BOOK REVIEWS

Correspondence concerning reviews should be addressed to the Book Review Editor, Professor William Kruskal, Department of Statistics, University of Chicago, Chicago 37, Illinois.

SAMUEL S. WILKS, *Mathematical Statistics*. John Wiley and Sons, New York and London, 1962. $12.95, £5.13s. xvi + 644 pp.

Review 1, by WASSILY HOEFFDING

University of North Carolina

The general nature of this book is well described in the following quotation from the author's preface: "... I have made a selection of basic material in mathematical statistics in accordance with my own preferences and prejudices, with inclinations toward trying to make a unified and systematic presentation of classical results of mathematical statistics, together with some of the more important contemporary results in a framework of modern probability theory, without going into too many ramifications." An early version of some of the material was issued under the same title in 1943 in lithoprinted form by the Princeton University Press. The book is intended for readers with good undergraduate backgrounds in mathematics. It starts out with a brief account of the foundations of modern probability theory, followed by chapters on distribution functions, mean values and moments, sequences of random variables, characteristic and generating functions, and special distributions. The statistical part begins with sampling theory and asymptotic sampling theory, followed by three chapters on statistical estimation (linear, nonparametric, and parametric) and two on hypothesis testing (parametric and nonparametric). The final chapters deal with sequential analysis, statistical decision functions, time series, and multivariate statistical theory. There are over 400 problems most of which are very helpful to the student and a good bibliography of 19 pages (which serves also as an author index).

As the quotation from the preface indicates, the emphasis is on "classical" rather than on more recent results. It is, of course, debatable which results are important enough to be included in a book which covers so vast an area. I think it would have been better if more attention had been given to those developments which have yielded fairly general and systematic methods for constructing statistical procedures with desirable properties and which bring out connections between seemingly unrelated topics. Books and papers containing important results of this kind are mentioned but often without enough indication of the content to arouse the reader's interest. These remarks apply especially to the treatment of estimation and hypothesis testing. The chapter on parametric statistical estimation deals mainly with the Cramér-Rao inequality, maximum likelihood

estimators (defined as unique roots of the maximum likelihood equations which maximize the likelihood), and confidence sets. Sufficient statistics are treated in a way which does not fully bring out the fruitfulness of the concept. The terms sufficient statistic and sufficient estimator are used interchangeably. By implication, if the parameter vector has r components, a sufficient statistic (or "set of sufficient statistics") is a random vector with r components (see, e.g., p. 356). One consequence is that it is difficult for the reader to grasp the full force of the Blackwell-Rao theorem (Theorem 12.2.3). As the theorem is stated, one has to know a sufficient *estimator* of θ in order to improve on an arbitrary unbiased estimator. The reader does not get a real idea of the powerful method of obtaining optimal unbiased estimators which involves the notion of a complete class of distributions. (An unsuccessful attempt to define the notion is made on p. 393.) Bayes and minimax estimators are not considered. The close relation between confidence sets and tests of hypotheses and Neyman's optimality criterion for a confidence set are not mentioned. The chapter on testing parametric statistical hypotheses contains a version of the Neyman-Pearson lemma (with an irrelevant restriction and incorrectly stated), Wald's extension to the case of a composite null hypothesis, and a discussion of uniformly most powerful tests; the rest of the chapter is devoted to the likelihood ratio test. No mention is made of methods for constructing most powerful unbiased tests, of tests of Neyman structure and of invariant tests. The chapters dealing with nonparametric estimation and hypothesis testing present mainly examples of particular tests and confidence sets and the reader may easily get the impression that no useful general methods for obtaining optimal nonparametric procedures exist. On pp. 462–3 the notion of a most powerful randomization test is briefly (and not very clearly) discussed. Then it is stated (p. 463) that "in order to make progress those who have utilized the method of component randomization in constructing nonparametric tests . . . have borrowed test functions . . . from parametric testing theory." This is accurate only insofar as the early history of these tests is concerned. It is not mentioned that many of the test functions borrowed from parametric theory, including those which are offered as examples, later have been shown (by Lehmann and Stein [4]) to have optimal properties similar to that which the author just discussed. Similarly out of date is the assertion on p. 466 that "it has been found necessary" to use rank tests suggested by analogous parametric testing problems. On p. 429 and again on p. 430 the author warns the reader not to confuse a nonparametric statistical hypothesis with a parametric statistical hypothesis. It would have been more instructive to emphasize that general methods for constructing good statistical procedures exist which have been successfully applied to both parametric and nonparametric problems (see, e.g., the books by Lehmann [3] and Fraser [2]). In the brief chapter on decision functions the author misses the opportunity to relate the Bayes and minimax solutions of a two-decision problem with the closely connected results on hypothesis testing in Section 13.2.

