CONSISTENCY AND UNBIASEDNESS OF CERTAIN NONPARAMETRIC TESTS

By E. L. Lehmann

University of California, Berkeley

1. Summary. It is shown that there exist strictly unbiased and consistent tests for the univariate and multivariate two- and k-sample problem, for the hypothesis of independence, and for the hypothesis of symmetry with respect to a given point. Certain new tests for the univariate two-sample problem are discussed. The large sample power of these tests and of the Mann-Whitney test are obtained by means of a theorem of Hoeffding. There is a discussion of the problem of tied observations.

2. Introduction. The purpose of the present paper is to investigate the existence and various properties of strictly unbiased and of consistent tests for testing certain nonparametric hypotheses. The problems that will be considered are the two-sample and k-sample problem, the hypothesis of independence and the hypothesis of symmetry with respect to a given point.

A sequence of tests is said to be consistent against a certain class of alternatives if for each alternative the power of the test tends to one as the sample sizes tend to infinity. A test will be said to be strictly unbiased if the power for each alternative exceeds the level of significance.

Consistency being a rather weak property, which one would expect most sequences of tests to satisfy for the class of alternatives for which they are designed, it is important to obtain some more detailed information concerning the power of the various tests under consideration. Because of the tremendous variety of the alternatives it seems fairly hopeless to get a comprehensive view of the achievements of most tests when the samples are small. This in spite of the fact that it is occasionally possible to write down the power explicitly (for example in the simplest cases of the tests discussed by Mathisen [1]). On the other hand, the large sample distribution of a number of test statistics may be found by means of the asymptotic theorems of Hoeffding [2]. Asymptotically, the power then usually involves only a few parameters and a large sample comparison of various different tests becomes possible.

3. Two-sample problem: specific classes of alternatives. We shall discuss in detail only one of the problems mentioned, the two-sample problem, and indicate only briefly certain extensions to the other problems. In the two-sample problem one is given independent samples $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ from populations with unknown cumulative distribution functions F and G respectively, and it is desired to test the hypothesis $F = G$. In this connection various classes of alternatives are possible.

It may, for example, be known that unless $F = G$, the Y's tend to be larger than the X's. For this problem it has been proposed as a test to consider the

number of pairs X_i, Y_j for which $X_i < Y_j$, and to reject the hypothesis if this number is too large. This test was proved consistent by Mann and Whitney [3] against the alternatives that

$$F(t) > G(t) \quad \text{for all} \quad t. \tag{3.1}$$

Actually their proof shows that the test is consistent[1] against all alternatives for which $P(Y_j > X_i) > \frac{1}{2}$.

We shall now prove that this test is also unbiased against the alternatives satisfying (3.1).[2] This is true not only for this test but also for those proposed by Thompson [4] and for tests based on randomisation of such statistics as $\bar{y} - \bar{x}$. In fact we have

THEOREM 3.1. *Let w be any similar region for testing H: $F = G$ on the basis of $X_1, \cdots, X_m; Y_1, \cdots, Y_n$. Suppose w is such that $(x_1, \cdots, x_m; z_1, \cdots, z_n) \varepsilon w$ and $y_i \geqq z_i$ for $i = 1, \cdots, n$ implies $(x_1, \cdots, x_m; y_1, \cdots, y_n) \varepsilon w$. Then the test is unbiased against all continuous alternatives F, G satisfying* (3.1).

PROOF. Suppose that $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are independent and all have the same c.d.f. F and that G is such that (3.1) holds. Then we shall construct $Y_i = f(Z_i)$ such that $Y_i > Z_i$ for $i = 1, \cdots, n$ and such that the Y's have c.d.f. G. Thus the probability of $(X_1, \cdots, X_m; Z_1, \cdots, Z_n) \varepsilon w$ equals the level of significance, α say, while the probability of $(X_1, \cdots, X_m; Y_1, \cdots, Y_n) \varepsilon w$ equals the power of the test against the alternative (F, G). But since $Y_i > Z_i$ for all i, the test rejects for the X's and Y's whenever it rejects for the X's and Z's, and hence the power is $\geqq \alpha$.

The function f is easily defined by the equation

$$G(f(z)) = F(z).$$

(When this does not define $f(z)$ uniquely, it does not matter which of the possible definitions is used.) That $y = f(z) > z$ follows from assumption (3.1).

The theorem as stated refers only to tests in which no randomisation is allowed, but the extension to randomised tests is immediate. Also, as we shall show later, the assumption of continuity of F and G may be omitted.

Theorem 3.1 may be used also to widen the applicability of the tests to which it refers. So far, we have taken the hypothesis to state that X and Y have the same distribution. This formulation may arise, for example, when one is faced with the question whether a treatment, known to be harmless, has a beneficial effect: Either it has no effect so that $F = G$, or it has a good effect. If, on the other hand, the comparison is between two different treatments one may wish to test hypothesis H' that Y tends to be smaller than X, against the alternatives that it tends to be larger. The hypothesis would then be

$$H': F(t) \leqq G(t) \quad \text{for all} \quad t.$$

---

[1] This was also noticed by van Dantzig who points it out in a paper "On the consistency and the power function of Wilcoxon's two sample test," to be published in *Proc. Roy. Inst. Acad. Sci.*, 1951.

[2] For alternatives (F, G) differing only in location this was proved by Van der Vaart [26].

