Chapter 2

BASIC MATHEMATICS FOR QOS

To understand QoS in packet networks, it is important to understand not only the mechanism of providing QoS but also the performance behavior that is produced by the QoS mechanism. This chapter reviews some of the basic mathematics that is needed in the analysis of QoS performance in packet networks. The following topics are reviewed in this chapter:

- probability
- random variables
- stochastic processes
- queuing theory

From the author's experience of teaching, students generally considered the mathematical concepts and disciplines such as probability theory, random variables and stochastic processes to be too abstract and hard to apply to real problems.[1-3] One of the purposes of this chapter is to explain the abstract concepts in layman's terms as much as possible so that they can be applied to real problems such as QoS.

1. PROBABILITY THEORY

1.1 Random experiments, outcomes and events

A random experiment is an experiment that produces random outcomes. For example, throwing a die is a random experiment in which each trial produces a random outcome from six possible outcomes, i.e., faces with one through six spots. The word "experiment" implies that the random situation

under consideration is controlled. However, the word may also be used in a broad sense to mean any random situation that produces random outcomes, let us say, a nature's experiment.

A trial is a single instantiation of a random experiment. If a die is thrown ten times, there would be ten trials. The key concept to note here is that each trial produces exactly one outcome.

Another term frequently used in probability is a random "event." A random event is a higher level outcome that may depend on multiple experiments and multiple outcomes of the experiments. For example, consider a game consisting of two random experiments, "throwing a die" and "throwing a coin." A player is to throw the die twice and the coin once. A player who gets the face with one spot in both die-throwings and a "head" in the coin-throwing wins the grand prize. In this game, the random "event" of interest is "winning the grand prize." This event would "occur," if the trials produce the following outcomes: one spot in both of the die-throwings and a "head" in the coin-throwing. In this example, the event depends on multiple experiments and multiple outcomes.

In set theory, a set is defined by the elements contained in the set, e.g., a set of all integers, a set of all even integers, and a set of positive numbers. Using set theory, an event is defined as a set containing the outcomes that make the event happen. For example, in the die-throwing experiment, an event called "face with an even number of spots" may be defined by a set denoted by say E as follows: E= *{"two," "four," "six"},* where *"two" "four"* and *"six"* denote the number of spots on the face of the die.

A random event defined by a set containing a single outcome is referred to as an "elementary event." For example, in the die throwing example, there are six possible random outcomes: *"one," "two," "three," "four," "five,"* and *"six"*. If each of these possible outcomes is defined to be an event, the six possible outcomes produce six elementary events: *{"one"}, {"two"}, {"three"}, {"four"}, {"five"},* and *{"six"}.*

The distinction between the outcome, e.g., *"one,"* and the event, e.g., *{"one"},* is significant and fundamental in the construct of probability theory because, as we shall see in Section 1.3, probability is defined for an event given the probabilities of the underlying random outcomes. *"One"* is an element of a set, whereas *{"one"}* is a set containing one element, *"one."* The probabilities of elementary events would then be equal to the probabilities of the random outcomes.

1.2 Definition of probability

What is probability? Mathematicians attempted to define this seemingly simple term without much success in reaching a consensus for a long time

until Kolmogorov presented his celebrated theory referred to as the "axiomatic approach." The power of the axiomatic approach is in its simplicity.

First, consider the debate that went on before Kolmogorov. A probability was defined as a frequency of occurrence. Consider 1,000 trials in the coin throwing experiment. If the head shows up 400 times, it is concluded that the "probability" of a head is 0.4. The dilemma of this definition of probability is that unless the coin is thrown many times and the outcomes are observed, there is no way of telling the probability.

Some would say that the probability of head should be 0.5 but then others would argue that, unless the coin is minted "perfectly" with identical sides, no one can say that its probability is 0.5 even though it may be "close," etc., etc. Mathematicians had difficulty overcoming the arguments such as this and, as a result, probability theory could not be developed into a useful discipline that could be applied to practical problems.

Most reasonable persons could agree, deep in their hearts, that it should be good enough to take the probability of, for example, a particular face in die throwing is 1/6 and move on to solve other probability problems associated with die throwing. If the 1/6 probability for a face is accepted, then one can find, for example, the probability of a face with an even number of spots, which would be 0.5, etc. With the frequency definition of probability, this simple solution would not be possible. Such an approach is possible because human beings are given this innate capability of *a priori* reasoning.

Kolmogorov presented this simple idea based on *a priori* reasoning that freed everyone interested in probability from the endless arguments. His approach is referred to as the "axiomatic probability theory" and is based on set theory and measure theory. His idea was that there was no need to determine whether a coin was minted perfectly to discuss its probability. He simply turned the table around and asserted that one could "assign" probabilities to the outcomes based on the *a priori* knowledge of the outcomes and let the probabilities initially assigned be the starting point for developing more complex probability theory just like accepting 1/6 as the probability of a face in die throwing.

The key concept is in the word "assign." In this approach, probability "begins" with the assignment of it based on one's own judgment about the likelihood of the outcome. In the axiomatic approach, one can start with "assigning" 1/6 each as the probability of a face in the die-throwing experiment. Once this initial assignment of probability is "accepted" (as an axiom, so to speak), it is now possible to solve all kinds of complex and interesting probability problems associated with die-throwing.

For example, what is the probability of getting an even number of spots? Since the 1/6 probability is "accepted," one can proceed to find its answer, which is 0.5. What is the probability of getting a face with more than four spots? Since either five or six spots would make this event happen, the answer would be 2/6.

1.3 Axiomatic approach to probability

A mathematical system, e.g., linear algebra, set theory, and group theory, is simply an artifact that is useful because it provides a structure for drawing meaningful inferences. The axiomatic probability theory is such a mathematical system.

Consider a random experiment with n possible outcomes, $\xi_1, \xi_2, \ldots, \xi_n$. The probability space S is defined as the set of all possible random outcomes of a random experiment as follows:

$$S = \{ \xi_1, \xi_2, \ldots\ldots, \xi_n \} \tag{2-1}$$

A "measure" is "assigned" to each outcome, ξ_i. This measure is referred to as "probability." Denote this measure by p_i. The measure chosen is a real number between 0 and 1 as follows:

$$0 \le p_i \le 1 \tag{2-2}$$

$$p_i = P(\xi_i) = \textit{probability of random outcome } \xi_i \tag{2-3}$$

The word "probability" was difficult to define because of the attempts to define its meaning semantically and in some instances philosophically. In the axiomatic probability theory, its definition is simply a "measure" that is assigned to an outcome. In fact, this measure does not have to be a number between 0 and 1. It can be a number between 0 and 100 or any number for that matter without changing the axiomatic theory. It is conventional though to use a number between 0 and 1 as a probability measure.

An axiom is a statement accepted as a truth or a rule as a basis of inference. Given the probability space S of (2-1) and the probability measures of the random outcomes of the experiment of (2-3), the axiomatic probability theory is based on the following three simple axioms:

$$\textit{Axiom I} \quad P(A) \ge 0 \tag{2-4}$$

$$\textit{Axiom II} \quad P(S) = 1 \tag{2-5}$$

$$\textit{Axiom III} \quad \textit{If } A \cap B = \{\phi\}, P(A \cup B) = P(A) + P(B) \tag{2-6}$$

In the above equations, S is a set referred to as the probability space defined earlier. A and B are subsets of S and define the random events of interest. Since A and B define the events, they are sometimes simply referred to as "events." S is also a set and, as such, also an event. Since S includes all possible outcomes, any outcome will make S happen and so S is referred to as a certain event. Similarly, $\{\phi\}$ is a set that contains no element. No outcome will make $\{\phi\}$ happen, and $\{\phi\}$ is referred to as an impossible event. Two set operations are used in these axioms. $A \cap B$ is an intersection of A and B, a set of elements belonging to both A and B. $A \cup B$ is a union of A and B, a set of elements belonging to either A or B.

Axiom I states that any event defined in the probability space is assigned a non-negative measure or probability. This is simply an agreement to start the theory. It is entirely possible in the axiomatic theory to use negative numbers for probability as long as that is agreed to at the beginning of the framework because probability is simply nothing more than a numerical measure in the axiomatic theory. However, it would be cumbersome to think in negative numbers when one considers probability.

Axiom I defines the starting point of development of a probabilistic framework of a random experiment under consideration. First, define the elementary events $\{\xi_i\}$ and assign probabilities to them, $P(\{\xi_i\})$. Note the distinction between $P(\{\xi_i\})$ and $P(\xi_i)$. The former is the probability of the elementary event $\{\xi_i\}$ and the latter, that of a random outcome ξ_i. It is important to note that the starting point of the axiomatic framework, i.e., Axiom I, is $P(\{\xi_i\})$ and not $P(\xi_i)$.

Axiom II states that the probability of the space S is one. The space S is a set that contains all possible outcomes under consideration and it would be reasonable to accept as a basic truth that the probability of all possible outcomes is one.

In effect, Axiom II simply states that the probability of certainty is one. One may then ask what about the probability of impossibility, i.e., a null event. Don't we need an axiom, say Axiom IIa that states $P(\{\phi\}) = 0$? It can be shown that the three axioms cover this axiom and adding it would be superfluous because it can be derived from Axioms II and III as follows.

From set theory, the union of the space S and the null set $\{\phi\}$ is the space S and the intersection of the space S and the null set $\{\phi\}$ is the null set $\{\phi\}$:

$$S \cup \{\phi\} = S \tag{2-7}$$

$$S \cap \{\phi\} = \{\phi\} \tag{2-8}$$

From Equation (2-7), it follows that:

$$P(S) = P(S \cup \{\phi\}) \tag{2-9}$$

Equation (2-8) satisfies the condition for Axiom III. Hence, from Axiom III and Equation (2-9), it follows that:

$$P(S) = P(S \cup \{\phi\}) = P(S) + P(\{\phi\}) \tag{2-10}$$

From Axiom II and Equation (2-10), it follows that:

$$P(S) = P(S \cup \{\phi\}) = P(S) + P(\{\phi\}) = 1 \tag{2-11}$$

Finally, from Equation (2-11), it follows that:

$$P(\{\phi\}) = 1 - P(S) = 0 \tag{2-12}$$

Note that Axiom I states $P(A) \geq 0$ but it does not include $P(A) \leq 1$. Once again, the reason is because it can be derived from other axioms and including $P(A) \leq 1$ would be superfluous.

Example 1

A box contains a total of 10 balls of different colors as follows: two white balls, three red balls and five black balls. A player is to withdraw a ball, and, if the ball withdrawn is either red or black, the player wins a piece of candy. What is the probability of winning a piece of candy by playing this game?

Solution

There are eight red or black balls out of a total of 10 balls, and so the probability of winning the grand prize is 0.8. This is a simple problem and one can get the answer quickly in the head without going through the rigor of axiomatic formulation.

However, we shall formulate and solve this problem using the axiomatic approach to illustrate how a probability problem can be formulated and solved systematically. For more complex problems, the disciplined way of dealing with the problem using the axiomatic approach is helpful.

First define the random experiment. There are two alternative ways of defining the space and random outcomes for this problem. Either method should yield the same answer.

Formulation 1. A more direct way of formulation is to define the outcomes of ball drawing like the outcomes of die throwing. Imagine that the individual balls can be distinguished (e.g., by numbering them) as the faces of a die are distinguished. Then there are ten possible outcomes with an equal probability as follows:

$$S = \{ \xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8, \xi_9, \xi_{10} \} \tag{2-13}$$

$$p_i = P(\xi_i) = 1/10; \qquad i = 1, \ldots, 10 \tag{2-14}$$

where ξ_1 and ξ_2 are drawing a white ball, ξ_3, ξ_4 and ξ_5, a red ball and ξ_6 through ξ_{10}, a black ball.

The next step is to define the event. The event of interest is "winning a candy" and is defined as a set denoted by *W*. In set theory, a set is defined by its members or a member is "qualified" to be included in the event set, if it makes that event happen. *W* in turn depends on the following two events:

$$R = \text{"ball withdrawn is red"} = \{ \xi_3, \xi_4, \xi_5 \} \tag{2-15}$$

$$B = \text{"ball withdrawn is black"} = \{ \xi_6, \xi_7, \xi_8, \xi_9, \xi_{10} \} \tag{2-16}$$

Since $\{\xi_i\}$'s are mutually exclusive, i.e., $\{\xi_i\} \cap \{\xi_j\} = \{\phi\}$ *for i, j = 3 - 8,* it follows that:

$$R = \{ \xi_3, \xi_4, \xi_5 \} = \{ \xi_3 \} \cup \{ \xi_4 \} \cup \{ \xi_5 \}$$

$$= [\{ \xi_3 \} \cup \{ \xi_4 \}] \cup \{ \xi_5 \} \tag{2-17}$$

Applying Axiom III twice, it follows that:

$$P(R) = P(\{ \xi_3, \xi_4, \xi_5 \}) = P([\{ \xi_3 \} \cup \{ \xi_4 \}] + P(\{ \xi_5 \})$$

$$= P(\{\xi_3\}) + P(\{\xi_4\}) + P(\{\xi_5\}) = 0.3 \tag{2-18}$$

Similarly,

$$P(B) = P(\{\{\xi_6, \xi_7, \xi_8, \xi_9, \xi_{10}\}\}) = 0.5 \tag{2-19}$$

W would occur if the ball withdrawn is either red or black: W would occur if either R or B occurs. Since R and B are mutually exclusive events, it follows that:

$$R \cap B = \{\phi\} \tag{2-20}$$

$$W = R \cup B \tag{2-21}$$

Hence, from Axiom III, it follows that:

$$P\{W\} = P\{R \cup B\} = P\{R\} + P\{B\} = 0.3 + 0.5 = 0.8 \tag{2-22}$$

Formulation 2. As long as the axiomatic approach is followed, different definitions of outcomes are possible. The above formulation can be simplified by defining the experimental outcomes as the colors of the balls as follows:

$$S = \{\xi_w, \xi_r, \xi_b\} \tag{2-23}$$

where ξ_w, ξ_r and ξ_b are random outcomes of white, red and black color.

