# RECURSIVE GENERATION OF THE DISTRIBUTION OF THE MANN-WHITNEY-WILCOXON U-STATISTIC UNDER GENERALIZED LEHMANN ALTERNATIVES

By Roger A. Shorack

*U. S. Navy Electronics Laboratory, San Diego*

**0. Summary.** Let $X$ and $Y$ be independent random variables having strictly increasing continuous distribution functions $F$ and $G$ respectively. It is shown that for hypotheses of the form $H: F = G$ versus $K_1: G = F^k$, $k > 0$ or $H: F = G$ versus $K_2: G = 1 - (1 - F)^k$, $k > 0$, the alternative distribution of the MWW $U$-statistic can be generated by a recursive formula analogous to the one used by Mann and Whitney in [2] to generate the null distribution of $U$.

**1. Introduction.** In the literature the class of alternatives $K_1$ is usually referred to as Lehmann alternatives. For the purposes of this paper we will refer to the class of alternatives $K_1$ as $L_1$ alternatives and to the class of alternatives $K_2$ as $L_2$ alternatives. The two classes taken together will be called generalized Lehmann alternatives. Note that if $k$ is a positive integer, $F^k$ and $1 - (1 - F)^k$ are respectively the df's of the maximum and minimum of $k$ independent identically distributed random variables with common df $F$.

Let $x_1, \cdots, x_n, y_1, \cdots, y_m$ be independent observations on $X$ and $Y$. Let $U$ be the number of times $y_j < x_i$, $i = 1, \cdots, n; j = 1, \cdots, m$. It is well known that $U$ can also be expressed in terms of the $Y$-ranks. Specifically, if $S_1, \cdots, S_m$ are the $Y$-ranks with respect to the combined sample then $U = mn + m(m + 1)/2 - \sum_{j=1}^{m} S_j$. In [1] Lehmann considers the power of the MWW $U$-test for the hypotheses $H: F = G$ versus $K: G = g(F)$ for certain $g$ and, for the cases $g(v) = v^k$, derives the expression

$$(1) \quad P(S_1 = s_1, \cdots, S_m = s_m) = k^m \Gamma(s_j + jk - j)\Gamma(s_{j+1})/\binom{m+n}{n}\Gamma(s_{j+1} + jk - j)\Gamma(s_j).$$

A similar expression could easily be obtained for $g(v) = 1 - (1 - v)^k$. We will refer to this expression as $(1')$ but will not bother to exhibit it here. To obtain the power against $K_1(K_2)$ one sums (1) ((1')) over the set of points $\{(s_1, \cdots, s_m)\}$ comprising the critical region.

In [3] Savage also gives a fairly complicated expression by which one can obtain the power of the $U$-test against ordinary Lehmann alternatives.

In the present paper it is shown that the distribution of $U$ under generalized Lehmann alternatives can be generated by a simple recursive formula. The formula is easily programmed for automatic machine computation. If the generation is being done by hand it is not necessary to produce the entire distribution

Received 26 April 1965; revised 23 August 1965.

for most power calculations. The expression to be given is easier to use than either of the previously mentioned expressions in most cases.

**2. General distribution of $U$ and the power of the $U$ test.** Mann and Whitney [2] give the distribution of $U$ when $F = G$ as (using their notation)

$$(2) \qquad p_{nm}(u) = [m/(m+n)]p_{n,m-1}(u) + [n/(m+n)]p_{n-1,m}(u-m).$$

Consider now $L_1$ alternatives; that is $G = F^k$, $k > 0$. First let $k = a/b$, $a$ and $b$ positive integers, and let $g(v) = F^{-1}[F^{1/b}(v)]$, where $F^{-1}$ is the inverse function of $F$. Define $X_i^* = g(X_i)$, $i = 1, \cdots, n$ and $Y_j^* = g(Y_j)$, $j = 1, \cdots, m$. Then $X_i^*$ has df $F^b$ and $Y_j^*$ has df $F^a$. Since the transformation is strictly increasing the distribution of $U$ is unaffected.

Let $W_{11}, \cdots, W_{b1} : \cdots : W_{1n}, \cdots, W_{bn}$ and $Z_{11}, \cdots, Z_{a1} : \cdots : Z_{1m}, \cdots, Z_{am}$ be independent and identically distributed random variables with df $F$. Then $X_i = \max_{1 \leqq t \leqq b} W_{ti}$ has df $F^b$, $i = 1, \cdots, n$ and $Y_j = \max_{1 \leqq t \leqq a} Z_{tj}$ has df $F^a, j = 1, \cdots, m$. Replace $W_{i1}$ by $-1$'s, $\cdots$, $W_{in}$ by $-n$'s, $i = 1, \cdots, b$ and replace $Z_{j1}$ by 1's, $\cdots$, $Z_{jm}$ by $m$'s, $j = 1, \cdots, a$. Define $A_{bn,am}(u)$ to be the number of distinguishable configurations of $-b$'s, $\cdots$, $-1$'s, 1's, $\cdots$, $a$'s having $U = u$. Then

$$(3) \qquad p_{nm}(u) = A_{bn,am}(u)/[(bn+am)!/(b!)^n(a!)^m].$$

Considering configurations for which the right extremity is first negative and then positive we obtain the recursion formula

$$(4) \qquad A_{bn,am}(u) = n\binom{bn+am-1}{b-1}A_{b(n-1),am}(u-m) + m\binom{bn+am-1}{a-1}A_{bn,a(m-1)}(u).$$

