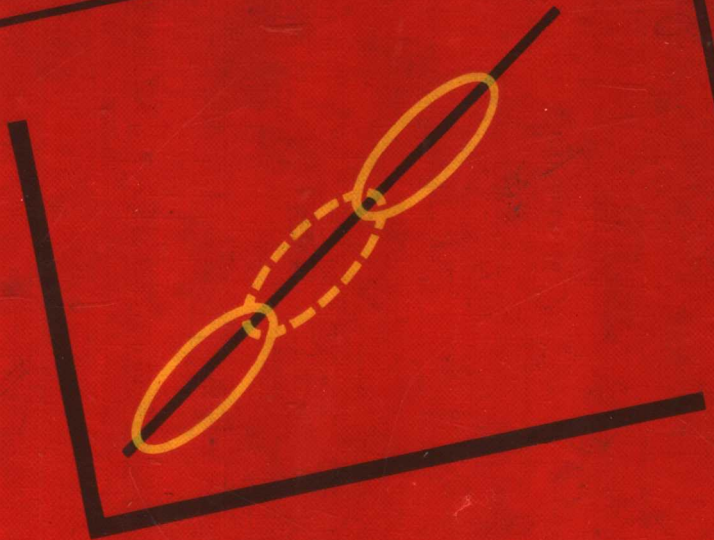


5344

Sung C. Choi

# Introductory Applied Statistics in Science



# Introductory Applied Statistics In Science

Sung C. Choi

*Washington University  
St. Louis*

***Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632***

*Library of Congress Cataloging in Publication Data*

CHOI, SUNG C

Introductory applied statistics in science.

Bibliography: p.

Includes index.

1. Mathematical statistics. 2. Biometry.

3. Science—Statistical methods. I. Title.

QA276.C474 519.5'02'45 77-22869

ISBN 0-13-501619-3

© 1978 by Prentice-Hall, Inc., Englewood Cliffs, N.J. 07632

All rights reserved. No part of this book  
may be reproduced in any form or  
by any means without permission in writing  
from the publisher.

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

The outline of this book evolved from lecture notes and materials, which were used in one-semester courses in applied statistics in the School of Engineering and Applied Science, and in several biostatistics courses in the School of Medicine at Washington University, St. Louis, Missouri.

It is designed to be a text in elementary applied statistics for students in science, mainly engineering and biomedical fields, and a compact reference text for the applied statistician and researcher. As a textbook it can be covered in a single semester if Sec. 9.6–9.10, Chap. 10, together with about eight additional sections, at the discretion of instructors, are omitted. Chap. 9 and 10 deal with relatively more complex topics, and are presented primarily for additional reading and reference. The prerequisite of this book is a standard algebra course and some of the basic ideas of calculus. Those readers without an introduction to calculus can omit the parts where calculus is used without loss of continuity.

In writing this book, an attempt was made to motivate the readers with the applicability and usefulness of statistical methods in the real world. Throughout the book, the methods are illustrated with real or realistic examples. Some slant toward medical examples perhaps reflects the author's interest and the availability of data. In addition, special sets of problems, many of them based on real data, follow each of the first nine chapters. It is hoped that those materials most frequently required or useful in applications are included here, though to some extent the importance of the materials depends on the field of applications. Although it is an applied statistics text, some effort is made to give those interested readers the rationality of the methods wherever possible.

It is suggested that the starred sections and starred problems be omitted for a shorter term course or for a strictly application-oriented course. In addi-

tion, the following alternative sequences come to mind for even shorter courses or for readers with different backgrounds and different interests.

- (a) Chapter 1 (1.1 through 1.4), Chap. 3, Chap. 5 (5.1, 5.5, 5.6, 5.8), Chap. 6 through Chap. 9.
- (b) Chapter 4, Sec. 3.6, Chap. 5 (5.1, 5.5, 5.6, 5.8), Chap. 6 through Chap. 9.
- (c) Chapter 5 (5.1, 5.5, 5.6, 5.8), Chap. 6 through Chap. 9.

The author is indebted to various publishers for permission to reproduce and adopt tables and figures as acknowledged in each table and figure. He is grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint Table III from their book, *Statistical Tables for Biological, Agricultural and Medical Research*, (6th edition, 1974).

He also wishes to express his appreciation to Sara Sanders for preparing the manuscript, and to many unnamed individuals, but particularly to Barbara Hixon for making helpful comments and suggestions. Finally, the author would like to thank the editorial staff of Prentice-Hall, Inc. for their friendly cooperation in the production of this book.