Then from the problem, the probabilities of the random outcomes can be assigned as follows:

$$P(\xi_w) = 0.2; \quad P(\xi_r) = 0.3; \quad P(\xi_b) = 0.5 \tag{2-24}$$

W would occur if ξ_r or ξ_b shows up. Hence,

$$W = \{\xi_r, \xi_b\} = \{\xi_r\} \cup \{\xi_b\} \tag{2-25}$$

Since $\{\xi_r\} \cap \{\xi_r\} = \{\phi\}$, from Axiom III and Equations (2-24) and (2-25), it follows that:

$$P\{W\} = P(\{\xi_r, \xi_b\}) = P(\{\xi_r\} \cup \{\xi_b\}) = P(\{\xi_r\}) + P(\{\xi_b\})$$

$$= 0.3 + 0.5 = 0.8 \qquad (2\text{-}26)$$

2. RANDOM VARIABLES

2.1 Definition

It is conventional to denote a random variable by a bold letter and a deterministic variable or a fixed value by a regular letter. For example, a random variable may be denoted by $\boldsymbol{x}$ and a fixed value that $\boldsymbol{x}$ can take, by x.

A random variable (RV) $\boldsymbol{x}$ is a function of a random outcome $\boldsymbol{\xi}$ of a random experiment that maps a random outcome to a real value: $\boldsymbol{x}(\boldsymbol{\xi})$. As discussed in Section 1, random outcomes could be any objects. It can be the faces of a die in die throwing, the colors of the balls in the ball drawing, etc. Random outcomes could also be real numbers, discrete or continuous. A number can just be an object of a set. A term "real line" is used to denote the set of all real numbers, i.e., the continuum, from $-\infty$ to $+\infty$. Since the real line is a continuum, discrete points are also included in the set. Figure 2-1 illustrates the mapping from $\boldsymbol{\xi}$ to x on the real line.

In the die throwing experiment, the space is a set of six possible outcomes:

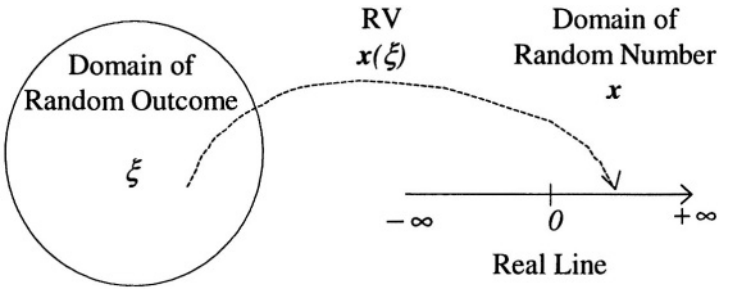

Figure 2-1. Mapping from $\boldsymbol{\xi}$ to x on the real line by $\boldsymbol{x}(x)$.

$$S = \{\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6\} \tag{2-27}$$

These six outcomes are not numbers; they are simply objects that constitute the set *S*. An RV is a function that relates these objects to real numbers. An RV must first be defined just as a function must be defined. Let us now define a random variable $\boldsymbol{x}$ that maps the six objects of *S* to a set of real numbers, i.e., onto the real line. To illustrate the concept, suppose that a player is paid \$1 to \$6 depending on the number of spots on the face as follows:

Random Outcome, ξ	Payoff $\boldsymbol{x}(\xi)$
One spot	\$1
Two spots	\$2
Three spots	\$3
Four spots	\$4
Five spots	\$5
Six spots	\$6

In this example, the RV $\boldsymbol{x}(\xi)$ maps the six outcomes to real numbers representing payoffs. Since, in this case, the random outcomes are (non-numerical) objects, there is no convenient way of expressing the functional relationship $\boldsymbol{x}(\xi)$. The best way of "defining" the RV is by a table such as the one above.

Having introduced this basic concept of RV, we now extend the concept to a little more abstract situation. Suppose now that the space of random outcomes *S* is itself the real line:

$$S = \{x \mid x \varepsilon (-\infty, +\infty)\} \tag{2-28}$$

In set theory, the above expression is read as follows: "*S* is a set of *x,* where *x is* a member (as denoted by $\boldsymbol{\varepsilon}$) of an interval of real numbers from $-\infty$ to $+\infty$." It can also be specified that *x* is an integer. In that case, *S* is a set of all integers from $-\infty$ to $+\infty$." An RV $\boldsymbol{x}$ can now be defined as a function on *S* that maps *x of S* to *x*, $\boldsymbol{x}(x) = x$. This is illustrated in Figure 2-2.

It could be less confusing, if the real numbers of *S* were denoted by a different symbol such as *y*; however, this would be even more confusing because then *y* and *x* can take on different values. For now, consider $\boldsymbol{x}(x) = x$ to read as follows: "RV $\boldsymbol{x}$ maps *x* of *S* to itself *x."* In most situations of random variables that we are familiar with, this is the definition tacitly used.

For example, when we say that the temperature in a certain area is a random variable $\boldsymbol{x}$, we cannot possibly mean that the random temperature is a result of multitudes of random outcomes of the nature. It may be possible

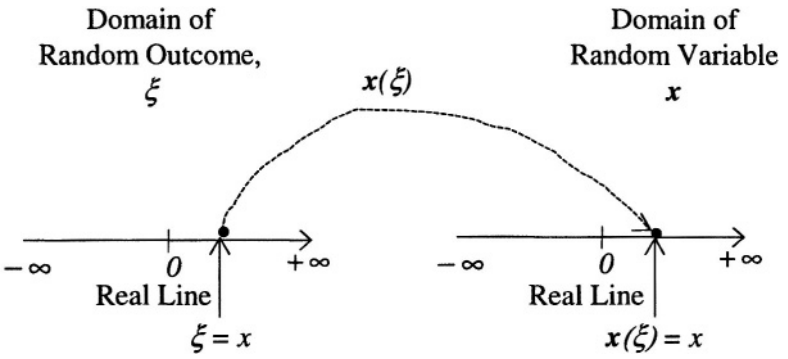

Figure 2-2. Mapping from $\xi=x$ on real line to x on real line by $x(\xi)$.

to do so in certain circumstances. In most cases, however, the way we interpret the random temperature x is as follows. We measure the temperature, i.e., perform a "trial," and take its reading as a random outcome. We then take this random outcome as the value of the random variable, i.e., $x(x) = x$.

Suppose now that the domain of random outcomes ξ is the real line, a continuum. A continuous random variable x is a function of random outcome ξ that maps the specific value x of the random outcome ξ from the continuum of the real line $(-\infty, +\infty)$ (i.e., $\xi = x$) to itself x. Figure 2-2 illustrates this definition: the continuous random variable x maps random outcomes, i.e., $\xi = x$, on the real line to the same value x on the real line, i.e., mapping from x to x by $x(\xi)$.

Finally, an RV may be defined to map multiple outcomes to a single number, i.e., many to one; however, an RV cannot map a single outcome to multiple numbers.

2.2 CDF and pdf

Let x be a random variable (RV). Its cumulative distribution function (CDF) is defined as follows:

$$F_x(x) = P\{x \le x\} \tag{2-29}$$

$P\{x \le x\}$ reads: "the probability that the RV x will be less than a value x."

The probability density function (pdf) of x is defined as follows:

$$f_x(x) = \frac{dF_x(x)}{dx} \tag{2-30}$$

From the above definition, $F(x)$ can also be given by the following integral:

$$F(x) = \int_{-\infty}^{x} f(z)dz \tag{2-31}$$

Conceptually, it is easier to interpret the pdf in the following way. Consider the probability that the random variable x will lie in the small interval between x and $x+\Delta x$. From the definition of the CDF $F(x)$, this probability is obtained as follows:

$$P\{x \prec x \leq x + \Delta x\} = P\{x \leq x + \Delta x\} - P\{x \leq x\}$$

$$= F(x + \Delta x) - F(x) = \int_{-\infty}^{x+\Delta x} f(z)dz - \int_{-\infty}^{x} f(z)dz = \int_{x}^{x+\Delta x} f(z)dz \approx f(x)\Delta x \tag{2-32}$$

From the above, we have:

$$f(x) \approx \frac{P\{x \prec x \leq x + \Delta x\}}{\Delta x} \tag{2-33}$$

Taking the limit, we have:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \prec x \leq x + \Delta x\}}{\Delta x} \tag{2-34}$$

From the above, we see that the pdf $f(x)$ is the probability that x will lie in a small interval of length Δx divided by the interval length Δx as Δx becomes infinitesimally small. This is illustrated in Figure 2-3.

The word "density" refers to the fact that the small probability $P\{x < x \leq x + \Delta x\}$ is normalized by the interval length Δx.

For a discrete random variable x, the pdf is given by:

$$f(x) = \sum_i P\{x = x_i\}\delta(x - x_i) = \sum_i p_i \delta(x - x_i) \tag{2-35}$$

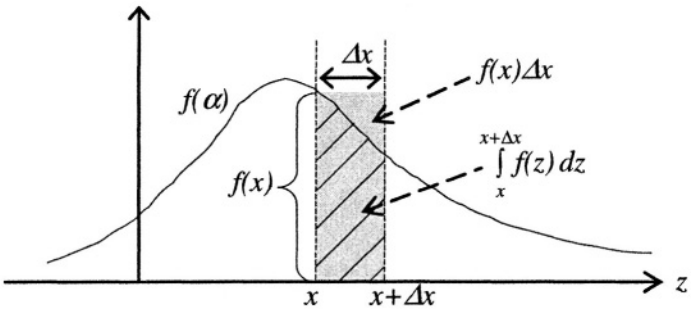

Figure 2-3. Definition of the probability density function (pdf).

$$\text{where } p_i = P\{ x = x_i \} \quad i \varepsilon \{integers\} \tag{2-36}$$

$$\int_a^b \delta(x - x_i)dx = 1 \quad if\ a \prec x_i \leq b$$

$$= 0 \quad if\ b \prec x_i\ or\ x_i \leq a \tag{2-37}$$

The impulse function, $\delta(x)$, as defined above, has the following property. It produces a value when it is integrated, and, without the integration, $\delta(x)$ is undefined. If the integration interval $a \sim b$ includes x_i, the integration of $\delta(x - x_i)$ over this interval is *1*; if x_i lies outside of the integration interval, the integration of $\delta(x - x_i)$ over the interval is zero.

The impulse function is a mathematical artifact convenient for expressing mathematically the pdf of a discrete random variable x, as given by Equation (2-35). For the pdf $f(x)$ of a discrete random variable x defined in terms of the impulse function $\delta(x)$, it is possible to express the CDF $F(x)$ as the integral of $f(x)$ as follows:

$$F(x) = \int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} \sum_i p_i \delta(z - x_i)dz$$

$$= \sum_i \{ \int_{-\infty}^{x} p_i \delta(z - x_i)dz \} = \sum_{i=i_{min}}^{i=i_{max}} p_i \tag{2-38}$$

where $p_{i_{min}} = P\{x = x_{i_{min}}\}$ (2-39)

$x_{i_{min}}$ = *smallest discrete value that* x *can take, which is* $\leq x$ (2-40)

$p_{i_{min}} = P\{x = x_{i_{max}}\}$ (2-41)

$x_{i_{max}}$ = *largest discrete value that* x *can take, which is* $\leq x$ (2-42)

2.3 Mean and variance

Consider the ball drawing game of Example 1 discussed earlier. Define an RV x as the payoffs of the game as follows: \$10 if a white ball is drawn, \$20 for a red ball, and \$30 for a black ball. What is the amount of money a player can expect to win by playing this game?