The first term on the right hand side comes from the configurations with a negative element last. Each of the elements $-1, \cdots, -n$ are a possibility. The factor $\binom{bn+am-1}{b-1}$ is the number of ways of putting the remaining $b-1$ indistinguishable integers (which match the right extremity) into the $b(n-1)+am+1$ cells formed by the other integers. The argument $u - m$ is a result of the fact that the value of the statistic $U$ is reduced by $m$ for each of these configurations if the right extremity and all elements matching it are deleted. A similar argument establishes the second term on the right hand side. From (3) it is apparent that the division of (4) by $(bn+am)!/(b!)^n(a!)^m$ yields

$$(5) \qquad p_{nm}(u) = [n/(km+n)]p_{n-1,m}(u-m) + [km/(km+n)]p_{n,m-1}(u).$$

Note that (5) reduces to (2) if $k = 1$.

Now suppose $G = F^k$, $k > 0$, $k$ *irrational*. Let $Q_{F,k}(u)$ be the df of the probability distribution given by (5) for $k > 0$ but $k$ not necessarily rational. Define the distance between two df's $F$ and $G$ as $d(F, G) = \sup_x |F(x) - G(x)|$. Then $d(Q_{F,k}, Q_{F,k'}) = \max_{u=0,1,\cdots,nm} |Q_{F,k}(u) - Q_{F,k'}(u)|$. Let $T_F(k)$ map the pair $(F, F^k)$ onto the df of $U$ when $X$ has df $F$ and $Y$ has df $F^k$. We have already shown that for positive rational $k$ $T_F(k) = Q_{F,k}$. Using (1) it is straightforward to show that $T_F$ is uniformly continuous in the sense that given $\varepsilon > 0$ there exists $\delta > 0$

such that if $0 < |k - k'| < \delta$ then $d(Q_{F,k}, Q_{F,k'}) < \varepsilon$. Let $k > 0$ be irrational and let $\{r_n\}_{n=1}^{\infty}$ be a sequence of positive rationals such that $\lim_n r_n = k$. Then $\lim_n T_F(r_n) = \lim_n Q_{F,r_n}$. But by (5) $\lim_n Q_{F,r_n}$ is easily seen to be $Q_{F,k}$. Since $T_F$ is continuous at $k$, $\lim_n T_F(r_n) = T_F(\lim_n r_n) = T_F(k)$. Thus $T_F(k) = Q_{F,k}$. Hence (5) holds for all $k > 0$.

The derivation for $L_2$ alternatives with rational $k$ is almost identical except we now use the transformation $g(v) = F^{-1}[1 - (1 - F(v))^{1/b}]$ to obtain $X_i^*$ and $Y_j^*$ and take $X_i = \min_{1 \leqq t \leqq b} W_{ti}$ and $Y_j = \min_{1 \leqq t \leqq a} Z_{tj}$. Then $X_i$ has df $1 - (1 - F)^b$ and $Y_j$ has df $1 - (1 - F)^a$. By considering configurations for which the *left* extremity is first negative and then positive we obtain the recursion formula

$$(4') \quad A_{bn,am}(u) = n\binom{bn+am-1}{b-1}A_{b(n-1),am}(u) + m\binom{bn+am-1}{a-1}A_{bn,a(m-1)}(u - n).$$

and finally

$$(5') \quad p_{nm}(u) = [n/(km + n)]p_{n-1,m}(u) + [km/(km + n)]p_{n,m-1}(u - n).$$

The proof for irrational $k$ remains unchanged except that (1′) is used in place of (1).

**3. Application.**

EXAMPLE. Consider the exponential distribution; that is let $F_\theta(x) = 1 - \exp[-x/\theta]$, $x > 0, \theta > 0$. Then for $\nu > 0$, let $F_\nu(x) = 1 - [1 - F_\theta(x)]^{\theta/\nu}$. For hypotheses of the form $H: F = G$ versus $K: F = F_\theta$, $G = F_\nu$ the power of the MWW $U$-test can be evaluated *exactly*. Thus we can compute (for example) the power function of a one sided MWW test of $H: \theta/\nu = 1$ versus $K: \theta/\nu < 1$.

**4. Conclusion.** It appears as though one could also generate the alternative distribution of $U$ recursively when $G = \sum_{i=1}^{k} a_i F^i$, $\sum a_i = 1$, but the method of proof used here becomes quite cumbersome to carry through.

For $m$, $n$ of even moderate size the recursive formula presented here is somewhat easier to use than Lehmann's procedure and in addition yields the entire alternative distribution as a matter of course. It is of course possible to generate the entire alternative distribution using (1) or (1′) but for large $m$, $n$ it is troublesome to have to determine what set of $Y$-ranks $\{(s_1, \cdots, s_m)\}$ is associated with a given value of $U$. It should also be noticed that the result of the present note is obtained by more elementary methods than is (1).

The recursive method is also much easier to use than the one given by Savage in [3].

**5. Acknowledgment.** I wish to thank the referee for suggesting the combinatorial method of proof to replace the more cumbersome one given by the author in the original manuscript.

REFERENCES

[1] LEHMANN, E. L. (1953). The power of rank tests. *Ann. Math. Statist.* **24** 24–43.

[2] MANN, H. B. and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18** 50–60.

[3] SAVAGE, I. R. (1956). Contributions to the theory of rank order statistics—the two sample case. *Ann. Math. Statist.* **27** 590–615.