On Characterization of Two-Sample U-Statistics

E. Schechtman*
Department of Industrial Engineering & Management
Ben Gurion University of the Negev
Beer Sheva, Israel
email: ednas@bgumail.bgu.ac.il

G. Schechtman
Department of Mathematics
The Weizmann Institute
Rehovot, Israel
email: gideon@wisdom.weizmann.ac.il

Abstract

A verifiable condition for a symmetric statistic to be a two-sample U-statistic is given. As an illustration, we characterize which linear rank statistics with two-sample regression constants are U-statistics. We also show that invariance under jackknifing characterizes a two-sample U-statistic.
KEY WORDS: Jackknife, kernel, linear rank statistic.

*corresponding author

1. INTRODUCTION.

Let $(X_1, \ldots, X_n; Y_1, \ldots, Y_m)$ be two independent random samples from distributions with c.d.f's $F(x)$ and $G(y)$, respectively. Let θ be a parameter and suppose that (r, t) are the smallest sample sizes for which there exists a function h, symmetric in its r X's and t Y's, such that
$E(h(X_1, \ldots, X_r; Y_1, \ldots, Y_t)) = \theta$, for every F, G in some family of distributions. Then, the two-sample U-statistic with kernel h, and of degree (r, t), is, for $n \geq r$, $m \geq t$,

$$\begin{aligned} U_{n,m} &= U(X_1, \ldots, X_n; Y_1, \ldots, Y_m) \\ &= \binom{n}{r}^{-1} \binom{m}{t}^{-1} \sum_{(n,r)} \sum_{(m,t)} h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_t}), \end{aligned}$$

where $\sum_{(n,r)}(\sum_{(m,t)})$ denotes summation over all distinct subsets $i_1, \ldots, i_r$ of the integers $1, \ldots, n$ ($j_1, \ldots, j_t$ of the integers $1, \ldots, m$).

U-statistics were first studied by Hoeffding in [1], and were further investigated by many authors. (See, for example, [2], [3], [4]). Once a statistic is known to be a U statistic, there are many desirable properties it possesses, which are already known and ready to use. However, the question faced by the researcher is - is the test statistic a U-statistic? An answer to this question, for a one-sample statistic was provided by Lenth in [5]. The purpose of this paper is to extend Lenth's results to two-sample U-statistics. In his paper, Lenth gives an easily verifiable condition for a one sample symmetric statistic to be a U-statistic, and shows that invariance under jackknife characterizes one-sample U-statistics. In the case of a two-sample statistic, the jackknife procedure is less obvious, since the term "leave one out" is not at all clear - one X at a time, then one Y at a time? One pair (X, Y) at a time?. The main result of this paper, Theorem 1 provides a verifiable condition under which a statistic is a two-sample U-statistic. The condition can also be useful in finding the kernel and the degree of the statistic. As a corollary we show that, using Arvesen's jackknifing procedure [6], invariance under jackknifing characterizes a two-sample U-statistic. These resultss, as well as an additional version of the first one, are discussed in section 2. In section 3 we illustrate the use of the conditions assuring that a two-sample statistic is a U-statistic by characterizing which linear rank statistics with two-sample regression constants (in [2], p.252, terminology) are U-statistics: We show that among this class of statistics, the **only** ones which are U-statistics are

linear transformations of the Mann-Whitney statistic. The proof here is more involved than the ones in section 2, which are not very complicated.

2. THE MAIN RESULTS

Let $S_{n,m}(X_1,\ldots,X_n;Y_1,\ldots,Y_m); n \geq r, m \geq t$ be a sequence of symmetric two-sample statistics (that is, symmetric in the X's and symmetric in the Y's). Following Arvesen's jackknife procedure, denote

(i) $S^0_{n,m}$ = the original statistic, based on all observations

(ii) $S^{-i,0}_{n-1,m}$ = the statistic, after leaving X_i out, for $n > r$

(iii) $S^{0,-j}_{n,m-1}$= the statistic, after leaving Y_j out, for $m > t$.