There is of course no nontrivial similar region for this problem, however any region w satisfying the condition of Theorem 3.1 and such that $P(w) = \alpha$ whenever $F \neq G$ clearly will be of size α for testing H' i.e. $P(w)$ will be $\leqq \alpha$ whenever $F(t) \leqq G(t)$ for all t.

Returning to the Mann-Whitney test, let us define V by

$$mnV = \text{number of pairs } X_i, Y_j \text{ with } X_i < Y_j. \tag{3.2}$$

It was shown by Mann and Whitney that V is asymptotically normally distributed when $F = G$ and $m, n \to \infty$ in an arbitrary manner. From a result of Hoeffding (Theorem 7.3 of [2]) it follows that asymptotic normality holds also when $F \neq G$ provided m/n remains constant as $m, n \to \infty$.

We shall apply Hoeffding's theorem to prove asymptotic normality not only of V, but of a large class of statistics connected with the two-sample problem. We begin by stating Hoeffding's theorem, somewhat specialised and with slight changes of notation:

Let $Z_1, \cdots, Z_n$ be independently, identically distributed chance vectors with real components, let $s \leqq n$ and let $\phi(Z_1, \cdots, Z_s)$ be a real valued symmetric function of its s arguments such that $E[\phi(Z_1, \cdots, Z_s)]^2 < \infty$, and let us write

$$E\phi(Z_1, \cdots, Z_s) = \theta.$$

Let

$$U_n = \binom{n}{s}^{-1} \Sigma\phi(Z_{\alpha_1}, \cdots, Z_{\alpha_s}),$$

where the summation extends over all subscripts $1 \leqq \alpha_1 < \cdots < \alpha_s \leqq n$, and let

$$U_n' = U_n + R_n,$$

where R_n is a random variable for which

$$E(nR_n^2) \to 0 \quad \text{as} \quad n \to \infty.$$

Then $\sqrt{n}(U_n' - \theta)$ is asymptotically normally distributed. Further, if we put

$$\psi(z_1) = E[\phi(z_1, Z_2, \cdots, Z_s) - \theta],$$

the limiting distribution of $\sqrt{n}(U_n' - \theta)$ is nondegenerate provided $E[\psi(Z_1)]^2 > 0$.

We can now state

THEOREM 3.2. *Let $X_1, \cdots, X_m$; $Y_1, \cdots, Y_n$ be independently distributed with* c.d.f.'s *F, G respectively. Let $t(x_1, \cdots, x_r, y_1, \cdots, y_r)$ be symmetric in the x's alone and in the y's alone. Suppose that*

$$Et(X_1, \cdots, X_r, Y_1, \cdots, Y_r) = \theta(F, G) = \theta,$$

$$E[t(X_1, \cdots, X_r, Y_1, \cdots, Y_r)]^2 = M < \infty.$$

Let $m/n = c$, and let n be sufficiently large so that $r \leqq m, n$. Define

$$U_n' = \binom{m}{r}^{-1}\binom{n}{r}^{-1} \Sigma t(X_{\alpha_1}, \cdots, X_{\alpha_r}, Y_{\beta_1}, \cdots, Y_{\beta_r}),$$

where the summation is extended over all subscripts $1 \leqq \alpha_1 < \cdots < \alpha_r \leqq m$; $1 \leqq \beta_1 < \cdots < \beta_r \leqq n$. *Then, as* $n \to \infty$, $\sqrt{n}(U'_n - \theta)$ *is asymptotically normally distributed.*

PROOF. For the sake of simplicity we shall give the proof only in the case $m = n$. Let $Z_i = (X_i, Y_i)$ and define

$$\phi(Z_1, \cdots, Z_{2r}) = \binom{2r}{r}^{-1} \Sigma t(X_{i_1}, \cdots, X_{i_r}, Y_{j_1}, \cdots, Y_{j_r})$$

summed over all sets of indices for which $(i_1 < \cdots < i_r, j_1 < \cdots < j_r)$ is a permutation of $(1, \cdots, 2r)$. Further let

$$U_n = \binom{n}{2r}^{-1} \Sigma\phi(Z_{\gamma_1}, \cdots, Z_{\gamma_{2r}})$$

summed over all γ's such that $1 \leqq \gamma_1 < \cdots < \gamma_{2r} \leqq n$.

Clearly $\binom{n}{r}^2 U'_n$ is the sum of all possible t-terms, while $\binom{2r}{r}\binom{n}{2r}U_n$ is the sum of only those t-terms in which the X's and Y's have no common subscript. Hence, since $\binom{n}{2r}\binom{2r}{r} = \binom{n}{r}\binom{n-r}{r}$, we have

$$U'_n = \binom{n-r}{r}\binom{n}{r}^{-1} U_n + \binom{n}{r}^{-2} W_n,$$

where W_n is a sum of $\left[\binom{n}{r}\binom{n}{r} - \binom{n}{r}\binom{n-r}{r}\right]$ t-terms, and we can write

$$U'_n = U_n + D_n,$$

where

$$D_n = \left[\binom{n-r}{r} - \binom{n}{r}\right]\binom{n}{r}^{-1} U_n + \binom{n}{r}^{-2} W_n.$$

Since for any real numbers $t_1, \cdots, t_k$ we have $(t_1 + \cdots + t_k)^2 \leqq k(t_1^2 + \cdots + t_k^2)$ we see that

$$E(D_n^2) \leqq \left[\binom{n}{r} - \binom{n-r}{r}\right]^2 \binom{n}{r}^{-2} M + \left[\binom{n}{r} - \binom{n-r}{r}\right]\binom{n}{r}^{-3} M.$$

But, as $n \to \infty$, $\sqrt{n}\left[\binom{n}{r} - \binom{n-r}{r}\right]\binom{n}{r}^{-1} \to 0$. Hence $E(nD_n^2) \to 0$ and the result follows.

