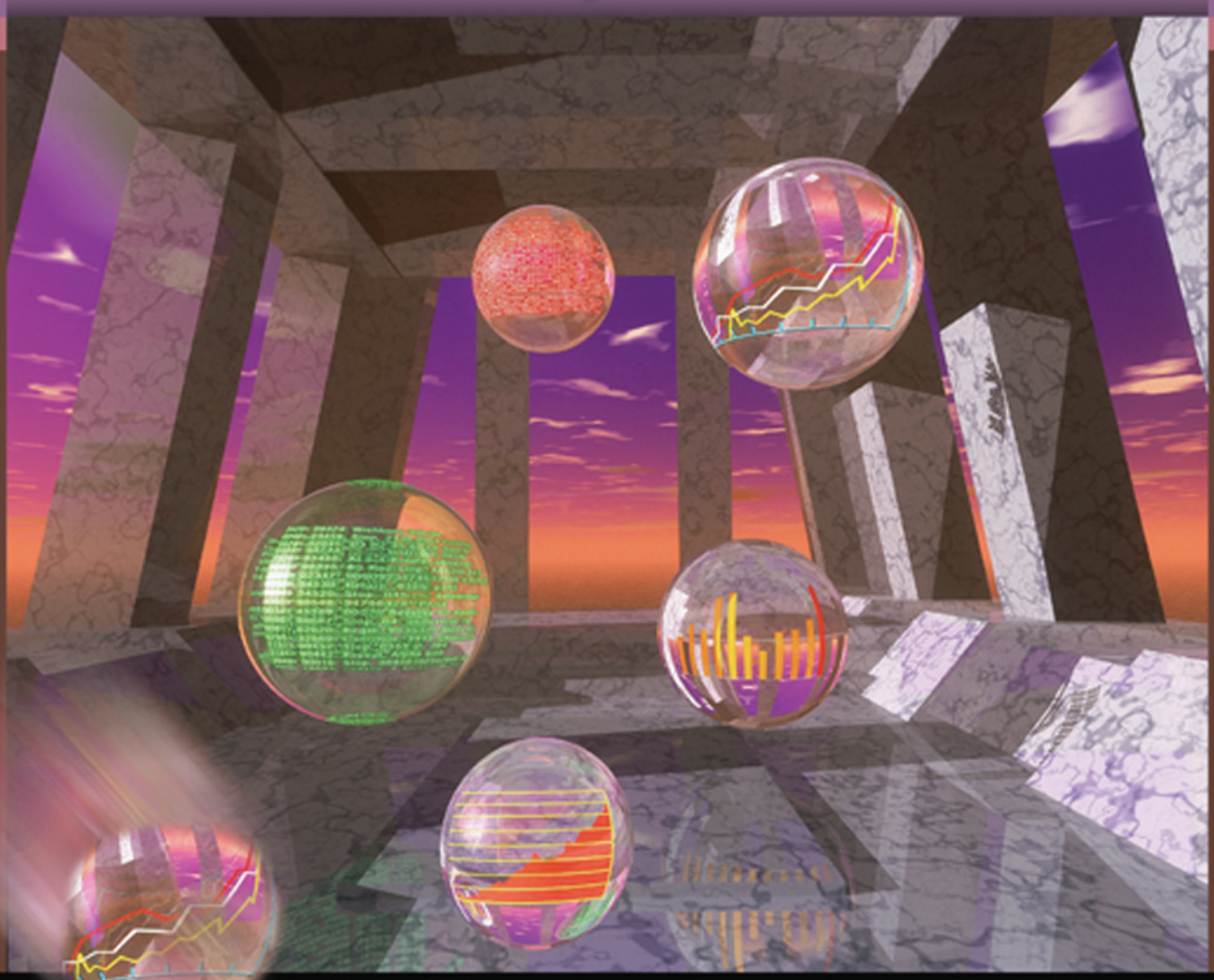
ELSEVIER
ACADEMIC
PRESS

Advanced Statistics from an Elementary Point of View



MICHAEL J. PANIK

Advanced Statistics from an Elementary Point of View

Advanced Statistics from an Elementary Point of View

Michael J. Panik

University of Hartford

ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Acquisition Editor: Tom Singer
Project Manager: Sarah Hajduk
Marketing Manager: Linda Beattie
Cover Design: Eric DeCicco
Cover Image: Getty Images
Composition: Cepha Imaging Pvt Ltd
Cover Printer: Phoenix Color
Interior Printer: The Maple-Vail Book Manufacturing Group

Elsevier Academic Press
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data

