Statistics

A Model for Uncertainty

Frank P. Soler
Chris W. Avery

Statistics:

A Model for Uncertainty

Frank P. Soler and Chris W. Avery
De Anza College
Cupertino, California

KENDALL/HUNT PUBLISHING COMPANY
4050 Westmark Drive Dubuque, Iowa 52002

Revised Printing 1997

Library of Congress Catalog Number: 96-77385

ISBN 0-7872-7069-3

Printed in the United States of America
10 9 8 7 6 5 4 3

Preface

Statistics: A Model for Uncertainty is a general, yet comprehensive, introduction to statistics for students at two-year and four-year colleges and universities and for advanced high school students preparing for the AP Examination in Statistics. In this preface we will briefly describe the book in order to give teachers and students an overview of its content and direction.

The driving principle behind the book is to bring out statistical concepts and ideas while keeping computational and algorithmic driven activities to a minimum. We assume that the students have passed a second course in Algebra. Both, student and teacher, are given ample opportunity to engage in the business of learning and teaching. Many of the examples and problems in the book are intended to stimulate discussion of basic statistical issues. We are fully aware that many of these issues are open in nature, that is, there are no unique solutions or approaches. However, there are underlying concepts and ideas that should be part of any approach. Throughout, we emphasize the latter.

Many statistical procedures are influenced by probabilistic thought. As such, there is no escaping the study of probability. We acknowledge that the issue of how to present probability in a course of this nature will be the subject of discussion for years to come. Research in learning probability documents the fragility of probabilistic concepts. To this end, we have included very little formal probability. The chapter on probability (Chapter 3) deals with elementary ideas about a sample space, simple events, and compound events. The examples and problems we have given follow the motto: "if it can't be done with a Venn diagram, tree diagram, or a table, then it is not done..." Mutually exclusive, independence, and conditioning are presented in order to facilitate their later role in the treatment of inference.

We focus on the idea of a random variable and develop it early. We do this without any formal probability. We view the random variable as a tool that facilitates the study of random phenomena. Chapters 1 and 2 get the student thinking in terms of uncertain (rather than deterministic) outcomes. The emphasis is in gathering data to study these phenomena and in organizing and displaying the data in order to obtain useful information about them.

Chapter 4 covers discrete random variables. It starts by introducing expectation and variance and continues with Bernoulli trials and coverage of the major discrete distributions (Binomial, Geometric, Pascal, Poisson, and Hypergeometric). We do this without using computational formulas, with the exception of tables for the Binomial and very simple computations for the Geometric. We are mostly interested in having the student recognize different ways of thinking about discrete phenomena. All but the Binomial distribution may be skipped without loss of continuity.

The transition from discrete to continuous random variables occurs in Chapter 5. The key distribution here is the Uniform. We emphasize the idea of probability as an area. We cover two additional distributions: Exponential and Triangular. We recognize that most of this chapter can be easily skipped without any loss of continuity.

The Normal distribution is the subject of Chapter 6. We feel that a good understanding of this distribution goes a long way toward paving the study of other continuous distributions (i.e., t, F, Chi-square).

Chapter 7 introduces the fundamental idea of the sampling distribution of a statistic. This leads to the Central Limit Theorem. We work with both versions, averages and sums. We thought it would be appropriate to introduce the Law of Large Numbers as a way of interpreting the population mean in terms of a long run relative frequency.

Chapter 8 marks the beginning of inference. It gives an overview of estimation and proceeds to develop the methodology for constructing confidence intervals.

Chapter 9 gives a formal introduction to hypotheses testing. We develop the p-value, tying it to previously developed material, and then we talk about decision making, including the Type I and Type II error probabilities, and issues of significance (fixed and practical).

Chapters 10–13 explore additional hypotheses testing situations using the Student's-t distribution, Chi-square and F. We devote some material to examining assumptions about these distributions and pointing out pitfalls and checkpoints.

Chapter 14 introduces bivariate data, starting with the scatterplot and continuing with correlation, outliers, influential points, and least-squares regression.

Many examples and problems throughout the book are real. We cite references where appropriate. In many instances, the applications come from the authors' experience and expertise. In such cases, no references are given.

