MODIFIED RANDOMIZATION TESTS FOR NONPARAMETRIC HYPOTHESES[1]

By Meyer Dwass

Northwestern University and *Stanford University*

1. Introduction and summary. Suppose $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are $m + n = N$ independent random variables, the X's identically distributed and the Y's identically distributed, each with a continuous cdf. Let

$$z = (z_1, \cdots, z_m, z_{m+1}, \cdots, z_N) = (x_1, \cdots, x_m, y_1, \cdots, y_n)$$

represent an observation on the N random variables and let

$$u(z) = (1/m) \sum_{i=1}^{m} z_i - (1/n) \sum_{i=m+1}^{N} z_i = \bar{x} - \bar{y}.$$

Consider the $r = N!$ N-tuples obtained from $(z_1, \cdots, z_N)$ by making all permutations of the indices $(1, \cdots, N)$. Since we assume continuous cdf's, then with probability one, these r N-tuples will be distinct. Denote them by $z^{(1)}, \cdots, z^{(r)}$, and suppose that they have been ordered so that

$$u(z^{(1)}) \geqq \cdots \geqq u(z^{(r)}).$$

Notice that since

$$\bar{x} - \bar{y} = (1/m) \sum_{i=1}^{N} z_i - (N/m)\bar{y} = (N/n)\bar{x} - (1/n) \sum_{i=1}^{N} z_i,$$

the same ordering can be induced by choosing $u(z) = c\bar{x}$ or $u(z) = -c\bar{y}$ for any $c > 0$.

Assuming that the cdf's of X_1, Y_1 are of the form $F(x)$, $F(x - \Delta)$ respectively, Pitman [2] suggested essentially the following test of the hypothesis H' that $\Delta = 0$. Select a set of k $(k > 0)$ integers $i_1, \cdots, i_k$, $(1 \leqq i_1 < \cdots < i_k \leqq r)$. If the observed z is one of the points $z^{(i_1)}, \cdots, z^{(i_k)}$, reject H', otherwise accept. When H' is true, the type one error does not depend on the specific form of the distribution of the X's and the Y's and is in fact equal to k/r. The choice of the rejection set $i_1, \cdots, i_k$ should depend on the alternative hypothesis. For instance, if the experimenter wants protection against the alternative that the "X's tend to be larger than the Y's," then the labels $1, \cdots, k$ might be reasonable. For the alternative that the "X's tend to be smaller than the Y's" the analogous procedure is to use the other tail, $r - k + 1, \cdots, r$. Against both alternatives,

Received, January 31, 1956.

[1] This work was supported in part by an Office of Naval Research contract at Stanford University.

TABLE 1

Under each α heading, the left-hand column is computed from (8) and the right-hand column from a normal approximation. Computations were made only for those values of s such that $d + 1 = \alpha(s + 1)$ is an integer.

$\alpha^{-1}A(\alpha)$

s	α							
	.01		.02		.05		.10	
19					.642		.743	
39					.736		.815	
49			.636				.834	
59					.782		.848	
79					.810		.868	
99	.634	.618	.732	.726	.829	.827	.881	.881
119					.843	.842	.892	.891
149			.778	.775			.903	.902
199		.725		.804		.877		.915
299		.774		.840		.900		.931
499		.824		.876		.922		.946
999		.875		.912		.945		.962

a two-tail procedure could be used. Lehmann and Stein have shown in [1] that in the class of all tests (of size $\alpha = k/r$) of the hypothesis

H: the distribution of $X_1 \cdots, X_m, Y_1, \cdots, Y_n$ is invariant under all permutations,

the single-tail test based on $1, \cdots, k$ is uniformly most powerful against the alternatives that F_1 is an $N(\theta, \sigma)$ cdf, F_2 is an $N(\theta + \Delta, \sigma)$ cdf, $\Delta < 0$; the test based on $r - k + 1, \cdots, r$ is uniformly most powerful for $\Delta > 0$.

