose  $R_1 = 0.73$  is to be generated, then

### 1. Graphically

First locate  $R_1 = 0.73$  on vertical axis, draw a horizontal line segment until it hits a 'jump' in cdf and then drop a perpendicular to horizontal axis to get the generated variate.

### 2. Table – lookup procedure

First find the interval in which  $R_1$  lies, In general for  $R = R_1$ ,

if

$$F(x_{i-1}) = r_{i-1} < R \leq r_i = F(x_i)$$

then

Set  $X_1 = x_i$

Here  $r_0 = 0$ ,  $x_0 = -\infty$ , while  $x_1, x_2, \dots, x_n$  are possible values of random variable and

$$r_k = p(x_1) + p(x_2) + \dots + p(x_k), \quad k = 1, 2, \dots, n$$

For this example

$$n = 3, \quad x_1 = 0, \quad x_2 = 1, \quad x_3 = 2, \quad \text{hence } r_1 = 0.5, \quad r_2 = 0.8, \quad r_3 = 1.0$$

Since  $r_1 = 0.5 < R_1 = 0.73 \leq r_2 = 0.8$

Set  $X_1 = X_2 = 1$

Therefore, generation scheme is summarized as

$$X = \begin{cases} 0, & R \leq 0.5 \\ 1, & 0.5 < R \leq 0.8 \\ 2, & 0.8 < R \leq 1.0 \end{cases}$$

Example 5.3 (Discrete Uniform Distributions)

Consider discrete uniform distribution on  $(1, 2, \dots, k)$  with

pmf  $\Rightarrow p(x) = 1/k, x = 1, 2, \dots, k$

cdf,

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{k}, & 1 \leq x < 2 \\ \frac{2}{k}, & 2 \leq x < 3 \\ \vdots & \vdots \\ \frac{k-1}{k}, & k-1 \leq x < k \\ 1, & k \leq x \end{cases}$$

Let us consider

$x_i = i$

$r_i = p(1) + p(2) + \dots + p(x_i) = F(x_i)$

$$F(x_i) = \frac{i}{k}$$

$$F(x_{i-1}) = \frac{x_i - 1}{k} = \frac{i - 1}{k}$$

By using inequality,  $F(x_{i-1}) = r_{i-1} < R \leq r_i = F(x_i)$ , generated random number  $R$  satisfies

$$r_{i-1} = \frac{i-1}{k} < R \leq r_i = \frac{i}{k} \quad (5.6)$$

Then  $X$  is generated by setting  $X = i$ . Now the above inequality (5.6) can be solved for  $i$

$$i - 1 < Rk \leq i$$

$$Rk \leq i < Rk + 1$$

This yields to a formula for generating  $X$ , i.e.

$$X = \lceil Rk \rceil$$

(rounds up the values of  $Rk$ )

$$\text{For example, } X_1 = \lceil 7.8 \rceil = 8$$

#### Example 5.4

Consider discrete distribution with pmf given by

$$p(x) = \frac{2x}{k(k+1)}, \quad x = 1, 2, \dots, k$$

For integer values of  $x$  in the range  $\{1, 2, \dots, k\}$

cdf,

$$F(x) = \sum_{i=1}^x \frac{2i}{k(k+1)}$$

$$= \frac{2}{k(k+1)} \sum_{i=1}^x i$$



$$= \frac{2}{k(k+1)} \frac{x(x+1)}{2}$$

$$= \frac{x(x+1)}{k(k+1)}$$

Generate R and use inequality  $F(x_{i-1}) = r_{i-1} < R \leq r_i = F(x_i)$ , such that

$$F(x-1) = \frac{(x-1)x}{k(k+1)} < R \leq \frac{x(x+1)}{k(k+1)} = F(x)$$

$$\Rightarrow \underbrace{(x-1)x < k(k+1)R \leq x(x+1)}$$

To get 'x' in terms of R consider first inequality

$$(x-1)x = k(k+1)R$$

$$\rightarrow x^2 - x - k(k+1)R = 0$$

The above equation is in form of quadratic equation. So, the solution is obtained by using quadratic formula,

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

With  $a=1$ ,  $b=-1$ ,  $c=-k(k+1)R$

$$x = \frac{-(-1) \pm \sqrt{1 - 4(1)(-k(k+1)R)}}{2(1)}$$

$$x = \frac{1 + \sqrt{1 + 4(k^2 + k)R}}{2} \quad [\text{Considering only positive roots}]$$

By rounding up, the solution is  $X = \lceil x \rceil$

$$\therefore X = \left\lceil \frac{1 + \sqrt{1 + 4(k^2 + k)R}}{2} \right\rceil$$

Consider the geometric distribution with pmf

$$p(x) = p(1-p)^x, \quad x = 0, 1, 2, \text{ where } 0 < p < 1$$

cdf,

$$F(x) = \sum_{i=0}^x p(1-p)^i$$

$$= p \sum_{i=0}^x (1-p)^i$$

$$= p [1 + (1-p) + (1-p)^2 + (1-p)^3 + \dots + (1-p)^x]$$

$$= p \frac{1 - (1-p)^{x+1}}{1 - (1-p)} \quad \left( \sum_{k=0}^n (a-b)^k = \frac{a - (1-b)^{n+1}}{a - (1-b)} \right)$$

$$= 1 - (1-p)^{x+1}$$

Generate  $R$  and use inequality  $F(x_{i-1}) = r_{i-1} < R \leq r_i = F(x_i)$ , such that

$$F(x-1) = 1 - (1-p)^x < R \leq 1 - (1-p)^{x+1} = F(x), \quad 0 < R < 1$$

$$(1-p)x + 1 \leq 1 - R < (1-p)x$$

$$(x+1) \ln(1-p) \leq \ln(1-R) < x \ln(1-p)$$

÷ by  $\ln(1-p)$

$$(x+1) \leq \frac{\ln(1-R)}{\ln(1-p)} < x$$

$1-p < 1$  implies  $\ln(1-p) < 0$ , so that

$$\frac{\ln(1-R)}{\ln(1-p)} - 1 \leq x < \frac{\ln(1-R)}{\ln(1-p)}$$

Thus,  $X = x$  for that integer value of  $x$  satisfying inequality. By using round up function

$$X = \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil$$

For a geometric variate  $X$ , assume values  $\{q, q+1, q+2 \dots\}$  with pmf

$$p(x) = p(1-p)^{x-q}, \quad (x = q, q+1, \dots)$$

Such a variate,  $X$  can be generated as

$$X = q + \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil$$

Note- Commonly  $q = 1$

Generate 3 values from a geometric distribution on the range  $\{X \geq 1\}$  with mean 2. Such a geometric distribution has pmf  $p(x) = p(1-p)^{x-1}$ , where  $x = 1, 2, \dots$  with mean  $1/p = 2$ .

[Random numbers  $R_1 = 0.932$ ,  $R_2 = 0.105$ ,  $R_3 = 0.687$  from table A.2]

$$p = 1/2$$

$$q = 1$$

$$p(x) = p (1 - p)^{x-1}$$

$$X = q + \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil$$

$$X_1 = 1 + \left\lceil \frac{\ln(1-0.932)}{\ln(1-0.5)} - 1 \right\rceil = \lceil 3.8 \rceil = 4$$

$$X_2 = 1 + \left\lceil \frac{\ln(1-0.105)}{\ln(1-0.5)} - 1 \right\rceil = \lceil 0.166 \rceil = 1$$

$$X_3 = 1 + \left\lceil \frac{\ln(1-0.687)}{\ln(1-0.5)} - 1 \right\rceil = \lceil 1.67 \rceil = 2$$

### Acceptance – Rejection Technique



Consider a method for generating random variates  $X$ , uniformly distributed between  $1/4$  and  $1$ . The procedure is as follows:

Step1 – Generate a random number  $R$

Step2 – a) If  $R \geq \frac{1}{4}$ , accept  $X = R$  and then go to step3

b) If  $R < \frac{1}{4}$ , reject  $R$ , and return to step1

Step3 – If another uniform random variate on  $[\frac{1}{4}, 1]$  is needed, repeat from beginning at step1, otherwise stop.

