STATISTICAL INFERENCE IN THE NON-PARAMETRIC CASE[1]

By Henry Scheffé

Princeton University

CONTENTS

1. Introduction. In most of the problems of statistical inference for which we possess solutions the distribution function is assumed to depend in a known way on certain parameters. The values of the parameters are unknown, and the problems are to make inferences about the unknown parameter values. We refer to this as the parametric case. Under it falls all the theory based on normality assumptions.

Only a very small fraction of the extensive literature of mathematical statistics is devoted to the non-parametric case, and most of this is of the last decade. We may expect this branch to be rapidly explored however: The prospects of a theory freed from specific assumptions about the form of the population distribution should excite both the theoretician and the practitioner, since such a theory might combine elegance of structure with wide applicability. The process of development will no doubt inspire some mathematical attacks of considerable abstractness. There are already signs that more number-theoretic problems and measure-theoretic problems will enter our subject through this door, and perhaps even some topological ones. Some ability to think in terms of

[1] Parts of this paper were used in an invited address given under the title "Statistical inference when the form of the distribution function is unknown" before the meeting of the Institute of Mathematical Statistics on September 12, 1943 in New Brunswick, N. J.

functionals, function spaces, and metrization of function spaces will be useful in attempting general theories of "best" tests and estimates. Toward such abstract phases of the development the attitude of the practical statistician should be one of tolerance, for the new theory already promises to give him many new tools which are both simpler and of wider use.

While the maturity of the non-parametric theory is still in the future, it is well to remark that its beginnings go relatively far back. Of our most famous tests, such as Pearson's χ^2-test, Student's test, and Fisher's analysis of variance tests, the oldest concerns a non-parametric problem: In 1900 Karl Pearson proposed his χ^2-criterion to test the goodness of fit of a theoretical distribution to observations, and in 1911 he extended his χ^2-method to the problem of two samples. The first of these problems may be regarded as non-parametric if the choice of the theoretical distribution is not based on calculations from the data, and the second is without doubt a non-parametric problem. R. A. Fisher treated an analysis of variance problem non-parametrically at least as early as 1925, for in the first edition of his *Statistical Methods for Research Workers* we find the sign test. General formulations of the problems of statistical inference, and criteria for "good" and "best" solutions[2] have been advanced by R. A. Fisher, Neyman, E. S. Pearson, and Wald. These general theories were all strictly parametric until 1941 when Wald proposed one sufficiently broad to cover the non-parametric case.

We now introduce some notation to which we shall adhere throughout this paper. Statistical inferences are based on measurements. The total number of measurements will always be denoted by n. We conceive of n random variables $X_1, X_2, \cdots, X_n$ on which the measurements are made. The domain of each X_j can always be taken to be a set of real numbers. If vector random variables occur, the X_j will denote components. The cumulative distribution function (c.d.f.) of the random variables will be written $F_n(x_1, x_2, \cdots, x_n)$, —— this is the probability that all $X_j \leq x_j$ simultaneously. The c.d.f. F_n is then always defined in a complete n-dimensional Euclidean space W, called the sample space; W is the space of points $E = (x_1, x_2, \cdots, x_n)$. The sample point with the random coordinates $X_1, \cdots, X_n$ will be denoted by $\mathbf{E}$.

In describing the validity of specific non-parametric tests and estimates in the sequel it will be convenient to refer to the following classification[3] of *univariate* c.d.f's $F(x)$: Ω_0 is the class of all F. Ω_2 is the class of all continuous F. Ω_3 is the class of all absolutely continuous F, that is, those F for which there exists a probability density function $f(x)$, so that

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi.$$

Ω_4 consists of all F which may be written in the above form with f continuous.

[2] For a bibliography see [22].

[3] The notation follows [31].

PART I. NON-PARAMETRIC TESTS

2. The randomization method of obtaining similar regions. In any problem of statistical inference it is assumed that the c.d.f. F_n of the measurements is a member of a given class Ω of n-variate distribution functions; we write $F_n \in \Omega$. Ω is called the class of admissible F_n. If Ω is a k-parameter family of functions the problem is called parametric, otherwise, non-parametric. A statistical hypothesis H is a statement that $F_n \in \omega$, where ω is a given subclass of Ω. A test of the hypothesis H consists of choosing a Borel region w in the sample space W and rejecting H if and only if the sample point $\mathbf{E}$ falls in w; w is called the critical region of the test.

The choice of the critical region w is usually[4] made as follows: A positive constant α (ordinarily about .01 or .05) is chosen and called the significance level of test. If regions w exist for which $Pr\{\mathbf{E} \in w \mid F_n\}$—the probability that the sample point $\mathbf{E}$ fall in w, calculated from the c.d.f. F_n—is equal to α for all $F_n \in \omega$, then the choice of critical region is usually limited to this class. Such regions are very important in the theory of testing hypotheses, and it is convenient to have a name for them: Following the terminology of Neyman [22] in the parametric case we shall call them *similar* to the sample space with respect to all F_n in ω, or more briefly, similar regions. A similar region is then a region w for which $Pr\{\mathbf{E} \in w \mid F_n\}$ is the same constant for all $F_n \in \omega$. The advantage of using similar regions as critical regions is that the risk of rejecting the hypothesis when it is true (type I error) is controlled: no matter what member of ω the unknown F_n happens to be, the probability of rejection of the (true) hypothesis is exactly α. We remark here that the problem of the existence and structure of similar regions in the parametric case has been treated only under very heavy restrictions and must be considered still mostly unsolved, whereas we shall see later that in the non-parametric case it promises to be relatively simple.

When similar regions exist for a chosen α there is usually a large family of them. Ideally the choice of the critical region w from the family of similar regions would be based on a complete knowledge of two functionals of F_n for $F_n \in \Omega - \omega$, that is, for those F_n corresponding to the various admissible ways in which the hypothesis can be false: the first, the probability of rejection (of avoiding a type II error), namely $Pr\{\mathbf{E} \in w \mid F_n\}$, called the power function of w, and the second, the relative importance of rejection in the concrete situation in which the test is to be applied. In other words, one would like to choose the w with the power function "best" for the very specific problem at hand. However, little has been done along this line in the non-parametric case, and, as we shall note below, the choice of w from the family of similar regions is usually made by means of a statistic chosen on intuitive grounds.

A general method of obtaining similar regions, which we shall call the *randomization method*, will now be described. The credit for originating this method and envisioning its wide applicability belongs to R. A. Fisher, who first

[4] Another approach to the choice of critical region will be described in section 13.

used it in 1925 [3]. Consider the set S of permutations on the coordinates $x_1, x_2, \cdots, x_n$, which leave invariant all the c.d.f's F_n in ω. Suppose the number of permutations in the set S is s; then s divides $n!$. Now define for any point E in W a corresponding set $\{E'\}$ of s points obtained by making on the coordinates of E the permutations of the set S. The value of the c.d.f. F_n is then the same at all s points E' generated by E, for all $E \in W$ and all $F_n \in \omega$. The s points of $\{E'\}$ will be distinct unless the point E lies in a certain region W_0; W_0 depends on the set S of permutations determined by the class ω, and will always be contained in the union of all diagonal hyper-planes $x_i = x_j$ $(i \neq j)$. A critical region w is constructed by the randomization method by choosing a positive integer $q < s$, and for every E not in W_0, putting q points of the corresponding set $\{E'\}$ in w and the remaining $s - q$ points outside w, by any rule whatever, just so w is a Borel set. We shall also say that a Borel set w is obtained by the randomization method if it has the structure just described except on a (Borel) subset w_0 of w having the property $Pr\{\mathbf{E} \in w_0 \mid F_n\} = 0$ for all $F_n \in \omega$. It may be shown by the methods used elsewhere [31] by the writer that if ω is a class of continuous c.d.f's then the region w thus obtained is a similar region with $\alpha = q/s$; furthermore, that under mild restrictions (roughly, that the boundary of w be a sufficiently "thin" set), at least for certain classes ω, this is the *only* method of obtaining similar regions.

One might call the set $\{E'\}$ of points corresponding to E the subpopulation of points "equally likely" under the null hypothesis H, but we shall call $\{E'\}$ simply the *subpopulation* corresponding to E. The decision as to which q of the s points of the subpopulation are to be put into the critical region w is usually made with the aid of a statistic T chosen on an intuitive basis. By a statistic T we mean of course a function of the sample only, not depending on the c.d.f. F_n, thus $T(\mathbf{E}) = T(X_1, \cdots, X_n)$. For a suitably chosen q, the q points of the subpopulation $\{E'\}$ giving $T(E')$ values in a certain range—usually the q largest or q smallest values—are put into w, and these q values are then called the "significant" values.

Before proceeding further let us consider an example illustrating all the definitions we have introduced thus far. Suppose that on the basis of a sample of m pairs (X_i, Y_i), $i = 1, 2, \cdots, m$, from a bivariate population with unknown c.d.f $G(x, y)$ we wish to test the independence of the random variables X, Y. To fit our general notation write $Y_i = X_{i+m}$. Assuming only that the sample is random, we have, with $n = 2m$, that the c.d.f. of the sample point $\mathbf{E}$ is of the form

$$F_n(x_1, \cdots, x_n) = \prod_{i=1}^{m} G(x_i, x_{i+m}).$$

Now suppose we know or are willing to assume further that the unknown c.d.f. $G(x, y)$ of the population is in a certain class $\Omega_\nu^{(2)}$ of bivariate c.d.f.'s, where $\Omega_\nu^{(2)}$ is the bivariate analogue of the class Ω_ν of univariate c.d.f.'s defined in section

1; thus if we knew the unknown $G(x, y)$ were continuous, we would have $G \in \Omega_2^{(2)}$. The class Ω of admissible F_n is then

$$\Omega = \left\{F_n \mid F_n = \prod_{i=1}^{m} G(x_i, x_{i+m}); G \in \Omega_\nu^{(2)}\right\},$$

where the notation $\{F_n \mid F_n$ of the form $\mathfrak{F}\}$ denotes the class of all F_n of the form $\mathfrak{F}$. The hypothesis of independence may now be expressed as $H: F_n \in \omega$, where the subclass ω of Ω is

$$\omega = \left\{F_n \mid F_n = \prod_{i=1}^{m} F^{(1)}(x_i) \prod_{j=m+1}^{2m} F^{(2)}(x_j); F^{(k)} \in \Omega_\nu, k = 1, 2\right\}.$$

The set S of permutations which leave all $F_n \in \omega$ invariant is obtained by making all possible permutations of the first m coordinates $x_1, \cdots, x_m$ among themselves, and of the second m coordinates $x_{m+1}, \cdots, x_{2m}$ among themselves. The total number s of permutations in S is thus $(m!)^2$. Making these permutations on the coordinates of any point E in W, we get the set $\{E'\}$ of $(m!)^2$ points. The points of $\{E'\}$ are distinct unless E lies in the region W_0 defined as the union of all hyperplanes $x_i = x_j$ where $i \neq j$ and i, j are both in the set of integers $1, 2, \cdots, m$ or else both in the set $m + 1, \cdots, 2m$. Pitman [28] applied the randomization method to this problem, using as the statistic $T(E)$ the numerical value of the (sample) Pearsonian correlation coefficient,

$$T(E) = \left|\sum_{i=1}^{m} x_i x_{i+m}\right| \Big/ \left(\sum_{i=1}^{m} x_i^2 \sum_{j=m+1}^{2m} x_j^2\right)^{\frac{1}{2}},$$

the large values of T being the significant ones. We note that $T(E)$ takes on at most $m!$ different values over the subpopulation. What we previously called a "suitably chosen" q would be in the present case a multiple of $m!$, and the choice of significance level $\alpha = q/s$ would then be limited to multiples of $1/m!$.

