NON-PARAMETRIC STATISTICAL INFERENCE

J. WOLFOWITZ
COLUMBIA UNIVERSITY

1. Introduction

In most statistical problems treated in the literature a datum of the problem is the information that the various distributions of the chance variables involved belong to given families of distribution functions (d.f.'s) completely specified except for one or more parameters. Non-parametric statistical inference is concerned with problems where the d.f.'s are not specified to such an extent, and where their functional form is unknown. This does not preclude some knowledge of the d.f.'s; for example, we may know that they are continuous, uni-modal, bi-modal, and the like.

It is clear that the more information that is available from which to draw inferences the more decisive can our conclusions be, that is, the confidence regions may have smaller average size, the statistical tests will have greater power, and the like. Hence if the functional forms of the d.f.'s are known or if there is good ground for assuming them, it is a loss not to make use of this information. Where this information is not at hand, statistical inference must properly proceed without it. In the latter event the criticism of some statisticians that non-parametric tests are "inefficient" is not valid, because "efficiency" (in the colloquial sense) implies thorough use of available resources, and it cannot be inefficient not to make use of unavailable information. Statistical efficiency must be appraised in the context of available information and, except where uniformly most powerful procedures exist, with respect to specific alternatives.

In the present paper we shall describe briefly a few recent advances in non-parametric theory. Readers who expect a complete, unified theory such as may be found in the analysis of variance will be disappointed; it is impossible to present such a theory because none exists. What has been accomplished thus far is only a series of small advances in various directions. It is as if, faced by a hardy opponent, one lashed out in all directions and succeeded in penetrating the enemy's armor slightly in several places. The analogy is correct to the further extent that many problems in non-parametric inference are of considerable difficulty. It is the hope of the author of the present paper to arouse the interest of the readers of these proceedings, to acquaint them with a few of the developments, and to enlist their aid in solving the multitude of outstanding problems.

2. Estimation

Let $x_1, \cdots, x_n$ be n independent observations on the chance variable X, about whose cumulative d.f., $f(x)$, nothing is known except that it is continuous. We shall discuss the problem of estimating $f(x)$.

It is intuitively obvious that the "best" estimate of $f(x)$ is the sample (cumulative) d.f., $\phi(x)$, a step-function defined as follows: $n\phi(x)$ is the number of observations among $x_1, \cdots, x_n$ which are less than x. Such an estimate, however, suffers from the same disadvantages as a point estimate of a parameter. It does not associate with the estimate an appraisal of its accuracy, so to speak; a distinction must be made, for example, between estimates obtained from ten or from a thousand observations. Since the estimating step-function cannot be the true d.f., the value of the estimate lies in this: it implies, with a certain confidence coefficient α (before the experiment, with a probability α), that some "neighborhood" of $\phi(x)$ contains $f(x)$. The purpose of a theory of estimation is to clarify and make precise the meaning of "neighborhood."

This same problem occurs in the estimation of a parameter, and the theory of estimation was put on a rigorous foundation by Neyman [9][1] in a manner as ingenious as it is simple. For each value of the parameter θ which completely characterizes the d.f. to be estimated (the limitation to one parameter is made here for convenience only, not of necessity), an "acceptance" region $A(\theta)$ in the sample space is assigned so that the probability of this region is α when θ is the parameter value. Then the "neighborhood" (confidence region) is simply the totality of all θ for which the observed sample point lies in the region of acceptance. Further restrictions on the regions $A(\theta)$ are needed in order that the confidence regions have special desirable forms (e.g., that they be intervals). These restrictions are far from sufficient to determine uniquely the regions $A(\theta)$. Although many valuable results have been obtained bearing on the problem of choice of $A(\theta)$, many outstanding problems remain.

Let us now return to the problem of estimation of $f(x)$. Our object will be to construct regions $A(f)$ such that, when $f(x)$ is the d.f. of X, the probability is α that $\phi(x)$ will lie entirely in $A(f)$. Then the confidence region (to which we earlier gave the intuitive designation "neighborhood") is the totality of all f for which the observed $\phi(x)$ lies in $A(f)$. Strictly speaking, this totality is a region only in the function space. However, no confusion will result from designating the set of functions as a region, just as no confusion results from calling a set of points which estimate a parameter a "region," which it need not always be in the precise mathematical definition of region. The problems which therefore confront us are the following:

1. How to construct $A(f)$ for a given continuous f, such that the probability that $\phi(x) \subset A(f)$ is α;

2. How to choose such $A(f)$ that the confidence region $R(\phi)$ shall be constructible and shall be practically constructible [if the totality of all f for which $\phi(x) \subset A(f)$ can only be envisaged conceptually without being practically accessible, its value in practice is limited];

3. To determine further conditions on the $A(f)$ which will ensure such desirable properties as minimum average size in some suitable sense, and/or unbiassedness, and the like.

We shall now discuss a method which yields an answer (though of course not the only possible one) to the first two problems.

[1] Boldface numbers in brackets refer to references at the end of the paper (see p. 112).

Let $f(x)$ be any continuous cumulative d.f. Let $\delta_1(x)$ and $\delta_2(x)$ be non-negative continuous functions defined in the closed interval [0,1]. Define the functions $l_1(x)$ and $l_2(x)$ as follows:

$$l_1(x) \equiv f(x) + \delta_1[f(x)],$$
$$l_2(x) \equiv f(x) - \delta_2[f(x)].$$

It will be noticed that the graphs of $l_1(x)$ and $l_2(x)$ form a "belt" which encloses $f(x)$. Subject to certain conditions on $\delta_1(x)$ and $\delta_2(x)$ which we shall discuss in a moment, we wish to determine the probability that $\phi(x)$ will lie in this belt, that is, that

$$l_2(x) \leqq \phi(x) \leqq l_1(x), \text{ for all } x.$$

Our intention is, after this probability has been determined, so to manipulate $\delta_1(x)$ and $\delta_2(x)$ that this probability is the prescribed α. We intend then to let the region within this belt [more properly, the totality of all $\phi(x)$ which are such that $l_2(x) \leqq \phi(x) \leqq l_1(x)$] constitute the acceptance region $A(f)$. The confidence region $R(\phi)$ will then consist of all f for which ϕ lies in the belt $A(f)$. We shall see that this confidence region will also be determined by a belt enclosing $\phi(x)$, that is, two step-functions $\phi_1(x)$ and $\phi_2(x)$, such that

$$\phi_2(x) \leqq \phi(x) \leqq \phi_1(x),$$

and $R(\phi)$ is the totality of all continuous (cumulative) d.f.'s which are such that at every point x they lie between $\phi_2(x)$ and $\phi_1(x)$.