Let us now consider the application of this theorem to the Mann-Whitney statistic. We define

$$t(x, y) = \begin{cases} 1 & \text{if } y > x, \\ 0 & \text{if } y \leqq x. \end{cases}$$

Then $U_n' = V_{m,n}$ and asymptotic normality follows since $E\psi^2(X, Y) \leqq 1$. It remains to check under what conditions $E\psi^2(Z_1) > 0$. Since we have $s = 2r = 2$,

$$2\psi(z_1) = P(Y_2 > x_1) + P(y_1 > X_2) - 2P(Y > X).$$

Hence $E\psi^2(Z) = 0$ is equivalent to $F(Y) - G(X) =$ constant with probability 1, or $P(Y > x) + P(y > X) =$ constant except on a set of points (x, y) that has probability zero. It is easy to see that this is satisfied if and only if $P(Y > X)$ is 1 or 0.

So far we have considered the hypothesis H: $F = G$ against the alternatives that the Y's tend to be larger than the X's. As a second example we shall consider testing H, or even the wider hypothesis H' that F and G differ only in location (i.e., that $F(x) = G(x + d)$ for some d), against the alternative that the Y's are more spread out than the X's (in a sense to be defined below). In analogy with the Mann-Whitney test let $W_{m,n}$ be the proportion of quadruples X_i, X_j, Y_k, Y_l for which $|Y_l - Y_k| > |X_j - X_i|$. We reject H if $W_{m,n}$ is too large. This test is unbiased against all alternatives (F, G) for which $F(x_1) = G(y_1), F(x_2) = G(y_2)$ implies $|x_1 - x_2| < |y_1 - y_2|$. The test is consistent against the wider class of alternatives for which $P(|Y' - Y| > |X' - X|) > \frac{1}{2}$ where X, X', Y, Y' are independently distributed with c.d.f. F, G, respectively. The proof of unbiasedness is quite analogous to the one given previously, and we shall therefore omit it.

We shall however indicate the proof of consistency, and refer in this connection to the closely related remarks by Hoeffding [5] on the construction of consistent sequences of tests.

We first state for reference the following trivial

LEMMA 3.1. *Let $\theta = f(F, G)$ be a real valued function such that $f(F, F) = \theta_0$ for all (F, F) in a class $\mathcal{C}_0$. Let $T_{m,n} = t_{m,n}(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ be a sequence of real valued statistics such that $T_{m,n}$ tends to θ in probability as* min $(m, n) \to \infty$. *Suppose that $f(F, G) > \theta_0 (\neq \theta_0)$ for all (F, G) in a class $\mathcal{C}_1$. Then the sequence of tests which reject when $T_{m,n} - \theta_0 > C_{m,n}$ (when $|T_{m,n} - \theta_0| > C'_{m,n}$) is consistent for testing H: $\mathcal{C}_0$ at every fixed level of significance against the alternatives $\mathcal{C}_1$.*

For proof one need only to notice that a fixed level of significance $\neq 0$ implies that $C_{m,n} \to 0$ ($C'_{m,n} \to 0$) as $m, n \to \infty$.

In the applications we have in mind, $E(T_{m,n})$ is usually independent of m and n, and is easy to find. On the other hand some work is required to determine $\sigma^2(T_{m,n})$. It is therefore of interest to notice that the evaluation of $\sigma^2(T_{m,n})$ is frequently not necessary to prove consistency. To this end we shall state the following lemma, which is a generalisation of a theorem of Halmos [6], and which follows easily from Theorem 5.1 of [7]. A simple proof will be given in [8].

LEMMA 3.2 (Lehmann-Scheffé). *Let $f(F, G)$ be a real valued function defined for all continuous* c.d.f.'s *F and G. There exists at most one function $t_{m,n}$ such that $t_{m,n}(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ is symmetric in the first m and in the last n arguments and is an unbiased estimate of $f(F, G)$ for all continuous (or even ab-*

solutely continuous) c.d.f.'s F, G. *If such a function* $t_{m,n}$ *exists, (and has finite variance), it has among all unbiased estimates of* $f(F, G)$ *uniformly smallest variance.*

For the application to be made here we need the slightly stronger statement that the conclusion of the Lemma remains valid if $t_{m,n}(X_1, \cdots, X_m ; Y_1, \cdots, Y_n)$ is an unbiased estimate of $f(F, G)$ for all continuous c.d.f.'s F, G for which

$$P(|Y' - Y| > |X' - X|) > \tfrac{1}{2}.$$

This generalization follows immediately from the proof of the Lemma given in [8].

