System Modeling and Simulation

Onte Sl no.	Topic	Details
1	*Introduction to Simulation*	*1.1 Advantages and Disadvantages of Simulation* *1.2 When Simulation is the Appropriate Tool* *1.3 When Simulation is not Appropriate* *1.4 Areas of Application* *1.5 Systems and System Environment* *1.6 Components of a System* *1.7 Discrete and Continuous Systems* *1.8 Model of a System* *1.9 Steps in a Simulation Study*
2	*Simulation of Queuing and Inventory Systems*	*2.1 Introduction to Queuing Systems* *2.2 Characteristics of Queuing Systems* *2.3 Queuing Notation* *2.4 Simulation of Queuing Systems* *2.5 Simulation of Inventory Systems*
3	*Statistical Models*	*3.1 Review of Terminology and Concepts* *3.2 Useful Statistical Models* *3.3 Discrete Distributions* *3.4 Continuous Distributions* *3.5 Poisson Process* *3.6 Empirical Distributions*
4	*Random-Number Generation*	*4.1 Random Numbers* *4.2 Techniques for Generating Random Numbers* *4.3 Tests for Random Numbers*
5	*Random Variate Generation*	*5.1 Inverse Transform Technique* *5.2 Acceptance-Rejection Technique*
6	*Input Modeling*	*6.1 Data Collection* *6.2 Identifying the Distribution with Data* *6.3 Parameter Estimation* *6.4 Goodness - of - Fit Tests* *6.5 Selecting Input Models without Data* *6.6 Multivariate and Time-series Input Models*
7	*Verification and Validation of Simulation Models*	*7.1 Model building, Verification of models* *7.2 Verification of Simulation Models* *7.3 Calibration and Validation of Models*
8	*Appendix*	*Tables*

Unit 1
Introduction to Simulation

Simulation is the imitation of the operation of a real-world process or system over time. Simulation involves the generation of an artificial history of a system, and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system that is represented.

Simulation is an indispensable problem-solving methodology for the solution of many real-world problems. Simulation is used to describe and analyze the behavior of a system and aid in the design of real systems. Both existing and conceptual systems can be modeled with simulation.

1.1 Advantages and Disadvantages of Simulation

Simulation has several advantages and some disadvantages in various applications of which few are listed by Pegden, Shannon and Sadowski [1995].

The advantages are:

1. New policies, operating procedures, information flows and so on can be explored without distracting ongoing operations of the real system.
2. New hardware designs, physical layouts, transportation systems etc can be tested without assigning resources for their acquisition.
3. Time can be compressed or expanded allowing for a speed-up or slow-down of the phenomenon (clock is self-control).
4. Hypotheses about how and why certain event occurs can be tested for feasibility.
5. Insight can be obtained about interaction and importance of variables to the performance.

6. Bottleneck (critical) analysis can be performed to discover where work in process, materials, information and so on is delayed.
7. A simulation study helps in understanding how the system operates.
8. "What if" questions can be answered, that is useful for designing new systems.

The disadvantages are:

1. Special training is required for Model building. It is an art that is learnt overtime and through experience.
2. Simulation results can be difficult to interpret, as most simulation outputs are basically random variables (random inputs).
3. Simulation modeling and analysis can be time consuming and expensive.

1.2 When Simulation is the Appropriate Tool

The situations, under which Simulation is the appropriate (apt) tool to use, have been discussed by many authors from Naylor to Banks [1996]. Simulation can be used for the following purposes

1. A simulation model helps us to gain knowledge and suggest improvements in the system under investigation.
2. Simulation enables the study and experimentation with the internal interactions of a complex system or subsystem within a complex system.
3. Informational, organizational and environmental changes can be simulated and their effects can be observed.
4. By changing simulation inputs and observing resulting outputs, valuable insight may be obtained into which variables are important and how variables interact.
5. Simulation can be used as a pedagogical (rules related to educational measures) device to reinforce analytical solution methodologies.
6. Simulation can be used with new design and policies before implementation.
7. Simulating different capabilities for a machine can help to determine the requirements.
8. Simulation models designed for training, makes learning possible without the cost and disruption of on-the-job learning.
9. A plan can be visualized with animated simulation.
10. The modern system (factory, wafer fabrication plant, service organization) is too complex that its internal interaction can be treated only by simulation.

1.3 When Simulation is not Appropriate

Banks and Gibson [1997] gave ten rules for determining when simulation is not appropriate.

1. Simulation should not be used when the problem can be solved using common sense. For example if customers arrive randomly at an average rate of 50/hour and

are served at a mean rate of 10/hour, then to determine minimum number of servers, simulation is not required. Just compute 50/10=5 servers.

2. Simulation should not be used if the problem can be solved analytically.
3. Simulation should not be used if it is easier to perform direct experiments.
4. Not to use simulation, if the costs exceed the savings.

5 & 6 .Simulation should not be performed if the resources or time are not available.

7. If no data is available, not even estimates then simulation is not advised as it requires data, sometimes lots of data.

8 This rule is concerned with the ability to verify and validate the model. If there is not enough time or personnel not available, simulation is not appropriate.

9. If mangers have unreasonable expectation –say, too much too soon- or power of simulation is overestimated, then simulation may not be appropriate.
10. If system behavior is too complex or can't be defined then it is not appropriate.

1.4 Areas of Application

- Manufacturing Applications
- Semiconductor Manufacturing
- Construction Engineering and project management
- Military application
- Logistics, Supply chain and distribution application
- Transportation modes and Traffic
- Business Process Simulation
- Health Care
- Automated Material Handling System (AMHS)

1.5 Systems and System Environment

System

A system is defined as an aggregation or assemblage of objects joined in some regular interaction or interdependence.

System Environment

A system is often affected by the changes occurring outside the system. Such changes are said to occur in system environment.

Example: Factory system that makes and assembles parts into a product (Fig 1.1)

.

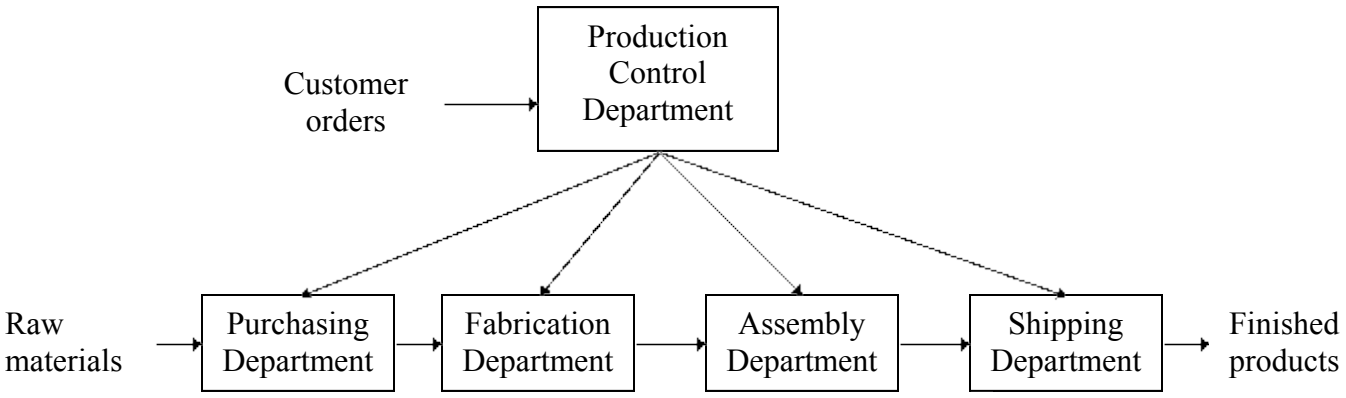


Fig 1.1 A factory system

The factors controlling the arrival of orders may be considered to be outside the influence of the factory and therefore part of the environment. If the demand grows in the market, production increases in the factory. Hence market is system environment.

1.6 Components of a System

Entity - An object of interest in the system.
Attribute -The property of an entity.
Activity - A time period of specified length.
State - A collection of variables that describe the system in any time.
Event - An instantaneous occurrence that might change the state of the system.

The table 1.1 lists few examples for the above mentioned components of a system.

System	Entities	Attributes	Activities	Events	State variables
Bank	Customers	Balance , Credit status	Depositing, withdrawal	Arrival, Departure	No. of busy tellers, No. of customers waiting
Production	Machines	Speed, Capacity	Welding, stamping	Breakdown	Status of machine (busy, idle or down)
Communication	Messages	Length, Destination	Transmitting	Arrival at destination	Number waiting to be transmitted

Table 1.1 Examples of Systems and Components

Other terminologies

Endogenous - Activities and events occurring with the system
Exogenous - Activities and events occurring with the environment
Open System- A system for which there is exogenous activity.
Closed System- A system for which there is no exogenous activity.

1.7 Discrete and Continuous Systems

A **discrete system** is one in which the state variable(s) change only at a discrete set of points in time.
Example: Banking system- The number of customers (state variable) in the bank, changes only when a customer arrives or the service provided to a customer is completed. The fig 1.2 shows the customers changing only at discrete points of time.

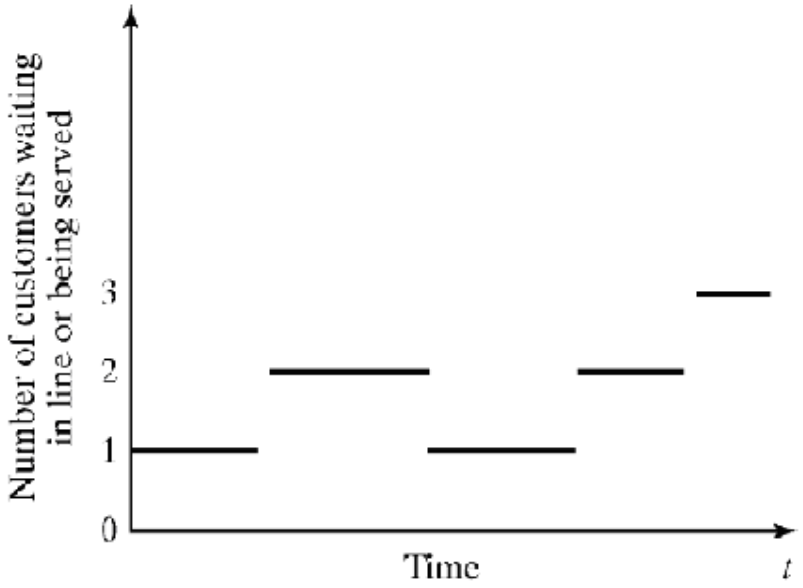


Fig 1.2 Discrete Systems

A **continuous system** is one in which the state variables change continuously over time.
Example: Head of water behind a dam. After a rain storm, water flows into the lake behind the dam. Water is drawn from dam for flood control and to generate electricity. Fig 1.3 shows how the head of water behind the dam (state variable) changes.

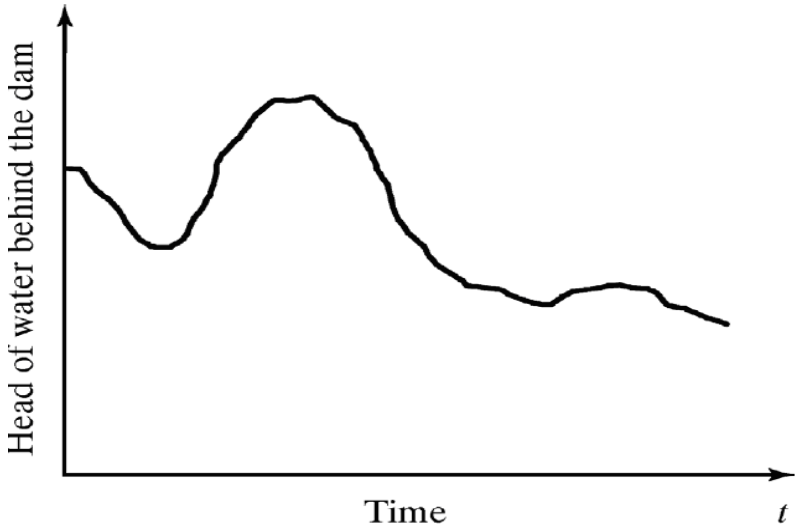


Fig 1.3 Continuous Systems

1.8 Model of a System

A **Model** is defined as a representation of a system for the purpose of studying the system. In simple words it is also defined as *simplification of reality*. Model constructs a conceptual framework that describes a system. It is necessary to consider those aspects of systems that affect the problem under investigation (unnecessary details to be removed).

1.8.1 Types of models

Models can be classified as Physical and Mathematical Models.

Physical Model is a smaller or larger physical copy of an object. Physical models allow visualization, from examining the model of information about the thing the model represents. A model can be a physical object such as an architectural model of a building.

Mathematical Model uses symbolic notation and mathematical equations to represent a system. It is further classified as static and dynamic models.

Static Mathematical Model gives the relationship between the system attributes when the system is in equilibrium. For example in market model there is balance between supply and demand for commodity and both factors depend upon price.

Dynamic Mathematical Model allows the change of system attributes to be derived as a function of time. The derivation may be with analytical solution or by numerical computation.

Simulation model is a particular type of mathematical model of a system. It is defined as a mathematical representation of the essential characteristics of a real-world system or situation, which can be used to predict future behavior under a variety of different conditions.

The process of developing a simulation model involves defining the situation or system to be analyzed, identifying the associated variables, and describing the relationships between them as accurately as possible

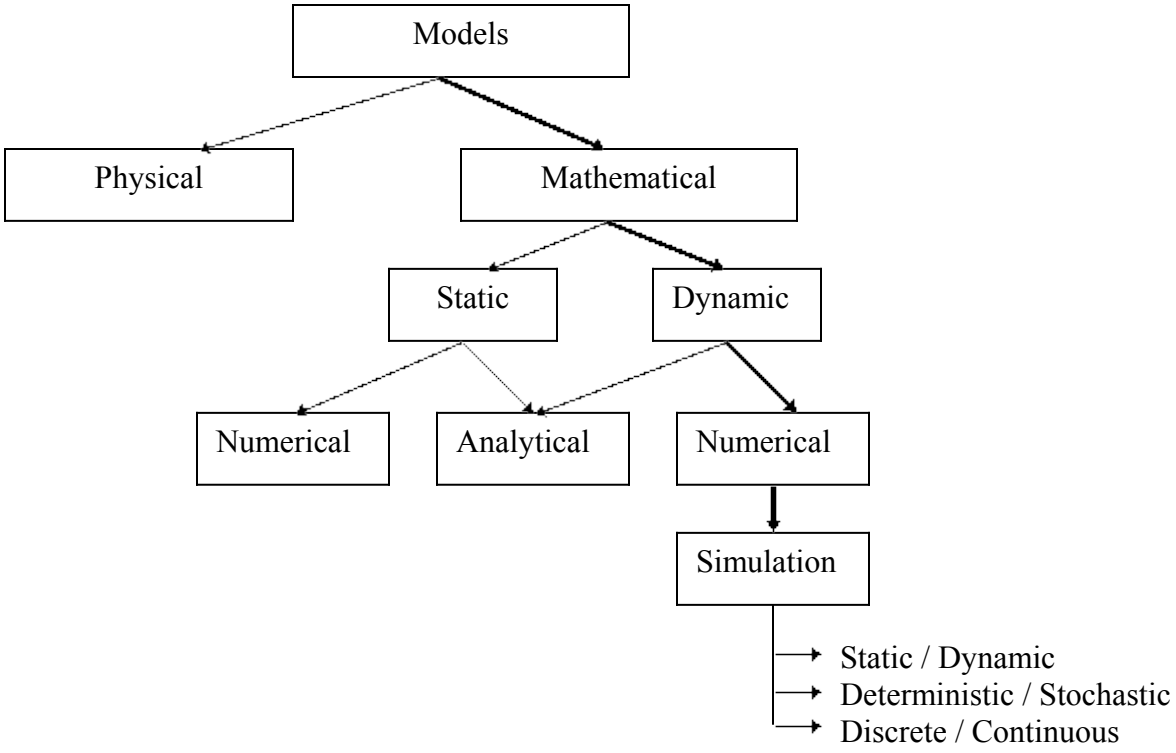


Fig 1.4 Types of Models

1.8.2 Characterizing a simulation model

1. Static or Dynamic Simulation model

A static simulation model also called as Monte Carlo simulation represents a system at particular point in time.

Dynamic simulation model represent systems as they change over time.

2. Deterministic or Stochastic Simulation model

Deterministic models do not contain random variables. They have a known set of inputs resulting in a unique set of outputs.

Example- Classes conducted in college at particular time.

Stochastic models have one or more random variables as inputs. These inputs lead to random outputs.

Example- Arrival of customers to bank.

3. Continuous or Discrete Simulation model

Continuous Simulation model, the state of system changes at all point of time.

Discrete Simulation model, the state of system changes only at discrete set of time.

1.9 Steps in a Simulation Study

The set of steps in simulation study helps a model builder in thorough understanding of the system. The fig 1.5 refers to the detailed steps in study of simulation.

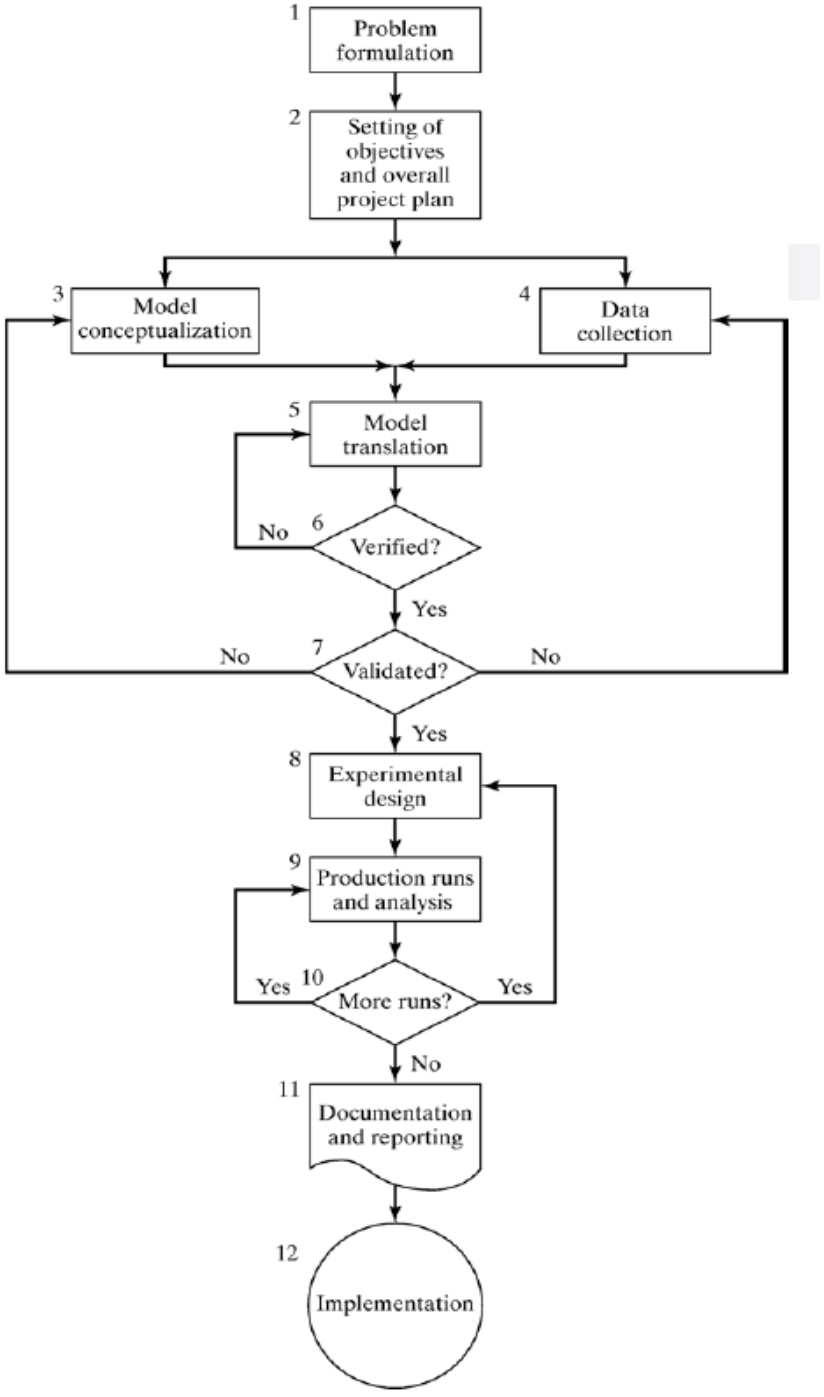


Fig 1.5 Steps in a simulation study

Problem formulation

The study begins with defining the problem statement. It can be developed either by the analyst or client. If the statement is provided by client, then the analyst must take extreme care to ensure that the problem is clearly understood. If a problem statement is prepared by the simulation analyst, it is important that the client understand and agree with the formulation. Even with all of these precautions, it is possible that the problem will need to be reformulated as the simulation study progresses.

Setting of objectives and overall project plan
Another way to state this step is "prepare a proposal." The objectives indicate the questions to be answered by the simulation study. Whether the simulation is appropriate or not is to be decided at this stage. The overall project plan should include a statement of the alternative systems and a method for evaluating the effectiveness of these alternatives. The plan includes number of personnel, number of days to complete the task, stages in the investigation, output at each stage, cost of the study and billing procedures, if any.

Model conceptualization
Model is a simplification of reality. The real-world system under investigation is abstracted by a conceptual model. It is recommended that modeling begins with simple model and grows until a model of appropriate complexity has been achieved. For example, consider the model of a manufacturing and material handling system. The basic model with the arrivals, queues and servers is constructed. Then, add the failures and shift schedules. Next, add the material-handling capabilities. Finally, add the special features. Constructing an excessive complex model will add to the cost of the study and the time for its completion, without increasing the quality of the output. Maintaining client involvement will enhance the quality of the resulting model and increase the client's confidence in its use.

Data collection
This step involves in gathering the desired input data. The data changes over the complexity of model. Data collection takes a huge amount of total time required to perform simulation. It should be started at early stages together with model building. The collection of data should be relevant with the objectives of study.

Model translation
The conceptual model constructed in Step 3 is coded into a computer recognizable form, an operational model. The suitable simulation language is used.

Verified?
Verification is with respect to the operational model. Is it performing properly? If the input parameters and logical structure of model are correctly represented in computer, then verification is completed.

Validated?
Validation is the determination, that the model is an accurate representation of the real system. This is done by calibration of model –an iterative process of comparing model to the actual system behavior. This process is repeated until model accuracy is acceptable.

Experimental design
The alternatives to be simulated must be determined. For each scenario that is to be simulated, decisions need to be made concerning the length of the simulation run, the number of runs (also called replications), and the length of initialization period.

Production runs and analysis
Production runs, and their subsequent analysis, are used to estimate measures of performance for the system design that are being simulated.

More runs?
After the completion on the analysis of runs, the simulation analyst determines if additional runs are needed and any additional experiments should follow.

Documentation and reporting
There are two types of Documentation: Program and Progress. Program documentation is necessary for numerous reasons. If the program is going to be used again by the same or different analysts, it may be necessary to understand how the program operates. This will enable confidence in the program so that the client can make decisions based on the analysis. Also, if the model is to be modified, this can be greatly facilitated by adequate documentation. Progress reports provide a chronology of work done and decisions made. It is the written history of a simulation project. The result of all the analysis should be reported clearly and concisely. This will enable the client to review the final formulation.

Implementation
If the client has been involved throughout the study period, and the simulation analyst has followed all of the steps rigorously, then the likelihood of a successful implementation is increased.

The simulation model building process shown in the fig 1.5 can be divided to four phases
Phase 1-Problem formulation
Setting of objectives and overall design
Phase 2- Model conceptualization
Data collection
Model translation
Verification
Validation
Phase 3- Experimental design
Production runs and analysis
Additional runs
Phase 4- Documentation and reporting
Implementation

MONTE CARLO METHOD

A Monte Carlo method is a stochastic technique that involves use of random numbers and probability statistics to solve the problems. The term Monte Carlo Method was coined by S. Ulam and Nicholas Metropolis in reference to games of chance, a popular attraction in Monte Carlo, Monaco (Hoffman, 1998; Metropolis and Ulam, 1949). This method can be used in many areas from economics, nuclear physics to regulating the flow of traffic. To call something a "Monte Carlo" experiment, all you need to do is use random numbers to examine some problem.

The Monte Carlo method is just one of many methods for analyzing uncertainty propagation, where the goal is to determine how random variation, lack of knowledge, or error affects the sensitivity, performance, or reliability of the system that is being modeled.

Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. This method is often used when the model is complex, nonlinear, or involves more than just a couple of uncertain parameters. A simulation can typically involve over 10,000 evaluations of the model, a task which in the past was only practical using super computers.

Unit 2

Simulation of Queuing and Inventory Systems

2.1 Introduction to Queuing Systems

Simulation is often used in the analysis of queuing models. In a simple but typical queuing model, shown in Fig 2.1, customers arrive from time to time and join a queue or waiting line, are eventually served and finally leave the system.

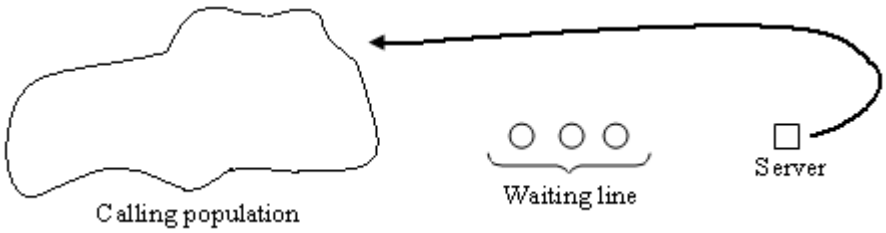


Fig 2.1 Queuing system

The term "customer" refers to any type of entity that can be viewed as requesting "service" from the system. Therefore many service facilities like production systems, repair and maintenance facilities, communications and computer systems and transport and material handling systems can be viewed as queuing systems. Typical measures of system performance include server utilization (percentage of time a server is busy), length of waiting lines and delays of customers.

2.2 Characteristics of Queueing Systems

The key elements of a queuing system are the *customers* and *servers*.