St. Louis, Missouri

SUNG C. CHOI

# Contents

Preface ix

Introduction 1

## 1. Random Variables and Probability 3

- 1.1 Introduction 3
- 1.2 Sample space, random variable and probability 3
- 1.3 Addition theorem 7
- 1.4 Conditional probability, independent events  
and multiplication theorem 9
- 1.5 Theorem on total probability 12
- 1.6 Bayes theorem 13
- 1.7 Summary 16
- Problems 16

## 2. Description of Random Variables 19

- 2.1 Introduction 19
- 2.2 Theoretical distribution of variables 19
- 2.3 Simple description of theoretical distributions 23
- \*2.4 More about mean and variance of variables 26
- 2.5 Summary 28
- Problems 28

<b>3. Some Important Theoretical Distributions</b>	<b>31</b>
3.1 Introduction	31
3.2 Bernoulli distribution and binomial distribution	32
3.3 Geometric distribution	34
3.4 Poisson distribution	35
3.5 Exponential distribution	37
3.6 Normal distribution	39
3.7 Lognormal distribution	47
3.8 Other distributions related to normal distribution	49
3.8.1 <i>t</i> -distribution	49
3.8.2 $\chi^2$ -distribution	50
3.8.3 <i>F</i> -distribution	51
3.9 Summary	54
Problems	54
<b>4. Organization of Data and Descriptive Statistics</b>	<b>58</b>
4.1 Introduction	58
4.2 Remarks on sample and population	58
4.3 Processes in statistical inference	59
4.4 Sample frequency distributions	60
4.5 Graphical representation of sample frequency distribution	62
4.6 Relation between sample frequency distribution and theoretical distribution	63
4.7 Measure of location	64
4.8 Measure of variation	66
4.9 Coefficient of variation	68
4.10 Summary	69
Problems	70
<b>5. Statistical Inference : Principles and Methods</b>	<b>73</b>
5.1 Introduction	73
*5.2 Basic principles of point estimation	74
*5.3 Method of point estimation	77
*5.4 Basic principles of confidence interval	79
5.5 Basic ideas of hypothesis testing	80
5.5.1 <i>Rationale of tests</i>	81
5.5.2 <i>One-sided and two-sided tests</i>	82
5.5.3 <i>Risk involved in decision based on tests</i>	83
5.6 Tests concerning mean of normal distributions with known variance	85

*5.7	Power and choice of test	89
5.8	$P$ value	91
5.9	Problem of sample size	93
	Problems	95
<b>6.</b>	<b>Estimation and Testing Hypotheses: Frequency Data</b>	<b>97</b>
6.1	Introduction	97
6.2	Inference about a proportion	98
6.2.1	Testing the hypothesis about a proportion	98
6.2.2	Confidence interval for a proportion	100
6.3	Comparison of two proportions	101
6.4	Goodness of fit test	102
6.5	Two-by-two table	104
6.6	Contingency tables	108
6.7	Comparison of several proportions	111
6.8	Two-by-two table for paired observations	113
*6.9	Test of hypothesis and sample size	115
6.9.1	Sample size for testing a proportion	116
6.9.2	Sample size for comparing two proportions	117
6.10	Summary	118
	Problems	118
<b>7.</b>	<b>Estimation and Testing Hypotheses: Measurement Data</b>	<b>123</b>
7.1	Introduction	123
7.2	Variance of a normal distribution	124
7.2.1	Confidence interval for $\sigma^2$	124
7.2.2	Test concerning a variance	125
7.3	Mean of a normal distribution with unknown variance	127
7.3.1	Confidence interval for $\mu$	127
7.3.2	Test concerning a mean: One-sample $t$ -test	128
7.4	Comparison of the variances of two normal distributions	130
7.4.1	Confidence interval for $\sigma_1^2/\sigma_2^2$	131
7.4.2	Test for the equality of two variances	131
7.5	Comparison of means of two normal distributions:	
	Independent case	133
7.5.1	Pooled standard deviation when $\sigma_1 = \sigma_2$	133
7.5.2	Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$	133
7.5.3	Testing the equality of two means when $\sigma_1 = \sigma_2$ :	
	Two-sample $t$ -test	134
7.5.4	Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 \neq \sigma_2$	136
7.5.5	Testing the equality of two means when $\sigma_1 \neq \sigma_2$ :	
	Aspin-Welch test	137



