CURRENT TRENDS IN PROGRAMMING METHODOLOGY

VOLUME III

Software Modeling

K. MANI CHANDY and RAYMOND T. YEH, *Editors*

CURRENT TRENDS
IN
PROGRAMMING METHODOLOGY

VOLUME III

Software Modeling

K. MANI CHANDY and RAYMOND T. YEH, *Editors*

PRENTICE-HALL, INC., Englewood Cliffs, New Jersey 07632

Library of Congress Cataloging in Publication Data

Main entry under title:

Current trends in programming methodology.

Includes bibliographies.
Includes indexes.
CONTENTS: v. 1. Software specifications and design.
—v. 3. Software modeling.
1. Programming (Electronic computers) I. Yeh,
Raymond Tzuu-Yau (date)
QA76.6.C87 001.6'42 76-46467
ISBN 0-13-195727-9 (v. 3)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PRENTICE-HALL INTERNATIONAL INC., *London*
PRENTICE-HALL OF AUSTRALIA PTY. LIMITED, *Sydney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*
PRENTICE-HALL OF SOUTHEAST ASIA PTE. LTD., *Singapore*
WHITEHALL BOOKS LIMITED, *Wellington, New Zealand*

PREFACE

Performance is an important aspect of programs. Programs serve specific functions, such as computing payrolls or guiding space ships, and these functions have no value if the service is either too slow or too unreliable. We recognize that, in general, it is important to study the performance of the total system (such as spaceships) in which programs play only a part in mission accomplishment. However, in this book our emphasis is on studying the performance of computing systems, and the interface with the larger system is represented by a description of the load offered to the computing system.

Designers of computing systems must have techniques which help them predict system performance because it is too expensive to build a system and later decide that its performance is unsatisfactory. Values have to be selected for a very large number of parameters in any computing system design and system designers need techniques to help them search the large parameter space. This book is concerned with *models* to help designers search large parameter spaces rapidly and also to predict system performance.

Many computer scientists are now aware that performance modeling can have significant practical value in their work. However, a significant portion of computer scientists have only a vague notion of modeling techniques and often presume (incorrectly) that these techniques are very complex and extremely hard to understand. This book reinforces the idea that performance modeling has practical utility and dispels the notion that the modeling area is complex. Professional computer scientists, as well as professors and students will find this book helpful in getting a clear introduction to the use of models in computing system performance analysis. A wide range of models is covered and the relevance of the models is emphasized. Models are discussed in sufficient depth that the reader can immediately grasp key issues and apply the models to his work.

Models such as those discussed in this book have been used for some time in diverse disciplines, though their utility in computing systems design has not been apparent until recently. The general area of modeling has been called Operations Research. Recently, a significant number of operations researchers have become interested in applications of their discipline to computer system design and tuning. There also seems to be increasing agreement among operations researchers that for models to have significant value it is at least as important to gain expertise in the systems being modeled as it is to have expertise in the theory of modeling. This book will prove valuable to operations researchers, management scientists, statisticians, systems engineers, and applied mathematicians who are interested in applying their techniques to analyze computer systems performance. This book shows how different types of models can be usefully applied to computing systems analysis.

Chapter 1 introduces the reader to statistical analysis of performance data at a very basic level. Statistical analysis is the key to performance modeling because without experimental data there cannot be good models and analysis is required to understand large volumes of raw data. This chapter contains clear descriptions of recent advances in the statistical analysis of performance data. Many breakthroughs in this area were discovered by the authors of this chapter.

Discrete-event simulation methods have been used to model systems for a very long time. The question that is most often asked relating to simulation is: How does one know whether a simulation has been run long enough to have a high degree of confidence in simulation results? Even though discrete-event simulation is the most widely used performance modeling tool, very few computer scientists know how probability theory can be used in answering the crucial question regarding confidence in simulation results. Chapter 2 uses the theory of regenerative processes to address this and other related questions. The author of this chapter is a pioneer in this field.

Simulations suffer from the fact that they may require an enormous amount of computation time to obtain meaningful results. Analytic methods based on queueing theory generally require much less computation. It is sometimes helpful for a designer to use an interactive analysis package which rapidly provides estimates of performance metrics. Simulations are generally unsatisfactory in such an interactive environment. Chapter 3 discusses recent results in congestion models of computing systems. Models based on networks of queues have been particularly useful in computing systems analysis. The author of Chapter 3 is a leading figure in the area of queueing network models.