The method of randomization is seen to exploit whatever symmetry properties the F_n in ω possess as a class. A special case of the general method is the *method of ranks*. This gives regions of an especially simple form defined by certain inequalities on the coordinates. Probably the only case in which the method of ranks will ever be used is when the F_n in ω have the following special kind of symmetry: Suppose they are completely symmetrical in each of certain subsets of the coordinates, say t sets of $n_1, n_2, \cdots, n_t$ coordinates, respectively, where $\sum_{i=1}^{t} n_i = n$. We may assume the coordinates numbered so that F_n is completely symmetrical in the set $x_{p_i+1}, x_{p_i+2}, \cdots, x_{p_i+n_i}$ $(p_i = \sum_{j=1}^{i-1} n_j; i = 2, 3, \cdots, t; p_1 = 0)$, for all $F_n \in \omega$. The set S of permutations is thus generated by making all $n_i!$ permutations on the n_i coordinates $x_{p_i+1}, \cdots, x_{p_i+n_i}$ $(i = 1, \cdots, t)$, so that the total number of permutations in S is $s = n_1!\, n_2! \cdots n_t!$.

Corresponding to the i-th set of coordinates in which F_n is symmetrical, let us divide the sample space W up into $n_i!$ regions defined by

$$x_{p_i+1} < x_{p_i+2} < \cdots < x_{p_i+n_i}$$

and the $n_i! - 1$ other inequalities obtained by permuting the subscripts in the above. Denote these regions by $w_{i,k}$ $(k = 1, \cdots, n_i!)$. Let

$$w_{k_1,k_2,\cdots,k_t} = w_{1,k_1} \cap w_{2,k_2} \cap \cdots \cap w_{t,k_t},$$

that is, $w_{k_1,k_2,\cdots,k_t}$ is the part of W common to the regions $w_{1,k_1}, w_{2,k_2}, \cdots, w_{t,k_t}$. This process divides the sample space W up into s disjoint regions $w_{k_1,k_2,\cdots,k_t}$, which we shall now denote simply by w_σ $(\sigma = 1, \cdots, s)$. The set $\{w_\sigma\}$ of regions covers all of the sample space W except the region W_0 on which certain coordinates become equal. We shall say that the sample point $\mathbf{E}$ has the σ-th ranking, R_σ, if $\mathbf{E}$ falls in w_σ. We may then speak of a random variable $R = R(\mathbf{E})$, the "ranking", taking on the s possible values R_σ, or the "tied" ranking R_0 if $\mathbf{E} \in W_0$.

A critical region w is constructed by the method of ranks by taking w to be the union of q of the regions w_σ. Those rankings R_σ corresponding to the q regions w_σ constituting the critical region w, will be called the significant rankings. Any statistic $T(E)$ used as the criterion to decide which are the significant rankings now becomes a function of the ranking R only, $T(E) = U(R)$. We may regard the method of ranks as a simplification of the problem of testing statistical hypotheses, in which the infinite n-dimensional sample space W is replaced by a finite space of $s + 1$ points R_σ. If Ω is a class of continuous F_n we may ignore the point R_0 since then $Pr\{R = R_0\} = 0$.

In the problem of independence, which we have used before to illustrate the definitions of this section, the method of ranks was applied by Hotelling and Pabst [9], who took as the statistic $U(R)$ the numerical value of the Spearman coefficient of rank correlation, large values being significant.

The method of randomization yields similar regions if ω is a class of continuous functions. What will the method get us if we drop the continuity restriction? In this case we can no longer ignore the possibility that the sample point $\mathbf{E}$ fall in the exceptional region W_0, for we do not have $Pr\{\mathbf{E} \in W_0\} = 0$. We owe to Pitman [27] the following idea: We continue to use the subpopulation $\{E'\}$ and a chosen statistic $T(E)$ as above, but instead of separating the points of $\{E'\}$ into two classes (significant points and non-significant points) by means of $T(E)$ we now add a third class of "doubtful" points.[5] If the s points of the set $\{E'\}$ are not distinct they are to be counted according to their multiplicities under the process of applying the permutations of the set S to the coordinates of E. Suppose that the large values of T are significant. Number the s points of $\{E'\}$ so that $T(E'_1) \geq T(E'_2) \geq \cdots \geq T(E'_s)$. If $T(E'_q) > T(E'_{q+1})$ we call $E'_1, \cdots, E'_q$ significant, and the rest non-significant. However if $T(E'_q) = T(E'_{q+1})$, we term all points E' with $T(E') = T(E'_q)$ doubtful, points E' for which $T(E') > T(E'_q)$, significant, and points E' with $T(E') < T(E'_q)$, non-significant. This process divides the sample space W up into three regions instead of the customary

[5] Instead of the terms significant, non-significant, doubtful, Pitman uses discordant, concordant, neutral.

two, namely, a rejection region w_R, an acceptance region w_A, and a doubtful region w_D. It is a special case of the following procedure: For every set $\{E'\}$ define positive integers $m_R = m_R(\{E'\})$ and $m_A = m_A(\{E'\})$ such that $m_R \leq q$, $m_A \leq s - q$, and put m_R of the points E' in w_R, m_A of the points E' in w_A, and the remaining $s - m_A - m_R$ of the points E' in w_D, in any way so that w_R and w_A are Borel regions. When any E' is assigned to w_R or w_A it is to be counted according to its multiplicity as defined above, if $\{E'\}$ contains less than s distinct points. It may be shown that with $\alpha = q/s$, $Pr\{\mathbf{E} \in w_R \mid F_n\} \leq \alpha$ and $Pr\{\mathbf{E} \in w_A \mid F_n\} \leq 1 - \alpha$ for all $F_n \in \omega$, that is, whenever H is true.

Before closing this section on the method of randomization, we mention a few difficulties which frequently arise when it is applied. Except for very small samples the calculation determining whether or not the observed value E_0 of the sample point $\mathbf{E}$ belongs to the significant points of the subpopulation $\{E_0'\}$ generated by E_0, is usually extremely tedious. In such cases the author of the test often gives an approximation to the discrete distribution of his statistic $T(\mathbf{E})$ over the subpopulation $\{E'\}$ by means of some familiar continuous distribution for which tables are available, the laborious exact calculation by enumeration then being replaced by the computation of a few moments (that is, values of certain homogeneous polynomials in the observed coordinates) and the use of existing tables of percentage points of the continuous distribution.[6] Barring some papers where the method of ranks is used, the justification of these approximations is never satisfactory from a mathematical point of view, the argument being based on a study of the behavior of two, or at most four, moments. The only exception to the last statement appears to be a very recent paper by Wald and Wolfowitz [42], which may point the way to genuine derivations of asymptotic distributions for the non-rank case of the randomization method. We shall distinguish between derivations of asymptotic distributions and arguments based on two or four moments by saying that a distribution is "proved" in the former case and "fitted" in the latter.

Another difficulty arises, most noticeably in the method of ranks, out of the possibility of equality of the observed coordinates. In the distribution theory this is usually avoided by assuming ω to be a class of continuous c.d.f's, so that $Pr\{\mathbf{E} \in W_0 \mid F_n\} = 0$ for all $F_n \in \omega$, but in practice, since the measurements are usually made to about three significant figures, ties *do* occur in the sample. While some scattered work has been done on this question there is need for a thorough general treatment.

In some of the work that has been done on particular non-parametric tests

[6] In many cases the approximate test obtained by fitting a familiar distribution is found to coincide with widely used tests based on normality assumptions. In such cases if the fitting is asymptotically correct the following remarks are justified: (1) If the non-parametric test is used in a case where we hesitate to assume normality but normality actually exists, the non-parametric test is asymptotically as efficient as the older test assuming normality. (2) If normality is assumed when it does not exist, no error is incurred asymptotically when the older test is used.

it is not very clear just what the null hypothesis H is. Two situations often occur: Suppose $H: F_n \in \omega$ is the hypothesis we actually wish to test at significance level α. Let w be the chosen critical region and ω_w the class of F_n for which $Pr\{\mathbf{E} \in w \mid F_n\} = \alpha$. The two situations are (*i*) ω is a proper subset of ω_w, and (*ii*) ω_w is a proper subset of ω. Of these (*i*) seems less objectionable, for then the probability of a type I error (rejecting H when true) is strictly α, but the probability of accepting H is the same when certain alternatives ($F_n \in \omega_w - \omega$) are true as when H is true. In case (*ii*) the probability of a type I error is not α unless F_n is in the subclass ω_w of ω; thus there might be a much higher probability than α of rejecting H when it is true, if the true $F_n \in \omega - \omega_w$. To illustrate situation (*i*) consider K. Pearson's χ^2-test for goodness of fit of a theoretical distribution $F_0(x)$ to a sample $\mathbf{E}$. Suppose $\mathbf{E}$ is from a univariate population whose true c.d.f. is $F(x)$. If F has the property that for the intervals I_j defined in section 3, $\int_{I_j} dF = \int_{I_j} dF_0$, $j = 1, 2, \cdots, N$, then the probability of rejection is the same as when the hypothesis is true. An example of (*ii*) might occur if we wish to test whether the *means* of two univariate populations are the same. If we use one of the tests of section 4 in which the probability of rejection is calculated under the assumption that the *distributions* of the populations are the same, then we do not know that the probability of a type I error is α, for the samples might come from two populations with the same mean but different distributions.