To answer this question, the probabilities of drawing the three colors need to be determined as follows:

$$P\{white\} = \frac{2}{10} = 0.2;\ P\{red\} = \frac{3}{10} = 0.3;\ P\{black\} = \frac{5}{10} = 0.5 \quad (2\text{-}43)$$

The expected amount of payoff is calculated by:

$$E\{x\} = (\$10 \times 0.2) + (\$20 \times 0.3) + (\$30 \times 0.5) = \$23 \quad (2\text{-}44)$$

The expected value is also referred to as two other common terms, "mean" and "average." The term average is used because if the player plays the game long enough performing many "trials," then the average winning, which is determined by dividing the total amount of money won by the number of trials, should approach the expected value, i.e.,

$$\frac{x_1 + x_2 + \ldots + x_N}{N} \rightarrow \$23 \text{ as } N \rightarrow \infty \quad (2\text{-}45)$$

where N is the number of times of playing, and x_i is the i^{th} payoff. In general, the expected value of a discrete random variable x taking on the values of x_i with the probability p_i, $i = 1, 2, \ldots, N$ is:

$$E\{x\} = \sum_{i=1}^{N} x_i p_i \tag{2-46}$$

$$\text{where } p_i = P\{x = x_i\} \quad i = 1, 2, \ldots, N \tag{2-47}$$

To extend the above concept to a continuous random variable x as defined in Figure 2-2, imagine a similar game in which a player takes a measurement from the real line $(-\infty, +\infty)$ and receives a payoff equal to the measurement: payoff x, is x, i.e., $x(\xi = x) = x$. Now consider a small interval of width Δx from x to $x+\Delta x$ on the real line of x domain and a random payoff x falling in this interval as shown in Figure 2-4. The value of $x(\xi)$ in this interval is somewhere between x and $x+\Delta x$, i.e., $x \le x(\xi) \le x+\Delta x$, and so is approximately equal to x, if Δx is small enough. In fact, Δx can be made as small as necessary to make the value of $x(\xi)$ as close to x as possible.

The expected value of the payoff for this small interval is therefore approximately equal to x times the probability that x will fall in this interval as follows:

$$E\{\text{payoff of } x \text{ falling in } (x, x+\Delta x)\} \approx xP\{x \prec x \le x + \Delta x\} \tag{2-48}$$

$$\approx xf(x)\Delta x. \tag{2-49}$$

By taking the limit,

$$E\{\text{payoff of } x \text{ falling in } (x, x+\Delta x)\} = \lim_{\Delta x \to 0} xf(x)\Delta x \tag{2-50}$$

Hence,

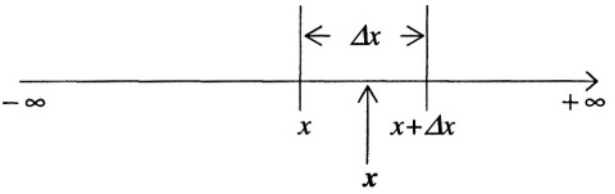

Figure 2-4. Random variable x falling in Dx.

$$E\{x\} = \lim_{\Delta x \to 0} \sum xf(x)\Delta x = \int_{-\infty}^{+\infty} xf_x(x)dx = \eta_x \tag{2-51}$$

The variance of a random variable x is a measure of the variability of x around its mean, η_x. It is the expected value of the square of the difference between the random variable x and its mean η_x as follows:

$$\sigma_x^2 = E\{(x - \eta_x)^2\} = \int_{-\infty}^{+\infty} (x - \eta_x)^2 f_x(x)dx \tag{2-52}$$

The difference is squared because the magnitude of the variation rather than its direction is of primary interest. From the above, the following equation is derived:

$$\sigma_x^2 = E\{x^2\} - \eta_x^2 \tag{2-53}$$

The square root of the variance is the standard deviation:

$$\sigma_x = \sqrt{\sigma_x^2} \tag{2-54}$$

2.4 The normal distribution

Two of the most widely used and important distributions are the normal or Gaussian distribution and the Poisson distribution. The normal random variable x is a continuous random variable with the following pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\eta)^2/2\sigma^2} \tag{2-55}$$

where η is the mean of x and σ is the standard deviation of x.

The CDF of a normal random variable x is the integral of $f(x)$ as follows:

$$F(x) = \int_{-\infty}^{x} f(z)dz = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(\alpha-\eta)^2/2\sigma^2} dz \tag{2-56}$$

The normal CDF given by the above integral is tabulated in mathematical tables. It can be shown that the mean and variance of the normal random variable x with the above pdf is η and σ^2. It is significant that, if an RV x is

normal, its pdf can be completely determined by two parameters, mean and variance.

2.5 The Poisson distribution

A Poisson random variable x is a discrete random variable with the following pdf:

$$f(x) = \sum_{k=0}^{\infty} p_k \delta(x - k) \tag{2-57}$$

$$\text{where } p_k = P\{ x = k \} = e^{-\lambda} \frac{\lambda^k}{k!} \tag{2-58}$$

In Equation (2-58), k is an integer taking on a value from *0* to infinity. Putting Equations (2-57) and (2-58) together, the Poisson pdf is given by:

$$f(x) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta(x - k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \delta(x - k) \tag{2-59}$$

The mean and variance of the Poisson random variable x with the parameter λ are both found to be λ as follows:

$$\eta = E\{ x \} = \lambda \qquad \sigma^2 = E\{ x^2 \} - \eta^2 = \lambda \tag{2-60}$$

The Poisson pdf is defined by a single parameter λ. It is significant that, if an RV x is Poisson, its pdf can be completely determined by a single parameter, λ. More will be discussed on the Poisson pdf and the parameter λ later in this chapter.

3. STOCHASTIC PROCESSES

3.1 Definition of a stochastic process

A random variable x is a static variable defined on random outcomes, "static" in the sense that time is fixed for an RV: an RV is a function of random outcome ξ, $x(\xi)$, but time is not an argument of this function.

A stochastic process $x(t)$ is the random variable x extended into another dimension t. To be rigorous in notation, we might write a stochastic process as $x(\xi, t)$, where x is a function of two variables, time t and random outcome ξ. For simplicity, however, $x(\xi, t)$ is written $x(t)$ as the random variable $x(\xi)$ is written x.

To define it in more general terms, a stochastic process is a set of random variables arranged in time as follows:

$$x(t) = \{ x(t) \mid t \varepsilon (-\infty, +\infty) \} \tag{2-61}$$

$$x(t) = \{ x(t) \mid t \varepsilon (t_1, t_2, ..., t_N) \} \tag{2-62}$$

$$x(t) = \{ x(t) \mid t \varepsilon (t_a, t_b) \} \tag{2-63}$$

If the interval (t_a, t_b) is a continuum of time, the stochastic process is referred to as a continuous process; if the interval (t_a, t_b) is a set of discrete time points, t_i's, the stochastic process is referred to as a discrete process.

3.2 CDF and pdf of stochastic process

A useful concept to remember is that, once the time t is fixed at a specific value, say t^*, the stochastic process yields a random variable; that is, $x(t^*)$ is a random variable. Consider a stochastic process $x(t)$. Let us fix the time t to t^* and consider the stochastic process at the instant t^*. At time t^*, $x(t^*)$ is a random variable. Consider the CDF defined earlier for this random variable $x(t^*)$:

$$F(x) = P\{ x(t^*) \leq x \} \tag{2-64}$$

Now let us take a leap and fix time t to an arbitrary value, say t, and write the above equations as follows:

$$F(x) = P\{ x(t) \leq x \} \tag{2-65}$$

Once we fix time t to an arbitrary value t, the first-order CDF of $x(t)$ is a function of time t as follows:

$$F(x,t) = P\{ x(t) \leq x \} \tag{2-66}$$

The first-order refers to the fact that one random variable is considered, i.e., random variable defined at one time point. The first order pdf of $x(t)$ is given by:

$$f(x,t) = \frac{\partial F(x,t)}{\partial x} \tag{2-67}$$

The mean of $x(t)$ is

$$\eta(t) = E\{x(t)\} = \int_{-\infty}^{+\infty} x f(x,t) dx \tag{2-68}$$

Now consider two time points t_1 and t_2 and the two random variables defined for these time points, $x(t_1)$ and $x(t_2)$. The statistics considered for these two random variables is referred to as the second-order statistics, and the joint CDF and joint pdf for $x(t_1)$ and $x(t_2)$ are as follows:

$$F(x_1, x_2; t_1, t_2) = P\{x(t_1) \le x_1, x(t_2) \le x_2\} \tag{2-69}$$

$$f(x_1, x_2; t_1, t_2) = \frac{\partial^2 F(x_1, x_2; t_1, t_2)}{\partial x_1 \partial x_2} \tag{2-70}$$

3.3 Autocorrelation and cross-correlation

An important concept in stochastic processes is the autocorrelation function. It is a measure of the correlation between two random variables defined at two time points for the same stochastic process. The pre-fix "auto" signifies that the correlation is considered for the same process. Later, the cross-correlation function defines the same between two different processes.

For a real process $x(t)$, consider the two real random variables defined for two time points, $x(t_1)$ and $x(t_2)$. The autocorrelation function of $x(t)$, denoted by $R(t_1, t_2)$, is defined as the expected value of the product of $x(t_1)$ and $x(t_2)$, where t_1 and t_2 are variables:

$$R_{xx}(t_1, t_2) = E\{x(t_1)x(t_2)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_1 x_2 f(x_1, x_2; t_1, t_2) dx_1 dx_2 \tag{2-71}$$

For $t_1 = t_2 = t$, we have

$$R_{xx}(t,t) = E\{x(t)x(t)\} = E\{x^2(t)\} \tag{2-72}$$

In general, for a complex process $x(t)$, the autocorrelation of $x(t)$ is given by:

$$R_{xx}(t,t) = E\{x(t)x^*(t)\} \tag{2-73}$$

where $x^*(t)$ is the complex conjugate of $x(t)$. The autocovariance of a real process $x(t)$, denoted by $C_{xx}(t_1, t_2)$, is given by:

$$C_{xx}(t_1,t_2) = E\{[(x(t_1) - \eta_x(t_1))][(x(t_2) - \eta_x(t_2))]\}$$

$$= R_{xx}(t_1,t_2) - \eta_x(t_1)\eta_x(t_2) \tag{2-74}$$

The variance of $x(t)$ is then given by:

$$var\{x(t)\} = \sigma^2_{x(t)} = E\{[(x(t) - \eta_x(t))]^2\} \tag{2-75}$$

$$= C_{xx}(t,t) = R_{xx}(t,t) - \eta_x(t)^2 = E\{x^2(t)\} - \eta(t)^2 \tag{2-76}$$

In general, for a complex process $x(t)$, the autocovariance of $x(t)$ is given by:

$$C_{xx}(t_1,t_2) = E\{[(x(t_1) - \eta_x(t_1))][(x^*(t_2) - \eta^*_x(t_2))]\} \tag{2-77}$$

$$= R_{xx}(t_1,t_2) - \eta_x(t_1)\eta^*_x(t_2) \tag{2-78}$$

The correlation coefficient $r(t_1, t_2)$ of a process $x(t)$ *is* given by:

$$r(t_1,t_2) = \frac{C_{xx}(t_1,t_2)}{\sqrt{C_{xx}(t_1,t_1)C_{xx}(t_2,t_2)}} \tag{2-79}$$

The cross-correlation is a measure of the correlation between two random variables defined at two time points from two different stochastic processes, $x(t)$ and $y(t)$. Consider the two random variables defined for two time points, t_1 for the real process $x(t)$ and t_2 for the real process $y(t)$: $x(t_1)$ and $y(t_2)$. The cross-correlation of real processes $x(t)$ and $y(t)$, denoted by $R_{xy}(t_1, t_2)$, is defined as the expected value of the product of $x(t_1)$ and $y(t_2)$, where t_1 and t_2 are variables:

$$R_{xy}(t_1,t_2) = E[x(t_1)y(t_2)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_1 y_2 f_{xy}(x_1, y_2; t_1, t_2)dx_1 dy_2 \quad (2\text{-}80)$$

For $t_1 = t_2 = t$, we have

$$R_{xy}(t,t) = E[x(t)y(t)] \quad (2\text{-}81)$$

In general, for two complex processes $x(t)$ and $y(t)$, the cross-correlation of $x(t)$ and $y(t)$ is given by:

$$R_{xy}(t_1,t_2) = E[x(t_1)y^*(t_2)] \quad (2\text{-}82)$$

The cross-covariance of real processes $x(t)$ and $y(t)$, denoted by $C_{xy}(t_1, t_2)$, is defined by:

$$C_{xy}(t_1,t_2) = E\{[(x(t_1) - \eta_x(t_1))][(y(t_2) - \eta_y(t_2))]\} \quad (2\text{-}83)$$

$$= R_{xy}(t_1,t_2) - \eta_x(t_1)\eta_y(t_2) \quad (2\text{-}84)$$

In general, for complex processes $x(t)$ and $y(t)$, their cross-covariance is given by:

$$C_{xy}(t_1,t_2) = E\{[(x(t_1) - \eta_x(t_1))][(y^*(t_2) - \eta_y^*(t_2))]\} \quad (2\text{-}85)$$

$$= R_{xy}(t_1,t_2) - \eta_x(t_1)\eta_y^*(t_2) \quad (2\text{-}86)$$

3.4 The normal process

A stochastic process $x(t)$ is normal, if the n random variables defined for n arbitrarily selected time points, $x(t_1), x(t_2), \ldots, x(t_n)$, are jointly normal for any n. For $n = 1$, the first-order statistics of the normal process is:

$$f(x,t) = \frac{1}{\sigma(t)\sqrt{2\pi}} e^{-(x-\eta(t))^2/2\sigma^2} \tag{2-87}$$

For $n = 2$, the second-order statistics of the normal process is:

$$f(x_1, x_2; t_1, t_2) = \frac{1}{2\pi\sigma_1(t)\sigma_2(t)\sqrt{1 - r^2(t_1, t_2)}}$$

$$\times exp[-\frac{1}{2[1 - r^2(t_1, t_2)]}(\frac{x_1^2}{\sigma_1^2(t)} - 2r(t_1, t_2)\frac{x_1 x_2}{\sigma_1(t)\sigma_2(t)} + \frac{x_2^2}{\sigma_2^2(t)})] \tag{2-88}$$

3.5 Statistical characterization of a stochastic process

How does one go about characterizing a stochastic process $x(t)$ statistically? How does one know that a stochastic process $x(t)$ is "completely" characterized statistically? What statistical information or data characterizes a stochastic process $x(t)$ completely? These are the same question phrased differently. The key word is "completely." Unless the ultimate, i.e., "complete," statistical information is defined, it would be hard to determine what to search for and when to stop collecting data.