One special feature of the book is the attention that is given to sampling from finite populations. Results in this field, some of which appear to be new or have

not appeared in textbooks, are scattered throughout the book. Worth noting is the concept of a reproductive c.d.f., or rather a parametrized family of c.d.f.'s $F(x; \theta)$. The family is reproductive with respect to θ if the sum of two independent random variables having the respective c.d.f.'s $F(\cdot\ ;\ \theta_1)$ and $F(\cdot\ ;\ \theta_2)$ has the c.d.f. $F(\cdot\ ;\ \theta_1 + \theta_2)$. (But it is ambiguous to say, as is done on p. 158, that the normal distributions $N(\mu, \sigma^2)$ are reproductive with respect to both μ and σ^2; they are reproductive with respect to (μ, σ^2).) Some new nomenclature is introduced. The name Dirichlet distribution (of which the beta distribution is a special case) may well find general acceptance. New notations are used for some classical distributions such as Bi (n, p) for binomial, Be (a, b) for beta, Po (μ) for Poisson.

Unfortunately the first printing of the book contains an unusually large number of errors, ranging from false theorems and inadequate proofs to a variety of minor inaccuracies.[1] In addition, some topics are presented in a needlessly difficult or awkward form which detracts the reader's attention from the essentials. While most chapters are free of major defects, those on parametric estimation and parametric hypothesis testing contain so many that extensive corrections are needed before they can be used. I will list the more important errors and shortcomings that I have noticed or that have been brought to my attention and only a few of the lesser ones.

Theorem 4.3.8 is false.[2] It is stated as follows. "Let $(x_1, x_2, \cdots)$ be a stochastic process such that $(f_1(x_1, \theta), f_2(x_1, x_2, \theta), \cdots)$ is a stochastic process which converges in probability uniformly with respect to θ in (θ', θ'') to a finite number $g(\theta)$, where $g(\theta)$ is continuous at $\theta = \theta_0$ in (θ', θ''). Let $(\theta_1^*(x_1), \theta_2^*(x_1, x_2), \cdots)$ be a stochastic process which converges in probability of θ_0. Then $f_1(x_1, \theta_1^*(x_1))$, $f_2(x_1, x_2, \theta_2^*(x_1, x_2))$, $\cdots)$ converges in probability to $g(\theta_0)$." Counterexample: Let x_1 be uniform on $(0, 1)$; the other x_i do not enter. Let $f_n(x_1, \cdots, x_n, \theta) = \theta + \exp[-n(x_1 - n\theta)^2]$ and $\theta_n^*(x_1, \cdots, x_n) = x_1/n$. The assumptions of 4.3.8 are satisfied with $g(\theta) = \theta$, $\theta_0 = 0$, but $f_n(x_1, \cdots, x_n, \theta_n^*) = (x_1/n) + 1$ converges to $1 \neq g(\theta_0)$. The theorem is repeatedly used in Chapters 12 and 13, in particular on pp. 362, 374, 384 and 410. Theorem 4.3.4 states (correctly) that convergence in probability implies convergence in distribution. Theorem 4.3.5 says (again correctly) that if x_n converges in probability to the random variable x and g is a continuous function then $g(x_n)$ converges in probability to $g(x)$. On p. 105 it is said that "in view of 4.3.4" versions of 4.3.5 (and of some related theorems) can be obtained by requiring only convergence in distribution rather than convergence in probability. Indeed such versions (with "converges in probability" replaced by "converges in distribution" both in the hypothesis and in the conclusion) are later used, e.g., in the proof of Theorem 9.3.2, but it is hardly obvious that they can be obtained "in view of 4.3.4".[3] In derivations of

[1] Professor Wilks has informed the reviewer that errors and inaccuracies will be corrected in the next printing.

[2] This was pointed out to me by J. F. Hannan.

