A FURTHER APPROXIMATION TO THE DISTRIBUTION OF WILCOXON"S STATISTIC IN THE GENERAL CASE[1]

by

R. M. Sundrum

University of Rangoon
and Institute of Statistics, University of North Carolina

Institute of Statistics
Mimeograph Series No. 89
Limited Distribution
January 1954

---

[1]This research was supported by the United States Air Force, through the Office of Scientific Research of the Air Research and Development Command.

Bibliographical Control Sheet

1. O.A.: Institute of Statistics, North Carolina State College of the University of North Carolina

   M.A.: Office of Scientific Research of the Air Research and Development Command

2. O.A.: CIT Report No. 3
   M.A.: OSR Technical Note

3. A FURTHER APPROXIMATION TO THE DISTRIBUTION OF WILCOXON'S STATISTIC IN THE GENERAL CASE
   (UNCLASSIFIED)

4. Sundrum, R. M.

5. January, 1954

6. 13

7. None

8. AF 18(600)-458

9. RDO No. R-354-20-8

10. UNCLASSIFIED

11. None

12. In an earlier paper, the author obtained the variance of Wilcoxon's two-sample test statistic and used it to find the large sample power of Wilcoxon's test against certain alternatives. In this paper, general expressions are derived for the third and fourth moments of the statistic; these are found to depend on relatively few parameters. A short table is given of the values of these parameters under normal alternatives.

A Further Approximation to the Distribution of Wilcoxon's Statistic in the General Case

by R. M. Sundrum
University of Rangoon
and Institute of Statistics, University of North Carolina

Summary. In an earlier paper, the author obtained the variance of Wilcoxon's two-sample test statistic and used it to find the large sample power of Wilcoxon's test against certain alternatives. In this paper, general expressions are derived for the third and fourth moments of the statistic; these are found to depend on relatively few parameters. A short table is given of the values of these parameters under normal alternatives.

1. Introduction. Wilcoxon's test (Wilcoxon, 1945) is a distribution-free test of the hypothesis that two samples $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ are independent random samples from the same population. It is based on the statistic

$$U = \sum_{i=1}^{m} \sum_{k=1}^{n} d_{ik} \qquad (1)$$

where

$$d_{ik} = 1 \quad \text{if} \quad X_i < Y_k$$

$$= 0 \text{ otherwise.}$$

In the general case, we have

$$E(U) = mnp \quad \text{where } p = \Pr(X < Y) \qquad (2)$$

$$\sigma^2(U) = mn \left\{ (p-p^2) + (m-1)(q-p^2) + (n-1)(r-p^2) \right\} \quad (3)$$

where

$$q = \Pr.(X_1 < Y;\ X_2 < Y)$$

and

$$r = \Pr.(X < Y_1;\ X < Y_2)$$

(Pitman, 1948; Sundrum, 1953.) In the null case, when X and Y have the same distribution, $p = 1/2$; $q = r = 1/3$; and these formulae become

$$E(U) = \frac{1}{2} mn, \quad (4)$$

$$\sigma^2(U) = \frac{mn(m+n+1)}{12} . \quad (5)$$

As the statistic tends to be asymptotically normally distributed, a knowledge of its mean and variance is sufficient to obtain the large sample distribution. However, for medium-sized samples, where the need for distribution-free methods is greatest, it is useful to have the values of its third and fourth moments so as to get a closer approximation to the sampling distribution based, e.g. on the Edgeworth asymptotic expansion derived from the normal distribution. These moments have been obtained for the null case by Mann and Whitney (1947) and Haldane and Smith (1948):

$$\mu_3(U) = 0 \qquad (6)$$

$$U_4(U) = \frac{mn}{240}\ (m+n+1)(5m^2n + 5mn^2 + 3mn - 2m^2 - 2n^2 - 2m - 2n) \qquad (7)$$

In this paper, general expressions are derived for the third and fourth moments, giving the above results as special cases.