To discuss this concept, let us start with a simple random variable x. The complete statistical information of x is its CDF or pdf. The CDF or the pdf of x represents all the data one can possibly have statistically for x. If you have the pdf of x, you can derive the mean, the variance, and the higher order moments of x. However, the converse is in general not true unless the random variable is known or assumed to be a certain kind, e.g., normal, Poisson, etc. The statistical moments of x are not in general sufficient to derive the CDF or pdf of x. For certain types of random variables, e.g., the normal random variable x, however, the mean and variance of x are sufficient to determine the pdf of x.

Let us now carry this discussion to the stochastic process $x(t)$. Recall that $x(t)$ is a set or collection of random variables in time t. To characterize $x(t)$, proceed as follows. First, pick n arbitrary time points, $t_1, t_2, \ldots, t_n$. For

these n time points, we now have n random variables: $x(t_1), x(t_2), \ldots\ldots, x(t_n)$. For these n random variables, consider the n^{th}-order statistics, i.e., n^{th}-order joint CDF or joint pdf, as follows:

$$F(x_1, x_2, \ldots\ldots, x_n; t_1, t_2, \ldots\ldots, t_n)$$

$$= P\{x(t_1) \le x_1, x(t_2) \le x_2, \ldots\ldots, x(t_n) \le x_n\} \qquad (2\text{-}89)$$

$$f(x_1, x_2, \ldots\ldots, x_n; t_1, t_2, \ldots\ldots, t_n)$$

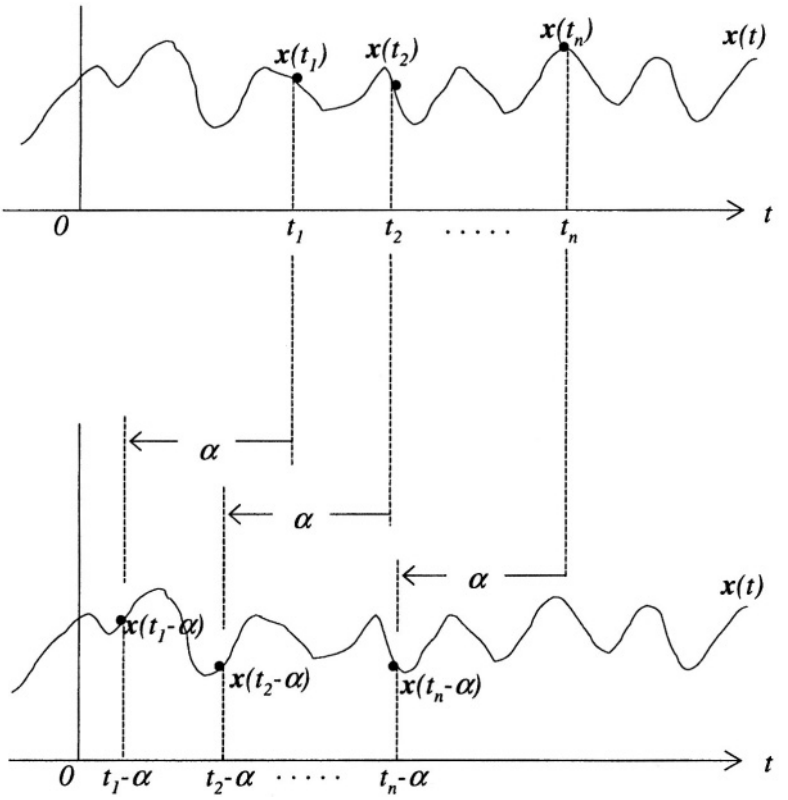

Figure 2-5. Strict sense stationary (SSS).

$$= \frac{\partial^n F(x_1, x_2, \ldots\ldots, x_n; t_1, t_2, \ldots\ldots, t_n)}{\partial x_1 \partial x_2 \ldots\ldots\ldots \partial x_n} \qquad (2\text{-}90)$$

To define a stochastic process $x(t)$ completely statistically, one must

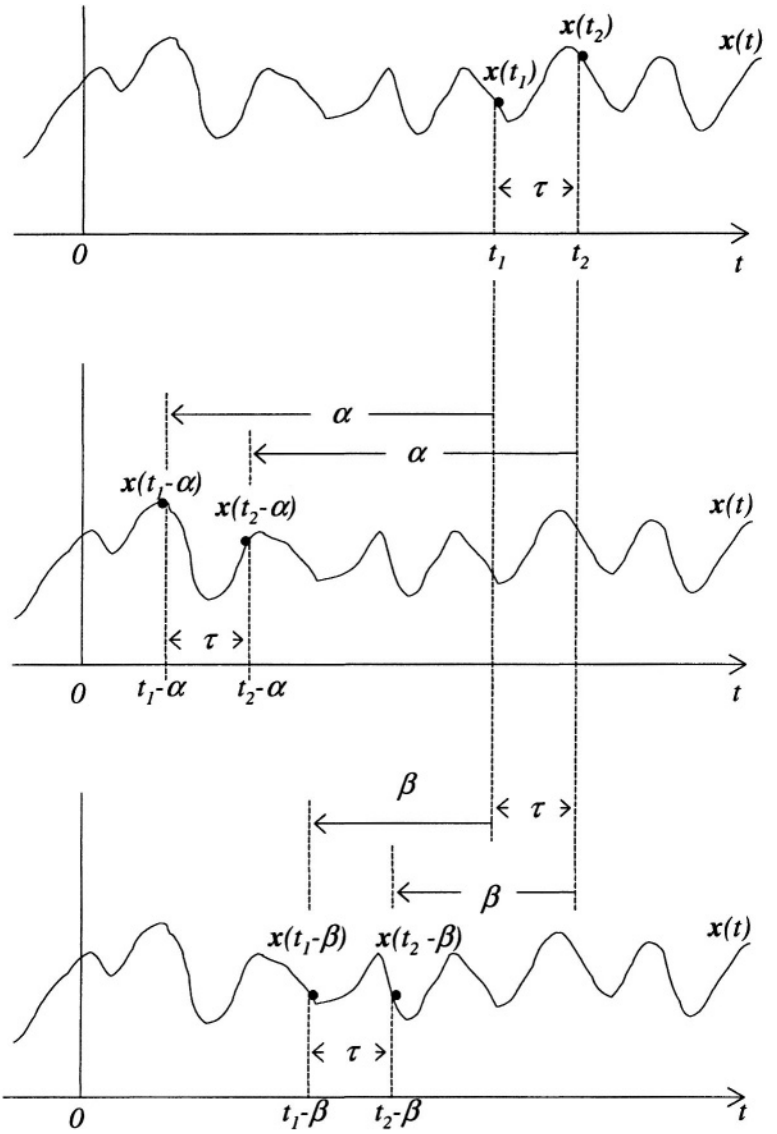

Figure 2-6. Strict sense stationarity.

obtain the n^{th}-order joint CDF or pdf defined above for an arbitrary n and for an arbitrary set of n time points. For a continuous process $x(t)$, one can see that it is not possible to complete the process of data collection to satisfy this definition. First of all, n can be indefinitely large. Secondly, for a given n, there are an infinite number of possibilities of choosing the n time points.

Nevertheless, this ideal definition of "statistical characterization" of a stochastic process $x(t)$ provides a framework for statistical investigators of a random process: i) to determine how to go about collecting data and ii) to determine when to stop collecting data, i.e., how much data is enough.

3.6 Stationarity

A very important concept in stochastic processes is *stationarity*. In practice, unless the stochastic process under consideration is stationary, it is in general intractable for analysis or simulation. If a process is non-stationary, therefore, it can be divided into a number of sub-processes defined over smaller sub-time intervals over which the sub-processes are either stationary or approximately stationary. The analysis of the original non-stationary process can then be performed by analyzing the sub-processes and synthesizing the results of the sub-processes. Two types of stationarity are defined: strict sense stationarity (SSS) and wide sense stationarity (WSS). SSS is stronger (or harder to meet) than WSS.

3.6.1 Strict sense stationarity (SSS)

A stochastic process $x(t)$ is strict sense stationary (SSS) if its statistical characterization defined in Section 3.5 is invariant to a shift of the origin. This simple definition has the following profound implication. Suppose that,

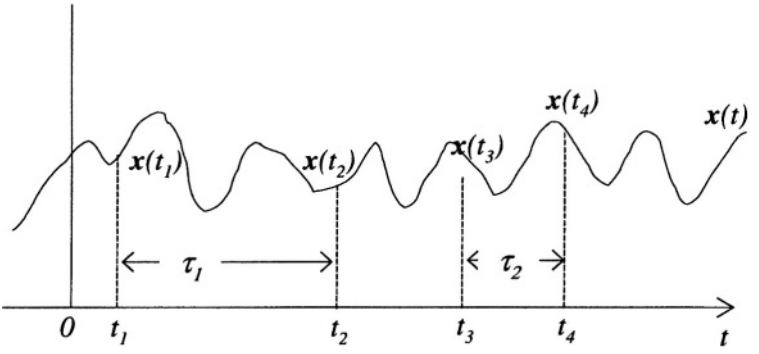

Figure 2-7. SSS.

to characterize $x(t)$, a set of n time points and the corresponding n random variables are selected: $x(t_1), x(t_2), \ldots, x(t_n)$. Next, shift the original n random variable to the left in time by the same amount α and consider the resulting n random variables: $x(t_1 - \alpha), x(t_2 - \alpha), \ldots, x(t_n - \alpha)$. Figure 2-5 shows the original n RV's and the n shifted RV's

To satisfy the above definition of SSS, the two sets of n random variables must have the same n^{th}-order joint pdf as follows:

$$f(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)$$

$$= f(x_1, x_2, \ldots, x_n; t_1 - \alpha, t_2 - \alpha, \ldots, t_n - \alpha) \tag{2-91}$$

Just as the complete statistical characterization of $x(t)$ is empirically not possible, establishing strict sense stationarity empirically is not impossible because the above equality must be established for any n and for any arbitrary n time points, and, finally, for any value of α. If $x(t)$ is SSS satisfying the above general equation, the following properties can be derived. For all values of α:

$$f(x;t) = f(x;t-\alpha) \tag{2-92}$$

The above equation indicates that the 1st order pdf of SSS $x(t)$ does not change as time t is varied. This means that $f(x;t)$ of an SSS $x(t)$ is independent of t: that is, for all values of t, $x(t)$ has the same pdf.