The term *customer* can refer to people, machines, trucks, mechanics, patients, airplanes, e-mail, cases, orders, or dirty clothes-anything that arrives at a facility and requires service.

The term *server* might refer to receptionists, mechanics, tool-crib clerks, medical personnel, automatic storage and retrieval machines (e.g., cranes), runaways at an airport, automatic packers, etc which provides the requested service.

2.2.1 The calling population

The population of potential customers, referred to as the *calling population*, may be assumed to be *finite* or *infinite*.

Finite calling population model
The arrival rate to the queuing system does depend on the number of customers being served and waiting.

A more typical example is that of five tire-curing machines serviced by a single worker. The machines are the "customers" who arrive at the instant they automatically open. The worker is the "server", who "serves" an open machine as soon as possible. When all five are closed and instant a machine opens and requires a service, the arrival rate decreases. At those times when all five are open (so four machines are waiting for service while the worker is attending the other one), the arrival rate is zero; that is, no arrival is possible until the worker finishes with a machine, in which case it returns to the calling population and becomes a potential arrival. But if arrival rate is defined as the expected number of arrivals in the next unit of time, then it becomes clear that this expectation is largest when all machines could potentially open in the next unit of time.

Infinite calling population model
The arrival rate (i.e., the average number of arrivals per unit time) is not affected by the number of customers who have left the calling population and joined the queuing system. When the arrival process is homogeneous over time (e.g., there are no "rush hours"), the arrival rate is usually assumed to be constant.

Examples of *infinite population* include the potential customers of the restaurant, bank, or other similar service facility and also very large group of machines serviced by a technician. In systems with large population of potential customers, the calling population is usually assumed to be infinite.

2.2.2 System capacity

In many queuing systems there is a limit to the number of customers that may be in the waiting line or system. When a system has limited capacity, a distinction is made between the arrival rate (i.e., the number of arrivals per time unit) and the effective rate (i.e., the number who arrive and enter the system per time unit).

For example,
1. *(Limited capacity)* - An automatic car wash may have room only for 10 cars to enter the mechanism. It may be too dangerous or illegal for cars to wait in the street. An arriving customer who finds the system full does not enter but returns immediately to the calling population.
2. *(Unlimited capacity)* - Some systems, such as concert ticket sales for students, may be considered as having unlimited capacity. There are no limits on the number of students allowed to wait to purchase tickets.

2.2.3. The arrival process

The arrival process for infinite population models is usually characterized in terms of interarrival times of successive customers. Arrivals may occur at scheduled times or random times.

- The most important model for *random arrivals* is the Poisson arrival process. Poisson arrival process is used as a model for the arrival of people to restaurants, driving banks and other service facilities like the arrival of telephone calls to a telephone exchange, etc.

- Second important class of intervals is the *scheduled arrivals*. In this case the inter arrival times may be constant, or constant plus or minus a small random amount to represent early or late arrivals. For example, Patients to a physicians office, scheduled airline flight arrivals to an airport.

- A third situation occurs when at least one customer is assumed to be always present in the queue so that the server is never idle because of lack of customers. For example, a customer may represent raw material for a product and sufficient raw material is assumed to be always available.

For finite population models the arrival process is characterized in a different manner. We define a customer as pending, when that customer is outside the queuing system and a member of calling population; a run time of a given customer is the length of time from departure from the queuing system until that customers next arrival to the queue. Let

$A_1^{(i)}$, $A_2^{(i)}$,… be the successive run times of customers *i* and let $S_1^{(i)}$, $S_2^{(i)}$,… be the corresponding successive system times i.e. $S_n^{(i)}$ is the total time spent in the system by customer *i* during the n^{th} visit. For example, a tire curing machine is pending when it is closed and curing a tire. It becomes not pending, the instant it opens and demands service from the worker. The following fig 2.2 illustrates these concepts for machine 3.

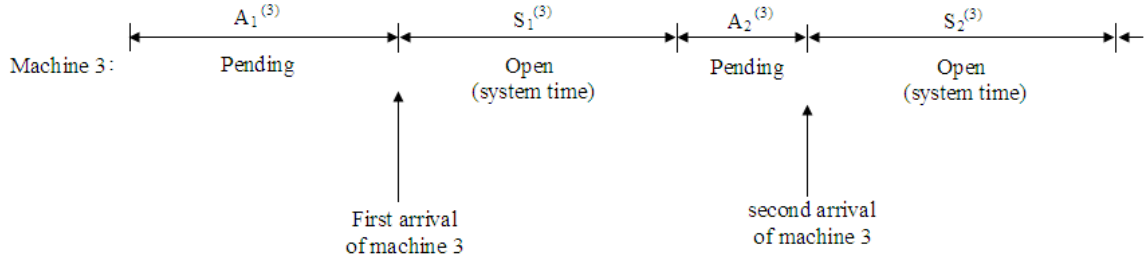


Fig 2.2 Arrival process for a finite population model

Suppose, if it is assumed that all machines are pending at time 0, the first arrival to the system occurs at time $A_1 = \min \{A_1^{(1)}, A_1^{(2)}, A_1^{(3)}, A_1^{(4)}, A_1^{(5)}\}$. If $A_1 = A_1^{(2)}$, then machine 2 is the first arrival (i.e., the first to open) after time 0. Here, the arrival rate is not constant but is a function of the number of pending customers.

2.2.4 Queue behavior and queue discipline

Queue behavior refers to customer actions while in a queue waiting for service to begin. There is a possibility that the incoming customers may

- *Balk*: leave when they see that the line is too long.
- *Renege*: leave after being in the line when they see that the line is moving too slow.
- *Jockey*: move from one line to another if they think they have chosen a slow line.

Queue discipline refers to the logical ordering of customers in a queue and determines which customer is chosen for service when the server becomes free.
Queue disciplines can be:

- FIFO (first in, first out)
- LIFO (last in, first out)
- SIRO (service in random order)
- SPT (shortest processing time first)
- PR (priority service)

In a job shop, queue disciplines are some times based on due dates and expected processing time for a given type of job.

2.2.5 Service times and the service mechanism

The service times of successive arrivals are denoted by S_1, S_2, and S_3… They may be constant or of random duration. In case of random, $\{S_1, S_2, \ldots\}$ is usually characterized as a sequence of independent and identically distributed random variables.

The distributions like exponential, weibull, gamma, etc can be used as models for service times. Sometimes services may be identically distributed for all customers of a given type or class or priority, while customers of different types may have different service time distributions. Sometimes, service times depend upon the time of day or length of waiting line.

A queuing system consists of a number of service centers and interconnecting queues. Each service center consists some number of severs *c,* working in parallel. Parallel service mechanisms are either single server (c=1), multiple servers ($1 < c < \infty$) or unlimited servers ($c=\infty$). A self service facility is usually characterized as having an unlimited number of servers.

Example 2.1
Consider a discount warehouse where customers may either serve themselves or wait for one of the 3 clerks and finally leave after paying a single cashier. The system is represented by the flow diagram in fig 2.3.The sub system consisting of queue 2 and service center 2 is shown in more detail in the fig 2.4.

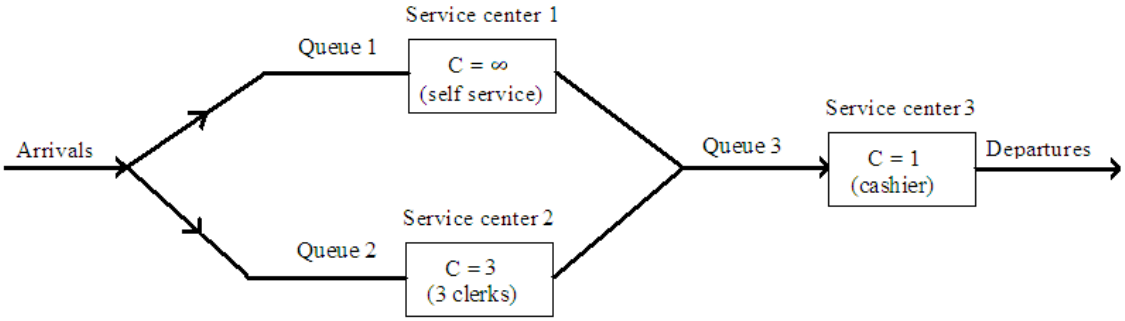


Fig 2.3 Discount warehouse with three service centers

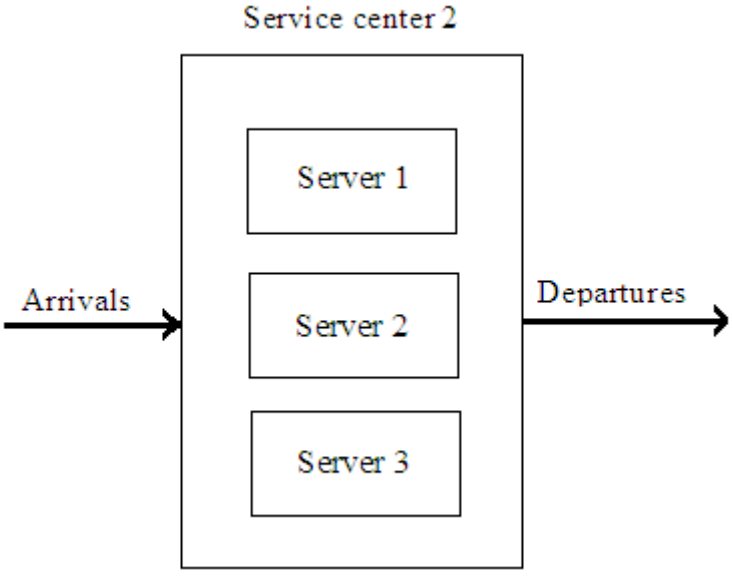


Fig 2.4 Service center 2, with c=3 parallel servers

2.3 Queuing Notation

The notation is based on the format A / B / c / N / K.
A → Interarrival-time distribution
B→ Service time distribution
[Common symbols for A and B include M (exponential or Markow), D (constant or deterministic), E_k (Erlang of order k), and GI (general independent).]
c→ Number of parallel servers
N→ System capacity
K→ Size of the calling population

For example, M / M / 1 / ∞ / ∞ are often shortened to M / M / 1. The tire curing system can be initially represented by G / G / 1 / 5 / 5.

2.4 Simulation of Queuing Systems

In a single channel queue the calling population is infinite. Arrivals for service occur one at a time in a random fashion; once they join the waiting line, they are eventually served. The system capacity has no limit, the units are served in the order of their arrival usually in FIFO for a single server or channel. Arrivals and services are defined by the distribution of the time between arrivals and the distribution of the service times respectively. Some concepts of queuing system are:

State of the system – The number of units in the system and the status of the server-busy or idle
Event – Set of circumstances that cause an instantaneous change in the system. There are only 2 possible events that can affect the state of the system. They are the entry of a unit in the system (the arrival event) or the completion of service on a unit (the departure unit).
Simulation Clock – used to track simulated time.

The queuing system includes the server, the unit being serviced and the units in the queue. If a unit has just completed the service the simulation proceeds as shown in the fig 2.5

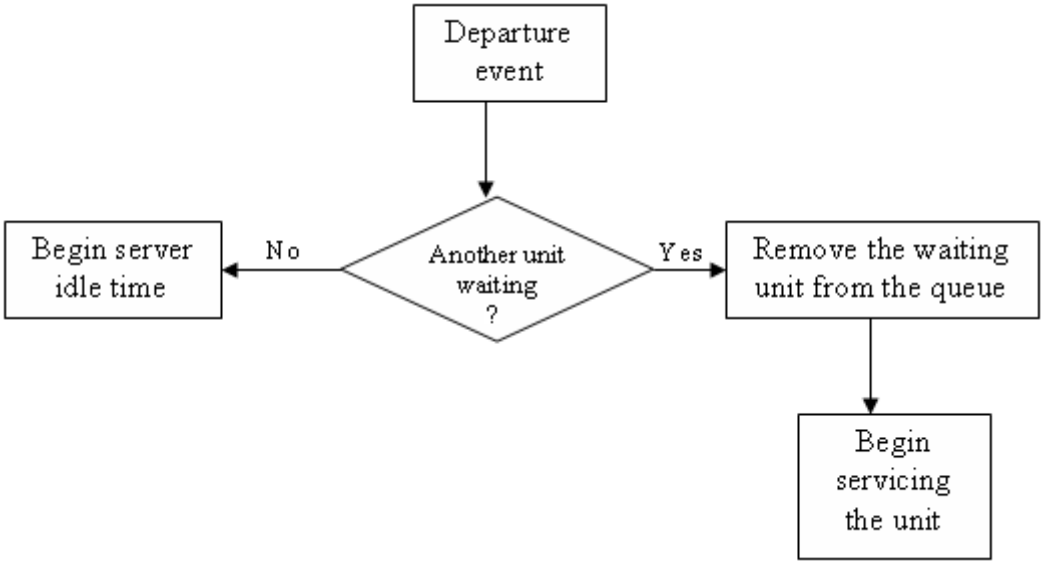


Fig 2.5 Service-just-completed flow diagram

The arrival event occurs when the unit enters the system. The flow diagram for the arrival event is shown in the fig 2.6.

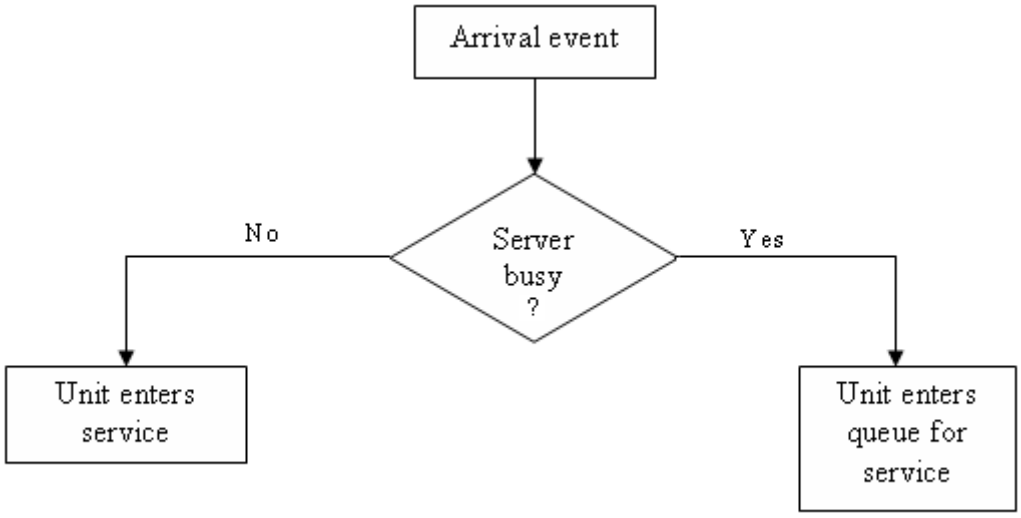


Fig 2.6 Unit-entering-system flow diagram

The unit may find the server either idle or busy; therefore it begins service immediately or enters the queue for the server. This course of action is shown in fig 2.7

		Queue status	
		Not empty	Empty
Server status	Busy	Enter queue	Enter queue
	Idle	Impossible	Enter service

Fig 2.7 Potential unit actions upon arrival

After the completion of the service the server may become idle or remain busy with the next unit. This relationship of the server outcomes to the status of the queue is shown in fig 2.8

		Queue status	
		Not empty	Empty
Server outcomes	Busy		Impossible
	Idle	Impossible	

Fig 2.8 Server outcomes after service completion

If the queue is not empty another unit will enter the server and it will be busy. If the queue is empty, the server will be idle after a service is completed. These two possibilities are shown in the shaded portion of the fig 2.8

Simulation clock times for arrivals and departures are computed in a simulation table customized for each problem. In simulation, events usually occur at random times. The randomness needed to imitate real life is made possible through the use of random numbers. Random numbers are distributed uniformly and independently and the interval (0, 1). Random digits are uniformly distributed on the set {0, 1, 2,…, 9}. Random digits can be used to form random numbers by selecting the proper number of digits for each random number and placing a decimal point to the left of the value selected.

Example 2.2 (Single channel queue simulation problem)

A small grocery store has only one check out counter. Customers arrive at this check out counter at random from 1-8 minutes apart. Each possible value of inter arrival time has the same probability of occurrence as shown in the table 2.1. The service times varies from 1-6 min with the probabilities shown in the table 2.2. The problem is to analyze the system by simulating the arrival and service of 20 customers.

Time between arrivals(minutes)	Probabili-ty	Cumulative probability	Random-digit assignment
1	0.125	0.125	001-125
2	0.125	0.250	126-250
3	0.125	0.375	251-375
4	0.125	0.500	376-500
5	0.125	0.625	501-625
6	0.125	0.750	626-750
7	0.125	0.875	751-875
8	0.125	1.000	876-000

Table 2.1 Distribution of time between arrivals

Service time (minutes)	Probabili-ty	Cumulative probability	Random-digit assignment
1	0.10	0.10	01-10
2	0.20	0.30	11-30
3	0.30	0.60	31-60
4	0.25	0.85	61-85
5	0.10	0.95	86-95
6	0.05	1.00	96-00

Table 2.2 Service time distributions

- Obtain the random digits from the table of random digits (refer appendix). Since the probabilities in the table 2.1 are accurate to 3 significant digits, three-place random numbers will suffice.
- Similarly for table 2.2 two-place random numbers will suffice. It is necessary to list only 19 random numbers to generate times between arrivals (for 20 customers) as the first arrival is assumed to occur at time 0.
- The third column in tables 2.1 and 2.2 contains cumulative probability for the distribution.
- The rightmost column contains the random digit assignment in table 2.1; the first random-digit assignment is 001-125. There are 1000 three digit values possible (001 through 000). The probability of time-between-arrivals of 1 minute is 0.125 and 125 of the 1000 random digit values are assigned to such an occurrence.
- In the table 2.3, the time between arrivals is determined. The first random digits are 913. To obtain the corresponding time between arrivals, find in fourth column of table 2.1 the assignment in which 913 lies and read 8 minutes from the first column of the table.
- Similarly service times for all 20 customers are generated in table 2.4 with the aid of table 2.2.

Customer	Random digits	Time between arrivals (minutes)	Customer	Random digits	Time between arrivals (minutes)
1	-	-	11	109	1
2	913	8	12	093	1
3	727	6	13	607	5
4	015	1	14	738	6
5	948	8	15	359	3
6	309	3	16	888	8
7	922	8	17	106	1
8	753	7	18	212	2
9	235	2	19	493	4
10	302	3	20	535	5

Table 2.3 Time–between–Arrivals determination

Customer	Random digits	Service time (minutes)	Customer	Random digits	Service time (minutes)
1	84	4	11	32	3
2	10	1	12	94	5
3	74	4	13	79	4
4	53	3	14	05	1
5	17	2	15	79	5
6	79	4	16	84	4

7	91	5	17	52	3
8	67	4	18	55	3
9	89	5	19	30	2
10	38	3	20	50	3

Table 2.4 Service times generated

The simulation table for a single channel queue is shown in table 2.5. The first customer is assumed to arrive at time 0. Service begins immediately and finishes at time 4. The customer was in the system for 4 minutes.

After the first customer, the subsequent rows in the table are based on the random numbers for interarrival time and service time and the completion time of the previous customer. For example, the second customer arrives at time 8 and thus the server was idle for 4 min.

Skipping down to the fourth customer, it is seen that this customer arrived at time 15 but could not be served until time 18. This customer had to wait in the queue for 3 minutes. This process continues for all 20 customers.

A Customer	B Time since last arrival(mins)	C Arrival time	D Service time (mins)	E Time service begins	F Time customer waits in queue (mins)	G Time service ends	H Time customer spends in system (mins)	I Idle time of server (mins)
1	-	0	4	0	0	4	4	0
2	8	8	1	8	0	9	1	4
3	6	14	4	14	0	18	4	5
4	1	15	3	18	3	21	6	0
5	8	23	2	23	0	25	2	2
6	3	26	4	26	0	30	4	1
7	8	34	5	34	0	39	5	4
8	7	41	4	41	0	45	4	2
9	2	43	5	45	2	50	7	0
10	3	46	3	50	4	53	7	0
11	1	47	3	53	6	56	9	0
12	1	48	5	56	8	61	13	0
13	5	53	4	61	8	65	12	0
14	6	59	1	65	6	66	7	0
15	3	62	5	66	4	71	9	0
16	8	70	4	71	1	75	5	0
17	1	71	3	75	4	78	7	0
18	2	73	3	78	5	81	8	0
19	4	77	2	81	4	83	6	0
20	5	82	3	83	1	86	4	0
			68		56		124	18

Table 2.5 Simulation table for queueing problem

The following are the statistics

1. The average waiting time for a customer is 2.8 minutes. This is determined by

$$\text{Average waiting time (minutes)} = \frac{\text{total time customers wait in queue (minutes)}}{\text{Total numbers of customers}}$$

$$= 56 / 20 = 2.8 \text{ minutes}$$

2. The probability that a customer has to wait in the queue is 0.65. It is obtained by

$$\text{Probability (wait)} = \frac{\text{number of customers who wait}}{\text{total number of customers}}$$

$$= 13 / 20 = 0.65$$

3. The fraction of idle time of the server is 0.21, determined as follows

$$\text{Probability of idle server} = \frac{\text{total idle time of server (minutes)}}{\text{total run time of simulation (minutes)}}$$

$$= 18 / 86 = 0.21$$

The probability of the server being busy is the complement of 0.21 = 0.79

4. The average service time is 3.4 minutes, determined as

$$\text{Average service time (minutes)} = \frac{\text{total service time (minutes)}}{\text{total number of customers}}$$

$$= 68 / 20 = 3.4 \text{ minutes}$$

The result can be compared with the expected service time by finding the mean of the service time distribution using the equation

$$E(S) = \sum_{s=0}^{\infty} sp(s)$$

$$= 1(0.10) + 2(0.20) + 3(0.30) + 4(0.25) + 5(0.10) + 6(0.05) = 3.2 \text{ minutes}$$

The expected service time is slightly lower than the average service time in the simulation. The longer the simulation, the closer the average will be to E(S).

5. The average time between arrivals is 4.3 minutes, determined as

$$\text{Average time between arrivals(minutes)} = \frac{\text{Sum of all times between arrivals(minutes)}}{\text{number of arrivals} - 1}$$

= 82 / 19 = 4.3 minutes

One is subtracted from the denominator because the first arrival is assumed to occur at time 0.

The mean is given by

$$E(A) = \frac{a + b}{2}$$

= (1 + 8) / 2 = 4.5 minutes

The expected time between arrivals is slightly higher than the average. As the simulation becomes longer, the average value of time between arrivals will approach the theoretical mean E(A).

6. The average waiting time of those who wait is 4.3 minutes. This is obtained by

$$\text{Average waiting time of those who wait (minutes)} = \frac{\text{total time customers wait in queue(minutes)}}{\text{total number of customers who wait}}$$

= 56 / 13 = 4.3 minutes

7. The average time a customer spends in the system is 6.2 minutes. This can be determined in two ways.

First,

$$\text{Average time customer spends in the system (minutes)} = \frac{\text{total time customer spends in the system (minutes)}}{\text{total number of customers}}$$

=124 / 20 = 6.2 minutes

Second,

$$\text{Average time customer spends in the system (minutes)} = \text{Average time customer spends waiting in queue (minutes)} + \text{Average time customer spends in service (minutes)}$$

=2.8 + 3.4 = 6.2 minutes

Example 2.3 (The Able Baker Carhop problem)

Consider a drive-in restaurant where carhops take orders and bring food to the car. Cars arrive in the manner shown in the table 2.6. There are two carhops- Able and Baker. Able is better able to do the job and works a bit faster than Baker. The distribution of their service times is shown in tables 2.7 and 2.8.

This is an example for more than one service channels (2 servers). A simplifying rule is that Able gets the customer if both carhops are idle. Able has the seniority.

Time between arrivals (Minutes)	Probability	Cumulative probability	Random-Digit Assignment
1	0.25	0.25	01-25
2	0.40	0.65	26-65
3	0.20	0.85	66-85
4	0.15	1.00	86-00

Table2.6 Inter arrival Distribution of Cars

Service Time (Minutes)	Probability	Cumulative Probability	Random-Digit Assignment
2	0.30	0.30	01-30
3	0.28	0.58	31-58
4	0.25	0.83	59-83
5	0.17	1.00	84-00

Table2.7 Service Distribution of Able

Service Time (Minutes)	Probability	Cumulative Probability	Random-Digit Assignment
3	0.35	0.35	01-35
4	0.25	0.60	36-60
5	0.20	0.80	61-80
6	0.20	1.00	81-00

Table2.8 Service Distribution of Baker

The problem is to find how well the current arrangement is working. A simulation of 60 minutes of operation is made to estimate the system performance.

Here there are more events:

- A customer arrives
- A customer begins service from Able
- A customer completes service from Able
- A customer begins service from Baker
- A customer completes service from Baker

The analyses of table 2.9 are as follows

- Of 62-minute period, Able was busy for 90% of the time.
- Baker was busy only for 69% of the time.
- The average waiting time for all customers (26) was only about 11/26=0.42 minute (25 seconds), which is very small.
- Nine have to wait for on an average of 11/9 = 1.22 minutes, which is quite low.
- On the whole the system seems to be well balanced. One server cannot handle all the customers and three servers would be probably too many. The waiting time can be reduced nearly to zero by adding an additional server but the cost would be quite high.

