NONPARAMETRIC STATISTICS*

I. Richard Savage
University of Minnesota†

0. INTRODUCTION

Statistical procedures whose validity does not depend on the underlying random variables having a special form, such as Gaussian, are known as distribution-free or nonparametric. The amount of published material on nonparametric statistics is increasing rapidly. A bibliography published five years ago listed 999 entries [37], and perhaps another 500 have appeared since. The use of nonparametric techniques has become widespread, and they appear now to be entering introductory statistics textbooks. Specialized courses in nonparametric inference are given in several universities, and meetings of statistics professional associations usually include sessions, and almost invariably include papers, on the subject.

Both applied and theoretical statisticians have felt a need to have this material brought together. In recent months Fraser [12] and Siegel have provided books aimed to help fill this need.[1] Siegel's book is designed for the research worker with a limited mathematical and statistical background, Fraser's for the advanced student of mathematical statistics. Thus, practice and theory are divided between the two books much as they are divided between chapters in Kendall's rank correlation book [19].

Nonparametric methods are needed in many fields, and can be applied in all. Siegel, by his choice of sub-title and examples, has directed his book to the attention of behavioral scientists. As will be brought out later, however, little behavioral science is involved in it, and the interest of the book for research workers will be the same whether they are interested in behavioral science or not.

Siegel's book contains nine chapters and twenty-one tables. The first three chapters review principles of testing hypotheses, the theory of scaling, and the advantages and disadvantages of nonparametric methods. Other chapters treat (4) one sample, (5) two matched samples, (6) two independent samples, (7) k matched samples, (8) k independent samples, and (9) multivariate samples. Each of these chapters offers several alternative techniques. For each technique, the discussion is organized in a standard pattern, covering the intuitive appeal of the technique, the distribution of the test statistic under the null hypothesis (separately for small and large samples), and completely worked examples. This organization is a natural one and makes it easy to find specific material.

In spite of the need for a book on nonparametric statistics for the research worker, I cannot recommend this one. The conception is good, but the execution

* An invited review of Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill Book Company, 1956. Pp. xvii, 312. $6.50.

† This article was written while the author was at the Center for Advanced Study in the Behavioral Sciences.

[1] A review of Fraser's book appears elsewhere in this issue.

is bad. The bases for these judgments, together with some constructive suggestions, can be summarized under the following four headings, around which the rest of this article is organized:

(1) *Scope.* The only form of statistical inference considered is testing hypotheses of "no difference." Neither other types of hypotheses nor estimation is considered. Furthermore, the exposition of the principles of testing "no difference" hypotheses is frequently faulty; in particular, the treatment of power is never adequate.

(2) *Organization.* There is frequent, undesirable repetition. Many of the procedures are introduced several times with different names and in slightly different contexts. Usually the fact of repetition is not made clear. No effort is made to show how the procedures can be generalized or used in new situations. The dual treatment of large and small samples is unnecessary. Frequently-occurring topics, such as ties, limit theorems, and randomization to achieve a desired level of significance, are not treated in a unified manner.

(3) *Behavioral Science.* Articulation with behavioral science is shallow. The examples are sketchy. The relation of statistical techniques to the substance of research is not properly presented.

(4) *Alternative Sources.* There are readily available collections of material on nonparametric statistics that are comparable in scope and sounder in presentation.

1. SCOPE

1a. *Forms of Inference.* While tests of significance have dominated the literature of nonparametric inference, there have been other important developments, for example, tolerance intervals and confidence intervals. The recent elementary text of Wallis and Roberts,[2] for example, explains how confidence intervals can be generated from tests and applies this in the nonparametric case [48, pp. 461–3, 593–8]. Nonparametric test statistics frequently constitute estimates of important parameters: the sign test statistic estimates the probability of a positive response; the Wilcoxon two sample test statistic (rather, a known function of it) estimates the probability that an individual selected at random from one population will have a larger score than an individual selected at random from a second population; the Kendall rank correlation statistic estimates the difference between the probabilities of concordance and discordance.

Siegel does not explain why his interest is confined to tests of significance; to make measurements and then ignore their magnitudes would ordinarily be pointless. Exclusive reliance on tests of significance obscures the fact that statistical significance does not imply substantive significance.

The tests given by Siegel apply only to null hypotheses of "no difference." In research, however, null hypotheses of the form "Population A has a median at least five units *larger* than the median of Population B" arise. Null hypotheses of no difference are usually known to be false before the data are collected

[2] The Wallis-Roberts book was published only about four months before Siegel's, so presumably was not available to him.

[9, p. 42; 48, pp. 384–8]; when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science.

1b. *Influence of scale of measurement in choosing a test.* The question of what test statistic may be used appropriately on particular sets of data is repeatedly mishandled. The erroneous statement is frequently made that arithmetic operations should not be performed if the results are not meaningful in terms of the phenomenon the data measure. (See pp. 21–30 and sections marked "ii. Statistical Test" in each example.) This implies that I.Q. scores should never be added, since two I.Q.'s of 50 are not like one of 100 (on p. 139, however, "aggression scores" are added). I know of no reason to limit statistical procedures to those involving arithmetic operations consistent with the scale properties of the observed quantities. Thus, in working with I.Q. scores, the Fisher-Pitman randomization test is certainly acceptable even if the t test is not. The randomization test has the desired level of significance, and for many alternatives has good power.