Panik, Michael J.
Advanced statistics from an elementary point of view / Michael Panik.
p. cm.
Includes bibliographical references and index.
ISBN 0-12-088494-1 (acid-free paper)
1. Mathematical statistics—Textbooks. I. Title.
QA276.P224 2005
519.5—dc22 2005009834

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 13: 978-0-12-088494-0
ISBN 10: 0-12-088494-1

For all information on all Elsevier Academic Press Publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America
06 07 08 09 9 8 7 6 5 4 3 2 1

To the Memory of
Frank C. Grella
Friend and Colleague

Contents

16 Bivariate Linear Regression and Correlation 669

Appendix A 717

Preface

This book is intended as an introduction to probability and statistical inference for junior- or senior-level students in one- or two-semester courses offered by departments of mathematics or statistics. It can also serve as the foundation text for first-year graduate students in disciplines such as engineering, economics, and the physical and life sciences, where considerable use of statistics is the norm.

No previous study of probability or statistical inference is assumed. The only prerequisite is the standard introductory course in the calculus. Review sections dealing with set algebra, functions, and basic combinatories are included for your convenience.

A strength of this book is that it is highly readable. Great care has been taken to fully develop statistical concepts and definitions. Detailed explanations of theorems, tests, and results are offered without compromising the rigor and integrity of the subject matter. An objective of this work is to get students to concentrate on the statistics without being overwhelmed by the calculations. Students who have used this book should be well on their way to thinking like a statistician when it comes to problem solving- and decision- making.

An important feature of this text is the considerable attention given to sampling distributions, point and interval estimation, parametric and distributional hypothesis testing, and linear regression and correlation. These topics typically constitute the heart and soul of most statistics courses, and this book has been written with this notion in mind.

This book can be used at a variety of levels. If theorem content but not theorem proof is important, then the general flow of the various chapters can be followed and the task-oriented/applications exercises found at the end of each chapter can be selectively chosen. However, if proofs and derivations are an integral part of the course, then the exercises that address the same can be attempted. The motivation underlying the execution of each proof as well as step-by-step details necessary for its completion are offered in the *Instructor's Manual*. So although the main text is certainly not devoid of proofs (it engages the reader in proofs that are more or less constructive or that reinforce the conceptual notions and definitions at hand), the more complex and mathematically challenging proofs

are available as standalone items and presented without impeding the continuity of presentation of the basic material.

After a review in Chapter 2 of some basic descriptive concepts, Chapter 3 develops the rudiments of probability theory. The latter is a key chapter since it sets the stage for the study of a broad range of inferential statistical techniques that follow. Chapter 4 treats general univariate probability distributions in considerable detail, and Chapter 5 does the same for general bivariate probability distributions. Chapters 6 and 7 introduce you to a variety of important specific discrete and continuous probability distributions, respectively. The bivariate normal distribution is also introduced in Chapter 7.

Chapter 8 exposes you to the concept of random sampling and the sampling distribution of a statistic (including those of the mean, proportion, and variance). Laws of large numbers and a Central Limit Theorem are carefully developed and explained. Chapter 9 deals with a set of derived distributions (chi-square, t, and F) and revisits the sampling distribution of the variance under the normality assumption.

Point estimation is the topic of Chapter 10. Small-sample as well as large-sample properties of point estimators are covered along with a variety of techniques for finding good point estimators. This is a critical chapter in that you are introduced to the Cramér-Rao lower bound, the Fisher-Neyman Factorization Theorem, and the theorems of Lechmann-Scheffé and Rao-Blackwell. The methods of least squares, maximum likelihood, and best linear unbiased estimation are fully explored. The method of moments technique is also addressed in the chapter exercises.

Chapter 11 introduces you to the construction of a variety of single-sample and two-sample confidence intervals. Independent populations as well as paired comparisons are considered. In addition, the joint estimation of a family of population parameters is conducted using the Bonferroni method.

Parametric statistical hypothesis testing is the topic of Chapter 12. Great care is taken to develop the preliminaries. That is, issues such as statistical hypothesis formulation, the research question, varieties of decision outcomes, errors in testing, devising tests, types of tests, and so on, are treated in detail before any actual testing is undertaken. Determining the best test for a statistical hypothesis, the power of a test, and generalized likelihood ratio tests are included to complete the hypothesis testing methodology. Various one-sample and two-sample hypothesis tests are conducted, where the latter involve both independent and dependent populations. Hypothesis tests for Spearman's rank correlation coefficient round out the chapter. Throughout all of the presentation, the appropriate reporting of hypothesis testing results is emphasized.

