UNIT 2. STATISTICAL MODELS IN SIMULATION

While modeling real-world phenomena, in some of the situations, the actions of the entities within the system can be predicted completely. Whereas, in certain situations, variation may occur by chance and cannot be predicted. However, some statistical model may describe the behavior of such systems. An appropriate model can be developed by sampling the phenomenon. Then through the trained guesses, the model builder can:

- select a known distribution form
- make an estimate of the parameters of this distribution
- test to see the goodness of fit.

2.1 Review of Terminology and Concepts

2.1.1 Descrete Random Variables

Let X be a random variable. If the number of possible values of X is finite, then X is called as discrete random variable. The possible values of X are listed as $x_1, x_2, ...$ The possible values that a discrete random variable X can take are called as range space of X and is denoted by R_X. The probability that a random variable X takes the value x_i is denoted by $p(x_i)$. For any discrete random variable X, the number $p(x_i)$ must satisfy the following two conditions:

1. $0 \leq p(x_i) \leq 1, \forall i$
2. $\sum_{i=1}^{\infty} p(x_i) = 1$

The collection of pairs $(x_i, p(x_i))$, $i = 1, 2, ...$ is called as probability distribution of X and $p(x_i)$ is called as ***probability mass function (pmf)*** of X.

Example 1: The number of people arriving to a bank on a particular day is a random variable, say X. Then, the range space $R_X = 0, 1, 2, ...$

Example 2: Consider an experiment of tossing a single die. It is obvious that, the possible outcome may any one of the numbers from 1 to 6. Thus,
The random variable $X =$ the value that appears on the top of a die.
The range space $R_X = 1, 2, 3, 4, 5, 6.$
One can observe that the probability of getting any number on top is $\frac{1}{6}$.

Example 3: Consider an experiment of tossing two dice at a time. Let the random variable X be the sum of the numbers which appears on top of the dice. The possible numbers that may appear on top of two dice would be a set of ordered pair:

$$
\begin{aligned}
S = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\
&(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\
&\dots\dots\dots\dots\dots\dots\\
&(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}
\end{aligned}
$$

So, the range space of X contains the possible sums of the ordered pairs of the set S. Thus, $R_X = 2, 3, 4, ..., 12.$

Prepared By: Dr. Chetana Hegde, Associate Professor, RNS Institute of Technology, Bangalore - 98
Email: chetanahegde@ieee.org

2.1.2 Continuous Random Variables

If the range space R_X of the random variable X, is an interval or a collection of intervals, then X is called as continuous random variable. For a continuous random variable X, the probability that X lies in the interval [a, b] is given by-

$$P(a \leq X \leq b) = \int_a^b f(x).dx$$

The function f(x) is called as Probability Density Function (PDF) of random variable X. The PDF satisfies following conditions:

(i) $0 \leq f(x) \leq 1 \quad \forall x \in R_X$

(ii) $\int_{R_X} f(x) dx = 1$

(iii) $f(x) = 0$. if x is not in R_X.

NOTE:

1. For any specified value x_0, the probability is zero, because -

$$\int_{x_0}^{x_0} f(x).dx = 0$$

2. As $P(X = x_0) = 0$, the following equation holds good:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b)$$
$$= P(a < X < b)$$

2.1.3 Cumulative Distribution Function

The probability that a random variable X having the value less than or equal to x is called as Cumulative distribution function (CDF) and is denoted by $F(x)$. That is,

$$F(x) = P(X \leq x).$$

If X is a discrete random variable, then

$$F(x) = \sum_{\forall i} P(x_i)$$

If X is a continuous random variable, then

$$F(x) = \int_{-\infty}^{x} f(x).\,dx$$

Some of the properties of CDF are:

(i) F is a nondecreasing function. That is, if $a < b$, then $F(a) \leq F(b)$

(ii) $\lim_{x \to \infty} F(x) = 1$

(iii) $\lim_{x \to -\infty} F(x) = 0$

NOTE: In most of the cases, probability of random variable X in a specified range is calculated using CDF. That is,

$$P(a < X \leq b) = F(b) - F(a) \quad \forall a < b.$$

2.1.4 Expectation

Expectation of a random variable x is also known as mean (denoted by μ). The expectation $E(x)$ for a discrete random variable is –

$$E(x) = \sum_{\forall i} x_i \cdot P(x_i)$$

For a continuous random variable,

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

$E(x)$ is also called as 1^{st} moment of x. The n^{th} moment of x is computed as –

$$E(x) = \sum_{\forall i} x_i^n \cdot p(x_i) \quad , \text{ if } x \text{ is discrete}$$

$$E(x) = \int_{-\infty}^{\infty} x^n f(x)\, dx \quad , \text{ if } x \text{ is continuous.}$$

The variance of x denoted by $V(x)$ or σ^2 is given as –

$$V(x) = E[(x - E(x))^2]$$
$$= E(x^2) - [E(x)]^2$$

The standard deviation σ is –

$$\sigma = \sqrt{V(x)}$$

2.1.5 The Mode

The mode is one of the measures of central tendency used for describing several statistical models. In case of discrete random variable, the mode is the value of variable which occurs most frequently. In other words, it is the value of a variable which has appeared more number of times in a given list. For a continuous random variable, mode is the value at which the pdf is maximized. Note that, mode may not be unique for a given list. If there are two modal values for a given random variable, then the distribution is called as ***bimodal distribution.***

2.2 USEFUL STATISTICAL MODELS

While developing simulation models, the analyst may require to generate random events, identify statistical distribution of those events and to use well-known statistical models. So, here, we will discuss various statistical models appropriate to various applications.

2.2.1 Queuing Systems

Queuing is nothing but a waiting line. For example, people are waiting in a queue to get a ticket. There may be one or more counters and the number of people waiting in a queue is dynamic. A queuing system is described its population, the nature of arrivals, the service mechanism, the system capacity and the queuing discipline.

In the queuing systems, inter-arrival and service time patterns will be given. The times between arrivals and the service times are probabilistic, in most of the cases. The distribution of time between arrivals and the distribution of the number of arrivals per time period are important in the simulation of queuing systems. The distributions that suits queuing system may depend on following situations:

- If the service time is completely random, the exponential distribution is suitable for simulation purpose. Example: The time required to get a task done at a government office is random. It may vary depending on type of the task, the person whom you approach etc.
- If the service time at the beginning of the system is low, and increases as the time passes by and again reduces at the end, then the Normal distribution is suitable. For example, the service time at the bank in the morning hours will be less, it increases afterwards and again decreases by the end of the day.
- In certain other situations, gamma and Weibull distributions are also used.

Inventory and Supply-Chain Systems

There are three major random variables in inventory and supply-chain systems:

- The number of units demanded per order or per time period
- The time between the demands
- The time duration between placing an order and receiving the items. This is called as lead time.

The Gamma distribution will suit for the random variable representing lead time. Even, Geometric, Poisson and Negative Binomial distributions will suit for random variables representing various types of demands.

2.2.2 Reliability and Maintainability

Time to failure has been modeled with various distributions like Exponential, Gamma and Weibull. Exponential distribution is used when there are random failures. For modeling standby redundancy, Gamma distribution is used. When there are large numbers of components in a system and the failure is dues of large number of defects, Weibull distribution is used.

2.2.3 Limited Data

In many situations, simulation modeling begins before the completion of data collection. For such a limited data, Uniform, Triangular and Beta distributions are useful. When an inter-arrival time or service time is random, the uniform distribution is used. The Triangular distribution is used when assumptions are made about the minimum, maximum and modal values of the random variable.

2.2.4 Other Distributions

Several other distributions like Bernoulli and Binomial distributions are helpful in discrete-system simulation. The hyper-exponential distribution is similar to the exponential distribution, but its greater variability might be helpful in certain cases.