The proof of consistency of the proposed test is now immediate. For let $W'_{m,n}$ be the proportion of quadruples for which the Y's are further apart than the X's among the independent quadruples $X_1, X_2, Y_1, Y_2 ; X_3, X_4, Y_3, Y_4 ; \cdots$. Then

$$E(W'_{m,n}) = E(W_{m,n}) = P(|Y' - Y| > |X' - X|),$$

and hence by Lemma 3.2,

$$\sigma^2(W_{m,n}) \leqq \sigma^2(W'_{m,n}). \tag{3.3}$$

But $\sigma^2(W'_{m,n})$ obviously tends to zero as $m, n \to \infty$.

We remark finally that the large sample distribution of $W_{m,n}$, by Theorem 3.1, is again normal. Degeneracy occurs only if either F or G are one-point distributions.

As a last problem in this section we shall consider the hypothesis $F = G$ against the combined class of alternatives that the Y's are larger than the X's or more spread out. In such a problem it seems important not only to decide whether F and G are equal but, in case the hypothesis is rejected, for which of the two possible reasons it is rejected or whether it is rejected for both of them. (See in this connection the discussion by Berkson [9]). Thus one is really dealing with a multidecision problem. One must decide between

d_0: Accepting the hypothesis $H: F = G$,

d_1: Rejecting H for the reason that the Y's are larger than the X's,

d_2: Rejecting H for the reason that the Y's are more spread out than the X's,

d_3: Rejecting H for both reasons.

It is desired to find a decision procedure under which the probability of taking decision d_0 is $1 - \alpha$ when $F = G$ while the probability of taking the appropriate of the decisions d_1, d_2, d_3 when the hypothesis is false tends to 1 as the sample sizes tend to infinity. Let us recall the statistics $V_{m,n}$ and $W_{m,n}$ introduced in connection with the previous problems and let us denote $E(V_{m,n})$ and $E(W_{m,n})$ by θ and η respectively. One may then accept H when $V_{m,n} \leqq a_{m,n}$, $W_{m,n} \leqq b_{m,n}$, or take one of the remaining three decisions according as to which one of the three complementary inequalities holds. The constants $a_{m,n}$ and $b_{m,n}$ are not completely determined by the equation

$$P(V_{m,n} \leqq a_{m,n}, W_{m,n} \leqq b_{m,n} \mid F = G) = \alpha.$$

One may specify some additional restriction, such as

$$P(V_{m,n} \leqq a_{m,n} \mid F = G) = P(W_{m,n} \leqq b_{m,n} \mid F = G).$$

It is easy to prove that the above procedure has the consistency property asked for. This follows from Lemmas 3.1 and 3.2 generalised to the case that the function $f(F, G)$ of these lemmas is vector valued instead of real valued. The function $t_{m,n}$ of Lemma 3.2 is then also vector valued and instead of its variance one may consider its ellipsoid of concentration (see [10] and Theorem 5.2 of [7]). In the present case we notice that $(V_{m,n}, W_{m,n})$ is a symmetric estimate of (θ, η) and hence has a uniformly smallest ellipsoid of concentration. But one can easily construct unbiased estimates of θ and η based on independent samples, whose concentration ellipsoid has both axes tending to zero as the sample sizes increase indefinitely and so consistency can be proved by the device used after Lemma 3.2.

4. Two-sample problem: general class of alternatives. In the present section we shall consider the problem of testing the hypothesis $F = G$ against the class of all continuous alternatives $F \neq G$. One might argue that this should not be treated as a hypothesis-testing problem. For Berkson's argument seems to apply: The question is not only whether or not the hypothesis is true. If it is false, it is necessary to decide what alternative hypothesis is correct. While in some situations, this criticism seems to be valid, there are others in which it does not seem to apply.

The two-sample problem may arise in the following two quite different settings.

A: Two production processes, treatments or populations are available, and it is desired to decide whether one is better than the other. In this case the populations F and G are in competition, and the main problem is that of ranking them. Here the notion of such a ranking automatically suggests some specific class or classes of alternatives to the hypothesis that the populations do not differ.

B: The two populations coexist. There is no question of which is preferable, but we wish to know whether the two can be treated as one. One may, for example, want to know whether the output of two different machines can be treated as a uniform product, or whether data obtained under two different experimental setups or by two different investigators may be pooled. These problems really are two-decision problems: The data can or can not be pooled. An explanation of why they can not be pooled is not necessarily of interest.

In connection with the present problem Wald and Wolfowitz [11] proposed as test statistic the total number of runs of the ordered x's and y's, the hypothesis to be rejected if the number of runs is too small. The authors proved their test consistent, under the assumption of constant ratio of sample sizes m/n, against alternatives of all shapes restricted only by mild assumptions, concerning existence and positiveness of the probability densities. It was also proved in their paper that the test statistic has an asymptotically normal distribution when the hypothesis is true. More recently Wolfowitz [12] proved that the limiting dis-

tribution is normal even when $F \neq G$, and obtained the asymptotic variance for this case. It follows from his results that the test is in general not consistent if $m/n \to 0$ or ∞. This is actually what one would expect since when m/n is sufficiently extreme the maximum number of runs will in general occur with near-certainty whether the hypothesis is true or false.