We have therefore to be able to determine the probability that $\phi(x)$ will lie within the belt formed by $l_1(x)$ and $l_2(x)$. One prospect which may cause some dismay is that this probability will depend on $f(x)$. If this were so, then the adjustment of $\delta_1(x)$ and $\delta_2(x)$ so that the probability of $A(f)$ shall be α would depend upon f and would give rise to serious complications. This, however, is not so, because of the significant fact that, in the definition of $l_1(x)$ and $l_2(x)$, δ_1 and δ_2 are functions of $f(x)$ and not of x. Consider any topologic (one-to-one and continuous) transformation of the real line into itself which takes a chance variable X into a chance variable Y. If the point y goes over into the point y', then the probability that $X < y$ is the same as the probability that $Y < y'$, and equals $f(y)$. Hence the functions $f(x)$, $l_1(x)$, and $l_2(x)$ are unchanged by this transformation except for the fact that each point x may receive a different "name" x'. The probability that $\phi(x)$ will fall within the belt formed by $l_1(x)$ and $l_2(x)$ depends only on the positions of $l_1(x)$ and $l_2(x)$ relative to $f(x)$, and these are unchanged. Hence the probability that $\phi(x) \subset A(f)$ depends only upon δ_1 and δ_2 and not at all upon $f(x)$. This is a considerable simplification.

Are we perfectly free to choose $\delta_1(x)$ and $\delta_2(x)$ (provided only that they be non-negative and continuous)? No. We intend to be able to "manipulate" $\delta_1(x)$ and $\delta_2(x)$ so that the probability that $\phi(x) \subset A(f)$ be any prescribed positive $\alpha < 1$. This means precisely that what we really want is a one-parameter family of couples of functions $\delta_1(x)$ and $\delta_2(x)$, each couple corresponding to a

value of α. In most problems of practical importance we shall be able to achieve this by enlarging or decreasing $\delta_1(x)$ and $\delta_2(x)$; for example, we may choose $\delta_1(x)$ and $\delta_2(x)$ both constant, or both a constant multiplied by x. However, certain restrictions on $\delta_1(x)$ and $\delta_2(x)$ are general. Thus, since $\phi(x)$ is monotonically non-decreasing, we can assume that $l_1(x)$ and $l_2(x)$ are also monotonically non-decreasing. If, for example, $s = l_1(z_2) < l_1(z_1)$ when $z_2 > z_1$, we may, without at all changing the probability that $\phi(x) \subset A(f)$, let $l_1(z_1) = s$. This is so because, from

$$\phi(z_1) \leqq \phi(z_2)$$
$$\phi(z_2) \leqq l_1(z_2)\,,$$

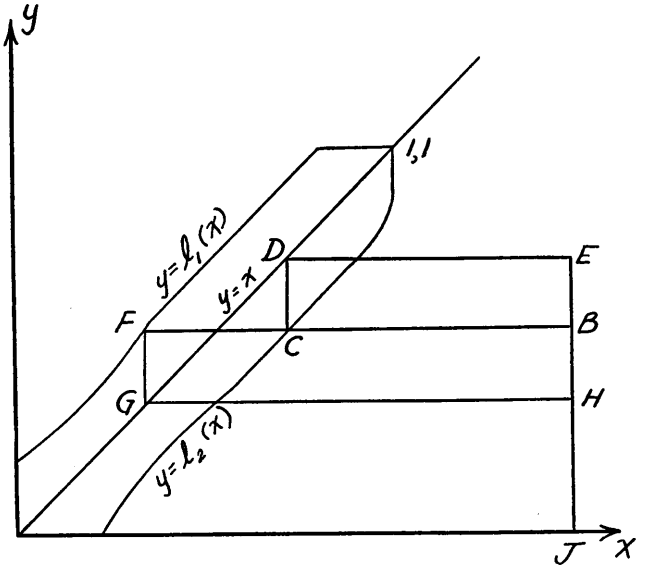
we obtain

$$\phi(z_1) \leqq l_1(z_2)\,.$$

Moreover, $l_1(x)$ must take the value one for some finite x for which $f(x) < 1$, since $\phi(x)$ does (with probability one). If $l_1(x)$ did not do this, then the probability that $\phi(x)$ should lie in the belt formed by $l_1(x)$ and $l_2(x)$ would be zero. For a similar reason, $l_2(x)$ must take the value zero at some x for which $f(x) > 0$.

Suppose now that $\delta_1(x)$ and $\delta_2(x)$ have been chosen so that the probability that $\phi(x) \subset A(f)$ is the prescribed α. How shall we, given $\phi(x)$, obtain $R(\phi)$?

Let us assume, for simplicity of exposition, that $l_1(x)$ and $l_2(x)$ are strictly monotonically increasing. By our definitions, l_1 and l_2 are really functions of $f(x)$, and $f(x)$ takes continuously all values from zero to one. If, for $x = t$, with t any arbitrarily chosen number, $\phi(t) = 1$, then the trivial upper bound on $f(t)$ is one. Suppose then that $\phi(t) < 1$. In the x,y-plane draw the graphs of $y = x$, $y = x + \delta_1(x)$, and $y = x - \delta_2(x)$ $(0 \leqq x \leqq 1)$. We have then the region $A(f)$ corresponding to the simple distribution $f(x) \equiv x$ $(0 \leqq x \leqq 1)$, $f(x) \equiv 0$ $(x < 0)$, $f(x) \equiv 1$ $(x > 1)$. Let B be the point $t,\phi(t)$. Let BC be the



intersection of the line $y = \phi(t)$ with $y = l_2(x)$ and let D be the intersection with $y = x$ of the vertical line through C. Finally, complete the rectangle $BCDE$. Then no d.f. which is such that its acceptance region A contains entirely the given $\phi(x)$ can have, at t, an ordinate higher than E. For if it did, then the point $[t_2,l_2(t)]$ would be higher than C. Hence $\phi(x)$ would be below $l_2(x)$ at $x = t$, in violation of the fact that it lies in the acceptance region of the d.f.

Similarly, let F be the intersection of $y = \phi(t)$ with $y = l_1(x)$, and G the intersection with $y = x$ of the vertical line through F. Complete the rectangle $BFGH$. Then by an argument similar to that above we can prove that a d.f. whose ordinate at $x = t$ is lower than H would have an acceptance region whose upper bound at $x = t$ would fall below B and hence would not include the observed $\phi(x)$.

It is obvious that, if either the intersection at C or that at F does not exist, E is at a distance one from the x-axis or H is on the x-axis, respectively.

Finally, it is clear that any d.f. whose ordinate at t lies between E and H is such that the observed $\phi(x)$ fulfills the condition $l_2(t) \leqq \phi(t) \leqq l_1(t)$.

We may therefore assert, with confidence coefficient α, that the actual d.f. of X has, at $x = t$, an ordinate which lies between E and H. What is more, if this procedure of obtaining the points E and H is repeated everywhere (for all x), we can assert, with confidence coefficient α, that at *every* x the ordinate of $f(x)$ lies between the corresponding E and H. Every continuous d.f. which lies entirely within these bounds is such that the observed $\phi(x)$ lies in its acceptance region.

The lengths of EJ and HJ are functions of BJ only. Since $\phi(x)$ is a step-function whose only values are multiples of 1/n, the procedure for obtaining E and H need be repeated at most $n + 1$ times. Denote the totality of points E by $\phi_1(x)$ and the totality of points H by $\phi_2(x)$. Then $\phi_1(x)$ and $\phi_2(x)$ are also step-functions which determine the desired region $R(\phi)$. The latter is thus practically constructible.