Discrete Distributions

Discrete random variables are used to explain random phenomena in which only integer values can occur. We will discuss following discrete distributions:

(i) Bernoulli Distribution
(ii) Binomial Distribution
(iii) Geometric "
(iv) Poisson "
(v) Negative Binomial "

(i) <u>Bernoulli Distribution</u>:

A random variable X taking only two values zero and one representing failure and success respectively is said to follow Bernoulli distribution. That is, in an experiment with n trials,

$$x_i = \begin{cases} 1, & \text{if } i^{th} \text{ experiment is success} \\ 0, & \text{if } '' \quad '' \quad '' \text{ failure} \end{cases}$$

The probability distribution is given by-

$$p(x_j) = \begin{cases} p, & \text{if } x_j = 1 \\ 1-p = q, & \text{if } x_j = 0 \end{cases}$$

<u>Mean</u>:

$$E(x) = \sum_{\forall i} x_i . p_i = 1.p + 0.q = p$$

<u>Variance</u>:

$$\sigma^2 = E[X^2] - [E(x)]^2$$
$$= \{1^2.p + 0^2.q\} - p^2 = p - p^2$$
$$= p(1-p) = pq$$

(ii) Binomial Distribution

Multiple (say, n) trials of Bernoulli distribution constitutes a Binomial distribution. Assume that an experiment is conducted n times. Each experiment may result in success or failure. So, series of results may be -

SSFSFFSSSFSFFSS.....

putting all success together,

$$\underbrace{SSSSSSS}_{x \text{ times}} \; \underbrace{FFFFFFFF}_{n-x \text{ times}}$$

If probability of success is denoted by p, and probability of failure as q, then the success is p^x and failures are q^{n-x}. Out of n experiments, x experiments may result in success. Keeping these information, one can say that the probability distribution is given as-

$$p(x) = \begin{cases} {}^nC_x \cdot p^x \cdot q^{n-x}, & x = 0, 1, 2 \dots n \\ 0, & \text{otherwise} \end{cases}$$

As Binomial random variate is nothing but a Bernoulli variate repeated n times, it is obvious that -

$$E(x) = np$$

$$V(x) = npq$$

Example

Qn: A production process manufactures computer chips on the average at 2% nonconforming (i.e. defect). Every day, a random sample of size 50 is taken from the process. If the sample contains more than two nonconforming chips, the process will be stopped. Compute the probability that the process is stopped by the sampling scheme.

Solution: Consider a sampling process as $n = 50$ Bernoulli trials with $p = 2\% = 0.02$.

Then, the total no of defective chips in the sample X would have a binomial distribution as –

$$p(x) = \begin{cases} {}^{50}C_x (0.02)^x (0.98)^{50-x}, & \text{if } x = 0, 1, 2, \ldots, 50 \\ 0, & \text{otherwise} \end{cases}$$

It is given that, the process will be stopped if more than 2 defective chips are found. Hence we need to compute $P(X > 2)$.

But, $P(X > 2) = 1 - P(X \leq 2)$

Now, $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

or $P(X \leq 2) = \sum_{x=0}^{2} {}^{50}C_x (0.02)^x \cdot (0.98)^{50-x}$.

$$\therefore P(X \le 2) = {}^{50}C_0 (0.02)^0 (0.98)^{50-0} + {}^{50}C_1 (0.02)^1 (0.98)^{50-1} + {}^{50}C_2 (0.02)^2 (0.98)^{50-2}$$

$$= 1 \cdot 1 \cdot (0.98)^{50} + 50 (0.02)(0.98)^{49} + 1225 (0.02)^2 (0.98)^{48}$$

$$= 0.3642 + 0.3716 + 0.1858$$

$$= 0.9216$$

Now,

$$P(X > 2) = 1 - P(X \le 2)$$

$$= 1 - 0.9216$$

$$= 0.0784$$

$$\approx 0.08$$

The probability that the production process is stopped on any day is 0.08.

NOTE : Mean and variance for the above random sample can be calculated as-

$$E(X) = np = 50 \cdot (0.02) = 1$$

$$V(X) = npq = 50 (0.02)(0.98) = 0.98$$

(iii) Geometric and Negative Binomial Distributions

A random variable X defined as the number of trials required to achieve first success is said to follow geometric distribution. That is, the event $\{X = x\}$ occurs when there are $x-1$ failures and then a success occurs. Each failure has a probability q (i.e. $1-p$) and success has a probability p. Thus,

$$P(FFF....FS) = q^{x-1} \cdot p$$

Thus, the probability distribution of X is —

$$p(x) = \begin{cases} q^{x-1} \cdot p, & \text{for } x = 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

