MATHEMATICAL STATISTICS:

Basic Ideas and Selected Topics

MATHEMATICAL STATISTICS

Basic Ideas and Selected Topics

PETER J. BICKEL
University of California, Berkeley

KJELL A. DOKSUM
University of California, Berkeley

HOLDEN-DAY, INC.
San Francisco
Düsseldorf Johannesburg London
Panama Singapore Sydney

MATHEMATICAL STATISTICS

Library of Congress Catalog Card Number: 76-8724
ISBN: 0-8162-0784-4

Printed in the United States of America

1234567890 09876

PREFACE

This book presents our view of what an introduction to mathematical statistics for students with a good mathematics background should be. By a good mathematics background we mean linear algebra and matrix theory, and advanced calculus (but no measure theory). Since the book is an introduction to statistics we need probability theory and expect readers to have had a course at the level of, for instance, P. Hoel, S. Port and C. Stone's *Introduction to Probability Theory*. Our appendix does give all the probability that is needed. However, the treatment is abridged with few proofs and no examples or problems.

We feel such an introduction should at least do the following:

(1) Describe the basic concepts of mathematical statistics indicating the relation of theory to practice.

(2) Give careful proofs of the major "elementary" results such as the Neyman-Pearson lemma, the Lehmann-Scheffé theorem, the information inequality and the Gauss-Markoff theorem.

(3) Give heuristic discussions of more advanced results such as the large sample theory of maximum likelihood estimates, and the structure of both Bayes and admissible solutions in decision theory. The extent to which holes in the discussion can be patched and where patches can be found should be clearly indicated.

(4) Show how the ideas and results apply in a variety of important subfields such as Gaussian linear models, multinomial models, and nonparametric models.

Although there are several good books available for this purpose we feel that none has quite the mix of coverage and depth desirable at this level. The work of Rao, *Linear Statistical Inference and Its Applications*, 2nd ed., covers most of the material we do and much more but at a more abstract level employing measure theory. At the other end of the scale of difficulty for books at this level is the work of Hogg and Craig, *Introduction to Mathematical Statistics*, 3rd ed. These authors also discuss most of the topics we deal with but in many instances do not include detailed discussion of topics we consider essential such as existence and computation of procedures and large sample behavior.

Our book contains more material than can be covered in two quarters. In the two quarter courses for graduate students in mathematics, statistics, the physical sciences and engineering that we have taught we cover the core Chapters 2 to 7 which go from modelling through estimation and testing to linear models. In addition we feel Chapter 10 on decision theory is essential and cover at least the first two sections. Finally we select topics from Chapter 8 on discrete data and Chapter 9 on nonparametric models.

Chapter 1 covers probability theory rather than statistics. Much of this material unfortunately does not appear in basic probability texts but we need to draw on it for the rest of the book. It may be integrated with the material of Chapters 2–7 as the course proceeds rather than being given at the start; or it may be included at the end of an introductory probability course which precedes the statistics course.

A special feature of the book is its many problems. They range from trivial numerical exercises and elementary problems intended to familiarize the students with the concepts to material more difficult than that worked out in the text. They are included both as a check on the student's mastery of the material and as pointers to the wealth of ideas and results that for obvious reasons of space could not be put into the body of the text.

Conventions: (i) In order to minimize the number of footnotes we have added a section of comments at the end of each chapter preceding the problem section. These comments are ordered by the section to which they pertain. Within each section of the text the presence of comments at the end of the chapter is signalled by one or more numbers, 1 for the first, 2 for the second, etc. The comments contain digressions, reservations and additional references. They need to be read only as the reader's curiosity is piqued.

(ii) Various notational conventions and abbreviations are used in the text. A list of the most frequently occurring ones indicating where they are introduced is given at the end of the text.

(iii) Basic notation for probabilistic objects such as random variables and vectors, densities, distribtuion functions and moments is established in the Appendix.