A practical shortcoming of this procedure is the great difficulty in enumerating the points $z^{(i)}$ and the evaluation of $u(z^{(i)})$ for each of them. For instance, even after eliminating those permutations which always give the same value of u, then for sample sizes $m = n = 5$, there are $\binom{10}{5} = 252$ permutations to examine, and for sample sizes $m = n = 10$, there are $\binom{20}{10} = 184{,}765$ permutations to examine. In the following section, we propose the almost obvious procedure of examining a "random sample" of permutations and making the decision to accept or reject H on the basis of those permutations only. Bounds are determined for the ratio of the power of the original procedure to the modified one. Some numerical values of these bounds are given in Table 1. The bounds there listed correspond to tests which in both original and modified form have size α, and for which the modified test is based on a random sample of s permutations drawn with replacement. These have been computed for a certain class of alterna-

tives which is described below. For simplicity, we have restricted the main exposition to the two-sample problem. In Section 5, we point out extensions to the more general hypotheses of invariance studied in [1].

2. Description of modified procedure. We first make some definitions. For any $z = (z_1, \cdots, z_N)$, let $T(z)$ be the set of all points obtained from z by permuting its coordinates. With probability one, all sets $T(z)$ contain $r = N!$ points $z^{(1)}, \cdots, z^{(r)}$ and we restrict our discussion to such sets. We also suppose that they are ordered in the manner described earlier. Define $R^{(i)}$ to be the union over all sets $T(z)$ of the points $z^{(i)}$, $(i = 1, \cdots, r)$. Evidently $R^{(1)}, \cdots, R^{(r)}$ are disjoint sets whose union is the whole sample space except for a set of probability zero. Let $P(i) = P(R^{(i)})$. Restricting ourselves to the case $\Delta < 0$, we describe the Pitman procedure given above in terms of a test φ, as follows:

$$\varphi(z) = \begin{cases} 1 & \text{if } u(z) \geqq u(z^{(k)}), \\ 0 & \text{if } u(z) < u(z)^{(k)}); \end{cases} \qquad (1 \leqq k \leqq r),$$

where $z^{(1)}, \cdots, z^{(r)}$ are the points of $T(z)$ and $\varphi(z)$ is the probability with which H is rejected when z is observed. Let $r^+ = r^+(z)$ be the number of $z^{(i)}$ in $T(z)$ such that $u(z^{(i)}) \geqq u(z)$. (Notice that $R^{(i)}$ is the event that $r^+(z) = i$.) Then the test described is equivalent to rejecting H when $r^+ \leqq k$ and accepting otherwise.

The modified procedure will be to make this decision on the basis of examining a random subset of $T(z)$. Specifically we describe a modified test φ_L as follows: Select at random s $(s < r)$ points of $T(z)$. For simplicity, we suppose the sampling from $T(z)$ is done with replacement. Let r'^+ equal the number of the s points for which $u(z^{(i)}) \geqq u(z)$. Then we define

$$\varphi_L^{(z)} = \begin{cases} 1 & \text{if } r'^+ \leqq d, \\ 0 & \text{if } r'^+ > d, \end{cases}$$

where $d(0 \leqq d \leqq s)$ is a predetermined integer. We point out that φ_L is a randomized test and that r'^+ depends not only on z but on the s points of $T(z)$ selected. Let

$$\Psi(t) = \sum_{i=0}^{d} \binom{s}{i} t^i (1-t)^{s-i}, \qquad (0 \leqq t \leqq 1).$$

The following is easily verified.

PROPOSITION 1.

$$E\varphi_L = \sum_{i=1}^{r} \Psi(i/r)P(i), \qquad E\varphi = \sum_{i=1}^{k} P(i). \tag{1}$$

REMARK. In particular, when H is true, then $P(i) = 1/r$ and for large r, $E_H\varphi_L$ is approximately equal to

$$\int_0^1 \Psi(t)\, dt. \tag{2}$$

In what follows, we always assume that

$$E_H\varphi_L = E_H\varphi = k/r.$$

Notice that $\Psi(t)$ is a nonincreasing function in (0, 1). This fact is used in deriving the following bounds in Propositions 2 and 3.