For the sake of simplicity of exposition we assumed that $l_1(x)$ and $l_2(x)$ were strictly monotonic and proceeded in a somewhat intuitive geometric fashion. The reader interested in a rigorous description is referred to Wald and Wolfowitz [23].

It is unnecessary to remark that it cannot be asserted that the unknown $f(x)$ is such that its graph lies in $R(\phi)$. The correct statement is that the foregoing procedure is such that, before the sample is obtained, the probability is α that $R(\phi)$ will "cover" $f(x)$.

In some problems only upper (or lower) confidence limits are of interest. In that event we let $\delta_1(x)$ [or $\delta_2(x)$] be identically zero. The result is a statement, valid with confidence coefficient α, that $f(x)$ is at most $\phi_1(x)$ [at least $\phi_2(x)$] for *all* x.

It is interesting to note that this method gives no information about the range of the chance variable X. On reflection this is perfectly reasonable. From the knowledge that $f(x)$ is continuous and the knowledge of the smallest and largest observations on X, we cannot conclude, for example, that no observa-

tion can be smaller than some lower bound. It is only on the basis of additional information, such as that of the functional form of $f(x)$, that such estimates could be made. Formally this can be seen as follows: We have already remarked earlier that $l_2(x)$ must be zero for all x less than some number t^* such that $f(t^*)$ is positive. If this were not so, then the probability that $\phi(x) \subset A(f)$ would be zero. However, when this is so, we have that $\phi_1(x) = f(t^*)$ for all x less than the smallest observed value. A similar argument applies to the large values of x.

How shall $\delta_1(x)$ and $\delta_2(x)$ be chosen? This is an important and probably difficult problem on which no results have as yet been published. It is likely that in studying the probability that the confidence region shall "cover" some specific d.f. other than $f(x)$ a fruitful procedure would be to introduce a metric in the space of continuous d.f.'s. The distance between two d.f.'s would then serve as a measure of the importance of the difference between them. If $f(x)$ is the d.f. of X, the probability that the confidence region shall include another d.f. is not likely, except under special conditions, to be the same for all d.f.'s at the same distance from $f(x)$, and attention will probably be directed to the greatest lower bound of these probabilities. If the metric is not to be arbitrary but is to serve as a yardstick of the importance of the deviations of the d.f.'s, it must follow from the practical circumstances of an actual problem in application, and will in general be different for different problems. This may prove a serious drawback.

An expeditious method is to choose $\delta_1(x)$ and $\delta_2(x)$ both constant. Suppose that this is done and both constants have the same value Δ. Then $\phi_1(x) = \phi(x) + \Delta$, $\phi_2(x) = \phi(x) - \Delta$, except where obvious changes are necessary to keep the boundaries from going above one or below zero. Hence if a table of Δ as a function of α were available, the construction of confidence regions would be almost immediate. Such a table for large sample sizes n was constructed by Kolmogoroff [5] and enlarged by Smirnoff [16]. Kolmogoroff proved that, when $\delta_1(x) \equiv \delta_2(x) \equiv \lambda/\sqrt{n}$, the probability that $\phi(x) \subset A(f)$ approaches

$$\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$$

as $n \to \infty$, uniformly in λ. Smirnoff proved a more general result which we shall mention later. The series in λ converges very rapidly. For $\alpha = 0.95$, λ is approximately 1.35, that is, Δ is very close to $1.35/\sqrt{n}$ for large samples.

An example of the construction of confidence regions using a small sample size ($n = 6$) and exact probabilities will be found in Wald and Wolfowitz [23]. The reader will also find there a method for finding the probability that $\phi(x)$ will be in $A(f)$, applicable to general $\delta_1(x)$ and $\delta_2(x)$ and finite sample sizes.

Finally, one remark about "jumps," which we shall discuss only for the case where $\delta_1(x)$ and $\delta_2(x)$ are both constant. Conceivably it could happen at some value of x, say ξ, that the lower bound of the confidence belt coming in from the right [i.e., $\phi_2(\xi+)$] is greater than the upper bound coming in from the left [i.e., $\phi_1(\xi-)$]. No continuous d.f. could then lie entirely in $R(\phi)$. However,

if $f(x)$ is continuous, then, except for an event of probability zero, no two observations will coincide, and hence $\phi(x)$ will not sustain a jump of more than $1/n$ at any point. Now $\delta_1(x) + \delta_2(x)$ is a constant not less than $1/n$, else no $\phi(x)$ could lie in $A(f)$, $\phi(x)$ being a step-function with all its jumps a multiple of $1/n$. Hence the probability of the situation where no continuous $f(x)$ could lie in $R(\phi)$ is zero. Of course, the actual observations can never really come from a strictly continuous distribution because of limitations of measurement. The difficulties which may sometimes arise can probably be obviated by some simple assumption about the distribution of values within intervals of a length equal to the unit of measurement.

We remarked earlier that, if the a priori information about $f(x)$ consists solely in the knowledge that it is continuous, the information about the "tails" of the distribution is of necessity limited. Hence we cannot estimate moments, since the latter depend critically upon the tails. Moreover, we have no reason for assuming that any of the moments exist. However, we can readily construct confidence intervals for the median or any other percentile of the distribution.

In order to see this, perform the transformation $Y = f(X)$. This transformation is continuous and one-to-one, except that entire intervals in X (one end of which may be at infinity) of zero probability are transformed into single points, a fact which causes no difficulty. Now the probability that the interval between the kth and k'th largest observations on Y (in a sample of n) shall contain the point $1/2$ is readily found, since Y is uniformly distributed between zero and one. Let us choose k and k' so that this probability is suitably close to the desired confidence coefficient; since k and k' may take only a finite number $\left[\binom{n}{2}\right]$ of values, some compromise may be necessary. Then it is clear that the interval between the kth and k'th largest observations on X is a confidence interval for the median. Other percentiles may be similarly treated. This method is due to W. R. Thompson [**19**].

Just as the probability that the interval between the kth and k'th largest observations will contain the median is independent of $f(x)$ when the latter is continuous, so, in exactly the same way, we see that the probability distribution of the chance variable defined as the probability assigned by $f(x)$ to the interval running from the kth largest x to the k'th largest x (a chance variable because the ends of the interval are chance variables) is independent of $f(x)$, depends only on k and k' (and n, of course), and may readily be found from the uniform distribution. In this way Wilks [**26**] was able to give non-parametric tolerance intervals, that is, pairs (k,k'), such that the probability is a prescribed α that the probability assigned by $f(x)$ to the interval will be at least a prescribed β. (The α is subject to the compromise mentioned above.)

Wilks [**27**] also solved the problem of tolerance intervals for two independent variables jointly, and Wald [**21**] solved it for two variables jointly, without the hypothesis of independence. Wilks [**26**] also solved the problem of non-parametric tolerance intervals which would include at least a specified pro-

portion of a second sample. In this work the only necessary assumption on $f(x)$ is that of continuity.