We would like to acknowledge our indebtedness to colleagues, students, and friends who helped us during the various stages (notes, preliminary edition, final draft), through which this book passed. E. L. Lehmann's wise advise has played a decisive role at many points. R. Pyke's careful reading of a next-to-final version caught a number of infelicities of style and content. Many careless mistakes and typographical errors in an earlier version were caught by D. Minassian who sent us an exhaustive and helpful listing. W. Carmichael, in proofreading the final version, caught more mistakes than both authors together. A serious error in Problem 2.2.5 was discovered by F. Scholz. Among many others who helped in the same way we would like to mention C. Chen, S. J. Chou, G. Drew, C. Gray, U. Gupta, P. X. Quang, and A. Samulon. Without Winston Chow's lovely plots Section 9.6 would probably not have been written and without Julia Rubalcava's impeccable typing and tolerance this text would never have seen the light of day.

We would also like to thank the colleagues and friends who inspired and helped us to enter the field of statistics. The foundation of our statistical knowledge was obtained in the lucid, enthusiastic and stimulating lectures of Joe Hodges and Chuck Bell, respectively. Later, we were both very much influenced by Erich Lehmann whose ideas are strongly reflected in this book.

Last and most important we would like to thank our wives Nancy Kramer Bickel and Maria Delasalas Doksum and our families for support, encouragement, and active participation in an enterprise which at times seemed endless.

Peter J. Bickel
Kjell Doksum

Berkeley
1976

CONTENTS

CHAPTER 1
SOME TOPICS IN PROBABILITY

In this chapter we will give some results in probability theory, which are essential in our treatment of statistics and which may not be treated in enough detail in some probability texts.

Measure theory will not be used. We make the blanket assumption that all sets and functions considered are measurable.

1.1. CONDITIONING BY A RANDOM VARIABLE OR VECTOR

The concept of conditioning is important in studying associations between random variables or vectors. In this section we present some results useful for prediction theory, estimation theory, and regression.

1.1.A. The Discrete Case

The reader is already familiar with the notion of the conditional probability of an event A given that another event B has occurred. If $\mathbf{X}$ and $\mathbf{Y}$ are discrete random

vectors possibly of different dimensions we want to study the conditional probability structure of $\mathbf{X}$ given that $\mathbf{Y}$ has taken on a particular value $\mathbf{y}$.

Define the *conditional frequency function* $p(\cdot\,|\,\mathbf{y})$ *of* $\mathbf{X}$ *given* $\mathbf{Y} = \mathbf{y}$ by,

$$p(\mathbf{x}\,|\,\mathbf{y}) = P[\mathbf{X} = \mathbf{x}\,|\,\mathbf{Y} = \mathbf{y}] = \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} \tag{1.1.1}$$

where p and $p_{\mathbf{Y}}$ are the frequency functions of $(\mathbf{X}, \mathbf{Y})$ and $\mathbf{Y}$. The conditional frequency function p is defined only for values of $\mathbf{y}$ such that $p_{\mathbf{Y}}(\mathbf{y}) > 0$. With this definition it is clear that $p(\cdot\,|\,\mathbf{y})$ is the frequency function of a probability distribution, since

$$\Sigma_{\mathbf{x}} p(\mathbf{x}\,|\,\mathbf{y}) = \frac{\Sigma_{\mathbf{x}} p(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} = \frac{p_{\mathbf{Y}}(\mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} = 1$$

by (A.8.11). This probability distribution is called the *conditional distribution of* $\mathbf{X}$ *given that* $\mathbf{Y} = \mathbf{y}$.

Example 1.1.1. Let $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are the indicators of a set of n binomial trials with probability of success p. Let $Y = \Sigma_{i=1}^n X_i$, the total number of successes. Then Y has a binomial, $\mathcal{B}(n, p)$, distribution and

$$p(\mathbf{x}\,|\,y) = \frac{P[\mathbf{X} = \mathbf{x},\ Y = y]}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}} \tag{1.1.2}$$

if the x_i are all 0 or 1 and $\Sigma x_i = y$.

Thus, if we are told we obtained k successes in n binomial trials, then these successes are as likely to occur on one set of trials as on any other. ■

Example 1.1.2. Let X and Y have the joint frequency function given by the table

TABLE 1.1.1

x \ y	0	10	20	$p_X(x)$
0	0.25	0.05	0.05	0.35
1	0.05	0.15	0.05	0.25
2	0.05	0.10	0.25	0.40
$p_Y(y)$	0.35	0.30	0.35	1

For instance, suppose Y is the number of cigarettes that a person picked at random from a certain population smokes per day (to the nearest 10), and X is a general

health rating for the same person with 0 corresponding to good, 2 to poor, and 1 to neither. We find for $y = 20$

x	0	1	2
$p(x\|20)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{5}{7}$

These figures would indicate an association between heavy smoking and poor health, since $p(2\,|\,20)$ is almost twice as large as $p_X(2)$. ■

The conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is easy to calculate in two special cases.