As a matter of fact, it is not always clear which operations are involved in particular test procedures; different computational procedures, one acceptable by Siegel's criterion and another unacceptable, may be logically equivalent and lead to identical results. The Wilcoxon signed-rank test and the Walsh test, for example, are essentially the same [44]. In the Wilcoxon test the observations are ranked from smallest to largest in absolute value, but in the Walsh test some of the observations are added together. Similarly, the Wilcoxon two-sample test and the Mann-Whitney test are equivalent; but the Mann-Whitney test is based only on comparisons of pairs of observations, while the Wilcoxon test requires the ranking of the observations from smallest to largest.

It is probably true, however, that sophisticated assumptions about the distribution of the observations are commonly unsupportable in situations where very few arithmetic operations with the data lead to numbers that are meaningful in terms of the phenomena under study.

1c. *Role of Normal Distribution.* Almost invariably, Siegel compares nonparametric procedures with the corresponding normal procedures based on t, χ^2, or F. (He never states explicitly the basis of the comparison; see sections marked "Power and Power-Efficiency" in the treatment of each procedure.) The value of nonparametric techniques lies in the fact that they may be used without making the assumption of normality. Thus, in order to evaluate these procedures their operating characteristics should be found for the largest possible class of distributions. Whenever possible, one should go further than just considering normal situations. Siegel's exclusive interest in normal alternatives weakens the book by excluding consideration in any detail of the power functions of the sign test, various tests related to contingency tables, and tests of randomness. This weakness is illustrated by: "Inasmuch as there is no parametric technique applicable to data measured in a nominal scale, it would be meaningless to inquire about the power-efficiency of the binomial test when used with nominal data" (p. 42). If the binomial test is the only procedure available, then, by definition, its power efficiency must be one.

In spite of strong interest in normal alternatives, the author never gives power functions—not even for normal alternatives[3]—,never describes such frequently used tests for normality as the χ^2 goodness of fit test with parameters estimated from the data, and never supplies any data as to the effects of non-normality on tests involving normality assumptions.

1d. *Power Efficiency.* "Power efficiency" is defined as the ratio of the sample sizes required by two test procedures in order to obtain the same power. In order to use this concept, however, modifications and details are necessary. The meaning of a power efficiency of .95 when one of the sample sizes is 6 (see pp. 42, 67, etc.) is not covered by this definition, since, for example, .95×6 and 6/.95 are not integers. Some notion of a fractional observation must be introduced. Also, the power efficiency depends on the alternative that is being investigated and the significance level. Usually, when *asymptotic* efficiency is meant, this is made clear; but we read: "The randomization test for matched pairs, because it uses all of the information in the sample, has a power-efficiency of 100 per cent" (p. 92; see also p. 156). This statement should be limited to the asymptotic case and to comparing the randomization procedure to the appropriate t test. The meaning of "uses all of the information" is vague.

1e. *Miscellaneous Remarks on Principles of Inference.* Following is a small sample intended to illustrate the nature of the erroneous, imprecise, or ambiguous assertions which permeate the book:

"Roughly speaking, the larger χ^2 is, the more likely it is that the observed frequencies did not come from the population on which the null hypothesis is based" (p. 43). The use of "likely" suggests an *a priori* distribution which is never mentioned. Although this line of reasoning is interesting, it runs counter to the formal theory of testing hypotheses adopted as the framework of this book.

"Notice that the p we found by the Kolmogorov-Smirnov test is smaller than .01, while that found by the χ^2 test is larger than .05. This difference gives some indication of the superior power of the Kolmogorov-Smirnov test" (p. 51). If the null hypothesis is true, no such statement is true.

"That is, if the tables which assume that $F_0(X)$ is continuous are used to test a hypothesis about a discontinuous variable, the test is a conservative one: if H_0 is rejected by that test we can have real confidence in that decision" (p. 59). What should be said is that if the procedure is used where ties can occur, then the level of significance is less than that shown in the table. "Real confidence" adds nothing.

At one point, the null and alternative hypotheses are formulated in terms of the test statistic instead of the underlying populations (p. 78).

A statement is made (p. 126) to the effect that a nonparametric procedure (in this case the Wilcoxon test) can be superior to its "parametric alternative." What is meant is that for some types of underlying distributions it is better (in the sense of more powerful) to use the Wilcoxon statistic than it is to use the t test. However, there is always a parametric procedure at least as good as the

[3] The only example of a power function is on page 10, where a curve is given purporting to be the power function of the two-sided test for the mean of a normal distribution with known variance. The drawing is, however, grossly inaccurate, and serves only to show the most general features of the shape.

Wilcoxon procedure if only because, in any parametric context, the Wilcoxon procedure can be regarded as parametric. (With regard to this point, see the recent paper by Hodges and Lehman [18].)

In discussing multiple tests of significance, the computations are made as if the tests were mutually independent (pp. 159–60). This is not the case in considering all contrasts in the analysis of variance. An additional reference to Scheffé [39] would have been helpful at this point.