PROPOSITION 2.

$$E\varphi_L \geqq \Psi(\alpha)E\varphi, \qquad (\alpha = k/r). \tag{3}$$

The above bound is quite weak. On the other hand, equality in (3) is attained only when $P(i) = 0$, $(i \neq k)$, $P(k) = 1$. It would not be unreasonable to say that the alternatives against which φ can be expected to be effective are those satisfying

$$P(1) \geqq P(2) \geqq \cdots \geqq P(r). \tag{4}$$

In particular, (4) is satisfied when the $P(i)$ are the probabilities induced by any simple alternative against which φ is most powerful for all $0 < \alpha < 1$. According to [1], this is true for the normal alternatives described in the introduction, uniformly for $\Delta < 0$. Hence, we shall next determine the value of inf $E\varphi_L / E\varphi$ over all $P(1), \cdots, P(r)$ satisfying (4), and such that φ, φ_L have size $\alpha = k/r$.

PROPOSITION 3. Suppose (4) is satisfied. Then

$$E\varphi_L \geqq k^{-1} \sum_{i=1}^{k} \Psi(i/r)E\varphi. \tag{5}$$

PROOF.

$$E\varphi_L / E\varphi = \sum_{i=1}^{r} \Psi(i/r)P(i) \Big/ \sum_{j=1}^{k} P(j)$$

$$= \sum_{i=1}^{k} \Psi(i/r)P(i) \Big/ \sum_{j=1}^{k} P(j) + \sum_{i=k+1}^{r} \Psi(i/r)P(i) \Big/ \sum_{j=1}^{k} P(j).$$

Hence, by replacing $P(i)$ with $P(i) / \sum_{j=1}^{k} P(j)$ for $i = 1, \cdots, k$, and with 0 for $i = k + 1, \cdots, r$, we do not increase the value of $E\varphi_L / E\varphi$ and we may as well assume at the outset that $P(k + 1) = \cdots = P(r) = 0$. Now by the monotonicity of Ψ, it is easy to see that subject to (4), $\sum_{i=1}^{k} \Psi(i/r)P(i) / \sum_{j=1}^{k} P(j)$ is minimized when $P(1) = \cdots = P(k) = 1/k$, which completes the proof.

REMARKS.

(a) It is evident from the proof that (5) holds if (4) is replaced by

$$P(1) \geqq \cdots \geqq P(k). \tag{4'}$$

(b) By (5), $E\varphi_L / E\varphi \geqq (r/k) \sum_{i=1}^{k} \Psi(i/r)/r$. For large r, $\sum_{i=1}^{k} \Psi(i/r)/r$ is approximately equal to $\int_0^\alpha \Psi(t)\, dt$; hence inf $E\varphi_L / E\varphi$ over all $P(i)$ satisfying (4') approximately equals

$$\alpha^{-1} \int_0^\alpha \Psi(t)\, dt. \tag{6}$$

Let $B[s, t]$ denote the number of successes in s independent binomial trials with t the probability of success in each. Then

$$P(B[s,t] \leqq d) = \Psi(t) = 1 - \frac{s!}{(s-d-1)!d!}\int_0^t u^d(1-u)^{s-d-1}\,du.$$

Let $A(t) = \int_0^t \Psi(u)\,du$. After integration by parts, we have

$$\begin{aligned} A(t) &= t\Psi(t) + s\binom{s-1}{d}\int_0^t u^{d+1}(1-u)^{s-d-1}\,du \\ &= t\Psi(t) + (d+1)/(s+1)P(B[s+1,t] \geqq d+2). \end{aligned} \tag{7}$$

By (2), $E_{B\varphi} = E_{B\varphi_L} = k/r$ is approximately equal to $A(1) = (d+1)/(s+1)$. Suppose d and s are chosen so that $(d+1)/(s+1) = k/r = \alpha$. Then by (7), the value of (6) is

$$\alpha^{-1}A(\alpha) = P(B[s,\alpha] \leqq d) + P(B[s+1,\alpha] \geqq d+2). \tag{8}$$

Some values of $\alpha^{-1}A(\alpha)$ are given in Table 1.

3. Concluding remarks for the two-sample problem.

(a) The main point is that instead of basing our decision on $\binom{m+n}{n}$ permutations of the observations, we can base it on a smaller number of permutations and the power of the modified test will be "close" to that of the most powerful nonparametric test. It may be argued that s still has to be ridiculously large. For instance, if $\alpha = .05$, $s = 10^3$, then (8) equals .945; and if $\alpha = .05$, $s = 10^4$, then (8) equals .98. However, the optimum test is usually completely impossible. For instance, if $m = 20$, $n = 20$, then $\binom{m+n}{n} > 10^{11}$, and if there were a machine that could check 10 permutations a second, the job would run something on the order of 1000 years. The point is, then, that an impossible test can be made at least possible, if not always practical.