3. Testing hypotheses

The division of the subject of the present lectures into estimation and testing of hypotheses is based more on expediency than on logical need. Naturally the relation between estimation and testing of hypotheses is a very intimate one; indeed, Neyman's idea of a confidence region is to take the totality of all parameter values not rejected by a set of tests of hypotheses. In spite of this, we shall find it convenient to retain this division.

The procedure of estimation described in the preceding section can of course be employed to test hypotheses about the population d.f. Suppose that it is desired to test the hypothesis that the cumulative d.f., $f(x)$, of a chance variable X on which n random independent observations have been made is a given d.f., $F(x)$. If it is not required that $f(x)$ be determined by a finite number of parameters, the problem is a non parametric one. Let us assume that the $\delta_1(x)$ and $\delta_2(x)$ of the previous section have been suitably chosen. Then a reasonable test would consist in seeing whether $\phi(x)$ is entirely contained in $A(F)$, the notation of the previous section being retained. If $\delta_1(x)$ and $\delta_2(x)$ are constant and n is large, Kolmogoroff's result is applicable, and the test would consist of seeing whether

$$\sup |F(x) - \phi(x)| \sqrt{n}$$

exceeds the Kolmogoroff λ which corresponds to the chosen confidence coefficient.

A closely allied problem is that of "two samples," that is, to test whether two samples, of size n_1 and n_2, respectively, came from the same population. The test that naturally suggests itself in connection with the previous discussion on estimation is that based on

$$\sup |\phi(x) - \phi'(x)| ,$$

where $\phi(x)$ and $\phi'(x)$ are the two sample d.f.'s. When both samples are of large size the result of Smirnoff [16] is available; it states that the probability that

$$\sup |\phi(x) - \phi'(x)| \sqrt{n} \leqq \lambda , \qquad n = \frac{n_1 n_2}{n_1 + n_2} ,$$

approaches the same limiting function as that in Kolmogoroff's result cited above, namely,

$$\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$$

when n_1 and n_2 approach infinity at a constant ratio, no matter what the common continuous d.f. of the populations from which both samples were drawn.

Another attack on the problem of two samples was made by Wald and Wolfowitz [24]. Let the observations in both samples be arranged in order of size, and replace each observation by 1 or 2 according as it came from the first or second sample. The result is a sequence V of 1's and 2's which abstracts from the sample only the order relationships (in size) among the various observations. Under the null hypothesis (that both samples came from the same population) all V's have the same probability. When the two populations are different, some small intervals have greater probability under one distribution than under the other. This results in a diminution of the average total number of "runs." A run, for this purpose, is defined as a subsequence of consecutive 1's or 2's which cannot be extended in either direction. Hence the statistic proposed by these writers is the total number of runs, small values being critical. The distribution of this statistic is known (see Wald and Wolfowitz [24]). When n_1 and n_2 are not large, the critical points are not difficult to obtain, and an excellent table by Swed and Eisenhart [18] is available. As n_1 and n_2 approach infinity at a constant ratio, the distribution of the total number of runs approaches normality, so that no difficulty arises.

Wald and Wolfowitz [24], employing the Neyman-Pearson idea of the power function, extended the notion of consistency, due to R. A. Fisher. A precise description can be found in their paper; let us here content ourselves with a more suggestive and less rigorous description. A consistent estimate of a parameter is one such that, as the sample size increases indefinitely, the probability approaches one that the estimate will lie in an arbitrarily small neighborhood of the "true" value of the parameter being estimated. A test is consistent if it is based on the (proper) use of a consistent estimate. In this event the probability of rejecting an alternative hypothesis when it is false approaches one as the sample number increases indefinitely. A non-parametric test is called consistent if it possesses this latter property. It is easy to see that this definition includes the parametric one as a special case. It is important to note, however, that a test may be consistent with respect to one alternative, but not with respect to another. This is particularly important in non-parametric theory, but is true in the classical parametric theory as well. Thus, a test of the hypothesis that the mean of a normal distribution with unit variance is a given number, the test being based on one "tail" of the normal distribution, is consistent with respect to one set of alternatives and not consistent with respect to another set of alternatives.

The statistic U, the total number of runs, is shown by Wald and Wolfowitz [24] to provide a consistent test for the problem of two samples, with respect to all alternatives which satisfy some mild restrictions. The method of proof is essentially this: Let n_1 and n_2 increase at a constant ratio. It is shown that U/n_1 converges stochastically to its expected value when either the null or any alternative hypothesis (subject to the restrictions mentioned) is true. The expected value is explicitly obtained and is shown to be a maximum when both populations are identical. The proof rests essentially on the following lemma (see [24], Theorem II, lemma 2):

Let $\mathrm{f(x)}$ *and* $\mathrm{g(x)}$ *be the cumulative distribution functions of the populations from which the* $\mathrm{n_1}$ *and* $\mathrm{n_2}$ *random independent observations, respectively, are obtained. Suppose that:*

$$\begin{aligned}
&a)\quad f(x) \equiv 0 && (x < 0)\\
&\qquad f(x) \equiv x && (0 \leqq x \leqq 1)\\
&\qquad f(x) \equiv 1 && (x > 1)\\
&b)\quad g(x) \equiv 0 && (x \leqq 0)\\
&\qquad g(x) \equiv 1 && (x \geqq 1)
\end{aligned}$$

c) The derivative $\mathrm{g'(x)}$ *of* $\mathrm{g(x)}$ *exists, is continuous and positive everywhere in the interval* $0 \leqq \mathrm{x} \leqq 1$

d) $\mathrm{n_1/n_2} = \mathrm{c}$, *a constant.*

Then

$$\lim_{n_1 \to \infty} E(U/n_1) = 2\int_0^1 \frac{g'(x)}{c + g'(x)}\,dx$$

$$\lim_{n_1 \to \infty} \sigma^2(U/n_1) = 0\,.$$

The writer has recently improved this result as follows:

$$\lim_{n_1 \to \infty} \sigma^2(U/\sqrt{n_1}) = 4\left[\int_0^1 \frac{cg'^2}{(c+g')^3}\,dx + \int_0^1 \frac{g'(c^3+g'^3)}{(c+g')^4}\,dx - \left(\int_0^1 \frac{g'^2}{(c+g')^2}\,dx\right)^2 - c^3\left(\int_0^1 \frac{g'}{(c+g')^2}\,dx\right)^2\right].$$

Also it can be shown that under these conditions the distribution of U approaches the normal distribution. Using these facts, we can obtain the power of the test based on U for alternatives subject to some slight (from the statistical point of view) restrictions when the sample sizes are large.