(i) If $\mathbf{X}$ and $\mathbf{Y}$ are independent $p(\mathbf{x}\,|\,\mathbf{y}) = p_{\mathbf{X}}(\mathbf{x})$ and the conditional distribution coincides with the marginal distribution.

(ii) If $\mathbf{X}$ is a function of $\mathbf{Y}$, $h(\mathbf{Y})$, then the conditional distribution of $\mathbf{X}$ is degenerate, $\mathbf{X} = h(\mathbf{y})$ with probability 1.

Both of these assertions follow immediately from Definition (1.1.1).

Two important formulae follow from (1.1.1) and (A.4.5). Let $q(\mathbf{y}\,|\,\mathbf{x})$ denote the conditional frequency function of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Then,

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}\,|\,\mathbf{y})p_{\mathbf{Y}}(\mathbf{y}) \tag{1.1.3}$$

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{q(\mathbf{y}\,|\,\mathbf{x})p_{\mathbf{X}}(\mathbf{x})}{\Sigma_{\mathbf{z}} q(\mathbf{y}\,|\,\mathbf{z})p_{\mathbf{X}}(\mathbf{z})} \qquad \text{Bayes' Rule} \tag{1.1.4}$$

whenever the denominator of the right-hand side is positive.

Equation (1.1.3) can be used for model construction. For instance, suppose that the number Y of defectives in a lot of N produced by a manufacturing process has a $\mathscr{B}(N, \theta)$ distribution. Suppose the lot is sampled n times without replacement and let X be the number of defectives found in the sample. We know that given $Y = y$, X has a hypergeometric, $\mathscr{H}(y, N, n)$, distribution. We can now use (1.1.3) to write down the joint distribution of X and Y

$$P[X = x,\ Y = y] = \binom{N}{y}\theta^y(1-\theta)^{n-y}\frac{\binom{y}{x}\binom{N-y}{n-x}}{\binom{N}{n}} \tag{1.1.5}$$

where the combinatorial coefficients $\binom{a}{b}$ vanish unless a, b are integers with $b \leq a$.

We can also use this model to illustrate (1.1.4). Since we would usually only observe X, we may want to know what the conditional distribution of Y given $X = x$ is. By (1.1.4) this is,

$$P[Y = y\,|\,X = x] = \binom{N}{y}\theta^y(1-\theta)^{n-y}\binom{y}{x}\binom{N-y}{n-x}/c(x)$$

where $c(x) = \Sigma_y \binom{N}{y}\theta^y(1-\theta)^{n-y}\binom{y}{x}\binom{N-y}{n-x}$. This formula simplifies to (see Problem 1.1.11) the binomial probability,

$$P[Y = y \mid X = x] = \binom{N-n}{y-x}\theta^{y-x}(1-\theta)^{N-n-(y-x)}. \tag{1.1.6}$$

1.1.B. Conditional Expectation for Discrete Variables

Suppose that X is a random variable with $E(|X|) < \infty$. Define the *conditional expectation of X given* $\mathbf{Y} = \mathbf{y}$, written $E(X \mid \mathbf{Y} = \mathbf{y})$, by

$$E(X \mid \mathbf{Y} = \mathbf{y}) = \Sigma_x x p(x \mid \mathbf{y}). \tag{1.1.7}$$

Note that by (1.1.1), if $p_{\mathbf{Y}}(\mathbf{y}) > 0$,

$$\Sigma_x |x| p(x \mid \mathbf{y}) \leqq \Sigma_x |x| \frac{p_X(x)}{p_{\mathbf{Y}}(\mathbf{y})} = \frac{E(|X|)}{p_{\mathbf{Y}}(\mathbf{y})}. \tag{1.1.8}$$