Chapter 13 involves a collection of nonparametric hypothesis tests. Here, both single-sample and two-sample tests are executed. Comparisons between parametric and non-parametric tests are frequently made as successive tests are developed.

Testing goodness of fit is the thrust of Chapters 14 and 15. Chapter 14 treats distributional hypotheses (via chi-square Kolmogorov-Smirnov, Lilliefors, and

Shapiro-Wilk procedures) and Chapter 15 employs contingency tables to test for independence, homogeneity, and uniformity among a set of proportions. Issues concerning strength of association are also explored.

Chapter 16 offers an extremely detailed discussion of bivariate regression and correlation. Topics treated include the assumptions underlying the strong vs. weak classical linear regression models, the Gauss-Markov Theorem, least squares estimation, hypothesis test for the population parameters, confidence bands, prediction, decomposition of the sample variation in the dependent variable, the correlation model, and inferences about the population correlation coefficient. Embellishments of the basic regression model (e.g., dummy variables, nonlinearities, etc.) are found in the chapter exercises.

Many individuals have helped to make this work possible. A debt of gratitude is owed to each of them. First and foremost is my wife, Paula, whose support, patience, and encouragement helped sustain me throughout all the writing and rewriting. I would also like to thank the Department of Economics, Finance and Insurance, the Barney School, and the University's Coffin Grant Committee for financial assistance. Additionally, I am grateful to Alice Schoenrock and to a whole host of graduate assistants, particularly Amlan Datta, who helped with many aspects of the preparation of the basic manuscript and accompanying tables. Special accolades go to Marilyn Baleshiski who expertly typed the final draft of the entire manuscript.

A special note of thanks is extended to the Editorial and Production Departments at Elsevier. Acquisition Editors Barbara Holland and Tom Singer, along with the Project Manager, Sarah Hajduk, made the entire publication process quite painless. Their kindness and professionalism are deeply appreciated. Furthermore, the following reviewers generously offered many valuable comments about the manuscript: John Travis, Mississippi College; Laura McSweeney, Fairfield University; Eric Slud, University of Maryland at College Park; Tom Short, Indiana University of Pennsylvania; Cristopher Mechlin, Murray State University; Pierre Grillet, Tulane University; and Mohammad Shakil, Miami Dade University. I appreciate their insight and constructive suggestions.

1 Introduction

1.1 Statistics Defined

Broadly defined, statistics encompasses the theory and methods of collecting, organizing, presenting, analyzing, and interpreting data sets so as to determine their essential characteristics. Although the collection, organization, and presentation of data will be addressed frequently throughout the text, primary emphasis will be placed upon the analysis of data and the interpretation of the results. Underlying the analysis of data is the vast mathematical apparatus of abstract concepts, theorems, formulae, algorithms, and so on that constitute the statistical tools that we will employ to study a data set. In this regard, our goal is to develop *a kit of tools* with which to analyze and interpret data. These tools will then enable us to build a framework for good decision making.

1.2 Types of Statistics

There are essentially two major categories of statistics: descriptive and inductive. *Descriptive statistics* includes any treatment of data designed to *summarize* their essential features. Here we are interested in arranging data in a readable form; for example, we may construct tables, charts, and graphs; and we can compute percents, averages, rates of change, and so on. In this regard, we do not go beyond the data at hand.

With *inductive statistics*, we are concerned with making estimates, predictions or forecasts, and generalizations. Since induction is the process of reasoning from the specific to the general, the essential characteristic of inductive statistics is termed *statistical inference*—the process of inferring something about *the whole* from an examination of only *the part*. Specifically, this process is carried out through *sampling*; that is, a representative subgroup of items is subject to study (the part) and the conclusions derived therefrom are assumed to characterize the entire group (the whole). Moreover, since an exhaustive *census* (or complete enumeration) of the whole is not being undertaken, so that our conclusions are hostage to the characteristics of those items actually comprising the part, some

level of error will most assuredly taint our conclusions about the whole. Hence we must accompany any generalization about the whole by a measure of the uncertainty of the inference made. Such measures can be calculated once the rudiments of probability theory are covered.

For instance, suppose we want to gain some insight into the relative popularity of a collection of candidates for the presidency of the United States. Can we realistically poll each and every individual of voting age (the whole)? Certainly not. But we can, using scientific polling techniques, elicit the preferences of only a small segment of all potential voters (the part). Hence, we may possibly conclude from this exercise that candidate A is the choice of 64% of *all* eligible voters. But this conclusion is not couched in absolute certainty. Some margin of error emerges since only a sample of individuals was taken. Hence we would accompany the 64% figure with a statement reading something like: the degree of precision of our estimate is ±3% with a 95% reliability level. The notions of *precision* and *reliability* will play a key role in the development of our inferential techniques.