A	B	C	D	E	F	G	H	I	J	K	L
						Able			Baker		
Customer no.	Random digits of arrival	Time between arrivals	Clock time of interval	Random digits for service	Time service begins	Service time	Time service ends	Time service begins	Service time	Time service ends	Time in queue
1	-	-	0	95	0	5	5				0
2	26	2	2	21				2	3	5	0
3	98	4	6	51	6	3	9				0
4	90	4	10	92	10	5	15				0
5	26	2	12	89				12	6	18	0
6	42	2	14	38	15	3	18				1
7	74	3	17	13	18	2	20				1
8	80	3	20	61	20	4	24				0
9	68	3	23	50				23	4	27	0
10	22	1	24	49	24	3	27				0
11	48	2	26	39	27	3	30				1
12	34	2	28	53				28	4	32	0
13	45	2	30	88	30	5	35				0
14	24	1	31	01				32	3	35	1
15	34	2	33	81	35	4	39				2
16	63	2	35	53				35	4	39	0
17	38	2	37	81	39	4	43				2
18	80	3	40	64				40	5	45	0
19	42	2	42	01	43	2	45				1
20	56	2	44	67	45	4	49				1
21	89	4	48	01				48	3	51	0
22	18	1	49	47	49	3	52				0
23	51	2	51	75				51	5	56	0
24	71	3	54	57	54	3	57				0
25	16	1	55	87				56	6	62	1
26	92	4	59	47	59	3	62				0
						56			43		11

Table 2.9 Simulation table for carhop example

2.5 Simulation of Inventory Systems

Inventory system is one of the important classes of simulation problems. This inventory system has a periodic review of length *N*, at which time the inventory is checked. At each review, an order is made to bring the inventory up to the level *M* (Maximum inventory). At the end of first review period, an order quantity Q_1 is placed. The lead time (length of time between the placement and receipt of an order) is zero.

A simple inventory system is shown in fig 2.9. Demand is shown as being uniform over the time period in fig 2.9. Actually demands are not usually uniform, they fluctuate over time. One possibility is that demands all occur at the beginning of the cycle. Another is that the lead time is random of some positive length.

In the second cycle the amount in inventory drops below zero, indicating shortage. In fig 2.9 these units are backordered; when the order arrives, the demands for the backordered items are satisfied first.

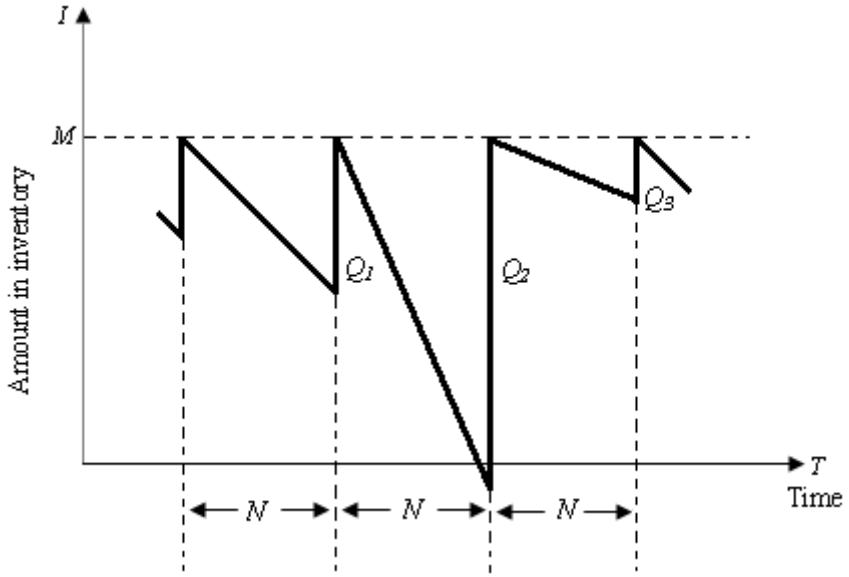


Fig 2.9 Probabilistic order-level inventory system

To avoid shortages, a buffer, or safety, stock would need to be carried.
Carrying stock in inventory might affect the cost attribute. Some of the scenarios are

- The costs can be interest paid on the funds borrowed to buy the items, renting of storage space, hiring guards etc.
- Carrying high inventory causes more frequent reviews and consequently, more frequent purchases which lead to ordering cost.
- Carrying short inventory may cause loss of good will to customers.

Larger inventories decreases the possibilities of shortages but these costs must be traded off in order to minimize the total cost of an inventory system. The total cost (profit) of an inventory system is the measure of performance which is affected by changing M and N.

In an (M, N) inventory systems, the events that may occur are: the demand for items in the inventory, the review of the inventory position and the receipt of an order at the end of each review period. When the lead time is zero, as in fig 2.9, the last two events occur simultaneously.

Example 2.4 (The newspaper seller's problem)
A classical inventory problem concerns the purchase and sales of newspapers. The paper seller buys the paper for 33 cents each and sells them for 50 cents each. Newspapers not sold at the end of the day are sold as scrap for 5 cents each. Newspapers can be purchased in bundles of 10. Thus, the paper seller can buy 50, 60 and so on. There are 3 types of newsdays, 'good', 'fair' and 'poor' with probabilities of 0.35, 0.45, and 0.20 respectively. The distribution of papers demanded on each of these days is given in table 2.10. The problem is to determine the optimal number of papers the newspaper seller should purchase. Simulate demands for 20 days and record profits from sales each day.

Table 2.10 Distribution of newspapers demanded

Demand	Probability Distribution		
	Good	Fair	Poor
40	0.03	0.10	0.44
50	0.05	0.18	0.22
60	0.15	0.40	0.16
70	0.20	0.20	0.12
80	0.35	0.08	0.06
90	0.15	0.04	0.00
100	0.07	0.00	0.00

The profits are given by the following relationship

$$\text{Profit} = \left[\left(\begin{matrix}\text{Revenue}\\ \text{from sales}\end{matrix} \right) - \left(\begin{matrix}\text{Cost}\\ \text{of newspapers}\end{matrix} \right) - \left(\begin{matrix}\text{Lost profit from}\\ \text{excess demand}\end{matrix} \right) + \left(\begin{matrix}\text{Salvage from sale}\\ \text{of scrap papers}\end{matrix} \right) \right]$$

Note→ 1$ = 100 cents

- The revenue from sales is 50 cents for each paper sold. The cost of newspapers is 33 cents for each paper purchased.
- The lost profit from excess demand is 50 – 33 = 17 cents for each paper demanded that could not be provided.
- The salvage value of scrap papers is 5 cents each.
- Tables 2.11 and 2.12 provide the random digits for the types of newsdays and the demands for those days.

- To solve this problem by simulation, requires setting a policy of buying a certain number of papers each day, then simulating the demands for papers over the 20-day time period to determine the total profit. Here simulation is carried out for 70 newspapers and is shown in the table 2.13.

- The policy is changed to other values and the simulation is repeated until the best value is found.

Type of Newsday	Probability	Cumulative probability	Random-digit assignment
Good	0.35	0.35	01-35
Fair	0.45	0.80	36-80
Poor	0.20	1.00	81-00

Table 2.11 Random - digit assignment for type of newsday

Demand	Cumulative Distribution			Random – Digit assignment		
	Good	Fair	Poor	Good	Fair	Poor
40	0.03	0.10	0.44	01-03	01-10	01-44
50	0.08	0.28	0.66	04-08	11-28	45-66
60	0.23	0.68	0.82	09-23	29-68	67-82
70	0.43	0.88	0.94	24-43	69-88	83-94
80	0.78	0.96	1.00	44-78	89-96	95-00
90	0.93	1.00	1.00	79-93	97-00	
100	1.00	1.00	1.00	94-00		

Table 2.12 Random – Digit assignments for newspapers demanded

- On day 1 the demand is 60 newspapers. The revenue from the sale of 60 newspapers is $30. Ten newspapers are left at the end of the day. The salvage value at 5 cents each is 50 cents. The profit for the first day is

  Profit = $30.00 - $23.10 – 0 + $0.50 = $7.40

- On 5th day the demand is greater than the supply. The revenue from sales is $35, since only 70 papers are available under this policy. An additional 20 papers could have been sold. Thus a lost profit of $3.40 (20 * 17 cents) is assessed. The daily profit is determined as follows

  Profit = $35.00 - $23.10 - $3.40 + 0 = $8.50

- The profit for the 20-day period is the sum of the daily profits $174.90. It can also be computed from the totals for the 20 days of simulation as

  Total profit = $645.00 - $462.00 - $13.60 + $5.50 = $174.90.

This simulation is repeated by changing the values for the policy (buying number of newspapers) like 60, 80 and so on. The best policy is obtained by comparing all the profits.

Day	Random digits for type of Newsday	Type of Newsday	Random digits for demand	Demand	Revenue from sales	Lost profit from excess demand	Salvage from sale of scrap	Daily profit
1	94	Poor	80	60	$30.00	-	$0.50	$7.40
2	77	Fair	20	50	25.00	-	1.00	2.90
3	49	Fair	15	50	25.00	-	1.00	2.90
4	45	Fair	88	70	35.00	-	-	11.90
5	43	Fair	98	90	35.00	$3.40	-	8.50
6	32	Good	65	80	35.00	1.70	-	10.20
7	49	Fair	86	70	35.00	-	-	11.90
8	00	Poor	73	60	30.00	-	0.50	7.40
9	16	Good	24	70	35.00	-	-	11.90
10	24	Good	60	80	35.00	1.70	-	10.20
11	31	Good	60	80	35.00	1.70	-	10.20
12	14	Good	29	70	35.00	-	-	11.90
13	41	Fair	18	50	25.00	-	1.00	2.90
14	61	Fair	90	80	35.00	1.70	-	10.20
15	85	Poor	93	70	35.00	-	-	11.90
16	08	Good	73	80	35.00	1.70	-	10.20
17	15	Good	21	60	30.00	-	0.50	7.40
18	97	Poor	45	50	25.00	-	1.00	2.90
19	52	Fair	76	70	35.00	-	-	11.90
20	78	Fair	96	80	35.00	1.70	-	10.20
					$645.00	$13.60	$5.50	$174.90

Table 2.13 Simulation table for purchase of 70 newspapers

Example 2.5 (Simulation of an (M, N) Inventory system)
Suppose that the maximum inventory level M, is 11 units and the review period N is 5 days. The problem is to estimate by simulation, the average ending units in inventory and the number of days when a shortage occurs. The distribution of the number of units demanded per day is shown in table 2.14. The lead time is shown in the table 2.15. Assume that the orders are placed at the close of business and are received for inventory at the beginning of business as determined by the lead time.

Demand	Probability	Cumulative probability	Random-digit assignment
0	0.10	0.10	01 - 10
1	0.25	0.35	11 - 35
2	0.35	0.70	36 - 70
3	0.21	0.91	71 - 91
4	0.09	1.00	92 - 00

Table 2.14 Random – Digit assignments for daily demand

Lead Time (days)	Probability	Cumulative probability	Random-digit assignment
1	0.6	0.6	1 - 6
2	0.3	0.9	7 - 9
3	0.1	1.0	0

Table 2.15 Random – Digit assignments for lead time

- Only five cycles are shown.
- The random digit assignments for daily demand and lead time are shown in rightmost columns of tables 2.14 and 2.15. The simulation results are shown in table 2.16.
- The simulation has been started with inventory level at 3 units and an order of 8 units scheduled to arrive in 2 days time.
- The order for 8 units is available on the morning of third day of first cycle, raising the inventory level from 1 to 9 units.
- Demands during the remainder of first cycle reduced the ending inventory level to 2 units on fifth day. Thus an order for 9 units was placed. The lead time for this order was 1 day. The order of 9 units was added to inventory on the morning of day 2 of cycle 2.

Based on 5 cycles of simulation, the average ending inventory is approximately 3.5(88 / 25) units. On 2 of 25 days a shortage condition existed.

To make an estimate of the mean units in ending inventory, many cycles would have to be simulated and also changing M and N values.

Cycle	Day	Beginning inventory	Random digits for demand	Demand	Ending inventory	Shortage quantity	Order quantity	Random digits for lead time	Days until order arrives
	1	3	24	1	2	0	-	-	1
	2	2	35	1	1	0	-	-	0
1	3	9	65	2	7	0	-	-	-
	4	7	81	3	4	0	-	-	-
	5	4	54	2	2	0	9	5	1
	1	2	03	0	2	0	-	-	0
	2	11	87	3	8	0	-	-	-
2	3	8	27	1	7	0	-	-	-
	4	7	73	3	4	0	-	-	-
	5	4	70	2	2	0	9	0	3
	1	2	47	2	0	0	-	-	2
	2	0	45	2	0	2	-	-	1
3	3	0	48	2	0	4	-	-	0
	4	9	17	1	4	0	-	-	-
	5	4	09	0	4	0	7	3	1
	1	4	42	2	2	0	-	-	0
	2	9	87	3	6	0	-	-	-
4	3	6	26	1	5	0	-	-	-
	4	5	36	2	3	0	-	-	-
	5	3	40	2	1	0	10	4	1
	1	1	07	0	1	0	-	-	0
	2	11	63	2	9	0	-	-	-
5	3	9	19	1	8	0	-	-	-
	4	8	88	3	5	0	-	-	-
	5	5	94	4	1	0	10	8	2
					88				

Table 2.16 Simulation table for (M, N) inventory system

Unit 3
Statistical Models

3.1 Review of Terminology and Concepts

3.1.1 Random variables

A random variable is a rule that assigns a number to each outcome of an experiment. These numbers are called values of random variable. Random variables are usually denoted by X.

Ex 1. If a die is rolled out, the outcome has a value from 1 through 6.
2. If a coin is tossed, the possible outcome is head ‘H’ or tail ‘T’.

There are two types of random variables

1. Discrete random variable
2. Continuous random variable

Discrete Random Variable

A discrete random variable takes only specific, isolated numerical values. The variables which take finite numeric values are called as *Finite discrete random variables* and which takes unlimited values are called as *Infinite discrete random variables*. The examples are shown in the table 3.1.

Random Variable	Values	Type

Flip a coin three times; *X* = the total number of heads.	{0, 1, 2, 3}	Discrete Finite There are only four possible values for *X*.
Select a mutual fund; *X* = the number of companies in the fund portfolio.	{2, 3, 4, ...}	Discrete Infinite There is no stated upper limit to the size of the portfolio.

Table 3.1 Examples for discrete random variables

Let

$X \rightarrow$ discrete random variable

$R_X \rightarrow$ possible values of X, given by range space of X.

$x_i \rightarrow$ individual outcome in R_X.

A number $p(x_i) = P(X = x_i)$ gives the probability that the random variable equals the value of x_i. The number $p(x_i)$, i=1, 2, 3 … must satisfy two conditions

1. $p(x_i) \geq 0$, for all the values of i
2. $\sum_{i=1}^{\infty} p(x_i) = 1$

The collection of pairs $(x_i, p(x_i))$ i.e. a list of probabilities associated with each of its possible values is called probability distribution of X. $p(x_i)$ is called **probability mass function** (pmf) of X.

Example 3.1

Consider the experiment of tossing a single die, defining X as the number of spots on up face of die after a toss.

Solution

N=total number of observations = 21

The discrete probability distribution is given by

x_i	1	2	3	4	5	6
$P(x_i)$	1/21	2/21	3/21	4/21	5/21	6/21

The conditions also are satisfied, i.e.

1. $p(x_i) \geq 0$, for i = 1,2,…6
2. $\sum_{i=1}^{\infty} p(x_i) = 1/21 + 2/21 + \ldots + 6/21 = 1$

The distribution is shown graphically in fig 3.1

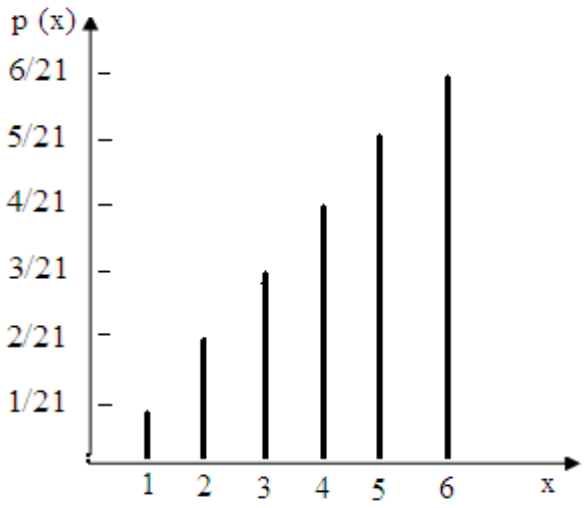


Fig 3.1 pmf for loaded die example

Continuous random variable

Continuous Random variable takes any values within a continuous range or an interval. The example is tabulated in table 3.2.

Random Variable	Values	Type
Measure the length of an object; X = its length in centimeters.	Any positive real number	Continuous The set of possible measurements can take on any positive value

Table 3.2 Example for continuous random variables

For a continuous random variable X, the probability that X lies in the interval [a, b], is given by

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)\, dx \qquad (3.1)$$

The function f(x) is called **probability density function** (pdf) of random variable X.

The pdf must satisfy the following conditions

1. $f(x) \geq 0$, for all x in R_X
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$ (total area under graph is 1)
3. $f(x) = 0$, if x is not in R_X

For any specified value x_0, $P(X = x_0) = 0$ since

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

Since $P(X = x_0) = 0$, the following equation also hold:

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

The graphical interpretation of equation 3.1 is shown in fig 3.2

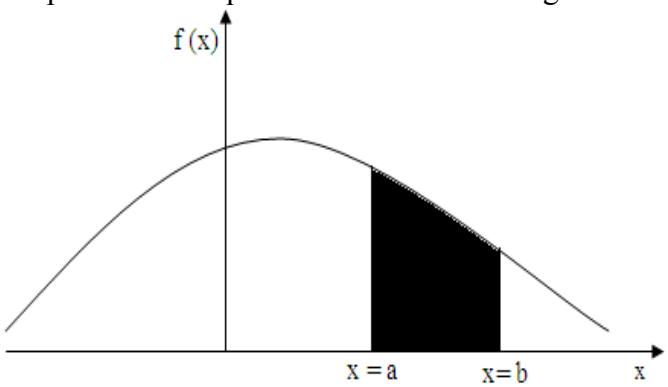


Fig 3.2 Graphical interpretation of P (a < X < b)

Example 3.2

The life of a laser- ray device used to inspect cracks in aircraft wings is given by X, continuous random variable, assuming $x \geq 0$.The pdf of lifetime in years is ,

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Determine the life of device between the interval [2, 3].

Solution

The probability that the life of laser ray device between 2 and 3 years is determined from

$$P(2 \leq X \leq 3) = \frac{1}{2} \int_2^3 e^{-x/2}\, dx$$

$$= \frac{1}{2} \left[-2 e^{-x/2} \right]_2^3$$

$$= -e^{-3/2} + e^{-1}$$

$$= -0.223 + 0.368 = 0.145$$

3.1.2 Cumulative distribution function (cdf)

The cdf is denoted by F(x), measures the probability that the random variable X is less than or equal to *x*, i.e. $F(x) = P(X \leq x)$.

If X is discrete, then

$$F(x) = \sum_{\text{all } x_i \leq x} p(x_i)$$

If X is continuous, then

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Some properties of cdf are

1. F is a non decreasing function. If a < b, then F (a) ≤ F (b)
2. $\lim_{x \to \infty} F(x) = 1$
3. $\lim_{x \to -\infty} F(x) = 0$

Note - All probability questions about X can be answered in terms of cdf. For example P (a < X ≤ b) = F (b) – F (a), for all a < b.

Example 3.3

The life of a laser- ray device used to inspect cracks in aircraft wings is given by X, continuous random variable, assuming x ≥ 0.The pdf of lifetime in years is ,

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

1. Determine the probability that the device will last for less than 2 years.
2. Determine the probability that the life of laser ray device is between 2 and 3 years.

Solution

1. The probability that the device will last for less than 2 years is,
   $P(0 \leq X \leq 2) = F(2) - F(0)$
   $= 1 - e^{-1}$
   $= 0.632$
2. The probability that the life of laser ray device between 2 and 3 years is
   $P(2 \leq X \leq 3) = F(3) - F(2)$
   $= (1 - e^{-3/2}) - (1 - e^{-1})$
   $= -e^{-3/2} + e^{-1}$
   $= -0.233 + 0.368$
   $= 0.145$

3.1.3 Expectation

If X is a random variable, the expected value of X, denoted by E (X) is defined as

$$E(X) = \sum_{\text{all } i} x_i\, p(x_i) \qquad \text{if X is discrete}$$

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

E (X) is also referred as Mean μ or first moment of X.

E (X^n), $n \geq 1$ [n^{th} moment of X] is computed as

$$E(X^n) = \sum_{\text{all } i} x_i^n \, p(x_i) \qquad \text{if X is discrete}$$

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

Variance of random variable X is denoted by V(X) or Var(X) or σ^2, is defined as

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

Standard deviation is given by

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

Note

- ✓ Mean E (X) is a measure of central tendency of a random variable.
- ✓ Variance V(X) is a measure of variation of possible values of X around the mean E (X).

Example 3.4
Find the mean, variance and standard deviation of die- tossing experiment.

Solution
N=21

x_i	1	2	3	4	5	6
p (x_i)	1/21	2/21	3/21	4/21	5/21	6/21

$$\text{Mean} = E(X) = 1\left(\frac{1}{21}\right) + 2\left(\frac{2}{21}\right) + \ldots + 6\left(\frac{6}{21}\right) = 4.33$$

$$E(X^2) = 1^2\left(\frac{1}{21}\right) + 2^2\left(\frac{2}{21}\right) + \ldots + 6^2\left(\frac{6}{21}\right) = 21$$

Variance = $V(X) = 21 - (4.33)^2 = 2.22$

$$\text{Standard Deviation} = \sigma = \sqrt{2.22} = 1.49$$

3.1.4 Mode

In case of discrete, Mode is the value of random variable that occurs most frequently. In case of continuous, the mode is the value at which pdf is maximized.

3.2 Useful statistical models

During the conduct of simulation, numerous situations arise where an investigator choose to introduce probabilistic events. For example

In Queuing systems ─ inter arrival and service times are probabilistic.
In Inventory models ─ time between demand and lead time may be probabilistic.
In Reliability model ─ time to failure may be probabilistic.

In each case, the simulation analyst desires to generate random events and use known statistical models if the distribution can be found. Some of the systems and the chosen statistical models are discussed.

Queuing system

In queuing examples, interarrival and service-time patterns are given. The times between the arrivals and service time are always probabilistic. Service times may be constant or probabilistic.

- If service times are completely random, exponential distribution is often used.
- If service time is constant, but some random variability causes fluctuations in positive or negative way then normal distribution is used.
  For example- The time it takes for lathe to traverse a 10cm shaft should be always same, but the material may have slight difference in hardness, causing different processing times.
- If there are more large service times, then weibull distribution is a better model.
- To model interarrival and service times ─ exponential, gamma and weibull distributions are used

The differences between these distributions involve the location of modes of pdf's and the shape of their tails for large and small times.

Mode –	Exponential distribution	– at origin
	Gamma and Weibull distribution	– at some point (≥ 0)
Tail –	Gamma and Exponential distribution	– long
	Weibull distribution	– declines more or less rapidly

Inventory systems

It has three random variables

1. Number of units demanded per order or per time period
2. Time between demands
3. Lead time (time between placing an order and receiving receipt of that order)

- In simple mathematical model, demand is constant over time and lead time is zero or constant.
- In realistic cases (Simulation models), demand occurs randomly in time and number of units demanded each time is also random.
- The geometric distribution has its mode at unity, given that atleast one demand has occurred.
- If demand data are characterized by a long tail, Negative Binomial distribution is appropriate.
- The Poisson distribution is often used to model the demand.

Reliability and maintainability

- If only random failures occur, the time-to-failure distribution may be modeled as Exponential.
- If most failure are due to wear, Normal distributions may be appropriate.
- To describe time-to-failure for some types of components, lognormal distribution is found to be applicable.
- When there are number of components in a system and failures is due to serious large number of defects, then Weibull distribution is extensively used.

Limited data

In many situations, simulations begin before data collection is completed. Three distributions have application to incomplete/ limited data

- Uniform distribution can be used when an interarrival or service time is known to be random but no information is immediately available about the distribution.
- Triangular distribution can be used when assumptions are made about minimum, maximum and modal values of random variable.
- Beta distribution.

3.3 Discrete distributions

Discrete random variables are used to describe random phenomena. The four distributions are discussed.

3.3.1 Bernoulli trials and Bernoulli distribution

Consider an experiment consisting of *n* trials each of which is success or failure.

Let

$X_j = 1$, if j^{th} experiment results in success

$X_j = 0$, if j^{th} experiment results in failure

The *n* Bernoulli trails are called *Bernoulli process*, if trails are independent, each trail has only two possible outcomes (success, failures) and the probability of success remains constant from trail to trail.

Thus

$$p(x_1, x_2, \ldots x_n) = p_1(x_1).\, p_2(x_2) \ldots p_n(x_n)$$

$$p_j(x_j) = p(x_j) = \begin{cases} p, & x_j = 1,\ j = 1, 2, \ldots, n \\ 1 - p = q, & x_j = 0,\ j = 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

This is called Bernoulli distribution.

Mean of X_j is given by

$$E(X_j) = 0.q + 1.p = p$$

and Variance,

$$V(X_j) = [(0^2 .q) + (1^2 .p)] - p^2 = p(1-p)$$

3.3.2 Binomial distribution

The random variable X denotes the number of successes in *n* Bernoulli trails has a binomial distribution given by p(x).

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

i.e. probability of a particular outcome with all success (S) occurring in first *x* trails followed by *n-x* failures (F).

$$P(\overbrace{SSS\ldots\ldots SS}^{x \text{ of these}}\ \overbrace{FF\ldots\ldots FF}^{n-x \text{ of these}}) = p^x q^{n-x}$$

where q = 1 – p,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Mean is computed by considering X as a sum of *n* independent Bernoulli random variables each with mean *p*. Then

$$X = X_1 + X_2 + \ldots + X_n$$

$$E(X) = p + p + \ldots + p = np$$

Variance p (1 - p) = pq, for *n* independent variables

$$V(X) = pq + pq + \ldots + pq = npq$$

Example 3.5
A production process manufactures computer chips on the average at 2% non-conforming. Every day a random sample of size 50 is taken from the process. If the sample contains more than two non-conforming chips, the process will be stopped. Determine the probability that the process is stopped by the sampling scheme.

Solution
n = 50 Bernoulli trails
p = 2% = 0.02 (each trial)
q = 1 - p = 0.98
X→ Total number of non conforming chips in the sample
Binomial distribution

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

To compute, determine the probability that more than two non conforming chips are present in the sample

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

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{50}{x} (0.02)^x (0.98)^{50-x}$$

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

$$= 0.92$$

$$P(X>2) = 1 - 0.92 = 0.08$$

Therefore the probability that the process is stopped on any day by sampling scheme is approximately 0.08.