Another test suitable for this problem is that of Smirnov [13] based on the maximum difference between the two sample cumulative distribution functions. For the given samples $X_1, \cdots, X_m$; $Y_1, \cdots, Y_n$ let

$$\phi_m(t) = \phi(X_1, \cdots, X_m; t) = \frac{1}{m} \text{ (number of } X\text{'s} \leqq t),$$

$$\psi_n(t) = \psi(Y_1, \cdots, Y_n; t) = \frac{1}{n} \text{ (number of } Y\text{'s} \leqq t),$$

be the two sample cumulative distribution functions. It follows from a theorem of Glivenko-Cantelli [14], that $\sup_t |\phi_m(t) - F(t)|$ and $\sup|\psi_n(t) - G(t)|$ tend to zero in probability as $\min(m, n) \to \infty$. From this it is easily seen that $\sup|\phi_m(t) - \psi_n(t)|$ is a consistent estimate of $\sup|F(t) - G(t)|$, and hence that Smirnov's test is consistent against all alternatives $F \neq G$ as $\min(m, n) \to \infty$. A different proof of this fact was given recently by Massey [25].

The large sample distribution of $\sup|\phi_m(t) - \psi_n(t)|$ was obtained by Smirnov, for the case that $F = G$, a simpler proof having recently been given by Feller [15] (see also Doob [16] and Smirnov [17]). Although the large sample distribution is not known when $F \neq G$, Massey [25] obtained a lower bound for the power of Smirnov's test, which may permit comparing this test with others.

While these two generally consistent tests are known for the two-sample problem, very little work has been done on the existence of unbiased tests for this or other nonparametric problems. Mann [18] proved unbiasedness of a test for randomness against a certain class of trends. Hoeffding [5] proved the nonexistence for the hypothesis of independence of unbiased critical regions based on ranks, corresponding to certain very small levels of significance.

As far as the two-sample problem is concerned, Smirnov's test is easily shown to be biased on the basis of an example given by Massey for the problem of goodness of fit. On the other hand, it seems very possible that the Wald-Wolfowitz run test is unbiased whenever the two samples are of equal size. We have not proved this but shall now construct a test for the two sample problem that is strictly unbiased.

LEMMA 4.1 *Let X, X'; Y, Y' be independently drawn from populations with continuous cumulatives F, G respectively, and let us denote for any random variables U, U'; V, V' the event* $\max(U, U') < \min(V, V')$ *by $U, U' < V, V'$. Then*

$$p = P((X, X' < Y, Y') + (Y, Y' < X, X')) = \tfrac{1}{3} + 2\int (F - G)^2 \, d\left(\frac{F + G}{2}\right),$$

and hence p attains its minimum value $\frac{1}{3}$ if and only if $F = G$.

PROOF. Since F and G are continuous,

$$p = \int (1 - F)^2\, dG^2 + \int (1 - G)^2\, dF^2 = 2 + \int (F^2\, dG^2 + G^2\, dF^2)$$

$$- 4 \int FG\, d(F + G) = 2 + \int d(F^2 G^2) - 4 \int FG\, d(F + G)$$

$$= 3 - 2 \int [(F+G)^2 - (F - G)^2]\, d\left(\frac{F + G}{2}\right)$$

$$= \tfrac{1}{3} + 2 \int (F - G)^2\, d\left(\frac{F + G}{2}\right).$$

To prove the second part of the lemma, we must show that $\Delta = \int (F - G)^2\, d(F + G) = 0$ implies $F = G$. Now $\Delta = 0$ implies $F(x) = G(x)$ except possibly on a set N such that $\int_N dF = \int_N dG = 1$. Suppose that $F(x_1) \neq G(x_1)$, $G(x_1) - F(x_1) = \eta > 0$ say. Then by continuity there exists $x_0 < x_1$ such that $G(x_0) = F(x_0) + \eta/2$ and $F(x) < G(x)$ for $x_0 \leqq x \leqq x_1$. Since $G(x_1) - G(x_0) > 0$, it follows that $\Delta > 0$.

It is now clear that there exists a strictly unbiased test of $H: F = G$ if $m, n \geqq 2$. For we can consider the number of quadruples $X_{2i-1}, X_{2i}; Y_{2i-1}, Y_{2i}$ for which either the two X's fall below the two Y's or vice versa. These may be regarded as the successes in independent trials with probability $p = \frac{1}{3} + 2\Delta$ of success, and the problem reduces to that of testing $H: p = \frac{1}{3}$ against alternatives $p > \frac{1}{3}$.

The unbiased test just described has the pleasant property that its power is a strictly increasing function of $\Delta = \int (F - G)^2\, d\,\frac{F + G}{2}$, which seems a reasonable measure of the degree of difference of F and G. On the other hand one would not expect this test to be very efficient. More reasonable use of the data seems to be made if one modifies the test in the direction of the Mann-Whitney test described earlier. One would then compare each pair of X's with each pair of Y's, and reject H if among the $\binom{m}{2}\binom{n}{2}$ possible quadruples $X_i, X_j; Y_k, Y_l$ it happened too frequently that both X's lie on the same side of both Y's.

This test is no longer unbiased, but it is still consistent as follows from the argument given in the previous section. Further, the test retains the property that the statistic on which it is based provides an unbiased estimate, in fact the minimum variance unbiased estimate, of the quantity $\int (F - G)^2\, d\,\frac{F + G}{2}$. Finally, it is again easily seen that the distribution of the test statistic is approximately normal, degeneracy occurring only if $P(Y > X)$ equals 1 or 0.