We struggled with the usual format of including a problem set after each section. We decided against this. Our thinking was that most chapters are fairly brief, with plenty of examples for the student to work through. The one advantage to having a single problem set at the end of each chapter is the ability to immediately incorporate concepts from the entire chapter throughout most of the problem set.

We encourage the use of calculators. In fact, it would be impossible to read through these materials without a handy calculator. Any calculator that is capable of accumulating sums will be sufficient. The best way to check this is to make sure that keys such as $\bar{x}$, s, σ, appear on the calculator.

Algebra Review

As part of the Appendix, we have included a brief Algebra Review. This is done in terms of problems to be solved. Complete solutions are included.

Software Supplement

There are many places in the book where we mention the use of a computer or "appropriate software". While the exposition and problems are free of computer software, we recognize that having access to pedagogically sound statistical software is advantageous. We recommend *Statistics,* by The Math Lab. It is available for the IBM or compatibles and the Macintosh in 3.5 inch disk format. For students who wish additional practice, this package, complete with a User's Manual, contains an extensive **tutorial**, covering all of the major topics in this book, complete with random generation of problems, hints, and a self-scoring system. Additionally, it allows for extensive **simulations** using fourteen distributions, producing empirical and theoretical graphs, and allowing for the computation of probabilities and critical values. Extensive demonstrations of difficult statistical issues are also included. For instance, in Analysis of Variance, the user can easily simulate the trade-off between sample size and magnitude of the unknown population variance. Also, it is enlightening to demonstrate the statement of the Central Limit Theorem by sampling from any number of distributions and comparing the shape and statistics of the corresponding sampling distribution of averages or sums to that of the underlying population.

Acknowledgements

We wish to acknowledge a former student of ours, Teck Ky, for his assistance in providing solutions to many of the problems in the problem sets. Lenore DeSilets, a mathematics instructor at De Anza College, reviewed parts of the original manuscript and provided us with useful comments. James Lum, of San Jose State University and adjunct faculty at De Anza College, pointed out areas of difficulty and ambiguous wording while teaching from our first draft. A special thanks to Professor Lum for sharing with us his keen understanding of statistical issues. Carol Olmstead, also a mathematics instructor at De Anza College, taught from the first printing and provided us with numerous suggestions on how to improve the exposition.

Most of all, we are indebted to the many students who, since 1986, have learned statistics from our previous textbook. Their triumphs have inspired us. Their struggles drove us to produce this book. We are equally grateful to all those instructors, especially at De Anza College, who have taught from our materials and used our software over the last nine years. We have learned much from them. Many of their comments and constructive criticisms are very much a part of this work.

Chris W. Avery and Frank P. Soler
June 1997

COMPUTER SOFTWARE

TUTORIAL, SIMULATION, GRAPHICS

Computer software is available as a supplement to **Statistics: A Model for Uncertainty**. Thoroughly tested at community colleges, universities, and high schools, this software is composed of a comprehensive tutorial package, emphasizing basic conceptual and computational elements of the major topics in an elementary statistics course, along with powerful simulation, graphics, and demonstration capabilities.

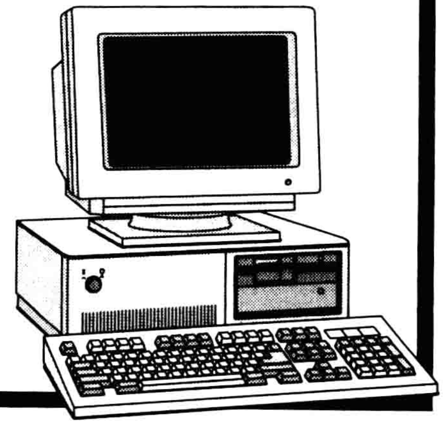
OUTSTANDING FEATURES:

- Random problems are generated for major topics with three levels of difficulty
- Students' scores are recorded and score pages can be printed
- The simulation component provides a convenient way to demonstrate some of the most difficult concepts of the course. (i.e., the Central Limit Theorem)
- Demonstration capabilities allow for the generation of data from 14 probability distributions and accompanying graphic demonstrations. (i.e., overlapping of theoretical and empirical results)

Not intended to be a "number cruncher", this software package involves your students in exploring and discovering the "why and how" of statistics at an elementary, yet highly conceptual, level. The authors have also written a set of laboratory assignments using the simulation and graphics component of the software, which are available from the Math Lab, upon adoption, with permission to reproduce them for classroom use only.

