Jack Carl Kiefer

Introduction to Statistical Inference

Springer-Verlag

World Publishing Corporation

Jack Carl Kiefer

Introduction to Statistical Inference

Edited by
Gary Lorden

With 60 Illustrations

Springer-Verlag
World Publishing Corporation

Gary Lorden
California Institute of
Technology
Pasadena, CA 91125

Reprinted by World Publishing Corporation, Beijing, 1989
for distribution and sale in The People's Republic of China only

AMS Classification: 62–01, 62FXX

Library of Congress Cataloging in Publication Data
Kiefer, Jack
Introduction to statistical inference.
(Springer texts in statistics)
Includes index.
1. Mathematical statistics. I. Lorden, Gary.
II. Title. III. Series.
QA276.16.K54 1987 519.5 86-31638

ISBN 0-387-96420-7 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-96420-7 Springer-Verlag Berlin Heidelberg New York
ISBN 7-5062-0292-1

Preface

This book is based upon lecture notes developed by Jack Kiefer for a course in statistical inference he taught at Cornell University. The notes were distributed to the class in lieu of a textbook, and the problems were used for homework assignments. Relying only on modest prerequisites of probability theory and calculus, Kiefer's approach to a first course in statistics is to present the central ideas of the modern mathematical theory with a minimum of fuss and formality. He is able to do this by using a rich mixture of examples, pictures, and mathematical derivations to complement a clear and logical discussion of the important ideas in plain English.

The straightforwardness of Kiefer's presentation is remarkable in view of the sophistication and depth of his examination of the major theme: How should an intelligent person formulate a statistical problem and choose a statistical procedure to apply to it? Kiefer's view, in the same spirit as Neyman and Wald, is that one should try to assess the consequences of a statistical choice in some quantitative (frequentist) formulation and ought to choose a course of action that is verifiably optimal (or nearly so) without regard to the perceived "attractiveness" of certain dogmas and methods.

Besides the centrally important decision-theoretic concepts such as admissibility and Bayes and minimax procedures, Kiefer is concerned with providing a rational perspective on classical approaches, such as unbiasedness, likelihood-based methods, and linear models. He also gives thorough consideration to useful ways of reducing and simplifying statistical problems, including sufficiency, invariance (equivariance), and asymptotics. Many of these topics are taken up in Chapter 4, which provides both an overview and a foundation for developments in later chapters, especially Chapters 6 and 7 on sufficiency and estimation, respectively. Chapter 5, on linear unbiased estimation, as well as Chapters 8 and 9, on hypothesis testing and confidence intervals, cover quite a bit of "standard statistical method"; although many insights and helpful ideas are provided, beginners who want to learn the usual techniques of Student's t-tests, F-tests, and analysis of variance will likely find the treatment too brief. Supplementary use of a standard text is recommended for this purpose. The first three chapters are short and introductory; Chapter 1, in particular, gives a clear and challenging view of what Kiefer has in mind. Appendix A lists some notational conventions.

Following Jack Kiefer's death in 1981, many of his colleagues, friends, and former students were eager to see his course notes published as a book. The statistical community owes its thanks to the late Walter Kaufmann-Bühler, of Springer-Verlag, for pursuing the project, and to Jack's wife, Dooley, for agreeing to it. In editing the notes, I have made only a few minor changes in the narrative, concentrating instead on correcting small errors and ensuring the internal consistency of the language, notation, and the many cross-references. The result, I believe, fully preserves the style and intentions of Jack Kiefer, who communicated his persistent idealism about the theory and practice of statistics with characteristic warmth and enthusiasm to generations of students and colleagues. May this book inspire future generations in the same way that Jack's teaching inspired all of us.

Gary Lorden
Pasadena, California

Contents

CHAPTER 1

Introduction to Statistical Inference

A typical problem in probability theory is of the following form: A sample space and underlying probability function are specified, and we are asked to compute the probability of a given chance event. For example, if $X_1, \ldots, X_n$ are independent Bernoulli random variables with $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p$, we compute that the probability of the chance event $\left\{\sum_{i=1}^{n} X_i = r\right\}$, where r is an integer with $0 \le r \le n$, is $\binom{n}{r} p^r (1-p)^{n-r}$.