It has been recognized for some time that a key problem in utilizing modeling theory is the paucity of computer scientists with experience in modeling (and operations researchers with experience in computer sciences). Another key problem is that there is generally relatively little time to develop and analyze models; people who demand performance estimates usually should have requested the estimates several months earlier. It is therefore crucial that performance modelers use as many standard, "canned" analysis programming packages as possible. It is also extremely important that modelers have formal languages to describe models to each other and to analysis packages. Chapter 4 discusses a versatile system, called RESQ, which addresses both these problems. There is enough detail in this chapter to help the user build his own versatile analysis package. RESQ is the best analysis package in existence today. The authors are leaders in language development for modeling as well as in queueing models and simulation.

Graph models are widely used in computing systems analysis. Graphs are used in the analysis of programs, recognition of parallelism, scheduling resources, and a variety of other areas. Chapter 5 starts with basic concepts and then discusses recent advances in this area. The reader will obtain a clear understanding of graph models, a thorough knowledge of the application of these models in certain areas of computer sciences, and an appreciation of the value of graph models in all areas of computer sciences. Expository papers by the author of this chapter are among the most widely quoted in this area.

Graph theory is a part of the discipline of combinatories. Chapter 6 discusses applications of combinatories which are not emphasized in the previous chapter. Combinatoric models are used in the analysis of algorithms, resource scheduling, systems design, and in designing communication network topology. Recent developments in complexity theory, particularly the notion of NP-completeness, are of great importance to the professional computer scientist as well as to theoreticians. Computer professionals often refer to some problems as being *tractable* and to others as being *intractable* or *hard*. They develop *optimal* or *exact* algorithms to solve certain problems and *heuristics* to solve others; the former type of algorithm is preferred but the latter type may be used for hard problems. It is vitally important for computer professionals to be aware of the precise definitions of these terms; it is also important that they understand the complexity of the problems that they are dealing with so that they select problem-solving methods most likely to yield satisfactory solutions. Chapter 6 introduces the reader to these important concepts. The authors have made substantial contributions in this area and also have a remarkable capacity for clear exposition.

The problem of scheduling resources is considered in Chapters 5 and 6. However, there is an important aspect of resource scheduling, *deadlocks*, which deserves an entire chapter to itself. Deadlocks arise in concurrent systems when no process can proceed any further because they cannot obtain the resources they require for further processing. The key issues are those of recognizing states which may lead to deadlock and managing resources to avoid (or reduce the possibility of) deadlock. Though resource management has been studied in several branches of engineering and business, the problem of deadlocks has received the most attention in computer sciences. Chapter 7, written by the leader in this field, introduces the reader to the problem of deadlocks with the aid of clear examples and illustrations. This chapter will prove useful to computer scientists and to people concerned with resource management in general.

The performance of computing systems depends in large part upon the effective management of memory. It is crucial that program behavior be understood and that memory be allocated to those programs which are most likely to contribute to effective system performance. Chapter 8 is a comprehensive, lucid discussion of the problem of optimally managing memory. It is written by a pioneer in operating systems theory, who also identified the *working set* principle, which is the key concept in understanding program behavior with reference to memory.

Chapter 9 is an introduction to mathematical programming. No prior acquaintance with mathematical programming is required to understand this chapter though a basic knowledge of linear algebra is assumed. The goal of this chapter is to provide the reader who is unfamiliar with this area with a clear, thorough understanding of the key concepts of optimization theory. This chapter is designed to help the reader (a) to identify situations in which mathematical programming techniques will help in computer systems design problems, (b) formulate systems design problems as mathematical programs, and (c) communicate with operations researchers. This chapter emphasizes intuitive understanding at the expense of rigor. The chapter relies on illustrations and examples in elucidating key concepts.

This book could be used as a text in a beginning graduate course in computer systems modeling. A familiarity with introductory probability theory and simulation is assumed. The text should be supplemented with descriptions of case studies and measurement techniques found in the proceedings of recent workshops such as those sponsored by IFIP working group 7.3. on computer systems performance modeling, measurement and evaluation and ACM SIGMETRICS. Case studies should be selected depending upon the specific interests of students; for instance, case studies can be drawn from computer networks, data-base systems, or multiprogrammed operating systems. The modeling issues descussed in this text apply to a variety of cases. This book, supplemented with case study material, is planned to be used in first-year graduate courses in the Computer Sciences Department, University of Texas at Austin.