Mean number of non conforming chips with sample size 50 is

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

Variance is

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

3.3.3 Geometric distribution

The geometric distribution is related to a sequence of Bernoulli trails. The random variable X is defined as number of trails to achieve the first success. The distribution of X is given by

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

The event {X = x} occurs when there are *x-1* failures followed by a success. Each failure has an associated probability *p*. Thus

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

Mean,

$$E(X) = \frac{1}{p}$$

Variance,

$$V(X) = \frac{q}{p^2}$$

Example 3.6

40% of assembled bubble-jet printers are rejected at inspection station. Find the probability that the first acceptable printer is third one inspected.

Solution
q =40% =0.4
q = 1- p → p = 1 - 0.4 = 0.6
$p(x) = q^{x-1} p$
$p(3) = (0.4)^2 (0.6) = 0.096$
Approximately 10% of cases are first accepted

3.3.4 Poisson distribution

Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The pmf is given by

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

where $\alpha > 0$

Mean and variance,

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

The cdf is given by,

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

Most queuing systems characteristics such as arrival and departure processes are described by a poisson distribution.

Example 3.7
A computer terminal repair person is beeped each time there is a call for service. The number of beeps per hour is known to occur in accordance with a poisson distribution with a mean of $\alpha = 2$ per hour.
1. Determine the probability of three beeps in next hour.
2. Determine the probability of two or more beeps in an hour period.

Solution
$\alpha = 2$
$x = 3$
1. The probability of three beeps in next hour

$$p(3) = \frac{e^{-2} 2^3}{3!} = \frac{(0.135)(8)}{6} = 0.18$$

The same results can be obtained from table A.4 by

F (3) – F (2) = 0.857 – 0.677 = 0.18

2. The probability of two or more beeps in an hour period

p (2 or more) = 1 – p(0)- p(1)
= 1- F (1)
= 1- 0.406
= 0.594

3.4 Continuous distribution

Continuous random variables are used to describe the random phenomena. The distributions are described below.

3.4.1 Uniform distribution

A random variable X is uniformly distributed on interval (a, b), if pdf is given by,

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

cdf,

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

Note that

$$P(x_1 < X < x_2) = F(x_2) - F(x_1) = \frac{x_2 - x_1}{b - a}$$

The probability is proportional to the length of interval for all x_1 and x_2 satisfying $a \leq x_1 < x_2 \leq b$.

Mean,

$$E(X) = \frac{a+b}{2}$$

Variance,

$$V(X) = \frac{(b-a)^2}{12}$$

Uniform distribution plays a vital role in simulation. Random numbers, uniformly distributed between 0 & 1, provide a means to generate random events.

Example 3.8

A bus arrives every 20 minutes at a specified stop beginning at 6.40am and continues till 8.40am. A certain passenger does not know the schedule, but arrives randomly (uniformly distributed) between 7am and 7.30am every morning. What is probability that the passenger waits more than 5 minutes for a bus?

Solution

The passenger waits for more than 5 minutes only if his/her arrival time is between 7am and 7.15am or between 7.20am and 7.30am.

If X is a random variable, which denotes the number of minutes past 7am that the passenger arrives, then the probability is given by

$$P(0 < X < 15) + P(20 < X < 30)$$

Now, X is uniform random variable on (0, 30). Therefore the desired probability is

$$F(15) - F(0) + F(30) - F(20)$$

$$= 15/30 - 0 + 1 - 20/30 = 5/6$$

3.4.2 Exponential distribution

A random variable X is said to be exponentially distributed with parameter $\lambda > 0$ if

The pdf,

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

The density functions are shown in the fig 3.3

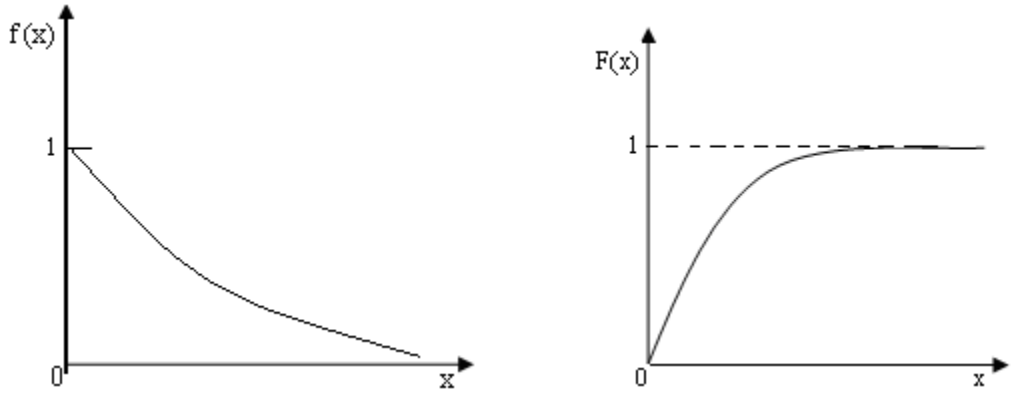


Fig. 3.3 Exponential density function and cumulative distribution function

Mean,

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

Variance,

$$V(X) = \frac{1}{\lambda^2}$$

Therefore Mean and Standard deviation are equal.

The cdf ,

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

- This distribution is used to model interarrival times and service times when they are completely random. In this case λ is a rate: arrival per hour or services per minute
- It is also used to model the lifetime of a component that fails instantaneously, such as light bulb, then λ is failure rate.

Example 3.9
Suppose the life of an industrial lamp, in thousands of hours is exponentially distributed with failure rate λ=1/3 (one failure every 3000 hours, on average)

1. Determine the probability that lamp last longer than its mean life of 3000 hours.
2. Determine the probability that the lamp last between 2000 and 3000 hours.

Solution

λ=1/3

1. The probability that the lamp will last longer than its mean life is given by
$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) \\ &= 1 - F(3) \\ &= 1 - (1 - e^{-3/3}) \\ &= 0.368 \end{aligned}$$
2. The probability that the lamp will last between 2000 and 3000 hours is
$$\begin{aligned} P(2 \leq X \leq 3) &= F(3) - F(2) \\ &= (1 - e^{-3/3}) - (1 - e^{-2/3}) \\ &= -0.368 + 0.513 \\ &= 0.145 \end{aligned}$$

Note: Exponential distribution is memory less.

3.4.3 Normal distribution

A random variable X with mean μ ($-\infty < \mu < \infty$) and variance σ^2 ($\sigma^2 > 0$) has a normal distribution, if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty$$

The Notation $X \sim N(\mu, \sigma^2)$ indicates that random variable X is normally distributed with mean μ and variable σ^2. The normal pdf is represented in fig 3.4

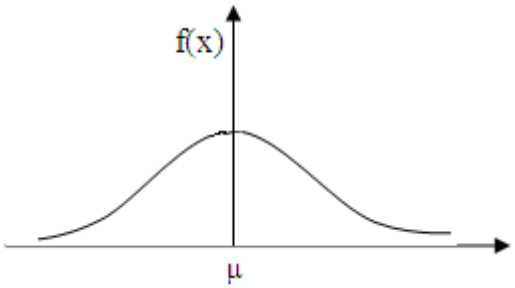


Fig 3.4 pdf of normal distribution

Some of the special properties of normal distribution are

1. $\lim_{x \to -\infty} f(x) = 0$ and $\lim_{x \to +\infty} f(x) = 0$
   The value of f(x) approaches zero as *x* approaches negative or positive infinity.
2. $f(\mu - x) = f(\mu + x)$, pdf is symmetric about μ.
3. Maximum value of pdf occurs at x = μ [Mean & mode are equal].

The cdf of normal distribution is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt$$

Since the above equation is in closed form, it is not possible to evaluate. So a transformation variable, z = (t - μ) / σ, allows the evaluation to be independent of μ and σ.

If $X \sim N(\mu, \sigma^2)$, let Z = (X - μ) / σ, to obtain

$$F(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

$$= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$$= \int_{-\infty}^{(x-\mu)/\sigma} \Phi(Z)\, dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

The pdf ,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

The above is the pdf of normal distribution with mean 0 and variance 1.Thus Z ~ N (0, 1) and it is said that Z has standard Normal Distribution. It is shown in the fig 3.5

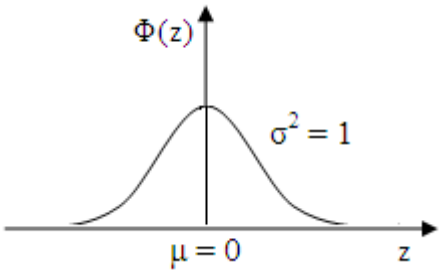


Fig 3.5 The pdf of standard normal distribution

The cdf for Standard Normal Distribution is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Example 3.10
X ~ N (50, 9). Determine F(56).

Solution

X ~ N (50, 9) → X is normally distributed with mean value 50 and variance 9

F (56) = P (X ≤ 56)

=Φ [(56 -50)/3] [Φ (x - μ)/σ]

=Φ (2)

=0.9772 (using table A.3)

The fig 3.6(a) shows pdf of X ~ N(50, 9) and x_0 = 56. The shaded portion is the desired probability. The fig 3.6(b) shows Standard Normal Distribution Z ~ N (0, 1) with value 2 marked since x_0 = 56 is 2σ greater than the mean.

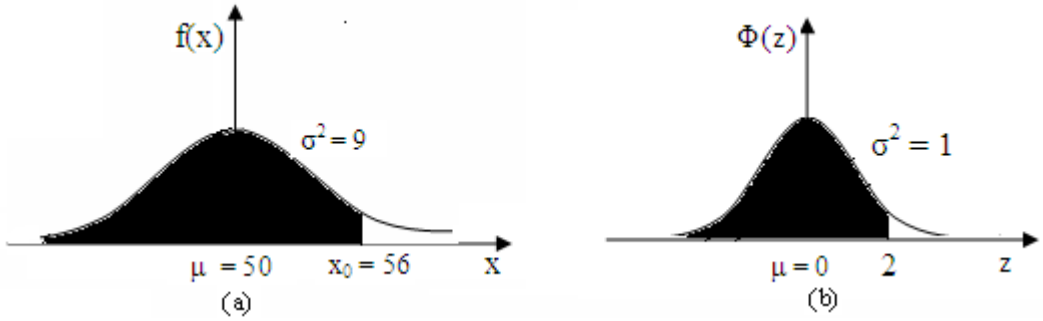


Fig 3.6 Transforming to the standard normal distribution

Example 3.11
The time required to load a vessel X is distributed N (12, 4). Determine the probability that the vessel will be loaded in less than 10 hours.

Solution

$\mu = 12$

$\sigma^2 = 4$

F (10) =?

F (10) $= \Phi [(10 - 12)/2]$

$= \Phi (-1)$

$= 0.1587$ [where $\Phi (-1) = 1 - \Phi (1)$]

$\Phi(1) = 0.8413$, the complement of 0.8413 or 0.1587 is contained in the tail.

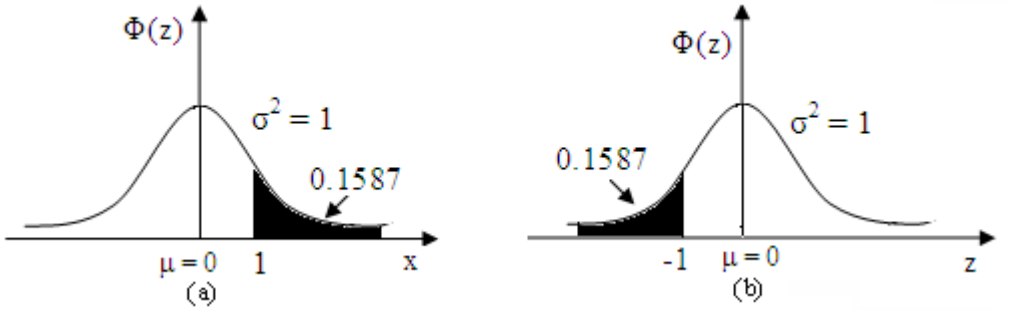


Fig 3.7 Using the symmetry property of the normal distribution

The fig 3.7(a) shows the shaded portion of standard normal distribution. The fig 3.7(b) shows the symmetry property, to determine the shaded region to be $\Phi (-1)$.

Example 3.12

The time to pass through a queue to begin self-service at a cafeteria has been found to be N(15, 9). Determine the probability that an arriving customer waits between 14 and 17 minutes.

Solution

N (15, 9) => $\mu = 15, \sigma^2 = 9$

$P (14 \leq X \leq 17) = F (17) - F (14)$

$= \Phi [(17 - 15)/3] - \Phi [(14 - 15)/3]$

$= \Phi [0.667] - \Phi [-0.333]$

$= 0.3780$

If fig 3.8(a) represents the probability F(17) – F(14) then fig 3.8(b) represents the equivalent probability $\Phi [0.667] - \Phi [-0.333]$ for standard normal distribution.

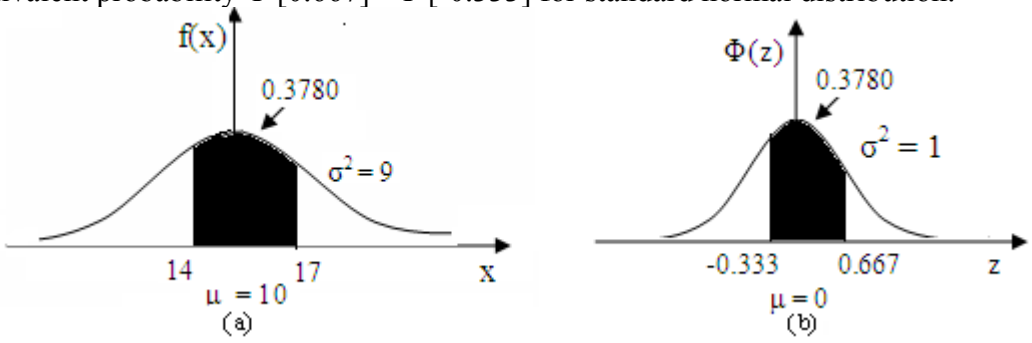


Fig 3.8 Transformation to the standard normal for cafeteria problem

3.4.4 Weibull distribution

The random variable X has a weibull distribution if its pdf has

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

The three parameters in distribution are

- $v \rightarrow$ location parameter ($-\infty < v < \infty$)
- $\alpha \rightarrow$ Scale parameter ($\alpha > 0$)
- $\beta \rightarrow$ Shape parameter ($\beta > 0$)

When $v = 0$, pdf becomes

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

When $v = 0$ & $\beta = 1$, pdf is reduced to

$$f(x) = \begin{cases} \frac{1}{\alpha} e^{-x/\alpha}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

which is an exponential distribution with parameter $\lambda = 1/\alpha$.

Mean,

$$E(X) = V + \alpha \Gamma\left(\frac{1}{\beta} + 1\right)$$

Variance,

$$V(X) = \alpha^2 \left[\Gamma\left(\frac{2}{\beta} + 1\right) - \left[\Gamma\left(\frac{1}{\beta} + 1\right)\right]^2\right]$$

where

$$\Gamma(\beta) = \int_0^{\infty} x^{\beta-1} e^{-x}\, dx, \quad \text{If } \beta \text{ is an integer then } \Gamma(\beta) = (\beta - 1)!$$

The cdf ,

$$F(x) = \begin{cases} 0, & x < v \\ 1 - \exp\left[-\left(\frac{x-v}{\alpha}\right)^{\beta}\right], & x \geq v \end{cases}$$

Example 3.13
The time it takes for an aircraft to land and clear the runway at a major airport has a weibull distribution with v = 1.34 minutes, β = 0.5 and α = 0.04 min. Determine the probability that an incoming airplane will take more than 1.5 minutes to land and clear the runway.

Solution
V = 1.34mins
β = 0.5
α = 0.04 min
The probability than an incoming airplane will take more than 1.5 minutes is

$$P(X > 1.5) = 1 - P(X \leq 1.5)$$
$$= 1 - F(1.5) = 1 - \exp\left[-\{(1.5 - 1.34)/(0.04)\}^{0.5}\right]$$
$$= 1 - e^{-2} = 1 - 0.135 = 0.865$$

3.4.5 Triangular distribution

A random variable X with triangular distribution has its pdf,

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, & a \leq x \leq b \\ \dfrac{2(c-x)}{(c-b)(c-a)}, & b < x \leq c \\ 0, & \text{elsewhere} \end{cases}$$

where a ≤ b ≤ c. The mode occurs at x = b.

A triangular pdf and the representation of height is shown in figure 3.9

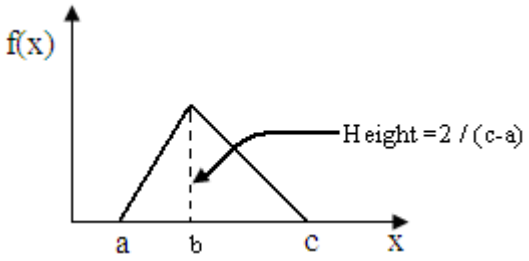


Fig 3.9 pdf of the triangular distribution

Mean = E(X) = (a + b + c) / 3

Mode = b = 3 E(X) – (a + c)

Since a ≤ b ≤ c,

$$\frac{2a + c}{3} \leq E(X) \leq \frac{a + 2c}{3}$$

The cdf,

$$F(x) = \begin{cases} 0, & x \leq a \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, & a < x \leq b \\ 1 - \dfrac{(c-x)^2}{(c-b)(c-a)}, & b < x \leq c \\ 1, & x > c \end{cases}$$

Example 3.14
The central processing requirements for a program that will execute, have a triangular distribution with a = 0.05 second, b = 1.1 seconds and c = 6.5 seconds. Determine the probability that the CPU requirement for a random number is 2.5 seconds or less.

Solution
a = 0.05 second
b = 1.1 seconds
c = 6.5 seconds
P (X ≤ 2.5) = F (2.5) =?

Interval → (0.05, 1.1) plus that portion in (1.1, 2.5) i.e. 1.1 < 2.5 ≤ 6.5 (b < x ≤ c)

Therefore F (2.5) = 1- [(6.5 – 2.5)^2 / {(6.5 – 0.05) (6.5 – 1.1)}] = 0.541

Thus the probability is 0.541 that the CPU requirement is 2.5 seconds or less.

3.4.6 Lognormal distribution

A random variable X has a lognormal distribution, if its pdf is,

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\sigma x} \exp\left[-\dfrac{(\ln x - \mu)^2}{2\sigma^2}\right], & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\sigma^2 > 0$

Mean of lognormal random variable is

$$E(X) = e^{\mu + \sigma^2/2}$$

Variance of lognormal random variable is

$$V(X) = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$$

The parameter μ and σ^2 are not the mean and variance of lognormal. These parameters come from the normal distribution. When Y has $N(\mu, \sigma^2)$, then $X= e^Y$ has a lognormal distribution with parameters μ and σ^2. If mean and variance of lognormal are known to be μ_L and σ^2_L, then the parameter μ and σ^2 are given by

$$\mu = \ln\left(\frac{\mu_L^2}{\sqrt{\mu_L^2 + \sigma_L^2}}\right)$$

$$\sigma^2 = \ln\left(\frac{\mu_L^2 + \sigma_L^2}{\mu_L^2}\right)$$

3.5 Poisson Process

The arrivals of jobs at a job shop, arrival of aircraft at a runway, arrival of calls at a switchboard are some of the random events. These events can be described by a counting function N(t), for all $t \geq 0$.The counting function will represent the number of events that occurred in interval [0, t]. Time zero is the point at which the observation begins. For each interval [0, t] the value N(t) is an observation of a random variable (N(t) is assumed to be integer).

The counting process $\{N(t), t \geq 0\}$ is said to be a poisson process with mean rate λ, if it satisfies the following assumptions.

1. Arrivals occur one at a time.
2. $\{N(t), t \geq 0\}$ has stationary increments. The distribution of number of arrivals between *t* and *t* + *s* depends only on length of interval *s* and not on starting point *t*. Thus arrivals are completely random
3. $\{N(t), t \geq 0\}$ has independent increments.

If arrivals are according to the poisson process, it can be shown that probability N(t) is equal to *n* i.e.

$$P[N(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad \text{for } t \geq 0 \text{ and } n = 0, 1, 2, \ldots$$

Comparing poisson pmf with the above equation, N(t) has poisson distribution with parameter $\alpha = \lambda t$.

Thus mean and variance is given by

$$E(N(t) = \alpha = \lambda t = V[N(t)]$$

For any times *s* and *t*, such that $s < t$, the assumption - stationary increments implies random variable N(t) –N(s) representing number of arrivals in interval *s* to *t* is also poisson distributed with mean λ (t - s).

Thus

$$P[N(t) - N(s) = n] = \frac{e^{-\lambda(t-s)}[\lambda(t-s)]^n}{n!} \quad \text{for} \quad n = 0, 1, 2, \ldots$$

and

$$E[N(t) - N(s)] = \lambda(t - s)$$
$$= V[N(t) - N(s)]$$

Now, consider the time at which an arrival occur in poisson process. Let the first arrival occur at time A_1, the second at time A_1+A_2 and so on. Thus $A_1, A_2, \ldots$ are successive inter arrival times. It is depicted in fig 3.10

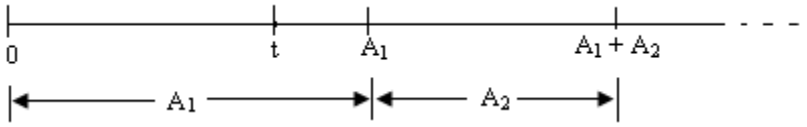


Fig 3.10 Arrival process

Since first arrival occurs after time t and no arrivals in interval [0, t], it is seen that $\{A_1 > t\} = \{N(t) = 0\}$

Therefore

$$P(A_1 > t) = P[N(t) = 0] = e^{-\lambda t}$$

The probability that the first arrival will occur in [0, t] is given by

$$P(A_1 \leq t) = 1 - e^{-\lambda t}$$

which is the cdf of exponential distribution with λ . Hence A_1 is exponentially distributed with mean $1/\lambda$ and also $A_1, A_2, \ldots$ inter arrival times are exponentially distributed and independent with mean $1/\lambda$.

Therefore poisson process can be also defined as if inter arrival times are exponentially and independently distributed then number of arrivals by time t, say N(t) meets three assumptions.

Properties of Poisson process

Suppose, each time an event occurs is classified as either type I or type II event. Further that each event is classified as type I event with probability p and type II event with probability $1-p$, independently of all other events. The two properties are

1. Random splitting

Let $N_1(t)$ be random variable denoting number of type I event, $N_2(t)$ for type II event. $N(t) = N_1(t) + N_2(t)$. $N_1(t)$and $N_2(t)$ are both poisson processes having rates λp and $\lambda(1-p)$ is shown in fig 3.11

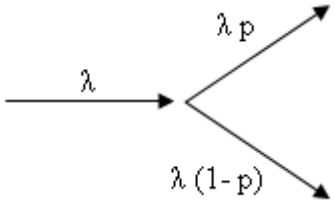


Fig 3.11 Random splitting

2. Pooled process

It is the process of pooling two arrival streams. If $N_i(t)$ are random variables representing independent poisson process with rates λ_i then $N(t) = N_1(t) + N_2(t)$ is a poisson process with rate $\lambda_1 + \lambda_2$, shown in fig 3.12

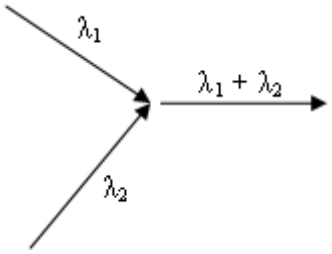


Fig 3.12 Pooled process

3.6 Empirical Distributions

An empirical distribution can be either continuous or discrete. It is used when it is impossible or unnecessary to establish random variable that has any particular known distribution.

Example 3.15 (Discrete)

Customers at a local restaurant arrive at a lunch time in groups ranging from 1 to 8 persons. The number of persons per party in the last 300 groups has been observed and summarized in table 3.3 below

Arrivals per party	Frequency	Relative frequency	Cumulative relative frequency
1	30	0.10	0.10
2	110	0.37	0.47
3	45	0.15	0.62
4	71	0.24	0.86
5	12	0.04	0.90
6	13	0.04	0.94
7	7	0.02	0.96
8	12	0.04	1.00

Table 3.3 Arrivals per party distribution

Relative frequency and cumulative relative frequency are calculated.

Relative frequency = 30/300 = 0.10, 110/300 = 0.37…
The histogram of data that were gathered depicted in fig 3.13

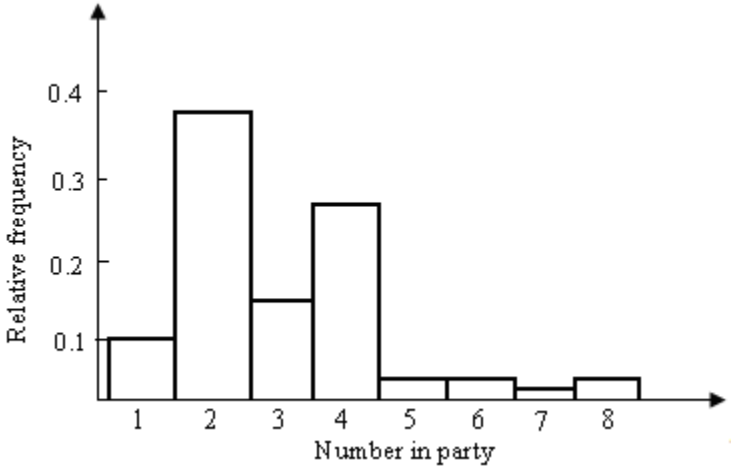


Fig 3.13 Histogram of party size

Fig 3.14 provides a cdf of data, nothing but empirical distribution of given data.