In a typical problem of statistics it is not a single underlying probability law which is specified, but rather a *class* of laws, any of which may *possibly* be the one which actually governs the chance device or experiment whose outcome we shall observe. We know that the underlying probability law is a member of this class, but we do not know which one it is. The object might then be to determine a "good" way of guessing, on the basis of the outcome of the experiment, which of the *possible* underlying probability laws is the *one* which actually governs the experiment whose outcome we are to observe.

EXAMPLE 1.1. We are given a coin about which we know nothing. We are allowed to perform 10 independent flips with the coin, on each of which the probability of getting a head is p. We do not know p, but only know that $0 \le p \le 1$. In fact, our object might be, on the basis of what we observe on the 10 flips, to guess what value of p actually applies to this particular coin. We might be professional gamblers who intend to use this coin in our professional capacity and who would clearly benefit from knowing whether the coin is fair or is strongly biased toward heads or toward tails. The outcome of the experiment cannot tell us with complete sureness the exact value of p. But we can try to make an accurate guess on the basis of the 10 flips we observe.

There are many ways to form such a guess. For example, we might compute [number of heads obtained in 10 tosses]/10 and use this as our guess. Instead, we might observe how many flips it took to get the first head, guessing p to be 0 if no head comes up in 10 tosses, and to be 1/[number of toss on which first head appears] if at least one head occurs. A third possibility is to guess p to be $\pi/8$ or 2/3, depending on whether the number of heads in 10 tosses is odd or even. Fourth, we might ignore the experiment entirely and always guess p to be 1/2. You can think of many other ways to make a guess based on the outcome of the experiment.

Which of these ways of forming a guess from the experimental data should we actually employ? The first method suggested may seem reasonable to you. Perhaps the second one does not seem very unreasonable either, since one would intuitively expect to get the first head soon if p were near 1 and later (or not at all) if p were near 0. The third method probably seems unreasonable to you; a *large* or *small* number of heads might indicate that p is near 1 or 0, respectively, but the knowledge only that the number of heads is *even* seems to say very little about p. The fourth method, which makes no use of the data, may also seem foolish.

Having formed some intuitive judgment as to the advisability of using any of these four guessing schemes, you may be asked to *prove that method* 1 *is better than method* 3. You will probably be unable even to get started on a proof, no matter how many of the classical works and elementary textbooks on statistics you consult. Yet, all of these books will tell you to use method 1, and in fact you will probably not find a reference to *any* other method! At best you will find some mention of a "principle of unbiased estimation" or a "principle of maximum likelihood" (don't worry now about what these mean) to justify the use of method 1, but this only transfers the inadequacy of the discussion to another place, since you will not find any real justification of the use of these "principles" in obtaining guessing methods. In fact, you will not even find any satisfactory discussion of what is meant by a "good" method, let alone a proof that method 1 really *is* good.

Surely it should be possible to make precise what is meant by a "good" guessing method and to determine which methods are good. It is possible. Such a rational approach to the subject of statistical inference (that is, to the subject of obtaining good guessing methods) came into being with the work of Neyman in the 1930s and of Wald in the 1940s. Almost all of the development of guessing methods before that time, and even a large amount of it in recent years, proceeded on an intuitive basis. People suggested the use of methods in many probabilistic settings because of some intuitive appeal of the methods. We shall see that such intuitive appeal can easily lead to the use of very bad guessing methods.

It is unfortunate that the modern approach to statistical inference is generally ignored in elementary textbooks. Books and articles developed in the modern spirit are ordinarily written at a slightly more advanced mathematical level. Nevertheless, it is possible to discuss all of the important ideas of modern statistical inference at an accessible mathematical level, and we shall try to do this. First we must describe the way in which a statistics problem is specified. This description can be made precise in a way that resembles the precise formulation of probability problems in terms of an underlying sample space, probability function, and chance events whose probabilities are to be computed.

