COMPUTER SCIENCE

UNIT – I

APPLIED PROBABILITY AND OPERATIONS RESEARCH

RANDOM PROCESSES, PROBABILITY DISTRIBUTIONS, QUEUING MODELS AND SIMULATION, TESTING OF HYPOTHESIS, DESIGN OF EXPERIMENTS.

RANDOM PROCESSES

In probability theory, random process is a collection of random variables, representing the evolution of some system of random values over time. This is the probabilistic counterpart to a deterministic process (or deterministic system). Instead of describing a process which can only evolve in one way (as in the case, for example, of solutions of an ordinary differential equation), in a stochastic or random process there is some indeterminacy: even if the initial condition (or starting point) is known, there are several (often infinitely many) directions in which the process may evolve.

A random variable (RV) is a rule (or function) that assigns a real number to every outcome of a random experiment, while a random process is a rule (or function) that assigns a time function to every outcome of a random experiment.

CLASSIFICATION OF RANDOM PROCESSES

Depending on the continuous or discrete nature of the state space S and parameter set T, a random process can be classified into four types:

(i) If both T and S are discrete, the random process is called a discrete random sequence. For example, if X_n represents the outcome of the n^{th} toss of a fair dice, then $\{X_n, n \geq 1\}$ is a discrete random sequence, since T = {1, 2, 3,........} and S ={1, 2, 3, 4, 5, 6}.

(ii) If T is discrete and 5 is continuous, the random process is called a continuous random sequence. For example, if X_n represents the temperature at the end of the nth hour of a day, then $\{Xn, 1 \leq n \leq 24\}$ is a continuous random sequence, since temperature can take any value in an interval and hence continuous.

(iii) If T is continuous and S is discrete, the random process is called a discrete random process. For example, if X(t) represents the number of telephone calls received in the interval (0, t) then {X(t)} random process, since S = {0, 1, 2, 3, …}.

(iv) If both T and S are continuous, the random process is called acontinuous Random process. For example, if X(r) represents the maximum temperature at a place in the interval (0, t), {X(t)} is a continuous random process. In the names given above, the word 'discrete' or 'continuous' is used to refer to the nature of S and the word 'sequence' or 'process' is used to refer to the nature of T.

(I) MARKOV CHAINS AND MARKOV PROCESSES

Important classes of stochastic processes are Markov chains and Markov processes. A Markov chain is a discrete-time process for which the future behaviour, given the past and the present, only depends on the present and not on the past. A Markov process is the continuous-time version of a Markov chain. Many queueing models are in fact Markov processes. This chapter gives a short introduction to Markov chains and Markov processes focussing on those characteristics that are needed for the modelling and analysis of queueing problems.

(II) PROCESS WITH INDEPENDENT INCREMENTS

If, for all choices of t2, •••, t„ such that t, < t2 < t3 < tn the random variables X(t2) –X(t_1), X(t3) – X(t2), •••, X(tn)-X(tn–1,) are independent, then the process (X(t)} is said to be a radom process with independent increments.

If T= {0, 1, 2, •••} is the parameter set for {Xn}, then {Zn}, where Z0 =Xn and Zn = Xn - Xn–1 |, is a random sequence with independent increments if the RVs Z0, Z_1, Z_2, •••, are independent.

Two processes with independent increments play an important role in the theory of random processes. One is the Poisson process that has a Poisson distribution for the increments and the other is the Wiener process with a Gaussian distribution for the increments.

(III) STATIONARY PROCESSES

If certain probability distribution or averages do not depend on t, then the random process {X(t)} is called stationary. A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

PROBABILITY DISTRIBUTIONS

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level. A probability distribution is a table or an equation that links each possible value that a random variable can assume with its probability of occurrence.

Some practical uses of probability distributions are:

To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests. For univariate data, it is often useful to determine a reasonable distributional model for the data.

DISCRETE PROBABILITY DISTRIBUTIONS

The probability distribution of a discrete random variable can always be represented by a table. For example, suppose you flip a coin two times. This simple exercise can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of heads that result from the coin flips. The variable X can take on the values 0, 1, or 2 and X is a discrete random variable.

The table below shows the probabilities associated with each possible value of X. The probability of getting 0 heads is 0.25; 1 head is 0.50; and 2 heads is 0.25. Thus, the table is an example of a probability distribution for a discrete random variable.

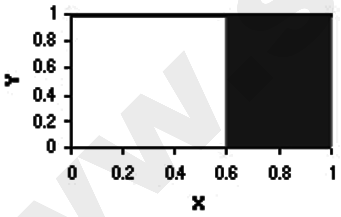
Number of heads, x	Probability, P(x)
0	0.25
1	0.50
2	0.25

CONTINUOUS PROBABILITY DISTRIBUTIONS

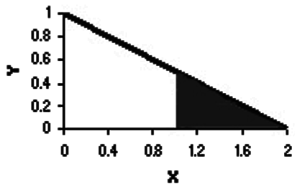
The probability distribution of a continuous random variable is represented by an equation, called the **probability density function** (pdf). All probability density functions satisfy the following conditions:

- The random variable Y is a function of X; that is, y = f(x).
- The value of y is greater than or equal to zero for all values of x.
- The total area under the curve of the function is equal to one.

The charts below show two continuous probability distributions. The chart on the left shows a probability density function described by the equation y = 1 over the range of 0 to 1 and y = 0 elsewhere. The chart on the right shows a probability density function described by the equation y = 1 - 0.5x over the range of 0 to 2 and y = 0 elsewhere. The area under the curve is equal to 1 for both charts.



y = 1



y = 1 – 05x

The probability that a continuous random variable falls in the interval between *a* and *b* is equal to the area under the pdf curve between *a* and *b*. For example, in the first chart above, the shaded area shows the probability that the random variable X will fall between 0.6 and 1.0. That probability is 0.40. And in the second chart, the shaded area shows the probability of falling between 1.0 and 2.0. That probability is 0.25.