In sum, if we only want to summarize or present data or just catalog facts, then we will use descriptive statistics. But if we want to make inferences based on sample data or make decisions in the face of uncertainty, then we must rely on inductive statistical methods.

1.3 Levels of Discourse: Sample vs. Population

Let us further elaborate on the concepts of *the whole* and *the part*. To set the stage for this discussion let us define an *experiment* as any process of observation. It may involve something like quality control inspection of electric motors or simply monitoring the flow of various types of motor vehicles along a particular street on a given day and over a given time interval. Next, a *variable* (denoted as X) can be thought of as any observable or measurable magnitude. Here X may depict an individual's set of exam scores, height (in inches) of sixth graders in a particular school system, weight (in pounds) of dressed turkeys, elapsed time (in minutes) taken to perform a certain task, and so on.

In this regard, let us define the sample (the part) as those things or events that both happened and were observed. It is drawn from the population (the whole), which involves things or events that happened but were not necessarily observed. Equivalently, we may think of the *population* as representing all conceivable observations on some variable X, whereas a *sample* is simply a subset of the population, that is, a given collection of observations taken from the population (see Figure 1.1). By convention, for a finite population, let us depict the population size by N and the sample size by n, where obviously $n \leq N$. Moreover, if X depicts a population variable, then the items within the population can be listed as $X : X_1, X_2, X_3, \ldots, X_N$ or as $X : X_i,\ i = 1, \ldots, N$. And if X represents a sample variable, then the collection of sample items can be written as $X : X_1, X_2, \ldots, X_n$ or $X : X_i,\ i = 1, \ldots, n$.

A sample is of finite size, but a population can be finite or infinite. That is, the collection of all individuals who have read part or all of this book can be thought

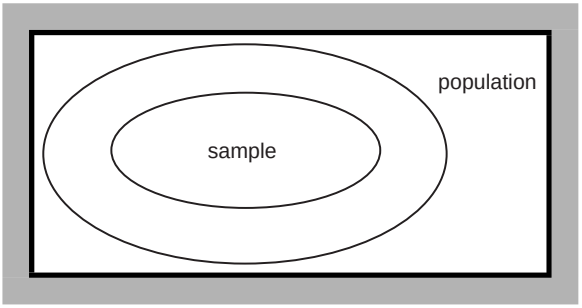


Figure 1.1 A sample as a subset of a population.

of as constituting a finite population, whereas an infinite population can easily be generated by sampling with replacement. For instance, if a population consists of N items and we *sample without replacement*, then the item obtained on any draw is set aside before the next item is chosen. But if we *sample with replacement* (on each draw we select an item from the population and then replace it before the next one is chosen), then clearly we can effectively sample in this fashion forever; that is, we can operate *as if* the population is infinite even though in reality it is not. So under sampling with replacement, a member of the population can appear more than once in the sample; in sampling without replacement, we disregard any item that has already been chosen for inclusion in the sample.

A few additional points merit our attention. First, it is important to mention that oftentimes inductive statistics is described as *decision making under uncertainty*. For our immediate purposes, note that an important source of uncertainty is the concept of *randomness*. That is, if the outcome of an experiment is not predictable, then it is said to occur in a random fashion. Next, we stated earlier that in sampling, a representative group of items is desired. In this regard, we may deem a sample as representative if it is *typical*, that is, it adequately reflects the attributes of the population. Moreover, we will frequently engage in the process of random sampling. Specifically, a sample is *random* if each and every item in the population has an equal (and known) chance of being included in the sample.

It is important to remember that the process of random sampling does not guarantee that a representative sample will be obtained. Randomization is likely but by no means certain to generate a representative sample. This is because randomization gives the same chance of selection to every sample of a given size—representative ones as well as nonrepresentative ones.

We end this section by noting that the term *parameter* (denoted θ) is used to represent any descriptive measure of a population, whereas a *statistic* (denoted $\hat{\theta}$) is any descriptive measure of a sample. Here $\hat{\theta}$ serves as an *estimator* of (the unknown) θ; that is, $\hat{\theta}$ is some function of the sample values used to discern the level of θ. The $\hat{\theta}$ actually obtained by the estimation process will be called an *estimate* of θ. If $\hat{\theta}$ represents a single numerical value, then it is termed a *point estimator* of θ. An *interval estimator* of θ enables us to state just how confident

we are, in terms of probability, that θ lies within some range of values. We shall return to these notions in Chapters 10 and 11.