Example 1.1.3. Suppose X and Y have the joint frequency function of Table 1.1.1 above. We find

$$E(X \mid Y = 20) = 0 \cdot \tfrac{1}{7} + 1 \cdot \tfrac{1}{7} + 2 \cdot \tfrac{5}{7} = \tfrac{11}{7} = 1.57.$$

Similarly, $E(X \mid Y = 10) = \frac{7}{6} = 1.17$ and $E(X \mid Y = 0) = \frac{3}{7} = 0.43$. Note that in the health versus smoking context, we can think of $E(X \mid Y = y)$ as the mean health rating for people who smoke y cigarettes a day. ■

Let $g(\mathbf{y}) = E(X \mid \mathbf{Y} = \mathbf{y})$. The random variable $g(\mathbf{Y})$ is written $E(X \mid \mathbf{Y})$ and is called the *conditional expectation of X given* $\mathbf{Y}$.†

As an example we calculate $E(X_1 \mid Y)$ where X_1 and Y are given in Example 1.1.1. We have,

$$E(X_1 \mid Y = i) = P[X_1 = 1 \mid Y = i] = \frac{\binom{n-1}{i-1}}{\binom{n}{i}} = \frac{i}{n}. \tag{1.1.9}$$

The first of these equalities holds because X_1 is an indicator. The second follows from (1.1.2), since $\binom{n-1}{i-1}$ is just the number of ways i successes can occur in n trials with the first trial being a success. Therefore,

$$E(X_1 \mid Y) = \frac{Y}{n}. \tag{1.1.10}$$

The conditional distribution of a random vector $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ corresponds

†We shall follow the convention of also calling $E(X \mid \mathbf{Y})$ any variable which is equal to $g(\mathbf{Y})$ with probability 1.

to a single probability measure P_y on $(\Omega, \mathcal{A})$. Specifically, define for $A \in \mathcal{A}$,

$$P_y(A) = P(A|[\mathbf{Y} = \mathbf{y}]) \text{ if } p_{\mathbf{Y}}(\mathbf{y}) > 0. \tag{1.1.11}$$

This P_y is just the conditional probability measure on $(\Omega, \mathcal{A})$ mentioned in (A.4.2). Now the conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is the same as the distribution of $\mathbf{X}$ if P_y is the probability measure on $(\Omega, \mathcal{A})$, Therefore, the conditional expectation *is* an ordinary expectation with respect to the probability measure P_y. It follows that all the properties of the expectation given in (A.10.3)–(A.10.8) hold for the conditional expectation given $\mathbf{Y} = \mathbf{y}$. For example,

$$E(\alpha X_1 + \beta X_2|\mathbf{Y} = \mathbf{y}) = \alpha E(X_1|\mathbf{Y} = \mathbf{y}) + \beta E(X_2|\mathbf{Y} = \mathbf{y}) \tag{1.1.12}$$

identically in y for any X_1, X_2 such that $E(|X_1|)$, $E(|X_2|)$ are finite. Since the identity holds for all y we have,

$$E(\alpha X_1 + \beta X_2|\mathbf{Y}) = \alpha E(X_1|\mathbf{Y}) + \beta E(X_2|\mathbf{Y}). \tag{1.1.13}$$

This process can be repeated for each of (A.10.3)–(A.10.8) to obtain analogous properties of the conditional expectation. An important example of this type of argument is given in Section 1.6.

In two special cases we can calculate conditional expectations immediately. If X and $\mathbf{Y}$ are independent and $E(|X|) < \infty$, then

$$E(X|\mathbf{Y}) = E(X). \tag{1.1.14}$$

This is clear by (i).

On the other hand by (ii)

$$E(h(\mathbf{Y})|\mathbf{Y}) = h(\mathbf{Y}). \tag{1.1.15}$$

The notion implicit in (1.1.15) is that given $\mathbf{Y} = \mathbf{y}$, $\mathbf{Y}$ acts as a constant. If we carry this further we have a relation which we shall call the *substitution theorem for conditional expectations*:

$$E(q(\mathbf{X}, \mathbf{Y})|\mathbf{Y} = \mathbf{y}) = E(q(\mathbf{X}, \mathbf{y})|\mathbf{Y} = \mathbf{y}), \tag{1.1.16}$$

for all y such that $p_{\mathbf{Y}}(\mathbf{y}) > 0$ and $q(\mathbf{X}, \mathbf{Y})$ has a finite expectation. Assertion (1.1.16) is immediate, since

(1.1.17)

$$P[q(\mathbf{X}, \mathbf{Y}) = a|\mathbf{Y} = \mathbf{y}] = P[q(\mathbf{X}, \mathbf{Y}) = a, \mathbf{Y} = \mathbf{y}|\mathbf{Y} = \mathbf{y}] = P[q(\mathbf{X}, \mathbf{y}) = a|\mathbf{Y} = \mathbf{y}]$$

for any a.