The mean and variance are given by —

$$E(X) = 1/p$$

and $V(X) = q/p^2$

The negative Binomial distribution indicates the number of trials required until k^{th} success. If Y is a random variate with negative binomial distribution, then.

$$p(y) = \begin{cases} {}^{y-1}C_{k-1} \cdot p^k \cdot q^{y-k}, & y = k, k+1, k+2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

As negative binomial random variable Y is sum of k independent geometric random variables, we can say that,

$$E(Y) = k/p \quad \text{and}$$

$$V(Y) = k \cdot q/p^2$$

Example:

At an inspection section, 40% of the assembled ink-jet printers are rejected. Find the probability that the first acceptable printer is the 3rd one inspected.

Solution: It is given that, the probability of rejection is 40%. That is,

$q = 0.4$

$\therefore p = 1 - q = 0.6$

We would like to find the probability of success at the 3rd trial. It indicates that, first two trials were rejected. With this information, we can say that the probability distribution is Geometric distribution.

So, $P(X=3) = q^{3-1} \cdot p$

$= (0.4)^2 \cdot (0.6)$

$= 0.096$

Now, assume we need to compute the probability that the 5th inspected printer is the second acceptable printer. That is, out of first 5 trials, 5th trial is a success and out of previous 4 trials, one trial was success. Then, to compute this probability, Negative binomial distribution.

$P(X=5) = {}^{5-1}C_{2-1} \, (0.4)^{5-2} \cdot (0.6)^2$

$= 0.1382$

(iv) Poisson Distribution

A discrete random variable x is said to follow Poisson distribution, if it has pmf as -

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

Here, $\alpha > 0$.

For poisson random variate, both mean and variance are same. That is -

$$E(X) = V(X) = \alpha$$

The CDF of poisson variate is given by -

$$F(x) = \sum_{i=0}^{x} \frac{e^{-\alpha} \alpha^i}{i!}$$

Example:

A computer repair person is beeped each time there is a call for service. The no of beeps per hour follows Poisson distribution with mean = 2 per hour. Compute the probability that -

(i) there will be 3 beeps in the next hour.

(ii) there will be 2 or more beeps in 1 hour.

Solution: The random variate X indicating no of beeps per hour has poisson distribution with $\alpha = 2$.

(i) We need to compute the probability of getting 3 beeps.

$$\therefore P(x=3) = \frac{e^{-2}2^3}{3!}$$

$$= \frac{0.1353 \times 8}{6}$$

$$= 0.18$$

NOTE: The solution for above problem can be found even using Poisson CDF and by referring an existing Poisson CDF Table. (It is given at the end of prescribed text book. Also, can be found in internet).

So, using CDF, it would be -

$$F(3) - F(2) = 0.857 - 0.677$$

$$= 0.18$$

Here, $F(3)$ indicates 'at the most 3 beeps' and $F(2)$ indicates 'at the most 2 beeps'.

(ii) Probability of 2 or more beeps in one hour. ie $P(x \geq 2)$ has to be computed.

But, $P(x \geq 2) = 1 - P(x < 2)$

$$= 1 - F(1)$$

$$= 1 - 0.406 \quad \text{(using Poisson table)}$$

$$= 0.594$$

Table A.4 Cumulative Poisson Distribution

	α = Mean											
x	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	x
0	.990	.951	.905	.819	.741	.670	.607	.549	.497	.449	.407	0
1	1.000	.999	.995	.982	.963	.938	.910	.878	.844	.809	.772	1
2		1.000	1.000	.999	.996	.992	.986	.977	.966	.953	.937	2
3				1.000	1.000	.999	.998	.997	.994	.991	.987	3
4						1.000	1.000	1.000	.999	.999	.998	4
5									1.000	1.000	1.000	5

	α = Mean											
x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	x
0	.368	.333	.301	.273	.247	.223	.202	.183	.165	.150	.135	0
1	.736	.699	.663	.627	.592	.558	.525	.493	.463	.434	.406	1
2	.920	.900	.879	.857	.833	.809	.783	.757	.731	.704	.677	2
3	.981	.974	.966	.957	.946	.934	.921	.907	.891	.875	.857	3
4	.996	.995	.992	.989	.986	.981	.976	.970	.964	.956	.947	4
5	.999	.999	.998	.998	.997	.996	.994	.992	.990	.987	.983	5
6	1.000	1.000	1.000	1.000	.999	.999	.999	.998	.997	.997	.995	6
7					1.000	1.000	1.000	1.000	.999	.999	.999	7
8									1.000	1.000	1.000	8