K. MANI CHANDY
RAYMOND T. YEH

Austin, Texas

CONTENTS

3

4

5

CHAPTER 1

STATISTICAL METHODS IN COMPUTER PERFORMANCE ANALYSIS

Y. BARD

M. SCHATZOFF
IBM Cambridge Scientific Center
Cambridge, Massachusetts

1.1. INTRODUCTION

Application of statistical methodology to the study of computer system performance seems quite natural in light of our ability to generate very large volumes of data on many variables, whether by hardware or software instrumentation, and our need to understand what such data have to say about the system from which they are derived. Statistics provides tools for summarizing the salient features of data, searching for meaningful patterns, detecting instances of anomalous behavior, understanding complex relationships among many variables, and designing efficient experiments for determining the effects of factors that can be controlled.

The types of data generally collected for computer systems usually include utilizations of various system resources such as the central processing unit, I/O devices, and channels; counts of events such as various classes of program interrupts; and tracing of program paths initiated by specified events. The level of detail of data collection depends on the purposes for which the data are to be used. At one level, the purpose might be to provide overall descriptions of system, subsystem, or program behavior. At another level, the objective might be to answer specific questions pertaining to certain aspects of these behavioral phenomena. Or, we might wish to predict, control, or improve system performance under stipulated conditions.

Statistics provides a useful methodology for approaching problems at all of these levels of inquiry. Because of the stochastic nature of events occurring in the operation of a computer system, it is natural to think of describing the resulting data in terms of probability distributions. Such descriptions entail questions of fitting distributions to data, testing goodness of fit, estimating relevant parameters of fitted distributions, and statistical testing of hypotheses concerning such parameters.

In Section 1.2, we shall outline some of the basic notions of probability and statistical inference to be used in subsequent sections. Techniques such as regression and correlation analysis, which apply to the study of the joint distributions of two or more variables, are treated in Section 1.3. Concepts of design and analysis of experiments, pertaining to situations in which one or more factors can be controlled by the investigator, are dealt with in Section 1.4. Finally, in Section 1.5, we shall examine the applicability of design of experiments to the problem of optimizing or tuning computer systems.

As a prelude to the ensuing sections, we shall introduce the idea of a data matrix, in which each row contains observations on a number of variables, taken at a given time. Successive rows then represent observations taken at successive points in time. Determination of the appropriate lengths of sampling intervals is itself a question of some importance, for which there is no definitive answer. The choice depends on factors such as impact of the measurements on system performance, activity level of the system, degree of time dependence in the data, and desired accuracy of results. Exploratory data analysis can usually provide reasonable guidelines as to an appropriate choice of sampling interval.

1.2. STATISTICAL PRELIMINARIES

1.2.1. Random Variables and Probability Distributions

The data for a single variable, as recorded in a particular column of a data matrix, may be viewed as a sample drawn from a hypothetical population consisting of all possible values that the variable might assume. If the data generating mechanism were allowed to run indefinitely, we could observe the frequencies with which observations fall in various intervals. Thus, the characteristics of the hypothetical population could be described in terms of a probability distribution, from which one could calculate the proportion of all possible observations lying within any specified interval. Usually, we are unable to observe the entire population and must confine our inferences about the population to observations taken on a sample drawn from it. Procedures for selecting samples are the subject of sampling theory, whereas methods for making statements about the nature of the underlying population derive from principles of statistical inference.

Let us denote a random variable by X and a possible value of the variable by x. The distribution function $F(x)$ for the random variable X is defined by the following relation: $F(x) = \text{prob}\,(X \leq x) = \int_{-\infty}^{x} dF(y)$. With this definition in hand we can readily express the probability that X lies within a prescribed interval (a, b) by $\text{prob}\,(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} dF(x)$. If $F(x)$ is a continuous function, then

$dF(x) = f(x)\,dx$, and $f(x)$ is said to be the density function of x. In the case of a discrete-valued random variable, the foregoing Stieltjes integral reduces to a summation. For example, if X is defined only on the nonnegative integers, then $F(x) = \sum_{y=0}^{x} f(y)$, where $f(x)$, the probability that $X = x$, is called the frequency function of x.

The mathematical expectation or mean of a random variable is defined by $\mu = E(X) = \int_{-\infty}^{\infty} x\,dF(x)$. This is the center of gravity of the distribution, since $E(X - E(X)) = 0$. The variance or second central moment of a random variable is defined by $V(X) = E(X - \mu)^2 = EX^2 - \mu^2$. It provides a measure of the dispersion, or spread of the distribution. In general, the kth central moment of a distribution is defined by $\mu_k = E(X - \mu)^k$. It is straightforward to verify that the kth moment of CX, where C is a constant, is equal to $C^k\mu_k$ and that $E(X - \theta)^2$ is minimized when $\theta = \mu$.