3. Goodness of fit. Randomness. The non-parametric case of testing goodness of fit is the following: On the basis of a sample $\mathbf{E}$ from a population with c.d.f. $F(x)$ known to be a member of some Ω_ν, we wish to test whether $F = F_0$, where F_0 is a given c.d.f. The class of admissible c.d.f.'s F_n is

$$\Omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F(x_i);\ F \in \Omega_\nu\right\},$$

and the hypothesis H specifies that $F_n \in \omega$, where

$$\omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F_0(x_i)\right\}.$$

K. Pearson's χ^2-test [25] consists of choosing an integer N, dividing the x-axis up into a set $\{I_j\}$ of disjoint intervals ($j = 1, 2, \cdots, N$), and using as statistic $T(E)$ the Pearsonian chi square,

$$\chi_P^2 = \sum_{j=1}^{N} [m_j - \mathfrak{E}(m_j)]^2/\mathfrak{E}(m_j),$$

where m_j is the number of observed coordinates of $\mathbf{E}$ in I_j, and $\mathfrak{E}(m_j) = n\int_{I_j} dF_0$. Large values of χ_P^2 are regarded as significant. Exact significance

levels for χ_P^2 could be obtained by considering its distribution over the subpopulation $\{E'\}$ generated by the sample. This process would lead to the multinomial distribution of the m_j mentioned in the usual derivations of the asymptotic distribution of χ_P^2 (for $n \to \infty$ with N fixed). Pearson himself found this asymptotic distribution, namely the χ^2-distribution with $N - 1$ degrees of freedom. In studying the problem of a "best" choice of the set $\{I_j\}$ of intervals, Mann and Wald [17] adopted a non-parametric treatment, with $\nu = 2$ for the class Ω_ν above.

Another test not depending on a choice of intervals I_j could be made by using confidence belts for F as described in section 9 and rejecting H at the α level of significance if the graph of F_0 is not covered by the belt with confidence coefficient $1 - \alpha$.

The problem of randomness is usually non-parametric; in the univariate case the class ω of this problem is identical with the class Ω of the preceding. The index ν and the class Ω for the problem of randomness would depend on the specific situation in which it arises. With two exceptions [42, 52], all tests of randomness proposed thus far have been functions of runs in the sample. Two kinds of runs have been considered, runs up and down, and runs above and below the median [1, 4, 14, 19, 32, 44, 51]. We note that the set S of permutations determined by ω is the set of all $n!$ permutations on the n coordinates of E. Suppose now $\nu = 2$. The proof [31] that all similar regions w have the randomization structure applies to this problem. On the other hand such a region w has the property $Pr\{\mathbf{E} \in w \mid F_n\} = \alpha$ for any F_n which is completely symmetrical in the coordinates. Difficulty (i) discussed at the end of section 2 now arises if Ω contains such symmetrical alternatives. The definition of an appropriate class $\Omega - \omega$ of alternatives and the question of the power of tests against the alternatives make the problem of randomness a difficult one. Beyond these few remarks we refer the reader to an expository paper by Wolfowitz [51] devoted to the problem in the previous issue of this journal, and to a paper by Wald and Wolfowitz [42] in the present issue. The latter paper is one of the exceptions, previously mentioned, not based on the method of ranks.

4. The problem of two samples. Suppose $X_1, \cdots, X_{m_1}$ and $Y_1, \cdots, Y_{m_2}$ are samples from univariate populations with c.d.f's $F(x)$ and $G(x)$ respectively, where we assume $F, G \in \Omega_\nu$, and that we wish to test the hypothesis that $F = G$. Write $Y_i = X_{i+m_1}$, so that with $n = m_1 + m_2$ we have

$$\Omega = \left\{F_n \mid F_n = \prod_{i=1}^{m_1} F(x_i) \cdot \prod_{j=m_1+1}^{n} G(x_j);\ F, G \in \Omega_\nu\right\},$$

$$\omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F(x_i);\ F \in \Omega_\nu\right\}.$$

The set S of permutations determining the subpopulation $\{E'\}$ consists of all $n!$ permutations on the n coordinates of E. The writer has shown [31] that no

similar regions exist in this case if $\nu = 0$, while if $\nu = 2$, 3, or 4 a similar region necessarily has the randomization structure.

The first non-parametric attack on this problem was given [26] by K. Pearson. The x-axis is divided up into intervals $I_1, \cdots, I_N$ as in section 3. Let m_{j1} and m_{j2} be the number of measurements from the first and second samples, respectively, falling in I_j, so that $\sum_{j=1}^{N} m_{jk} = m_k$, $k = 1, 2$. The statistic $T(E)$ used is

$$\chi^2_{P'} = (m_1 m_2)^{-1} \sum_{j=1}^{N} (m_1 m_{j2} - m_2 m_{j1})^2/(m_{j1} + m_{j2}),$$

with large values significant. In view of the remarks at the end of the last paragraph it would be necessary to calculate the distribution of $\chi^2_{P'}$ over the subpopulation $\{E'\}$ in order to get a similar region. Pearson found the asymptotic distribution of $\chi^2_{P'}$ under the null hypothesis to be the χ^2-distribution with $N - 1$ degrees of freedom.

A solution based on the method of randomization was proposed by Pitman [27]; the special case of this solution for $m_1 = m_2$ was published a little earlier by R. A. Fisher [6]. Pitman employed the numerical value of the difference of the sample means as statistic,

$$T(E) = \left| \sum_{i=1}^{m_1} x_i/m_1 - \sum_{j=m_1+1}^{n} x_j/m_2 \right|,$$

large values being significant. He fitted an incomplete Beta-distribution to the subpopulation distribution of his $T(E)$, and noted that this approximation gave a result identical with the usual t-test valid when the population distributions $F(x)$ and $G(x)$ are assumed normal with equal variances.

Turning now to tests based on the method of ranks, we mention here that one for the case $m_1 = m_2$ was given by R. A. Fisher as early as 1925, namely the "sign test" or "binomial series test" [3]. We may (and Fisher did) regard this as a test of a less restrictive hypothesis, and shall describe it in section 6. Between 1938 and 1940 several tests employing ranks were proposed for the problem of two samples. The earliest of these, by W. R. Thompson [36], was shown to be inconsistent (section 11) with respect to certain alternatives $F_n \in \Omega - \omega$ by Wald and Wolfowitz [40]. These authors used as statistic $U(R)$ the total number of runs in a sequence V of n elements constructed as follows: Rank the measurements of the combined sample in order of increasing magnitude. According as the j-th measurement in this rank order is from the first or second sample, put the j-th element of the sequence V equal to 1 or 2. In this test small values of the statistic $U(R)$ are regarded as significant. The test is now quite practicable (for $\nu = 2$) for certain ranges of m_1 and m_2. For m_1 and m_2 both ≤ 20, tables by Swed and Eisenhart [34] give the 1% and 5% significant values of $U(R)$. Wald and Wolfowitz proved that for $n \to \infty$ with $k = m_1/m_2$ fixed, the distribution of $U(R)$ is asymptotically normal with mean $2m_1/(1 + k)$ and variance $4km_1/(1 + k)^3$. Swed and Eisenhart computed that for $m_1 = m_2$ this

gives a very satisfactory approximation outside the range of their tables. However, further computation needs to be done on the accuracy of this approximation for $m_1 \neq m_2$ and one of them >20.

Another test based on ranks was advanced by Dixon [2], using as statistic $U(R)$ the random variable

$$C^2 = \sum_{j=1}^{m_1+1} [(m_1 + 1)^{-1} - n_j/m_2]^2,$$

where the integers n_j are defined thus: Let $Z_1 \leq Z_2 \leq \cdots \leq Z_{m_1}$ denote the measurements of the first sample arranged in rank order. Then n_j is the number of measurements in the second sample falling in the interval (Z_{j-1}, Z_j), where we define $Z_0 = -\infty$, $Z_{m_1+1} = +\infty$. Large values of C^2 are significant. Dixon tabulated the 1%, 5%, and 10% significant values of C^2 for $m_1, m_2 = 2, 3, \cdots, 10$; for larger m_1, m_2 he fitted a χ^2-distribution.

A paper by Smirnoff [33, 16] suggests the following as a statistic $U(R)$: Let $G_{m_1}(x)$ and $G_{m_2}(x)$ be the "empirical distribution functions" of the first and second samples, that is, $m_i G_{m_i}(x)$ is the number of measurements in the i-th sample $\leq x$ $(i = 1, 2)$ and take[7]

$$U(R) = (m_1^{-1} + m_2^{-1})^{-1} \sup_x |G_{m_1}(x) - G_{m_2}(x)|$$

with large values significant. Smirnoff showed that if $\nu = 2$ the asymptotic distribution of his statistic $U(R)$ is a certain c.d.f. $\Phi(\lambda)$, previously introduced by Kolmogoroff [15]. More specifically, let $\Phi_{m_1,m_2}(\lambda) = Pr\{U(R) \leq \lambda; \nu = 2\}$. Then if $n \to \infty$ with m_1/m_2 fixed, we have $\Phi_{m_1,m_2}(\lambda) \to \Phi(\lambda)$. The definition of $\Phi(\lambda)$ and references to tables of $\Phi(\lambda)$ are given in section 9. If instead of assuming $\nu = 2$ we take $\nu = 0$, the risk of type I errors may be controlled to the extent that $Pr\{\text{rejecting } H\} \leq \alpha$ for all $F_n \in \omega$, by employing Smirnoff's theorem stating $Pr\{U(R) \leq \lambda; \nu = 0\} \leq \Phi_{m_1,m_2}(\lambda)$, where $\Phi_{m_1,m_2}(\lambda)$ is defined above.

A test for the problem of two samples obtained by Wolfowitz by a modification of the likelihood ratio procedure will be discussed in section 12. When $m_1 = m_2$ the non-parametric analysis of variance tests of the "randomized blocks" type described in section 6 might also be used to test the more restricted hypothesis considered in this section.

The non-parametric problem of k samples has been attacked by Welch [46], who used the method of randomization with the analysis of variance ratio as statistic $T(E)$, and by Wolfowitz [50] with his modified likelihood ratio method.

In this as in all the other sections where several solutions of the same problem of statistical inference are described, the question as to the relative merits of the various solutions arises, and in every case the question is as yet mostly or entirely unanswered. The only easy conclusion about the tests of this section would seem to be that the tests of K. Pearson and Pitman are not consistent with

[7] We use the notations *sup* and *inf* respectively for least upper bound and greatest lower bound.

respect to certain subclasses of the admissible alternatives, according to the definition of section 11.