Step1 generates a new random number  $R$  at each time of execution. Step 2a is an Acceptance and step 2b is a rejection. When the condition is finally satisfied, desired random variate  $X[\frac{1}{4}, 1]$  can be computed ( $X = R$ ). This procedure is proved to be correct by recognizing the accepted values of  $R$ , the conditioned values.  $R$  conditioned on event  $\{R \geq \frac{1}{4}\}$ , have the desired distribution.

To show this

$$\text{if } \frac{1}{4} \leq a < b \leq 1$$

then

$$P\left[\frac{a < R \leq b}{\frac{1}{4} \leq R \leq 1}\right] = \frac{P(a < R \leq b)}{P\left(\frac{1}{4} \leq R \leq 1\right)} = \frac{b-a}{\frac{3}{4}} \quad [\text{Probability for uniform distribution from } b \text{ to } a]$$

This is the probability distribution of  $R$ , given that  $R$  is between  $\frac{1}{4}$  and 1 is the desired distribution.

$$\therefore \text{if } \frac{1}{4} \leq R \leq 1, \text{ set } X = R$$

The efficiency of an acceptance rejection technique depends on rejections i.e. minimum the number of rejections, maximum the efficiency.

This technique is illustrated by Poisson distribution for  $r^{\text{th}}$  generation of random variates.

## Poisson Distribution

Poisson defined the equation for the  $k$  events in  $t$  time as

$$P(N(t)=k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$



Figure 5.5a: Poisson Distribution.

Where  $N \rightarrow$  Poisson random variate

$N$  can be interpreted as number of arrivals from Poisson arrival process in one unit of time.

If there are no event in time  $t$

$$P(N(t)=0) = \frac{e^{-\lambda t} (\lambda t)^0}{0!} = e^{-\lambda t} = R \text{ (Random number for No event)}$$

Or we can define  $k$  event in time  $t$  and no event between  $t$  to  $r$  as shown in figure 5.5b



Figure 5.5b: Poisson Distribution.

$$P(N(t+r)=k/N(t)=k)$$

$$P(N(t+r)-N(t)=0) = e^{-\lambda t}$$

Equation for random number for the event A1 for the figure 5.5c is

$$R = A1 e^{-\lambda t}$$

$$A1(-\lambda t) = \ln R$$

$$A1 = \frac{-\ln R}{\lambda t} > 1$$


Figure 5.5c: Poisson Distribution.

Equation for random number for the event A1 for the figure 5.5d

$$R = A1 e^{-\lambda t}$$

$$A1(-\lambda t) = \ln R$$


$$A1 = \frac{-\ln R}{\lambda t} < 1$$


Figure 5.5d: Poisson Distribution.

The above two equations can be re-written for the k events which are happens before and after  $t = 1$ .

$$\sum_{i=1}^n \frac{-\ln R}{\lambda t} < 1 < \sum_{i=1}^{n+1} \frac{-\ln R}{\lambda t}$$

The probability of claims will be defined as

Procedure for generating a Poisson random variate N

Step1 – Set  $k = 0$  (where k is number of events),  $P = 1$

Step2 – Generate random number  $R_{n+1}$  and replace P by  $P \cdot R_{n+1}$

Step3 – If  $P < e^{-\alpha}$ , then accept  $N = n$

Else reject the current  $n$ , increase  $n$  by one and return to step2.

Note

If  $P \geq e^{-\alpha}$  in step3, then  $n$  is rejected and generation process must proceed through at least one more trial.

If  $N = n$ , then  $n+1$  random number are required, so the average number is given by  $E(N + 1) = \alpha + 1$ .

Generate three Poisson variates with mean  $\alpha = 0.2$  Random Numbers are 0.4357, 0.4146, 0.8353, 0.9952, 0.8004

$$e^{-\alpha} = e^{-0.2} = 0.8187$$

Step1 – Set  $n = 0$ ,  $P = 1$

Step2 –  $R1 = 0.4357$ ,  $P = 1 * 0.4357 = 0.4357$

Step3 – Since  $P < e^{-\alpha}$  i.e.  $0.4357 < 0.8187$ , Accept  $N = 0$ .

Step1 – Set  $n = 0$ ,  $P = 1$

Step2 –  $R1 = 0.4146$ ,  $P = 1 * 0.4146 = 0.4146$

Step3 – Since  $P < e^{-\alpha}$  i.e.  $0.4146 < 0.8187$ , Accept  $N = 0$ .

Step1 – Set  $n = 0$ ,  $P = 1$

Step2 –  $R1 = 0.8353$ ,  $P = 0.8353$

Step3 – Since  $P \geq e^{-\alpha}$ , reject  $n = 0$ , return to step2 with  $n = 1$ .

Step2 –  $R2 = 0.9952$ ,  $P = P.R2 = 0.8353 * 0.9952 = 0.8312$

Step3 – Since  $P \geq e^{-\alpha}$ , reject  $n = 1$ , return to step2 with  $n = 2$ .

Step2 –  $R3 = 0.8004$ ,  $P = P.R3 = 0.8312 * 0.8004 = 0.6654$

Step3 – Since  $P < e^{-\alpha}$ , accept  $N = n = 2$ .

The calculations are summarized below

n	$R_{n+1}$	p	Accept/Reject	Result
---	-----------	---	---------------	--------



0	0.4357	0.4357	$P < e^{-\alpha}$ (accept)	$N = 0$
0	0.4146	0.4146	$P < e^{-\alpha}$ (accept)	$N = 0$
0	0.8353	0.8353	$P \geq e^{-\alpha}$ (reject)	—
1	0.9952	0.8312	$P \geq e^{-\alpha}$ (reject)	—
2	0.8004	0.6654	$P < e^{-\alpha}$ (accept)	$N = 2$

To generate 3 Poisson variates, it took 5 random numbers, R. If 1000 Poisson variates with  $\alpha = 0.2$ , has to be generated then  $1000(\alpha + 1) = 1200$  random number are required.

### Exercises:

Generate 200 variates  $X_i$  with distribution exponent distribution ( $i = 1$ )

Derive mean and derivation equation of parameter estimation.

Draw Q-Q plot for the following data

j	Value	j	Value	j	Value	j	Value
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.9	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

Five observations of fire crew response times (in minutes) to incoming alarms are collected to be used in a simulation investigating possible alternative staffing and crew-scheduling policies. The data are: 1.76, 1.49, 0.73, 2.48, 2.29. Generate empirical fire crew response time.

Generate uniformly distributed random numbers between 1/4 and 1, by using Acceptance-Rejection technique

Test whether the 2nd, 8th, 14th and so on, numbers in the sequence are auto-correlated. The level of significance is 0.05.

0.14   0.17   0.23   0.28   0.89   0.31   0.64   0.28   0.83   0.93  
 0.99   0.15   0.33   0.35   0.91   0.31   0.40   0.77   0.74   0.38  
 0.65   0.39   0.55   0.46   0.45   0.68   0.19   0.36   0.69   0.87

Derive auto-correlation for the random number given below

0.13   0.71   0.63   0.28   0.85   0.81   0.60   0.88   0.63   0.73  
 0.95   0.75   0.83   0.65   0.91   0.41   0.60   0.27   0.75   0.85  
 0.68   0.47   0.55   0.33   0.96   0.38   0.69   0.76   0.49   0.87

## Chapter-6

### INPUT MODELLING

Input data modelling is the one that provides the driving force for a simulation model. Before using any data as input to a system, it

undergoes many processes and these processes with the input data is called input modelling.

There are four steps in the development of a useful model of input data:

Collect data from the real system of interest. This often requires a substantial time and resource commitment.

Identify a probability distribution to represent the input process. When data are available, this step typically begins with the development of a frequency distribution of the data.

Choose parameters that determine a specific instance of the distribution family.