The test can be expressed in a form more convenient for computation in terms of the ranks of one of the sets of variables. Let $r_1 < r_2 < \cdots < r_n$ be the ranks

of the n Y's among the totality of $m + n$ observations, and denote by $Q_{m,n}$ the number of quadruples X_i, X_j, Y_k, Y_l for which both X's lie on the same side of both Y's. Then it is easily seen that

$$Q_{m,n} = \sum_{k=1}^{n} \left[(n-k) \binom{r_k - k}{2} + (k-1) \binom{m - r_k + k}{2} \right].$$

From this it follows by easy computation that

$$2Q_{m,n} = (n-1) \sum_{k=1}^{n} r_k^2 - 2(n+m-2)\Sigma k r_k - (n-2m+1)\Sigma r_k + (n+2m-3) \frac{n(n+1)(2n+1)}{6} + (n+m^2-3m+1) \frac{n(n+1)}{2}.$$

It may perhaps be worth noting that the first of the two tests described in this section can also be used as the basis of a sequential test of the two sample problem. This is clear since the problem is simply reduced to that of testing a simple binomial situation against a one-sided class of alternatives. The sequential probability ratio test to which one is led in this manner of course is again unbiased and has a power function that is strictly increasing in $\int (F-G)^2 \, d\frac{F+G}{2}$.

The measure of discrepancy

$$\int (F-G)^2 \, d\frac{F+G}{2}$$

utilised in the tests of the present section, suggests using $\int (\phi_m - \psi_n) \, d\left(\frac{\phi_m + \psi_n}{2}\right)$ as test statistic. It should be pointed out that tests of this kind have been studied in connection with the closely related problem of goodness of fit by Cramér [19] and von Mises [20]. In the present case, let us denote the x's and y's in order of magnitude by $x^{(1)} < x^{(2)} < \cdots < x^{(m)}$; $y^{(1)} < y^{(2)} < \cdots < y^{(n)}$, let m_1 be the number of x's $< y^{(1)}$, m_2 the number of x's between $y^{(1)}$ and $y^{(2)}$ etc., and define $n_1, n_2, \cdots$ analogously. Then it is easily seen that

$$\int (\phi_m - \psi_n)^2 \, d(\phi_m + \psi_n) = \frac{1}{n}\left[\left(\frac{m_1}{m} - \frac{1}{n}\right)^2 + \left(\frac{m_1+m_2}{m} - \frac{2}{n}\right)^2 + \cdots + \left(\frac{m_1+\cdots+m_n}{m} - 1\right)^2\right] + \frac{1}{m}\left[\left(\frac{n_1}{n} - \frac{1}{m}\right)^2 + \cdots\right].$$

Tests of this type have been proposed by Dixon and by Mood [21], but have not been studied thoroughly.

Finally it should be mentioned that one might also try the method of randomisation, which has been considered by Pitman [22] and others in connection

with specific classes of alternatives, for the present problem. One statistic which may be suitable for this purpose if $m = n$ is $\sum_{i=1}^{n} (Y^{(i)} - X^{(i)})^2$.

5. Discontinuous distributions. So far, we have assumed F and G to be continuous. This assumption is obviously not satisfied in practice, and we must therefore consider the difficulties introduced by discontinuities. (These difficulties were investigated in connection with various estimation problems by Scheffé and Tukey [23]).

Let us restrict our attention to rank tests and introduce the convention that tied observations are ordered at random. Thus if $X_{i_1} = \cdots = X_{i_s} = Y_{j_1} = \cdots = Y_{j_t}, s + t = r$, we perform an experiment with $r!$ possible and equally probable outcomes. We then establish a 1:1 correspondence between the $r!$ possible orderings of r objects and these $r!$ outcomes, and treat the X's and Y's as if they had occurred in the order indicated by this experiment. If the X's and Y's have the same distribution it is then clear that the distribution of any rank statistic of the X's and Y's is what it would be if this common distribution were continuous since in both cases each possible ordering of the X's and Y's is again equally probable.

In order to see that various unbiasedness results of the preceding and following sections remain valid, we state the following

LEMMA 5.1. *Let* $\mathcal{E} = \mathcal{E}(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$ *be a random event depending only on the ranking of the* X*'s and* Y*'s. Suppose that* F *and* G *may have discontinuities and that in case of ties the event* $\mathcal{E}$ *is defined by ordering the tied observations at random. Then there exist continuous* c.d.f.'s $F^* = F^*(F, G)$ *and* $G^* = G^*(F, G)$ *such that*

$$P_{F,G}(\mathcal{E}) = P_{F^*,G^*}(\mathcal{E}).$$

and that $F^* = G^*$ *if and only if* $F = G$.

PROOF. We shall only give the construction of F^*, G^*; the remainder of the proof then follows easily.

Consider the (denumerable) totality of points that are points of discontinuity of either F or G, and suppose these points have been numbered: $x_1, x_2, \cdots$. Consider first x_1 and define two new c.d.f.'s F_1, G_1 as follows:

$$\begin{aligned} F_1(x) &= F(x + \tfrac{1}{4}) \quad \text{if} \quad x < x_1 - \tfrac{1}{4} \\ &= F(x_1^-) + \frac{x - (x_1 - \frac{1}{4})}{\frac{1}{2}} [F(x_1) - F(x_1^-)] \quad \text{if} \quad |x_1 - x| \leqq \tfrac{1}{4} \\ &= F(x - \tfrac{1}{4}) \quad \text{if} \quad x > x_1 + \tfrac{1}{4}. \end{aligned}$$

G_1 is defined analogously in terms of G. What this construction does is to push F and G apart at x_1 symmetrically by a total amount of $\frac{1}{2}$, and to distribute the probability at x_1 uniformly over the gap thus created.