7.6	Comparison of means of two normal distributions: Paired case	138
7.6.1	Confidence interval for $\mu_1 - \mu_2$ in paired data	139
7.6.2	Testing the equality of two means: Paired t-test	139
7.7	Robustness, transformation and nonparametric tests	141
7.7.1	Transformation	141
7.7.2	Nonparametric tests	143
7.8	Rank test for comparing two populations: Independent case	144
7.9	Rank test for comparing two populations: Paired case	146
7.10	Other nonparametric tests	148
7.10.1	Median test	148
7.10.2	Sign test	149
7.11	Large sample test for the mean	150
*7.12	Test of hypothesis and sample size	152
7.12.1	Sample size for one-sample test of means	152
7.12.2	Sample size for two-sample test of means	153
7.13	Summary	155
	Problems	155
8.	Regression and Correlation Analysis	160
8.1	Introduction	160
8.2	Simple linear regression	161
8.3	Inferences about regression coefficients	165
8.4	Inferences about predicted value	168
8.5	Correlation coefficient	169
8.6	Inference about correlation coefficient	171
8.7	Analysis involving more than one $x$	174
8.7.1	Correlation matrix	174
8.7.2	Partial correlation coefficient	175
8.7.3	Multiple regression	177
*8.8	Non-linear regression model	179
8.9	Contingency coefficient	180
8.10	Biserial correlation	181
8.11	Rank correlation	183
8.12	Summary	185
	Problems	185
9.	Analysis of Variance	191
9.1	Introduction	191
9.2	Models and assumptions in analysis of variance	191

9.3	Dot notation for representing means	193
9.4	One-way classification	193
9.5	Multiple comparisons—LSD test	198
*9.6	Two-way classification	200
9.6.1	<i>Crossed classification with replication</i>	201
9.6.2	<i>Crossed classification with no replication</i>	206
9.6.3	<i>Nested classification</i>	208
*9.7	Miscellaneous remarks on the analysis of variance	210
9.7.1	<i>Low F-ratios</i>	210
9.7.2	<i>Missing Data</i>	211
9.7.3	<i>Greater number of classifications</i>	213
*9.8	Applications of analysis of variance to regression	214
*9.9	Analysis of covariance	216
	Problems	221

## 10. Computer Analysis     224

10.1	Introduction	224
10.2	Components of a computer system and programs	224
10.3	Preparing data for computer analysis	227
10.4	Case studies	230

## References     238

## Appendix 1 Subscripts and Summations     240

## Appendix 2 Tables and Figures     242

## Appendix 3 Answers to Selected Problems     266

## Index     271

# Introduction

Statistics is the scientific method of collecting information in a form of numerical data and drawing conclusions by analyzing the information. Consider, for example, the following problems:

1. deciding whether or not a certain game is fair;
2. estimating the number of fish in a lake;
2. determining the unemployment rate;
4. deciding whether or not a drug is effective;
5. comparing mileage obtained using several different brands of gasoline;
6. testing the possible relation between the length of the "life-line" on the hand and life expectancy;
7. deciding whether or not cigarette smoking causes cancer; or
8. estimating the yield of wheat for different amounts of a standard fertilizer applied.

In each of the above problems, the only practical scientific approach is to perform some sort of experiment or survey and base the solution on the information obtained. But what kind of information and how much? And after we have the information, what do we do with it to solve the problem? Statistics deals with answering these kinds of questions by specific techniques.

Statistics usually consists of four broad processes, although there are not always clear boundaries between them: collection, organization, analysis of numerical data, and the decision process.

Collecting data is the process of obtaining measurements or counts after some sort of experiment or survey is conducted. Valid conclusions can result only from properly collected data.

Organization of data is the process of preparing and presenting the collected data in a form suitable for description as well as for further analysis.

## **2 Introduction**

Analysis of data is the process of performing certain calculations and evaluations in order to extract relevant and pertinent information buried in the data.

The decision process is the task of interpreting and reaching valid conclusions based on the analysis of the data and the mathematical theory of probability. The analysis of data and the decision process form the main portion of this book.

## Random Variables and Probability

### 1.1 Introduction

In many scientific studies, we deal with experiments that are repetitive in nature or that can be conceived as being repetitive. For example, if we toss a coin ten times, we may want to know the chance that no head will appear, one head will appear, two heads will appear, and so on. If a sample of 100 electronic tubes is selected from a shipment and each tube is tested, it may be desirable to estimate the proportion of defective tubes in the lot. In a medical investigation, if the survival time of a group of sick mice receiving a placebo and a second group receiving a certain medication is recorded, we may want to examine the effectiveness of the treatment in terms of survival time. These are illustrations of experiments that can be carried out actually or conceptually. In the study of probability, we are concerned with the derivation of the laws and rules of chance related to the outcomes of experiments.

### 1.2 Sample Space, Random Variable, and Probability

Consider an experiment of tossing a coin. In the experiment, there are only two possible outcomes, a head or a tail. It is convenient to represent head and tail by certain letters, for example,  $H$  for head and  $T$  for tail, although such letters can be arbitrary. For an experiment of rolling a die, there are six possible outcomes, most conveniently represented as  $1, 2, \dots, 6$ .