CHAPTER 2

Specification of a Statistical Problem

We shall begin by giving a formulation of statistics problems which is general enough to cover many cases of practical importance. Modifications will be introduced into this formulation later, as they are needed.

We suppose that there is an experiment whose outcome the statistician can observe. This outcome is described by a random variable or random vector X taking on values in a space S. The probability law of X is unknown to the statistician, but it is known that the distribution function (df) F of X is a member of a specified class Ω of df's. Sometimes we shall find it convenient to think of Ω not as a collection of df's F but rather as a collection of real or vector parameter values θ, where the form of each possible F is specified completely in terms of θ, so that a knowledge of the value taken on by the label θ is equivalent to a knowledge of F. The members of Ω are sometimes referred to as the *possible states of nature*. To emphasize which of the *possible* values of F or θ is the unknown one which actually *does* govern the experiment at hand, this value is sometimes referred to as the "*true* value of F or θ," or the "*true* state of nature."

EXAMPLE 2.1(a). In Example 1.1 (Chapter 1), we can let $X = (X_1, X_2, \ldots, X_{10})$, where X_i is 1 or 0 according to whether the i^{th} toss is a head or a tail. S consists of the 2^{10} possible values of X. The class Ω of possible df's (or probability laws) of X consists of all df's for which the X_i are independently and identically distributed with $P(X_i = 1) = \theta = 1 - P(X_i = 0)$, where $0 \le \theta \le 1$ (θ was called p in Example 1.1). It is convenient to think of these df's as being labeled by θ. It is easier to say "$\theta = 1/3$" than to say "the X_1 are independently and identically distributed with $P\{X_1 = 1\} = 1 - P\{X_1 = 0\} = 1/3$." Of course, the description "$\theta = 1/3$" can be used *only* because we have carefully specified the meaning of θ and the precise way in which it determines F.

It is implied in our discussion of specifying a statistical problem that by conducting an experiment the statistician can obtain some information about the "actual state of nature" F in Ω. It can often be shown by using the law of large numbers or other asymptotic methods (like those discussed later in Chapter 7) that if one could repeat the experiment an infinite number of times (independent replications), then one could discover F exactly. In reality the statistician performs only a finite number of experiments. The more experiments made, the better is the approximation to performing an infinite number of experiments.

It is frequently the case that statisticians or experimenters proceed as follows. We decide on an experiment and specify the statistical problem (as described later) in terms of the experiment. We then decide whether the results that would be obtained would be accurate enough. If not, we decide on a new experiment (often involving only more replications, like flipping the coin more) and work out the specification of a new statistical problem.

The behavior, just described, of a statistician, is often called *designing an experiment*. It is an aspect of statistics which we will not consider further. We will in the sequel be mostly concerned with questions of how best to use the results of experiments already decided upon.

Returning to our general considerations, we must introduce some notation. We will want to compute various probabilities and expectations involving X and must somehow make precise which member of Ω is to be used in this computation. Thus, in our Example 1.1, if we say, "the probability that $\Sigma X_i = 4$ is .15," under which of the possible underlying probability laws of X has this probability been computed? To eliminate confusion, we shall in general write

$$P_{F_0}\{A\} \quad \text{or} \quad P_{\theta_0}\{A\}$$

to mean "the probability of the chance event A, computed when the underlying probability law of X is given by $F = F_0$ (or by $\theta = \theta_0$)". Thus, in our Example 1.1 we would write

$$P_\theta\left\{\sum_{i=1}^{10} X_i = 4\right\} = \binom{10}{4}\theta^4(1-\theta)^6.$$