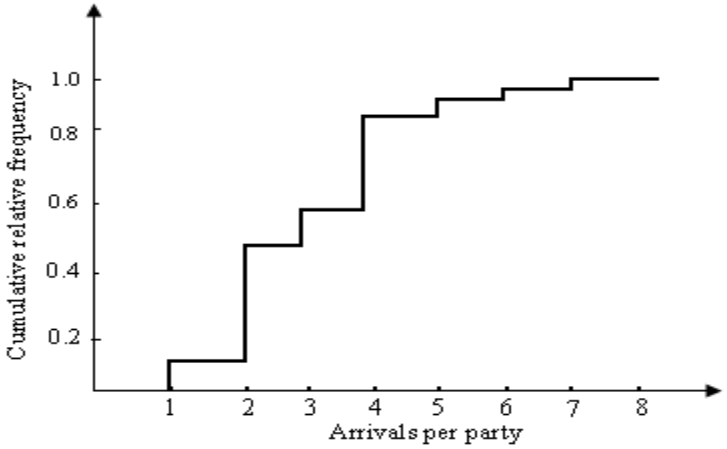


Fig 3.14 Empirical cdf of party size

Example 3.16 (Continuous)
The time required to repair a conveyor system which has suffered a failure has been collected for the last 100 instances with results shown in the table 3.4

Interval (hours)	Frequency	Relative frequency	Cumulative frequency
$0 < x \leq 0.5$	21	0.21	0.21
$0.5 < x \leq 1.0$	12	0.12	0.33
$1.0 < x \leq 1.5$	29	0.29	0.62
$1.5 < x \leq 2.0$	19	0.19	0.81
$2.0 < x \leq 2.5$	8	0.08	0.89
$2.5 < x \leq 3.0$	11	0.11	1.00

Table 3.4 Repair times for conveyor

It is seen that between 0 and 0.5 hour, there are 21 instances, between 0.5 and 1.0 hours 12 instances and so on. The empirical cdf is shown in fig 3.15

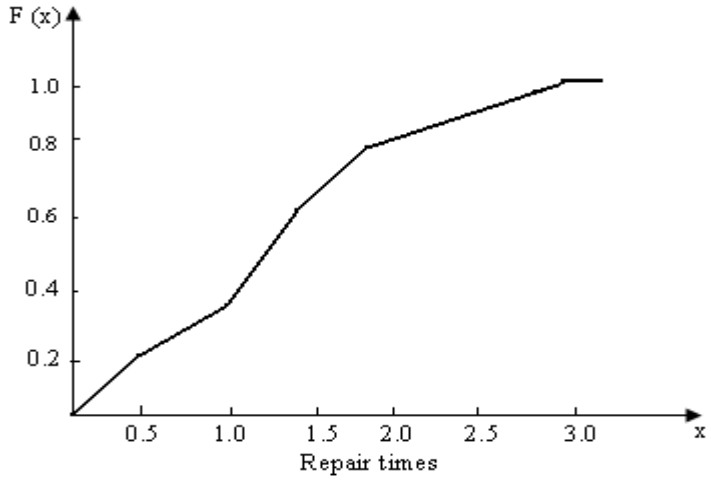


Fig 3.15 Empirical cdf for repair times

The curve is formed by connection of points in the form [x, F(x)]
i.e. (0, 0), (0.5, 0.21), (1.0, 0.33)…

Unit 4
Random-Number Generation

4.1 Random Numbers

A random number is a number generated by a process, whose outcome is unpredictable, and which cannot be sub sequentially reliably reproduced. Random numbers are the basic building blocks for all simulation algorithms.

4.1.1 Properties of Random Numbers

The two important statistical properties are:
1. Uniformity
2. 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

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

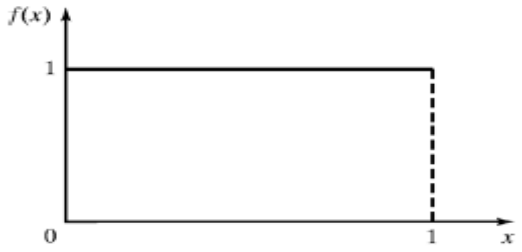


Fig 4.1 The pdf for random numbers

The expected value of each R_i is given by

$$E(R) = \int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

The variance is given by

$$V(R) = \int_0^1 x^2\,dx - [E(R)]^2$$
$$= [x^3/3]_0^1 - (1/2)^2 = 1/3 - 1/4$$
$$= 1/12$$

The consequences of uniformity and independence properties are:

1. 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.
2. The probability of observing a value in a particular interval is independent of previous values drawn.

4.1.2 Generation of Pseudo-Random Numbers

The word “Pseudo” is used, because generating numbers using a known method removes the potential for true randomness. Since the sequence of numbers is deterministic they are referred to as "pseudo-random".

Goal: To produce a sequence of numbers between [0, 1] that simulates, or replicates, the ideal properties of random numbers (RN).

The important considerations in selecting RN routines/methods are

1. Fast and cost effective
2. Portable to different computers
3. Have sufficient long cycle
4. Replicable
5. Closely approximate the ideal statistical properties of uniformity and independence.

4.2 Techniques for Generating Random Numbers

The most widely used techniques for generating random numbers are:

1. Linear Congruential Method (LCM)
2. Combined Linear Congruential Generators (CLCG)

4.2.1 Linear Congruential Method

The most widely used technique for generating random numbers, initially proposed by Lehmer [1951]. This method produces a sequence of integers, X_1, X_2 … between 0 and *m-1* by following a recursive relationship:

$$X_{i+1} = (aX_i + c) \bmod m, \quad i = 0,1,2,\ldots$$

The multiplier — The increment — The modulus

The initial value X_0 is called seed. The selection of the values for *a, c, m,* and X_0 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

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

Example 4.1
Use linear congruential method to generate sequence of random numbers with
$X_0 = 27$, a = 17, c = 43, and m = 100.

Solution
Random numbers (R_i)

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

Equations → $X_{i+1} = (a X_i + c) \bmod m$, $R_i = X_i / m$, i=1,2,…..

$X_1 = (17 * 27 + 43) \bmod 100 = 2$, $R_1 = 2 / 100 = 0.02$
$X_2 = (17 * 2 + 43) \bmod 100 = 77$, $R_2 = 77 / 100 = 0.77$
$X_3 = (17 * 77 + 43) \bmod 100 = 52$, $R_3 = 52/ 100 = 0.52$
⋮
Hence the numbers are generated.

The secondary properties to generate random numbers include maximum density and maximum period.
1. Maximum density means values assumed by R_i, i = 1, 2,… leave no large gaps on the interval [0,1].
Problem: The values generated from $R_i = X_i / m$, is discrete on integers instead of continuous.

Solution: A very large integer for modulus *m*.
2. Maximum Period
To achieve Maximum density and avoid cycling, the generator should have largest possible period. Most digital computers use a binary representation of numbers. Speed and efficiency are aided by a modulus *m*, to be (or close to) a power of 2.Maximal period is achieved by proper choice of *a, c, m* and X_0. The different cases are

- For *m* a power of 2, say $m = 2^b$ and $c \neq 0$, the longest possible period is $\mathbf{P = m = 2^b}$, provided that *c* is relatively prime to *m* and **a = 1 + 4k**, where *k* is an integer.
- For *m* a power of 2, say $m = 2^b$ and $c = 0$, the longest possible period is $\mathbf{P = m / 4 = 2^{b-2}}$, which is achieved provided that the seed X_0 is odd and the multiplier *a*, is given by **a = 3 + 8k** or **a = 5 + 8k**, for some k = 0, 1...
- For *m* a prime number and c = 0, the longest possible period is **P = m - 1**, which is achieved provided that the multiplier *a*, has the property that the smallest integer *k* such that $a^k - 1$ is divisible by *m* is k = m - 1,

Example 4.2
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.

Solution
c=0 (multiplicative congruential method), $m = 2^6 = 64$ and a=13 → (a=5+8*1=13) so 'a' is in the form 5+8k with k=1.
So therefore the maximal period p= m / 4= 64 / 4=16 for odd seeds i.e. for X_0=1 and 3

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

i	X_i $X_0 = 1$	X_i $X_0 = 2$	X_i $X_0 = 3$	X_i $X_0 = 4$
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	

Seed

Table 4.1 Period determination using various seeds

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

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

is uniformly distributed on the integers 0 to $m_1 - 2$.

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

Let $X_{i,1}, X_{i,2}, \ldots X_{i,k}$ be i^{th} output from *k* different multiplicative congruential generators, where the j^{th} generator has prime modulus m_j and multiplier a_j is chosen so that the period is $m_j - 1$. Then the j^{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,

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

The maximum possible period for a generator is

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

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

Example 4.3
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

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

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

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

4.3 Tests for Random Numbers

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

Testing for Uniformity
The hypotheses are as follows

$H_0 : R_i \sim U [0, 1]$
$H_1 : R_i \quad U [0, 1]$

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

$H_0 : R_i \sim$ independently
$H_1 : R_i$ 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 α must be stated.

$$\text{Level of significance } \alpha = \frac{\text{probability of rejecting the test}}{\text{probability of accepting the test}}$$
$$= P \text{ (reject } H_0 \mid H_0 \text{ true)}$$

Frequently, α 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.

1) **Frequency test** - Compares the distribution of set of numbers generated to a uniform distribution by using the Kolmogorov-Smirnov or the chi-square test.
2) **Runs test** - Tests the runs up and down or the runs above and below the mean by comparing the actual values to expected values. The statistic for comparison is the chi-square test.
3) **Autocorrelation test** - The correlation between numbers is tested and compares the sample correlation to the expected correlation of zero.
4) **Gap test -** Counts the number of digits that appear between repetitions of a particular digit and then uses the Kolmogorov-Smirnov test to compare with the expected size of gaps.
5) **Poker test** - Treats the numbers grouped together as a poker hand. Then the hands obtained are compared to what is expected using the chi-square test.

4.3.1 Frequency Tests

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

1. Kolmogorov-Smirnov test
2. Chi-Square test

1. 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
$S_N(x) \rightarrow$ Empirical cdf
$N \rightarrow$Total number of observations
$R_1, R_2 \ldots R_N \rightarrow$ Samples from Random generator
$D \rightarrow$ Sample statistic
$D_\alpha \rightarrow$ Critical value

By definition,

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

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

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^{th} smallest observation, so that

$$R_{(1)} \leq R_{(2)} \leq \ldots\ldots \leq R_{(N)}$$

Step 2 – Compute

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

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

Step 3 – Compute

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

Step 4 – Determine the critical value D_α, from the table A.8 for the specified significance level α and the given sample size N.

Step 5

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

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

Example 4.4

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 $\alpha = 0.05$.

Solution

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

Arrange R_i from smallest to largest

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
Step 2 -	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

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

$D^- = \max \{R_i – [(i-1)/N]\}$

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

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

$D_{\alpha} = D_{0.05} = 0.565$ (from table A.8)

$D < D_{\alpha} \rightarrow 0.26 < 0.565$

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

The calculations in the above table are depicted in the fig 4.2, where empirical cdf $S_N(x)$ is compared to uniform cdf F(x). It is seen that D^+ is the largest deviation of $S_N(x)$ above F(x) and D^- is the largest deviation of $S_N(x)$ below F(x).

Fig 4.2 Comparison of F(x) and $S_N(x)$

2. 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 → observed number in i^{th} class
E_i → expected number in i^{th} class
n → number of classes

For uniform distribution E_i is given by

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

Where N →Total number of observations

It can be shown that the sampling distribution χ_0^2 is approximately the chi-square distribution with n-1 degrees of freedom (i.e., $\chi_0^2 \leq \chi_{\alpha,n-1}^2$)

Example 4.5
Use chi-square test with α=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
0.6 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

Solution

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

$X^2_{0.05,\,9}$ =16.9 (check the table A.6 –using α, n-1)

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

Therefore null hypothesis of uniform distribution is not rejected.

Interval	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$X_0^2 = (O_i-E_i)^2 / E_i$
0.01 - 0.10	8	10	-2	4	0.4
0.11 - 0.20	8	10	-2	4	0.4
0.21 - 0.30	10	10	0	0	0.0
0.31 - 0.40	9	10	-1	1	0.1
0.41 - 0.50	12	10	2	4	0.4
0.51 - 0.60	8	10	-2	4	0.4
0.61 - 0.70	10	10	0	0	0.0
0.71 - 0.80	14	10	4	16	1.6
0.81 - 0.90	10	10	0	0	0.0
0.91 - 1.00	11	10	1	1	0.1
	100	100	0		$X_0^2 = 3.4$

Table 4.2 Computation of chi-square test

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 $N \geq 50$.

4.3.2 Runs Tests

Run - The succession of similar events preceded and followed by a different event is called as run.

Run length - Number of events that occur in the run.

Example 4.6: Tossing coin
Consider the sequence of tossing a coin 10 times: H T T H H T T T H T

No.	Run length	Run
1	1	H
2	2	T T
3	2	H H
4	3	T T T
5	1	H
6	1	T

There are two possible concerns in run tests. They are
1. Number of runs- Run up and down & Runs above and below mean
2. Length of runs

1. Runs up and down

a. **Up run**-Sequence of numbers each of which is succeeded by larger number is called as up run.

b. **Down run**-Sequence of numbers each of which is succeeded by smaller number is called as down run.

c. If a number is followed by a larger number then it denoted by '+'. If followed by smaller number then by '-'.

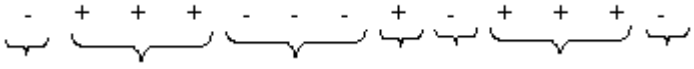
To illustrate the above, consider the sequence of numbers

0.87 0.15 0.23 0.45 0.69 0.32 0.30 0.19 0.24 0.18 0.65 0.82 0.93 0.22

The up run and down run are marked as

$^{-}$0.87 $^{+}$0.15 $^{+}$0.23 $^{+}$0.45 $^{-}$0.69 $^{-}$0.32 $^{-}$0.30 $^{+}$0.19 $^{-}$0.24 $^{+}$0.18 $^{+}$0.65 $^{+}$0.82 $^{-}$0.93 0.22

The sequence of '+' and '-'are



It has 7 runs, first run of length one, second run of length three, third run of length 3, fourth run with one, fifth run with one, sixth run with three and seventh run with one. There are three up runs and four down runs.

If N is number of numbers in sequence, then maximum numbers of runs is N-1 and minimum number of runs is one.

If 'a' is total number of runs in a random sequence, Mean is given by

$$\mu_a = \frac{(2N-1)}{3}$$

Variance,

$$\sigma_a^2 = \frac{16N - 29}{90}$$

For N > 20, the distribution of 'a' is reasonably approximated by a normal distribution, $N(\mu_a, \sigma_a^2)$. This approximation is used to test the independence of numbers from a generator.

The test statistic is obtained by subtracting the mean from observed number of runs 'a' and dividing by standard deviation, i.e. Test statistic is given by,

$$Z_0 = \frac{a - \mu_a}{\sigma_a}$$

Substituting μ_a and σ_a in above equation, we get

$$Z_0 = \frac{a - [(2N-1)/3]}{\sqrt{[(16N-29)/90]}}$$

Where $Z_0 \sim N(0, 1)$

The null hypothesis is accepted when $-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2}$, where α is the level of significance. The critical values and rejection region is shown in the fig 4.3.

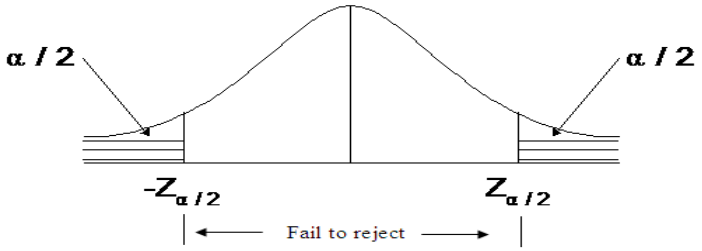


Fig 4.3 Accept null hypothesis

Example 4.7

Based on runs up and runs down, determine whether the following sequence of 40 numbers is such that the hypothesis of independence can be rejected or accepted where $\alpha = 0.05$.

0.41	0.68	0.89	0.94	0.74	0.91	0.55	0.62	0.36	0.27
0.19	0.72	0.75	0.08	0.54	0.02	0.01	0.36	0.16	0.28

0.18 0.01 0.95 0.69 0.18 0.47 0.23 0.32 0.82 0.53
0.31 0.42 0.73 0.04 0.83 0.45 0.13 0.57 0.63 0.29

Solution

The sequence of runs up and down is as follows:

\+ + + - + - + - - - + + - + - - + - +
\- - + - - + - + + - - + + - + - - + + -

No. of runs → a = 26

N = 40

$\mu_a = \{2(40) - 1\} / 3 = 26.33$

$\sigma_a^2 = \{16(40) - 29\} / 90 = 6.79$

$Z_0 = (26 - 26.33) / \sqrt{(6.79)} = -0.13$

Critical value → $Z\alpha_{/2} \rightarrow Z_{0.025} = 1.96$ (from table A.3)

$-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2} \rightarrow -1.96 \leq -0.13 \leq 1.96$

Therefore independence of the numbers cannot be rejected, we accept null hypothesis.

Disadvantage of runs up and down

- ✓ Insufficient to review the independence of group of numbers

2. Runs above and below the mean

Runs are described with above/below the mean value. A '+' sign is used to indicate above mean and '-'sign for below the mean.

To illustrate the above, consider the sequence of 2-digit random numbers

0.40 0.84 0.75 0.18 0.13 0.92 0.57 0.77 0.30 0.71
0.42 0.05 0.78 0.74 0.68 0.03 0.18 0.51 0.10 0.37

Mean = (0.99+0.00)/2 = 0.495

The runs above and below mean are marked as

-0.40 +0.84 +0.75 -0.18 -0.13 +0.92 +0.57 +0.77 -0.30 +0.71
-0.42 -0.05 +0.78 +0.74 +0.68 -0.03 -0.18 +0.51 -0.10 -0.37

The sequence of '+' and '-'are

\- + + - - + + + - + - - + + + - - + - -

There are 11 runs, of which 5 are above mean and 6 runs below mean.

Let $n_1 \rightarrow$ No. of individual observations above mean
$n_2 \rightarrow$ No. of individual observations below mean
b → Total number of runs
N→ Maximum number of runs, where $N = n_1 + n_2$

The mean is given by

$$\mu_b = \frac{2n_1n_2}{N} + \frac{1}{2}$$

Variance

$$\sigma_b^2 = \frac{2n_1n_2\,(2n_1n_2 - N)}{N^2\,(N-1)}$$

For either n_1 or n_2 greater than 20, b is approximately normally distributed. The test statistic is obtained by subtracting mean from number of runs 'b' and dividing by standard deviation i.e.

$$Z_0 = \frac{b - (2n_1n_2/N) - 1/2}{\left(\frac{2n_1n_2\,(2n_1n_2 - N)}{N^2\,(N-1)}\right)^{1/2}}$$

The null hypothesis is accepted when $-Z_{\alpha/2} \le Z_0 \le Z_{\alpha/2}$, where α is the level of significance.

Example 4.8

Based on runs above and below mean, determine whether the following sequence of 40 numbers is such that the hypothesis of independence can be rejected or accepted where $\alpha = 0.05$.

0.41	0.68	0.89	0.94	0.74	0.91	0.55	0.62	0.36	0.27
0.19	0.72	0.75	0.08	0.54	0.02	0.01	0.36	0.16	0.28
0.18	0.01	0.95	0.69	0.18	0.47	0.23	0.32	0.82	0.53
0.31	0.42	0.73	0.04	0.83	0.45	0.13	0.57	0.63	0.29

Solution

Mean= 0.495

The sequence of runs above and below mean is as follows:

\- + + + + + + + - - - + + - + - - - - -

\- - + + - - - - + + - - + - + - - + + -

$n_1 = 18$

$n_2 = 22$

$N = n_1 + n_2 = 40$

b = 17

$\mu_b = [\{2(18)\,(22)\}\,/\,40] + (1/2) = 20.3$

$\sigma_b^2 = [2\,(18)\,(22)\,\{(2)\,(18)\,(22) - 40\}] / [(40)^2\,(40-1)] = 9.54$

Since $n_2 > 20$, normal approximation is accepted.

$Z_0 = (17 - 20.3) / \sqrt{(9.54)} = -1.07$

Critical value $\rightarrow Z\alpha_{/2} \rightarrow Z_{0.025} = 1.96$ (from table A.3)

$-Z_{\alpha/2} \le Z_0 \le Z_{\alpha/2} \rightarrow -1.96 \le -1.07 \le 1.96$

Therefore hypothesis of independence cannot be rejected on the basis of this test.

Disadvantage of runs above and below mean

- ✓ If two numbers are below mean, two numbers are above mean and so on. Then the numbers are dependent.

3. Runs test: length of runs

Let Y_i be the number of runs of length i, in a sequence of N numbers. For an independent sequence,

The expected value of Y_i for runs up and down is given by

$$E(Y_i) = \frac{2}{(i+3)!}[N(i^2+3i+1)-(i^3+3i^2-i-4)], \quad i \leqslant N-2$$

$$E(Y_i) = \frac{2}{N!}, \quad i = N-1$$

For runs above and below mean, the expected value of Y_i is given by

$$E(Y_i) = \frac{Nw_i}{E(I)}, \quad N > 20$$

Where w_i, the approximate probability that a run has length i, is given by

$$w_i = \left(\frac{n_1}{N}\right)^i \left(\frac{n_2}{N}\right) + \left(\frac{n_1}{N}\right)\left(\frac{n_2}{N}\right)^i, \quad N > 20$$

And E (I), the approximate expected length of a run, is given by

$$E(I) = \frac{n_1}{n_2} + \frac{n_2}{n_1}, \quad N > 20$$

The approximate expected total number of runs (of all lengths) E (A), is given by

$$E(A) = \frac{N}{E(I)}, \quad N > 20$$

The appropriate test is chi-square test with O_i, the observed number of runs of length i. The test statistic is given by

$$X_0^2 = \sum_{i=1}^{L} \frac{[O_i - E(Y_i)]^2}{E(Y_i)}$$

Where L = N – 1 for runs up and down
L = N for runs above and below mean.

If null hypothesis of independence is true then X^2_0 is approximately chi-squared distributed with L-1 degrees of freedom.

Example 4.9

Given the sequence of numbers, can the hypothesis that the numbers are independent be rejected on the basis of length of runs up and down at $\alpha = 0.05$?

0.30	0.48	0.36	0.01	0.54	0.34	0.96	0.06	0.61	0.85
0.48	0.86	0.14	0.86	0.89	0.37	0.49	0.60	0.04	0.83
0.42	0.83	0.37	0.21	0.90	0.89	0.91	0.79	0.57	0.99
0.95	0.27	0.41	0.81	0.96	0.31	0.09	0.06	0.23	0.77
0.73	0.47	0.13	0.55	0.11	0.75	0.36	0.25	0.23	0.72
0.60	0.84	0.70	0.30	0.26	0.38	0.05	0.19	0.73	0.44

Solution

N = 60

The sequence of + and – are as follows

\+ - - + - + - + + - + - + + - + + - + - + - - + - + - - +
\- - + + + - - - + + - - - + - + - - - + - + - - - + - + + -

The length of runs in the sequence is as follows

1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 2, 1, 1,
1, 2, 1, 2, 3, 3, 2, 3, 1, 1, 1, 3, 1, 1, 1, 3, 1, 1, 2, 1

Calculate O_i

Run Length, i	1	2	3	4
Observed Runs, O_i	26	9	5	0

The expected value of Y_i,

For run length one,

$$E(Y_1) = \frac{2}{(1+3)!}[60(1^2 + 3(1) + 1) - (1^3 + 3(1)^2 - 1 - 4)] = 25.08$$

Run length two,

$$E(Y_2) = \frac{2}{(2+3)!}[60(2^2 + 3(2) + 1) - (2^3 + 3(2)^2 - 2 - 4)] = 10.77$$

Run length three

$$E(Y_3) = \frac{2}{(3+3)!}[60(3^2 + 3(3) + 1) - (3^3 + 3(3)^2 - 3 - 4)] = 3.04$$

$$\therefore \quad E(Y_1) + E(Y_2) + E(Y_3) = 38.89$$

We find mean (runs up and down)

$$\mu_a = \frac{2N - 1}{3} = \frac{2(60) - 1}{3} = 39.67$$

Expected value, when $i \geq 4$

$$\mu_a - \sum_{i=1}^{3} E(Y_i) = 39.67 - 38.89 = 0.78$$

To find X^2_0, the calculations and procedures are shown in table 4.3

Run length (i)	Observed number of runs (O_i)	Expected number of runs $E(Y_i)$	$\frac{[O_i - E(Y_i)]^2}{E(Y_i)}$
1	26	25.08	0.03
2	9	10.77	
3	5 } 14	3.82 } 14.59	} 0.02
4	0	0.78	
-	40	39.67	$X_0^2 = 0.05$

Table 4.3 Length of runs up and down: X_0^2 test

$X^2_{0.05,\,1} = 3.84$

$X_0^2 < X^2_{0.05,\,1} = 0.05 < 3.84$

Therefore hypothesis of independence is accepted.

Example 4.10

Given the sequence of numbers can the hypothesis that the numbers are independent be rejected on the basis of length of runs above and below mean at α = 0.05?

0.30	0.48	0.36	0.01	0.54	0.34	0.96	0.06	0.61	0.85
0.48	0.86	0.14	0.86	0.89	0.37	0.49	0.60	0.04	0.83
0.42	0.83	0.37	0.21	0.90	0.89	0.91	0.79	0.57	0.99
0.95	0.27	0.41	0.81	0.96	0.31	0.09	0.06	0.23	0.77
0.73	0.47	0.13	0.55	0.11	0.75	0.36	0.25	0.23	0.72
0.60	0.84	0.70	0.30	0.26	0.38	0.05	0.19	0.73	0.44

Solution

N = 60

Mean = (0.99+0.00)/2 = 0.495

The sequence of + and – are as follows

\- - - - + - + - + + - + - + + - - + - + - + - - +
\+ + + + + + - - + + - - - - + + - - + - + - - - +
\+ + + - - - - - + -

$n_1 = 28$

$n_2 = 32$

$N = n_1 + n_2 = 60$

The length of runs in the sequence is as follows

4, 1, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 7, 2, 2, 4, 2, 2, 1, 1, 1, 3, 4, 5, 1, 1

Calculate O_i

Run Length, i	1	2	3	≥4
Observed Runs, O_i	17	8	1	5

The probabilities of runs of various lengths w_i are as follows

$$w_1 = \left(\frac{28}{60}\right)^1 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^1 = 0.498$$

$$w_2 = \left(\frac{28}{60}\right)^2 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^2 = 0.249$$

$$w_3 = \left(\frac{28}{60}\right)^3 \frac{32}{60} + \frac{28}{60}\left(\frac{32}{60}\right)^3 = 0.125$$

The expected length of run E (I) is

$$E(I) = \frac{28}{32} + \frac{32}{28} = 2.02$$

The expected number of runs of various lengths is

$$E(Y_1) = \frac{N w_1}{E(I)} = \frac{60(0.498)}{2.02} = 14.79$$

$$E(Y_2) = \frac{N w_2}{E(I)} = \frac{60(0.249)}{2.02} = 7.40$$

$$E(Y_3) = \frac{N w_3}{E(I)} = \frac{60(0.125)}{2.02} = 3.71$$

The expected total number of runs is

$$E(A) = \frac{N}{E(I)} = \frac{60}{2.02} = 29.7$$

For $i \geq 4$,

$$E(A) - \sum_{i=1}^{3} E(Y_i) = 29.7 - 25.9 = 3.8$$

To find X_0^2, the calculations and procedures are shown in table 4.4

Run length (i)	Observed number of runs (O_i)	Expected number of runs $E(Y_i)$	$\frac{[O_i - E(Y_i)]^2}{E(Y_i)}$

1 2 3 ≥ 4	17 8 1 5 } 6	14.79 7.40 3.71 3.80 } 7.51	0.33 0.05 } 0.30
-	31	29.70	$X_0^2 = 0.68$

Table 4.4 Length of runs above and below mean: X_0^2 test

$$X^2_{0.05,2} = 5.99$$

$$X^2_0 < X^2_{0.05,2} = 0.68 < 5.99$$

Therefore hypothesis of independence is accepted.