To order, contact: **The Math Lab, 10893 Leavesley Place, Cupertino, CA 95014**

Call (408) 257-7217
Fax (408) 252-2686



Quantity Ordered:	1-4	5-20	21+
Cost/Diskette:	$16	$14	$12

(Networking and site licenses are available.
Arrangements made on an individual basis.)

Table of Contents

Chapter 4 Discrete Probability Distributions

Chapter 5 Continuous Probability Distributions

Chapter 6 The Normal Distribution

Chapter 7 Sampling Distributions, the Central Limit Theorem, and The Law of Large Numbers

Chapter 8 Estimation and Confidence for Means and Proportions

Chapter 9 Testing Statistical Hypotheses and Inference as Decision

Chapter 10 The Student's–t Distribution

Chapter 11 Testing Means and Proportions about Two Populations

Chapter 12 The Chi-Square Distribution and its Applications

Chapter 13 The F Distribution and its Applications

Chapter 14 Bivariate Data: Correlation and Regression

Index

Chapter 1

Uncertainty, Randomness and Data

1.1 The need to model uncertainty

Some phenomena have predictable outcomes: drop an object from a known height and the time it takes to fall can be precisely predicted (banning small measurement errors) from known physical equations. However, there are other types of phenomena that are not so predictable. For instance, take a fair coin and flip it. We cannot predict whether it will come up heads or tails. That is, the outcome is *uncertain*. Yet, note that coin flipping is not haphazard or chaotic. Intuition indicates that if we flip a fair coin a large number of times, the proportion of heads (or tails) will be very close to one-half. This long term regularity is not a theoretical construct, it is an observed fact.

Quantifying uncertainty involves the related topics of *data* and *chance*.[1] Uncertain situations appear everywhere in the world around us. Each of the following is an example of an uncertain situation. For each one, think of how data and chance are a part of the phenomena.

- The closing price of Apple stock at the end of the next business day
- The number of people who will buy the new exercise machine built by NordicTrack
- The winner of the Baseball World Series for the current season
- The number of dart throws before the thrower hits the bulls' eye
- The improvement shown by patients for a given dosage of a new medication
- The numbers that will come up in the next lottery drawing

1.2 Random Variables

Phenomena with predictable outcomes are said to be *deterministic*. We typically use mathematical functions to describe and quantify deterministic phenomena. Here are some examples of deterministic situations and their associated functions.

- The area of a rectangular plot of land 30 ft. by 50 ft.
  Area function = length•width = 30•50 = 1500 square ft.

[1] **Statistics** deals with the study of data while **Probability** deals with the study of chance

- The total distance traveled after 4 hours at a constant rate of 15 miles per hour.
  Distance function = rate•time = 15•4 = 60 miles.

- The profit made if revenues are $1 million and costs are $0.75 million.
  Profit function = revenue – cost = 1 million – 0.75 million = $0.25 million.

Phenomena with unpredictable outcomes are classified as *random.* A **random variable** is a function with an uncertain outcome.[2] Outcomes may be numerical in nature or exhibit an inherent non-numerical characteristic. The following are examples of random phenomena and some of their possible outcomes.

(1) X = number of cars passing a toll booth during a fixed time period
Possible values of X: X = 50; X = 300.

(2) Y = type of computer a company will use in 5 years
Possible values of Y: Y = Apple; Y = IBM; Y = Compaq.

(3) X = the length, in minutes, of the next phone call to your most significant other
Possible values of X: X = 1; X = 25; X = 1.2.

(4) W = the height, in inches, of the students in your Statistics class
Possible values of W: W = 62; W = 76.9.

(5) An urn contains 3 red marbles and 7 yellow marbles. One marble at a time is selected, without replacement, until a red marble appears. X = the number of marbles selected.
Possible values of X: X = 2; X = 5. Note that for convenience, it is customary to denote the case X = 2 by the outcome: **YR.** This means that a yellow marble was chosen first and then a red marble was chosen. The case X = 5 is denoted by the outcome: **YYYYR.**

We will study three different types of random variables: *quantitative discrete*, *quantitative continuous* and *qualitative* (also called *categorical* or *attribute.*) A *discrete* random variable keeps track of **counts.** A *continuous* random variable obtains its values from **measuring.** Outcomes associated with discrete and continuous random variables are numerical. The set of all possible outcomes for a situation involving uncertainty is called the *range* of a random variable. For each of the five examples directly above, the type of random variable and range are as follows:

[2] It is customary to denote random variables using capital letters from the tail end of the alphabet. For example, W, X, Y, Z, etc. This notation will be most prominent later in the book.