In applying the Kruskal-Wallis rank analysis of variance to the weights of pigs at birth the null hypothesis is rejected (p. 192). The data involve litters of various sizes. It is concluded that the rejection of the null hypothesis implies that the birth weights vary on the litter sizes. A more justifiable conclusion is that the litters have different average birth weights. When examining the same data Kruskal and Wallis [21] avoid this temptation to specify the cause for rejecting the null hypothesis and Snedecor [42, 4th ed., p. 238], from whom the data originate, explicitly warns the reader against it.

"The Spearman r_s and the Kendall τ are equally powerful in rejecting H_0, inasmuch as they make equivalent use of the information in the data" (p. 223). Actually, the Spearman and Kendall procedures for small samples do not have the same power. The statement applies only approximately for large samples. The expression "equivalent use" is not defined.

2. ORGANIZATION

Wisely, there is no attempt to cover all nonparametric procedures, but, unwisely, the existence of other procedures is not indicated, except at one place (p. 194). Frequently several procedures are introduced where an elementary and enlightening argument would show them to be equivalent. The following section contains a list of these equivalent procedures, and the section after that discusses the value of showing the equivalences.

2a. *Multiple Appearance of Procedures.* As has been pointed out, both the one sample signed-rank Wilcoxon test and the Walsh median test are included and treated separately when they are essentially the same. The material on the Walsh test adds five unnecessary (and misleading) pages. Also included are both the Friedman analysis of variance and the Kendall-Smith coefficient of concordance, although they are equivalent. The description of the Kendall measure takes eleven pages. The sign test is introduced several times: as the binomial test on page 36, as the McNemar test on page 63, and as the sign test on page 68. The use of chi-square for $r \times k$ contingency tables is duplicated unnecessarily: one treatment discusses two independent samples (p. 104), the other, k independent samples (p. 175). Two tests based on runs are introduced. They are the total number of runs above and below the median and the Wald-Wolfowitz total number of runs for the two-sample problem: under the null hypothesis these have the same distributions.

It is true that each time a test is introduced it is in a different setting and slightly different aspects of it are considered. Because the book may be used as a handbook some repetition is perhaps desirable, or at least unavoidable. But to omit any unified treatments, to repeat the details each time, and to make such switches as that between the Friedman and Kendall-Smith forms of the

Friedman test, is to introduce needless confusion and render the book almost useless as a textbook.

Part of the duplication can perhaps be explained, but not excused, by noting that several of the procedures have been published several times by authors who did not know of the earlier work or did not understand its relation to their own. That Wilcoxon's name should appear only once, and then only in the references, in the section on the Wilcoxon two-sample procedure is, however, scarcely explainable. (Incidentally, on p. 120 of this discussion, Wilcoxon's formula for computing the test criterion appears with no mention of Wilcoxon.) Another historical point is the neglect of the Wallis paper [46] in preference to the simultaneous paper of Kendall and Smith [20] when talking about the correlation ratio using ranks. If more attention had been paid either to the Wallis paper or to Friedman's later paper [14], the relation between the Friedman and the Kendall-Smith procedures would have been exploited.

2b. *Flexibility of Nonparametric Procedures.* The duplicate presentation of procedures and techniques causes not only excessive size and "cook-bookishness," but—more important—it keeps the book from conveying any appreciation of the flexibility of nonparametric procedures. Flexibility is well illustrated by the sign test. There are many reasons for the popularity of this test: the assumptions underlying the test are relatively weak; it is easy to compute the test statistic; the distribution of the test statistic is available for both the null and alternative hypotheses when working with small or large samples; and for many testing situations the test has fairly good power when compared to other tests available. These reasons all contribute, but the real popularity of the sign test comes from its usefulness in many and diverse experimental situations. It can be used whenever a dichotomized variable is being studied, such as diseased or not diseased, successful or not successful, male or female, change from good to bad or from bad to good. Often experimental data can be put in dichotomized form even if they did not originate in that form. Applying the sign test to numerical data, we replace the numbers by the two classes above and below the hypothetical median. When some frequencies on a scale with several values are small, it may be desirable to reduce the scale to two well-represented points. If, for example, on the four-point scale "like very much," "like," "dislike," and "dislike very much," the two extremes are seldom used, the scale can be replaced by the dichotomy "like" and "dislike." Thus, a large variety of experimental data can be cast into a form where the sign test is applicable. The sign test is applicable to other situations involving dichotomies even if the hypothesis does not involve a fifty-fifty split. Examples of other than fifty-fifty splits arise in genetics and in multiple-response questions with only one correct answer. In many such cases, estimation of the underlying parameter of the dichotomy is important. Then such tools as the Clopper and Pearson [6] charts for confidence intervals, not covered by Siegel, are useful. Finally, the sign test involves working with binomial variables and is in that sense "parametric." Hence, the use of sequential procedures and decision theory is possible. (Sequential nonparametric procedures, such as presented by Romani [36], are not covered in Siegel's book.)