	α = Mean											
x	2.2	2.4	2.6	2.8	3.0	3.5	4.0	4.5	5.0	5.5	6.0	x
0	.111	.091	.074	.061	.050	.030	.018	.011	.007	.004	.002	0
1	.355	.308	.267	.231	.199	.136	.092	.061	.040	.027	.017	1
2	.623	.570	.518	.469	.423	.321	.238	.174	.125	.088	.062	2
3	.819	.779	.736	.692	.647	.537	.433	.342	.265	.202	.151	3
4	.928	.904	.877	.848	.815	.725	.629	.532	.440	.358	.285	4
5	.975	.964	.951	.935	.916	.858	.785	.703	.616	.529	.446	5
6	.993	.988	.983	.976	.966	.935	.889	.831	.762	.686	.606	6
7	.998	.997	.995	.992	.988	.973	.949	.913	.867	.809	.744	7
8	1.000	.999	.999	.998	.996	.990	.979	.960	.932	.894	.847	8
9		1.000	1.000	.999	.999	.997	.992	.983	.968	.946	.916	9
10				1.000	1.000	.999	.997	.993	.986	.975	.957	10
11						1.000	.999	.998	.995	.989	.980	11
12							1.000	.999	.998	.996	.991	12
13								1.000	.999	.998	.996	13
14									1.000	.999	.999	14
15										1.000	.999	15
16											1.000	16

continues...

Table A.4 Continued

x	α = Mean 6.5	7.0	7.5	8.0	9.0	10.0	12.0	14.0	16.0	18.0	20.0	x
0	.002	.001	.001									0
1	.011	.007	.005	.003	.001							1
2	.043	.030	.020	.014	.006	.003	.001					2
3	.112	.082	.059	.042	.021	.010	.002					3
4	.224	.173	.132	.100	.055	.029	.008	.002				4
5	.369	.301	.241	.191	.116	.067	.020	.006	.001			5
6	.527	.450	.378	.313	.207	.130	.046	.014	.004	.001		6
7	.673	.599	.525	.453	.324	.220	.090	.032	.010	.003	.001	7
8	.792	.729	.662	.593	.456	.333	.155	.062	.022	.007	.002	8
9	.877	.830	.776	.717	.587	.458	.242	.109	.043	.015	.005	9
10	.933	.901	.862	.816	.706	.583	.347	.176	.077	.030	.011	10
11	.966	.947	.921	.888	.803	.697	.462	.260	.127	.055	.021	11
12	.984	.973	.957	.936	.876	.792	.576	.358	.193	.092	.039	12
13	.993	.987	.978	.966	.926	.864	.682	.464	.275	.143	.066	13
14	.997	.994	.990	.983	.959	.917	.772	.570	.368	.208	.105	14
15	.999	.998	.995	.992	.978	.951	.844	.669	.467	.287	.157	15
16	1.000	.999	.998	.996	.989	.973	.899	.756	.566	.375	.221	16
17		1.000	.999	.998	.995	.986	.937	.827	.659	.469	.297	17
18			1.000	.999	.998	.993	.963	.883	.742	.562	.381	18
19				1.000	.999	.997	.979	.923	.812	.651	.470	19
20					1.000	.998	.988	.952	.868	.731	.559	20
21						.999	.994	.971	.911	.799	.644	21
22						1.000	.997	.983	.942	.855	.721	22
23							.999	.991	.963	.899	.787	23
24							.999	.995	.978	.932	.843	24
25							1.000	.997	.987	.955	.888	25
26								.999	.993	.972	.922	26
27								.999	.996	.983	.948	27
28								1.000	.998	.990	.966	28
29									.999	.994	.978	29
30									.999	.997	.987	30
31									1.000	.998	.992	31
32										.999	.995	32
33										1.000	.997	33
34											.999	34
35											.999	35
36											1.000	36