$$f(x;t) = f(x) \tag{2-93}$$

This means that an SSS $x(t)$ has a constant mean and a constant variance:

$$\eta_x(t) = \eta_x; \qquad \sigma_x^2(t) = \sigma_x^2 \tag{2-94}$$

For an SSS $x(t)$, the following is true for all values of α:

$$f(x_1, x_2; t_1, t_2) = f(x_1, x_2; t_1 - \alpha, t_2 - \alpha) \tag{2-95}$$

In the above equation, let $\tau = t_2 - t_1$, and rewrite it as follows:

$$f(x_1, x_2; t_1, t_2) = f(x_1, x_2; t_1 - \alpha, t_2 + t_1 - t_1 - \alpha)$$

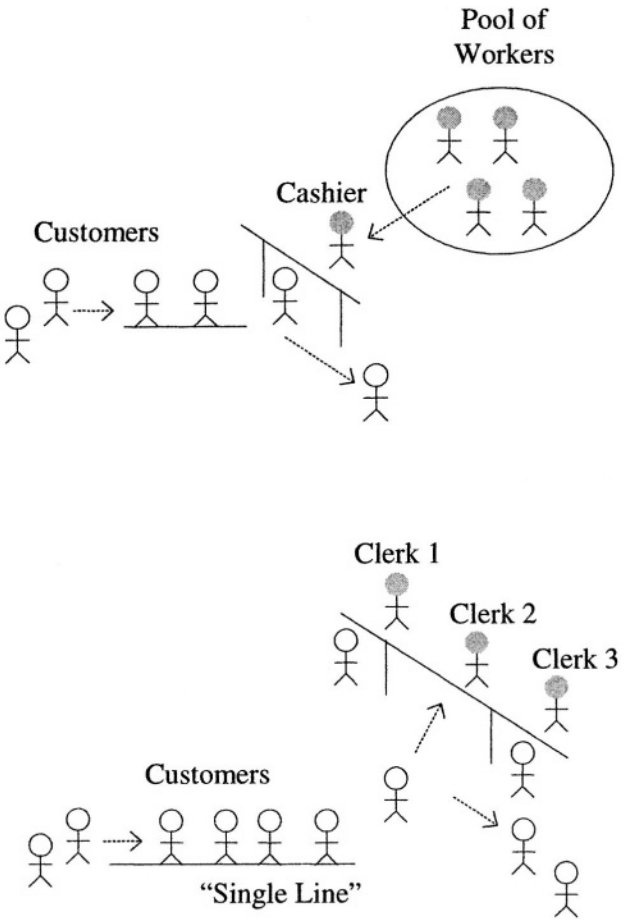

Figure 2-8. Real-life queuing example 1.

$$= f\{x_1, x_2; t_1 - \alpha, t_1 + (t_2 - t_1) - \alpha\}$$

$$= f(x_1, x_2; t_1 - \alpha, t_1 - \alpha + \tau) = f(x_1, x_2; \tau) \qquad (2\text{-}96)$$

In Figure 2-6, $x(t_1)$ and $x(t_2)$ are shifted to the left by α and β, respectively, keeping the distance between t_2 and t_1 at a constant τ. In this case, the 2nd-order pdf stays the same for both α and β; that is, as long as the distance between the two time points τ is constant, the amount of shift does

not change the 2nd-order pdf: the 2nd-order pdf of $\{x(t_1+\alpha), x(t_2+\alpha)\}$ and that of $\{x(t_1+\beta), x(t_2+\beta)\}$ are the same.

Figure 2-7 shows two pairs of time points and the random variables defined for those four time points: $\{x(t_1), x(t_2)\}$; $\{x(t_3), x(t_4)\}$. The distances between t_1 and t_2 and between t_2 and t_3 are τ_1 and τ_2, respectively. If $\tau_1 \neq \tau_2$, the 2nd-order pdf's of the two pairs of random variables are in general not equal as follows:

$$f(x_1, x_2; t_1, t_2) \neq f(x_3, x_4; t_3, t_4) \quad \text{if } \tau_1 \neq \tau_2 \tag{2-97}$$

where $\tau_1 = t_2 - t_1; \tau_2 = t_4 - t_3$.

3.6.2 Wide sense stationarity (WSS)

A stochastic process $x(t)$ is wide sense stationary (WSS) if the following two conditions are met. First, the mean of $x(t)$ is constant, i.e., independent of t:

$$\eta_x(t) = E\{x(t)\} = \eta_x \tag{2-98}$$

The second condition for WSS is that the autocorrelation function $R(t_1, t_2)$ is a function of the difference between t_2 and t_1, i.e., τ, only:

$$R_{xx}(t_1, t_2) = E\{x(t_1)x^*(t_2)\} = R_{xx}(t_1 - t_2) = R_{xx}(\tau) \tag{2-99}$$

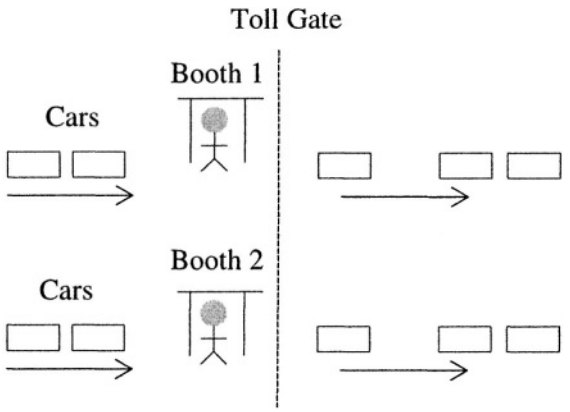

Figure 2-9. Toll gate example.

4. QUEUING THEORY BASICS

The QoS mechanisms implemented in packet routers and switches use various types of queuing discipline. Some basic understanding of queuing theory will help the reader appreciate the QoS performance analysis. This section reviews the following topics:

- Real-life examples of queuing
- Random arrivals
- Random service times
- Utilization factor
- Queuing system metrics
- *M/M/1* queue

4.1 Real-life examples of queuing

Consider several examples of queuing situations that we all experience in our everyday lives. In Figure 2-8, two real life examples of queuing are shown. A customer comes to a place, say, a fast food restaurant, to be served, joins the queue, and, when his turn comes, receives the service and leaves the place. The customer is served by a cashier. Behind the cashier, however, a team of workers help provide the service. How fast the service is provided can be controlled by controlling the number of workers in the pool. If the service is too slow, the restaurant manager can hire more workers and add them to the pool; if the manager considers that operation is too expensive, the manager can reduce the work force. In the former case, the customer service would improve; in the latter case, it would deteriorate.

The second example of Figure 2-8 is a typical bank example. There is a single line of customers and multiple tellers are serving the line. Whenever a teller becomes available the customer at the Head of Queue (HoQ) moves forward to the teller, receives the service and leaves.

The two cases shown in Figure 2-8 are similar and can be modeled by a single line single server queue. In the bank example, the multiple tellers can be considered a single server, i.e., a single pool of tellers, serving the single line. Once again, how fast the service can be provided can be controlled by the number of tellers serving the line.

Figure 2-9 shows another example that we experience at toll gates. A car comes to a toll gate and "randomly" selects one of the lanes for a toll booth, pays the toll and leaves. If all of the booths are assumed to be equal in service rates and other characteristics and that the cars select a booth randomly, the toll gate example can also be modeled as a single-line single-server queue.

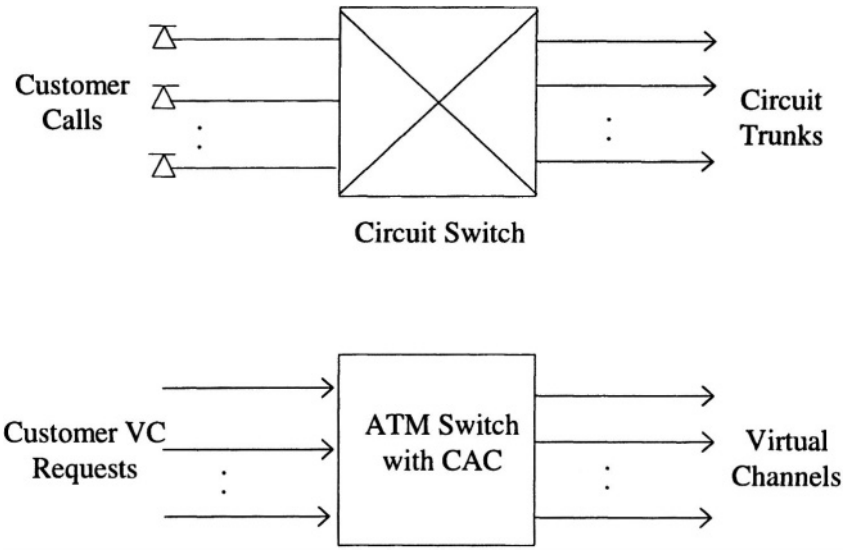

Figure 2-10. Network switching examples.

Figure 2-10 shows two switching examples: telephone calls served by trunks at a circuit switch and virtual connection requests served by an ATM switch. Customers' telephone lines are served by a central office. When a customer attempts to make a call going outside of the serving central office, the switch first tries to find an idle trunk for the call. If an idle trunk is available, the call is place on that trunk. If no trunk is available, the customer gets an "all trunks busy" signal.

In the ATM example in the figure, a virtual connection request arrives at

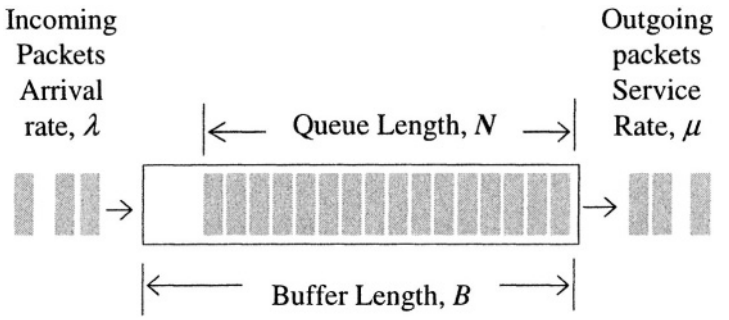

Figure 2-11. Packets arriving at a packet switch.

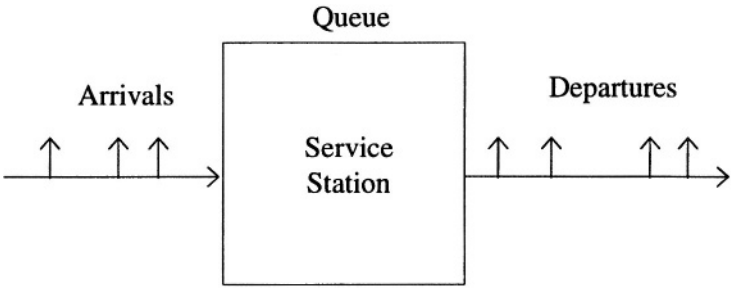

Figure 2-12. Queuing system.

the ATM switch. A Connection Admission Control (CAC) algorithm implemented in the switch determines whether there is enough bandwidth available at the output port to accept the request. If the answer is affirmative, the virtual connection request is accepted; otherwise, it is rejected. This type of queuing system is analyzed by the Erlang B and C systems. The Erlang systems will be discussed in detail in Chapter 3 and also in Chapter 6 for the ATM CAC.

Finally, Figure 2-11 shows the incoming packets that are put into a buffer in an IP router. The packet scheduler determines which packets are sent out from the output port. Various types of packet schedulers are treated in detail in Chapter 4.

The following are some of the typical questions that would be of interest in various types of queuing situations:

- How long would the customer line be?
- How long would a customer wait?
- How quick would the service be?

Queuing theory is a mathematical discipline that addresses this type of questions.

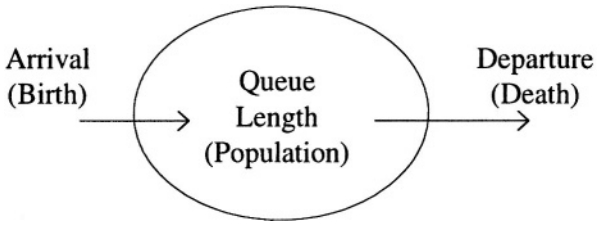

Figure 2-13. Birth-death process model.

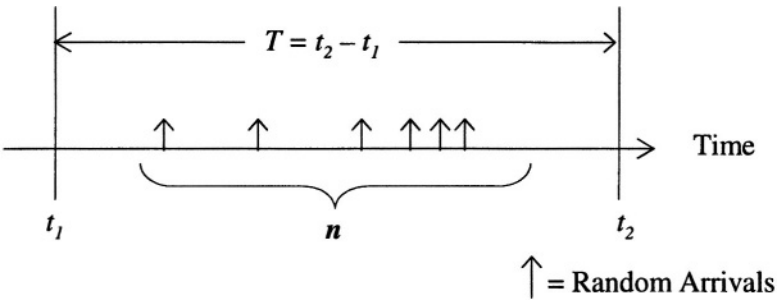

Figure 2-14. Packets arriving at a packet switch.

4.2 Definition of queuing system

In queuing theory, a place or "entity" where some kind of service is rendered is generically referred to as a "service station;" the word "customer" refers to any object that arrives at a service station to get a service. For example, a packet in Figure 2-11 is a "customer" in this sense.

In the examples described in Section 4.1, the following common features may be noted. A customer comes to a "service station," joins a line or "queue" and waits for his/her turn. When the turn comes, the customer receives the service and departs the service station. Figure 2-12 shows a queuing system. A queuing system is a mathematical model of the situation where customers arrive randomly at a service station, get service and depart the service station. A queuing system is defined by probabilistic characterizations of:

- Arrival pattern
- Service mechanism
  - When the service is available
  - How many customers can be served at a time
  - How long the service takes
- The queue-discipline
- The method by which a customer is selected for service

4.3 Birth-death process model

Consider the following examples:

- People being born and dying
- People arriving and leaving in a park

In these examples, the world and the park may be considered a queue; a baby's birth and a person's entrance into the park, an arrival; a person's death and a person's leaving the park, a service completion and departure; the size of the population of the world or the park, the queue length ***N***; and the amount of time a person spends in the world, i.e., age, or in the park, queuing delay ***d***, etc.