[3] It is true that if $x_n \to x$ in distribution and g is continuous then $g(x_n) \to g(x)$ in distri-

certain asymptotic distributions (e.g., Theorems 9.3.1, 12.3.3, 13.4.3) almost certain convergence is used where only convergence in probability is needed. Similarly, on p. 96 the study of "random variables having infinitely many components" is inadequately motivated by the need to determine limiting distributions of functions of n-dimensional random variables as $n \to \infty$. It may be judged convenient to talk of limit distributions in the framework of an infinite-dimensional sample space but it is certainly not necessary. Problem 5.14 is wrong.

The Edgeworth expansion of the distribution function of a sample sum (Theorem 9.4.1) is formally obtained under the sole assumption that certain moments are finite, but the claimed order of the remainder term is incorrect for lattice distributions. Corollary 9.4.1a and the related Problem 9.9 (which involves a lattice distribution) are also in error.

The proof of Theorem 10.2.2 (a minimum variance result of Halmos) is quite inaccurate. (In (10.2.6) delete the expectation signs to avoid confusion; in the condition for equality C is not a constant but a symmetric function of the sample point.) In section 10.9 linear estimators for means of stratified finite populations are considered. The strata sizes are $N_g = Np_g$, the strata means μ_g and the strata variances σ_g^2, $g = 1, \cdots, m$. It is then stated (p. 318) that the results "can be extended in an obvious manner to stratified sampling from an infinite population provided p_g and σ_g^2 converge to positive values and μ_g to finite values $\cdots$ as $N_1, \cdots, N_m \to \infty$. These extensions are left as exercises to the reader." The only reference to infinite populations seems to be on pp. 214–5, where simple random sampling from an infinite population is said to refer to n independent, identically distributed random variables. To regard random sampling from an infinite population as a limiting case of random sampling from a finite population may be intuitively appealing but the relation is not quite obvious. The meaning of the limits of the p_g is not clear to me.

After deriving the confidence interval for a quantile ξ_p in terms of order statistics, the author treats the case of large samples (p. 331) in a strange way. It would have been natural to observe that $x_{(k_1)} < \xi_p < x_{(k_1+k_2)}$ is equivalent to $k_1 \leqq n_1 < k_1 + k_2$, where n_1 is binomially distributed, and to apply the normal approximation. Instead, a detour is made via confidence limits for the parameter p (which is known here) and the resulting confidence bounds for ξ_p are order statistics whose ranks are random variables. On p. 334 Robbins' result on tolerance limits is misstated. The results attributed on pp. 336 and 339 to Birnbaum and Tingey and to Dempster have been published by Smirnov [5] in 1944.

Chapters 12 and 13 (parametric statistical estimation and testing parametric statistical hypotheses) are the weakest part of the book. The beginning of Chap-

bution. L. J. Savage has pointed out to me that for real-valued random variables this can be deduced from Theorems 4.3.5 and 4.3.4 by using the following device. For any distribution function F (continuous on the right) define $F^{-1}(u)$ as $\inf \{y \mid F(y) \geqq u\}$, $0 < u < 1$. Let $x, x_1, x_2, \cdots$ be random variables with respective distribution functions $F, F_1, F_2, \cdots$. Let U be uniformly distributed on (0, 1). Then the random variables $x' = F^{-1}(U)$ and $x_n' = F_n^{-1}(U)$ have the respective distribution functions F and F_n. Moreover, if $x_n \to x$ in distribution, then $x_n' \to x'$ in probability. This easily implies the stated result.

ter 12 is rough going for the reader. In the reviewer's opinion the proper setting for the treatment of these topics would be to assume a family of distributions absolutely continuous with respect to a fixed measure. Instead, the distribution function $F(x; \theta)$ (with x and θ real) is first left arbitrary and by formal differentiation such equations as $\int_{-\infty}^{\infty} [(\partial/\partial\theta) \log dF(x; \theta)]\, dF(x; \theta) = 0$ (eq. (12.1.2)) are obtained. It is then stated that, under specified regularity conditions (essentially, $\partial F/\partial\theta$ exists), "it is evident" that $(\partial/\partial\theta) \log dF(x; \theta)$ (denoted by $S(x, \theta)$) is defined as