	Type of Random Variable	Comment	Range
(1)	Quantitative discrete	We would *count* the number of cars	Non-negative counting numbers
(2)	Qualitative or categorical	Computer makers represent *categories*	Name of manufacturer
(3)	Quantitative continuous	Time is *measured*	Positive real numbers
(4)	Quantitative continuous	Height is *measured*	Positive real numbers
(5)	Quantitative discrete	We *count* the number of marbles selected	1,2,3,4,5,6,7,8 (Why not 9 or 10?)

1.3 Definition of key terms

Besides random variables, the following terms will be used extensively throughout the remainder of this book.

Experiment: A planned activity that yields meaningful results

Population: The complete collection of objects, items, persons, or things under study

Parameter: A statement or measurable characteristic pertaining to the population

Sample: That part of the population from which information is actually gathered

Statistic: A statement or measurable characteristic pertaining to the sample

Data: Values of the random variable produced by the experiment.

We illustrate all of the above terms with the following two examples.

Example 1

We wish to estimate the average age of all the homeowners of the city of San Francisco.

Experiment: the process of obtaining the ages of San Francisco homeowners

Population: all San Francisco homeowners

Random Variable: age of homeowners

Parameter: the average age of all San Francisco homeowners

Sample: all homeowners around Golden Gate Park

Statistic: the average age of those homeowners around Golden Gate Park

Data: counting numbers representing age. That is, age is a *discrete* random variable.

Example 2

We want to know the largest number of heads that come up when 3 fair coins are simultaneoulsy flipped. For each coin, let H = heads comes up and T = tails comes up. Thus, the outcome **HTH** denotes that the 1st and 3rd coins came up heads and the 2nd coin came up tails.

Experiment: the act of simultaneously flipping 3 fair coins

Population: all possible outcomes when 3 fair coins are simultaneously flipped

Random Variable: number of heads that come up when 3 fair coins are simultaneously flipped

Parameter: the largest number of heads that can come up when 3 fair coins are simultaneously flipped (i.e., 3)

Sample: Suppose we repeat the experiment 5 times: TTH, HHT, TTT, HTT, THT

Statistic: 2 (note this is the largest value of the random variable for the 5 outcomes in the sample)

Data: 1,2,0,1,1 (note these are the values of the random variable for the outcomes in the sample)

1.4 Data and sampling

Data are gathered in order to study the behavior of random variables. The gathering, analysis, and representation of data are major components of the field of *Statistics*. Securing data that are meaningful and trustworthy takes a great deal of planning. The data that statisticians are most interested in stems from carefully planned experiments. We classify the different types of data in the same manner in which we classified random variables. That is, data may be quantitative (numerical: discrete, continuous) or qualitative (categorical). The following are examples of experimental situations where the random variable of interest produces numerical or categorical outcomes.

Example 3

Experiment: obtain the age, in years, of randomly selected students from the student body

Random variable: age of the students

Possible outcomes: discrete data such as: 20, 35, 18

Example 4

Experiment: add the number on the faces that come up when two fair dice are rolled

Random variable: sum of the number on the faces that come up

Possible outcomes: discrete data such as: 2, 6, 10, 12

Example 5

Experiment: determine which type of bread is eaten more frequently for lunch at a prominent San Francisco sandwich shop

Random variable: different types of bread

Possible outcomes: qualitative data such as: wheat, sourdough

Example 6

Experiment: determine the useful life, in hours, of a set of neon light bulbs

Random variable: number of hours before neon light bulbs burn out

Possible outcomes: quantitative continuous data such as: 1005.7, 950, 2375

Often we wish to extend the conclusions we draw from analyzing sample data to some larger group of individuals (i.e., the population). If the data don't fairly represent the larger group, our conclusions will not apply to the larger group. In order to gather data that are trustworthy, careful attention must be paid to **sampling**. The whole idea of sampling is to study a part of the population.

The following two examples illustrate the danger of believing results of experiments conducted with poorly produced data.