Theorem 1 *Let $S_{n,m}$ be a sequence of symmetric statistics, and suppose that for $n > r$, $m > t$,*

$$S_{n,m}(X_1,\ldots,X_n;Y_1,\ldots,Y_m) = \frac{(n-1)\sum_{i=1}^n S^{-i,0}_{n-1,m} + (m-1)\sum_{j=1}^m S^{0,-j}_{n,m-1}}{n(n-1)+m(m-1)}. \tag{1}$$

with boundary conditions

$$S_{n,t}(X_1,\ldots,X_n;Y_1,\ldots,Y_t) = n^{-1}\sum_{i=1}^n S^{-i,0}_{n-1,t} \qquad \text{for } n > r \tag{2}$$

$$S_{r,m}(X_1,\ldots,X_r;Y_1,\ldots,Y_m) = m^{-1}\sum_{j=1}^m S^{0,-j}_{r,m-1} \qquad \text{for } m > t. \tag{3}$$

Then, $S_{n,m}$ is a two-sample U-statistic of degree at most (r,t), i.e., $S_{n,m}$ is a U-statistic of degree (r',t') with $r \geq r', t \geq t'$. Conversely, any two-sample U-statistic satisfies (1).

Proof By the one-sample theorem ([5], proposition 2) and by (2), for all $n > r$ and for fixed $Y_1,\ldots,Y_t$,

$$S_{n,t}(X_1,\ldots,X_n;Y_1,\ldots,Y_t) = \binom{n}{r}^{-1}\sum_{(n,r)} S_{r,t}(X_{i_1},\ldots,X_{i_r};Y_1,\ldots,Y_t),$$

where the sum $\sum_{(n,r)}$ is taken over all distinct subsets $i_1, \ldots, i_r$ of the integers $1, \ldots, n$. Similarly, for all $m > t$ and for fixed $X_1, \ldots, X_r$,

$$S_{r,m}(X_1, \ldots, X_r; Y_1, \ldots, Y_m) = \binom{m}{t}^{-1} \sum_{(m,t)} S_{r,t}(X_1, \ldots, X_r; Y_{j_1}, \ldots, Y_{j_t}).$$

We shall prove, by induction on $n + m$, that

$$\begin{aligned} &S_{n,m}(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \\ &\quad = \binom{n}{r}^{-1} \binom{m}{t}^{-1} \sum_{(n,r)} \sum_{(m,t)} S_{r,t}(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_t}) \end{aligned} \tag{4}$$

for all n, m such that $n \geq r$ and $m \geq t$. Assume (4) holds for all $n \geq r, m \geq t$ with $n + m \leq s$ and let $n \geq r, m \geq t$ with $n + m = s + 1$. If $n = r$ or $m = t$, then this follows from the one sample theorem, as indicated at the beginning of this proof. Otherwise, since $n + m - 1 = s$, we get from the induction hypothesis, that

$$\begin{aligned} &(n(n-1) + m(m-1)) S_{n,m}(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \\ &\quad = (n-1) \sum_{i=1}^{n} S_{n-1,m}^{-i,0} + (m-1) \sum_{j=1}^{m} S_{n,m-1}^{0,-j} \\ &\quad = \frac{n-1}{\binom{n-1}{r}\binom{m}{t}} \sum_{i=1}^{n} \sum_{(n^{-i},r)} \sum_{(m,t)} S_{r,t} + \frac{m-1}{\binom{n}{r}\binom{m-1}{t}} \sum_{j=1}^{m} \sum_{(n,r)} \sum_{(m^{-j},t)} S_{r,t} \end{aligned} \tag{5}$$

where $\sum_{(n^{-i},r)}$ denotes summation over all distinct subsets $i_1, \ldots, i_r$ of the integers $1, \ldots, i-1, i+1, \ldots, n$ and similarly $\sum_{(m^{-j},t)}$ denotes summation over all distinct subsets $j_1, \ldots, j_t$ of the integers $1, \ldots, j-1, j+1, \ldots, m$.

Given $i_1 < i_2 < \ldots < i_r$ in $\{1, 2, \ldots, n\}$ there are exactly $n - r$ elements of $\{1, 2, \ldots, n\}$ which do not appear in the sequence $(i_1, i_2, \ldots, i_r)$. Thus the double summation $\sum_{i=1}^{n} \sum_{(n^{-i},r)}$ is the same as $(n-r) \sum_{(n,r)}$. Similarly, $\sum_{j=1}^{m} \sum_{(m^{-j},t)} = (m-t) \sum_{(m,t)}$. We thus get that

$$\begin{aligned} &(n(n-1) + m(m-1)) S_{n,m}(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \\ &\quad = \left(\frac{(n-1)(n-r)}{\binom{n-1}{r}\binom{m}{t}} + \frac{(m-1)(m-t)}{\binom{n}{r}\binom{m-1}{t}} \right) \sum_{(n,r)} \sum_{(m,t)} S_{r,t}. \end{aligned}$$