(b) For some alternatives, the efficiency of the modified test may be better than the bound in (8) would indicate, since we would often expect a *strictly* decreasing sequence in (4).

(c) For moderate size s there may be reasonable hand-computing procedures. A possibility is the following: Enter each of the $m + n$ observations on a separate card. Perform s "random shufflings." For each shuffle, sum the first m entries and record.

(d) An open problem which may be worth investigating, at least empirically, is the following: For what value of s is the modified test already better than some given *rank order* test, or in particular, than the rank order test which is best against the alternative under consideration?

4. Generalizations. Lehmann and Stein [1] have studied randomization tests in a general framework. We do not describe here the most general setup, but rather one to which the results of the earlier sections are adaptable. Suppose

$(Z_1, \cdots, Z_N) = Z$ are N random variables and there is a partition of the sample space of points $(z_1, \cdots, z_N) = z$ into classes of equivalent points. For instance, in the Pitman example, two points are equivalent if the coordinates of one can be obtained from the other by a permutation. For simplicity, we suppose that with probability one, each equivalence class $T(z)$ contains a finite number, r, of points. Let H be the hypothesis that the distribution of $(Z_1, \cdots, Z_N) = Z$ is, for any z, invariant over all the points of $T(z)$. (This is stated here in a somewhat unrigorous way. For the correct statement and for the necessary measurability assumptions, see [1].) A test of H is a function φ which assigns to each point z a number, $\varphi(z)$ between zero and one representing the probability of rejecting H when z is observed. If

$$\sum_{z' \in T(z)} \varphi(z') = \alpha r$$

identically in z, then φ is a similar size-α test for testing H. Lehmann and Stein have shown in [1] that under quite general circumstances, a most powerful and similar size-α test of H against a simple alternative is given by ordering the points of $T(z)$, so that

$$u(z^{(1)}) \geqq \cdots \geqq u(z^{(r)}),$$

and setting

$$\varphi(z) = \begin{cases} 1 & \text{if } u(z) > u(z^{(1+[\alpha r])}), \\ a & \text{if } u(z) = u(z^{(1+[\alpha r])}), \\ 0 & \text{if } u(z) < u(z^{(1+[\alpha r])}), \end{cases}$$

where u is an appropriately chosen function and $a = a(z)$ is uniquely determined to provide a size-α test. We assume that the random variable $u(Z)$ has a continuous cdf and that the size of the test is k/r where k is an integer $(1 \leqq k \leqq r)$. The effect of this assumption is to eliminate ties and to provide a nonrandomized test with probability one. We can now describe a modified test procedure exactly as was done in the two-sample case above. There is no reason to suppose that s items are to be selected at random and with replacement from the set $T(z)$ when z is observed, however. Any "lot acceptance" plan for deciding whether or not $r^+(z) \leqq k$ would be appropriate; for instance, the elements of $T(z)$ can be selected without replacement or sequentially, etc. Let

$$\Psi(t) = P\{\text{deciding } r^+ \leqq k \mid r^+/r = t\}.$$

Notice that this coincides with the definition in the special case studied previously. Now Proposition 1 goes through in exactly the same way as before. If $\Psi(u)$ is a nonincreasing function in (0, 1) (which is a negligible restriction), then Propositions 2 and 3 also go through exactly as before. It is also true that for large r, $\int_0^1 \Psi(t)\, dt$ is practically equal to α and that the lower bound on the efficiency of φ_L versus φ, under the condition (4′) is practically equal to $\alpha^{-1} \int_0^\alpha \Psi(t)\, dt$,

but in general these quantities may be more difficult to compute than they were for the earlier special case.

REFERENCES

[1] E. L. LEHMANN AND C. STEIN, "On the theory of some nonparametric hypotheses," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 28–45.

[2] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations," *J. Roy. Stat. Soc.*, Vol. 4 (1937a), pp. 119–130.