Not all nonparametric procedures are as flexible as the sign test. The double dichotomy based on two natural classifications, however, is the same as the Brown and Mood [5] model for testing that two populations have the same median; and it is natural to generalize it to testing that two populations have the same lower quartile or other centile. The reason the total number of runs in two samples and the total number of runs above and below the median have the same distribution under the null hypothesis can be seen as follows: In considering runs above and below the median, form two samples, the first containing the serial numbers corresponding to observations below the median, and the second composed of the serial numbers corresponding to observations above the median. The serial numbers occurring in the two samples are not observations on mutually independent random variables but the dependence is "symmetric." Under these circumstances, the procedure may be applied to the two samples of serial numbers. Since we already have tables of the distribution of runs for the case of unequal sample sizes, it is desirable to point out that there are analogous procedures to runs above and below the sample median, such as runs above and below the hypothetical median, or runs above and below the sample quartile. By this time the reader would be curious about the possibilities of using runs when he has more than two kinds of elements and a brief treatment of runs of several kinds of elements would be desirable [28], [47], [48]. It would also help the reader avoid confusion to tell him that there is another useful type of run, runs up and down [24, 30]. The Fisher-Pitman randomization technique [34] is a general procedure and should be treated as such. Several uses are made of it, but it is not pointed out that it is always usable, and, in particular, it is neglected in the chapter on correlation.

2c. *Special Problems.* The discussion of each procedure includes several short paragraphs according to a standard outline. The sections on large samples are almost the same as those for small samples, except for a different table of significance levels. Both tables could easily be included in the same discussion, and using the same example would serve to compare the large- and small-sample techniques. Although there is usually a paragraph on power efficiency—that is, the comparative power of a test relative to some other—this is never related to the examples; and there is no discussion of the absolute power of a test—for example, whether a particular experiment was designed so that interesting results could reasonably be expected with samples of the sizes used. The separation in the examples of the "statistical test," "sampling distribution," and "region of rejection" gives rise to several steps with almost no separate content. It destroys the idea of a test of significance as a decision procedure.

Certain problems recur frequently, yet there is no place where they are treated in a unified manner. Thus, there is no mention of the particular types of central limit theorems useful in nonparametric work (see Fraser [12] Chap. 6). There is no unified treatment of tied observations.[4] The only discussion of ancillary randomization to attain a specified level of significance from an otherwise discontinuous test is for the 2×2 table (p. 102). Such a discussion is

[4] The most extensive available discussion of ties is by Putter [35]. This covers the sign test and the Wilcoxon one-sample signed-rank test.

needed for the first example in the book, where a test at the .01 level is proposed and a test at the .0032 level is used. There is no general discussion of continuity corrections, although they are introduced several times.

3. BEHAVIORAL SCIENCE

The function of the examples in this book is to give detailed computations showing the steps in applying the procedures. Computational techniques such as checks and shortcuts, however, are not discussed. The examples do not give evidence as to which of the procedures have been found particularly useful in applications. The discussion surrounding the examples fails to suggest the role of statistical techniques in the completed research project. The emphasis is always on making isolated conclusions from a limited part of the data.

3a. *Depth of Examples.* The vocabulary of certain behavioral sciences is used superficially in some of the examples, but no specialized knowledge of the subject matter is required to understand them. Most of the examples are unrelated to each other so that neither the subject matter nor the interrelations of statistical methods in different aspects of a research study are developed between examples. Within an example, the analysis related to the subject is not carried to a sufficient depth to give any real understanding of even the specific application of a statistical test. Thus, in the cross-cultural example begun on page 112 (the data are from Whiting and Child [49]), a "random sample of cultures" is required. The specification of the population of cultures and the method of selecting a sample are not discussed. This is, in fact, an extremely difficult task that has not been satisfactorily handled in any cross-cultural research, but until it is, tests of significance remain meaningless.

The data in the example begun on page 39 are treated as if they arose from one sample. Actually they came from two samples, the second set of observations having been collected after a significant result was obtained on the first set [1]. If two samples are combined when it is known that one of them is significant, the second test of significance is invalid, since there is a higher probability of rejecting the null hypothesis (when it is true) than the nominal level of significance indicates. How to combine the information of the two samples in an experiment of this kind must depend on the particular situation. One possibility is to use the first sample for exploratory purposes, i.e., find interesting hypotheses from it, and estimate the sample size required for testing these hypotheses. Then the experimenter can use the second sample for making inferences.

In the example begun on p. 54 the data used in making a test of randomness are only a portion of the relevant data in the source [40]. The experiment involved observing a group of children in pairs at different times. Of primary interest is the order in which the children appear, i.e., the first child, the second child, etc. In the example, the method of assigning orders to pairs of children observed simultaneously is not indicated. The children are observed twice, occasionally with the second observation on a particular child before the first on another child. Only the data from the first appearances are used in the example. The analysis of the example is formally valid, in that the significance levels are actually those claimed, but it would not satisfy a sensible experimenter, because half of the data are neglected.

3b. *Applications of Nonparametric Procedures.* That nonparametric procedures have a role special to the behavioral sciences is, I believe, false. They are important for the behavioral scientist, but most of them have been developed and used in other fields, for example, engineering, production, cloud seeding, medicine, and agriculture. On the other hand, if a particular technique has not been used and is of no theoretical interest, it should not be included in a book on applications.[5] Actually, it is not easy to find applications of all of the techniques in one substantive field. However, Siegel has failed to use extensive data on the use of nonparametric procedures which have been collected by Savage [37], Blum and Fattu [3], and Bradley and Duncan [4].