2. Third and Fourth Moments in the General Case.

From (1) we have

$$U^3 = \left\{ \sum_{i=1}^{m} \sum_{k=1}^{n} d_{ik} \right\}^3$$

which can be expressed as the sum of m^3n^3 terms. These can be grouped into ten types of terms, involving the following expectations:

$$p = E(d_{ik}) = E(d_{ik}^2) \quad \text{etc.}$$

$$q = E(d_{ik}d_{jk}) = E(d_{ik}^2 d_{jk}) \text{ etc.}$$

$$r = E(d_{ik}d_{i\ell}) \text{ etc.}$$

$$s = E(d_{ik}d_{jk}d_{mk})$$

$$t = E(d_{ik}d_{i\ell}d_{if})$$

$$u = E(d_{ik}d_{i\ell}d_{jk}). \qquad (8)$$

(i,j,m all different

k,ℓ,f all different.)

Term	Expectation	Number of terms. mn times
d_{ik}^3	p	1
$d_{ik}^2 d_{jk}$	q	3(m-1)
$d_{ik}^2 d_{i\ell}$	r	3(n-1)
$d_{ik}^2 d_{j\ell}$	p^2	3(m-1)(n-1)
$d_{ik}d_{jk}d_{mk}$	s	(m-1)(m-2)
$d_{ik}d_{i\ell}d_{if}$	t	(n-1)(n-2)
$d_{ik}d_{i\ell}d_{jk}$	u	6(m-1)(n-1)
$d_{ik}d_{jk}d_{mf}$	qp	3(m-1)(m-2)(n-1)
$d_{ik}d_{i\ell}d_{jf}$	rp	3(m-1)(n-1)(n-2)
$d_{ik}d_{j\ell}d_{mf}$	p^3	(m-1)(m-2)(n-1)(n-2)

Similarly,

$$U^4 = \left\{ \sum_{i=1}^{m} \sum_{k=1}^{n} d_{ik} \right\}^4$$

can be written as the sum of $m^4 n^4$ terms. These terms can be grouped into thirty-three types of terms, involving in addition to (8) the expectation terms:

$$v = E(d_{ik}d_{i\ell}d_{jk}d_{j\ell})$$

$$w = E(d_{ik}d_{i\ell}d_{jk}d_{m\ell})$$

$$x = E(d_{ik}d_{i\ell}d_{jk}d_{jf})$$

$$y = E(d_{ik}d_{i\ell}d_{jk}d_{mk})$$

$$z = E(d_{ik}d_{i\ell}d_{jk}d_{if})$$

$$a = E(d_{ik}d_{jk}d_{mk}d_{nk})$$

$$b = E(d_{ik}d_{i\ell}d_{if}d_{ig}) \qquad (9)$$

(i,j,m,n all different,

k,ℓ,f,g all different.)

The classification of these terms, as in the above table, is quite simple and is omitted here. Collecting terms together, we get

$$\mu^3(U) = (6p^3 + 6u - 6pq - 6pr)m^2n^2$$
$$+ (2p^3 + s - 3pq)m^3n + (2p^3 + t - 3pr)mn^3$$
$$+ (9pq + 6pr + 3q - 3s - 6u - 3p^2 - 6p^3)m^2n$$
$$+ (9pr + 6pq + 3r - 3t - 6u - 3p^2 - 6p^3)mn^2$$
$$+ (4p^3 + 3p^2 + p + 6u + 2s + 2t - 6pq - 6pr - 3q - 3r)mn \quad (10)$$