A few remarks on tests in the problem of two samples which are not consistent may be of interest at this point. Two proofs of lack of consistency are available in the literature, one by Wald and Wolfowitz [**24**] and the other by Bowker [**1**]. Let us examine the test (proposed by Mathisen [**6**]) which Bowker proved not consistent. One form of the test is as follows: Observe the number m of observations of the second sample whose values are less than the median of the first sample. When m is too large or too small (in a precisely defined way), the hypothesis of identity of d.f.'s is to be rejected. This test is not consistent if among the alternatives to the null hypothesis we admit any large part (of course this is precisely defined in the papers cited) of the totality of all couples of non-identical continuous d.f.'s, but is consistent if the alternatives are, for example, confined to couples of d.f.'s such that one is a translation of the other. The essence of the proof (of non-consistency) is this, that any two continuous d.f.'s, which have a common median and coincide in any small neighborhood of the median, will show, in the limit, identical behavior with respect

to m. Hence the two d.f.'s may differ ever so drastically (except in the neighborhood of their medians), the sample numbers may increase without limit, and yet the probability of rejecting the null hypothesis (which is false) may be made to differ from the size of the test by a number arbitrarily small. (When one d.f. is a translation of the other, the two medians cannot coincide.) Another version of this test uses quartiles instead of medians; this test can be proved not consistent in a similar manner. Indeed, tests which are based on the relative behavior of the d.f.'s at a fixed finite number of points are not consistent unless the alternatives are suitably restricted, whereas a statistic like U, the total number of runs, depends, intuitively speaking, on the relative behavior of the two d.f.'s along the entire real line. The χ^2-test for goodness of fit or for the problem of two samples is also inconsistent if the alternatives are not too restricted and if the location of the class intervals does not vary with n.

The efficiency of a statistical test depends upon its power function and is in general relative to specific alternatives. Relative efficiency of estimation could be defined, borrowing an idea of Neyman's, on the basis of the probability that the confidence region will include some specific d.f. not that of the population; this definition would also be relative to a specific alternative distribution.

The question naturally arises: How much of the available information is utilized by these various tests? We shall not attempt here to give the phrase "available information" a precise meaning. The statistic U made use only of the order relations among the observations, and the purely metric properties were not used. This need not constitute an adverse criticism, because on the basis of the sole a priori information that the two d.f.'s are continuous no "information" may really be lost. Similarly the statistic m, the number of observations in the second sample less than the median of the first, does not utilize many properties of the two samples. The fact that one of these statistics is consistent and the other is not implies that the one which is not consistent does not utilize all the available information.

In order to clarify the situation let $A^* = a_1, a_2, \cdots, a_{n_1+n_2}$ be the $n_1 + n_2$ observations from both populations, the first n_1 being from the first population. Let us assume that no two observations are equal; since the d.f.'s are continuous, this event has probability one. The fundamental idea of the following is due to R. A. Fisher [2], [3]. Under the null hypothesis all permutations of the elements of A^* have the same conditional probability, when the totality of observations is that of A^*. Consequently each of the $\binom{n_1+n_2}{n_1}$ possible divisions of the $n_1 + n_2$ elements of A^* into two samples of size n_1 and n_2, respectively, is equally likely under the null hypothesis. The test procedure then selects l of these divisions to constitute the critical region, where

$$l\left[\binom{n_1+n_2}{n_1}\right]^{-1} = \beta,$$

the level of significance (size of the critical region). Let A^* be considered a point in an $n_1 + n_2$ dimensional Euclidean space. Consider all the "conjugate" points obtained by permuting the coördinates of A^*. Let $\pi_1, \cdots, \pi_t$ be the $t = (n_1 + n_2)!$ permutations of the first $(n_1 + n_2)$ integers, and let R_{π_i} be the region in $(n_1 + n_2)$ space with the property that if the coördinates of any point in it are arranged in ascending order the sequence of their subscripts will be π_i. Then A^* and each of its conjugate points lies in a different R_{π_i}. Scheffé [13] has proved a result which essentially is that, subject to some slight restrictions, the method of randomization is the most general one which will yield (similar) regions of prescribed size for any common admissible d.f.

We are now in a position to see that, in one sense, requiring the statistic to be a function only of the ranks of the observations (of the order relations among them) is unnecessary from the point of view of obtaining similar regions. The general method requires only that a proper number of each set of conjugate points shall lie in the critical region, but these points need not always lie in the same R_{π_i}. The restriction to ranks requires that several entire regions R_{π_i} constitute the critical region. In behalf of methods based on ranks we may give the following argument, besides that of ease of application: If the observations represent scientific data, say, and statisticians agree on what is an optimum test, we should expect two scientists, examining separately the same material, to come to the same conclusion. Since their results should not depend on the accident of choice of a scale of measurement, the critical region should be invariant under at least linear transformations. All that is assumed about the underlying d.f.'s is their continuity, and this property is invariant under topologic transformation of the real axis into itself. Invariance of the critical region under topologic transformation requires that the statistic be a function of the ranks of the observations.

Using Fisher's randomization idea described above, Pitman [12] proceeded as follows: Let c and d denote, respectively, the means of the first and second samples, and let $w = c - d$. Consider the distribution of w over the set of all possible divisions of the $n_1 + n_2$ observations into two groups, containing n_1 and n_2 elements, respectively. All the attainable values of w are equally likely. Large values of $|w|$ are the critical values, taken in sufficient number to make the critical region of the proper size; the latter must be a multiple of $\left[\binom{n_1 + n_2}{n_1}\right]^{-1}$, so some compromise may be necessary.

For sizable n_1 and n_2 the arithmetic of Pitman's test becomes very burdensome. Pitman (*op. cit.*) gives the details of fitting a Pearson curve in order to find an approximate critical point. The parameters of the curve are of course functions of the observations. Pitman's conjecture that the distribution in large samples is approximately normal was rigorously proved, under mild restrictions, by Wald and Wolfowitz [22]. This means that Pitman's test and the test based on the assumption that the two d.f.'s are normal with the same variance are related as follows: (*a*) they are the same in the limit; (*b*) if the hypothesis of normality with equal variances is true, the classical test is, no

doubt, more efficient than Pitman's for small samples; (*c*) Pitman's test always has the correct size, whether normality holds or not.

A remark about the proof of the limiting normality of Pitman's statistic may be of interest. The problem involved is somewhat different from the usual sort of obtaining the limiting distribution, because the universe is that of the totality of all divisions of the $(n_1 + n_2)$ observations into two groups, n_1 and n_2 in number, respectively. A general theorem on the limiting distribution of linear forms in a universe of permutations of the observations was proved by Wald and Wolfowitz [**22**]. The method is to show that the moments approach those of the normal distribution, and some restrictions are imposed. The proof is related to that of the limiting normality of the rank correlation coefficient, due to Hotelling and Pabst [**4**], and indeed includes the latter result as a special case.

Another application of the randomization method to the problem of two samples from multivariate populations was made by Wald and Wolfowitz [**22**]. They employed Hotelling's generalized T in the population of permutations of the observations to test the hypothesis that two bivariate (or, in general, multivariate) d.f.'s are identical, the alternatives being restricted to the case where one d.f. can be obtained from the other by a translation. For reasons of simplicity they made use of a monotonic function of Hotelling's T, but the test is exactly as if the latter were literally employed. Its distribution under the null hypothesis that the d.f.'s are identical is over the population of permutations of the observations. For small samples it can be obtained by enumeration of the permutations. For large samples the distribution is proved by these same authors [**22**] to be approximately the χ^2 distribution, subject to some restrictions on the d.f.

In connection with the above-mentioned theorems on limiting distributions it would be desirable to investigate the rapidity of approach to the limiting distribution, which has not been done. The situation from the practical computing point of view is perhaps worst for samples of medium size, which are too large for enumeration and are too small for the limiting theorems to apply.