Several of Siegel's examples are artificial or trivial. This is true of the examples of the one-sample Kolmogorov-Smirnov test (p. 49) and the runs above and below the median test (p. 56). Goodman [16] has presented a good example of the Kolmogorov-Smirnov test and Wallis [47] and Wolfowitz [53] have discussed several interesting applications of the runs test. The example for the one sample chi-square goodness of fit test involves horse racing, despite the fact that χ^2 is the most frequently used nonparametric procedure in the behavioral sciences [3].

A striking illustration of the use of nonparametric statistics in research, specifically in experimental psychology, appears in Volume 45 (1953) of the *Journal of Experimental Psychology.* Of the 70 papers in this volume, at least 12 involve applications of nonparametric procedures, not counting chi-square tests of goodness-of-fit or contingency-tables. Presumably the procedures are being used still more now.

4. ALTERNATIVE SOURCES

The books and articles discussed in this section are alternative presentations of nonparametric statistics designed for an audience of the same statistical maturity as that of Siegel's book.

4a. *Textbooks.* From all that has been said already, it might still be supposed that Siegel has collected more nonparametric material than appears elsewhere. This is not in fact the case. The book by Walker and Lev [45], which includes a chapter by Moses on nonparametric procedures, covers nearly everything contained in the Siegel book (and, of course, it contains much more in the general field of statistics). Topics in the Siegel book which are omitted in the Walker-Lev text are the Fisher-Pitman randomization procedures, Kendall's tau, Cochran's Q test, and the Moses test. Omission of the randomization procedures would be a serious shortcoming in a theoretical treatise, but in practice these procedures are rarely used because of computational difficulties; their chief role in theory is to justify the use of "normal theory" even in nonnormal situations.[6] Walker and Lev present tables or approximate distributions for each procedure.

The Walker and Lev book has more in common with Siegel than others I examined. Dixon and Massey [7] include several nonparametric procedures

[5] The Wald-Wolfowitz test of total number of runs is a case in point. Siegel's example (pp. 138–141) for illustrating this procedure originally did not use it. For the limitations of the procedure see Mood [29] and Lehmann [23]. Siegel himself implies that the test should not be used when he writes: "The runs test also guards against all kinds of differences, but it is not as powerful as the Kolmogorov-Smirnov test" (p. 158).

[6] Randomization procedures were first introduced by R. A. Fisher [8, sec. 21] for this very purpose.

with the necessary tables for applying them. Also included in their book are many procedures based on order statistics. These procedures, although parametric in that they depend on the assumption of normality, are related to common nonparametric procedures in being easy to apply. Wallis and Roberts [48] consider at least one "deep" example (sec. 2.8.4) involving nonparametric techniques, include nonparametric confidence intervals, and present the Goodman and Kruskal [17] rational analysis of contingency tables.

4b. *Articles.* Many of the theoretical papers in nonparametric inference use a surprisingly small amount of mathematics and are accessible to the applied statistician. The survey articles by Smith [41],[7] Moses [31], and Blum and Fattu [3], can be used as basic reading in nonparametric inference; for many of the techniques, these papers contain sufficient information to make applications without finding other references. For many of the techniques, expository material has appeared which is particularly helpful when detailed information is required. In particular there are source papers by Friedman [13] and by Kruskal and Wallis [27] on the nonparametric analysis of variance. In Kendall's book on rank correlation the mathematics and applied material are well separated. The writings of Wilcoxon [50, 51] are most pleasant even for the beginning student. The technical material on the Kolmogorov-Smirnov procedures is difficult, but there are good expositions by Goodman [16], Massey [26], Birnbaum [2], and Miller [27].

In conclusion, Siegel's book stands as the first attempt to give a full dress presentation of nonparametric procedures in a form suitable for the research worker. The material presented is arranged so that it may be easily found. All other aspects of the book are open to severe criticism.[8]

APPENDIX

The following minor comments may help readers of the book, and will help substantiate my evaluation of it. All comments regarding the tables are presented together, regardless of where the tables are discussed in the text.

A1. *Text*

P. 8. The references suggested for reading in decision theory will be too advanced for most users of this book. Alternatives are Girshick [15], L. J. Savage [38], Luce and Raiffa [25], and Williams [52].

P. 8. "Associated probability" is used here and defined on p. 11 as the probability of an event at least as "extreme" as the one observed. The meaning of "extreme" is not made clear for the two-sided case.

P. 13. The statement that "the region of rejection is a region of the sampling distribution" can be understood by noting that the "sampling distribution" is defined by Siegel as the possible values of the test statistic, i.e., the sample space. His "sampling distribution" is not a distribution.

P. 15. In the formula for binomial probabilities, "P" is the probability of "success" and "Q" is the probability of "failure." It should be noted that $Q=1-P$. Here and on p. 71 Siegel is being irrelevantly general since he is interested only in the case $P=Q=\frac{1}{2}$.

P. 20. When the assumptions underlying a test are not met, not only the level of significance, but also the power, will be difficult to find.