Under certain conditions of random arrivals and departures (i.e., services), e.g., Markov chain conditions, a queuing system can be modeled as a special class called the "birth-death" process. The birth-death process is used commonly in population studies, biology, etc. The birth-death process is considered to be most analytically tractable. An important class of queue referred to as the *M/M/1* queue is the "birth-death" process and will be discussed later.

4.4 Arrival rate

4.4.1 Definition

Figure 2-13 illustrates random arrivals. Consider random arrivals in the time interval from t_1 to t_2. The interval length is $T = t_2 - t_1$. The arrival rate is a long term average of the number of arrivals per unit time. Its mathematical symbol is λ and its unit is time^{-1}, and is given by the following equation, where *n* is the number of arrivals in the interval of length *T*.

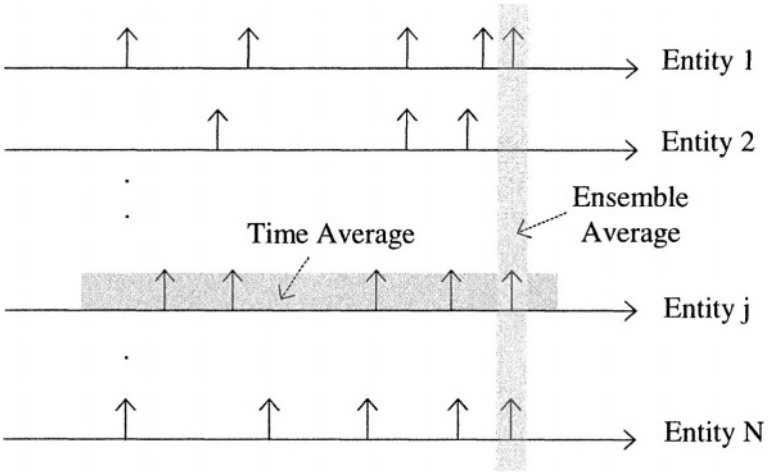

Figure 2-15. Time average and ensemble average.

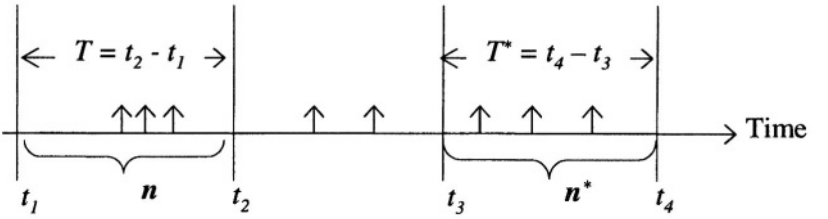

Figure 2-16. Illustration of stationarity.

$$\lambda = \frac{n}{T}$$

Example 2

The number of packets arriving at a packet switch during the one-minute period of from 9:00 *a.m.* to 9:01 *a.m.* has been counted over 30 days. The total count is 42 million packets. What is the packet arrival rate during this period?

Solution

$$\lambda = \frac{42 \times 10^6}{30 \times 1\,min} = 1.4 \times 10^6 / min$$

4.4.2 Empirical determination of arrival rate

There are two types of statistical averages:

- "Time average"
- "Ensemble average"

Figure 2-15 illustrates the two types of averages. The time average is the average taken over a period of time for the same physical entity. The ensemble average is the average taken over a number of physical entities (of the same kind) at a fixed time.

To illustrate these two types of averages, consider the toll gate example of Figure 2-9. Suppose that there are 10 lanes (or toll booths) at a toll plaza. Cars arrive at the toll plaza randomly and select a lane (or toll booth) randomly.

In this example, a time average of the arrival rate is obtained by selecting one toll booth, say Lane 1, counting the number of cars arriving at that

particular toll booth over a long period of time, say from *09:00 a.m.* to *10:00 a.m.,* and dividing the total count by the total measurement interval. This is a time average of arrivals per hour. Dividing this by 60, one can get a time average of arrivals per minute.

An ensemble average of the arrival rate is obtained by fixing a short interval, say one minute interval from a fixed time, say *09:00 a.m. – 09:01 a.m.,* counting the number of cars arriving at all 10 lanes, and dividing the total count by 10. This gives an ensemble average of arrivals per minute for 9:00 a.m.

It is important to understand the following two key concepts for measurement conditions:

- Stationarity
- Ergodicity

4.4.3 Stationarity

Section 3.6 defines Stationarity. To apply the concept of Stationarity to the measurements of arrival rate, consider two non-overlapping time intervals of equal length, T and T^*, shown in Figure 2-16. The random numbers of arrivals in these two intervals are denoted by $\boldsymbol{n}$ and $\boldsymbol{n}.^*$ If the arrival process is SSS, the two random variable $\boldsymbol{n}$ and $\boldsymbol{n}^*$ have the same statistics, e.g., same CDF and pdf. If the process is WSS, $\boldsymbol{n}$ and $\boldsymbol{n}^*$ may not have the same pdf, but they have the same mean: $\eta_n = \eta_{n^*}$.

In general, Stationarity would be harder to assume for a longer measurement interval. Examples of stationary arrivals may be car traffic during a one-hour rush hour period and telephone call traffic during a one-hour busy hour period. Examples of non-stationary arrivals may be traffic

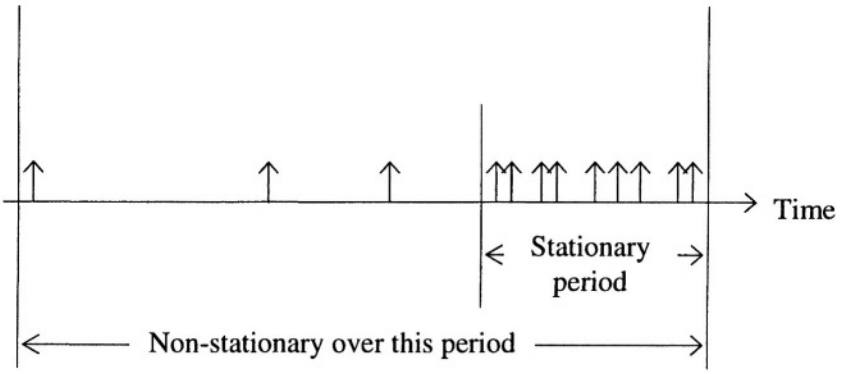

Figure 2-17. Example of stationary and non-stationary arrivals.

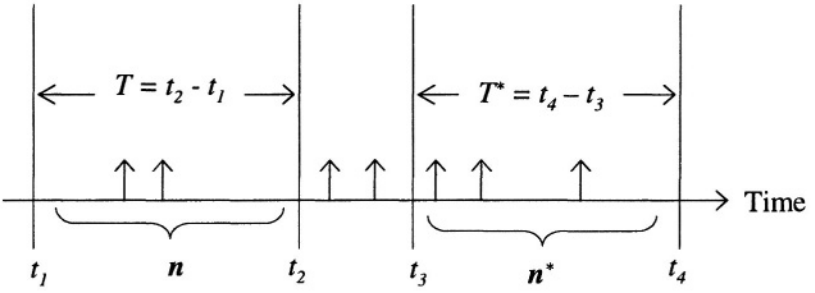

Figure 2-18. Poisson arrival.

over a 24-hour period and telephone call traffic over a 24 hour period. Figure 2-17 illustrates these examples.

4.4.4 Ergodicity

A complete definition and discussion of ergodicity are beyond the scope of this book. In this section, however, ergodicity is discussed for the mean or "average." A random process is "mean ergodic" if its time average is equal to its ensemble average. For example, to empirically determine the telephone call arrival rate λ for a central office in an area, consider the two types of averages determined as follows.

Suppose that the ensemble average is determined by taking the average across, say 10 randomly selected central offices in the area for a short fixed time interval, say, a busy hour from *12:00 p.m. – 01:00 p.m.* The time average is determined by taking the average at a single randomly selected central office, say office 7, over a longer time interval, say 24 hours. If mean ergodicity holds, the two methods should produce the same average. The ergodicity cannot hold for a non-stationary arrival process: stationarity is a necessary condition for ergodicity.

In the example above, if telephone traffic is non-stationary over the 24-hour period, it would be unreasonable to expect that the average over the 10 offices taken over a short one-hour interval and that taken over the 24-hour interval would be the same. The telephone office traffic engineering is based on busy hour statistics.

4.4.5 The Poisson Arrival

A Poisson arrival process is a random process in which the probabilities of the number of random arrivals in two non-overlapping intervals T and T^*,

$n(T)$ and $n(T^*)$, are independent. In a Poisson arrival process, the probability of future arrivals is not affected by previous arrivals.

A Poisson arrival process must satisfy the following conditions:

- The reservoir of arrivals is infinite.
- Stationarity.
- Ergodicity.

Poisson arrivals are often referred to as "pure random arrivals." A Poisson arrival process is an idealized mathematical model. The applicability of a Poisson model must be evaluated for each application under consideration for reasonableness.

The Poisson distribution is defined by a single parameter λ. Given λ, the probability of k arrivals in a time interval of length T is given as a function of T by the following equation:

$$P(k \text{ in } T) = e^{-\lambda T} \frac{(\lambda T)^k}{k!} \tag{2-100}$$

The unit of λ and T must be consistent in time. For example, if $\lambda = 120/\text{hour}$ in the above equation, T needs to be expressed in hours. If T is expressed in minutes, λ must be converted to 120/60 minutes, which yields 2/minute.

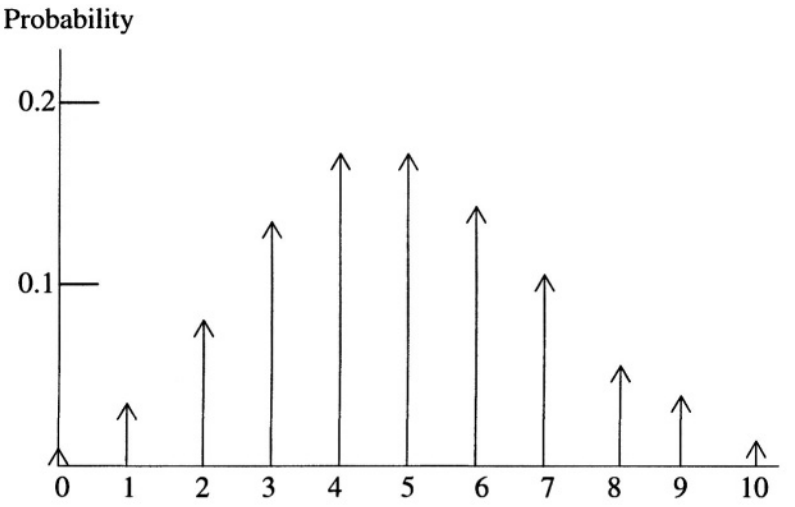

Figure 2-19. Poisson probability distribution.

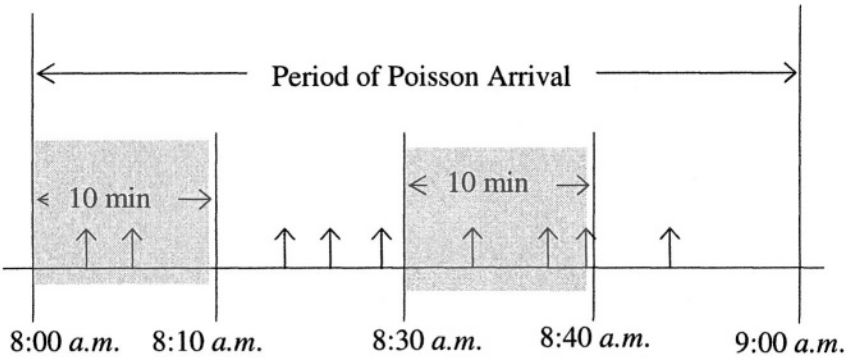

Figure 2-20. Poisson arrival.

Example 3

Consider packets arriving at a packet switch at the following arrival rate: $\lambda = 6 \times 10^6 / min$. Assuming a Poisson arrival, what is the probability that two packets will arrive in a 10-μsec interval?