We shall extend the use of this notation to write

$$p_{F_0;Y} \quad \text{or} \quad p_{\theta_0;Y}$$

or

$$f_{F_0;Y} \quad \text{or} \quad f_{\theta_0;Y}$$

for the probability mass or density function of a random variable Y (which has been defined as a function of X) when the probability law of X is given by $F = F_0$ (or $\theta = \theta_0$). Thus, in the preceding example, if $Y = \sum_{i=1}^{10} X_i$, we would write

$$p_{\theta;Y}(4) = \binom{10}{4}\theta^4(1-\theta)^6.$$

We shall carry this notation over to expectations, writing

$$E_{F_0}g(X) \quad \text{or} \quad E_{\theta_0}g(X)$$

and

$$\text{var}_{F_0}\, g(X) \quad \text{or} \quad \text{var}_{\theta_0}\, g(X)$$

for the expectation and variance of $g(X)$ when the probability law of X is given by $F = F_0$ (or $\theta = \theta_0$). In our example, we would thus write

$$E_\theta Y = 10\theta, \qquad \text{var}_\theta Y = 10\theta(1-\theta).$$

We now describe the next aspect of a statistics problem which must be specified. This is the collection of possible actions which the statistician can take, or of possible statements which can be made, at the conclusion of the experiment. We denote this collection by D, the *decision space*, its elements being called *decisions*. At the conclusion of the experiment the statistician actually only chooses *one* decision (takes one action, or makes one statement) out of the possible choices in D. The statistician must make such a decision.

EXAMPLE 2.1(a) (continued). In our previous example, the statistician was required to guess the value of θ. In this case, we can think of D as the set of real numbers d satisfying $0 \le d \le 1$:

$$D = \{d: 0 \le d \le 1\}.$$

Thus, the decision ".37" stands for the statement "my guess is that θ is .37."

EXAMPLE 2.1(b). Suppose our gambler does not want to have a numerical guess as to the value of θ but only requires to know whether the coin is fair, is biased toward heads, or is biased toward tails. In this case the space D consists of three elements,

$$D = \{d_1, d_2, d_3\},$$

where d_1 stands for "the coin is fair," d_2 stands for "the coin is biased toward heads," and d_3 stands for "the coin is biased toward tails."

Note that in both of these examples, D can be viewed as the collection of possible answers to a question that is asked of the statistician. ("What do you guess θ to be?" "Is the coin fair, biased toward heads, or biased toward tails?")

EXAMPLE 2.1(c). An even simpler question would be "Is the coin fair?" We would have

$$D = \{d_1, d_2\},$$

where d_1 means "yes (the coin is fair)" and d_2 means "no (the coin is biased)."

To see how D might be regarded as the collection of possible *actions* rather than *statements*, suppose the U.S. Mint attempted to discourage gamblers from unfairly using its coins, by issuing only coins which were judged to be fair. In this case, the Mint's statistician, after experimenting with a given coin, would either throw it into a barrel to be shipped to a bank (d_1) or throw it back into the melting pot (d_2). Actually, we shall see that, in the general theoretical approach to statistics, the practical meaning of the various elements of D as statements or actions does not affect the development at all: once S, Ω, D, and the function W (to be described later) have been specified, we would reach the same conclusions for any interpretation of the physical meaning of the elements of D. In two different applications for both of which a given specification of S, Ω, D, and W is appropriate, we will reach the same conclusion regarding what statistical procedure ("guessing method") to use, regardless of what the applications are. The same statistical technique would be used in conjunction with agronomical or astronomical work, provided the possible outcomes of the experiment (S), possible probability laws (Ω), mathematical representation of the space of possible decisions (D), and loss function (W) are the same in the two settings. The fact that d_1 stands for one physical statement in the agronomical application and for another in the astronomical setting does not matter.

EXAMPLE 2.1(d). Suppose that our gambler does not merely want a guess as to the value of θ which governs the coin at hand, but rather a statement of an interval of values which is thought to include θ. In this case we can think of each element d of D as being an interval of real numbers from d' to d'' inclusive, where $0 \leq d' \leq d'' \leq 1$, so that D is the set of all such intervals; or, equivalently, we can think of each d as being the ordered pair (d', d'') of endpoints of an interval. In either case, the decision $d = (d', d'')$ stands for the statement "My guess is that $d' \leq \theta \leq d''$." A guess in this form, such as "I guess that $.35 \leq \theta \leq .4$," may not seem to be so precise as the statement "I guess that $\theta = .37$," the latter form being that encountered in Example 2.1(a). We shall see later that in practice it is often more desirable to give a guess in the form of the present example than in the form of Example 2.1(a), because of the possibility of attaching to an interval of values a simple measure of confidence in the statement that the underlying value θ does fall in the stated interval.