Let prob (A) be the probability of the event A. Two random variables X and Y are said to be statistically independent if prob $[(X \in I)$ and $(Y \in J)] =$ prob $(X \in I) \times$ prob $(Y \in J)$ for all intervals I and J. Let A be the event $X \in I$ and B the event $Y \in J$. Then the conditional probability of the event B occurring, given that A has occurred, is defined as prob $(B|A) =$ [prob $(A$ and $B)] \div$ prob (A). Thus, if X and Y are statistically independent, prob $(B|A) =$ prob (B).

Two events A and B are said to be mutually exclusive if prob $(A$ or $B) =$ prob (A) + prob (B). In general, prob $(A$ or $B) =$ prob (A) + prob (B) − prob $(A$ and $B)$. When A and B are mutually exclusive, prob $(A$ and $B) = 0$.

Example 1

Suppose the interarrival times of messages arriving at the CPU of a data processing system are exponentially distributed with density function $f(t) = \lambda e^{-\lambda t}$, $t > 0$. Then the mean and variance of t are given, respectively, by $\mu = \int_0^\infty t\lambda e^{-\lambda t}\,dt = 1/\lambda$ and $\sigma^2 = \int_0^\infty t^2\lambda e^{-\lambda t}\,dt - (1/\lambda)^2 = 1/\lambda^2$.

Example 2

When interarrival times are exponentially distributed, it can be shown that the number of arrivals x in any interval of length t has a Poisson distribution, defined by the probability mass function

$$f(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \qquad x = 0, 1, \ldots$$

The mean and variance of the Poisson distribution are both equal to λt.

The two preceding distributions, which are very important in the study of queueing problems, are derived from the following two postulates:

1. The distribution of the number of occurrences depends only on the length of the interval and not on its position in time.
2. The number of occurrences in different intervals are statistically independent.

For any two random variables X and Y, $E(X+Y)=E(X)+E(Y)$. The covariance of X and Y is defined by

$$\text{cov}(X, Y) = E(X-\mu_X)(Y-\mu_Y)$$

where μ_X and μ_Y are the expectations of X and Y, respectively. If X and Y are statistically independent, then

$$\text{cov}(X, Y) = E(X-\mu_X)\cdot E(Y-\mu_Y) = 0$$

and

$$V(X+Y) = V(X) + V(Y)$$

These properties extend to any number of mutually independent random variables $X_1, \ldots, X_n$, so that

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

and

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 V(X_i)$$

When the X_i are not mutually independent, we have

$$\text{cov}\left[\left(\sum_{i=1}^{n} a_i X_i\right), \left(\sum_{j=1}^{n} b_j X_j\right)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \,\text{cov}(X_i, X_j)$$

The correlation coefficient for two random variables X and Y is defined by

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{[V(X)\cdot V(Y)]^{1/2}}$$

Given a set of p random variables $X_1, \ldots, X_p$, the covariance matrix is defined as a $p \times p$ matrix whose (i, j) element is given by cov (X_i, X_j). The correlation matrix is defined in an analogous manner.

1.2.2. Sampling Theory

Suppose that we draw a random sample of n observations, $x_1, \ldots, x_n$ from an infinite population having mean μ and variance σ^2. The x_i are observations on mutually independent random variables. Any function $g(x_1, \ldots, x_n)$ of the observations is also a random variable. In particular, we shall be interested in functions such as the sample mean and variance. Let $\bar{x} = (\sum_{i=1}^{n} x_i)/n$ denote the sample mean. From the results of the previous section we have

$$E(\bar{x}) = \sum_{i=1}^{n} E\left(\frac{x_i}{n}\right) = \frac{n\mu}{n} = \mu$$

and

$$V(\bar{x}) = \sum_{i=1}^{n} V\left(\frac{x_i}{n}\right) = \sum_{i=1}^{n} \frac{\sigma^2}{n^2}$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

In many instances, it is a simple matter to derive not only the expectations and variances of means of independently distributed random variables but also their entire sampling distributions. The interpretation of a sampling distribution is as follows. Suppose that random samples of size n are drawn over and over again. Each such sample would result in a particular value of $\bar{x}$, so that these sample means may themselves be viewed as random variables having a common probability distribution. Two important theorems relating to the sampling distribution of $\bar{x}$ follow.

Theorem 1 The Law of Large Numbers

Let $X_1, \ldots, X_n$ be a sequence of mutually independent random variables with a common distribution having expectation μ. As $n \longrightarrow \infty$, the probability that $\bar{X}$ differs from μ by less than any arbitrarily prescribed value ϵ tends toward 1. In other words, the sample mean tends toward the mean of the parent population with probability 1.