Evaluate the chosen distribution and the associated parameters for goodness of fit. After performing many processes on data there is a possibility that, data may change its behavior and structure. To test the healthy condition goodness of fit is performed on data. Goodness of fit may be evaluated informally, via graphical methods, or formally, via statistical tests. The chi-square and the Kolmogorov-Smirnov tests are standard goodness-of-fit tests. If not satisfied that the chosen distribution is a good approximation of the data, then the analyst returns to the second step, chooses a different family of distributions, and repeats the procedure. If several iterations of this procedure fail to yield a fit between an assumed distributional form and the collected data, the empirical form of the distribution may be used.

## **Data Collection**

In this book many exercises are given at the end of each module, solution can be easily obtained by using the available (given) data in the question. Where as in real time the data collection is one of the biggest challenges. Even data are available, they have rarely been

recorded in a form that is directly useful for simulation modeling. Even when the model structure is valid, if the input data are inaccurately collected, inappropriately analyzed, the simulated output will be misleading and possibly damaging model.

Many lessons can be learned from an actual experience at data collection. The following suggestions might enhance and facilitate data collection,

A useful expenditure of time is in planning. This could begin by a practice or pre-observing session. Try to collect data while pre-observing. Devise forms for this purpose. It is very likely that these forms will have to be modified several times before the actual data collection begins.

Try to analyze the data as they are being collected. Figure out whether the data being collected are adequate to provide the distributions needed as input to the simulation. Find out whether any data being collected are useless to the simulation. There is no need to collect superfluous data.

Check data for homogeneity in successive time periods and during the same time period on successive days. For example, check for homogeneity of data from 2:00 P.M. to 3:00 P.M. and 3:00 P.M. to 4:00 P.M., and check to see whether the data are homogeneous for 2:00 P.M. to 3:00P.M. on Thursday and Friday.

Consider the possibility that a sequence of observations that appear to be independent actually has autocorrelation. Autocorrelation can exist in successive time periods or for successive customers. For example, the service time for the  $i^{\text{th}}$  customer could be related to the service time for the  $(i + n)^{\text{th}}$  customer.

## **Identifying the Distribution with Data**

Selecting the families of input distributions when data are available is also very important.

Histogram is useful method for identifying the shape of a distribution. A histogram is constructed as follows:

1. Divide the range of the data into intervals.
2. Label the horizontal axis to conform to the intervals selected.
3. Find the frequency of occurrences within each interval.
4. Label the vertical axis so that the total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

If the histogram is associated with discrete data, it should look like a probability mass function.

The number of cars arriving at the workshop in a 10-minutes period between 8:00 A.M. and 8:10 A.M. was monitored for five workdays over a 20-week period. The first entry in the figure 6.1 indicates that 12 days zero vehicles arrived, 10 days one vehicle arrived, and so on.

Days	Cars arrival in a particular period
12	0
10	1
14	2
17	3
10	4
08	5
07	6
05	7
05	8
03	9
03	10
01	11

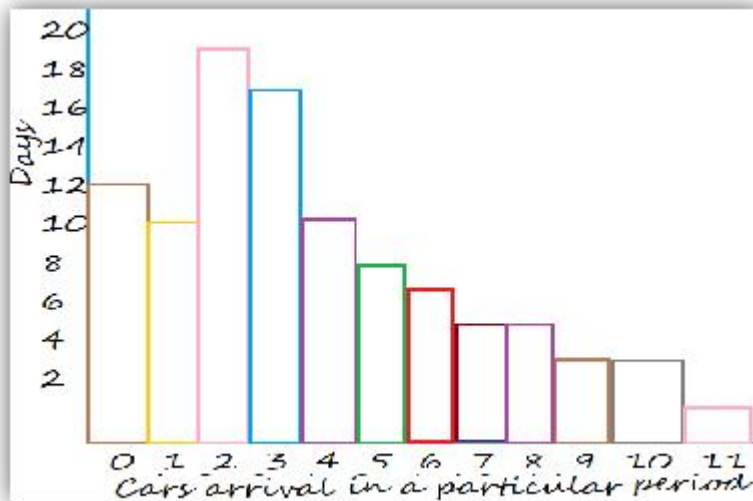


Figure 6.1: Histogram for the car's arrival between 8:00 AM to 8:10AM.

Life tests are performed on random sample of electronic chips at 1.5 times the nominal voltage and their lifetime in days are recorded

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.96
0.590	1.928	0.300	0.002	0.543

7.004    31.764    1.005    1.147    0.219  
 3.217    14.382    1.008    2.336    4.562

Lifetime is usually a continuous variable. Since the data is large from 0.002 day to 144.695 days, use intervals of width three results.

It is shown in figure 6.2.

Chip life (days)	Frequency
$0 \leq x_j < 3$	24
$3 \leq x_j < 6$	9
$6 \leq x_j < 9$	5
$9 \leq x_j < 12$	1
$12 \leq x_j < 15$	1
$15 \leq x_j < 18$	2
$18 \leq x_j < 21$	0
$21 \leq x_j < 24$	1
$24 \leq x_j < 27$	1
$27 \leq x_j < 30$	0
$30 \leq x_j < 33$	1
$33 \leq x_j < 36$	1
$\vdots$	$\vdots$
$42 \leq x_j < 45$	1
$\vdots$	$\vdots$
$57 \leq x_j < 60$	1
$\vdots$	$\vdots$
$78 \leq x_j < 81$	1
$\vdots$	$\vdots$
$144 \leq x_j < 147$	1

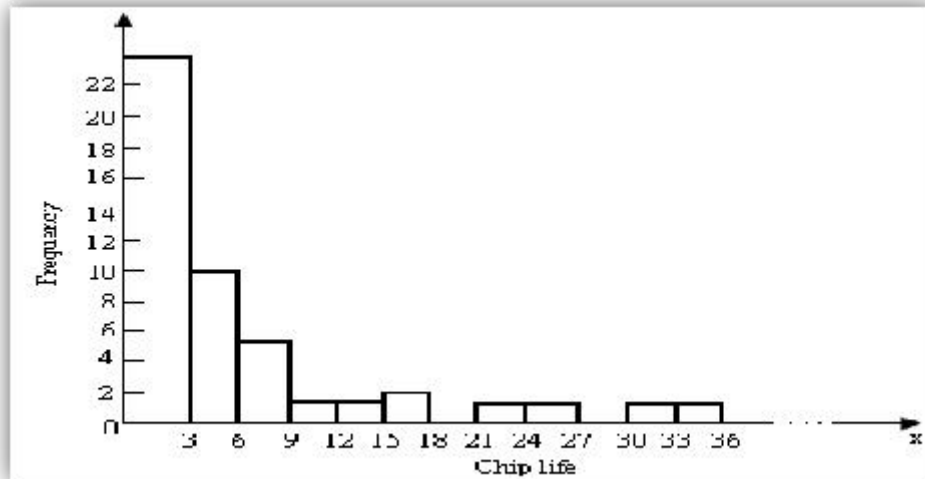


Figure 6.2: Histogram for chip life

## Selecting the Family of Distributions

The purpose of preparing a histogram is to infer a known pdf or pmf. A family of distributions is selected, on the basis of content being investigated with the shape of the histogram. The exponential, normal & Poisson distribution are frequently used, easy to analyze. Whereas gamma & Weibull distribution provide a wide array of shapes, difficult to analyze.

Some of the examples to select the distributions

**Binomial** - Models the number of successes in  $n$  independent trials with probability  $p$ .

Example: Number of defective chips found in  $n$  chips.

**Negative binomial** - Models the number of trials required to achieve 'k' successes.

Example: Number of chips that must be inspected to find 4 defective chips.

**Poisson** - Models number of independent events that occur in fixed amount of time.



Example: Number of customers arriving to a restaurant during 1 hour.

**Normal** - Models the process as sum of number of component processes

Example: Time to assemble a product is sum of times required for each assembly operation.

**Lognormal** -Models the distribution of process as product of number of component processes

Example: Rate of return on an investment.

**Exponential** -Models the time between independent events or process time which is memory less.

Example: Times between arrivals of large number of customers.

**Gamma** - Models nonnegative random variables, the gamma can be shifted away from 0 by adding a constant.

**Weibull** - Models the time to failure for components

Example: Time to failure for a disk drive. Exponential is the special case of Weibull

**Discrete or continuous uniform** - Models complete uncertainty, since all outcomes are equally likely. This distribution is often used when there are no data.