5. Independence. The classes Ω and ω defining the problem of independence have already been stated in section 2, in which we described Pitman's test [28] based on the randomization method and the use of $|r|$ as statistic $T(E)$, where r is the sample value of the Pearsonian correlation coefficient. Pitman fitted an incomplete Beta-distribution to the subpopulation distribution of r^2 and found the resulting approximation for $\nu = 2$ equivalent to the usual test employing the t-distribution and valid for the case of normality.

In section 2 we also mentioned the test earlier proposed by Hotelling and Pabst [9], which is based on the method ranks and employs the statistic $U(R) = |r'|$, where r' is the Spearman rank correlation coefficient. They proved that for $\nu = 2$ the distribution of r' is asymptotically normal if $F_n \in \omega$. Pitman's fitting of an incomplete Beta-distribution applies also to $(r')^2$, and Kendall, Kendall, and Smith [12] made numerical calculations indicating that this gives a better approximation than the normal distribution. Since r' is calculated from Σd^2, the sum of the squared rank differences, the latter may equivalently be used as the statistic $U(R)$, small and large values of Σd^2 being now both significant. Kendall, Kendall, and Smith [12] found the exact distribution of Σd^2 for the number of pairs $m \leq 8$. This work was anticipated by Olds [23], who calculated the exact distribution of Σd^2 for $m \leq 7$, and by fitting certain distributions for $m > 7$, gave a very useful table of the 1%, 2%, 4%, 10% and 20% significant values of Σd^2 for $m \leq 30$. It would be desirable to have these tables extended by inclusion of the 5% values.

M. G. Kendall [10] proposed another measure of rank correlation whose significant values are easier to calculate than those of Σd^2, but since the Olds' tables for the latter are available, Kendall's innovation does not seem to possess much practical advantage. Wolfowitz [50], using his modified likelihood ratio method, gave another test for independence and generalized it to the problem of independence of k random variables.

6. Analysis of variance. We suppose that we have $n = rc$ measurements arranged in a rectangular layout of r rows and c columns. The r rows might correspond to the blocks and the c columns to the varieties in an agricultural experiment. The null hypothesis H is that of "no difference" in the column effects. The measurement in the i-th row and j-th column is supposed to be on a random variable[8] X_{ij} with c.d.f. $F^{(ij)}(x) = Pr\{X_{ij} \leq x\}$. Let us assume at first that all the X_{ij} are independent. The joint c.d.f. of the random variables $X_{1j}, \cdots, X_{rj}$ of the j-th column is then

$$F^{(j)}(x_1, \cdots, x_r) = Pr\{x_{1j} \leq x_1, \cdots, x_{rj} \leq x_r\} = \prod_{i=1}^{r} F^{(ij)}(x_i).$$

[8] The double subscript notation is more convenient here than that used in section 2; after the class ω has been defined the reader will see that the numbers n_i used in section 2 to describe the symmetry of the $F_n \in \omega$ are all equal to c, and the X_{ij} of the present section coincides with the X_{p_i+j} of section 2.

The symbol F_n for the joint c.d.f. of all n random variables now denotes $F_n(x_{11}, \cdots, x_{1c}; \cdots; x_{r1}, \cdots, x_{rc})$. Ω is the class of all F_n of the form

$$F_n = \prod_{j=1}^{c} F^{(j)}(x_{1j}, \cdots, x_{rj}),$$

where $F^{(j)}$ is defined by the preceding equation, and all $F^{(ij)}$ are in a given class Ω_r. The hypothesis H states that the column distributions are all the same,

$$F^{(j)}(x_1, \cdots, x_r) = F^{(1)}(x_1, \cdots, x_r) \qquad (j = 2, 3, \cdots, c),$$

without specifying $F^{(1)}$. ω is thus the subclass of Ω comprising all F_n of the form

$$F_n = \prod_{j=1}^{c} F^{(1)}(x_{1j}, \cdots, x_{rj}).$$

The F_n in ω may be written

$$F_n = \prod_{i=1}^{r} \left\{ \prod_{j=1}^{c} F^{(i1)}(x_{ij}) \right\}.$$

Regarding the factor in braces for fixed i, we see that it is left unchanged by any permutation of the c coordinates $x_{i1}, \cdots, x_{ic}$. The set S of permutations is thus determined, and the subpopulation $\{E'\}$ consists of the $(c!)^r$ points obtained by permuting among themselves the first set of c coordinates, the second set of c coordinates, $\cdots$, the r-th set of c coordinates of $E = (x_{11}, \cdots, x_{1c}; \cdots; x_{r1}, \cdots, x_{rc})$.

The above argument leading to the subpopulation $\{E'\}$ of $(c!)^r$ points is based squarely on the assumed independence of the n random variables X_{ij}. Suppose now that the X_{ij} are not known to be independent, as may happen in agricultural experiments [24]. To make the discussion concrete suppose in the $r \times c$ layout we have been considering, the rows refer to blocks (of plots) and the columns to varieties, so that the random variable X_{ij} is the yield of the j-th variety on the i-th block. We owe to R. A. Fisher the method of including early in the experiment a random process which leads to the same "equally likely" subpopulation of points $\{E'\}$ obtained before in the case of independence. This physical process which he calls "randomization" then permits the construction of critical regions by the "method of randomization" in the sense we have been using the term.

To explain the experimental process of randomization we shall imagine another $r \times c$ layout and a random set of mappings of the two layouts onto each other. In each block there are c plots and we now assume these numbered from 1 to c, the numbering to be held fixed. The second layout refers to the plots; the rows again correspond to the blocks, but the columns now correspond to the number of the plot in the block, thus the i, j cell represents the j-th plot in the i-th block. Now consider all 1:1 correspondences or mappings between the two layouts so that the i-th row always maps onto the i-th row ($i = 1, \cdots, r$). There are $s = (c!)^r$ such mappings M_k ($k = 1, \cdots, s$). Suppose under the mapping M_k the i, t cell in the block-plot layout maps on the i, j_k cell of the block-variety

layout, where $j_k = j_k(i, t)$, and the i, j cell of the latter corresponds to the i, t_k cell of the former, $t_k = t_k(i, j)$. The physical randomization process consists of choosing the mapping M_k so that all s mappings have the same probability $1/s$ of being chosen. In other words, the randomized block pattern is selected in such a way that all the s possible patterns have equal probabilities of being adopted in the experiment. Now let $Y_{it}^{(k)}$ be the yield of the i, t plot if the variety assigned to it by the k-th pattern is planted there, and let $G^{(k)}(y_{11}, \cdots, y_{rc}) = Pr\{\text{all } Y_{ij}^{(k)} \leq y_{ij}\}$ be the joint c.d.f. of the $Y_{ij}^{(k)}$. In calculating the c.d.f. F_n of the X_{ij} associated with the first layout we must take account of the random process by which it is mapped onto the second:

$$\begin{aligned} F_n(x_{11}, \cdots, x_{rc}) &= Pr\{\text{all } X_{ij} \leq x_{ij}\} \\ &= \sum_{k=1}^{s} Pr\{\text{all } X_{ij} = Y_{i,t_k(i,j)}^{(k)}\}\ Pr\{Y_{i,t_k(i,j)}^{(k)} \leq x_{ij}\} \\ &= \sum_{k=1}^{s} s^{-1} G^{(k)}(x_{1,t_k(1,1)}, \cdots, x_{r,t_k(r,c)}). \end{aligned}$$

Ω consists of all F_n of the above form with $G^{(k)}$ in a given class, say $\Omega_\nu^{(n)}$. The hypothesis H of "no difference" of varieties asserts that the yields of the plots do not depend on the varieties planted on them, that is, that all $G^{(k)}$ are the same, $G^{(k)} = G^{(1)}$, without specifying $G^{(1)}$. ω is the subclass of Ω whose members are of the form

$$F_n = s^{-1} \sum_{k=1}^{s} G^{(1)}(x_{1,t_k(1,1)}, \cdots, x_{r,t_k(r,c)}).$$

It is now seen that any permutation in the set S previously considered merely rearranges the terms of the above sum, so that F_n remains invariant, and we have the same subpopulation $\{E'\}$ as before.

It is to be understood henceforth that either the X_{ij} are known to be independent or else an experimental randomization has been carried out as described above, so that in either case the above set $\{E'\}$ of $(c!)^r$ points is the "equally likely" subpopulation.

The first application in the literature of the randomization method is found in R. A. Fisher's "sign test" or "binomial series test" [3] for the case of randomized blocks with two columns ($c = 2$). Let D_i be the difference $X_{i1} - X_{i2}$. The statistic used is a function of the ranking only, namely the number of $D_i > 0$, small and large values being significant. For $\nu = 2$ its distribution under the null hypothesis is the binomial distribution with the n and p of the usual notation equal respectively to r and $\frac{1}{2}$. This test may be regarded as the special case when $c = 2$ of Friedman's rank method for analysis of variance to be described below.

Fisher later [5] proposed another test for the case $c = 2$ not based on ranks, and employing as statistic $T(E)$ the absolute value of the mean of the D_i defined above, with large values significant. The exact distribution of this statistic is very laborious to calculate unless r is very small, and K. R. Nair [20] pointed

out that the use of the numerical value of the median of the D_i (or one of the two central values when r is even) had the advantage of a very easily calculated distribution (if $\nu = 2$). The latter may be regarded as a modification of the rank method, the method of ranks being applied not in the $2r$-dimensionsl sample space as described in section 2 but in the r-dimensional space of the differences D_i. Nair also showed that the distributions of the range and of the midpoint of the range of the D_i are very simple.

From here on we consider the general case $c \geq 2$, but when we speak of distributions they will be understood to be for the case when the null hypothesis is true and $\nu = 2$. Welch [45] considered using as $T(E)$ the usual analysis of variance ratio appropriate to testing for "no difference" of column effects. He transformed this to another statistic and calculated two moments of its subpopulation distribution. The first moment always agrees with that obtained under "normal theory", that is for the case $X_{ij} = C_i + Z_{ij}$, where the C_i are constants and the Z_{ij} are independently normally distributed with the same variance and zero means, but the second moment depends on the subpopulation $\{E'\}$. Here the exact distribution of the statistic is of course in general much more tedious to calculate than in the previous case $c = 2$; an incomplete Beta-distribution was fitted by Welch. Welch anticipated Pitman [29] who obtained the same results and got besides the third and fourth moments of Welch's statistic.