4.3.3 Tests for Autocorrelation

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

For example

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

Even though the numbers satisfies all the previous test, still we find that every 5th, 10th, … numbers are larger numbers in that position. Hence the numbers are dependent. So this test is preferable.

The notations used are

ρ_{im} → autocorrelation

m → lag, the places where dependency exists. The numbers at these places can be smaller or larger.

i → starting position of observation

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

N →Total number of values in the sequence

$R_i, R_{i+m}, R_{i+2m} \ldots R_{i+(M+1)m}$ → Random numbers

The null hypothesis of independence is given by

$H_0 : \rho_{im} = 0$

$H_1 : \rho_{im} \neq 0$

$$\rho_{im} \approx \hat{\rho}_{im}$$

For large values the above is approximately normal if the values $R_i, R_{i+m}, R_{i+2m} \ldots R_{i+(M+1)m}$ are uncorrelated. The test statistic can be formed as

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

This is distributed normally with mean 0 and variance 1.

Where

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

Standard deviation of estimator is

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

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

If $\rho_{im} > 0$, the subsequence is exhibiting *positive autocorrelation*, which means high random numbers tends to be followed by high ones and low followed by low.
If $\rho_{im} < 0$, the subsequence is exhibiting *negative autocorrelation*, which means low random numbers tends to be followed by high ones and vice versa.

Example 4.11
Test whether the 3rd, 8th, 13th 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

Solution
$\alpha = 0.05$
i = 3 (beginning with 3rd number)
m = 5 (every five numbers)
N = 30
M= 4 (largest integer $3+(M+1)5 \le 30$)

Arranging the numbers to be observed with i = 3 & m = 5

0.28 0.33 0.27 0.05 0.36

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

$$= -0.1945$$

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

The test statistic,

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

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.

4.3.4 Gap Test

The Gap test measures the number of digits between successive occurrences of the same digit. A gap of length x occurs between the occurrences of some specified digit. The probability of gap is determined as

$$P\ (t \text{ followed by exactly x non-t digits}) = (0.9)^x\ (0.1),\ x=0, 1, 2,\ldots$$

For example
Length of gaps associated with the digit 3.

4, 1, 3, 5, 1, 7, 2, 8, 2, 0, 7, 9, 1, 3, 5, 2, 7, 9, 4, 1, 6, 3
3, 9, 6, 3, 4, 8, 2, 3, 1, 9, 4, 4, 6, 8, 4, 1, 3, 8, 9, 5, 5, 7
3, 9, 5, 9, 8, 5, 3, 2, 2, 3, 7, 4, 7, 0, 3, 6, 3, 5, 9, 9, 5, 5
5, 0, 4, 6, 8, 0, 4, 7, 0, 3, 3, 0, 9, 5, 7, 9, 5, 1, 6, 6, 3, 8
8, 8, 9, 2, 9, 1, 8, 5, 4, 4, 5, 0, 2, 3, 9, 7, 1, 2, 0, 3, 6, 3

There are eighteen 3's in list. Therefore 17 gaps, the first gap is of length 10, second gap is of length is length 7 and so on. We are interested in the frequency of gaps.
The probability of first gap is determined as $P(\text{gap of } 10)=(0.9)^{10}\,(0.1)$

The observed frequencies for all digits are compared to the theoretical frequency using the Kolmogorov-Smirnov test. The theoretical frequency distribution for random ordered digits is given by

$$P(\text{gap} \leqslant x) = F(x) = 0.1 \sum_{n=0}^{x} (0.9)^n = 1 - 0.9^{x+1}$$

Procedure for the test is as follows

Step 1 – Specify the cdf for the theoretical frequency distribution, $F(x) = 1 - 0.9^{x+1}$

Step 2 –Arrange the observed sample of gaps in a cumulative distribution with these same classes.

Step 3 – Find D, the maximum deviation between F(x) and $S_N(x)$ as
$D = \max |F(x) - S_N(x)|$

Step 4 – Determine the critical value D_α , from table A.8 for the specified value of α and the sample size N.

Step 5 –If the calculated value of D is greater than the tabulated value of D_α, the null hypothesis of independence is rejected.

Example 4.12
Based on the frequency with which gaps occur analyze the following 110 digits to test whether they are independent. Use α= 0.05.

4, 1, 3, 5, 1, 7, 2, 8, 2, 0, 7, 9, 1, 3, 5, 2, 7, 9, 4, 1, 6, 3
3, 9, 6, 3, 4, 8, 2, 3, 1, 9, 4, 4, 6, 8, 4, 1, 3, 8, 9, 5, 5, 7
3, 9, 5, 9, 8, 5, 3, 2, 2, 3, 7, 4, 7, 0, 3, 6, 3, 5, 9, 9, 5, 5
5, 0, 4, 6, 8, 0, 4, 7, 0, 3, 3, 0, 9, 5, 7, 9, 5, 1, 6, 6, 3, 8
8, 8, 9, 2, 9, 1, 8, 5, 4, 4, 5, 0, 2, 3, 9, 7, 1, 2, 0, 3, 6, 3

Solution

Digit	0	1	2	3	4	5	6	7	8	9
No. of Gaps	7	8	8	17	10	13	7	8	9	13

Total No. of gaps = 110 – 10 = 100

The numbers of gaps associated with the various digits are as follows in table 4.5

Gap Length	Frequency	Relative Frequency	Cumulative Frequency	F(x)	$\lvert F(x) - S_N(x) \rvert$
0-3	35	0.35	0.35	0.3439	0.0061
4-7	22	0.22	0.57	0.5695	0.0005
8-11	17	0.17	0.74	0.7176	0.0224
12-15	9	0.09	0.83	0.8147	0.0153
16-19	5	0.05	0.88	0.8784	0.0016
20-23	6	0.06	0.94	0.9202	0.0198
24-27	3	0.03	0.97	0.9497	0.0223
28-31	0	0.00	0.97	0.9657	0.0043

32-35	0	0.00	0.97	0.9775	0.0075
36-39	2	0.02	0.99	0.9852	0.0043
40-43	0	0.00	0.99	0.9903	0.0003
44-47	1	0.01	1.00	0.9936	0.0064

Table 4.5 Gap test example

The critical value of D is given by $D_{0.05} = 1.36 / \sqrt{100} = 0.136$

$D = \max |F(x) - SN(x)| = 0.0224$

$D < D_{0.05}$, Therefore null hypothesis of independence is not rejected.

4.3.5 Poker Test

Is based on the frequency with which certain digits are repeated in a series of number.

For example

0.255 0.577 0.331 0.414 0.828 0.909 0.303 0.001 …

Note: a pair of like digits appears in each number generated.

There are only three possibilities in a 3-digit numbers. They are

1. Individual numbers can all be different.

P(3 different digits) = P(2nd different from 1st) * P(3rd different from 1st & 2nd)
= (0.9) (0.8) = 0.72

2. Individual numbers can all be same.

P(3 like digits) = P(2nd digit same as 1st) * P(3rd digit same as 1st)
= (0.1) (0.1) = 0.01

3. There can be one pair of like digits.

P(exactly one pair) = 1 - 0.72 - 0.01 = 0.27

Example 4.13
A sequence of 1000 three-digit numbers has been generated and an analysis indicates that 680 have three different digits, 289 contain exactly one pair of like digits, and 31 contain three like digits. Based on the poker test, are these numbers independent?

Solution
Let $\alpha = 0.05$
The test is summarized in table 4.6

Combination, i	Observed frequency, O_i	Expected frequency, E_i	$\frac{[O_i - E_i]^2}{E_i}$
Three different digits	680	720	2.22
Three like digits	31	10	44.10
Exactly one pair	289	270	1.33
	------	------	-------
	1000	1000	$47.65 = \chi_0^2$

Table 4.6 Poker test results

$$X_0^2 > X_{0.05,2}^2 = 47.65 > 5.99$$

Therefore the independence of the numbers is rejected on the basis of this test.

Unit 5

<u>Random Variate Generation</u>

5.1 Inverse Transform Technique
- *5.1.1 Exponential Distributions*
- *5.1.2 Uniform Distributions*
- *5.1.3 Triangular Distributions*
- *5.1.4 Discrete Distribution*
  - *Empirical Distribution*
  - *Uniform Distributions*
  - *Geometric Distributions*

5.2 Acceptance-Rejection Technique
- *5.2 1 Poisson Distributions*

All the techniques for generating random variates assumes that the uniform (0, 1) random numbers R_1, R_2 … is readily available, where each R_i has
probability density function (pdf)

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

and cumulative distribution function (cdf)

$$F_R(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leqslant x \leqslant 1 \\ 1, & x > 1 \end{cases}$$

5.1 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.

5.1.1 Exponential Distribution

The exponential distribution has

pdf

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

cdf

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

Where λ→ Mean number of occurrences per unit time

Example 5.1
If interarrival times X_1, X_2, X_3... had an exponential distribution with rate λ, then λ can be interpreted as mean number of arrivals per unit time. For any *i*,

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

Where 1/λ → Mean interarrival time
Therefore a procedure can be developed for generating values X_1, X_2, X_3... 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

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

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

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

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

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 ... 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 …

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

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

5.1.2 *Uniform Distribution*

Let X → Random variable, uniformly distributed on interval [a, b].

R → Random number

A valid guess for generating X is given by

$$X = a + (b - a) R \qquad (5.3)$$

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

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

Step2 – Set

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

Step3 – Solve for X in terms of R, Then

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

Therefore $X = a + (b - a)\,R$

5.1.3 Triangular distribution

The pdf,

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

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

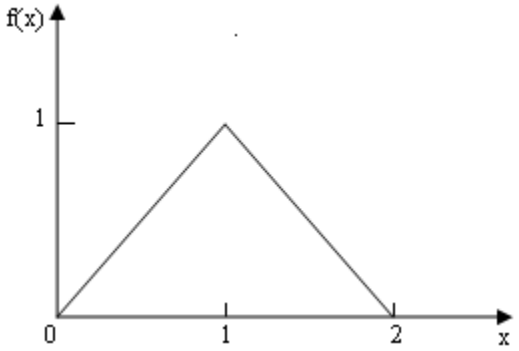


Fig 5.1 Density functions for triangular distribution.

The cdf,

$$F(x) = \begin{cases} 0, & x \leqslant 0 \\ \dfrac{x^2}{2}, & 0 < x \leqslant 1 \\ 1 - \dfrac{(2-x)^2}{2}, & 1 < x \leqslant 2 \\ 1, & x > 2 \end{cases}$$

For $0 \leq X \leq 1$,

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

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

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

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

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

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

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

Thus X is generated by

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

5.1.4 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.

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

Table 5.1 Distributions of number of shipments X

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

Table 5.2 Generating the discrete variate X

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,

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

The graphical representation of cdf is shown in fig 5.2

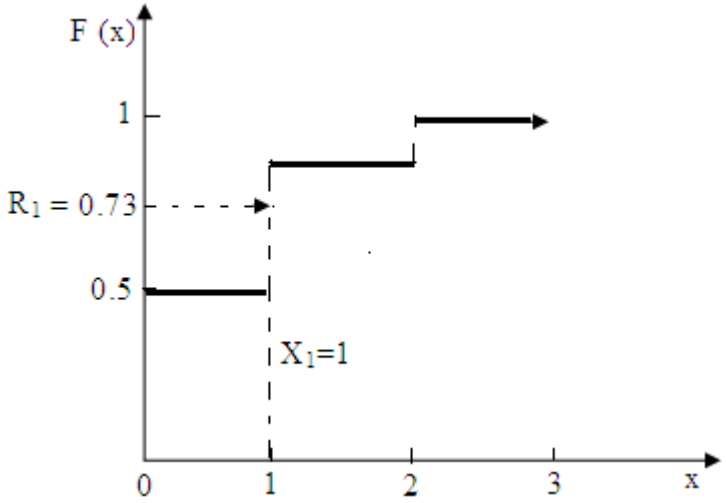


Fig 5.2 The cdf of number of shipments X

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 $x_1, x_2 \ldots x_n$ are possible values of random variable and

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

For this example
$n = 3$, $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, hence $r_1 = 0.5$, $r_2 = 0.8$, $r_3 = 1.0$

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

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

Example ***5.3 (Discrete Uniform Distributions)***

Consider discrete uniform distribution on (1, 2, … k) with

pmf => $p(x) = 1 / k$, $x = 1, 2, \ldots, k$

cdf,

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

Let us consider

$x_i = i$

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

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

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

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

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

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)

For example $X_1 = \lceil 7.8 \rceil = 8$

Example 5.4

Consider discrete distribution with pmf given by

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

For integer values of *x* in the range {1, 2, … 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

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

$$\longrightarrow \underbrace{(x-1)\,x < k\,(k+1)R} \leqslant x\,(x+1)$$

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 $X = \lceil x - 1 \rceil$

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

Example ***5.5 (Geometric Distribution)***

Consider the geometric distribution with pmf

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

cdf,

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

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

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

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

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

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

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

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

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

$\div$ by $\ln(1-p)$

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

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

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

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 …} with pmf

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

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

Example 5.6
Generate 3 values from a geometric distribution on the range {X ≥ 1} with mean 2. Such a geometric distribution has pmf $p(x) = p(1-p)^{x-1}$, where x= 1,2,… with mean 1/p = 2. [Random numbers $R_1 = 0.932$, $R_2 = 0.105$, $R_3 = 0.687$ from table A.2]

Solution
p = 1/2
q = 1

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

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

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

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

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

5.2 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 $R \geq ¼$, accept X = R and then go to step3
b) If $R < ¼$, reject R, and return to step1

Step3 – If another uniform random variate on [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[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 $\{R \geq ¼\}$, have the desired distribution.

To show this

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

then

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

This is the probability distribution of R, given that R is between ¼ 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^{th} generation of random variates.

5.2.1 Poisson Distribution

The Poisson distribution has pmf,

$$p(n) = P(N = n) = \frac{e^{-\alpha}\alpha^{n}}{n!}, \quad n = 0,1,2, \ldots$$

Where N→ Poisson random variate
α → mean, α > 0

N can be interpreted as number of arrivals from poisson arrival process in one unit of time. In exponential distribution, α is the mean number of arrivals per unit time. Thus there is a relationship between (discrete) poisson distribution and (continuous) exponential distribution. i.e. N = n if and only if

$$A_1 + A_2 + \ldots + A_n \leq 1 < A_1 + \ldots + A_n + A_{n+1} \qquad (5.7)$$

Where
$A_1, A_2, \ldots$→ inter arrival times of customers, exponentially distributed with rate α.

N = n, says exactly n arrivals during one unit of time but $A_1 + A_2 + \ldots + A_n \leq 1 < A_1 + \ldots + A_n + A_{n+1}$ says that n^{th} arrival occurred before time 1, whereas $(n + 1)^{st}$ arrival occurred after time 1. Therefore these two statements are equivalent. Equation (5.7) is simplified by using equation (5.2), we get

$$A_i = \left(-\frac{1}{\alpha}\right) \ln R_i \qquad (5.8)$$

Substitute equation (5.8) in equation (5.7), to obtain

$$\sum_{i=1}^{n} -\frac{1}{\alpha} \ln R_i \leq 1 < \sum_{i=1}^{n+1} -\frac{1}{\alpha} \ln R_i$$

Now multiply throughout by − α, which reverses the inequality sign and sum of logarithm to product of logarithm

$$\ln \prod_{i=1}^{n} R_i \geq -\alpha > \ln \prod_{i=1}^{n+1} R_i$$

Use the relation $e^{\ln x} = x$, then

$$\prod_{i=1}^{n} R_i \geq e^{-\alpha} > \prod_{i=1}^{n+1} R_i$$

This is equivalent to relation (5.7)

Procedure for generating a poisson random variate *N*

Step1 – Set n = 0, P = 1

Step2 – Generate random number R_{n+1} and replace P by P. 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 numbers are required, so the average number is given by E (N + 1) = α + 1.

Example 5.7
Generate three poisson variates with mean α =0.2 (Random Numbers are to be taken from Table A.2)

Solution

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

Step1 – Set n = 0, P = 1
Step2 – $R_1 = 0.4357$, 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 – $R_1 = 0.4146$, 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 – $R_1 = 0.8353$, P = 0.8353
Step3 – Since $P \geq e^{-\alpha}$, reject n = 0, return to step2 with n = 1.

Step2 – $R_2 = 0.9952$, P = P.R_2 = 0.8353 * 0.9952 = 0.8312
Step3 – Since $P \geq e^{-\alpha}$, reject n = 1, return to step2 with n = 2.

Step2 – $R_3 = 0.8004$, P = P.R_3 = 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 α =0.2, has to be generated then 1000(α + 1) =1200 random number are required.

Unit 6

Input Modeling

Input data provide the driving force for a simulation model. In queuing system the distribution of time between arrivals and service times are the input data. The distributions of demand and lead time are the input data for inventory system.

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

1. Data Collection
2. Identifying the distribution with data
3. Parameter Estimation
4. Goodness of fit tests

6.1 Data Collection

Data collection is one of the biggest tasks in solving a real problem. Collection of data requires a substantial time and resource commitment. The collection of data is from the real system of interest. In some situation it is not possible to collect the data. When data are not available, expert opinion and knowledge of the process must be used to make educated guess.

"GIGO or garbage - in, garbage - out" is the basic concept in computers. In case input data are not accurately collected and analyzed then simulation output data will result in misleading and possibly damage with increase in cost factor.

To enhance and facilitate data collection, the following suggestions are required.

1. A useful expenditure of time is in planning.
    - Data collection should start before observing the process. Devise the forms for this purpose. These forms have to be modified several times before actual data collection begins.
    - Check for unusual circumstances and consider how they will be handled.
    - If data is already present then plenty of time is required for converting the data into a usable format.
2. Try to analyze the data that are collected i.e. whether the data collected is adequate for simulation. Check the data that are useless for the system.
3. Try to combine homogeneous data sets; an initial test is to see if means of distribution are same.
4. Be aware of possibility of data censoring. Censoring → whether the part or whole data is accepted. Censoring results in long process time, so collect only relevant data.
5. Build a scatter diagram to determine the relationship between two variables. Scatter diagram – A dotted graph with one variable on x-axis and one variable on y-axis.
6. Observe the sequence of inputs for autocorrelation. If service time of i^{th} customer affects the 'i + n' customer then there is an autocorrelation.
7. Keep in mind the difference between the input data from output or performance data. Make sure to collect input data which represents the uncertain quantities that are largely beyond the control of system and will not be altered by changes made to improve the system.

6.2 Identifying the distribution with data

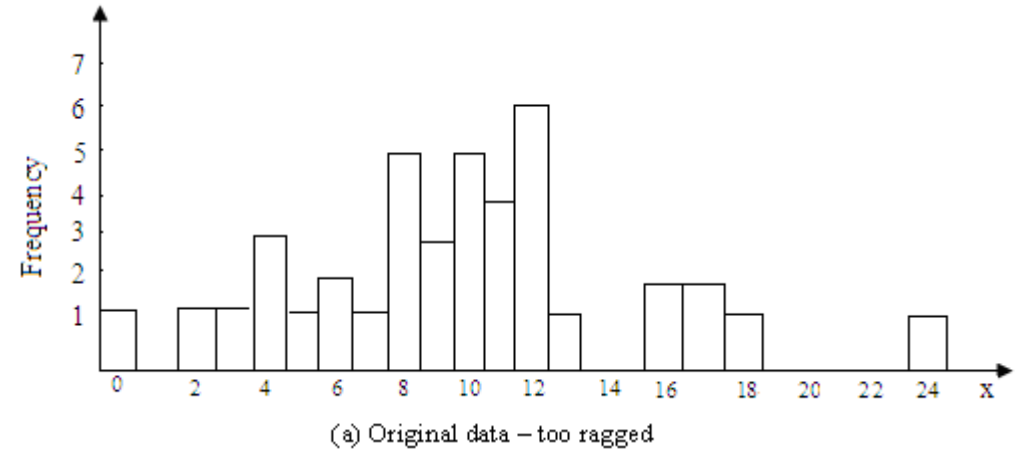
When the data is available, this step begins by developing a frequency distribution or histogram, of the data. Based on the frequency distribution and structural knowledge of the process, a family of distribution is chosen.

6.2.1 Histogram

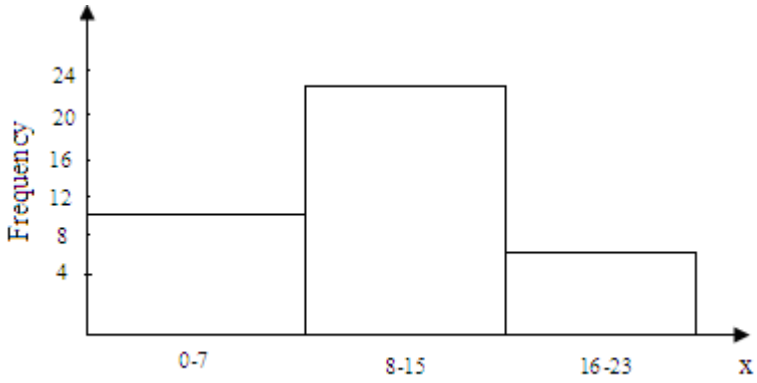
It is useful in identifying the shape of a distribution. The procedure for construction is as follows.

1. Divide the range of the data into intervals of equal width (unequal width may be used).
2. Label the horizontal axis to conform to the intervals selected.
3. Determine the frequency of occurrences within each interval.
4. Label the vertical axis. So that total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

The number of class intervals depends on number of observations and dispersion in the data. If intervals are too wide, histogram will be *coarse or blocky*. If intervals are too narrow, the histogram will be *ragged*. The fig 6.1 shows the different forms of histogram.



(a) Original data – too ragged



(b) Combining adjacent cells – too coarse

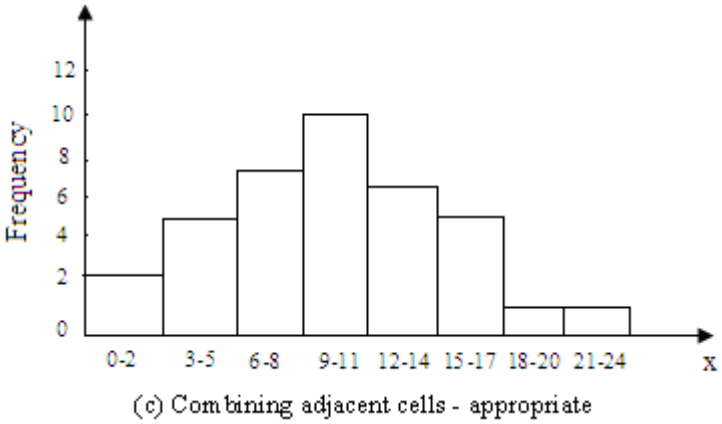


Fig 6.1 Ragged, coarse, and appropriate histograms

Histogram for continuous data corresponds to pdf of a theoretical distribution; if continuous, a line is drawn through the center point of each class interval frequency.

Histograms for discrete data corresponds to pmf of distribution, it has large number of data points so it should have a cell for each value in the range of data. If there are few points then combine adjacent cells to eliminate ragged appearance of histogram.

Example 6.1 (Discrete data)
The number of vehicles arriving at the corner of the road from 7 am to 7.05 am was monitored for five working days over 20 week period. The table 6.1 shows the resulting data with first entry – 12 periods each of 5 minutes during which zero vehicles arrived, 10 periods during which one vehicle arrived and so.

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

T able 6.1 Number of arrivals in a 5 minute period

The number of vehicles is a discrete variable; since there are sample data the histogram can have a cell for each possible value in the range of data. The resulting histogram is shown in fig 6.2

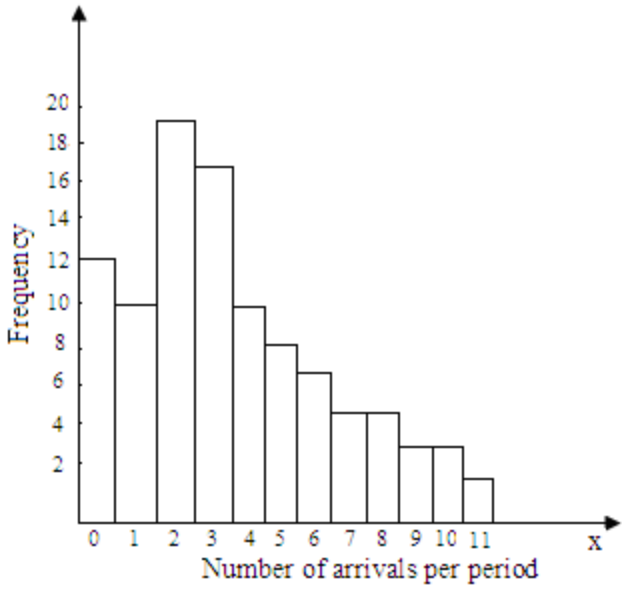


Fig 6.2 Histogram of number of arrivals per period

Example 6.2 (Continuous data)
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, as shown in table 6.2. It is represented in fig 6.3.