In all the foregoing the reader will no doubt have noticed the absence of any general method of estimation or of testing hypotheses which might correspond to general methods like that of maximum likelihood or the likelihood ratio. The question was specifically raised by E. S. Pearson [**11**], who pointed out that no general method was available for choosing those among the conjugate points which are to constitute the critical region. Statistics are usually chosen in analogy to those in use in classical parametric theory. This is true with Pitman's statistic described above, the rank correlation coefficient, the "randomized" serial correlation coefficient of Wald and Wolfowitz [**25**], and others. A small beginning toward the solution of this problem has been made by the writer [**28**].

The choice of critical region should without doubt be based, in the author's opinion, on the behavior of the power function of the test with respect to the important alternatives. However, the optimum behavior of the power

function of the parametric likelihood ratio of Neyman and Pearson has only recently been proved by Wald [20], although there has for a long time been little doubt that this idea is one of the most fruitful in all of theoretical statistics. What we shall now describe is an attempt to extend the likelihood ratio method to statistics based on ranks. It seems that the method, according to partial results obtained by the writer, can be extended to statistics of the general Fisher randomization type and need not be limited to rank statistics, and the writer hopes to publish some results on this in the future. It must be stressed that what is about to be described is only a beginning, and two fundamental problems to be solved will be described later.

The parametric likelihood ratio λ of Neyman and Pearson is defined as follows (see [10]): Its numerator is the maximum of the values of the likelihood assigned to the sample by the d.f.'s which satisfy the null hypothesis, and its denominator is the maximum of the values of the likelihood assigned to the sample by all admissible d.f.'s. Application requires that the distribution of λ shall be the same for all d.f.'s when the (composite) null hypothesis is true. Small values of λ are the critical values.

Let us return to the problem of two samples. Let the $n_1 + n_2$ observations be arranged in ascending order, and let V be the sequence of 1's and 2's obtained by replacing observations from the first sample by 1, and observations from the second sample by 2. Let f_1 and f_2 be the d.f.'s of the populations from which the first and second samples, respectively, were actually drawn. Let Ω, the totality of all admissible couples (F_1,F_2), be the set of all pairs of continuous d.f.'s. The null hypothesis states that (f_1,f_2) lies in ω, the subset of Ω which consists of all couples (F_1,F_2) both members of which are identical. Let us consider the maximum likelihood not of the observations in the sample directly, but of the sequence V derived from the sample, for the couples of d.f.'s in ω and Ω. When the null hypothesis is true, all possible V's have the same probability, which is a function solely of n_1 and n_2, and therefore independent of the sample. The maximum[2] probability assigned to V by the couples of d.f.'s in Ω is a function of V. The result is the following test based on the likelihood ratio: Let $V_1, V_2, \cdots, V_s$ be the sequence of all possible V's, arranged in descending order according to the maximum probability which each V can take under the couples of d.f.'s in Ω. The critical region is composed of all the V's in an initial subsequence $V_1, V_2, \cdots, V_l$, where l/s is the size of the critical region.

The procedure described above is applicable to other non-parametric problems as well; for more details the reader is referred to Wolfowitz [28]. Let us point out two of the gaps which must be filled in if this procedure is to be put to use. First, although the procedure is plausible it must be justified precisely; this is likely to be a very difficult task, judging from the difficulties encountered in the parametric theory and the larger classes of alternatives which must here be considered. Second, a constructive method of obtaining the maximum likelihood of V must be developed.

[2] Throughout we use the term "maximum" where "supremum" would be more appropriate. No confusion is caused thereby.

4. Conclusion

In the present paper we have tried to give a survey of a number of problems in non-parametric inference, preferring more or less intuitive and heuristic descriptions to a precise but perhaps less intuitively clear presentation. More detailed statements can be found in the various references, and an exhaustive account and bibliography is available in the excellent paper by Scheffé [14]. We have made no attempt at covering the entire field, omitting in particular considerable run theory, for which the reader is referred to Mood [7], Mosteller [8], Shewhart [15], Stevens [17], and Wolfowitz [29], among others. It seems appropriate to conclude by outlining a few major directions in which research in non-parametric inference may be expected to proceed.

The power functions of the more generally used tests should be obtained so that we can judge their efficiency against various alternatives. Very often the large-sample theory will be the one to be developed for reasons of expediency.

The subject of estimation should be cleared up, particularly in the details described in the second section.

General and constructive methods of obtaining critical regions should be developed in connection with the randomization method of Fisher. The validity of the extension of the likelihood ratio method should be investigated, and, if confirmed, the procedure should be generalized to other than rank statistics, and to a general class of problems. Constructive methods of obtaining the maximum likelihood will be needed.

Finally, it is necessary to develop a theory for the numerous and important situations where our a priori information tells us more than that the distributions are continuous or discrete, say, but falls short of telling us the functional forms of the distributions. Thus we may know that the d.f.'s involved are uni-modal, symmetric, or that their derivatives are bounded by a given constant. It is very likely that such a theory would be of greater practical importance than the one now existing. It is also likely to prove very difficult, which is perhaps why scarcely any results are available.

5. Appendix

Derivation of the formula for the large-sample variance of U *under an alternative hypothesis.*—In discussing the problem of two samples, we described the statistic U, the total number of runs of the two kinds of elements, and stated a new result, the formula for the large-sample variance of U under an alternative hypothesis. We shall now present the derivation of the formula. The alternative hypothesis which will be considered is really much more general than at first appears, since it is always possible, by a continuous transformation, to transform one of the distributions to the uniform distribution, and, by the method used by Wald and Wolfowitz [24] in the proof of their Theorem II, it is possible to extend the result to a large class of couples of continuous cumulative d.f.'s. By suitable manipulation and use of the formula, it is therefore possible to obtain the large-sample power of the statistic U for the alternatives of statistical importance. However, it must be emphasized that this is a result valid only for large-sample sizes.

Let X and Y be chance variables with the respective cumulative d.f.'s $f(x)$ and $g(x)$. Let n_1 and n_2 observations be made on X and Y, respectively. Arrange the $n_1 + n_2$ observations in ascending order, and replace each by 1 or 2, according as the observation is one on X or one on Y, respectively. Denote the sequence of 1's and 2's by $V = v_1, v_2, \cdots, v_{n_1+n_2}$. Then U can be defined as follows:

$$U = 1 + \sum_{j=2}^{n_1+n_2} (v_j - v_{j-1})^2 .$$

Suppose that:

a) $f(x) \equiv 0 \quad (x < 0)$
$f(x) \equiv x \quad (0 \leqq x \leqq 1)$
$f(x) \equiv 1 \quad (x > 1)$
b) $g(x) \equiv 0 \quad (x \leqq 0)$
$g(x) \equiv 1 \quad (x \geqq 1)$

c) The derivative $g'(x)$ of $g(x)$ exists, is continuous and positive everywhere in the interval $0 \leqq x \leqq 1$

d) $n_1/n_2 = c$, a constant.