$$S(x', \theta) = \lim_{x \to x'} \frac{\frac{\partial}{\partial\theta}[F(x'; \theta) - F(x; \theta)]}{F(x'; \theta) - F(x; \theta)}$$

where $x < x'$, provided the indicated limit exists. The interchange of the limit operations involved in this definition is not discussed. In the same context an integral $H(\theta, \theta') = \int_{-\infty}^{\infty} \log dF(x; \theta')\, dF(x; \theta)$ is first introduced in a formal way (eq. (12.1.6)) and then defined as the limit of certain approximating sums. It is not noticed that this limit equals $-\infty$ if, for instance, $F(x; \theta')$ is a continuous distribution function. The quantity $H(\theta, \theta')$ plays a prominent role in the treatment of the asymptotic distribution of the likelihood ratio (Sections 13.4 to 13.6), where also an analogous sum, $n^{-1}\sum_{\xi=1}^{n} \log dF(x_\xi, \theta)$, is repeatedly used. What is really needed here is the integral which, in the case $dF(x; \theta) = f(x; \theta)\, d\mu(x)$, is equal to $\int \log (f(x; \theta)/f(x; \theta'))f(x; \theta)\, d\mu(x)$ and which only under restrictive conditions can be written as $H(\theta, \theta) - H(\theta, \theta')$.

The treatment of point estimation suffers from several serious defects. On p. 351 an unbiased estimator for θ_0 (where θ_0 denotes the true value of θ) having finite variance is called an efficient estimator for θ_0 if no other unbiased estimator has a smaller variance. On p. 352 an unbiased estimator is said to be efficient if its variance equals the Cramér-Rao lower bound. This would be consistent with the definition on p. 351 if understood as a sufficient condition. But actually the latter statement is treated as another definition, e.g., in (12.2.15). Sometimes the reader does not know which of the two non-equivalent definitions is used, for instance in the definition of efficiency in (12.2.8). The statement of Problem 12.16 is not true with either definition. In Problem 12.10 the word exceeds should be replaced by equals. The distinction between local and uniform minimum variance unbiased estimators is not mentioned.

In Theorem 12.2.1 it is claimed that the Cramér-Rao inequality holds under the sole assumption that the distribution function of the sample is regular in its first θ-derivative in a neighborhood of the true θ value; the latter means that an equation analogous to the above-mentioned (12.1.2) is satisfied. At least the proof of this assertion is incorrect since conclusion (12.2.3) is not justified. (With this type of proof one needs regularity assumptions on the estimator as well as on the distribution function F. A result of Chapman and Robbins ([1], p. 584) implies that the inequality holds for any unbiased estimator if F satisfies a certain regularity condition). In the treatment of unbiased estimation of a vector param-

eter a lower bound, (12.6.5), for the variance of a linear combination of the components of the estimator, $\sum c_p' \tilde{\theta}_p$, is obtained which depends on arbitrary constants $c_1, c_2, \cdots$. Strangely enough the bound is not maximized with respect to these constants. (It is called a greatest lower bound for the variance of $\sum c_p' \tilde{\theta}_p$, which is not true in general even with the maximizing c_p.) Thus the covariance matrix of the estimator (if there is one) for which the (optimal) lower bound is attained is not derived explicitly.

Theorem 12.3.3 asserts the asymptotic normality of a maximum likelihood estimator under weak assumptions, but the proof depends on the false Theorem 4.3.8. On p. 362 the incorrect statement is made that the bias of a consistent estimator converges to zero. In the discussion of limiting asymptotic efficiency on pp. 363–4 the variance of the asymptotic distribution and the asymptotic value of the variance are confused.