The method of ranks was applied by Friedman [7] who employed as statistic $U(R)$ a quantity formed as follows: Rank each set of row entries X_{ij} (for fixed i) in ascending order of magnitude, and let r_{ij} be the rank of X_{ij}, so that $r_{i1}, \cdots, r_{ic}$ is a rearrangement of the integers $1, \cdots, c$. Let $\bar{r}_j$ be the mean rank of the j-th column, $\bar{r}_j = \sum_{i=1}^{r} r_{ij}/r$, and take for $U(R)$

$$U = C_{rc} \sum_{j=1}^{c} [\bar{r}_j - \mathcal{E}(\bar{r}_j)]^2,$$

where C_{rc} is a certain constant, and $\mathcal{E}(\bar{r}_j)$ is calculated under the null hypothesis. For Friedman's choice of C_{rc}, U may be rapidly computed from the equivalent formula

$$U = -3r(c + 1) + 12 \sum_{j=1}^{c} \left(\sum_{i=1}^{r} r_{ij} \right)^2 \Big/ [rc(c + 1)].$$

In his paper Friedman included a proof of Wilks' that U has asymptotically the χ^2-distribution with $c - 1$ degrees of freedom as $r \to \infty$. Kendall and Smith [13] fitted to a transform of U a Fisher z-distribution with continuity corrections, obtaining a better approximation for small r than the χ^2-distribution. Wallis [43] independently proposed the use of $\eta_r^2 = U/[r(c - 1)]$ as statistic and called it the rank correlation ratio. Friedman in a later paper [8] on the subject, using exact values he had calculated, together with the Kendall-Smith approximation, published tables[9] of the 1% and 5% significant values of U for $c = 3, 4, 5, 6,$

[9] In these tables our U, r, c are denoted respectively by χ_r^2, m, n.

7, and sufficiently many values of r so that for these c and any r the significant values of U are now easily available.

After the above lengthy discussion for the "randomized blocks" case of analysis of variance, it will perhaps suffice merely to mention that the "Latin square" case may be similarly attacked from the non-parametric point of view, and this has been considered by Welch [45], Pitman [29], and E. S. Pearson [24]. They have taken as the statistic the usual analysis of variance ratio, and the work of Welch and Pitman in calculating the first two moments of its subpopulation distribution is even more tedious than in the "randomized blocks" case.

Part II. Non-parametric Estimation

7. Classical results on point estimation. Throughout part II the symbol **E** will always denote a random sample $X_1, \cdots, X_n$ from a univariate population with c.d.f. $F(x)$, where F is an unknown member of a given class to be stated in each case. The c.d.f. of **E** is thus

$$F_n(x_1, \cdots, x_n) = \prod_{i=1}^{n} F(x_i).$$

The problems of estimation can be stated without reference to the class Ω of admissible F_n; Ω would be obvious in every case.

Let $\theta = \theta(F)$ be a real number determined by F (a functional of F) for F in a certain class of univariate c.d.f's. Thus θ might be the mean of the distribution, in which case θ would be defined for all F possessing a first moment. We shall not call θ a parameter in order to avoid confusion with the parametric case. R. A. Fisher's criteria of unbiasedness and of consistency for point estimation carry over without change from the parametric case. A statistic $T(E)$ is said to be an unbiased estimate of θ if $\mathfrak{E}(T) = \theta$. Write $\mathbf{E} = \mathbf{E}_n$ and $T = T_n$ to emphasize the sample size n, and assume that the statistic $T_n(\mathbf{E}_n)$ is defined for all n (or all $n >$ some n_0). Then we define $T_n(\mathbf{E}_n)$ to be a consistent estimate of θ if it converges stochastically to θ, that is, if $Pr\{| T_n - \theta | > h\} \to 0$ as $n \to \infty$, for every $h > 0$.

In the present paragraph it will be convenient to symbolize the class of F for which the i-th (absolute) moment exists; we denote it by $\Omega_{(i)} (i = 1, 2, \cdots)$. It is known[10] that a sufficient condition for the stochastic convergence of the sample mean $\bar{x}$ to the population mean is that $F \in \Omega_{(1)}$. Hence for all $F \in \Omega_{(1)}$, $\bar{x}$ is a consistent estimate of the population mean; furthermore it is unbiased. If we apply this result to the random variable $Y = X^2$, we find that for all $F \in \Omega_{(2)}$, $\sum_{i=1}^{n} x_i^2/n$ is a consistent unbiased estimate of the second moment of F about the origin. Similar statements may be made for higher moments. For $F \in \Omega_{(2)}$ one may show further that with Q defined as $\sum_{i=1}^{n} (x_i - \bar{x})^2$, the statistics Q/n and $Q/(n-1)$ are consistent estimates of the population variance, and the latter is unbiased.

[10] See, for example, J. L. Doob, *Annals of Math. Stat.*, Vol. 6 (1935), p. 163.

If there exists a number M such that $F(M) = \frac{1}{2}$, it is called the median of the distribution. The median $\tilde{x}$ of a sample of odd size is the central X_i when the X_i are arranged in order of magnitude; for a sample of even size we may take $\tilde{x}$ to be the average of the two central values. It may be shown[11] that $\tilde{x}$ is a consistent estimate of M for F in the subclass of Ω_3 for which the probability density function $f(x)$ is continuous at $x = M$ and $f(M) \neq 0$.

8. Confidence intervals for an unknown median, for the difference of medians. Arrange the sample in rank order and denote the result by $Z_1 \leq Z_2 \leq \cdots \leq Z_n$, where $Z_1, \cdots, Z_n$ is a rearrangement of $X_1, \cdots, X_n$. The joint distribution of the Z_i (or any subset of the Z_i) is well known [49] if $F(x)$ is restricted to Ω_4, which we now assume. From this distribution theory it is easy to show that for any positive integer $k < \frac{1}{2}n$, the probability that the random interval (Z_k, Z_{n-k+1}) cover the unknown population median M is

$$Pr\{Z_k \leq M \leq Z_{n-k+1}\} = 1 - 2I_{\frac{1}{2}}(n - k + 1, k),$$

where

$$I_x(p, q) = \int_0^x t^{p-1}(1 - t)^{q-1}\,dt \Big/ \int_0^1 t^{p-1}(1 - t)^{q-1}\,dt$$

is the incomplete Beta-distribution tabulated by K. Pearson. The practicability of estimating M by means of the above relation in the non-parametric case was noted first by W. R. Thompson [35]. It is not difficult to calculate tables giving, for various sample sizes n, the maximum k for which $Pr\{Z_k \leq M \leq Z_{n-k+1}\} \geq .95$ or .99. This has been done for $n = 6$ to 81 by K. R. Nair [21], who listed the maximum k as well as $n - k + 1$ and $I_{\frac{1}{2}}(n - k + 1, k)$, so that the exact confidence coefficient is available. Nair also gave asymptotic formulas which are very accurate for $n > 81$.

It is clear how confidence intervals for the difference $d = M_2 - M_1$ of the medians of two univariate populations with c.d.f's known only to be in Ω_4 might be obtained by combining two probability statements of the above kind: Let the desired confidence coefficient be $1 - \alpha$, and form confidence intervals of the above type for M_1 and M_2 with confidence coefficient $1 - \frac{1}{2}\alpha$; write them $Pr\{\underline{M}_i \leq M_i \leq \bar{M}_i\} \geq 1 - \frac{1}{2}\alpha$. Then $Pr\{\underline{M}_2 - \bar{M}_1 \leq d \leq \bar{M}_2 - \underline{M}_1\} \geq 1 - \alpha$. Solutions like this which are easily obtained by the combining method in many problems are in general not very efficient.

Some work of Pitman's [27] may be regarded as a solution of the problem of estimating the difference of medians (or other quantiles, or means) of two populations in a case essentially more restricted than the preceding, but more general than the corresponding parametric case in which the distributions are assumed to differ only in location. To describe the nature of Pitman's result,

[11] This follows from the asymptotic distribution of $\tilde{x}$. See, for instance, [49], and combine section 4.53 with Theorem (A), p. 134.

let us revert to the notation introduced at the beginning of section 4, but add to the assumption that F and G are in a known class Ω_ν the *restrictive assumption* that F and G differ only in location, that is, that $G(x) = F(x - d)$. The problem is the interval estimation of the unknown constant d. Define the random variables $Z_i = Y_i - d$. After noting that the $m_1 + m_2$ random variables $X_1, \cdots, X_{m_1}, Z_1, \cdots, Z_{m_2}$ are all independently distributed with the same c.d.f. F, Pitman was able to apply his results for the problem of two samples to show how functions $\underline{d}$ and $\bar{d}$ of $X_1, \cdots, X_{m_1}, Y_1, \cdots, Y_{m_2}$ could be calculated such that $Pr\{\underline{d} < d < \bar{d}\} \geq 1 - \alpha$ for $\nu = 0$, while for $\nu = 2$ the equality holds. After fitting an incomplete Beta-distribution Pitman found that the resulting approximate confidence intervals coincide with the well known ones employing the t-distribution and based on the assumption that F and G are normal with the same unknown variance.

9. Confidence limits for an unknown distribution function. Consider in an x, y-plane the graph g of the unknown c.d.f., g being the locus of the equation $y = F(x)$, and the possibility of covering g with random regions $\mathfrak{R}(\mathbf{E})$ depending on the sample $\mathbf{E}$. Wald and Wolfowitz [39] have shown how for given n and α it is possible in a large variety of ways to define regions $\mathfrak{R}(\mathbf{E})$ such that $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\}$, the probability that the random region $\mathfrak{R}(\mathbf{E})$ cover the unknown graph g, is $1 - \alpha$ for all $F \in \Omega_2$. Instead of describing their general method we shall limit ourselves to a special case. This is a very neat solution the necessary distribution theory for which was developed earlier by Kolmogoroff [15].