Solution

Use the Poisson pdf with the following values:

$$\lambda = 6 \times 10^6 \ /min = 1 \times 10^5 \ /sec = 0.1 / \mu \ sec . \qquad T = 10 \ \mu sec$$

Hence $\lambda T = 0.1 \times 10 = 1$. For $k = 3$, the Poisson pdf yields

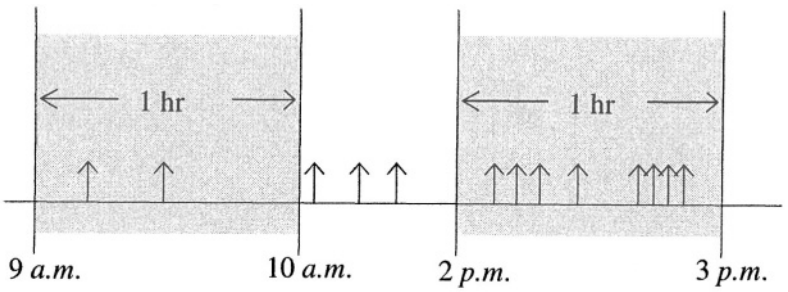

Figure 2-21. Non-Poisson arrival.

$$P\{3 \text{ in } 0.1\ \mu sec\} = e^{-1}\frac{(1)^3}{3!} = 0.06$$

Suppose that the arrival pattern is Poisson for the time interval from *8:00 a.m. – 9:00 a.m.* Under this assumption, probability distributions over the 10 minute period, e.g., from *08:00 - 08:10 a.m.* and that between *08:30 – 08:40 a.m.* are assumed to be independent and identically distributed (i.i.d). This is illustrated in Figure 2-20. On the other hand, a 10-minute interval taken from 8 – 9 *a.m.* and a 10-minute interval taken from, say, 2 – 3 *p.m.* may not have the same arrival characteristics, and stationarity would not hold over this stretched time period. In this case, the Poisson model would not apply. This is illustrated in Figure 2-21.

Consider the following two examples:

- People arriving at a bus station in a large city with millions of people
- People arriving at a bus station in a small community (with a few people)

Consider the significance of the difference between the "large" city and the "small" community. In the first case, the Poisson model may be applied because there is an infinite reservoir of arrivals. In the latter case, the Poisson model would not apply. For example, suppose that there are only five people in the community who take the bus. If five people have already arrived in T^*, the probability of an arrival in T is obviously zero, i.e., n and n^* are not independent. This violates one of the Poisson assumptions that the reservoir of arrivals is infinite.

Figure 2-22 illustrates inter-arrival times. The inter-arrival time, t, is a random variable and is defined as the time between two consecutive arrivals.

The inter-arrival times of Poisson arrivals are exponentially distributed with the following CDF and pdf, where λ is the arrival rate:

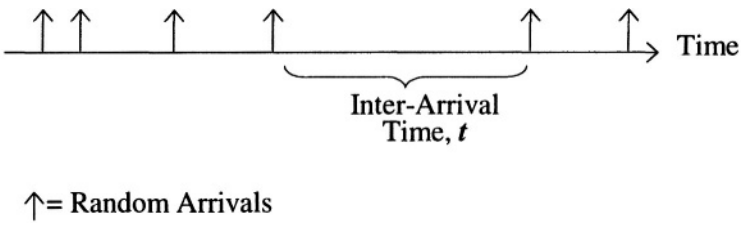

Figure 2-22. Inter-arrival time

$$CDF \quad F(t) = 1 - e^{-\lambda t} \qquad pdf \quad f(t) = \lambda e^{-\lambda t} \qquad (2\text{-}101)$$

By taking the expected value of t, the mean of the inter-arrival time is found to be $1/\lambda$ as follows:

$$\eta_t = E\{t\} = \int_{-\infty}^{+\infty} t f_t(t)dt = \int_{-\infty}^{+\infty} t(\lambda e^{-\lambda t})dt = \frac{1}{\lambda} \qquad (2\text{-}102)$$

Example 4

Assume a Poisson arrival with the arrival rate $\lambda = 10/\mu sec$. Find the probability that the inter-arrival time between consecutive arrivals is greater than $0.1\ \mu sec$.

Solution

$$P\{t > 0.1\ \mu sec\} = 1 - P\{t \le 0.1\ \mu sec\} = 1 - F(0.1)$$

$$= 1 - (1 - e^{-10 \times 0.1}) = e^{-1} = 0.37$$

4.4.6 Markov Modulated Poisson Process (MMPP)

The MMPP process $x(t)$ is a non-stationary process, which is composed of n separate Poisson processes, PP_i, $i = 1, \ldots, n$. While the process $x(t)$ is in "*state i*," $x(t)$ is Poisson process PP_i, with parameter λ_i. The process $x(t)$ moves or "makes transitions," between states at discrete points in time, t_m's. The state transitions are assumed to follow the Markov chain model. In a Markov chain, given that the process is in *state i* at time t_m, the probability that the process will transition into *state j* at the next time instant, t_{m+1}, is referred to as the transition probability p_{ij}. For an n-state MMPP, there are $n \times n$ transition probabilities, including the probability of staying in the current state, p_{ii}.

The $n \times n$ matrix of the transition probabilities is referred to as the transition probability matrix, Π, as follows:

$$\Pi = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix} \quad (2\text{-}103)$$

where the transition probability from *state i* to *state j*, p_{ij}, is the conditional probability defined as:

$$p_{ij} = P\{ x(t_{m+1}) \text{ in State } j \mid x(t_m) \text{ in State } i\} \quad (2\text{-}104)$$

Figure 2-23 shows an example of three-state MMPP. Figure 2-24 shows the three-state MMPP making transitions over time.

4.5 Service rate

The service rate is defined as the number of customers served in unit time. Its mathematical symbol is μ. Its unit is 1/time, i.e. time^{-1}.

$$\mu = \frac{m}{T} \quad (2\text{-}105)$$

where m is the number of customers served in the interval of length T.

The inverse of service rate, $1/\mu$, is the service time, which is the time expended to serve one customer. Assuming that the customers leave instantly after getting service, the departure rate is equal to the service rate.

Example 5

Referring to Figure 2-25, a packet switch has a single incoming port and five processors. Incoming packets can be routed to any idle processor. Each processor can serve on the average 1×10^6 packets per minute. What is the average service rate of the packet switch processor operation?

Solution

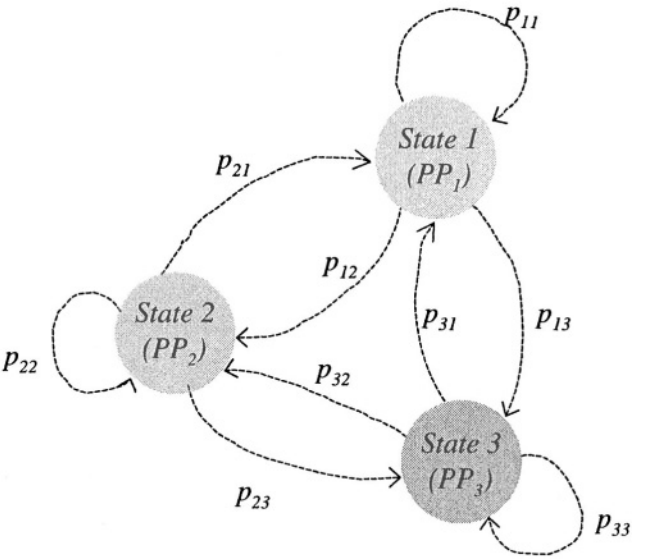

Figure 2-23. Three state Markov Modulated Poisson Process (MMPP).

$$\mu = \frac{1 \times 10^6 \times 5}{1\,min} = 5 \times 10^6 / min$$

Example 6

New processors are to be added to increase the service rate of the operation to $6x10^6$/min. Assume that the new processor can serve $0.5x10^6$ packets per minute. How many new processors need to be added?

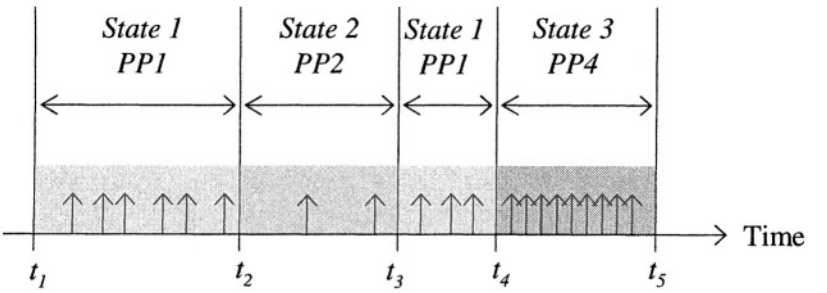

Figure 2-24. An example of MMPP in time.

Solution

Let the unknown number of new processors be *x*.

$$\mu = [(1 \times 10^6 \times 5) + (0.5 \times 10^6 \times x)]/_{min} = 6 \times 10^6/min$$

Solving for *x*, $x = 2$.

4.6 Utilization factor

Utilization factor is a measure of how fully the resource is used to meet the customer need. It is defined as the ratio of arrival rate to service rate. Its mathematical symbol is ρ, and ρ is dimensionless.

$$\rho = \frac{\lambda}{\mu} \tag{2-106}$$

In order for the queue to be stable, the service station should be able to serve customers at a faster rate than the customer arrival rate:

$$\mu \succ \lambda \tag{2-107}$$

In other words, in order for the queue to be stable, the utilization factor ρ should be less than one, i.e., less than 100% utilization of resources:

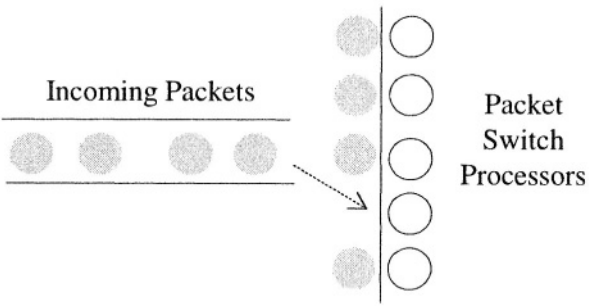

Figure 2-25. Example 5.

$$\rho = \frac{\lambda}{\mu} < 1 \tag{2-108}$$

In fact, the lower the utilization factor, the better the service to the customers. The lower the utilization factor, the shorter the queue length and the shorter the service delay (i.e., customer waiting).

Example 7

Consider a packet switch with the service rate $\mu = 5 \times 10^6$ *packets/min*. Assume that packets arrive at the packet switch at an arrival rate of $\lambda = 5 \times 10^4$ *packets/sec*. What is the utilization factor of the operation?

Solution

$$\rho = \frac{\lambda}{\mu} = \frac{5 \times 10^4 \times 60}{5 \times 10^6} = 0.6$$

4.7 Queuing system performance metrics

The metrics used for queuing system performance include:

- Queue length, *N*
- Delay or "waiting time" *d*

Note that both *N* and *d* are RV's.

4.7.1 Little's Theorem

A useful theorem on the relationship between the average queue length and the average queue delay is given by the Little's theorem as follows:

$$\eta_N = \lambda \times \eta_d \tag{2-109}$$

where λ is the arrival rate, $\eta_N = E\{N\}$ = *mean queue length* and $\eta_d = E\{d\}$ = *mean delay.*

A rigorous mathematical proof of the Little's theorem is beyond the scope of this book. In fact, only recently, this theorem was proven mathematically for arbitrary arrival and service time distributions.

A heuristic proof of the Little's theorem is illustrated in Figure 2-26. In the figure, suppose that Customer A arrives at the tail of queue (ToQ) at time

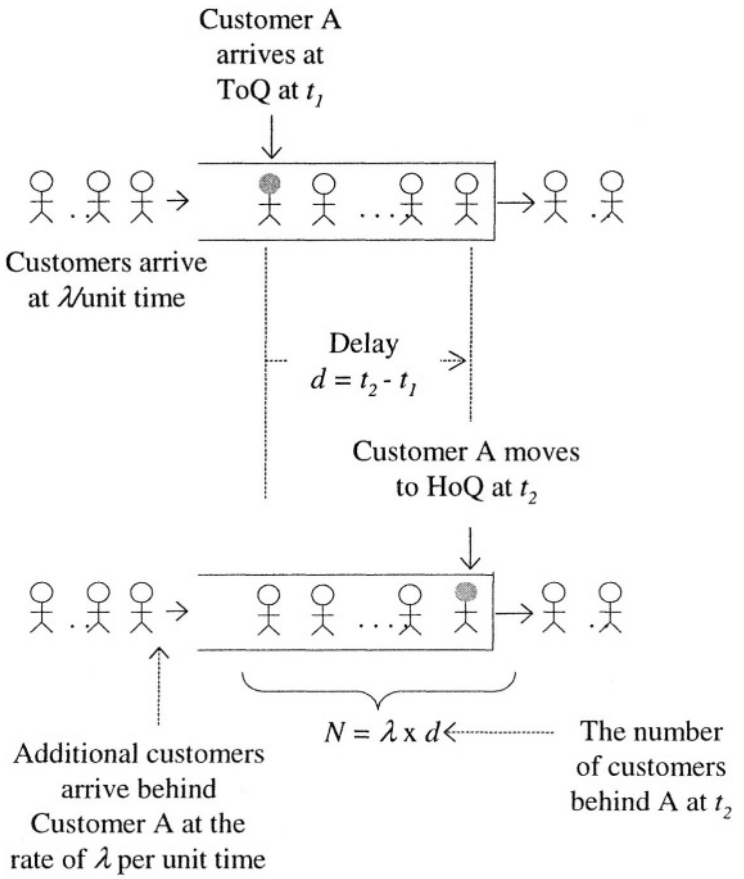

Figure 2-26. Heuristic proof of the Little's theorem.

t_1. Customer A moves to the head of the queue (HoQ) at time t_2. The amount of time Customer A spends in the queue is $d = t_2 - t_1$.