If we put $q(\mathbf{X}, \mathbf{Y}) = r(\mathbf{X})h(\mathbf{Y})$, where h is bounded and $r(\mathbf{X})$ has a finite expectation, we obtain by (1.1.16),

$$E(r(\mathbf{X})h(\mathbf{Y})|\mathbf{Y} = \mathbf{y}) = E(r(\mathbf{X})h(\mathbf{y})|\mathbf{Y} = \mathbf{y}) = h(\mathbf{y})E(r(\mathbf{X})|\mathbf{Y} = \mathbf{y}). \tag{1.1.18}$$

Therefore,

$$E(r(\mathbf{X})h(\mathbf{Y})|\mathbf{Y}) = h(\mathbf{Y})E(r(\mathbf{X})|\mathbf{Y}). \tag{1.1.19}$$

Another intuitively reasonable result is that the mean of the conditional means is the mean:

$$E(E(X|\mathbf{Y})) = E(X), \tag{1.1.20}$$

whenever X has a finite expectation. We refer to this as the *double expectation theorem.*

To prove (1.1.20) we write, in view of (1.1.7) and (A.10.5),

$$\begin{aligned} E(E(X|\mathbf{Y})) &= \Sigma_{\mathbf{y}} p_{\mathbf{Y}}(\mathbf{y})[\Sigma_x x p(x|\mathbf{y})] \\ &= \Sigma_{x,\mathbf{y}} x p(x|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) \\ &= \Sigma_{x,\mathbf{y}} x p(x,\mathbf{y}) = E(X). \end{aligned} \tag{1.1.21}$$

The interchange of summation used is valid, since the finiteness of $E(|X|)$ implies that all sums converge absolutely.

As an illustration, we check (1.1.20) for $E(X_1|Y)$ given by (1.1.10). In this case,

$$E(E(X_1|Y)) = E\left(\frac{Y}{n}\right) = \frac{np}{n} = p = E(X_1). \tag{1.1.22}$$

If we apply (1.1.20) to $X = r(\mathbf{X})h(\mathbf{Y})$ and use (1.1.19), we obtain the *product expectation formula:*

Theorem 1.1.1. If $h(\mathbf{Y})$ is bounded and $E(|r(\mathbf{X})|) < \infty$, then

$$E(r(\mathbf{X})h(\mathbf{Y})) = E(h(\mathbf{Y})E(r(\mathbf{X})|\mathbf{Y})). \tag{1.1.23}$$

Note that we can express the conditional probability that $\mathbf{X} \in A$ given $\mathbf{Y} = \mathbf{y}$ as

$$P[\mathbf{X} \in A|\mathbf{Y} = \mathbf{y}] = E(I_A(\mathbf{X})|\mathbf{Y} = \mathbf{y}) = \Sigma_{\mathbf{x}\in A} p(\mathbf{x}|\mathbf{y}).$$

Then by taking $r(\mathbf{X}) = I_A(\mathbf{X})$, $h = 1$ in Theorem 1.1.1 we can express the (unconditional) probability that $\mathbf{X} \in A$ as

$$P[\mathbf{X} \in A] = E(E(r(\mathbf{X})|\mathbf{Y})) = \Sigma_{\mathbf{y}} P[\mathbf{X} \in A|\mathbf{Y} = \mathbf{y}] p_{\mathbf{Y}}(\mathbf{y}). \tag{1.1.24}$$

For example, if X and Y are as in (1.1.5),

$$P[X \leqq x] = \Sigma_y \binom{N}{y} \theta^y (1-\theta)^{n-y} H_y(x)$$

where H_y is the distribution function of a hypergeometric distribution with parameters (y, N, n).