In section 3 above we have already mentioned the result as obtained by Wald and Wolfowitz [24] in their Theorem II, lemma 2. We wish now to prove that

$$\lim_{n_1 \to \infty} \sigma^2\left(\frac{U}{\sqrt{n_1}}\right) = 4\left[\int_0^1 \frac{cg'^2}{(c+g')^3}\,dx + \int_0^1 \frac{g'(c^3+g'^3)}{(c+g')^4}\,dx - \left(\int_0^1 \frac{g'^2}{(c+g')^2}\,dx\right)^2 - c^3\left(\int_0^1 \frac{g'}{(c+g')^2}\,dx\right)^2\right].$$

The result we wish to prove is an asymptotic one. This should be borne in mind in all that follows; in the interest of brevity we may omit explicit limit statements where no confusion is caused thereby below. The limiting process is with respect to n_1 approaching infinity; in the interest of typographical convenience we shall write simply n for n_1 where no confusion can result therefrom.

Let $\Delta = n^{-1/5}$ and let the interval $0 \leqq x \leqq 1$ be divided into subintervals of length Δ. Let t_i be the number of runs in the ith interval. Then $\sum_i t_i - U$ is at most $n^{1/5}$, since at most one new run is created in each subinterval by the partitioning. Since the variance of U is of order n and its covariance with the new runs created is of a smaller order, it follows that the limiting value of the variance of $n^{-1/2}\sum_i t_i$ is the same as that of $n^{-1/2}\,U$. We may therefore confine ourselves to investigating the first of these two.

In what follows we intend to let $n \to \infty$, so that $\Delta \to 0$, and shall omit writing terms of type $o(\Delta)$. Let ν represent a general normally distributed chance variable, with zero mean and unit variance; different ν will be distinguished by subscripts. The number of 1's which fall in the ith interval is the chance variable

$$n\Delta(1 + \sqrt{1/n\Delta}\,\nu_{1i})$$

and the number of 2's which fall in the ith interval is the chance variable

$$\frac{n\Delta}{c}\left(g_i' + \sqrt{\frac{cg_i'}{n\Delta}}\,\nu_{2i}\right),$$

where g_i' is the value of $g'(x)$ at, say, the midpoint of the ith interval. When these are the numbers of the two kinds of elements, the ratio c' of the number of elements 1 in the ith interval to the number of elements 2 in the ith interval is [always to within terms of type $o(\Delta)$]

$$c' = \frac{c}{g_i'}\left(1 + \frac{\nu_{1i}}{\sqrt{n\Delta}} - \sqrt{\frac{c}{ng_i'\Delta}}\,\nu_{2i}\right).$$

The (conditional) variance of t_i when these element numbers are fixed is therefore, using the limiting form of formula (13) of Wald and Wolfowitz [23],

$$4n\Delta\left(1 + \frac{1}{\sqrt{n\Delta}}\nu_{1i}\right)\frac{c'}{(1+c')^3} = 4n\Delta\,\frac{cg_i'^2}{(c+g')^3}\left(1 + \frac{1}{\sqrt{n\Delta}}\nu_{1i}\right)$$
$$\times\left(1 + \frac{1}{\sqrt{n\Delta}}\nu_{1i} - \sqrt{\frac{c}{ng_i'\Delta}}\,\nu_{2i}\right)\times\left(1 - \frac{3g'}{(c+g')}\left[\frac{c\nu_{1i}}{g_i'\sqrt{n\Delta}} - \frac{c^{3/2}\nu_{2i}}{g'^{3/2}\sqrt{n\Delta}}\right]\right).$$

Taking expected values with respect to the hitherto fixed number of elements, we have that the leading term of this contribution to the variance of t_i is

$$\frac{4n\Delta cg_i'^2}{(c+g_i')^3}.$$

The variance of t_i consists of the sum of the above and the expected value of the square of the difference between the conditional and absolute mean values of t_i. The former is

$$\frac{2n\Delta\left(1 + \frac{1}{\sqrt{n\Delta}}\nu_{1i}\right)}{1+c'} = \frac{2ng_i'\Delta}{c+g_i'}\left(1 + \frac{\nu_{1i}}{\sqrt{n\Delta}}\right)\left(1 - \frac{c}{c+g_i'}\left[\frac{\nu_{1i}}{\sqrt{n\Delta}} - \frac{\sqrt{c}\,\nu_{2i}}{\sqrt{ng_i'\Delta}}\right]\right).$$

The absolute mean value of t_i is of course

$$\frac{2ng_i'\Delta}{c+g_i'}.$$

The required contribution to the variance of t_i is therefore the expected value of

$$\frac{4ng_i'^2\Delta}{(c+g_i')^2}\left(\frac{g_i'\nu_{1i}}{c+g_i'} + \frac{c^{3/2}\nu_{2i}}{\sqrt{g_i'}\,(c+g_i')}\right)^2,$$

the terms of lower order being omitted. Since the number of elements 1 which fall in the ith interval is independent of the number of elements 2 which fall in

the ith interval, ν_{1i} and ν_{2i} are independently distributed. The required expected value is therefore

$$\frac{4ng_i'(c^3 + g'^3_i)\Delta}{(c + g_i')^4}.$$

The variance of t_i is obtained by summing the two parts obtained above. Summing the sum of these two parts for all the subintervals and passing to the limit, we have that the sum of the variances of the t_i is

$$4n\left[\int_0^1 \frac{cg'^2(x)}{[c + g'(x)]^3}dx + \int_0^1 \frac{g'(x)[c^3 + g'^3(x)]}{[c + g'(x)]^4}dx\right].$$

This is not yet the variance of U, because the various t_i are not independent and the covariances must be taken into account. We now proceed to do the latter.

We note that t_i and t_j $(i \neq j)$ are related only through the fact that the number of elements l $(l = 1, 2)$ which fall in the ith interval is correlated with the number of elements l which fall in the jth interval. Once the numbers of the different elements which fall in the two intervals are fixed, the chance variables t_i and t_j are independent, because they depend only on the relative position of the two kinds of elements. Let ν_{1j} and ν_{2j}, respectively, play the same role for the jth interval that ν_{1i} and ν_{2i} perform for the ith interval. From the well-known formula for the correlation between the numbers of observations falling into two classes of a multinomial distribution we have, always neglecting terms of type $o(\Delta)$,

$$E(\nu_{1i}\nu_{1j}) = -\Delta,$$
$$E(\nu_{2i}\nu_{2j}) = -\sqrt{g_i'g_j'}\,\Delta.$$

Also

$$E(\nu_{1i}\nu_{2i}) = E(\nu_{1i}\nu_{2j})$$
$$= E(\nu_{1j}\nu_{2j}) = E(\nu_{1j}\nu_{2i}) = 0,$$

since the two chance variables in any one parenthesis are independent.

Define the following quantities:

$\left.\begin{matrix} A_1 \\ B_1 \end{matrix}\right\}$ the absolute expected values of $\left\{\begin{matrix} t_i \\ t_j \end{matrix}\right\}$

$\left.\begin{matrix} A_2 \\ B_2 \end{matrix}\right\}$ the conditional expected values of $\left\{\begin{matrix} t_i \\ t_j \end{matrix}\right\}$ when the numbers of the different elements in the ith and jth intervals are fixed

$\left.\begin{matrix} A_3 \\ B_3 \end{matrix}\right\}$ the deviations of $\left\{\begin{matrix} t_i \\ t_j \end{matrix}\right\}$ from $\left\{\begin{matrix} A_2 \\ B_2 \end{matrix}\right\}$

$\sigma(t_it_j)$ = covariance between t_i and t_j.