Let $G_n(x)$ be the "empirical distribution function" of the sample: $nG_n(x)$ is the number of $X_i \leq x$. Define the random variable

$$D_n = \sqrt{n} \sup_x | F(x) - G_n(x) |;$$

and let $\Phi_n(\lambda)$ be the c.d.f. of D_n, $\Phi_n(\lambda) = Pr\{D_n \leq \lambda\}$. Kolmogoroff proved that $\Phi_n(\lambda)$ is independent of $F \in \Omega_2$, and that as $n \to \infty$, $\Phi_n(\lambda) \to \Phi(\lambda)$ uniformly in λ, where $\Phi(\lambda)$ is defined by the rapidly converging Dirichlet series

$$\Phi(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2\lambda^2).$$

A small table of values of the function $\Phi(\lambda)$ was given by Kolmogoroff [15], and a larger one by Smirnoff [33]. Define $\lambda_{n,\alpha}$ from $\Phi_n(\lambda_{n,\alpha}) = 1 - \alpha$, and λ_α from $\Phi(\lambda_\alpha) = 1 - \alpha$. Values of λ_α for $\alpha = .05, .02, .01, .005, .002, .001$ were listed by Kolmogoroff [16]. Now $1 - \alpha$ is the probability that

$$\sqrt{n} \sup_x | F(x) - G_n(x) | \leq \lambda_{n,\alpha}$$

if $F \in \Omega_2$. The above inequality is equivalent to

$$G_n(x) - \lambda_{n,\alpha}/\sqrt{n} \leq F(x) \leq G_n(x) + \lambda_{n,\alpha}/\sqrt{n} \qquad \text{(all } x\text{)}.$$

If we take as $\mathfrak{R}(\mathbf{E})$ the intersection of the region between the graphs of the functions $G_n(x) \pm \lambda_{n,\alpha}/\sqrt{n}$, with the strip $0 \leq y \leq 1$, we have $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\} =$

$1 - \alpha$. The values of $\lambda_{n,\alpha}$ have not been tabulated, but for practical purposes of determining an unknown c.d.f. one would usually require a large n, and the tabulated values of λ_α could then be used.

With $\Phi_n(\lambda)$ defined as the c.d.f. of D_n for $F \in \Omega_2$, Kolmogoroff has shown further that for $F \in \Omega_0$, $Pr\{D_n \leq \lambda\} \geq \Phi_n(\lambda)$. This gives the beautiful result that the above confidence belt is valid in the most general case where $F \in \Omega_0$. in the sense that for the above defined $\mathfrak{R}(\mathbf{E})$, $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\} \geq 1 - \alpha$.

10. Tolerance limits. An ingenious formulation and solution of a non-parametric estimation problem was given by Wilks [47]. Let us say that an interval (x', x'') covers a proportion π of a population with c.d.f. $F(x)$ if $F(x'') - F(x') = \pi$. In the notation of section 8, Wilks considered the proportion B covered by the interval (Z_k, Z_{n-m+1}) extending from the k-th smallest observation to the m-th largest, $B = F(Z_{n-m+1}) - F(Z_k)$. B is a random variable depending on the sample but is *not* a statistic since it depends also on the unknown c.d.f. $F(x)$. However, Wilks noted that the c.d.f. $G(b)$ of B is independent of $F \in \Omega_4$, in fact, for $0 < b < 1$,

$$G(b) = I_b(n - k - m + 1, k + m),$$

where $I_x(p, q)$ is defined in section 8. After k, m, a fixed proportion b, and a confidence coefficient $1 - \alpha$ have been chosen, the equation $G(b) = \alpha$ determines the sample size n for which we can then make the following assertion without any knowledge of F except that $F \in \Omega_4$: The probability is $1 - \alpha$ that in a sample size n the random interval (Z_k, Z_{n-m+1}) will cover at least 100 $b\%$ of the population.[12]

Wilks considered, among other extensions of his method, tolerance limits for multivariate distributions in which the variables are known to be independent, and the estimation of proportions in a second sample (instead of in the population) on the basis of a first sample [48]. The latter problem involves the calculation of $P(b; n, N, k, m)$, the probability that if a first sample of n is taken and then a second sample of N, a proportion b or more of the second sample will lie in the interval (Z_k, Z_{n-m+1}) determined from the first sample. Wilks' derivation of P requires the assumption that $F \in \Omega_4$, but a simple auxiliary argument (related to the method of randomization by ranks) will extend the validity to the case $F \in \Omega_2$: The complete set of $n + N$ variates is independently distributed, each with the same c.d.f. $F \in \Omega_2$. All $(n + N)!$ possible rankings (excluding the "tied" ranking R_0) as defined in section 2 then have the same probability $1/(n + N)!$. The fraction of these rankings for which the statement about proportions in the second sample is correct is a function of b, n, N, k, m only, and not of $F \in \Omega_2$, and this fraction is the desired P. Since P is the same for all $F \in \Omega_2$ it must of course coincide with the value calculated by Wilks for $F \in \Omega_4$. It would be desirable for practical purposes to extend the validity of the tolerance

[12] For fixed b, $G(b)$ of course takes on discrete values with n, so one would either choose the n giving $G(b)$ the nearest value to α or else the greatest value $\leq \alpha$.

limits of the first paragraph, concerning proportions in the population, at least to the case $F \in \Omega_3$. The extension to Ω_2 would follow immediately if the intuitively reasonable statement $1 - G(b) = \lim_{N\to\infty} P(b; n, N, k, m)$ could be justified for $F \in \Omega_2$.

The multivariate case when independence is not assumed was successfully attacked by Wald [38]. We shall describe here his solution for the bivariate case: Let (X_i, Y_i), $i = 1, \cdots, n$, be a sample from a population with bivariate c.d.f. $F(x, y) \in \Omega_4^{(2)}$, that is, F is of the form

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi, \eta)\, d\eta\, d\xi,$$

where $f(x, y)$ is continuous, but otherwise unknown. Plot the points (X_i, Y_i) in an x, y-plane and choose four (small) integers k_1, m_1, k_2, m_2. . Draw vertical lines (parallel to the y-axis) passing through the points with the k_1-th smallest and m_1-th largest abscissas. Considering *only* the $n - k_1 - m_1$ points inside these vertical lines (the probability of equal abscissas is zero), draw two horizontal lines passing through the points with k_2-th smallest and m_2-th largest ordinates. Let J be the rectangle bounded by the four lines and consider the proportion B of the population covered by the rectangle, $B = \int_J dF(x, y)$. Then the c.d.f. $G(b)$ of B is given by the previous formula in terms of the incomplete Beta-distribution with $k + m = k_1 + k_2 + m_1 + m_2$, and is thus independent of $f(x, y)$. Choose k_1, k_2, m_1, m_2, b, and α. Then the equation $G(b) = \alpha$ determines the sample size n for which the probability is $1 - \alpha$ that the random rectangle J will cover at least 100 $b\%$ of the population. Wald showed further how a series of rectangles instead of a single rectangle might advantageously be used in the case of highly correlated X, Y.

It would be most useful to have tables of n corresponding to $\alpha = .05$ and $.01$, some values of b close to unity, and a few small values of $k + m$, say, $k + m = 2, 4, \cdots, 2r$. The table could then be used for the univariate, bivariate, $\cdots$, r-variate cases with various choices of k_j, m_j, such that $\Sigma(k_j + m_j) = k + m$. Entries for $k + m = 4$ have been given by Wald [38, p. 55].

Part III. Toward a General Theory

11. The criterion of consistency. All the concepts of Part III have been carried over from, or suggested by, corresponding ones earlier developed for the parametric theory. Consistency of point estimation was defined in section 7. Wald and Wolfowitz [40] have generalized the notion of consistency to tests so that it is applicable in the non-parametric case. We have heretofore specified the hypothesis H and its admissible alternatives by means of classes of n-variate c.d.f's F_n. We now assume that H and its admissible alternatives can be framed as statements about one or more populations, independent of n. Thus in the problem of two samples (section 4) H may be taken as the statement that the c.d.f's F and G of the two populations are the same member of Ω_r, while the

admissible alternatives are statements that F and G are any two different members of Ω. Returning to the general case, we assume that a sequence of tests is under consideration, say, $\mathfrak{T}_1, \mathfrak{T}_2, \cdots$, such that as $j \to \infty$, the size of the sample in $\mathfrak{T}_j$ from each of the populations becomes infinite. The sequence $\{\mathfrak{T}_j\}$ may be called simply a "test" and is said to be consistent if the probability of rejection of H by $\mathfrak{T}_j$ approaches unity as $j \to \infty$ whenever an admissible alternative to H is true. It has been suggested [50] that consistency is a minimal requirement for a good test. In order to allow for the analogue of the "common best critical regions" in the parametric theory,[13] it would be better to define consistency with respect to any given subset of the admissible alternatives and then require consistency with respect to the subset appropriate to the specific situation in which the test is to be used.

Wald and Wolfowitz [40] proved that under certain restrictions on the admissible F, G in the problem of two samples their test based on runs (section 4) is consistent, while another previously proposed test is not. Judging from their work, we may expect that, while inconsistency proofs may be easy, consistency proofs will be difficult.

12. Likelihood ratio tests. A definition of the Neyman-Pearson likelihood ratio criterion[14] λ for testing the hypothesis H (we use the notation of section 2), which would yield the usual result in the parametric case, would be the following: Let $C(E;\delta)$ be a cube of edge 2δ in the sample space W with center at the point E and faces parallel to the coordinate hyperplanes, and let $P(E;\delta \mid F_n)$ be the "probability put into the cube by the c.d.f. F_n", that is, $P(E;\delta \mid F_n) = \int_{C(E;\delta)} dF_n$. Define

$$\lambda(E;\delta) = [\sup_{F_n \in \omega} P(E;\delta \mid F_n)]/[\sup_{F_n \in \Omega} P(E;\delta \mid F_n)],$$

$$\lambda = \lambda(E) = \lim_{\delta \to 0} \lambda(E;\delta).$$

This definition of λ is not useful in the non-parametric case as λ turns out in general to be independent of E; the reader may easily verify this for the problem of two samples (section 4).

Having seen now that the likelihood ratio does not carry over to the non-parametric case in an obvious way, we are in a position to appreciate a bold stroke by Wolfowitz [50]. He begins by limiting the critical regions to be considered to the relatively small class obtainable by the method of ranks (section 2). Let $R = R(\mathbf{E})$ be the ranking of the sample point $\mathbf{E}$, so that the random variable R takes on the possible values $R_0, R_1, \cdots, R_s$, and let $P(R_\sigma \mid F_n) = Pr\{R = R_\sigma \mid F_n\}$.

[13] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses", *Phil. Trans. Roy. Soc. London*, A, Vol. 231 (1933), pp. 289-337.

[14] J. Neyman and E. S. Pearson, *Biometrika*, Vol. 20A (1928), p. 264.

Then Wolfowitz takes the likelihood ratio to be the following function of the ranking R:

$$\Lambda(R) = [\sup_{F_n \in \omega} P(R \mid F_n)]/[\sup_{F_n \in \Omega} P(R \mid F_n)].$$

His modified likelihood ratio test then consists of applying the method of ranks (section 2) with $\Lambda(R)$ as the statistic, small values being regarded as significant. If Ω is a class of continuous F_n, all rankings $R \neq R_0$ have the same probability $1/s$ under the null hypothesis, while $P(R_0 \mid F_n) = 0$ for all $F_n \in \Omega$. Then the numerator of $\Lambda(R)$ is $1/s$, and we may thus use the denominator of $\Lambda(R)$ as statistic with large values significant. Wolfowitz' modification has one advantage we don't always find with the usual parametric method: it always leads to similar regions since it is a special case of the randomization method.