**Empirical** -It is used when no theoretical distributions are appropriate. Resample from the actual data collected.

### **Quantile - Quantile plots**

A quantile-quantile (q-q) plot is a useful tool for evaluating distribution fit, whereas histogram is not preferred for evaluating the fit of chosen distribution.

If  $X$  is a random variable with cdf  $F$  then  $q$ -quantile of  $X$  is that value  $\gamma$  such that

$$F(\gamma) = P(X \leq \gamma) = q, \quad 0 < q < 1$$

$$\gamma = F^{-1}(q)$$

Let:  $\{x_i, i = 1, 2, \dots, n\} \rightarrow$  Sample of data from X

$\{y_j, j = 1, 2, \dots, n\} \rightarrow$  Samples arranged in ascending order, where  $y_1 \leq y_2 \leq \dots \leq y_n$ .

$j \rightarrow$  ranking or order number

$j=1$  for smallest and  $j = n$  for largest.

The  $q - q$  plot is based on the fact that  $y_j$  is an estimate of  $(j - 1/2) / n$  quantile of x. i.e.

$y_j$  is approximately  $F^{-1}[(j - 1/2) / n]$ .

If F is a member of an appropriate family of distributions, then plot  $y_j$  versus  $F^{-1}[(j - 1/2) / n]$  will be approximately a Straight line

If F is a member of an appropriate family of distributions and has appropriate parameter values, then the line will have slope 1.

If the assumed distribution is not appropriate then points will deviate from a straight line.

A robot is used to install the doors on automobiles along an assembly line. It was thought that the installation time followed a normal distribution. The robot is capable of accurately measuring installation times. Samples of 20 installation times are tabulated below. (Values are in seconds)

99.79      99.56      100.17      100.33      100.26      100.41      99.98  
99.83

100.23    100.27    100.02      100.47      99.55      99.62      99.65  
99.82

99.96      99.90      100.06      99.85

$j = 1, 2, \dots, 20$

Sample mean = 99.99 seconds

Sample variance =  $(0.2832)^2$  seconds.

The observations are arranged in ascending order in table 6.1

Table 6.1: Computation of values

j	Value(y <sub>j</sub> )	$F^{-1}((j - 1)/2)/20$
1	99.55	0.03
2	99.56	0.08
3	99.62	0.13
4	99.65	0.18
5	99.79	0.23
6	99.82	0.28
7	99.83	0.33
8	99.85	0.38
9	99.90	0.43
10	99.96	0.48
11	99.98	0.53
12	100.02	0.58
13	100.06	0.63
14	100.17	0.68
15	100.23	0.73
16	100.26	0.78
17	100.27	0.83
18	100.33	0.88
19	100.41	0.93
20	100.47	0.98

The ordered observations  $y_j$  versus  $F^{-1}((j - 1/2)/20)$  is plotted, which is shown in figure 6.3.

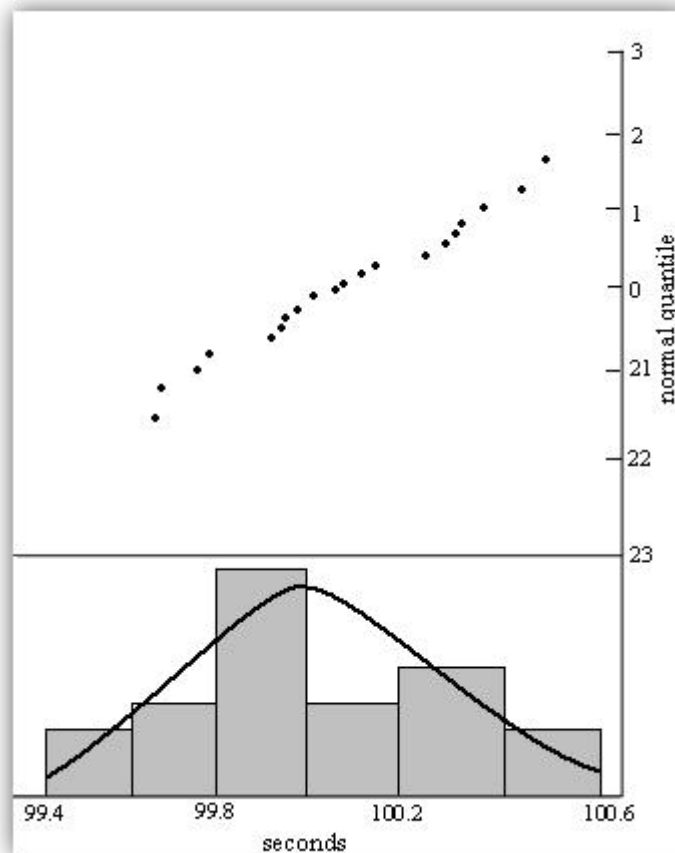


Figure 6.3: Histogram and q-q plot of the installation times

In evaluation of linearity of q-q plot, the following should be considered.

The observed values will never fall exactly on a straight line

The observed values are not independent, since they are ranked. The points will be scattered about the line.

The variances of extremes (largest & smallest values) are higher than variance in the middle of the plot. Greater discrepancies can be accepted at extremes. The linearity of points in the middle of plot is more important than linearity at the extremes.

## Parameter Estimation

The parameters are chosen to determine a specific instance of distribution family. When the data are available, these parameters can be estimated. Some estimators are described.

### Preliminary Statistics: Sample Mean and Sample Variance

The sample mean and sample variance are used to estimate the parameters of a hypothesized distribution.

The observations in a sample of size  $n$  are  $X_1, X_2, \dots, X_n$ .

1. If data are discrete or continuous raw data then sample mean is defined by

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

and sample variance is defined by

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - n \bar{X}^2}{n - 1}$$

2. If data are discrete and grouped in a frequency distribution then mean and variance is given by

$$\bar{X} = \frac{\sum_{j=1}^k f_j X_j}{n}$$

$$S^2 = \frac{\sum_{j=1}^k f_j X_j^2 - n \bar{X}^2}{n - 1}$$

Where  $f_j \rightarrow$  Observed frequency of value  $X_j$  of  $X$

$K \rightarrow$  Number of distinct values of  $X$

If data are discrete or continuous and have been placed in class interval, then mean and variance is

$$\bar{X} = \frac{\sum_{j=1}^c f_j m_j}{n}$$

$$S^2 = \frac{\sum_{j=1}^c f_j m_j^2 - n \bar{X}^2}{n - 1}$$

Where

$f_j \rightarrow$  Observed frequency

$m_j \rightarrow$  Midpoint of  $j^{\text{th}}$  interval

$c \rightarrow$  Number of class intervals

### Suggested Estimators

Numerical estimates of the distribution parameters are required to reduce the family of distributions to a specific distribution and to test the resulting hypothesis. The table 6.2 contains suggested estimators for distribution often used in simulation.

Table 6.2 Suggested estimators for distributions often used in simulation

Distribution	Parameter(s)	Suggested Estimators(s)
Poisson	$\alpha$	$\hat{\alpha} = \bar{X}$
Exponential	$\lambda$	$\hat{\lambda} = \frac{1}{\bar{X}}$

Gamma	$\beta, \theta$	$\hat{\beta}$ $\hat{\theta} = \frac{1}{\bar{X}}$
Normal	$\mu, \sigma^2$	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2$ (unbiased)
Lognormal	$\mu, \sigma^2$	$\hat{\mu} = \bar{X}$ (after taking $\ln$ of the data) $\hat{\sigma}^2 = S^2$ (after taking $\ln$ of the data)

**Note** - The parameter is denoted by  $\alpha$  and estimator is denoted by  $\hat{\alpha}$

The percentage rates of return on 10 investments in a portfolio are 18.8, 27.9, 21.0, 6.1, 37.4, 5.0, 22.9, 1.0, 3.1 and 8.3. Estimate the parameter of a lognormal model of this data.