Finally, it is easy to see that

$$\frac{(n-1)(n-r)}{\binom{n-1}{r}\binom{m}{t}} + \frac{(m-1)(m-t)}{\binom{n}{r}\binom{m-1}{t}} = \frac{n(n-1)+m(m-1)}{\binom{n}{r}\binom{m}{t}}.$$

This proves the first part of the theorem. The converse statement can be proved in a very similar way starting for example with the last equation in (5). ■

As a simple corollary to Theorem 1, we get a condition under which a two-sample jackknife estimator is a two-sample U-statistic.

Corollary *Let $S_{n,m}$ be a symmetric statistic, and let $S_{n,m}(JACK)$ denote its jackknifed version, according to Arvesen's procedure [6],*

$$\begin{aligned} &S_{n,m}(JACK) = \\ &\quad = \frac{\sum_{i=1}^{n}(nS^0_{n,m} - (n-1)S^{-i,0}_{n-1,m}) + \sum_{j=1}^{m}(mS^0_{n,m} - (m-1)S^{0,-j}_{n,m-1})}{n+m}. \end{aligned}$$

Then, $S_{n,m}$ is a two-sample U-statistic if and only if

$$S_{n,m}(JACK) = S^0_{n,m}.$$

Proof By [6],

$$\begin{aligned} &S_{n,m}(JACK) = \\ &\quad = \frac{\sum_{i=1}^{n}(nS^0_{n,m} - (n-1)S^{-i,0}_{n-1,m}) + \sum_{j=1}^{m}(mS^0_{n,m} - (m-1)S^{0,-j}_{n,m-1})}{n+m} \\ &\quad = \frac{(n^2+m^2)S^0_{n,m}}{n+m} - \frac{(n-1)\sum_{i=1}^{n} S^{-i,0}_{n-1,m} + (m-1)\sum_{j=1}^{m} S^{0,-j}_{n,m-1}}{n+m}. \end{aligned}$$

Therefore,

$$\begin{aligned} &S_{n,m}(JACK) - S^0_{n,m} = \\ &\quad = \frac{n^2+m^2-(n+m)}{n+m}\cdot \\ &\qquad \cdot \left(S^0_{n,m} - \frac{(n-1)\sum_{i=1}^{n} S^{-i,0}_{n-1,m} + (m-1)\sum_{j=1}^{m} S^{0,-j}_{n,m-1}}{n(n-1)+m(m-1)} \right), \end{aligned}$$

and we see that the condition $S_{n,m}(JACK) = S^0_{n,m}$ is equivalent to (1). ■

There are recursive conditions on $S_{n,m}$ other than (1) that characterize when $S_{n,m}$ is a U-statistic and some of them may be easier to use. The reason we have chosen (1) is that it fits nicely with jackknifing as one sees in the proof of the corollary above.

Here is another recursive characterization of two-sample U-statistics that will be easier to apply in the example we examine in the next section. One can prove Theorem 1' directly but, at this stage, it is easier to deduce it from Theorem 1.

Theorem 1' *Let $S_{n,m}$ be a sequence of symmetric statistics, and suppose that for any $n > r$, $m \geq t$,*

$$S_{n,m}(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \frac{1}{n} \sum_{i=1}^{n} S^{-i,0}_{n-1,m} \tag{6}$$

and for any $n \geq r$, $m > t$,

$$S_{n,m}(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} S^{0,-j}_{n,m-1}. \tag{7}$$

Then, $S_{n,m}$ is a two-sample U-statistic of degree at most (r, t), i.e., $S_{n,m}$ is a U-statistic of degree (r', t') with $r \geq r', t \geq t'$. Conversely, any two-sample U-statistic satisfies equations (6) and (7).