Table 6.2 Electronics chip data

Chip life (days)	Frequency
$0 \le x_j < 3$	23
$3 \le x_j < 6$	10
$6 \le x_j < 9$	5
$9 \le x_j < 12$	1
$12 \le x_j < 15$	1
$15 \le x_j < 18$	2
$18 \le x_j < 21$	0
$21 \le x_j < 24$	1
$24 \le x_j < 27$	1
$27 \le x_j < 30$	0
$30 \le x_j < 33$	1
$33 \le x_j < 36$	1
⋮	⋮
$42 \le x_j < 45$	1
⋮	⋮
$57 \le x_j < 60$	1
⋮	⋮
$78 \le x_j < 81$	1
⋮	⋮
$144 \le x_j < 147$	1

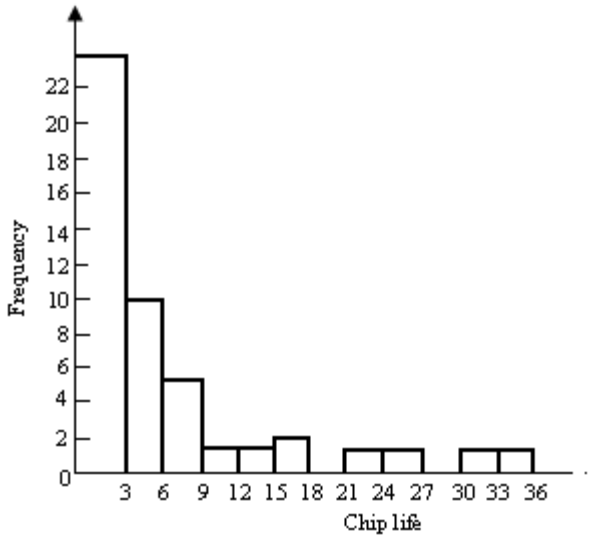


Fig 6.3 **Histogram of chip life**

6.2.2 ***<u>Selecting the Family of Distributions</u>***

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. Where as 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.

6.2.3 Quantile - Quantile plots

A quantile-quantile (q-q) plot is a useful tool for evaluating distribution fit, where as 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 γ such that

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

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

Let
$\{x_i, i = 1, 2, \ldots n\} \rightarrow$ Sample of data from X
$\{y_j \ j = 1, 2, \ldots n\} \rightarrow$ Samples arranged in ascending order, where $y_1 \leq y_2 \leq \ldots \leq y_n$.
$j \rightarrow$ 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.

Example 6.3 (Normal Q – Q plot)
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,…20
Sample mean =99.99 seconds
Sample variance = $(0.2832)^2$ seconds.

The observation are arranged in ascending order in table 6.3

j	Value(y_j)	$F^{-1}((j - 1/2)/20)$
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

Table 6.3 Computation of values

The ordered observations y_j versus $F^{-1}((j - 1/2)/20)$ is plotted, which is shown in fig 6.4

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

1. The observed values will never fall exactly on a straight line
2. The observed values are not independent, since they are ranked. The points will be scattered about the line.
3. 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.

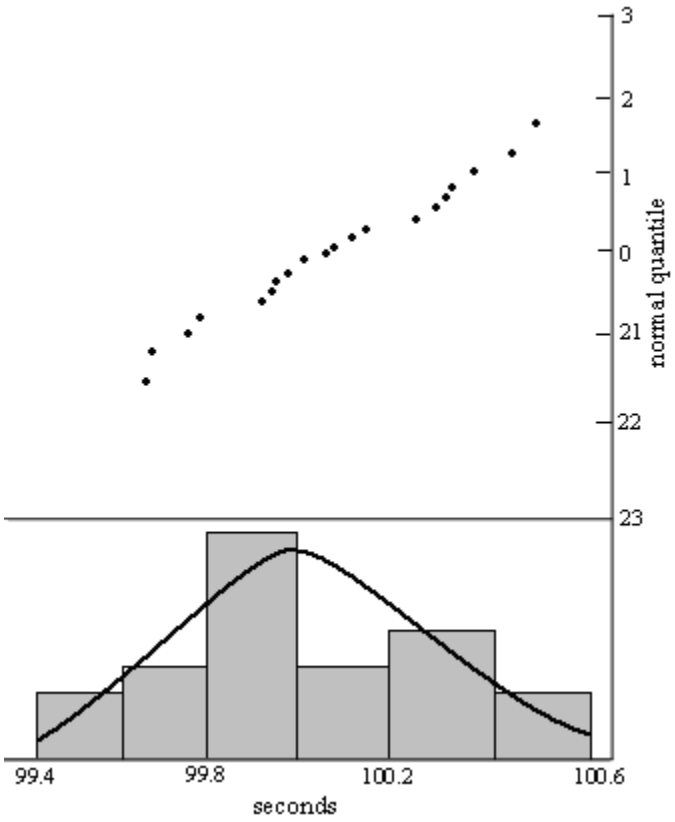


Fig 6.4 Histogram and q-q plot of the installation times

6.3 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.

6.3.1 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 $X_1, X_2, \ldots X_n$.

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

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

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→ Number of distinct values of X

4. 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
$m_j \rightarrow$ Mid point of j^{th} interval
c→ Number of class intervals

6.3.2 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 6.4 contains suggested estimators for distribution often used in simulation.

Distribution	Parameter(s)	Suggested Estimators(s)
Poisson	α	$\hat{\alpha} = \overline{X}$
Exponential	λ	$\hat{\lambda} = \frac{1}{\overline{X}}$
Gamma	β, θ	$\hat{\beta}$ $\hat{\theta} = \frac{1}{\overline{X}}$
Normal	μ,σ^2	$\hat{\mu} = \overline{X}$ $\hat{\sigma}^2 = S^2$ (unbiased)
Lognormal	μ,σ^2	$\hat{\mu} = \overline{X}$ (after taking ln of the data) $\hat{\sigma}^2 = S^2$ (after taking ln of the data)

Table 6.4 Suggested estimators for distributions often used in simulation

Note - The parameter is denoted by α and estimator is denoted by ά

Example 6.4 (lognormal distribution)
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.

Solution
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

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

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

6.4 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. It is especially important to understand the effect of sample size.

- If very little data are available, then goodness-of-fit test is unlikely to reject any candidate distributions.
- If very large data are available, then goodness-of-fit test is likely to reject all candidate distributions.

6.4.1 Chi-square Test

This test is valid for large sample sizes and used for both discrete and continuous distributional assumptions. It formalizes the spontaneous idea of comparing the histogram of data to the shape of candidate density or mass function.

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

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

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^{th} class interval.

χ_0^2 approximately follows chi-square distribution with k-s-1 degrees of freedom.

Where,

s → 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

The critical value $X^2_{\alpha,\ k\text{-}s\text{-}1}$ is found in table A.6. Null hypothesis H_0, is accepted if $X_0^2 < X^2_{\alpha, k-s-1}$

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.

If the distribution being tested is discrete, each value of random variable should be a class interval, unless it is required to combine adjacent class intervals to meet minimum expected cell-frequency requirements.

For discrete, if combining adjacent cells is not required, then

$$p_i = p(x_i) = P(X = x_i)$$

otherwise

p_i is sum of probabilities of appropriate adjacent cells.

If the distribution being tested is continuous, the class intervals are given by $[a_{i-1}, a_i)$, Where a_{i-1} and a_i are endpoints of i^{th} class interval

For continuous with assumed pdf f(x) or assumed cdf F(x), p_i is

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

Note

1. For discrete - Number of class intervals is determined by number of cells resulting after combining adjacent cells as necessary.
2. For continuous – Number of class intervals must be specified.

The table 6.5 helps in determining the number of class intervals for continuous data.

Sample size (n)	Number of Class Interval (k)
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
>100	√n to n/5

Table 6.5 Recommendations for number of class intervals for continuous data

Example 6.5

The vehicle arrival data is tabulated below

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

Since the histogram of the data, shown in fig 6.2 appears to follow Poisson distribution, the parameter $\hat{\alpha} = 3.64$ was determined.

Solution

Hypotheses

H_0 : Random variable is poisson distributed

H_1 : Random variable is not poisson distributed

pmf of poisson distribution

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

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 6.6

X_i	Observed Frequency, 0_i	Expected Frequency, E_i	$\frac{(0_i - E_i)^2}{E_i}$
0	12 } 22	2.6 } 12.2	} 7.87
1	10	9.6	
2	19	17.4	0.15
3	17	21.1	0.80
4	10	19.2	4.41
5	8	14.0	2.57
6	7	8.5	0.26
7	5	4.4	
8	5	2.0	
9	3 } 17	0.8 } 7.6	} 11.62
10	3	0.3	
11	1	0.1	
	100	100.0	27.68

Table 6.6 chi- square goodness-of-fit test

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

$X_0^2 > X^2_{0.05,5}$, Null hypothesis H_0 is not accepted

6.4.2 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.

Example 6.6
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

Solution

Since the histogram appears to follow exponential distribution. The parameter is given a

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

Hypotheses
H_o : Random variable is exponentially distributed
H_1 : 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^{th} interval, i = 1, 2, 3….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

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

Regardless of value of λ, 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 $0 \le x < 1.590$ second interval [1.590, 3.425) and so on. The values are computed and tabulated in table 6.7

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.∞)	6	6.25	0.01
	50	50	$38.72 = X_0^2$

Table 6.7 chi-square goodness-of-fit test

The degrees of freedom is k–s–1 = 8 – 1 – 1 = 6

At $\alpha = 0.05$, $X^2_{0.05,6} = 12.6$, Therefore $X_0^2 > X^2_{0.05,6}$, 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.

6.4.3 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 (α) errors than those specified.

Example 6.7
The interarrival times (minutes) are collected over 100-minute interval and are arranged in order of occurrence.

0.44	0.53	2.04	2.74	2.00	0.30	2.54	0.52	2.02	1.89	1.53	0.21
2.80	0.04	1.35	8.32	2.34	1.95	0.10	1.42	0.46	0.07	1.09	0.76
5.55	3.93	1.07	2.26	2.88	0.67	1.12	0.26	4.57	5.37	0.12	3.19
1.63	1.46	1.08	2.06	0.85	0.83	2.44	2.11	3.15	2.90	6.58	0.64

Solution
Hypotheses

H_0 : The inter arrival times are exponentially distributed .

H_1 : 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 $\{T_1, T_2, T_{3, \ldots}\}$ is exponential, arrival times are uniformly distributed on interval (0, T).

The arrival times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3$…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 $[T_1 / T, (T_1 + T_2) / T, \ldots]$. The resulting 50 points are

0.0044	0.0097	0.0301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3161	0.3356	0.3366	0.3508
0.3553	0.3561	0.3670	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6526	0.6845	0.7008	0.7154	0.7262	0.7468
0.7553	0.7636	0.7880	0.7982	0.8206	0.8417	0.8732	0.9022	0.9680	0.9744

$D^+ = 0.1054$
$D^- = 0.0080$
$D = \max \{D^+, D^-\} = \max \{0.1054, 0.0080\} = 0.1054$
Critical value → $D_{0.05} = 1.36/\sqrt{n} = 1.36/\sqrt{50} = 0.1923$ (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.

6.4.4 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

1. 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.
2. 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.
3. A summary statistic, like p-value is just a summary measure. It just tells where the lack of fit occurs.

6.5 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

1. **Engineering data**
   The values provided by manufacturers provide a starting point for input modeling by fixing a central value.
2. **Expert option**
   Talking to the experts who have experience with the process or similar processes. They can provide optimistic, pessimistic and most likely thoughts.
3. **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.
4. **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.

Example 6.8
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

Solution
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 6.8 summarizes the above information.

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

Table 6.8 Summary of sales information

6.6 Multivariate and Time Series Input Models

The variables may be related and if the variables appear in a simulation models as inputs, the relationship should be determined. When inputs exhibit dependence then *multivariate input models* are used. Example: Two random variables lead time and annual demand in inventory system.

Time series models are useful for representing a sequence of dependent inputs.
Example: Successive time between orders in a system.

The two measures of dependence are covariance and correlation.

6.6.*1 Covariance and correlation*

Let

X_1 X_2 → two random variables
$\mu_i = E(X_i)$ → mean of X_i
$\sigma_i^2 = var(X_i)$→ variance of X_i

Covariance and correlation are the measures of linear dependence between X_1 and X_2. In others words it indicates how well the relationship between X_1 and X_2 is described by the model

$$(X_1 - \mu_1) = \beta(X_2 - \mu_2) + \varepsilon$$

Where ε -> random variable with mean 0, independent of X_2

- If $(X_1 - \mu_1) = \beta(X_2 - \mu_2)$, then model is perfect
- If X_1 and X_2 are statistically independent, then $\beta = 0$, model is of no value
- Positive value of β indicates that X_1 and X_2 tends to be above or below their means
- Negative value of β indicates that X_1 and X_2 tends to be on opposite sides of their means

The covariance between X_1 and X_2 is defined as

$$cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E(X_1X_2) - \mu_1\mu_2$$

A value of cov $(X_1, X_2) = 0$ implies $\beta = 0$
Value of cov $(X_1, X_2) < 0$ implies $\beta < 0$
Value of cov $(X_1, X_2) > 0$ implies $\beta > 0$

The covariance can take any value between $-\infty$ and ∞. The correlation standardizes covariance to be between -1 and 1:

$$\rho = corr(X_1, X_2) = \frac{cov(X_1, X_2)}{\sigma_1 \sigma_2}$$

Value of corr $(X_1, X_2) = 0$ implies $\beta = 0$
Value of corr $(X_1, X_2) < 0 (> 0)$ implies $\beta < 0 (> 0)$

The closer ρ is to -1 or 1; stronger is the linear relationship between X_1 and X_2.

The sequence of random variables $X_1, X_2, \ldots$ that are identically distributed (same mean & variance) and may be dependent, such a sequence is called as ***Time series.***

cov (X_t, X_{t+h}) is called as lag-h autocovariance.
corr (X_t, X_{t+h}) is called as lag-h autocorrelation.

If the value of autocovariance depends only on *h* and not on *t*, then time series is called as *covariance stationary.*

The covariance stationary time series is denoted as

$$\rho_h = \text{corr}\,(X_t, X_{t+h})$$

6.6.2 Multivariate Input Models

If X_1 and X_2 are normally distributed then dependence between them can be modeled by bivariate normal distribution with parameter μ_1, μ_2, σ_1^2, σ_2^2 and $\rho = \text{corr}\,(X_1, X_2)$

To estimate ρ,
Let (X_{11}, X_{21}), (X_{12}, X_{22}),… $(X_{1n}, X_{2n}) \rightarrow n$ independent and identically distributed pairs

The sample covariance is

$$\widehat{\text{cov}}(X_1, X_2) = \frac{1}{n-1} \sum_{j=1}^{n} (X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2)$$

$$= \frac{1}{n-1} \left(\sum_{j=1}^{n} X_{1j}X_{2j} - n\bar{X}_1\bar{X}_2 \right)$$

$\bar{X}_1, \bar{X}_2 \longrightarrow$ sample means

The correlation is estimated by

$$\hat{\rho} = \frac{\widehat{\text{cov}}(X_1, X_2)}{\hat{\sigma}_1 \hat{\sigma}_2}$$

$\hat{\sigma}_1, \hat{\sigma}_2 \longrightarrow$ sample standard deviations

Example 6.9
The following data are available on demand and lead time for last 10 years. Determine the correlation.

Lead time	Demand
6.5	103
4.3	83
6.9	116
6.0	97
6.9	112
6.9	104
5.8	106
7.3	109
4.5	92
6.3	96

Solution

$$\overline{X}_1 = \mu_1 = \frac{\sum_{i=1}^{n} X_i}{n} = 6.14$$

$$\overline{X}_2 = \mu_2 = \frac{\sum_{i=1}^{n} X_i}{n} = 101.80$$

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}_1^2}{n-1} = \frac{386.44 - 10(37.7)}{9} = 1.04$$

$$\hat{\sigma}_1 = 1.02$$

$$\hat{\sigma}_2 = 9.93$$

To estimate the correlation,

$$\sum_{i=1}^{10} X_{1i} X_{2i} = 6328.5$$

$$\hat{cov} = [\, 6328.5 - (10)(6.14)(101.80) \,] / (10 - 1) = 8.66$$

$$\hat{\rho} = \frac{8.66}{(1.02)(9.93)} = 0.86$$

Therefore lead time and demand are strongly dependent

Algorithm to generate bivariate normal random variables

Step 1 – Generate Z_1 and Z_2, independent standard normal random variables

$Z_1 = (-2 \ln R_1)^{1/2} \cos(2\pi R_2)$

$Z_2 = (-2 \ln R_1)^{1/2} \sin(2\pi R_2)$

Step 2 – set $X_1 = \mu_1 + \sigma_1 Z_1$

Step 3 – set $X_2 = \mu_2 + \sigma_2 (\rho Z_1 + \sqrt{1-\rho^2}\, Z_2)$

6.6.3 Time Series Input Models

If $X_1, X_2, X_3, \ldots$ is a sequence of identically distributed but dependent and covariance – stationary random variables then there are number of time series models that can be used to represent the process.

Two models for describing time series are

1. AR(1) – Autoregressive order-1 model
2. EAR(1) – Exponential autoregressive order-1 model

AR(1) model

Consider the time-series model

$$X_t = \mu + \Phi(X_{t-1} - \mu) + \varepsilon_t, \; t = 2, 3, \ldots$$

where

$\varepsilon_2, \varepsilon_3, \varepsilon_4 \ldots \rightarrow$ independent and identically normally distributed with mean zero and variance σ_ε^2 and $-1 < \Phi < 1$

If X_1 is appropriately chosen then X_1, X_2 are normally distributed with mean μ, variance $\sigma^2_\varepsilon / (1 - \Phi^2)$ and $\rho_h = \Phi^h$, $h = 1,2,\ldots$ This time – series model is called Autoregressive order -1 model or AR(1).

Parameter $\Phi = \rho^1 = \text{corr}(X_t, X_{t+1})$ lag-1 autocorrelation

To estimate Φ, we first estimate the

$$\widehat{\text{cov}}(X_t, X_{t+1}) = \frac{1}{n-1} \sum_{t=1}^{n-1} (X_t - \bar{X})(X_{t+1} - \bar{X})$$

$$= \frac{1}{n-1} \left(\sum_{t=1}^{n-1} X_t X_{t+1} - (n-1)\bar{X}^2 \right)$$

Variance $\sigma^2 = \text{var}(X) \longrightarrow \hat{\sigma}^2$

Then

$$\hat{\Phi} = \frac{\widehat{\text{cov}}(X_t, X_{t+1})}{\hat{\sigma}_2} \quad [\because \hat{\Phi} = \hat{\rho}]$$

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2 (1 - \hat{\Phi}^2)$$

Algorithm to generate a stationary AR(1) time series *(*Given values of parameter Φ, μ, σ_ϵ^2)

Step1: Generate X_1 from normal distribution with mean μ and variance $\sigma_\epsilon^2/(1-\Phi^2)$
Set t=2

Step2: Generate ε_t from normal distribution with mean 0 and variance σ_ϵ^2

Step3: Set $X_t = \mu + \Phi (X_{t-1} - \mu) + \varepsilon_t$

Step4: Set t = t +1 and go to step2

EAR (1) Model

Consider the time-series model

$$X_t = \begin{cases} \Phi X_{t-1} & \text{, with probability } \Phi \\ \Phi X_{t-1} + \varepsilon_t & \text{, with probability } 1 - \Phi \end{cases}$$

for t =2, 3, …

Where $\varepsilon_2, \varepsilon_3 \rightarrow$ independent and identically exponentially distributed with mean $1/\lambda$ and $0 \leq \Phi < 1$

If initial values X_1 is chosen appropriately then $X_1, X_2, \ldots$.all are exponentially distributed with mean $1/\lambda$ and $\rho_h = \Phi^h$ for h=1, 2,… This time–series model is called **Exponential Autoregressive order-1model or EAR (1).**Only autocorrelation greater than 0, can be represented by this model. Estimation of parameter proceeds as AR (1) by setting

$\hat{\Phi} = \hat{\rho}$, lag –1 autocorrelation.

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

Algorithm to generate a stationary EAR (1) time series (Given values of parameter Φ and λ)

Step1: Generate X_1 from exponential distribution with mean $1/\lambda$, Set t = 2.

Step2: Generate U from uniform distribution on [0, 1]. If $U \leq \Phi$ then set $X_t = \Phi X_{t-1}$
Otherwise generate ε_t from exponential distribution with mean $1/\lambda$ and set $X_t = \Phi X_{t-1} + \varepsilon_t$

Step3: Set t = t +1 and go to step2.

Example 6.10
The stock brokerage would typically have a large sample of data. The following 20 times gaps between customer by and sell orders had been recorded (in seconds)

1.95 1.75 1.58 1.42 1.28 1.15 1.04 0.93 0.84 0.75
0.68 0.61 11.98 10.79 9.71 14.02 12.62 11.36 10.22 9.20

Estimate lag-1 autocorrelation,

Solution

$$\overline{X} = 5.2$$

$$\hat{\sigma}^2 = 26.7$$

$$\sum_{j=1}^{19} X_t \, X_{t+1} = 924.1$$

$$\hat{cov} = [924.1 - (20\text{-}1)\,(5.2)^2]/(20\text{-}1) = 21.6$$

$$\hat{\rho} = \frac{21.6}{26.7} = 0.8$$

Therefore we could model interarrival times as an EAR (1) process with

$$\hat{\lambda} = \frac{1}{5.2} = 0.192$$

and

$$\hat{\Phi} = 0.8$$

provided exponential distribution is good model for individual gaps.

Verification and Validation of the simulation model is one of the most important and difficult task carried out by the model developer, to work closely with end users throughout the period of development and to increase the model's credibility.

Validation is an integral part of the model development. The goal of validation is a two fold process:

1. To produce a model that represents true system behavior, this can be used as a substitute for the actual system, for the purpose of experimenting.
2. To increase the acceptance, credibility level of model, so that the model will be used by managers and other decision makers.

Conceptually, the verification and validation process consists of the following components.

1. Verification is concerned with **building the model right,** which is used in comparison of conceptual model to the computer representation.
2. Validation is concerned with **building the right model**, which is used to determine that a model is an accurate representation of a real system. It is achieved through the calibration of the model, an iterative process of comparing the model to actual system behavior. This process is repeated until model accuracy is judged to be acceptable.

7.1 Model building, Verification and Validation

The model building involves 3 steps

Step 1 - The first step in model building consists of

- Observing the real system
- Interactions among its various components
- Collecting data on its behavior

This step leads in understanding the system behavior. Persons familiar with the system or sub system should be questioned and gain the advantage of their special knowledge. As the development proceeds, new questions may arise and model developers will return to this step.

Step 2- The second step involves in the construction of a conceptual model. It includes the collection of assumptions of components, structure of the system and hypothesis on the values of model input parameters.

Step 3- The third step is the translation of operational model into a computer recognizable form - computerized model.

Model building is not a linear process instead the model builder goes back to these steps many times while building, verifying and validating the model. The fig 7.1 shows the model building process.

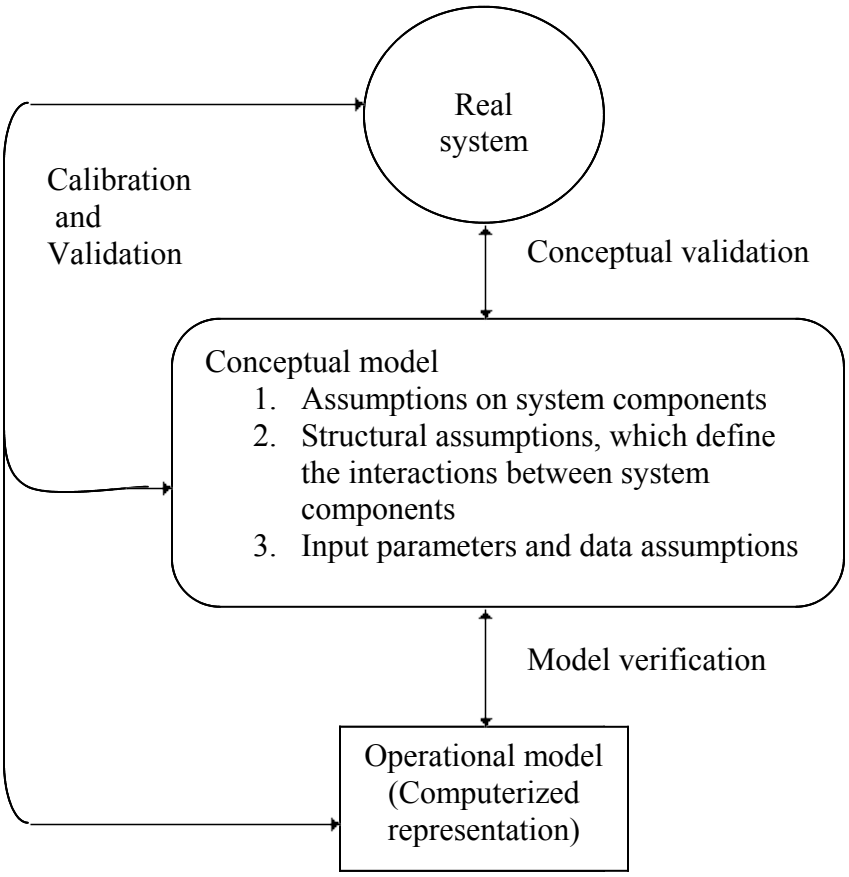


Fig 7.1 Model building, Verification and Validation

7.2 Verification of Simulation Models

The main purpose of verification is to assure that the conceptual model is accurately reflected in the operational model (computerized representation). The conceptual model involves some degree of abstraction about the operations of the system.