Natural log of the given data is

2.9, 3.3, 3.0, 1.8, 3.6, 1.6, 3.1, 0, 1.1 and 2.1

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n} = 2.25$$

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^n X_i^2 - n \bar{X}^2}{n - 1} = 1.3$$

### Goodness-of-Fit tests

Goodness-of-fit test provide helpful guidance for evaluating the suitability of a potential input model. If single distribution is selected, then other distributions are called candidate distributions.

The test procedure begins by arranging the  $n$  observations into a set of  $k$  class intervals or cells. The test statistic is given by

$$\chi^2_0 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

The expected frequency  $E_i$  for each class intervals is computed as

$$E_i = np_i$$

Where

$p_i \rightarrow$  theoretical, hypothesized probability associated with  $i^{\text{th}}$  class interval.

$$p(x) = \begin{cases} \frac{e^{-\alpha} \alpha^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

$\chi^2_0$  approximately follows chi-square distribution with  $k-s-1$  degrees of freedom.

Where,

$s \rightarrow$  Number of parameters of hypothesized distribution estimated by sample statistics.

The hypotheses are

$H_0$ : Random variable  $X$ , conforms to distribution assumption with parameter(s) given by estimate(s).

$H_1$ : Random variable  $X$  does not conform

Null hypothesis  $H_0$ , is accepted if  $\chi^2_0 < \chi^2_{\alpha, k-s-1}$

If an  $E_i$  value is too small, it can be combined with expected frequencies in adjacent class intervals. The corresponding observed



frequency  $O_i$  values should be also combined and  $k$  should be reduced by one for each cell combined.

The vehicle arrival data is tabulated below

Table 6.3: Vehicle arrival

Days	Cars arrival in a particular period
12	0
10	1
19	2
17	3
10	4
08	5
07	6
05	7
05	8
03	9
03	10
01	11

Since the histogram of the data, shown in table 6.3 appears to follow Poisson distribution, the parameter  $\hat{\alpha} = 3.64$  was determined.

Hypotheses

$H_0$ : Random variable is Poisson distributed

$H_1$ : Random variable is not Poisson distributed

pmf of Poisson distribution

$$p(x) = \begin{cases} \frac{e^{-\alpha} \alpha^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

The probabilities associated with various values of  $x$  are obtained by using above equation

$p(0) = 0.026$     $p(3) = 0.211$     $p(6) = 0.085$     $p(9) = 0.008$   
 $p(1) = 0.096$     $p(4) = 0.192$     $p(7) = 0.044$     $p(10) = 0.003$   
 $p(2) = 0.174$     $p(5) = 0.140$     $p(8) = 0.020$     $p(11) = 0.001$

The computations are shown in the table 6.4

Table 6.4: Chi-square chart.

$X_i$	Observed Frequency, $O_i$	Expected Frequency, $E_i$	$\frac{(O_i - E_i)^2}{E_i}$
0	12	2.6	7.37
1	10	9.6	0.15
2	19	17.4	0.80
3	17	21.1	4.41
4	10	19.2	2.57
5	8	14.0	0.26
6	7	8.5	
7	5	4.4	
8	3	2.0	11.62
9	3	0.8	
10	3	0.3	
11	1	0.1	
	100	100.0	27.68

Critical value  $X^2_{0.05, 5} = 11.1$

$X^2_0 > X^2_{0.05, 5}$ , Null hypothesis  $H_0$  is not accepted

### Chi-square test with equal probabilities

If a continuous distributional assumption is being tested, class intervals that are equal in probability should be used instead of equal in width of interval.

For equal probabilities

$$p_i = 1/k$$

$$E_i = np_i \geq 5$$

Substituting for  $p_i$ , we get

$$\frac{n}{k} \geq 5$$

Solving for k,

$$k \leq \frac{n}{5}$$

- In case of normal, exponential or Weibull distribution, the method is straight forward.
- For gamma or other certain distributions, the computation of end points for class intervals is complex and requires numerical integration for density function.

Life test were performed on a random sample of electronic chips at 1.5 times the nominal voltage, and their lifetime (or time to failure) in days was recorded:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

Since the histogram appears to follow exponential distribution. The

$$\hat{\lambda} = \frac{1}{\bar{X}} = 0.084$$

parameter is given a

## Hypotheses

Ho: Random variable is exponentially distributed

H1: Random variable is not exponentially distributed

The intervals must be of equal probability, so the end points of the class intervals must be determined. Number of intervals should be less than or equal to  $n / 5$ .

Here  $n = 50$ , so  $k \leq 50 / 5 \rightarrow k \leq 10$

Let  $k = 8$ , then each interval will have probability as  $p = 1/k = 1/8 = 0.125$

The cdf of exponential distribution is

$$F(a_i) = 1 - e^{-\lambda a_i}$$

Where  $a_i \rightarrow$  End point of  $i^{\text{th}}$  interval,  $i = 1, 2, 3, \dots, k$

$F(a_i) \rightarrow$  cumulative area from 0 to  $a_i$ .

$$F(a_i) = ip$$

$$ip = 1 - e^{-\lambda a_i}$$

$$e^{-\lambda a_i} = 1 - ip$$

Apply log to both sides, then

$$a_i = - \frac{1}{\lambda} \ln (1 - ip), \quad i = 0, 1, \dots, k$$

Regardless of value of  $\lambda$ , the above equation will always result in  $a_0 = 0$  and  $a_k = \infty$ .

With  $\hat{\lambda} = 0.084$ ,  $k = 8$ ,  $a$  is determined by

$$a_1 = - \frac{1}{0.084} \ln(1 - 0.125) = 1.590$$

$$a_2 = - \frac{1}{0.084} \ln(1 - (2)(0.125)) = 3.425$$

$$a_3 = 5.595$$

$$a_4 = 8.252$$

$$a_5 = 11.677$$

$$a_6 = 16.503$$

$$a_7 = 24.755$$

$$a_0 = 0$$

$$a_k = \infty$$

The first interval is  $[0, 1.590)$  that is  $0 \leq x < 1.590$  second interval  $[1.590, 3.425)$  and so on. The values are computed and tabulated in table 6.5

Table 6.5 Chi-square goodness-of-fit test

Class Interval	Observed Frequency, $O_i$	Expected Frequency, $E_i$	$\frac{(O_i - E_i)^2}{E_i}$
$[0, 1.590)$	19	6.25	26.01
$[1.590, 3.425)$	10	6.25	2.25
$[3.425, 5.595)$	3	6.25	0.81
$[5.595, 8.252)$	6	6.25	0.01
$[8.252, 11.677)$	1	6.25	4.41
$[11.677, 16.503)$	1	6.25	4.41
$[16.503, 24.755)$	4	6.25	0.81
$[24.755, \infty)$	6	6.25	0.01
	<u>50</u>	<u>50</u>	<u><math>38.72 = X_0^2</math></u>

The degrees of freedom is  $k - s - 1 = 8 - 1 - 1 = 6$

At  $\alpha = 0.05$ ,  $X_{0.05,6}^2 = 12.6$ , Therefore  $X_0^2 > X_{0.05,6}^2$ , null hypothesis is rejected.

## Disadvantages of using the chi-square test

Changing the number of classes and interval width affects the value of calculated and tabulated chi-square.

A hypothesis may be accepted when the data are grouped in one way but rejected if it is done in another way.

It requires the data to be placed in the class intervals. In case of continuous grouping is arbitrary.

### Kolmogorov-Smirnov goodness-of-fit test

Any continuous distributional assumption can be tested for goodness-of-fit using Kolmogorov-Smirnov test, while discrete distributional assumptions can be tested using gap test. This test is useful when sample sizes are small and when no parameters have been estimated from the data. The critical values in table A.8 are biased, they are too conservative. Conservative means that critical values will be too large, resulting in smaller type I ( $\alpha$ ) errors than those specified.

The interarrival times (minutes) are collected over 100-minute interval and are arranged in order of occurrence.

[illegible]

## Hypotheses

H0: The inter arrival times are exponentially distributed.

H1: The inter arrival times are not exponentially distributed.

The data were collected over the interval 0 to 100 minutes, so  $T = 100$  minutes. If the underlying distribution of inter arrival times  $\{T_1, T_2, T_3, \dots\}$  is exponential, arrival times are uniformly distributed on interval  $(0, T)$ .