Proof Clearly, conditions (6) and (7) imply (1), (2) and (3) and the conclusion of the first part of the theorem follows from Theorem 1. The converse statement is easy to check directly from the definition of a U-statistic. ■

3. AN APPLICATION

As an illustration of a possible use of theorem 1 or 1', we shall characterize below the U-statistics of the form $\sum_{l=1}^{n} e_{R_l}$, where $\{e_k\}_{k=1}^{n+m}$ is a sequence of real numbers (possibly depending on n and m), and R_l is the rank of $X_{(l)}$ among the $n+m$ observations $X_1, \ldots, X_n; Y_1, \ldots, Y_m$. (In [2], p. 252, this is referred to as a linear rank statistic, with two-sample regression constants). The best known U-statistic of this form is the Mann-Whitney statistic, written as a function of Wilcoxon's rank sum statistic, with $e_k = \frac{k}{nm} - \frac{n+1}{2nm}$; i.e., the statistic $MWW = \frac{1}{nm}(\sum_{i=1}^{n} R_i - \frac{n+1}{2}) = \frac{1}{nm} \sum_{i,j} I_{X_i < Y_j}$. Other important

statistics of this form are the Fisher-Yates normal score test and the one in Van der Waerden test (see [7], p. 96).

We shall show that $\sum_{l=1}^{n} e_{R_l}$ is a U-statistic, if and only if $e_k = a(\frac{k}{nm} - \frac{n+1}{2nm}) + b$ for some constants a and b. Let

$$S_{n,m} = \sum_{l=1}^{n} e_{R_l}^{n,m}.$$

Then, changing the order of summation, one gets that

$$\begin{aligned} \sum_{i=1}^{n} S_{n-1,m}^{-i,0} &= \sum_{l=1}^{n}[\sum_{i=1}^{l-1} e_{R_l-1}^{n-1,m} + \sum_{i=l+1}^{n} e_{R_l}^{n-1,m}] \\ &= \sum_{l=1}^{n}[(l-1)e_{R_l-1}^{n-1,m} + (n-l)e_{R_l}^{n-1,m}]. \end{aligned} \tag{8}$$

(Note that if $R_l - 1 = 0$ then $l - 1 = 0$.) One can also compute $\sum_{j=1}^{m} S_{n,m-1}^{0,-j}$ and show that it is equal to

$$\sum_{l=1}^{n}[(R_l - l)e_{R_l-1}^{n,m-1} + (m + l - R_l)e_{R_l}^{n,m-1}]. \tag{9}$$

(note that $R_l - l$ is the number of Y-s smaller than $X_{(l)}$ and $m + l - R_l$ is the number of Y-s larger than $X_{(l)}$.) If $e_k^{n,m} = a(\frac{k}{nm} - \frac{n+1}{2nm}) + b$ for some constants a and b independent of n and m and for all $n \geq r$, $m \geq t$ and $k \leq n + m$, then, using (8) and (9), it is quite easy to check that equations (6) and (7) are satisfied and then, by Theorem 1', $S_{n,m}$ is a U-statistic. (Of course, in this case the statistic is a simple transformation of the Mann-Whitney statistic and we can deduce the conclusion also from that.)

The deduction of the "only if" direction is more difficult. If $S_{n,m} = \sum_{l=1}^{n} e_{R_l}^{n,m}$ is a U-statistic then by Theorem 1',

$$S_{n,m} = \frac{1}{n} \sum_{i=1}^{n} S_{n-1,m}^{-i,0}$$

so

$$\sum_{l=1}^{n} e_{R_l}^{n,m} = \frac{1}{n} \sum_{l=1}^{n}[(l-1)e_{R_l-1}^{n-1,m} + (n-l)e_{R_l}^{n-1,m}]. \tag{10}$$

Let $1 \leq s \leq k-1 \leq n+m-1$. Consider any order of $X_1, \ldots, X_n; Y_1, \ldots, Y_m$ such that there are a Y, say $Y_{(p)}$, in the $(k-1)$-th place and an $X_{(s)}$ in the k-th place. Now, exchange the values of $Y_{(p)}$ and $X_{(s)}$, leaving all the rest unchanged, then $X_{(s)}$ is now in the $(k-1)$-th place. Consider equation (10) for these two arrangements and subtract the two equations to get

$$e_k^{n,m} - e_{k-1}^{n,m} = \frac{1}{n}[(s-1)(e_{k-1}^{n-1,m} - e_{k-2}^{n-1,m}) + (n-s)(e_k^{n-1,m} - e_{k-1}^{n-1,m})] \quad (11)$$

which holds for all $1 \leq s \leq k-1 \leq n+m-1$. (Note again that whenever e_{k-1} or e_{k-2} is not defined, its coefficient is zero).