P. 21. The material on measurement covers only the univariate case, and thus does not apply to the chapter on correlation.

[7] Coombs has an article on scaling in the same reference.

[8] Professor Siegel's comments on a draft of this article resulted in the removal of several errors. This does not imply that he agrees with this article in any way.

P. 32. "2. If sample sizes as small as $N=6$ are used, there is no alternative to using a nonparametric statistical test unless the nature of the population distribution is *known exactly*." It would be desirable to present data regarding the effects of non-normality. Departure from normality will, no doubt, disturb the significance levels of normal tests more for small samples then for large samples. On the other hand, with larger samples it is frequently desirable to work with smaller error rates; hence the comparison of small and large samples is not straightforward. The basis of "$N=6$" is unknown to me. For very small samples there is a special difficulty with nonparametric procedures, since their distributions are discrete. Thus, using the two-sided sign test with a sample of four, we must either have a significance level at least as large as $\frac{1}{8}$ or else use a randomized test. If a test at the ten per cent level is desired we would reject only if all of the signs are the same *and* we had a success with a randomization device with a success probability of .8.

P. 44. The "observations" referred to in computing the degrees of freedom are the observed frequencies in the various classes. This applies to p. 106, also.

P. 49. The example for the one-sample Kolmogorov-Smirnov test involves a discrete hypothetical distribution; thus the level of significance will be less than assumed. This is discussed on p. 59.

P. 55. For an even number of observations, "middlemost score" can be interpreted as the mean of the two middle observations.

P. 67. The difference in results between the binomial and chi-square tests depends not only on the coarseness of the chi-square tabulation but also on the fact that the chi-square test is an approximation. On p. 74 the difference is due only to coarseness.

P. 74. In Table 5.6, "rank with less frequent sign" should read "smaller sum of like-signed ranks."

P. 83. In a symmetrical distribution, the probabilities associated with negative deviations from the median are the same as the probabilities associated with corresponding positive deviations. The symmetry assumption is required for the Wilcoxon test as well as for the Walsh test.

P. 95. The examples described as "random sampling" actually involve "systematic sampling." Thus, in selecting every tenth house from a list of 1,000, there are only 10 possible samples. The usual interpretation of "random sampling" would be sampling without replacement, i.e., each of the $\binom{1000}{100}$ possible subsets of 100 from the original 1,000 would have the same probability of being the sample. If corner houses all have addresses ending with the same digit (or if, as is in fact common, about a quarter of them have addresses ending in one digit and a quarter another digit), a systematic sample will either heavily under- or over-represent corner houses.

P. 113. The "sampling distribution" can be stated before the experiment. The expected cell frequencies are known because all of the marginal totals are made equal in the Brown-Mood procedure.

P. 126. In case C, where $n_2 \geq 20$, the normal approximation cannot be used if n_1 is small. Fix and Hodges [11] have considered this possibility.

P. 128. The two sample Kolmogorov-Smirnov test is introduced here for discrete distributions, but it derives its nonparametric interest from the continuous case where exact significance levels are obtained.

P. 131. The occurrence of a chi-square variable in finding levels of significance of the one-sided Kolmogorov-Smirnov test is an interesting mathematical fact, but not related to the uses of chi-square that have occurred previously.

P. 137. The word "sign" is used here in the sense of "symbol," not plus or minus.

P. 172. Table 7.5 indicates that for the examples considered the two test procedures frequently give the same result. This, however, does not give information about the power since we are not told whether the null or alternative hypotheses were true.

P. 182. In Table 8.2 the headings presumably should be "8 or less," "9 or 10," and "11 or 12."

P. 201. The measure C is not directly comparable to the other measures of correlation in that they do not measure the same thing. In fact, all of the measures of correlation differ in what they measure and hence none of them are directly comparable.

P. 232. The lower limit of Rs_{av} is $-1/(k-1)$, not -1.

A2. *Tables*

All statistical tables are listed, whether or not I have comments, in order to describe the contents.

Table A (Normal distribution). The blanks in the lower right portion of the table are not explained, but actually indicate that the appropriate values do not differ appreciably from those given in the margin.

Table B (Critical value of t). It should have been explained that the sequence of the last five values for *df* facilitates interpolation [7].

Table C (Chi-square critical values). It should have been explained that if N (the *df*) >30 critical values may be approximated, for example by

$$\chi_p^2 = \tfrac{1}{2}(\sqrt{2N-1} + U_p)^2$$

where U_p is the corresponding one-sided critical value for a normal distribution.

Table D (Binomial distribution, $P=Q=\frac{1}{2}$). The blanks in the lower left portion indicate that the values are almost zero. The †'s in upper line indicate that the values are exactly 1.

Table E (One-sample Kolmogorov-Smirnov critical value). It is not made clear that this is a one-sided (upper-tail) table and the levels of significance are exact.

Table F (Critical value of runs). These tables are symmetric in n_1 and n_2 so that the values above the diagonals are redundant. In F_I the probability of getting not more runs than indicated is at most .025 and the probability of getting more than the number of runs indicated is more than .025. Thus the instructions in the heading are wrong for the Wald-Wolfowitz test. The level of significance that will be used if the instructions are followed will be .025 or slightly less, not .05 as claimed. The Dixon and Massey [7] runs table is similar to this one. (In their instructions on page 326 (1st edition), for the table corresponding to F_I, change "not more" to "at least." This is corrected in the 2nd edition.)