The arrival times  $T_1, T_1 + T_2, T_1 + T_2 + T_3 \dots$  are obtained by adding inter arrival times, then the arrival times are normalized to  $(0, 1)$  so that Kolmogorov-Smirnov test can be applied.

On interval  $(0, 1)$ , the points will be  $[T_1 / T, (T_1 + T_2) / T, \dots]$ . The resulting 50 points are

$$D^+ = \max \{(i / N) - R(i)\} 1 \leq i \leq N$$

$$D^- = \max \{R(i) - [(i - 1) / N]\} 1 \leq i \leq N$$

0.00	0.00	0.03	0.05	0.07	0.08	0.10	0.11	0.13	0.15
44	97	01	75	75	05	59	11	13	02
0.16	0.16	0.19	0.19	0.20	0.29	0.31	0.33	0.33	0.35
55	76	56	60	95	27	61	56	66	08
0.35	0.35	0.36	0.37	0.43	0.46	0.47	0.50	0.53	0.53
53	61	70	46	00	94	96	27	15	82
0.54	0.55	0.59	0.65	0.65	0.68	0.70	0.71	0.72	0.74
94	20	77	14	26	45	08	54	62	68
0.75	0.76	0.78	0.79	0.82	0.84	0.87	0.90	0.96	0.97
53	36	80	82	06	17	32	22	80	44

$$D^+ = 0.1054$$

$$D^- = 0.0080$$

$$D = \max \{D^+, D^-\} = \max \{0.1054, 0.0080\} = 0.1054$$

Critical value  $\rightarrow D_{0.05} = 1.36/\sqrt{n} = 1.36 / \sqrt{50} = 0.1923$  (from table A.8)

$D < D_{\alpha}$ . Therefore, interarrival times are exponentially distributed.

**Note** - A similar to Kolmogorov-Smirnov test is Anderson-Darling test. It is the test based on difference between empirical cdf and fitted cdf.

### **p-Values and “Best Fits”**

The p-value is the significance level at which one would just reject  $H_0$  for given value of test statistic. Therefore, a large p-value tends to indicate a good fit, while small p-value suggests a poor fit.

The p-value can be viewed as a measure of fit. This suggests we could fit every distribution at our disposal, compute a test statistic for each fit then choose the distribution that yields largest p-value.

Some points to remember

The software may know nothing about the physical basics of data and that information can suggest distribution families that are appropriate. The goal of input modeling is often to fill in gaps or smooth the data than finding an input model that conforms as closely as possible to the given sample.

Automated best-fit procedure tends to choose more flexible distributions (gamma, Weibull and exponential) because extra flexibility allows closer conformance to the data and better summary measure of fit. But close conformance to data may not always lead to most appropriate input model.

A summary statistic, like p-value is just a summary measure. It just tells where the lack of fit occurs.

### **Selecting Input Models without Data**

To develop a simulation model for demonstration purpose or preliminary study – before any process, data are available. In this case



modeler chooses input models and carefully checks the sensitivity of results to the chosen models.

There are many ways to obtain information, if data are not available. Few are mentioned below

### **Engineering data**

The values provided by manufacturers provide a starting point for input modeling by fixing a central value.

### **Expert option**

Talking to the experts who have experience with the process or similar processes. They can provide optimistic, pessimistic and most likely thoughts.

### **Physical or conventional limitations**

Many real processes have physical limits on performance (Ex. Computer data entry is faster than a person can type). Do not ignore obvious limits or bounds that narrow the range of input process.

### **The nature of process**

The choice of distribution should be after clear understanding of distributions.

When no data is available then uniform, triangular and beta distributions are used as input models. A useful refinement is obtained, when minimum, maximum and one or more breakpoints can be given. A breakpoint is an intermediate value and a probability of being less than or equal to that value.

For a product planning simulation, the sales volume of various products is required. The sales person responsible for product XYZ says that no fewer than 1000 units will be sold because of existing contracts, no more than 5000 units will be sold because of that is the entire market for the product. Based on experience she believes that there is

90% chance of selling more than 2000 units

25% chance of selling more than 3500 units

Only 1% chance of selling more than 4500 units

Minimum – 1000 units

Maximum – 5000 units

90% chance of selling more than 2000 units

10% = 0.10 chance of selling between 1000 and 2000 units

1% = 0.01 chance of selling more than 4500 units

25% = 0.24 chance of selling more than 3500 unit (because 1% chance of selling more than 4500 units).

Remaining 65% chance of selling between 2000 and 3500 units

The table summarizes the above information.

Interval (hours)	Frequency	Cumulative frequency
$1000 \leq x \leq 2000$		
$2000 < x \leq 3500$	0.10	0.10
$3500 < x \leq 4500$	0.65	0.75
$4500 < x \leq 5000$	0.24	0.99
	0.01	1.00

### Exercises:

Explain steps involved in development of useful model of input data.

Draw the histogram of vehicle arriving at intersection between 9:00AM and 9:10AM was monitored for 100 work days

Arrivals period:	0	1	2	3	4	5	6	7	8	9	10	11
Frequency:	14	10	13	14	10	6	8	5	5	4	9	2

The inter-arrival times (minutes) are collected over 100-minutes interval and are

0.4	0.5	2.0	2.7	2.0	0.3	2.5	0.5	2.0	1.8	1.5	0.2
4	3	4	4	0	0	4	2	2	9	3	1
2.8	0.0	1.3	8.3	2.3	1.9	0.1	1.4	0.4	0.0	1.0	0.7
0	4	5	2	4	5	0	2	6	7	9	6
5.5	3.9	1.0	2.2	2.8	0.6	1.1	0.2	4.5	5.3	0.1	3.1
5	3	7	6	8	7	2	6	7	7	2	9
1.6	1.4	1.0	2.0	0.8	0.8	2.4	2.1	3.1	2.9	6.5	0.6
3	6	8	6	5	3	4	1	5	0	8	4

arranged in order of occurrence. By applying Kolmogorov-Smirnov test find out the above time intervals are exponentially distributed.

The following data are available on demand and lead time for last 10 years. Determine the correlation.

Lead Time: 6.5 4.3 6.9 6.0 6.9 6.9 5.8 7.3 4.5

Demand: 103 83 116 97 112 104 106 109 92

With example explain different methods used in distribution concepts.

## Chapter-7

### Verification and Validation

#### Verification:

It verifies the program used for the simulation. The model developed by program is appropriate to the expectation or partial. It is tested by analyst. If partial, repeat the model program by changing the input features.

#### Validation:

Expert and user determine the developed model is accurate representation of real system or partial. If partial, repeat the model program by changing the input features. Validation is the overall process of comparing the model and its behavior to the real system.

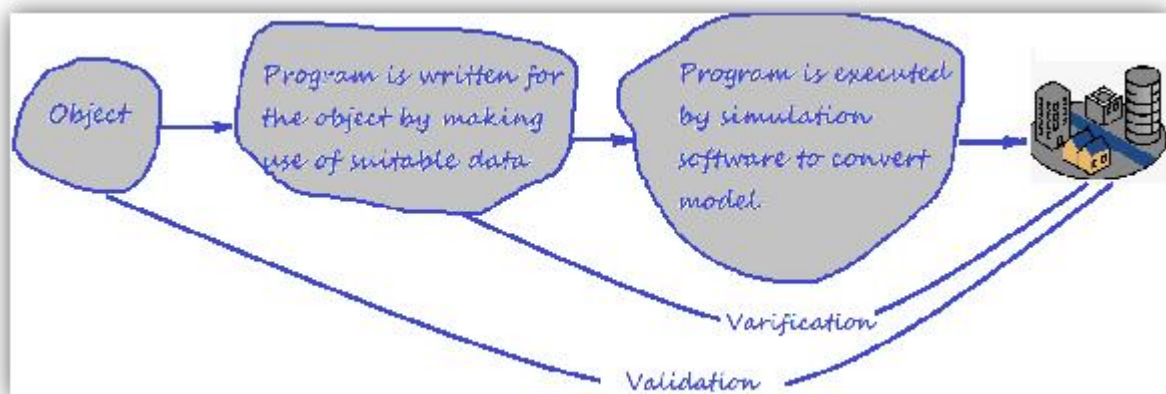


Figure7.1: Validation and verification.