$$\mu_4(U) = 3(q-p^2)^2m^4n^2 + 6(q-p^2)(r-p^2)m^3n^3 + 3(r-p^2)^2m^2n^4$$
$$+ (12qp^2 + a - 4sp - 3q^2 - 6p^4)m^4n + (12rp^2 + b - 4t\,p - 3r^2 - 6p^4)mn^4$$
$$+ (42rp^2 + 72qp^2 + 6qp + 12w + 12y - 42p^4 - 18q^2 - 18qr - 12sp - 48up - 6p^3)m^3n^2$$
$$+ (42qp^2 + 72rp^2 + 6rp + 12x + 12z - 42p^4 - 18r^2 - 18qr - 12tp - 48up - 6p^3)m^2n^3$$
$$+ (36p^4 + 18q^2 + 12qr - 72qp^2 - 36rp^2 + 24sp - 6a + 48up - 12w - 12y$$
$$+ 12p^3 - 18qp + 6s)m^3n$$
$$+ (36p^4 + 18r^2 + 12qr - 72rp^2 - 36qp^2 + 24tp - 6b + 48up - 12x - 12y + 12p^3 - 18rp + 6t)mn^3$$
$$+ (105p^4 + 42p^3 + 3p^2 + 33q^2 + 33r^2 + 54qr - 174qp^2 - 174rp^2 - 42pq - 42pr + 36sp + 36tp$$
$$+ 192up - 36w - 36x - 36y - 36z + 6v + 36u)m^2n^2$$
$$+ (132qp^2 + 108rp^2 - 66p^4 - 33q^2 - 36qr - 18r^2 - 44sp - 24tp + 11a - 144up + 36w + 24x + 36y$$
$$+ 24z - 6v - 36p^3 - 36u - 7p^2 + 54pq + 36pr - 18s + 7q)m^2n$$
$$+ (132rp^2 + 108qp^2 - 66p^4 - 33r^2 - 36qr - 18q^2 - 44tp - 24sp + 11b - 144up + 24w + 36x + 24y$$

$$+36z-6v-36p^3-36u-7p^2+54pr+36pq-18t+7r)mn^2$$

$$+ (36p^4+18q^2+24qr+18r^2-72qp^2-72rp^2+24sp+24tp-6a-6b+96up-24w-24x-24y-24z$$

$$+6v+24p^3+36u+7p^2-36pq-36pr+12s+12t-7q-7r+p)mn. \quad (11)$$

If we write $n = mk$ in the above formulae, the dominant terms in m are given by

$$\sigma^2(U) \sim m^3k \left\{ (q-p^2) + k(r-p^2) \right\}$$

$$\mu_3(U) = 0\ (m^4)$$

$$\mu_4(U) \sim 3m^6k^2 \left\{ (q-p^2) + k(r-p^2) \right\}^2$$

so that, as $m \longrightarrow \infty$ (k constant)

$$\beta_1 \longrightarrow 0; \qquad \beta_2 \longrightarrow 3 \qquad (12)$$

illustrating the asymptotic tendency to normality, under the conditions

$(q-p^2) > 0$; $(r-p^2) > 0$; $\frac{m}{n}$ constant as $m, n \longrightarrow \infty$.

3. Special Cases.

The formulae (10) and (11), though they seem complicated, depend only on a few parameters. In the first paper (Sundrum, 1953) it was shown that the parameters q and r are equal when X and Y have continuous symmetrical frequency functions, differing only in their location. Under the same condition, we also have

$$s = t; \quad w = x; \quad y = z; \quad a = b \qquad (13).$$

Under this condition, the formulae involve only eight parameters.

For evaluating these parameters, it is convenient to express them in terms of the probabilities of certain ordered arrangements of a given number of X's and Y's drawn at random from specified populations. In the following, we denote by Pr. (XYXY) the probability that when two X's and two Y's are drawn at random and arranged in ascending order of magnitude, they have the arrangement indicated in brackets; and so on. Then we have

$$\begin{aligned}
p &= \text{Pr. (XY)} \\
q &= \text{Pr. (XXY)} \\
s &= \text{Pr. (XXXY)} \\
u &= \text{Pr. (XXYY)} + \tfrac{1}{4}\ \text{Pr. (XYXY)} \\
v &= \text{Pr. (XXYY)} \\
a &= \text{Pr. (XXXXY)} \\
w &= \text{Pr. (XXXYY)} + \tfrac{1}{3}\ \text{Pr. (XXYXY)} \\
y &= \text{Pr. (XXXYY)} + \tfrac{1}{3}\ \text{Pr. (XXYXY)} + \tfrac{1}{6}\ \text{Pr. (XYXXY)} \qquad (14)
\end{aligned}$$

The expressions for r, t, b, x and z may be derived from those for q, s, a, w and y respectively by interchanging X and Y and reversing the order.