A point representing a possible outcome of a given experiment is called a *sample point*, and the set of all sample points is called the *sample space*. A numerically defined variable on a sample space is called a *random variable*. To be precise, a random variable is a numerical function defined on each element of the sample space. In practice, however, a random variable shall be conceived as a numerical value assigned to each element.

To be precise, let  $S = \{e_1, e_2, \dots, e_n\}$  denote the sample space, with each  $e_i$  representing a sample point. A random variable  $X$  is defined by assigning a real number  $x_i$  to each element  $e_i$  of  $S$ . Thus, we may write

$$X(e_i) = x_i.$$

In practice, the random variable defined by an investigator depends on the nature and purpose of the study and the criterion used, and it is usually denoted simply by  $X$  instead of  $X(e_i)$ . Also, it is often called the *variable* for the sake of simplicity.

**Example 1.1** Consider the experiment of rolling a die. The sample points can be conveniently represented by  $1, 2, \dots, 6$ , and the sample space is given by a set  $\{1, 2, \dots, 6\}$ . A random variable  $X$  can be defined as

$$X(1) = 1, X(2) = 2, \dots, X(6) = 6.$$

The relationship between the outcome and the random variable defined here is clearly illustrated in Fig. 1.1.







Outcome of experiment	Value of random variable $X$
	$X = 1$
	$X = 2$
	$X = 3$
	$X = 4$
	$X = 5$
	$X = 6$

Fig. 1.1 Relationship between outcome and random variable

Alternatively,  $X$  can be defined as follows if one is concerned only as to the outcome being an odd or an even number.

$$X(1) = X(3) = X(5) = 0, \text{ and } X(2) = X(4) = X(6) = 1,$$

or in many other ways. ▲

**Example 1.2** As the second example, a random variable might be defined as the number of heads appearing when two coins are tossed. Then,  $S = \{HH, HT, TH, TT\}$ , and

$$X(H, H) = 2, X(H, T) = X(T, H) = 1, \text{ and } X(T, T) = 0. \quad \blacktriangle$$

**Example 1.3** Consider the survival time of mice with a certain disease. The survival time can be any positive real number as is the sample point. The sample space is continuous and is given by a set  $\{t | 0 \leq t < T\}$ , where  $T$  is a large real number. A random variable  $X$  can be defined as the survival time of the animal itself; that is,

$$X(t) = t.$$

Alternatively, it can be defined as

$$X(t) = 0 \text{ if } 0 \leq t \leq 5 \text{ days,}$$

$$X(t) = 1 \text{ if } 5 < t \leq 18 \text{ days,}$$

$$X(t) = 2 \text{ if } t > 18 \text{ days,}$$

and again in many other ways.  $\blacktriangle$

Note that each outcome always determines one and only one value of the random variable  $X$ , but a given value of  $X$  may correspond to more than one outcome, although often assigned to only one outcome. Indeed, in many situations, the outcome of the experiment is already in the numerical form that we want to record and use as a random variable.

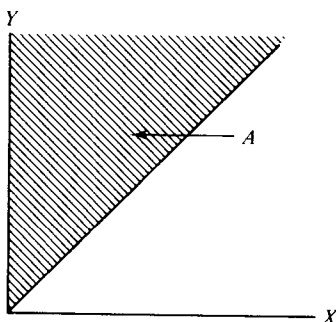
Returning to the sample space of an experiment, any subset of the sample space is called an *event*. The events shall be noted by capital letters, for example,  $A, B$ , etc. Suppose that the experiment related to the event  $A$  is performed. Every sample point belongs to either  $A$  or not to the event  $A$ . Only if the sample point belongs to  $A$ , can the event  $A$  then be said to have occurred. For example, for the experiment of rolling a die, let the event be defined as  $A = \{2, 4, 6\}$ . Then, we say that the event  $A$  has occurred when each roll of the die results in one of the three numbers, namely, 2, 4, and 6.

Given any event  $A$ , it is natural to consider the event that  $A$  does not occur, denoted by  $\bar{A}$ . Such an event is called the *complement* of  $A$  and consists of all elements in the sample space which are not in  $A$ . For example, if  $A = \{2, 4, 6\}$  in the experiment of rolling a die, then  $\bar{A} = \{1, 3, 5\}$ .