1.4 Levels of Discourse: Target vs. Sampled Population

The *target population* is defined as the population to be studied; it is the population about which information is desired. This is in contrast to the *sampled population*—the population from which the sample is actually obtained. This latter population concept is alternatively called the *sampling frame*, or just *frame* for short. Based upon these two population notions, consequently we can describe a sample (or study for that matter) as being *valid* if the target and sampled populations have similar characteristics. (Note that some items in the target population may not be a member of the frame.)

Continuing in this vein, an *elementary sampling unit* is an item in the frame and an *observation* is a piece of information possessed by an elementary sampling unit. A *sample*, then, is that portion of the sampled population actually studied. (Note also that a sample may be representative—it adequately reflects the attributes of the frame—but not valid.) So under random sampling, each item in the frame has the same chance of being chosen.

For example, suppose we want to develop a profile of the membership of the local country club. This is the target population. The sampled population or frame is the club's membership list. If the list is up to date, then the target and sampled populations coincide. However, if the list has not been updated recently, then the target and sampled populations may be widely disparate and thus the question of validity comes to the fore. Elementary sampling units are the individual members, whereas an observation consists of a particular data point or value of some characteristic of interest. In what follows we shall depict each characteristic by a separate variable. Hence an observation is the value of the variable representing a characteristic of an elementary sampling unit.

For instance, assume that a country club has 3000 ($= N$) members and that we want a sample of size 100 ($= n$). If the membership list is arranged alphabetically (which we can assume to be a random arrangement), then we can easily engage in a sampling process called *systematic random sampling*, with a *sampling cycle* of $30 = 3000/100$ ($= N/n$). That is, we start at the beginning of the list, select one member at random from the first 30 members listed, and then pick from that item on every thirtieth number for inclusion in the sample.

Which characteristic of a member (our elementary sampling unit) might we be interested in? We can possibly list them as sex, number of years as a member, number of years service as a board member, type of membership (e.g., founding, charter, regular), annual gift amount, major activity preferred (e.g., golf, tennis, swimming, bridge), and so on. As indicated earlier, each characteristic is represented by a variable and the value of the variable is an observation of the characteristic, as listed in Table 1.1.

Table 1.1

Country Club Membership Characteristics		
Characteristic	**Variable**	**Observation Values**
Sex	X_1	Male (1); Female (2)
Years	X_2	Number of years
Board	X_3	Number of years
Type	X_4	Founding (3); Charter (2); Regular (1)
Gift	X_5	$
Activity	X_6	Golf (1); Tennis (2); Swimming (3); Bridge (4)

1.5 Measurement Scales

What types or varieties of data have we defined in the preceding example on country club membership characteristics? We may generally refer to *data* as a collection of facts, values, observations, or measurements. So if our data consists of observations that can be *measured* (i.e., classified, ordered, or quantified), then at what level does the measurement take place? Here we are interested in the *forms* in which data is found or the *scales* on which data is measured. These scales, stated in terms of increasing information content, are classified as nominal, ordinal, interval, and ratio.

Let us first consider the *nominal* scale. Here nominal should be associated with the word *name* since this scale identifies categories. Observations on a nominal scale possess neither numerical values nor order. However, observations on this type of scale can be given numerical codes such as "0 or 1" or "$1, 2, 3, \ldots$." Variables X_1 and X_6 in Table 1.1 are nominal in nature and are termed *qualitative* or *categorical* variables. Note that when dealing with a nominal scale, the categories defined must be *mutually exclusive* (each item falls into one and only one category) and *collectively exhaustive* (the list of categories is complete in that each item can be classified). Since X_1 has only two categories, it is called a binary or *dichotomous* variable. Note also that a number code has been used to classify the members as either 1 (male) or 2 (female). These numbers serve only as *identifiers*; the magnitude of the differences between these numerical values is meaningless. The only valid operations for variables represented by a nominal scale are the determination of "$=$" or "$\neq$."

The *ordinal* scale (think of the word *order*) includes all properties of the nominal scale with the additional property that the observations can be ranked from the smallest to the largest or from the least important to the most important. (Note that nominal measurements cannot be ordered—all items are treated equally.) If the country club mentioned earlier is a hierarchical organization wherein founding members are more important or ranked higher (in terms of privileges) than charter members, which are in turn ranked above regular members, then X_4 is an ordinal (and thus qualitative) variable. (Although chartered is in some sense *better* than regular, the ranking does not indicate *how much better*.) That is, since the numerical values assigned to X_4 are 3, 2, and 1, these numbers