Theorem 2 The Central Limit Theorem

If the random variables in the above sequence have variance σ^2, then as $n \longrightarrow \infty$, $Z = n^{1/2}.\ (\bar{X} - \mu)/\sigma$ tends toward the normal distribution with mean 0 and variance 1, with the density function given by

$$f(z) = (2\pi)^{-1/2} \exp\left(\frac{-z^2}{2}\right)$$

This distribution will be denoted by the notation $N(0, 1)$, and $Z \sim N(0, 1)$ means that Z is distributed as a $N(0, 1)$ random variable. The $N(0, 1)$ distribution is a special case of the normal distribution with mean μ and variance σ^2, denoted $N(\mu, \sigma^2)$. The density function for the $N(\mu, \sigma^2)$ distribution is given by

$$f(x) = (2\pi)^{-1/2} \exp\left\{\frac{-[(x - \mu)/\sigma]^2}{2}\right\}$$

The normal distribution is completely described by its mean and variance. Linear transformations of normally distributed random variables are also normally distributed. In particular, if $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$. The normal distribution is symmetric about μ; the intervals $\mu \pm 1.96\sigma$ and $\mu \pm 2.576\sigma$ contain 95% and 99% of the distribution, respectively.

Although the central limit theorem is an asymptotic result, it can be shown that for a wide class of random variables it provides a good approximation for fairly small values of n, say on the order of 10 or 20. This result is extremely important and provides the basis for development of normal sampling theory. In many real-life applications, observed values of random variables may actually be sums of small random perturbations; furthermore, even when the observations themselves have an arbitrary distribution the sample means are approximately normally distributed.

Since the normal distribution plays such an important role in statistics, we shall list here some of its important properties. Derivations may be found in Fraser (1958) or Kendall and Stuart (1963).

1. If $X \sim N(0, 1)$, then $Y = X^2 \sim \chi_1^2$, where χ_1^2 denotes the chi-square distribution on one degree of freedom. Degrees of freedom refer to the number of independent components into which a random variable may be decomposed.
2. If $Y_1, \ldots, Y_k$ are independently distributed as $N(0, 1)$, then $\sum_{i=1}^k Y_i^2 \sim \chi_k^2$, where χ_k^2 denotes the chi-square distribution on k degrees of freedom.
3. If X and Y are independently distributed as $N(0, 1)$ and χ_k^2 random variables, respectively, then $t = k^{1/2}X/Y^{1/2}$ is distributed according to the Student t distribution on k degrees of freedom, denoted t_k.
4. If Y_1 and Y_2 are independently distributed as $\chi_{k_1}^2$ and $\chi_{k_2}^2$, respectively, then $\mathfrak{F} = (Y_1 \div k_1)/(Y_2 \div k_2)$ has the F distribution on k_1 and k_2 degrees of freedom, denoted F_{k_1,k_2}.

The above results provide the basis for distribution theory related to sampling from a normal distribution. Suppose that $x_1, \ldots, x_n$ are a random sample drawn from the $N(\mu, \sigma^2)$ distribution. Then

1. $\bar{x} = (1/n) \sum_{i=1}^n x_i \sim N(\mu, \sigma^2/n)$.
2. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 \sim \sigma^2 \chi_{n-1}^2$.
3. $\bar{x}$ and s^2 are independently distributed.
4. $(n^{1/2})(\bar{x} - \mu)/s^2 \sim t_{n-1}$.
5. $n(\bar{x} - \mu)^2/s^2 \sim F_{1,n-1}$.

Tables of the normal, t, χ^2, and F distribution functions may be found in Pearson and Hartley (1962) as well as in a number of statistics texts.

1.2.3. Statistical Inference

The foregoing results provide the distribution theory essentials for discussing aspects of statistical inference related to the normal distribution. By statistical inference, we refer to the processes by which we make inferences concerning the characteristics of a population based on samples drawn from that population. The subject of statistical inference is a controversial one from a philosophical viewpoint, giving rise to several schools of thought that differ from one another principally in their interpretation of probability. In this section, we shall deal only with the concepts of significance tests and confidence intervals, leaving philosophical discussions of statistical inference to others. More complete discussions of statistical inference are contained in Fraser (1958), Kendall and Stuart (1963), and Lehmann (1959).

Suppose that we draw a sample from some population for the purpose of learning about its characteristics, or possibly making a decision, based on what we have observed. The population may be viewed as having a probability distribution with one or more parameters whose values are unknown. We may be interested in estimating the values of these parameters or in testing hypotheses concerning their values. These notions are perhaps easiest to describe by means of simple examples.