We have

$$t_i = A_2 + A_3$$
$$t_j = B_2 + B_3$$

$$\sigma(t_i t_j) = E(t_i - A_1)(t_j - B_1)$$

$$E([A_2 - A_1] + A_3)([B_2 - B_1] + B_3)$$
$$= E[E^*([A_2 - A_1] + A_3)([B_2 - B_1] + B_3)],$$

where E^* denotes the conditional expected value taken when the chance variables ν_{1i}, ν_{2i}, ν_{1j}, ν_{2j} are held fixed (i.e., in the universe where the numbers of elements 1 and 2 which fall in the ith and jth intervals are held fixed). Because of the character of the dependence between t_i and t_j (described in the preceding paragraph), we have

$$E^*([A_2 - A_1] + A_3)([B_2 - B_1] + B_3) = [A_2 - A_1][B_2 - B_1],$$

since $E(A_3) = E(B_3) = 0$, and A_3 and B_3 are independently distributed. Now $A_2 - A_1$ has been shown to be given by

$$A_2 - A_1 = \frac{2g_i'\sqrt{n\Delta}}{(c + g_i')}\left(\frac{g_i'\nu_{1i}}{c + g_i'} + \frac{c^{3/2}\nu_{2i}}{\sqrt{g_i'}(c + g_i')}\right),$$

and similarly we have

$$B_2 - B_1 = \frac{2g_j'\sqrt{n\Delta}}{(c + g_j')}\left(\frac{g_j'\nu_{1j}}{c + g_j'} + \frac{c^{3/2}\nu_{2j}}{\sqrt{g_j'}(c + g_j')}\right).$$

Employing the formulas given above for the expected values of products of two ν, we obtain

$$\sigma(t_i t_j) = E[A_2 - A_1][B_2 - B_1]$$
$$= \frac{4g_i'g_j'n\Delta}{(c + g_i')(c + g_j')}\left(\frac{-g_i'g_j'\Delta}{(c + g_i')(c + g_j')} + \frac{-c^3\Delta}{(c + g_i')(c + g_j')}\right).$$

When the right member is summed over all pairs (i, j) we obtain

$$-4n\left[\left(\int_0^1 \frac{g'^2(x)}{[c + g'(x)]^2}dx\right)^2 + c^3\left(\int_0^1 \frac{g'(x)}{[c + g'(x)]^2}dx\right)^2\right].$$

Adding this to the sum of the variances of the t_i, we obtain the desired result.

REFERENCES

1. BOWKER, A. H. "Note on consistency of a proposed test for the problem of two samples," *Annals of Math. Stat.*, vol. 15 (1944), pp. 98–101.
2. FISHER, R. A. *Statistical Methods for Research Workers*, sec. 24, example 19. Oliver & Boyd, Edinburgh, 1925.
3. ———. *The Design of Experiments*, sec. 21. Oliver & Boyd, Edinburgh, 1935.
4. HOTELLING, H., 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.
5. KOLMOGOROFF, A. "Sulla determinatione empirica di una legge di distributione," *Giornale Ist. Ital. Attuari*, vol. 4 (1933), pp. 83–91.
6. MATHISEN, H. C. "A method of testing the hypothesis that two samples are from the same population," *Annals of Math. Stat.*, vol. 14 (1943), pp. 188–194.
7. MOOD, A. M. "The distribution theory of runs," *Annals of Math. Stat.*, vol. 11 (1940), pp. 367–392.
8. MOSTELLER, F. "Note on application of runs to quality control charts," *Annals of Math. Stat.*, vol. 12 (1941), pp. 228–232.
9. NEYMAN, J., "Outline of a theory of statistical estimation based on the classical theory of probability." *Philos. Trans. Roy. Soc. London*, Ser. A, vol. 236 (1937), pp. 333-380.
10. NEYMAN, J., and E. S. PEARSON. "On the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. Roy. Soc. London*, Ser. A, vol. 231 (1933), p. 289.
11. PEARSON, E. S. "Some aspects of the problem of randomization," *Biometrika*, vol. 29 (1937), pp. 53–64, and vol. 30 (1938), pp. 159–179.
12. PITMAN, E. J. G. "Significance tests which may be applied to samples from any populations," *Suppl. Jour. Roy. Stat. Soc.*, vol. 4 (1937), pp. 117–130.
13. SCHEFFÉ, H. "On a measure problem arising in the theory of non-parametric tests," *Annals of Math. Stat.*, vol. 14 (1943), pp. 227–233.
14. ———. "Statistical inference in the non-parametric case," *ibid.*, pp. 305–332.
15. SHEWHART, W. A. "Contribution of statistics to the science of engineering," in *Fluid Mechanics and Statistical Methods in Engineering* (University of Pennsylvania, Bicentennial Conference), pp. 97–124. Also *Bell Telephone System Technical Publications*, Monograph B-1319.
16. SMIRNOFF, N. "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bull. Math. Univ. Moscou*, Sérië internationale, vol. 2, fasc. 2 (1939).
17. STEVENS, W. L. "Distribution of groups in a sequence of alternatives," *Annals of Eugenics*, vol. 9 (1939), pp. 10–17.
18. SWED, F. S., and C. EISENHART. "Tables for testing randomness of grouping in a sequence of alternatives," *Annals of Math. Stat.*, vol. 14 (1943); pp. 66–87.
19. THOMPSON, W. R. "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.
20. WALD, A. "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, vol. 54 (1943), pp. 426–482.
21. ———. "An extension of Wilks' method for setting tolerance limits," *Annals of Math. Stat.*, vol. 14 (1943), pp. 45–55
22. WALD, A., and J. WOLFOWITZ. "Statistical tests based on permutations of the observations," *Annals of Math. Stat.*, vol. 15 (1944), pp. 358–372.
23. ———. "Confidence limits for continuous distribution functions," *ibid.*, vol. 10 (1939), pp. 105–118.
24. ———. "On a test whether two samples are from the same population," *ibid.*, vol. 11 (1940), pp. 147–162.
25. ———. "An exact test for randomness in the non-parametric case, based on serial correlation," *ibid.*, vol. 14 (1943), pp. 378–388.

26. WILKS, S. S. "Determination of sample sizes for setting tolerance limits," *Annals of Math. Stat.*, vol. 12 (1941), pp. 91–96.
27. ———. "Statistical prediction with special reference to the problem of tolerance limits," *ibid.*, vol. 13 (1942), pp. 400–409.
28. WOLFOWITZ, J. "Additive partition functions and a class of statistical hypotheses," *Annals of Math. Stat.*, vol. 13 (1942), pp. 247–279.
29. ———. "On the theory of runs with some applications to quality control," *ibid.*, vol. 14 (1943), pp. 280–288.