Putting $s = 1$, we get for all $2 \leq k \leq n+m$

$$e_k^{n,m} - e_{k-1}^{n,m} = \frac{n-1}{n}(e_k^{n-1,m} - e_{k-1}^{n-1,m}). \quad (12)$$

Equation (11) can also be written as

$$\begin{aligned} e_k^{n,m} - e_{k-1}^{n,m} \quad = \quad & \frac{n-1}{n}(e_k^{n-1,m} - e_{k-1}^{n-1,m}) \\ & + \frac{s-1}{n}[(e_{k-1}^{n-1,m} - e_{k-2}^{n-1,m}) - (e_k^{n-1,m} - e_{k-1}^{n-1,m})]. \quad (13) \end{aligned}$$

Using this for $s = 2$ together with (12) we get that, for all $3 \leq k \leq n+m$,

$$e_k^{n-1,m} - e_{k-1}^{n-1,m} = e_{k-1}^{n-1,m} - e_{k-2}^{n-1,m}. \quad (14)$$

Iterating (14), we get that for all $k \geq 3$

$$e_k^{n-1,m} - e_{k-1}^{n-1,m} = e_2^{n-1,m} - e_1^{n-1,m}. \quad (15)$$

Plugging this back into (12), we deduce that

$$e_k^{n,m} - e_{k-1}^{n,m} = \frac{n-1}{n}(e_2^{n-1,m} - e_1^{n-1,m}) \quad (16)$$

for all $k \geq 3$. Using (12) again, this time for $k = 2$ and $s = 1$ we get that (16) holds also for $k = 2$. Now put $\alpha = \frac{n-1}{n}(e_2^{n-1,m} - e_1^{n-1,m})$ and $\beta = e_1^{n,m} - \alpha$ and sum (16) for $k = 2, \ldots, l$ to get $e_l^{n,m} = \alpha l + \beta$.

We still need to show that $\alpha = \frac{a}{nm}$ and $\beta = -a\frac{n+1}{2nm} + b$ for some constants a and b which do not depend on n and m. Consider a sample in which all

the $X - s$ are equal to some constant (say 1) and all the $Y - s$ are equal to another, larger, constant (say 2). Then $S_{n,m}$ computed on this special sample is a constant independent of n and m. (If h is the kernel then $S_{n,m}$ is equal to $h(1, \ldots, 1; 2, \ldots, 2)$ where there are r 1-s and t 2-s.) On the other hand $S_{n,m} = \sum_{i=1}^{n}(\alpha i + \beta)$. Thus

$$\alpha \frac{n(n+1)}{2} + \beta n = c$$

for some absolute constant c. Similarly, using the value 2 for all the $X - s$ and 1 for all the $Y - s$ we get that

$$\alpha \frac{n(n+2m+1)}{2} + \beta n = d$$

for some absolute constant d. Put $a = d - c$ and solve for α to get $\alpha = \frac{a}{nm}$ and we can write $e_k^{n,m} = a(\frac{k}{nm} - \frac{n+1}{2nm}) + \gamma$ for a γ which may still depend on n and m. But, denoting by $W_{n,m}$ the U-statistic of the form we consider here with $e_k^{n,m} = a(\frac{k}{nm} - \frac{n+1}{2nm})$, we get that the constant γ which is the difference of $S_{n,m}$ and $W_{n,m}$ is also a U- statistic. This of course can happen only if γ does not depend on n and m.

Acknowledgement Gideon Schechtman was patially supported by a grant from the ISF.

REFERENCES

[1] W. Hoeffding, A Class of Statistics With Asymptotically Normal Distribution, Annals of Mathematical Statistics, 19, (1948), 293-325.

[2] R.H. Randles, and D.A. Wolfe, *Introduction to the Theory of Nonparametric Statistics*, New York: John Wiley (1979).

[3] R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, New York: John Wiley (1980).

[4] A.J. Lee, *U-Statistics Theory and Practice*, Marcel Dekker, Inc., New York (1990).

[5] R.V. Lenth, Some Properties of U Statistics, The American Statistician, 37,(1983), 311-313.

[6] J.N. Arvesen, Jackknifing U-statistics, Annals of Mathematical Statistics, 40, (1969), 2076-2100.

[7] E.L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, Inc., San Francisco (1975).