Some considerations in verification process are

1. The computerized model has to be checked by others, than its developer.
2. Make a flow diagram (logic flow), which includes operations of system, so that we know what operations takes place when the events occur.
3. Closely examine the output of model for reasonableness, under a variety of settings of input parameters.
4. 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.
5. As far as possible make the self documentation of computerized model.
6. If the computerized representation is animated, then check whether the animation reflects the real system.
7. The interactive run controller (IRC), debugger assists in finding and correcting the errors in the following ways.
    a. As the simulation progress, it can be monitored. This is achieved by advancing the simulation under a desired time and then display model information.
    b. 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.
    c. Selected model components values can be observed.
    d. Simulation can be temporarily suspended or paused, to view information and reassign values or redirect entities.
8. 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

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

2. **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.

1. **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.

2. **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.

3. **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.

Example 7.1
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 $\acute{L}_Q = 0.4375$ 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.

Definition of variables:
CLOCK = Simulation clock
EVTYP = Event type **(** start, arrival, departure, or stop**)**
NCUST = 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	EVTYP = 'Start'	NCUST = 0	STATUS = 0
CLOCK = 3	EVTYP = 'Arrival'	NCUST = 1	STATUS = 0
CLOCK = 5	EVTYP = 'Depart'	NCUST = 0	STATUS = 0
CLOCK = 11	EVTYP = 'Arrival'	NCUST = 1	STATUS = 0
CLOCK = 12	EVTYP = 'Arrival'	NCUST = 2	STATUS = 1
CLOCK = 16	EVTYP = 'Depart'	NCUST = 1	STATUS = 1

Table 7.1 Simulation of trace

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,

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

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

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

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.

7.3 Calibration and Validation of Models

Even though verification and validation are conducted simultaneously by the modeler, validation is the overall process of comparing the model and its behavior to the real system.

Calibration is the iterative process of comparing the model to the real system, making adjustments to the model, comparing the revised model to the reality, making additional adjustments and comparing again. The fig 7.2 shows the relationship of model calibration to the overall validation process.

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

1. **Subjective test** involves people, who are knowledgeable about one or more aspects of system, making judgment about the model and its output.
2. **Objective test** requires data on systems behavior and the corresponding data produced by the model.

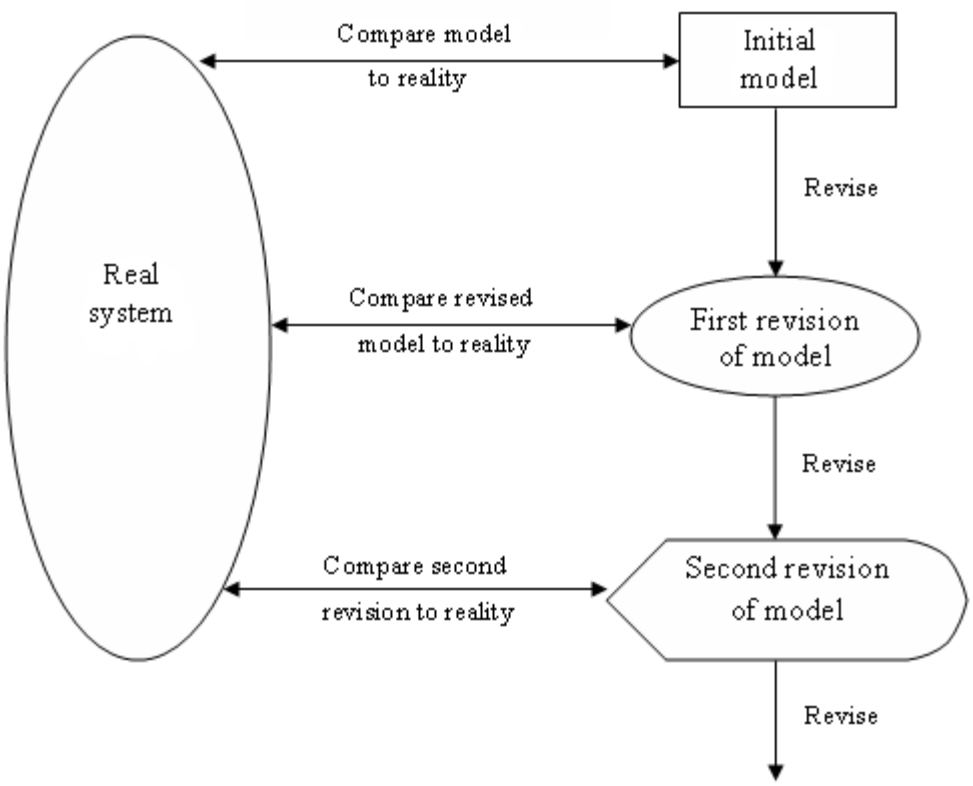


Fig 7.2 Iterative process of calibrating a model

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

1. Build a model that has high face validity.
2. Validate model assumptions.
3. Compare the model input–output transformation to corresponding input output transformation for real system.

7.3.1 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.

1. 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.
2. 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.

7.3.2 Validation of Model Assumptions

Model assumptions have two general classes

1. **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.

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

   Example – For a bank the data that were collected are
   1. Inter arrival times of customers during several 2 hours period of peak loading.
   2. Inter arrival times during a slack period.
   3. Service times for commercial accounts.
   4. 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 is 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
1. Identifying the appropriate probability distribution.
2. Estimating the parameters of hypothesized distribution.
3. 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.

7.3.3 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 non existent 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..

7.3.4 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.

7.3.5 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.

Table A.1 Random digits

39	65	76	45	45	19	90	69	64	61	20	26	36	31	62	58	24	57	14	97	95	06	70	99	00
73	71	23	70	90	65	97	60	12	11	31	56	34	19	19	47	83	75	51	33	30	62	38	20	44
72	20	47	33	84	51	67	47	97	19	98	40	07	17	66	23	05	09	51	80	59	78	11	52	69
75	17	25	69	17	17	95	21	78	58	24	33	45	77	48	69	81	84	09	29	93	22	70	45	80
37	48	79	88	74	63	52	06	34	30	01	31	60	10	27	35	07	79	71	53	28	99	52	01	64
02	89	08	16	94	85	53	83	29	95	56	27	09	24	43	21	78	55	09	82	72	61	88	73	61
87	18	15	70	07	37	79	49	12	38	48	13	93	15	96	41	92	45	71	51	09	18	25	58	94
98	83	71	70	15	89	09	39	59	24	00	06	41	14	20	14	36	59	25	47	54	45	17	24	89
10	08	58	07	04	76	62	16	48	68	58	76	17	14	86	59	53	11	52	21	66	04	18	72	87
47	19	56	37	31	71	82	13	50	14	27	55	10	24	92	28	04	67	53	44	95	23	00	84	47
93	05	31	03	07	34	18	04	52	35	74	13	39	35	22	68	95	23	92	35	36	63	70	35	31
21	89	11	47	99	11	20	99	45	18	76	51	94	84	86	13	79	93	37	55	98	16	04	41	67
95	18	94	36	97	27	37	83	28	71	79	57	95	13	91	09	61	87	25	21	56	11	20	32	44
97	08	31	55	73	10	65	81	92	59	77	31	61	95	46	20	44	90	32	64	26	99	76	75	63
69	26	88	86	13	59	71	74	17	32	48	38	75	93	29	73	37	32	04	05	60	82	29	20	25
41	47	10	25	03	87	63	93	95	17	81	83	83	44	49	77	45	85	50	51	79	88	01	97	30
91	94	15	63	62	08	61	74	51	68	92	79	43	83	79	29	18	94	51	23	14	85	11	47	23
80	06	54	18	47	08	52	85	08	40	48	40	35	94	22	72	65	71	08	86	50	03	42	99	36
76	72	77	63	99	89	85	84	46	06	64	71	06	21	66	89	37	20	70	01	61	65	70	22	12
59	40	24	13	75	42	29	82	23	19	06	94	79	10	08	81	30	15	39	14	81	83	17	16	33
63	62	06	34	41	79	53	36	02	95	94	61	09	43	62	20	21	14	68	86	84	95	48	46	45
78	47	23	53	90	79	93	96	38	63	34	85	52	05	09	84	43	01	72	73	14	93	87	81	40
87	68	62	15	43	97	48	72	66	48	53	16	71	13	81	59	97	50	99	52	24	62	20	42	31
47	60	92	10	77	26	97	05	73	51	88	46	38	00	58	72	68	49	29	31	75	70	16	08	24
56	88	87	59	41	06	87	37	78	48	65	88	69	58	39	88	02	84	27	83	85	81	56	59	38
22	17	68	65	84	86	02	22	57	51	68	69	80	95	44	11	29	01	95	80	49	34	35	96	47
19	36	27	59	46	39	77	32	77	09	79	57	92	36	39	89	74	39	82	15	08	50	94	34	74
16	77	23	02	77	28	06	24	25	93	22	45	44	84	11	87	80	61	65	31	09	71	91	74	25
78	43	66	71	61	97	66	63	99	61	80	45	67	93	82	59	73	19	85	23	53	33	65	97	21
03	28	28	26	08	69	30	16	09	05	53	58	47	70	93	66	56	45	63	79	45	56	20	19	47
04	31	17	21	56	33	63	99	19	87	26	72	39	27	67	53	77	57	69	93	60	61	97	22	61

Random digits

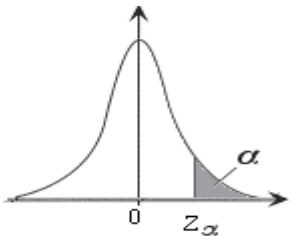
5965 2913 5612 6361 7075 5490 9626 4307 0840 7945
5801 9383 6173 8358 9236 5543 5811 5520 5814 7864
1223 5344 3649 6397 1678 4400 7715 7614 1209 7729
0220 2108 0784 8837 3916 0282 4490 3442 6471 6593
4131 9772 7594 8863 0874 1864 8117 6411 7012 2682
3074 5746 2723 5681 0989 8015 0818 5380 9981 3758
2939 6585 6658 7756 7916 9770 2868 2128 2665 2386
6003 5982 8829 2833 8160 2101 3365 4121 4522 8216
2039 2993 4362 6363 2914 4955 6364 5237 6456 5561
0176 2425 2968 3834 6077 4302 3499 9938 7231 2136
2161 1365 2764 7836 1584 2421 4247 2930 0783 9989
0407 1760 7048 1929 9034 0242 0753 4851 9465 0791
0055 7981 7760 2215 3323 4727 8884 8066 7965 3939
0726 2104 9164 6275 5464 4073 1715 3215 7883 8087
2475 9583 8713 1445 2702 4952 4307 5796 2913 0589
0686 1266 4341 9760 9608 5773 7394 9333 4752 8395
4223 4033 3734 8221 2055 5131 0065 1626 7742 5806
9596 5241 3230 3269 4836 9776 2894 5740 1557 2515
1581 5007 6906 8933 9981 3175 4979 4525 5334 6038
6558 6350 1273 6164 7125 1481 3084 1517 4748 0956
1974 7635 1129 0593 7963 3817 0148 1377 5165 6568
8671 4147 7231 3509 9032 4233 9087 3328 9044 3152
0979 6984 8428 7697 8859 5363 2984 2649 9244 7035
0635 0334 7219 7422 9571 1053 5954 4040 5777 2440
6686 8703 3451 1548 9797 0816 9342 0240 5814 9593
3878 6600 8703 9512 5588 2446 1842 0882 2024 7736
9869 8361 8090 8666 7540 6516 3343 7379 1140 5565
8969 4225 6202 8102 5691 8499 6466 7775 0721 9345
6339 8671 8023 3701 8250 0274 9339 5135 4475 7960
3187 5353 9213 1705 5580 1432 5962 8191 1676 5861
6142 5175 6497 9478 6278 8939 3902 0076 2004 9201
8286 5570 4400 3640 9650 5709 6855 3454 5397 9991
5531 0150 6376 0494 8239 1639 5611 5803 5645 0851
6357 6828 4497 2508 9084 7544 5964 3718 1007 9333
7376 2940 3503 3317 0465 2912 6500 3883 2539 6516
3060 1836 3740 7183 2965 3246 4028 5528 8607 5611
4767 1322 7035 6171 9065 2024 2318 5460 5571 2092
1550 2362 4356 9447 4196 1101 6479 3928 3321 3684
4956 5537 9056 3006 2066 7296 3018 3878 2927 9268
2504 8074 7591 9689 2755 3226 1726 9222 3633 9816
8328 3942 7243 1717 3592 9307 2738 3856 0684 9873
6227 3172 3764 9551 0426 6061 8384 5473 7418 8053
2946 2893 4927 2197 3452 6104 2255 2268 7063 1443
7574 3933 8021 2711 6276 7146 2391 1984 2962 3634
9042 6919 4140 4545 6873 3748 5053 8284 4120 1819
1839 7794 6640 0492 6833 0485 6422 5213 0394 2643
4861 2514 5827 7994 4041 9929 8055 3514 7126 4064
6051 9425 6381 7204 3938 3430 5952 2753 3471 5992
5306 1578 1198 6256 1865 5631 2852 1416 6313 4460
2521 8837 4158 5485 7726 4380 7901 6142 6385 6755

Table A.2 Random digits

94737	08225	35614	24826	88319	05595	58701	57365	74759
87259	85982	13296	89326	74863	99986	68558	06391	50248
63856	14016	18527	11634	96903	52146	53496	51730	03500
66612	54714	46783	61934	30258	61674	07471	67566	31635
30712	58582	05704	23172	86689	94834	99057	55832	21012
69607	24145	43886	86477	05317	30445	33456	34029	09603
37792	27282	94107	41967	21425	04743	42822	28111	09757
01488	56680	73847	64930	11108	44834	45390	86043	23973
66248	97697	38244	50918	55441	51217	54786	04940	50807
51453	03462	61157	65366	61130	26204	15016	85665	97714
92168	82530	19271	86999	96499	12765	20926	25282	39119
36463	07331	54590	00546	03337	41583	46439	40173	46455
47097	78780	04210	87084	44484	75377	57753	41415	09890
80400	45972	44111	99708	45935	03694	81421	60170	58457
94554	13863	88239	91624	00022	40471	78462	96265	55360
31567	53597	08490	73544	72573	30961	12282	97033	13676
07821	24759	47266	21747	72496	77755	50391	59554	31177
09056	10709	69314	11449	40531	02917	95878	71587	60906
19922	37025	80731	26179	16039	01518	82697	73227	13160
29923	02570	80164	36108	73689	26342	35712	49137	13482
29602	29464	99219	20308	82109	03898	82072	85199	13103
94135	94661	87724	88187	62191	70607	63099	40494	49069
87926	34092	34334	55064	43152	01610	03126	47312	59578
85039	19212	59160	83537	54414	19856	90527	21756	64783
66070	38480	74636	45095	86576	79337	39578	40851	53503
78166	82521	79261	12570	10930	47564	77869	16480	43972
94672	07912	26153	10531	12715	63142	88937	94466	31388
56406	70023	27734	22254	27685	67518	63966	33203	70803
67726	57805	94264	77009	08682	18784	47554	59869	66320
07516	45979	76735	46509	17696	67177	92600	55572	17245
43070	22671	00152	81326	89428	16368	57659	79424	57604
36917	60370	80812	87225	02850	47118	23790	55043	75117
03919	82922	02312	31106	44335	05573	17470	25900	91080
46724	22558	64303	78804	05762	70650	56117	06707	90035
16108	61281	86823	20286	14025	24909	38391	12183	89393
74541	75808	89669	87680	72758	60851	55292	95663	88326
82919	31285	01850	72550	42980	57518	01169	01786	98145
31388	26809	77258	99360	92362	21979	41319	75739	98082
17190	75522	15687	07161	99745	48767	03121	20046	28013
00466	88068	68631	98745	97810	35886	14497	90230	69264

Table A.3 Cumulative Normal Distribution

$$\Phi(Z_\alpha) = \int_{-\infty}^{Z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = 1 - \alpha$$



Z_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A.4 Cumulative Poisson distribution

	α = mean									
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.368
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4				1.000	1.000	1.000	0.999	0.999	0.998	0.996
5										0.999
6										1.000

	α = mean									
k	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7				1.000	1.000	1.000	1.000	0.999	0.999	0.999
8								1.000	1.000	1.000

	α = mean									
k	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0.122	0.111	0.100	0.091	0.082	0.074	0.067	0.061	0.055	0.050
1	0.380	0.355	0.331	0.308	0.287	0.267	0.249	0.231	0.215	0.199
2	0.650	0.623	0.596	0.570	0.544	0.518	0.494	0.469	0.446	0.423
3	0.839	0.819	0.799	0.779	0.758	0.736	0.714	0.692	0.670	0.647
4	0.938	0.928	0.916	0.904	0.891	0.877	0.863	0.848	0.832	0.815
5	0.980	0.975	0.970	0.964	0.958	0.951	0.943	0.935	0.926	0.916
6	0.994	0.993	0.991	0.988	0.986	0.983	0.979	0.976	0.971	0.966
7	0.999	0.998	0.997	0.997	0.996	0.995	0.993	0.992	0.990	0.988
8	1.000	1.000	0.999	0.999	0.999	0.999	0.998	0.998	0.997	0.996
9			1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
10							1.000	1.000	1.000	1.000

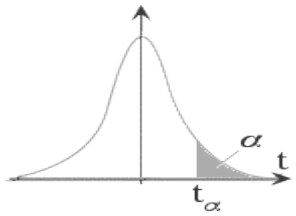
	α = mean									
k	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.045	0.041	0.037	0.033	0.030	0.027	0.025	0.022	0.020	0.018
1	0.185	0.171	0.159	0.147	0.136	0.126	0.116	0.107	0.099	0.092
2	0.401	0.380	0.359	0.340	0.321	0.303	0.285	0.269	0.253	0.238
3	0.625	0.603	0.580	0.558	0.537	0.515	0.494	0.473	0.453	0.433
4	0.798	0.781	0.763	0.744	0.725	0.706	0.687	0.668	0.648	0.629
5	0.906	0.895	0.883	0.871	0.858	0.844	0.830	0.816	0.801	0.785
6	0.961	0.955	0.949	0.942	0.935	0.927	0.918	0.909	0.899	0.889
7	0.986	0.983	0.980	0.977	0.973	0.969	0.965	0.960	0.955	0.949
8	0.995	0.994	0.993	0.992	0.990	0.988	0.986	0.984	0.981	0.979
9	0.999	0.998	0.998	0.997	0.997	0.996	0.995	0.994	0.993	0.992
10	1.000	1.000	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.997
11			1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
12								1.000	1.000	1.000

Table A.4 (continued)

	α= mean									
k	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.017	0.015	0.014	0.012	0.011	0.010	0.009	0.008	0.007	0.007
1	0.085	0.078	0.072	0.066	0.061	0.056	0.052	0.048	0.044	0.040
2	0.224	0.210	0.197	0.185	0.174	0.163	0.152	0.143	0.133	0.125
3	0.414	0.395	0.377	0.359	0.342	0.326	0.310	0.294	0.279	0.265
4	0.609	0.590	0.570	0.551	0.532	0.513	0.495	0.476	0.458	0.440
5	0.769	0.753	0.737	0.720	0.703	0.686	0.668	0.651	0.634	0.616
6	0.879	0.867	0.856	0.844	0.831	0.818	0.805	0.791	0.777	0.762
7	0.943	0.936	0.929	0.921	0.913	0.905	0.896	0.887	0.877	0.867
8	0.976	0.972	0.968	0.964	0.960	0.955	0.950	0.944	0.938	0.932
9	0.990	0.989	0.987	0.985	0.983	0.980	0.978	0.975	0.972	0.968
10	0.997	0.996	0.995	0.994	0.993	0.992	0.991	0.990	0.988	0.986
11	0.999	0.999	0.998	0.998	0.998	0.997	0.997	0.996	0.995	0.995
12	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.998
13			1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
14									1.000	1.000

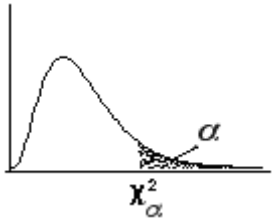
	α= mean									
k	5.2	5.4	5.6	5.8	6.0	6.2	6.4	6.6	6.8	7.0
0	0.006	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.001	0.001
1	0.034	0.029	0.024	0.021	0.017	0.015	0.012	0.010	0.009	0.007
2	0.109	0.095	0.082	0.072	0.062	0.054	0.046	0.040	0.034	0.030
3	0.238	0.213	0.191	0.170	0.151	0.134	0.119	0.105	0.093	0.082
4	0.406	0.373	0.342	0.313	0.285	0.259	0.235	0.213	0.192	0.173
5	0.581	0.546	0.512	0.478	0.446	0.414	0.384	0.355	0.327	0.301
6	0.732	0.702	0.670	0.638	0.606	0.574	0.542	0.511	0.480	0.450
7	0.845	0.822	0.797	0.771	0.744	0.716	0.687	0.658	0.628	0.599
8	0.918	0.903	0.886	0.867	0.847	0.826	0.803	0.780	0.755	0.729
9	0.960	0.951	0.941	0.929	0.916	0.902	0.886	0.869	0.850	0.830
10	0.982	0.977	0.972	0.965	0.957	0.949	0.939	0.927	0.915	0.901
11	0.993	0.990	0.988	0.984	0.980	0.975	0.969	0.963	0.955	0.947
12	0.997	0.996	0.995	0.993	0.991	0.989	0.986	0.982	0.978	0.973
13	0.999	0.999	0.998	0.997	0.996	0.995	0.994	0.992	0.990	0.987
14	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.996	0.994
15			1.000	1.000	0.999	0.999	0.999	0.999	0.998	0.998
16					1.000	1.000	1.000	0.999	0.999	0.999
17								1.000	1.000	1.000

Table A.5 Percentage of points of the students 't' Distribution with υ degrees of Freedom



υ	$t_{0.005}$	$t_{0.01}$	$t_{0.025}$	$t_{0.05}$	$t_{0.1}$
1	63.656	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
60	2.660	2.390	2.000	1.671	1.296
120	2.617	2.358	1.980	1.658	1.289
Infinity	2.576	2.326	1.960	1.645	1.282

Table A.6 Percentage points of the Chi-Square Distribution with υ degrees of Freedom



υ	$X^2_{.005}$	$X^2_{.010}$	$X^2_{.025}$	$X^2_{.050}$	$X^2_{.100}$
1	7.87944	6.63490	5.02389	3.84146	2.70554
2	10.59663	9.21034	7.37776	5.99146	4.60517
3	12.83816	11.34487	9.34840	7.81473	6.25139
4	14.86026	13.27670	11.14329	9.48773	7.77944
5	16.74960	15.08627	12.83250	11.07050	9.23636
6	18.54758	16.81189	14.44938	12.59159	10.64464
7	20.27774	18.47531	16.01276	14.06714	12.01704
8	21.95495	20.09024	17.53455	15.50731	13.36157
9	23.58935	21.66599	19.02277	16.91898	14.68366
10	25.18818	23.20925	20.48318	18.30704	15.98718
11	26.75685	24.72497	21.92005	19.67514	17.27501
12	28.29952	26.21697	23.33666	21.02607	18.54935
13	29.81947	27.68825	24.73560	22.36203	19.81193
14	31.31935	29.14124	26.11895	23.68479	21.06414
15	32.80132	30.57791	27.48839	24.99579	22.30713
16	34.26719	31.99993	28.84535	26.29623	23.54183
17	35.71847	33.40866	30.19101	27.58711	24.76904
18	37.15645	34.80531	31.52638	28.86930	25.98942
19	38.58226	36.19087	32.85233	30.14353	27.20357
20	39.99685	37.56623	34.16961	31.41043	28.41198
21	41.40106	38.93217	35.47888	32.67057	29.61509
22	42.79565	40.28936	36.78071	33.92444	30.81328
23	44.18128	41.63840	38.07563	35.17246	32.00690
24	45.55851	42.97982	39.36408	36.41503	33.19624
25	46.92789	44.31410	40.64647	37.65248	34.38159
26	48.28988	45.64168	41.92317	38.88514	35.56317
27	49.64492	46.96294	43.19451	40.11327	36.74122
28	50.99338	48.27824	44.46079	41.33714	37.91592
29	52.33562	49.58788	45.72229	42.55697	39.08747
30	53.67196	50.89218	46.97924	43.77297	40.25602

Table A.7 Percentage points of F distribution with α = 0.05

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105
inf	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307

df2/df1	12	15	20	24	30	40	60	120	INF
1	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	252.1957	253.2529	254.3
2	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19.4791	19.4874	19.4
3	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.5494	8.5
4	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.6877	5.6581	5.6
5	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.3985	4.3
6	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047	3.6
7	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674	3.2
8	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9
9	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475	2.7
10	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.5801	2.5
11	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.4480	2.4
12	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410	2.2
13	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524	2.2
14	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2229	2.1778	2.1
15	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141	2.0
16	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.0589	2.0
17	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.0584	2.0107	1.9
18	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.9681	1.9
19	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.9795	1.9302	1.8
20	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963	1.8
21	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.9165	1.8657	1.8
22	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.8894	1.8380	1.7
23	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.8648	1.8128	1.7
24	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.8424	1.7896	1.7
25	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.7684	1.7
26	2.1479	2.0716	1.9898	1.9464	1.9010	1.8533	1.8027	1.7488	1.6
27	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.7851	1.7306	1.6
28	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.7689	1.7138	1.6
29	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.7537	1.6981	1.6
30	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.7396	1.6835	1.6
40	2.0035	1.9245	1.8389	1.7929	1.7444	1.6928	1.6373	1.5766	1.5
60	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.5343	1.4673	1.3
120	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.4290	1.3519	1.2
inf	1.7522	1.6664	1.5705	1.5173	1.4591	1.3940	1.3180	1.2214	1.0

Table A.7 (continued)

Table A.8 Kolmogorov-Smirnov critical values

Degrees of Freedom (N)	$D_{0.10}$	$D_{0.05}$	$D_{0.01}$
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.642	0.708	0.828
4	0.564	0.624	0.733
5	0.510	0.565	0.669
6	0.470	0.521	0.618
7	0.438	0.486	0.577
8	0.411	0.457	0.543
9	0.388	0.432	0.514
10	0.368	0.410	0.490
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.433
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.328	0.392
17	0.286	0.318	0.381
18	0.278	0.309	0.371
19	0.272	0.301	0.363
20	0.264	0.294	0.356
25	0.24	0.27	0.32
30	0.22	0.24	0.29
35	0.21	0.23	0.27
Over 35	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$