In order to specify a statistical problem completely, we must state precisely how right or wrong the various possible decisions are for each possible underlying probability law F (or θ) of X. (We repeat that, in order to emphasize that we are speaking of that unknown value of θ, among all possible values of θ, which describes the distribution of X in the *particular experiment now at hand*, books and articles often refer to this value as the *true value of* θ. This true value may of course change when we move on to another experiment for which S and Ω are exactly the same—for example, when we look at another coin in Example 2.1.)

For example, in Example 2.1(c) it may be that, if the true value of θ is fairly close to .5, say $.495 < \theta < .505$, then we deem it correct to make decision d_1 ("the coin is fair") and incorrect to make decision d_2, since if $|\theta - .5| < .005$ the coin is judged to be close enough to fair for us to call it such. On the other hand, if $|\theta - .5| \geq .005$, we may feel that d_2 is to be thought of as correct and d_1 as incorrect. If we judge that making a correct decision causes us to incur no loss, whereas any incorrect decision causes us to incur the same positive loss L as any other incorrect decision, we can write down analytically the loss encountered if the true distribution of X is given by θ and if we make decision d, as follows:

$$W(\theta, d_1) = \begin{cases} 0 & \text{if } |\theta - .5| < .005, \\ L & \text{otherwise;} \end{cases}$$

$$W(\theta, d_2) = \begin{cases} L & \text{if } |\theta - .5| < .005, \\ 0 & \text{otherwise,} \end{cases}$$

Similarly, in any statistics problem we define $W(F, d)$ (or $W(\theta, d)$) to be the *loss incurred* if the *true distribution of X is given by F (or θ) and the statistician makes decision d*. The function W is one of the data of the problem; the loss $W(F, d)$ (or $W(\theta, d)$) which will be encountered if F (or θ) is the true distribution and decision d is made must be stated precisely for every possible F (or θ) in Ω and every possible decision d in D. The reason for this is that, as we shall see later, the choice of a good statistical procedure (guessing rule) depends on W. A procedure which is good for one loss function W may be very poor for another loss function which might be appropriate in other circumstances.

How do we obtain the loss function? In practice it is often difficult to judge the relative seriousness of the various possible incorrect decisions one might make for each possible true situation, so as to make W precise. In some settings W can be written down in units of dollars on the basis of economic considerations, but in other settings, particularly in basic scientific research, the losses (which might be in units of "utility") incurred by making various decisions are difficult to make precise. For example, if an astronomer makes several measurements (which include chance errors) on the basis of which he wants to publish a guess of the distance to a certain galaxy, what is the seriousness of misguessing the distance by 5 light-years, as compared to misguessing the distance by only 1 light-year?

Practically speaking, the essential fact is that there are many important practical settings where we can find *one* statistical procedure which is fairly good for any of a variety of Ws whose forms are somewhat similar. From this we can conclude that it may not be so serious if the astronomer's judgment as to the values of losses due to making various possible decisions is not exactly what it should be (that is, what it would be if the astronomer could foresee all possible future developments which might result from publishing this distance, such as winning the Nobel Prize, or getting fired, or the United States sending the first dog to a star, and could determine from these the disutilities of all bad guesses). For as long as the loss function the astronomer writes down has roughly the same shape and tendencies as the loss function he or she should be using, a statistical procedure which we determine to be good using the astronomer's loss function will also be good for the actual loss function he or she should be using. We will indicate what this means in an example:

EXAMPLE 2.2. Suppose the astronomer's measurements are independent random variables, each with normal $\mathcal{N}(\mu, \sigma^2)$ distribution, mean μ and variance σ^2, where μ is the actual distance to the star. If this distance is known to be at least h light-years but nothing else is known, we can write $\theta = (\mu, \sigma^2)$ and

$$\Omega = \{(\mu, \sigma^2) : \mu \geq h, \sigma > 0\}.$$

The decision, as in Example 1.1(a), is a real number:

$$D = \{d : d \geq h\},$$

the decision $d = 100$, meaning "I guess the distance μ to be 100 light-years." The loss function would probably reflect the fact that the larger the amount by which the astronomer misguesses μ, the larger the loss incurred. For example, we might have

$$W((\mu, \sigma^2), d) = |\mu - d|$$

or

$$W((\mu, \sigma^2), d) = (\mu - d)^2,$$

and you can think of many other forms of W which exhibit the same tendencies. The important practical fact is that *a certain statistical procedure (guessing method) which is good assuming one of these forms of* W *will also be good for many other* W's *of similar form.* We shall discuss this fact in detail later on.

Thus, in practice it will often suffice to make a rough judgment as to the form of W. *Such a rough judgment, and the resulting determination of a good statistical procedure, is to be preferred greatly to the classical approach to statistics wherein no attempt is made to determine* W, *but instead an "intuitively appealing" procedure is used.* We shall see that such intuition can easily lead to the use of very poor procedures. The use of procedures which do turn out to be good ones in certain settings is not logically justifiable on intuitive grounds, but only on the basis of a careful mathematical argument.

EXAMPLE 2.2 (continued). Although the same statistical procedure may be good for many different W's of similar form, such a procedure may be a poor one if we are faced with a W which is quite different, and this is one reason why even a rough judgment as to the form of W is better than the intuitive approach which does not consider W at all. For example, suppose that underguesses of the values of μ are more serious than overguesses (they may result in giving the dog too little fuel). This may be reflected in a loss function such as

$$W((\mu, \sigma), d) = \begin{cases} 2(\mu - d) & \text{if} \quad \mu \geq d, \\ d - \mu & \text{if} \quad d \geq \mu. \end{cases}$$

For such a loss function, we would use a different statistical procedure than we would for the loss functions exhibited previously, which were symmetric functions of $d - \mu$.

EXAMPLE 2.3. We shall now give an example to illustrate how W might be determined on economic grounds. A middleman buys barrels of bolts from a manufacturer at \$15 per barrel. He can sell them to customers at \$25 per barrel, with a guarantee that a barrel contains at most 2 percent defective bolts, the customer receiving his money back while keeping the barrel of bolts in the event that he finds more than 2 percent defective bolts in the barrel. (We assume the customer always discovers how many defective bolts there are.) The middleman also has the option of selling the barrels with no guarantee, at \$18 per barrel. He removes a random sample of 200 bolts from each barrel, inspecting the 200 bolts from each barrel for defectives and deciding on the basis of this inspection whether to make the decision d_1 to sell that barrel at \$25 with a guarantee or the decision d_2 to sell it at \$18 without a guarantee. (The selling prices are for the barrels with 200 bolts removed from each; the cost of inspection is included in the \$15; the inspection procedure destroys the inspected bolts, which therefore cannot be returned to the barrels.) Let θ be the (unknown) proportion of defectives in a given barrel after the sample of 200 has been removed. If $\theta \leq .02$, it would be best to ship the barrel out under a guarantee, for \$25 (that is, to make decision d_1); to ship it out at \$18 without a guarantee would result in making \$7 less profit. On the other hand, if $\theta > .02$ and the middleman makes decision d_1, he will have to refund the \$25 as guaranteed and will receive nothing for the bolts; by making decision d_2, he will receive \$18 in this circumstance, $\theta > .02$, too. His profit (revenue minus cost) on a barrel with proportion θ of defectives and on which he makes decision d is thus given by the following table:

		Proportion θ of defectives in barrel	
		$\leq .02$	$> .02$
Decision made	d_1	\$10	−\$15
on barrel	d_2	\$3	\$3

This being profit, we could take the values of the loss function $W(\theta, d)$ to be the *negative* of those just tabled. For example, $W(.05, d_1) = \$15$, and $W(.01, d_2) = -\$3$. Note that a negative loss is the same as a positive gain. Sometimes statisticians work not with the absolute loss as exemplified here, but rather with a quantity called the *regret*. For each F, the excess of the loss $W(F, d')$ over the *minimum possible loss* for this F (incurred by making