In the same way we now push F_1 and G_1 apart at the second discontinuity (in its new position) by a total amount of $1/2^2$ and distribute the amount of jump

uniformly over the gap, thus obtaining F_2 and G_2. Then the sequence F_1, F_2, $\cdots$ will converge to a continuous distribution F^* and analogously for the G's and F^*, G^* will have the desired properties.

It follows from this lemma that the unbiased test of the hypothesis $F = G$ discussed in Section 4 remains strictly unbiased when the assumption of continuity of F and G is dropped. On the other hand, the power is no longer such a simple function of F and G. In fact let $X, X'; Y, Y'$ denote as before independent random variables with distributions F and G respectively and denote by $X, X' < Y, Y'$ that this ordering occurred after randomisation of ties. Then it is not difficult to show that

$$P((X, X' < Y, Y') + (Y, Y' < X, X')) = \tfrac{1}{3} + 2\Delta',$$

where

$$3\Delta' = \int [(F - G)^2 + (F^- - G^-)^2 + (F - G)(F^- - G^-)]\, d\,\frac{F + G}{2}.$$

Here $F^-(x) = F(x^-)$, $G^-(x) = G(x^-)$.

6. Existence of unbiased tests for the hypothesis of independence and some other nonparametric problems. In this last section we shall briefly consider some more complicated nonparametric problems. Our aim is to prove for all these problems the existence of strictly unbiased and consistent tests. The problem is treated purely theoretically in that no effort is made to construct tests that make good use of the data and that are convenient to apply, but that instead the sole purpose is to exhibit tests possessing the properties asked for.

For the hypothesis of independence Hoeffding proposed a test that he proved consistent against all alternatives with continuous joint and marginal probability densities. In this connection he also considered the problem of unbiasedness and proved the nonexistence of unbiased critical regions based on rank for certain small levels of significance. This negative result seems to contradict those of the present section. This is however not so. Hoeffding restricted his attention to critical regions while we are here admitting also randomised tests. It should be pointed out in this connection that, while randomisation was used in previous sections only in a trivial manner, namely so as to get the exact level of significance, we shall here make very heavy use of this device. This could be avoided in part, however the tests would then become more complicated. Further if the problem is reduced, as is done here, to that of testing equality of two binomial p's, randomisation is needed to get an exactly similar test.

The hypothesis of independence states that the joint c.d.f. equals the product of the two marginal c.d.f.'s. Thus if $(X_i^{(1)}, X_i^{(2)})$, $i = 1, 2, \cdots$, are independently drawn from a bivariate distribution F, it is equivalent to the hypothesis that the pair $(X_1^{(1)}, X_1^{(2)})$ comes from the same bivariate population as the pair $(X_2^{(1)}, X_3^{(2)})$. It is therefore clear that if we can prove the existence of strictly unbiased and consistent tests for the bivariate two-sample problem, this will

imply the existence of tests with these properties for the hypothesis of independence. The same remark clearly applies to hypothesis of independence (both complete independence and independence of sets of variates) in more than two variables.

Consider now samples $X_i = (X_i^{(1)}, \cdots, X_i^{(k)})$ $i = 1, 2, \cdots$ and $Y_j = (Y_j^{(1)}, \cdots, Y_j^{(k)})$ $j = 1, 2, \cdots$ from two k-variate distributions F and G. The work of section 4 suggests utilising the expression

$$\int (F - G)^2 \, d\left(\frac{F+G}{2}\right) = \int (F^2 + G^2) \, d\left(\frac{F+G}{2}\right) - 2\int FG \, d\left(\frac{F+G}{2}\right).$$

All that is necessary is to construct events A and B such that

$$p_1 = P(A) = \int \frac{F^2 + G^2}{2} \, d\left(\frac{F+G}{2}\right),$$

$$p_2 = P(B) = \int FG \, d\left(\frac{F+G}{2}\right).$$

The hypothesis $H: F = G$ will then be reduced to $H': p_1 = p_2$ to be tested against alternatives $p_1 > p_2$. The events A and B may be defined as follows:

A: With probability $\frac{1}{2}$ observe either X_1, X_2 or Y_1, Y_2 and with probability $\frac{1}{2}$ observe either X_3 or Y_3. Denote the three variables that are observed by Z_1, Z_2, Z_3, and define A as the event

$$Z_1^{(i)}, Z_2^{(i)} \leqq Z_3^{(i)} \qquad \text{for } i = 1, \cdots, k.$$

B: Observe X_4, Y_4 and with probability $\frac{1}{2}$ either X_5 or Y_5. If the last of these variables is denoted by Z_5, define B as the event

$$X_4^{(i)}, Y_4^{(i)} \leqq Z_5^{(i)} \qquad \text{for } i = 1, \cdots, k.$$

It should be mentioned that instead of observing five random vectors some of which may be either X's or Y's, we could have obtained a test with the desired property based on ten observations, five X's and five Y's.

To complete the proof we must show that the hypotheses H and H' are really equivalent, that is, that $p_1 = p_2$ if and only if $F = G$. For the case that F and G are continuous this follows immediately by an argument similar to the one given in the univariate case, and it is easy to show it even without this restriction.