In applying his method to examples Wolfowitz finds it necessary to resort each time to an approximation in calculating his statistic $\Lambda(R)$. Instead of taking the "sup" over Ω as in the definition, he takes it instead over a subclass Ω' of Ω which lends itself more easily to calculation. Thus in the problem of two samples with $\nu = 2$, whereas Ω is the class defined in section 4 with F, G in Ω_2, the class Ω' is the subclass of Ω obtained by further limiting F, G as follows: The x-axis is divided up into a number of disjoint intervals, equal to the total number of runs in the sequence V defined in connection with the Wald-Wolfowitz test in section 4. If the j-th run in V is a run of 1's the restriction $G(x) = 0$ in the j-th interval is imposed, if the j-th run is a run of 2's, $F(x) = 0$ in the j-th interval. The intervals in which F, G are permitted to assign positive probability then correspond in order and number to the two kinds of runs. With this restriction the (twice) modified likelihood ratio statistic is found to be

$$\sum_i \sum_j (l_{ij} \log l_{ij} - \log l_{ij}!),$$

where l_{ij} is the number of elements in the j-th run of i's ($i = 1, 2$). Large values are significant. For large samples the asymptotic distribution of the statistic falls out as a special case of a general theorem of Wolfowitz.

In the same paper Wolfowitz obtained modified likelihood ratio tests for the problem of k samples and the problem of independence of two or more random variables.

In his examples the author states that the maximizing F_n in Ω' is "essentially the same" as the maximizing F_n in Ω, at least for the significant rankings R_σ and for large samples. The necessity of this approximation procedure is somewhat disturbing, as is the restriction to the method of ranks. Since it does not seem possible to give a definition of likelihood ratio tests sufficiently broad to include the non-parametric case, yet yielding the usual result in the parametric case, we are denied even the small comfort of saying that at least in special cases the method is known to yield optimum results. In some problems the set $\{R_\sigma\}$ of rankings, corresponding to the set $\{w_\sigma\}$ of regions in W which serves to separate the s points of the subpopulations $\{E'\}$ defined in section 2, is not

unique—consider for instance the problem of two samples when the populations are bivariate—and in such cases the method as defined above would not give a unique result. These remarks are intended to point the need for further investigation and cannot detract from the ingenuity of the method—the first general process that has been suggested for choosing one out of the welter of similar regions yielded by the randomization method.

13. Wald's formulation of the general problem of statistical inference. A formulation of the general problem of statistical inference broad enough to cover the non-parametric case, and including estimation and tests as well as statistical problems classifiable under neither of these headings, has been given by Wald [37]. This formulation extends certain concepts he had applied earlier[15] to the parametric case.

In this last section we shall permit ourselves a somewhat more abstract terminology and notation than before. As in section 2, $\mathbf{E} = (X_1, \cdots, X_n)$ will denote the sample; $F_n(E)$, its c.d.f.; W, the n-dimensional Euclidean space of E, the sample space; and Ω, the space of admissible F_n. Of central importance is a given class $\mathfrak{S}$ appropriate to the problem, $\mathfrak{S} = \{\omega_\beta\}$, whose members ω_β are (not necessarily disjoint) subsets of Ω, $\Omega = \bigcup_\beta \omega_\beta$. To every $\omega_\beta \in \mathfrak{S}$ there corresponds a hypothesis $H(\omega_\beta): F_n \in \omega_\beta$, so that there is a 1:1 correspondence between the members of the set $\mathfrak{S}$ and those of the set $\{H(\omega_\beta)\}$ of hypotheses. The general problem of statistical inference, according to Wald, is the choice of a decision function $\Delta(E)$ mapping W into $\mathfrak{S}$. For every $E \in W$ a decision function $\Delta(E)$ uniquely selects an element ω_β of $\mathfrak{S}$, $\omega_\beta = \Delta(E)$. Its statistical import is that when the sample point $\mathbf{E}$ equals E, we agree to accept the hypothesis $H(\omega_\beta)$ determined by $\Delta(E) = \omega_\beta$.

Before introducing any further definitions let us illustrate the preceding ones. In any problem of testing a hypothesis, the set $\mathfrak{S}$ has just two members ω_1 and ω_2 which we have heretofore denoted by ω and $\Omega - \omega$, respectively. The decision function $\Delta(E)$ then takes on just these two values, in fact, $\Delta(E) = \omega_2$ for E in the critical region w of the test, and $\Delta(E) = \omega_1$ for $E \in W - w$.

To illustrate the definitions in the case of point estimation, consider estimating the median M of a univariate population with c.d.f. $F(x)$. Ω would be the class of F_n of the form $\prod_{i=1}^n F(x_i)$ with, say, $F \in \Omega_4$ and $F'(M) \neq 0$ (which is sufficient to insure a unique M). The index β could now be identified with M, so that its domain is the real line, and $\omega_\beta = \{F_n \mid M(F) = \beta\}$. The classes ω_β would be disjoint in this case and each would contain an infinite number of F_n. The problem of estimating the unknown M may be said to be the choice of a decision function $\Delta(E)$: When $\mathbf{E} = E$ we accept $H(\omega_\beta): F_n \in \omega_\beta = \Delta(E)$, meaning in this case simply that we accept the statement that M equals the β determined by $\Delta(E)$.

---

[15] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 299–326.

Suppose next that instead of the point estimation of M just discussed we are interested in the interval estimation of M. We define Ω as above, and now take the index β to consist of a pair a, b of real numbers. An interval estimate $a \leq M \leq b$ may be regarded as an acceptance of the hypothesis $H(\omega_{a,b}): F_n \in \omega_{a,b}$, where $\omega_{a,b}$ is the subclass of Ω consisting of all F_n for which $M(F)$ lies in the interval $a \leq M \leq b$. The set $\mathfrak{S}$ now consists of all classes $\omega_{a,b}$ with $-\infty < a < b < +\infty$. Here as in the general case of interval estimation the classes ω_β of the set $\mathfrak{S}$ are not disjoint. The decision function $\Delta(E)$ adopted in section 8 is $\Delta(E) = \omega_{a,b}$ with $a = z_k$, $b = z_{n-k+1}$, where $z_1 \leq z_2 \leq \cdots \leq z_n$ is a rearrangement of the coordinates $x_1, \cdots, x_n$ of E.

An example of a problem neither of estimation nor testing would be the following: Let Ω be as above. Two real numbers A and B ($A < B$) are given and it is required to decide on the basis of the sample $\mathbf{E}$ to which of the three classes $-\infty < M < A$, $A \leq M \leq B$, $B < M < +\infty$ the unknown median M belongs. Here the set $\mathfrak{S}$ would consist of three disjoint classes $\omega_1, \omega_2, \omega_3$: where ω_1 is the subclass of Ω consisting of F_n with $M(F) < A$, etc.

We return now to the general case. Before defining a "best" decision function $\Delta = \Delta^*$, Wald asks that there be a given weight function $w(F_n, \omega_\beta)$ defined on the product space $\Omega \times \mathfrak{S}$. The weight function $w(F_n, \omega_\beta)$ is a real-valued function evaluating the loss involved in accepting $H(\omega_\beta)$, the statement that the unknown c.d.f. of $\mathbf{E}$ is a member of ω_β, when the unknown c.d.f. is actually F_n. If $F_n \in \omega_\beta$ we make no error in accepting $H(\omega_\beta)$, and in this case w is defined to be zero. Its value otherwise is required to be non-negative. In this theory the choice of the weight function is regarded as essentially *not* a mathematical problem, but the choice is to stem out of the very specific situation in which the statistical inference is to be made. In an industrial problem w might be the financial loss incurred when a certain kind of error is made.

After w is given, the decision functions Δ are to be restricted to the class for which $w(F_n, \Delta(E))$ is a Borel-measurable function of E for all $F_n \in \Omega$; note that w depends on E only through Δ, not through F_n. The expected value of w for a particular F_n is called the risk function; it depends of course on the decision function Δ and the weight function w as well as on F_n. Denote it by

$$r(\Delta, w \mid F_n) = \int_W w(F_n, \Delta(E))\, dF_n(E).$$

Since the true F_n is unknown, so in general will be the true value of the risk function associated with a particular decision function Δ. We might call

$$r(\Delta, w) = \sup_{F_n \in \Omega} r(\Delta, w \mid F_n)$$

the maximum risk associated with the decision function Δ. Wald defines Δ^* to be the "best" decision function relative to the weight function w if the maximum risk $r(\Delta, w)$ is minimum for $\Delta = \Delta^*$. He points out that the "best" decision

function might be defined as one which minimizes some weighted mean, taken over all $F_n \in \Omega$, of the risk function $r(\Delta, \mathfrak{w} \mid F_n)$, but that the above definition of the "best" decision function has certain advantages. Thus under certain restrictions on Ω and $\mathfrak{w}$, the risk function $r(\Delta^*, \mathfrak{w} \mid F_n)$ is independent of $F_n \in \Omega$, that is, we then know the exact value of the risk, regardless of what the true F_n may be. This is analogous to the desirable situations where confidence intervals are known, and the probability of a false statement (to the effect that the unknown quantity is in a given region when it is not) is then a constant independent of the unknown quantity.

Wald's theory is suggestive and formally very satisfying, but one would like to see some specific examples of its application to non-parametric cases. A discouraging aspect, not shared by the older Neyman-Pearson theory, lies in the very refinement that a decision function is declared best with respect to a very particular weight function $\mathfrak{w}$. An attractive possibility would be to impose a metric on Ω or on a related function space, and to let $\mathfrak{w}$ be the distance function. In the problem of two samples for example, after metrizing Ω_ν, the weight $\mathfrak{w}$ assigned to accepting H might be taken as the distance between F and G in the notation of section 4. A suitable choice of metric might yield a weight function appropriate to a large variety of situations. The difficulties of finding a distance function which is intuitively satisfactory and analytically tractable in calculating the risk function are no doubt formidable. The device of metrizing a space of distribution functions was used by Mann and Wald in a different connection [17], but their choice of distance function, while appropriate to their problem, would not be satisfactory here.