Theorem 13.2.1, a version of the Neyman-Pearson lemma, is incorrect as stated. (The condition $P(W_\alpha \mid \theta_0) = \alpha$ cannot in general be satisfied for any $\alpha \in (0, 1)$ with the stated definition of W_α.) Sections 13.4–13.7 deal with the asymptotic distribution of the likelihood ratio for testing the hypothesis $\theta = \theta_0$ under the assumption $\theta \in \Omega$, where Ω is first taken to be a real interval. Immediately (p. 408) the assumption $\theta \in \Omega$ is substituted by $\theta \in \Omega_0$ where Ω_0 is "some (open) interval containing θ_0." It is said that "we shall see that the only part of Ω which plays an essential role $\cdots$ is Ω_0", but this is not shown. As mentioned above, in these sections illegitimate use is made of the quantity $H(\theta_0, \theta)$ and the false Theorem 4.3.8 is applied. The treatment of the asymptotic power of the likelihood ratio test (Section 13.6) is unsatisfactory. Two tests of size α are compared whose critical regions are written as $u_n + nv_n > \chi_\alpha^2$ and $u_n^* + nv_n^* > \chi_\alpha^2$ and which are consistent against the alternative $\theta_1 \neq \theta_0$ as $n \to \infty$. To overcome the difficulty of comparing the power functions of two consistent tests, the original critical regions are substituted by $u_n + cv_n > \chi_\alpha^2$ and $u_n^* + cv_n^* > \chi_\alpha^2$, where c is a positive constant. The powers of the latter tests against the (fixed) alternative θ_1 are shown to tend to limits less than 1 as $n \to \infty$. It is not explained why the behavior of the substitute tests should have a bearing on the performance of the tests being studied. (A more intuitive approach would be to leave the tests unchanged and to let θ_1 approach θ_0 at a suitable rate as $n \to \infty$.)

The definition of ranks on p. 466 is not clear. Theorem 15.2.1 (the Wald-Blackwell equation for the expected sample size of a sequential test) is false. (An additional assumption such as $\mathcal{E}\,|x_i|$ bounded is needed.) The related Problem 15.1 is also incorrect. In the first lines of the proof of Theorem 15.6.1 (Stein's two-stage confidence interval) the crucial fact of the independence of two random variables is not adequately demonstrated. (There is also a minor slip in the statement of the theorem.) In the chapter on statistical decision functions the false statement is made on p. 507 that the set of all minimax solutions is a complete set of admissible decision functions. The last sentence in the example on p. 511 is not true. In Problems 16.2 and 16.4 the reader is asked to find the decision function which provides a Bayes solution against all possible *a priori* dis-

tributions; there is no such decision function. The proof of Theorem 17.3.1 on the spectral distribution of a stationary time series is incomplete. It is not shown that the approximating distributions F_M have a unique limit.

To sum up, although the material selected by the author is not always up-to-date, it acquaints the reader with the probabilistic background, the main branches and the basic problems of mathematical statistics. The presentation of the material in the first printing suffers from many errors and some other shortcomings which are mainly concentrated in the chapters on parametric estimation and hypothesis testing.

REFERENCES

[1] CHAPMAN, DOUGLAS G. and ROBBINS, HERBERT (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* **22** 581–586.
[2] FRASER, D. A. S. (1957). *Nonparametric Methods in Statistics*. Wiley, New York.
[3] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
[4] LEHMANN, E. L. and STEIN, C. (1949). On the theory of some nonparametric hypotheses. *Ann. Math. Statist.* **20** 28–45.
[5] SMIRNOV, N. V. (1944). Approximation of distribution laws of random variables from empirical data. (Russian.) *Uspehi Mat. Nauk* **10** 179–206.

Review 2, by D. R. Cox

Birkbeck College, London

Professor Wilks's new book in an important addition to the text-books on the mathematical aspects of the theory of statistics, giving an original and coherent account of a wide range of topics. The book is intended primarily as an introduction to the subject for those with a first degree in mathematics and the mathematical style, clear and unfussy, is nicely judged for this. In addition to the main group of readers for whom the book was written, it should be useful as a reference book on theory for applied statisticians with a good mathematical training.

The general plan of the book is as follows. The first 190 pages deal with probability, random variables and the special distributions of statistics. The next 90 pages cover sampling theory, both small sample and asymptotic. Then there are 200 pages on estimation and testing, in a way the core of the book. The final 130 pages are concerned with sequential analysis, decision theory, time series and multivariate analysis. Each of the 18 chapters has exercises, over 400 in all. These range from fairly simple problems based directly on material in the book to much more difficult exercises introducing important results not covered in the main text. The exercises are a very valuable part of the book.

This is a book on mathematical theory, not on statistical methodology and, in his Preface, Professor Wilks vigorously defends separating the two. Now it would be most unfair to criticize Professor Wilks for not producing a comprehensive treatise on all aspects of statistics. Nevertheless, the book would, I think, have been much strengthened by the inclusion both of more extended motivation of