It can also be explained with the help of an analogy with respect to checking the correctness of the program written as shown in figure 7.2.

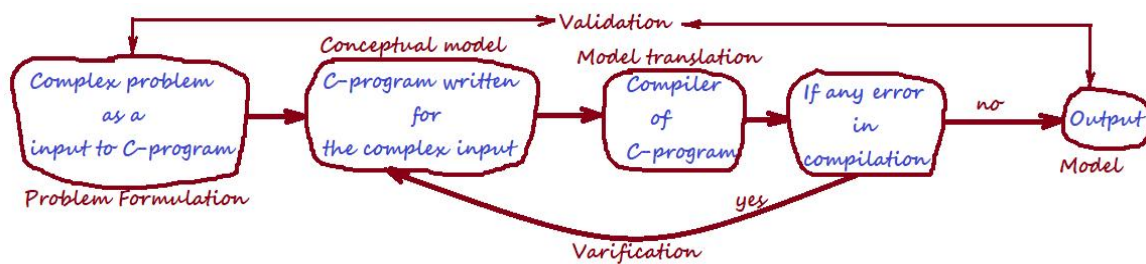


Figure 7.2: C program structure for validation and verification

The complex input is given to compiler through the C-program. The compiler compiles the program, if any errors in the program the user correct the errors and compiled again. This step is repeated till there are no-errors.

The first step in model building consists of observing the real system and the interactions among their various components and of collecting data on their behavior. But observation alone seldom yields sufficient understanding of system behavior. Persons familiar with the system, or any subsystem, should be questioned to take advantage of their special knowledge. Operators, technicians, repair and maintenance personnel, engineers, supervisors, and managers understand certain aspects of the system that might be unfamiliar to others. As model development proceeds, new questions may arise, and the model developers will return to this step of learning true system structure and behavior.

The second step in model building is the construction of a conceptual model-a collection of assumptions about the components and the structure of the system, plus hypotheses about the values of model input. The third step is the implementation of an operational model, usually by using simulation software and parameters. incorporating the assumptions of the conceptual model into the worldview and concepts of the simulation software.

## Verification of Simulation Model

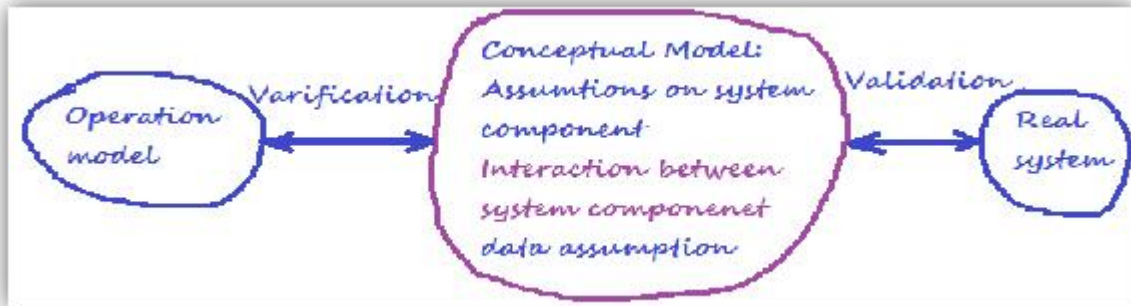


Figure 7.3: Verification of Simulation Model.

The purpose of model verification is to assure that the conceptual model is reflected accurately in the operational model. Verification asks the following question: Is the conceptual model accurately represented by the operational model, it is a model used for computerized representation? Validation asks conceptual model correctly interpreted as real time.

Some considerations in verification process are:

The computerized model has to be checked by others, then its developer.

Make a flow diagram (logic flow), which includes operations of system, so that we know what operations takes place when the events occur.

Closely examine the output of model for reasonableness, under a variety of settings of input parameters.

At the end of simulation, have the computerized representation that prints the input parameter. This is to confirm that these parameter values are not changed or modified.

As far as possible make the self-documentation of computerized model.

If the computerized representation is animated, then check whether the animation reflects the real system.

The interactive run controller (IRC), debugger assists in finding and correcting the errors in the following ways.

As the simulation progress, it can be monitored. This is achieved by advancing the simulation under a desired time and then display model information.

Focus on each or multiple line of logic that constitutes a procedure or a particular entity. For example, every time a specified entity becomes active, simulation will pause.

Selected model components values can be observed.

Simulation can be temporarily suspended or paused, to view information and reassign values or redirect entities.

Graphical interfaces are required to represent the model graphically, it simplifies the model understanding.

The standard statistics (average waiting time, average queue length etc.,

) are automatically collected in simulation language, which takes little time to display all statistics of interest.

Two sets of statistics that indicates the factor model reasonableness are

Current contents refer to the number of items in each components of the system at a given time.

Total count refers to the total number of items that have entered each component of the system by a given time.

Some possibilities of two test statistics are

If the current content in some parts of the system is high, then it indicates that a large number of entities are delayed and queue is unstable.

If the total count for some subsystem is zero, then no items entered the system.

If the current count and total count is equal to one then an entity has captured a resource but never freed it.

A careful evaluation is required to detect the mistakes in model logic. To help in error detection, it is best to adopt any of these verification processes.

**Common sense technique** - Forecasts a reasonable range of values of selected output statistics before making a run of the model. So, it reduces discrepancies and unusual output.

Documentation should contain brief comments, definitions of all variables and parameters and description of each major section of computerized model. Documentation is important as it provides a means to clarify the logic of a model.

**Trace** is more sophisticated technique. Trace is a detailed computer printout which gives the value of each variable in a program, every time that one of these variables changes in value. The purpose of trace is to verify the correctness of computer program by making detailed calculations (manual). To make this practical, a simulation with trace is usually restricted because of time factor. Selective trace is also carried out as required. For example, a selective trace could be set for specific locations in the computerized model.

In a single server queue model, an analyst made a run over 16 units of time and observed that time average length of waiting line was  $\bar{L}_Q = 0.4375$  customer, which is reasonably a short run. So, a detailed verification is required to be performed by analyst. The trace is shown in table 7.1

Table 7.1: Simulation of trace

Definition of variables:

CLOCK = Simulation clock

E = Event type (start, arrival, departure, or stop)

NC = Number of customers in system at time 'CLOCK'

STATUS = Status of server (1- busy, 0- idle)



State of system just after the Named Event Occurs:

CLOCK = 0      E = 'Start'      NC = 0      STATUS = 0

CLOCK = 3      E = 'Arrival'      NC = 1      STATUS = 0

CLOCK = 5      E = 'Depart'      NC = 0      STATUS = 0

CLOCK = 11      E = 'Arrival'      NC = 1      STATUS = 0

CLOCK = 12      E = 'Arrival'      NC = 2      STATUS = 1

CLOCK = 16      E = 'Depart'      NC = 1      STATUS = 1

The table 7.1 gives hypothetical printout from simulation time clock = 0 to 16 for single server queue. Note that at simulation time CLOCK = 3, the number of customers in the system is NCUST = 1, but the server status is idle (STATUS = 0). This is incorrect in logic, so an error is found. It should be rectified by using equation,

$$\bar{L}_Q = \frac{1}{T} \sum_{i=0}^{\infty} i T_i$$

The reader can verify that  $\bar{L}_Q$  is computed correctly from the data

$$\bar{L}_Q = \frac{(0-0)3 + (1-0)2 + (0-0)6 + (1-0)1 + (2-1)4}{3 + 2 + 6 + 1 + 4} = 0.4375$$

The computer value is correct according to the given status, but its value is indeed wrong as the attribute STATUS was not the correct value.

Of the three techniques, it is recommended that first two always to be carried out. The close examination of model output for reasonableness is especially valuable and informative.

## Calibration and Validation of Models

The figure 7.3 shows calibration, is the iterative process of comparing the model to the real system, making adjustments, additional adjustments and comparing simulation model to real system.