Also still lacking is any general theory relating the three concepts discussed in Part III. The following questions have been answered, at least for some specific examples, in the parametric case, but are still untouched in the non-parametric case: Are likelihood ratio tests consistent? Is there a simple weight function $\mathfrak{w}$ relative to which the likelihood ratio test becomes a "best" test, or asymptotically a "best" test? If a test is "best" relative to a given weight function, with respect to what set of alternatives is it consistent?

In conclusion let us emphasize the need for *constructive* methods of obtaining "good" and "best" tests and estimates in the non-parametric case. Recalling the history of the parametric case we may judge that half the battle was the definition of "good" and "best" statistical inference. Progress in the non-parametric case has been made in the direction of definition, mainly by carrying over or modifying criteria originally advanced for the parametric case. However, besides criteria for "good" and "best" tests and estimates, we have in the parametric case a large body of constructive theory which may be applied in particular examples to yield the optimum tests or estimates; thus we have the Fisher theory of maximum likelihood statistics for point estimation, and the constructive theorems of the Neyman-Pearson theory for the existence of critical regions of types A, A_1, B, B_1, and the related types of "best" confidence inter-

vals. The contrasting lack of any general constructive methods[16] at present challenges us in the non-parametric theory.

BIBLIOGRAPHY OF STATISTICAL INFERENCE IN THE NON-PARAMETRIC CASE

[1] W. E. CAMPBELL, "Use of statistical control in corrosion and contact resistance studies", *Bell Telephone System Technical Publications*, Monograph B-1350 (1942).
[2] W. J. DIXON, "A criterion for testing the hypothesis that two samples are from the same population", *Annals of Math. Stat.*, Vol. 11 (1940), pp. 199–204.
[3] R. A. FISHER, *Statistical Methods for Research Workers*, section 24, example 19, Oliver and Boyd, Edinburgh, 1925.
[4] R. A. FISHER, "On the random sequence", *Quart. J. Roy. Meteor. Soc.*, Vol. 52 (1926), p. 250.
[5] R. A. FISHER, *The Design of Experiments*, section 21, Oliver and Boyd, Edinburgh, 1935.
[6] R. A. FISHER, "Coefficient of racial likeness and the future of craniometry", *J. R. Anthrop. Inst.*, Vol. 66 (1936), pp. 57–63.
[7] M. FRIEDMAN, "The use of ranks to avoid the assumption of normality", *J. Amer. Stat. Ass.*, Vol. 32 (1937), pp. 675–701.
[8] M. FRIEDMAN, "A comparison of alternative tests of significance for the problem of m rankings", *Annals of Math. Stat.*, Vol. 11 (1940), pp. 86–92.
[9] H. HOTELLING AND M. R. PABST, "Rank correlation and tests of significance involving no assumptions of normality", *Annals of Math. Stat.*, Vol. 7 (1936), pp. 29–43.
[10] M. G. KENDALL, "A new measure of rank correlation", *Biometrika*, Vol. 30 (1938), pp. 81–93.
[11] M. G. KENDALL, "Note on the estimation of a ranking", *J. Roy. Stat. Soc.*, Vol. 105 (1942), pp. 119–121.
[12] M. G. KENDALL, S. F. H. KENDALL, AND B. B. SMITH, "The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times", *Biometrika*, Vol. 30 (1939), pp. 251–273.
[13] M. G. KENDALL AND B. B. SMITH, "The problem of m rankings", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 275–287.
[14] W. O. KERMACK AND A. G. MCKENDRICK, "Tests for randomness in a series of observations", *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 228–240.
[15] A. KOLMOGOROFF, "Sulla determinatione empirica di una legge di distributione", *Giornale Ist. Ital. Attuari*, Vol. 4 (1933), pp. 83–91.
[16] A. KOLMOGOROFF, "Confidence limits for an unknown distribution function", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 461–463.
[17] H. B. MANN AND A. WALD, "On the choice of the number of class intervals in the application of the chi square test", *Annals of Math. Stat.*, Vol. 13 (1942), pp. 306–317.
[18] H. C. MATHISEN, "A method of testing the hypothesis that two samples are from the same population", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 188–194.
[19] F. MOSTELLER, "Note on application of runs to quality control charts", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 228–232.
[20] K. R. NAIR, "The median in tests by randomization", *Sankhyā*, Vol. 4 (1940), pp. 543–550.
[21] K. R. NAIR, "Table of confidence interval for the median in samples from any continuous population", *Sankhyā*, Vol. 4 (1940), pp. 551–558. Second column heading on p. 556 should be changed from n to k.

[16] Wolfowitz' modified likelihood ratio method is a general constructive method of getting tests which we hope to be "good", but until its optimum properties are investigated, his method does not constitute an exception to the thesis of this paragraph. We remark however that the popularity of the usual parametric likelihood ratio method did not await Wald's recent proof of its optimum qualities (to appear in *Trans. of the Amer. Math. Soc.*).

[22] J. NEYMAN, "Basic ideas and some recent results of the theory of testing statistical hypotheses", sections 20, 21, *J. Roy. Stat. Soc.*, Vol. 105 (1942), pp. 292–327.

[23] E. G. OLDS, "Distribution of sums of squares of rank differences for small numbers of individuals", *Annals of Math. Stat.*, Vol. 9 (1938), pp. 133–148.

[24] E. S. PEARSON, "Some aspects of the problem of randomization", *Biometrika*, Vol. 29 (1937), pp. 53–64 and Vol. 30 (1938), pp. 159–179.

[25] K. PEARSON, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Phil. Mag.*, 5 ser., Vol. 50 (1900), pp. 157–175.

[26] K. PEARSON, "On the probability that two independent distributions of frequency are really samples from the same population", *Biometrika*, Vol. 8 (1911), pp. 250–254.

[27] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations", *Suppl. J. Roy. Stat. Soc.*, Vol. 4 (1937), pp. 117–130.

[28] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations. II. The correlation coefficient test." *Suppl. J. Roy. Stat. Soc.*, Vol. 4 (1937), pp. 225–232.

[29] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations. III. The analysis of variance test." *Biometrika*, Vol. 29 (1938), pp. 322–335.

[30] S. R. SAVUR, "The use of the median in tests of significance", *Proc. Ind. Acad. Sc.* (A), Vol. 5 (1937), pp. 564–576.

[31] H. SCHEFFÉ, "On a measure problem arising in the theory of non-parametric tests", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 227–233.

[32] W. A. SHEWHART, "Contribution of statistics to the science of engineering", Univ. of Pa. Bicentennial Conference. Volume on *Fluid Mechanics and Statistical Methods in Engineering*, pp. 97–124. Also *Bell Telephone System Technical Publications*, Monograph B-1319.

[33] N. SMIRNOFF, "On the estimation of the discrepency between empirical curves of distribution for two independent samples", *Bull. Math. Univ. Moscou, Série Int.*, Vol. 2, fasc. 2 (1939).

[34] F. S. SWED AND C. EISENHART, "Tables for testing randomness of grouping in a sequence of alternatives", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 66–87.

[35] W. R. THOMPSON, "On confidence ranges for the median and other expectation distributions for populations of unknown distribution form", *Annals of Math. Stat.*, Vol. 7 (1936), pp. 122–128.

[36] W. R. THOMPSON, "Biological applications of normal range and associated significance tests in ignorance of original distribution forms", *Annals of Math. Stat.*, Vol. 9 (1938), pp. 281–287.

[37] A. WALD, *On the principles of statistical inference*, Notre Dame, Indiana, 1942.

[38] A. WALD, "An extension of Wilks' method for setting tolerance limits", *Annals of Math. Stat.*, Vol. 14 (1943) pp. 45–55.

[39] A. WALD AND J. WOLFOWITZ, "Confidence limits for continuous distribution functions", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 105–118.

[40] A. WALD AND J. WOLFOWITZ, "On a test whether two samples are from the same population", *Annals of Math. Stat.*, Vol. 11 (1940), pp. 147–162.

[41] A. WALD AND J. WOLFOWITZ, "Note on confidence limits for continuous distribution functions", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 118–119.

[42] A. WALD AND J. WOLFOWITZ, "An exact test for randomness in the non-parametric case, based on serial correlation", *Annals of Math. Stat.*, this issue.

[43] W. A. WALLIS, "The correlation ratio for ranked data", *J. Amer. Stat. Ass.*, Vol. 34 (1939), pp. 533–538.

[44] W. A. WALLIS AND G. H. MOORE, *A significance test for time series*, Technical paper 1 (1941), National Bureau of Economic Research, New York.
[45] B. L. WELCH, "On the z-test in randomized blocks and Latin squares", *Biometrika*, Vol. 29 (1937), pp. 21–52.
[46] B. L. WELCH, "On tests of homogeneity", *Biometrika*, Vol. 30 (1938), pp. 149–158.
[47] S. S. WILKS, "Determination of sample sizes for setting tolerance limits", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 91–96.
[48] S. S. WILKS, "Statistical prediction with special reference to the problem of tolerance limits", *Annals of Math. Stat.*, Vol. 13 (1942), pp. 400–409.
[49] S. S. WILKS, *Mathematical Statistics*, Princeton University Press, 1943.
[50] J. WOLFOWITZ, "Additive partition functions and a class of statistical hypotheses", *Annals of Math. Stat.*, Vol. 13 (1942), pp. 247–279.
[51] J. WOLFOWITZ, "On the theory of runs with some applications to quality control", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 280–288.
[52] L. C. YOUNG, "On randomness in ordered sequences", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 293–300.

USEFUL DISTRIBUTION THEORY NOT INCLUDED ABOVE

[53] E. J. GUMBEL, "Les valeurs extrêmes des distributions statistiques", *Annales de l'Inst. H. Poincaré*, Vol. 5 (1935), pp. 89–114.
[54] W. O. KERMACK AND A. G. MCKENDRICK, "Some distributions associated with a randomly arranged set of numbers", *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp. 332–376.
[55] A. M. MOOD, "The distribution theory of runs", *Annals of Math. Stat.*, Vol. 11 (1940), pp. 367–392.
[56] N. SMIRNOFF, "Über die Verteilung des allgemeinen Gliedes in der Variationsreihe", *Metron*, Vol. 12, No. 2 (1935), pp. 59–81.
[57] N. SMIRNOFF, "Sur la dépendance des membres d'une séries de variations", *Bull. Univ. État Moscou, Sér. Int., Sect. A: Math. et Mécan.*, Vol. 1, fasc. 4, pp. 1–12.
[58] W. L. STEVENS, "Distribution of groups in a sequence of alternatives", *Annals of Eugenics*, Vol. 9 (1939), pp. 10–17.