Table G (Wilcoxon one-sample critical value). This table was originally adapted from a report by Tukey [44]. The levels of significance obtained by using the indicated values are as close as possible to the stated level of significance. Exact levels are not obtained for the test statistic, since it has a discrete distribution. The actual levels are sometimes above and sometimes below the stated values.

Table H (Walsh test critical value). Some of the exact probabilities here can be used immediately in finding the corresponding exact probabilities in Table G.

Table I (Contingency tables critical value). These are one-tailed probabilities and err on the "conservative" side as pointed out on page 99.

Table J (Two-sample Wilcoxon distribution). Values are given until the probabilities have accumulated to one half. (See the footnote on page 117).

Table K (Wilcoxon two-sample critical value). The entries in this table appear to be on the "conservative" side, i.e., the actual level of significance is not larger than the stated level of significance. The lower left could be left blank by symmetry. In Table K_{IV}, n_1 and n_2 are interchanged.

Table L (Two-sample Kolmogorov-Smirnov critical value). This table is for the case of two samples of equal size. Notice that for several sample sizes the one- and two-sided tests have the same critical values. Thus it "costs" no more to use a two-sided test in those cases. Clearly, some of the suggested values are on the safe side, but I have not checked all of them.

Table M (Two-sample Kolmogorov-Smirnov critical value).

Table N (Friedman analysis-of-variance distribution).

Table O (Kruskal-Wallis analysis-of-variance distribution).

Table P (Spearman's rank correlation critical value). The levels of significance appear to be on the "conservative" side. It should be noted that in the Olds [32, 33] articles the quantity whose distribution is given is $\sum d^2$, which is easier to compute than the rank correlation and is just as informative if only a test of significance is desired.

Table Q (Distribution of Kendall's tau).

Table R (Critical value of Kendall's measure of concordance). Entries in Table R may be obtained from those in Table N by multiplying those in N by $mn\ [n+1]/12$ where n is the number of treatments and m is the number of judges. The roles of k and N are interchanged in Table N and R.

REFERENCES

[1] Barthol, R. P., and Ku, Nani D. "Specific regression under a nonrelated stress situation," *American Psychologist*, 10 (1955), 482. (Abstract)

[2] Birnbaum, Z. W. "Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size," *Journal of American Statistical Association*, 47 (1952), 425–41.

[3] Blum, Julius R., and Fattu, Nicholas A. "Nonparametric methods," *Review of Educational Research*, 24 (1954), 467–87.

[4] Bradley, R. A., and Duncan, D. B. *Statistical Methods for Sensory Difference Tests of Food Quality, Bi-annual Report, No. 1*. Mimeographed report from Virginia Agricultural Experimental Station, Blacksburg, Virginia, December, 1950.

[5] Brown, G. W., and Mood, A. M. "On median tests for linear hypotheses," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, 159–66.

[6] Clopper, C. J., and Pearson, Egon S. "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, 26, (1934) 404–13.

[7] Dixon, W. J., and Massey, F. J. *An Introduction to Statistical Analysis*, New York: McGraw-Hill Book Co., 1st ed. 1951, 2nd ed. 1957.

[8] Fisher, R. A. *The Design of Experiments*, New York: Hafner Publishing Co., 1st ed. 1935, 5th ed, 1949.

[9] ———. *Statistical Methods and Scientific Inference*, New York: Hafner Publishing Co., 1956.

[10] Fisher, R. A., and Yates, Frank. *Statistical Tables for Biological, Agricultural and Medical Research*, 3rd ed. London: Oliver and Boyd, 1948.

[11] Fix, Evelyn, and Hodges, J. L., Jr. "Significance probabilities of the Wilcoxon Test," *Annals of Mathematical Statistics*, 26 (1955), 301–312.

[12] Fraser, D. A. S. *Nonparametric Statistics*, New York: John Wiley & Sons, Inc., 1957.

[13] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, 32 (1937), 675–701.

[14] ———. "A comparison of alternative tests of significance for the problem of *m* rankings," *Annals of Mathematical Statistics*, 11 (1940), 86–92.

[15] Girshick, Meyer A. "An elementary survey of statistical decision theory," *Review of Educational Research*, 24 (1954), 448–466.

[16] Goodman, Leo A. "Kolmogorov-Smirnov tests for psychological research," *Psychological Bulletin*, 51 (1954), 160–68.

[17] Goodman, Leo A., and Kruskal, William H. "Measures of association for cross classifications," *Journal of the American Statistical Association*, 49 (1954), 732–64.

[18] Hodges, J. L., Jr., and Lehmann, E. L. "The efficiency of some nonparametric competitors of the *t*-test," *Annals of Mathematical Statistics*, 27 (1956), 324–35.

[19] Kendall, M. G. *Rank Correlation Methods*, New York: Hafner Publishing Co., 1st ed. 1948, 2nd ed. 1955.