As has been stated, the random variable, or simply the variable, is a number assigned to the outcome of an experiment and, as might be expected, there is a relation between the variable and an event. In brief, the variable can define an event. Thus, sample points determined by  $\{a \leq X \leq b\}$ , where  $X$  is a variable and  $a$  and  $b$  are real numbers, always constitute an event. For example,  $\{X = 2\}$  in Example 1.2 determines the event that both coins show heads, and  $\{0 \leq X \leq 1\}$  the event characterized by "at most one head." In

**Example 1.3**, if the variable  $X$  is defined by  $X = t$ , then  $\{X > 3\}$  defines the event that the survival time of the animal is greater than three days, if day was the time scale. In statistical analyses, we shall deal with the observed values of  $X$  or the event defined by  $X$ . The main reason for, and advantage in, defining the random variable should be clear to the reader: it is much more convenient to work with a set of given numbers precisely defined on the outcome of an experiment or observation than to work with the outcome itself.

**Example 1.4** Consider the weight of couples. Let  $X$  and  $Y$  denote the weights of a husband and wife, respectively. Each sample point is given by  $(X, Y)$ ,  $X > 0$ ,  $Y > 0$ , although in reality  $X$  and  $Y$  are bound by certain values, and the sample space can be represented by the first quadrant of the  $X, Y$ -plane. The event  $A$ , “wife is heavier than husband,” for example, is given by the shaded region of Fig. 1.2.



**Fig. 1.2** Event  $A$  “Wife is heavier than husband”:  $X$  is weight of husband and  $Y$  the weight of wife

Many random variables can be defined; for instance, the random variable  $W$  can be the sum of a couple's weights, that is,  $W = X + Y$ . Next, the random variable  $D$  can be defined as the absolute difference between the weights,  $|X - Y|$ . It is also true that both  $X$  and  $Y$  are random variables. As a final example, let  $Z = 1$  if  $X < Y$  and  $Z = 0$  if  $X \geq Y$ . Then,  $Z$  is clearly a random variable, and  $\{Z = 1\}$  is characterized by the event  $A$  of Fig. 1.2. ▲

A basic and intuitive meaning of probability is most easily given when the sample space consists of a finite number of sample points, and when each sample point is equally likely to occur in a repetitive experiment. Suppose we are interested in a certain event  $A$  as to the likelihood that the event will occur in a single trial. The classical definition states: the *probability* of an event  $A$  is the ratio of the number of sample points in  $A$  to the total number of sample points. Thus, if  $P(A)$  denotes the probability of the event  $A$ , and if



$n_A$  and  $n$  denote the number of sample points in  $A$  and the total number of sample points respectively, then

$$P(A) = \frac{n_A}{n}. \quad (1.1)$$

For example, consider again the experiment of rolling a fair die with the event  $A$  defined as  $A = \{2, 4, 6\}$ . Since each of six numbers is equally likely to appear in the experiment,  $P(A) = \frac{3}{6} = \frac{1}{2}$ . Note that if a random variable is defined as  $X(1) = 1, X(2) = 2, \dots, X(6) = 6$ , then  $P(A) = P(X = 2, 4, \text{ or } 6)$ . On the other hand, if it is defined as  $X(1) = X(3) = X(5) = 0$  and  $X(2) = X(4) = X(6) = 1$ , then  $P(A) = P(X = 1)$ .

If the total number of sample points is infinite, the definition given by (1.1) is, of course, not appropriate. More generally, we may define the probability of the event  $A$  as the relative chance or likelihood that  $A$  occurs in a given experiment. This means roughly that the probability is the fraction of times the event  $A$  occurs if the experiment is repeated a large number of times under essentially identical conditions. This definition is somewhat ambiguous, but the more exact and broad meaning of the probability will become clear in the next chapter.

### 1.3 Addition Theorem

Suppose we have two events  $A$  and  $B$ . The event consisting of all sample points contained in  $A$  or  $B$ , or both, is called the *union* of  $A$  and  $B$ : it is written as

$$A \cup B.$$

The event consisting of all points contained in both  $A$  and  $B$  is called the *intersection* of  $A$  and  $B$ : it is written

$$A \cap B.$$

In Fig. 1.3 the event  $A \cup B$  is represented by the unshaded plus the shaded regions of  $A$  and  $B$ , while the event  $A \cap B$  is given by the shaded region only. Such a figure is known as a Venn diagram.

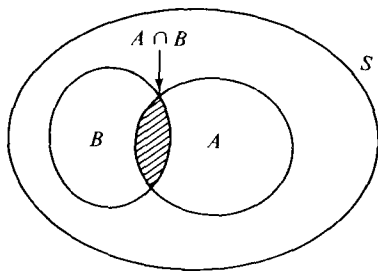


Fig. 1.3 Two events  $A$  and  $B$  within sample space  $S$