Comparing the model to reality is performed by either subjective test or objective test.

Subjective test involves people, who are knowledgeable about one or more aspects of system, making judgment about the model and its output.

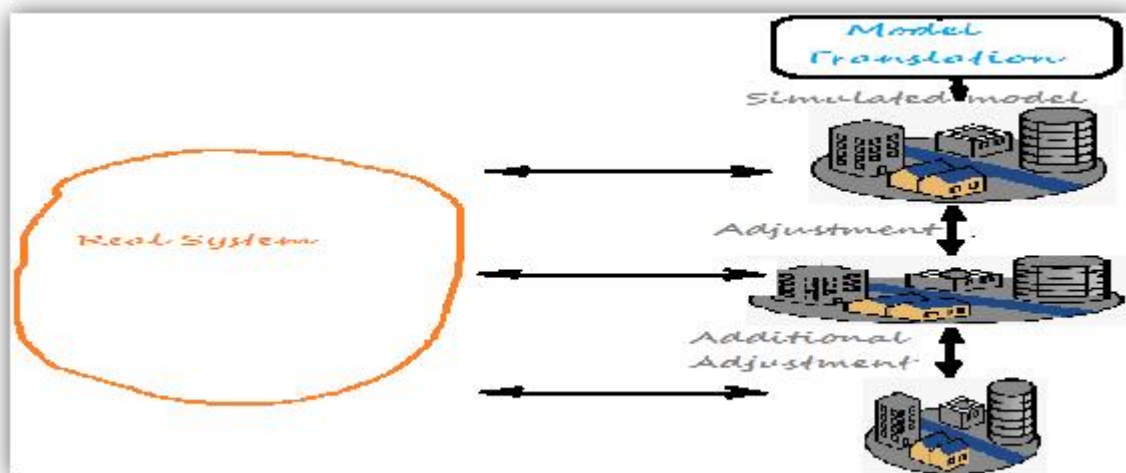


Figure 7.4: Iterative process of calibrating a model

Objective test requires data on systems behavior and the corresponding data produced by the model.

The iterative process of comparing model and real system, revising conceptual and operational model, is carried out until the model is judged accurate.

A possible criticism of calibration phase is to stop at the point where model has been “fit” to one data set. This can be overcome by collecting a new set of system data and using at final stage of

validation i.e. after the model has been calibrated using the original system data set, a final validation is done by using second system data set. In case of any discrepancy, the modeler has to return back to calibration phase and modify the model until it is acceptable.

Each revision of model involves cost, time and effort. The modeler must weigh increase in model accuracy versus the cost of increased validation effort. If the level of accuracy is not obtained within the budget constraints then accuracy level should be lowered or reject the model.

Naylor and finger [1967] proposed a three-step approach for validation

Build a model that has high face validity.

Validate model assumptions.

Compare the model input–output transformation to corresponding input output transformation for real system.

### **Face Validity**

The potential users of the model must be involved in the model construction from its conceptualization to implementation, to assure that the reality is built into the model through assumptions regarding system structure and reliable data.

The advantages of involving potential users are.

They can evaluate the model output for reasonableness and help in identifying the deficiencies. So, they are involved in the calibration process, as the model is iteratively improved.

The increase in the model's perceived validity or credibility helps the manager to trust the simulation results, a basis for decision making.

Sensitivity analysis can also be used to check a model's face validity – the model user is asked whether the model behaves in the expected way, when one or more input variables are changed. Based on

experience and observation on the real system, both model user and builder address the problem.

For most large-scale simulation models, many possible sensitivity tests are carried out as there are many input variables. The builder must choose the most critical input variables for testing if it is too expensive or time consuming.

### **Validation of Model Assumptions**

Model assumptions have two general classes

Structural assumptions involve questions of how the system operates simplifications and abstractions of reality.

Example: Customer queuing and service facility in a bank - Customers form a queue for each teller, they are served on first-come, first-serve basis. When there are many queues, customers may shift to other line that moves faster. The numbers of tellers are either fixed or variable. These structural assumptions should be verified by observations at regular interval with discussions between managers and tellers, regarding policies and implementation of these bank policies.

### **Data**

assumptions involve collection of reliable data and correct statistical analysis of the data.

Example – For a bank the data that were collected are

Inter arrival times of customers during several 2 hours period of peak loading.

Inter arrival times during a slack period.

Service times for commercial accounts.

Service times for personal accounts.

The reliability of data is verified by consultation with bank managers, who identify typical slack/rush time. When two or more data sets

collected are combined, objective statistical tests are performed for homogeneity of data.

Additional tests may be required for correlation in data. The analyst begins statistical analysis as soon as he is assured of dealing with a random sample.

The analysis consists of three steps

Identifying the appropriate probability distribution.

Estimating the parameters of hypothesized distribution.

Validating the model by goodness-of-fit tests (chi-square or Kolmogorov-Smirnov test) and by graphical methods.

**Note** - The use of goodness-of-fit tests is an important part of validation of model assumptions.

### **Validating Input-Output Transformations**

In this phase, the model is viewed as an input-output transformation i.e. model accepts values of input parameters and transforms these inputs into outputs measures of performance. The modeler collects two sets of data, one data set used at the time of developing and calibrating the model and the other if required at the final validation test.

In any case, the modeler should use the main responses of interest as criteria for validating a model. A necessary condition in this phase is, some version of system under study exists, so data can be collected (at least one set of input conditions), which might be useful to compare with model predictions. If system is in planning stage and no system operating data is collected, complete input output validation is not possible.

What about the validity of model of a nonexistent proposed system or model of existing system under new input conditions?

First, the responses of two models under similar input conditions will be used as criteria for comparison of existing and

## **Proposed System.**

Second, the proposed system is a modification of existing system in most cases. The modeler hopes that confidence in the model of existing system can be transferred to the model of new system. This transfer of confidence by modeler can be justified only if new model is relatively with minor modification of old model in terms of changes to computerized representation of the system.

Changes in computerized representation of the system, ranging relatively from minor to major includes.

Minor changes of single numerical parameters.

Example - speed of a machine, arrival rate of customers

Minor changes of statistical distribution.

Example - distribution of a service time or time to failure of a machine.

Major changes in logical structure of a subsystem.

Example - change in queue discipline for a waiting-line model.

Major changes of design in new system.

Example - computerized inventory control system.

If the changes are minor then it can be carefully verified and output from the new model is accepted with confidence. If a similar subsystem exists elsewhere, it may be possible to validate sub model that represents the subsystem and then integrate this sub model with other validated sub models to build a complete model, this is a partial validation of major changes.

There is no way to completely validate the input-output transformations of a model of non-existing system. The modeler should consider time and budget constraints and use as many validation techniques including input-output validation of subsystem models if operating data can be collected on such subsystems.

## **Input-Output Validation: Using Historical Input Data**

To conduct a validation test using historical input data, it is important that all input data ( $A_n$ ,  $S_n$ ....) and all system response data such as average delay ( $Z_2$ ) should be collected during the same time period. Otherwise the comparison of model to the system responses could be misleading – the responses depends on inputs and structure of the system or model.

Implementation of this technique for large system is difficult because of the need of simultaneous data collection. Some electronic counters and devices are used for ease of data collection. In this technique the modeler hopes that simulation will provide a replica of a real system, but to determine the level of accuracy both model builder's and model user's judgment is considered.

## **Input-Output Validation: Using a Turing Test**

The comparison of model output to system output can be carried out by persons who are knowledgeable about system behavior, when no statistical test is readily applicable.

For example: Suppose five reports of system performance over five different days are prepared and simulation output data are used to produce five fake reports. So, there are 10 reports exactly in same format and contains information as required by managers and engineers. These 10 reports are shuffled randomly and submitted to the engineer, to identify fake and real reports. If the engineer identifies fake reports, then the model builder questions the engineer and uses the information gained to improve the model. If the engineer cannot distinguish, then the modeler will conclude that this model is adequate. This type of validation test is called Turing test.

It provides a valuable tool in detecting model inadequacies and eventually in increasing model credibility.