[20] Kendall, M. G., and Smith, B. Babington. "The problem of *m* rankings," *Annals of Mathematical Statistics*, 10 (1939), 275–287.

[21] Kruskal, William H., and Wallis, W. Allen. "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, 47 (1952), 583–621.

[22] Lehmann, E. L. "Consistency and unbiasedness of certain nonparametric tests," *Annals of Mathematical Statistics*, 22 (1951), 165–79.

[23] ———. "The power of rank tests," *Annals of Mathematical Statistics*, 24 (1953), 23–42.

[24] Levene, Howard. "On the power function of tests of randomness based on runs up and down," *Annals of Mathematical Statistics*, 23 (1952), 35–56.

[25] Luce, Duncan, and Raiffa, Howard. *Conflict, Collusion, and Conciliation: A Survey and Critique of the Concepts and Results of Game and Related Decision Theories*. Unpublished manuscript, Center for Advanced Study in the Behavioral Sciences, 1956.

[26] Massey, Frank J., Jr. "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, 46 (1951), 68–78.

[27] Miller, Leslie H. "Table of percentage points of Kolmogorov statistics," *Journal of the American Statistical Association*, 51 (1956), 111–121.

[28] Mood, A. M. "The distribution theory of runs," *Annals of Mathematical Statistics*, 11 (1940), 367–92.

[29] ———. "On the asymptotic efficiency of certain nonparametric two-sample tests," *Annals of Mathematical Statistics*, 25 (1954), 514–522.

[30] Moore, Geoffrey H., and Wallis, W. A. "Time series significance tests based on signs of differences," *Journal of the American Statistical Association*, 38 (1943), 153–64.

[31] Moses, Lincoln E. "Nonparametric statistics for psychological research," *Psychological Bulletin*, 49 (1952) 122–43.

[32] Olds, Edwin G. "Distributions of sums of squares of rank differences for small numbers of individuals," *Annals of Mathematical Statistics*, 9 (1938), 133–48.

[33] ———. "The 5% significance levels for sums of squares of rank differences and a correction," *Annals of Mathematical Statistics*, 20 (1949), 117–18.

[34] Pitman, E. J. G. "Significance tests which may be applied to samples from any populations II. The correlation coefficient test," *Journal of the Royal Statistical Society*, (Series B), 4 (1937), 225–32.

[35] Putter, Joseph. "The treatment of ties in some nonparametric tests," *Annals of Mathematical Statistics*, 26 (1955), 368–86.

[36] Romani, Jose. "Tests no parametricos en forma secuencial," *Trabajos de Estadistica*, 7 (1956), 43–96.

[37] Savage, I. Richard. "Bibliography of nonparametric statistics and related topics," *Journal of the American Statistical Association*, 48 (1953), 844–906.

[38] Savage, L. J. "The theory of statistical decision," *Journal of the American Statistical Association*, 46 (1951), 55–67.

[39] Scheffé, H. "A method for judging all contrasts in the analysis of variance," *Biometrika*, 40 (1953), 87–104.

[40] Siegel, Alberta E. *The Effect of Film-mediated Fantasy Aggression on Strength of Aggressive Drive in Young Children*. Unpublished doctoral dissertation, Stanford University, 1955.

[41] Smith, K. "Distribution-free statistical methods and the concept of power efficiency." In L. Festinger and D. Katz (eds.), *Research Methods in the Behavioral Sciences*, New York: Dryden Press, 1953, 536–577.

[42] Snedecor, G. W. *Statistical Methods Applied to Experiments in Agricultural and Biology*, Ames, Iowa: Iowa State College Press, 4th ed. 1946, 5th ed. 1956.

[43] Swed, Frieda S., and Eisenhart, C. "Tables for testing randomness of grouping in a sequence of alternatives," *Annals of Mathematical Statistics*, 14 (1943), 66–87.

[44] Tukey, J. W. *The Simplest Signed-Rank Tests*. Mimeographed Report No. 17, Statistical Research Group, Princeton University 1949.

[45] Walker, Helen M., and Lev, Joseph, *Statistical Inference*, New York: Holt, 1953.

[46] Wallis, W. Allen. "The correlation ratio for ranked data," *Journal of the American Statistical Association*, 34 (1939), 533–38.

[47] ———. "Rough-and-ready statistical tests," *Industrial Quality Control*, 8 (1952), No. 5, 35–40.

[48] ——— and Roberts, Harry V. *Statistics: A New Approach*, Glencoe, Ill.: The Free Press, 1956.

[49] Whiting, J. W. M., and Child, I. L. *Child Training and Personality: A Cross-Cultural Study*, New Haven: Yale University Press (1953).

[50] Wilcoxon, Frank. "Individual Comparisons by ranking methods," *Biometrics*, 1 (1945), 80–83.

[51] ———. *Some rapid approximate statistical procedures*, American Cynamid Co., Stanford Research Laboratories (July 1949), 16 pp.

[52] Williams, J. D. *The Compleat Strategyst, Being a Primer on the Theory of Games of Strategy*, New York: McGraw Hill and Co. (1954).

[53] Wolfowitz, J. "On the theory of runs with some applications to quality control," *Annals of Mathematical Statistics*, 14 (1943), 280–88.