With a continuous distribution, there are an infinite number of values between any two data points. As a result, the probability that a continuous random variable will assume a particular value is always zero. For example, in both of the above charts, the probability that variable X will equal exactly 0.4 is zero.

Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.

Simulation studies with random numbers generated from using a specific probability distribution are often needed.

QUEUING, MODELS AND SIMULATION

Simulation is often used in the analysis of queuing models.A simple but typical model is the single-server queue system. In this model, the term "customer" refers to any type of entity that can be viewed as requesting "service" from a system. Some examples are production systems, repair and maintenance facilities, communications and computer systems, transport and material-handling systems etc. Queuing models, whether solved analytically or through simulation, provide the analyst with a powerful tool for designing and evaluating the performance of queuing systems. Typical measures of system performance are server utilization, length of waiting lines and delays of customers. Quite often, the analyst or the decision maker is involved in trade offs between server utilization and customer satisfaction in terms of line lengths and delays For relatively simple systems, these performance measures can be computed mathematically - at great savings in time and expense as compared with the use of a simulation model - but, for for realistic models of complex systems, simulation is usually required. Nevertheless, analytically tractable models, although usually requiring many simplifying assumptions, are valuable for rough-cut estimates of system performance.

CHARACTERISTICS OF QUEUING SYSTEMS

The key elements of queuing systems are the customers and servers.The term "customer" can refer to people, parts, trucks, e-mails etc. and the term "server" clerks, mechanics, repairmen, CPUs etc. Although the terminology employed will be that of a customer arriving to a server, sometimes the server moves to the customer; for example a repairman moving to a broken machine.

HYPOTHESIS TESTING

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

The method of hypothesis testing can be summarized in four steps.

1. To begin, we identify a hypothesis or claim that we feel should be tested. For example, we might want to test the claim that the mean number of hours that children in the United States watch TV is 3 hours.
2. We select a criterion upon which we decide that the claim being tested is true or not. For example, the claim is that children watch 3 hours of TV per week. Most samples we select should have a mean close to or equal to 3 hours if the claim we are testing is true. So at what point do we decide that the discrepancy between the sample mean and 3 is so big that the claim we are testing is likely not true? We answer this question in this step of hypothesis testing.
3. Select a random sample from the population and measure the sample mean. For example, we could select 20 children and measure the mean time (in hours) that they watch TV per week.
4. Compare what we observe in the sample to what we expect to observe if the claim we are testing is true. We expect the sample mean to be around 3 hours. If the discrepancy between the sample mean and population mean is small, then we will likely decide that the claim we are testing is indeed true. If the discrepancy is too large, then we will likely decide to reject the claim as being not true.

DESIGN OF EXPERIMENTS

The term experimental design refers to a plan for assigning experimental units to treatment conditions. A good experimental design serves three purposes.

CAUSATION: It allows the experimenter to make causal inferences about the relationship between independent variables and a dependent variable.

CONTROL: It allows the experimenter to rule out alternative explanations due to the confounding effects of extraneous variables (i.e., variables other than the independent variables).

VARIABILITY: It reduces variability within treatment conditions, which makes it easier to detect differences in treatment outcomes.

COMPLETELY RANDOMIZED DESIGN

The **completely randomized design** is probably the simplest experimental design, in terms of data analysis and convenience. With this design, participants are randomly assigned to treatments.

TREATMENT	
Placebo	**Vaccine**
500	500

A completely randomized design layout for the Acme Experiment is shown in the table to the right. In this design, the experimenter randomly assigned participants to one of two treatment conditions. They received a placebo or they received the vaccine. The same number of participants (500) were assigned to each treatment condition (although this is not required).

The dependent variable is the number of colds reported in each treatment condition. If the vaccine is effective, participants in the "vaccine" condition should report significantly fewer colds than participants in the "placebo" condition.

A completely randomized design relies on randomization to control for the effects of extraneous variables. The experimenter assumes that, on averge, extraneous factors will affect treatment conditions equally; so any significant differences between conditions can fairly be attributed to the independent variable.

RANDOMIZED BLOCK DESIGN

With a **randomized block design**, the experimenter divides participants into subgroups called **blocks**, such that the variability within blocks is less than the variability between blocks. Then, participants within each block are randomly assigned to treatment conditions. Because this design reduces variability and potential confounding, it produces a better estimate of treatment effects.

TREATMENT

Gender	Placebo	Vaccine
Male	250	250
Female	250	250

The table to the right shows a randomized block design for the Acme experiment. Participants are assigned to blocks, based on gender. Then, within each block, participants are randomly assigned to treatments. For this design, 250 men get the placebo, 250 men get the vaccine, 250 women get the placebo, and 250 women get the vaccine.

It is known that men and women are physiologically different and react differently to medication. This design ensures that each treatment condition has an equal proportion of men and women. As a result, differences between treatment conditions cannot be attributed to gender. This randomized block design removes gender as a potential source of variability and as a potential confounding variable.

MATCHED PAIRS DESIGN

Pair Treatment	Placebo	Vaccine
1	1	1
2	1	1
...	...	...
499	1	1
500	1	1

A matched pairs design is a special case of the randomized block design. It is used when the experiment has only two treatment conditions; and participants can be grouped into pairs, based on some blocking variable. Then, within each pair, participants are randomly assigned to different treatments.