It is clear that one can generalise further and instead of two samples consider s samples. For this purpose one may replace $\int (F - G)^2 \, d\left(\frac{F+G}{2}\right)$ for example by $\sum_{i=1}^{s} (F_i - \bar{F})^2 \, d\bar{F}$ where $\bar{F}$ is the average of the s c.d.f.'s. Alternatively, one may utilise the expression $\sum_{i<j} \int (F_i - F_j)^2 \, d\left(\frac{F_i + F_j}{2}\right)$.

As a last problem let us consider a sample $X_1, \cdots, X_n$ from an unknown

univariate c.d.f. F, assumed to be continuous. It is desired to test the hypothesis H of symmetry with respect to the origin, i.e., that $F(x) = 1 - F(-x)$ for all x. Smirnov [24] recently proposed $\max_x \{ | N^+(x) - N^-(x) | \}$ as a test statistic where $N^+(x)$, $N^-(x)$ denote the number of x's contained in the intervals $(0, x)$, $(-x, 0)$ respectively.

The work of Section 4 suggests considering 4 X's (X_i, X_j, X_k, X_l) and defining the following two events.

A: Exactly two of the four X's are positive.

B: If A is satisfied, and $X_i, X_j < 0 < X_k, X_l$, say, the event B is said to occur if neither

$$|X_i|, |X_j| < X_k, X_l \qquad \text{nor} \qquad X_k, X_l < |X_i|, |X_j| .$$

Then if $F(0) = p_0$, $P(A) = 6p_0^2q_0^2$ takes on its maximum value 3/8 if and only if $p_0 = 1/2$. Further, $P(B \mid A)$ takes on its maximum value 2/3 if and only if the conditional distribution F^* of $-X$ given $X < 0$ is the same as that, G^*, of X given $X > 0$. Thus

$$P(AB) = 6p_0^2 q_0^2 \left\{ \frac{2}{3} - 2 \int (F^* - G^*)^2 \, d\left(\frac{F^* + G^*}{2} \right) \right\}$$

takes on its maximum value 1/4 if and only if the hypothesis of symmetry holds.

If we apply this method to independent quadruples, we obtain a test that is strictly unbiased and consistent. If we apply it to all possible quadruples the test remains consistent and may be a reasonable test for the hypothesis in question. Hoeffding's theory can again be applied to the asymptotic distribution problem.

REFERENCES

[1] 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.

[2] W. Hoeffding, "A class of statistics with asymptotically normal distributions," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 293–325.

[3] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[4] 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.

[5] W. Hoeffding, "A non-parametric test of independence," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 546–557.

[6] P. R. Halmos, "The theory of unbiased estimation," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 34–43.

[7] E. L. Lehmann and H. Scheffé, "Completeness, similar regions, and unbiased estimation—Part I," *Sankhyā*, Vol. 10 (1950), pp. 305–340.

[8] E. L. Lehmann and H. Scheffé, "Completeness, similar regions, and unbiased estimation, Part II," unpublished.

[9] J. BERKSON, "Comments on Dr. Madow's 'Note on tests of departure from normality' with some remarks concerning tests of significance," *Jour. Am. Stat. Assn.*, Vol. 36 (1941), p. 539.

[10] H. CRAMÉR, "Contributions to the theory of statistical estimation," *Skandinavisk Aktuarietidskrift*, Vol. 29 (1946), p. 85.

[11] 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.

[12] J. WOLFOWITZ, "Non-parametric statistical inference," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1949.

[13] N. V. SMIRNOV, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bull. Math. Univ. Moscow, Serie Int.*, Vol. 2 (1939).

[14] M. FRÉCHET, *Recherches Théoriques Modernes sur la Théorie des Probabilités, Vol. 1*, p. 260, Gauthier-Villars, Paris, 1937.

[15] W. FELLER, "On the Kolmogorov-Smirnov limit theorems for empirical distributions," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 177–189.

[16] J. L. DOOB, "Heuristic approach to the Kolmogorov-Smirnov theorems," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 393–403.

[17] N. V. SMIRNOV, "Approximate laws of distribution of random variables from empirical data," *Uspehi Matem. Nauk*, Vol. 10 (1944), p. 179.

[18] H. B. MANN, "Non-parametric tests against trend," *Econometrica*, Vol. 13 (1945), pp. 245–259.

[19] H. CRAMÉR, "On the composition of elementary errors," *Skandinavisk Aktuarietidskrift*, Vol. 11 (1928), p. 13 and p. 141.

[20] R. VON MISES, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, Franz Deuticke, Leipzig, 1931.

[21] W. 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.

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

[23] H. SCHEFFÉ AND J. W. TUKEY, "Non-parametric estimation. I. Validation of order statistics," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187–192.

[24] N. V. SMIRNOV, "Sur un critère de symetrie de la loi de distribution d'une variable aléatoire," *C. R. Acad. Sci. URSS*, Vol. 56 (1947), p. 11.

[25] F. J. MASSEY, JR., "A note on the power of a non-parametric test," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 440–443.

[26] H. R. VAN DER VAART, "Some remarks on the power function of Wilcoxon's test for the problem of two samples. I," *Indagationes Math.*, Vol. 12 (1950), pp. 146–158.