If customers behind Customer A arrive at the ToQ at the rate of λ, if Customer A turns around at the HoQ and counts the customers standing behind, he would see on the average λ x η_d customers. That is equal to the average queue length η_N.

4.8 *M/M/1* queue

The *M/M/1* queue is a single server queue with Poisson arrivals with an arrival rate λ and an exponential service time of a service rate μ. The CDF and pdf of the service time t are given below:

$$CDF \quad F(t) = 1 - e^{-\mu t} \qquad pdf \quad f(t) = \mu e^{-\mu t} \tag{2-110}$$

The *M/M/1* queue satisfies the "birth-death" process and is analytically tractable. The *M/M/1* queue is considered an idealized queuing model.

The probability that there will be k customers in the queue in a steady state is given by the following equation:

$$p_k = (1-\rho)\rho^k \qquad k = 0, 1, 2, 3, \ldots \tag{2-111}$$

where

$$p_k = \text{probability of } k \text{ customers in the queue} = P\{N = k\} \tag{2-112}$$

$$\rho = \frac{\lambda}{\mu}; \qquad \mu \succ \lambda .$$

Table 2-1 tabulates p_k given by the above equation. Observe that p_k depends on λ and μ only through their ratio ρ, i.e., λ and μ are not independent variables for p_k.

Setting $k = 0$ in the above equation, the probability that the queue will be empty is given by:

$$p_0 = (1-\rho)\rho^0 = 1-\rho. \tag{2-113}$$

The probability that there will be at least k customers in the queue is given by

$$P\{N \geq k\} = \rho^k \qquad k = 0, 1, 2, 3, \ldots$$

Table 2-1. Probability of k customers in the queue as a function of ρ

ρ	0.5	0.6	0.7	0.8	0.9
k					
0	0.50	0.40	0.30	0.20	0.10
1	0.25	0.20	0.21	0.16	0.09
2	0.13	0.10	0.15	0.13	0.08
3	0.06	0.10	0.10	0.10	0.07
4	0.03	0.10	0.07	0.08	0.07
5	0.02	0.00	0.05	0.07	0.06

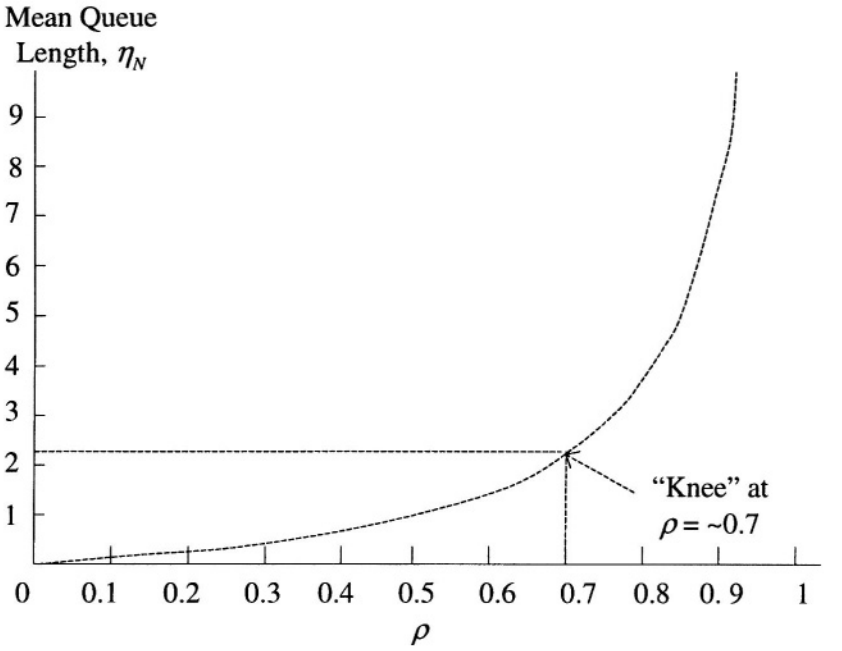

Figure 2-27. Mean delay,N, as a function of utilization factor, ρ

Since $P\{N \geq k\} + P\{N \prec k\} = 1$, the probability that there will be less than k customers in the queue is given by:

$$P\{N \prec k\} = 1 - \rho^k \qquad k = 0, 1, 2, 3, \ldots \tag{2-114}$$

The mean of the queue length, N, is given by:

$$\eta_N = \frac{\rho}{1-\rho}, \qquad \rho \prec 1 \tag{2-115}$$

For $\rho \geq 1$, $\eta_N \rightarrow \infty$. The variance of the queue length, N, is given by:

$$\sigma_N^2 = \frac{\rho}{(1-\rho)^2}, \qquad \rho \prec 1 \tag{2-116}$$

$$\sigma_N^2 = \frac{\rho}{(1-\rho)^2}, \qquad \rho < 1 \tag{2-116}$$

Observe that the variance of the queue length depends on λ and μ only through their ratio ρ, i.e., λ and μ are not independent variables.

Figure 2-27 plots the mean queue length, η_N, as a function of the utilization factor ρ. The mean queue length monotonically increases as the utilization factor increases, and goes to infinity at $\rho = 1$. The curve shows that, as ρ passes about 0.7, the mean queue length suddenly increases rapidly. There is a "knee" of the curve at about $\rho = 0.7$. A smart operator would be able to optimize the performance-to-cost tradeoff by recognizing the knee. For example, if the current operating point is slightly above the knee, say $\rho = 0.8$, the performance can be dramatically improved by bringing down ρ below 0.7 by adding a little more resources.

For example, suppose that a packet switch currently operates at a utilization factor $\rho = 0.9$. At this value of ρ, the average queue length would be nine packets. By decreasing ρ from 0.9 to 0.7, the average queue length can be dramatically cut down to 2.3, which will reduce the packet loss ratio. To improve the packet switch performance this way, more processors need to be added to reduce the utilization factor ρ.

To find the mean delay, η_d, consider the Little's theorem. Combining the Little's theorem of Equation (2-109) and Equation (2-115) yields:

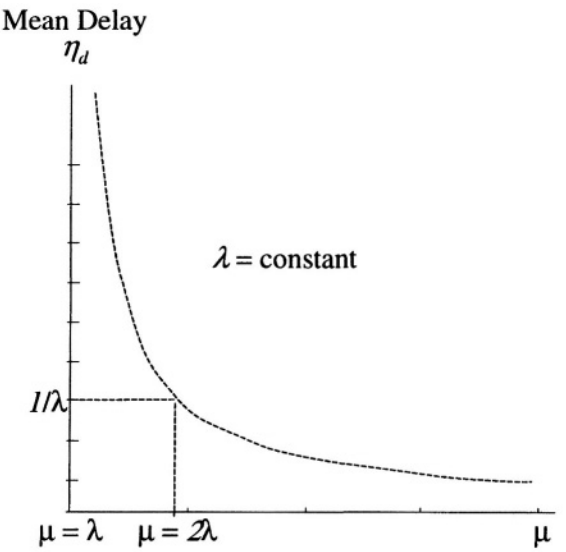

Figure 2-28. Mean delay, η_d, as a function of service rate, μ, for a fixed λ.

$$\eta_d = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda} \qquad (2\text{-}117)$$

Figure 2-28 plots the mean delay, η_d, as a function of μ for a constant value of λ. The figure shows that the mean delay is infinite if μ is equal to λ. The mean delay is $1/\lambda$ at μ equal to 2λ.

5. EXERCISES

5.1 Problems

1. A player is to throw a die, and, if the face with an even number of spots shows up, the player wins the prize. What is the probability of winning the prize? Formulate the problem and solve it using the axiomatic approach.

2. The number of switched virtual circuit requests arriving at an ATM switch during the period of from *9:00 a.m.* to *10:00 a.m.* has been counted over *30* days. The total count is 900 requests. What is the arrival rate during the *9:00 - 10:00 a.m.* period?

3. ATM cells arrive at an ATM switch at the rate of λ = 600 cells/minute. Assuming that the arrivals are Poisson, what is the probability that two cells will arrive in a 0.1-second interval? (Use the Poisson distribution; use $e \cong 2.72$; $e^{-1} \cong 0.37$.)

4. Packets arrive at a packet switch at the rate of 10×10^6 packets/sec from 8:00 *a.m.* to 9:00 *a.m.* and 5×10^6 packets/sec from 9:00 *a.m.* to 10:00 *a.m.* Would it be reasonable to assume that the arrival pattern over the 8:00 *a.m.*–10 *a.m.* period is Poisson and why?

5. Is the following statement true or false?

   - An arrival pattern of passengers at a train station over a certain period has been found to be non-stationary and ergodic.
   - A Poisson process is stationary and ergodic.

6. A packet switch has a single input port and five processors. Packets go to any available processor. Each processor can process 1.2×10^6 packets per second. What is the service rate of the packet switch operation?

7. Cars arrive at a toll booth at an average rate of five cars per minute. The cashier at a toll booth serves cars at the rate of 600 cars per hour. What is the utilization factor of the toll booth operation? What percentage of the time would the cashier be idle?

8. For an $M/M/1$ queue with $\rho = 0.7$, find the probability that the queue length will be between zero and two inclusively, i.e. $P\{0 \leq N \leq 2\}$; and the probability that the queue will be empty.

9. For an $M/M/1$ queue of $\rho = 0.4$ and $\lambda = 10/\text{hour}$, determine the following:
   - mean queue length
   - mean delay through the queue
   - service rate

   Suppose that the arrival rate doubles but the queue operates with the same service rate. Determine the delay.

5.2 Solutions

1. Die throwing experiment

$$S = \{\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6\}$$

$$p_i = P(\xi_i) = 1/6; \qquad i = 1, \ldots, 6$$

The event of interest is "winning the grand prize" and is defined as a set denoted by W as follows:

$$W = \text{"even number of spots"} = \{\xi_2, \xi_4, \xi_6\}$$

Since $\{\xi_2\}$, $\{\xi_4\}$, and $\{\xi_6\}$ are mutually exclusive, i.e., $\{\xi_2\} \cap \{\xi_4\} = \{\phi\}$, $\{\xi_4\} \cap \{\xi_6\} = \{\phi\}$, and $\{\xi_2\} \cap \{\xi_6\} = \{\phi\}$, it follows that $D = \{\xi_2\} \cup \{\xi_4\} \cup \{\xi_6\}$. From Axiom III, it follows that:

$$P(W) = P(\{\xi_4, \xi_5, \xi_6\}) = P(\{\xi_4\} \cup \{\xi_5\} \cup \{\xi_6\})$$

$$= P([\{\xi_4\} \cup \{\xi_5\}] \cup \{\xi_6\}) = P([\{\xi_4\} \cup \{\xi_5\}]) + P(\{\xi_6\})$$

$$= P(\{\xi_4\}) + P\{\xi_5\} + P\{\xi_6\} = 1/6 \times 3 = 0.5$$

2. 900/30 = 30/hr.

3. $k = 2$; $T = 0.1$ sec.; $\lambda = 600/\text{min} = 600/60 = 10/\text{sec}$.
$\lambda T = 10 \times 0.1 = 1$
$P\{2 \text{ in } 0.1 \text{ sec}\} = (e^{-1})[(1)^2/(2!)] = (0.37)/[(2)(1)] = 0.185$.

4. No, because non-stationary.

5. False; True.

6 $\mu = 1.2 \times 10^6 \times 5/\text{sec} = 6 \times 10^6/\text{sec}$.

7.

$\lambda = 5/\text{min}$; $\mu = 600/\text{hr} = 600/60 \text{ min} = 10/\text{min}$; $\rho = \lambda/\mu = 5/10 = 0.5$. 50% idle.

8.
From the table, $P\{0 \le N \le 2\} = P\{0\} + P\{1\} + P\{2\} = 0.3 + 0.21 + 0.15 = 0.66$.; $P\{\text{empty}\} = P\{0\} = 0.3$.

9. Mean queue length

$$\eta_N = \frac{\rho}{1-\rho} = \frac{0.4}{1-0.4} = (0.4/0.6) = 0.67$$

$$\text{Mean delay } \eta_D = \frac{N}{\lambda} = \frac{0.4/0.6}{10/60} = (0.4/0.6) \times (60/10) = 4 \text{ min}$$

$$\text{Service rate: } \textit{Given } \rho = \frac{\lambda}{\mu} = 0.4; \quad \textit{Hence } \mu = \frac{\lambda}{0.4} = \frac{10}{0.4} = 25/\text{hr}$$

Same service rate : $\mu' = \mu$; *double arrival rate :* $\lambda' = 2\lambda$

Delay: $\eta_{D'} = \frac{1}{\mu' - \lambda'} = \frac{1}{\mu - 2\lambda} = \frac{1}{25 - 20} = \frac{1}{5} hr = 12 min$

http://www.springer.com/978-0-387-23389-5

QoS in Packet Networks

Park, K.I.

2005, XIII, 243 p. 164 illus., Hardcover
ISBN: 978-0-387-23389-5