(a) Null case:

These probabilities can be obtained very simply in the null case from the consideration that all the permutations of the ordered sequence of X's and Y's are equiprobable. Then

$$p = \frac{1}{2} \qquad v = \frac{1}{6}$$

$$q = r = \frac{1}{3} \qquad a = b = \frac{1}{5}$$

$$s = t = \frac{1}{4} \qquad w = x = \frac{2}{15}$$

$$u = \frac{5}{24} \qquad y = z = \frac{3}{20} \qquad (15)$$

Substituting these values in (10) and (11), we get (6) and (7), thus providing a check on the algebra.

(b) Rectangular case:

In non-null cases, these values depend on the nature of the distribution functions of X and Y. In simple cases, they may be evaluated by direct integration. If, for example, X is uniformly distributed in the range 0 to 1, and Y is uniformly distributed in the range Δ to $1 + \Delta$ $(0 \leq \Delta \leq 1)$, we have

$$p = \frac{1}{2} + \Delta - \frac{\Delta^2}{2}$$

$$q = r = \frac{1}{3} + \Delta - \frac{\Delta^3}{3}$$

$$s = t = \frac{1}{4} + \Delta - \frac{\Delta^4}{4}$$

$$u = \frac{5}{24} + \frac{5}{6}\Delta + \frac{3}{4}\Delta^2 - \frac{5}{6}\Delta^3 + \frac{1}{24}\Delta^4$$

$$v = \frac{1}{6} + \frac{2}{3}\Delta + \Delta^2 - \frac{2}{3}\Delta^3 - \frac{1}{6}\Delta^4$$

$$a = b = \frac{1}{5} + \Delta - \frac{\Delta^5}{5}$$

$$w = x = \frac{2}{15} + \frac{2}{3}\Delta + \Delta^2 - \frac{\Delta^3}{3} - \frac{2}{3}\Delta^4 + \frac{\Delta^5}{5}$$

$$y = z = \frac{3}{20} + \frac{3}{4}\Delta + \frac{5}{6}\Delta^2 - \frac{1}{2}\Delta^3 - \frac{1}{4}\Delta^4 + \frac{1}{60}\Delta^5 \qquad (16)$$

(c) Normal Case:

However, when X and Y are normally distributed, it becomes difficult to evaluate these parameters. It may therefore be useful to have the following short table of these values. Let $E(Y) - E(X) = \theta$ and $\sigma^2(X) = \sigma^2(Y) = \sigma^2$. The values of the parameters are tabulated in terms of $h = \frac{\theta}{\sigma\sqrt{2}}$.

Table I

h	p	q=r	s=t	u	v	a=b
0	.500,000	.333,333	.250,000	.208,333	.166,667	.200,000
0.1	.539,828	.374,078	.287,723	.245,560	.199,750	.234,464
0.2	.579,260	.416,228	.327,876	.285,913	.236,363	.271,920
0.3	.617,911	.459,311	.370,057	.328,903	.276,243	.312,064
0.4	.655,422	.502,822	.413,791	.374,172	.319,002	.354,499
0.5	.691,463	.546,245	.458,548	.421,061	.364,141	.398.743
0.6	.725,747	.589,064	.503,761	.468,835	.411,065	.444,245

Table I (Continued)

h	w=x	y=z
0	.133,333	.150,000
0.1	.164,655	.182,975
0.2	.200,370	.219,830
0.3	.238,532	.258,792
0.4	.281,079	.301,779
0.5	.326,777	.347,537
0.6	.374,903	.396,223

In the former paper, it is shown how p and q = r may be obtained from published tables. Some of the values for the computation of the other parameters are based on certain elaborate tables computed by D. Teichroew on the electronic computing machine (SWAC) of the National Bureau of Standards, and kindly made available to me by Mr. I. Richard Savage. These consist of the probabilities of those arrangements in which a certain number of Y's lie to the right of a certain number of X's. The other values were computed by the Gaussian method of numerical quadrature, using nine ordinates evaluated at the zeros of the ninth order Hermite polynomial. A table of the zeros and weight factors of the first twenty Hermite polynomials is given by Salzer, Zucker and Capuano (1952.)

(d) Estimates from the Sample:

Where the form of the parent distribution functions is not known, we may use estimates of these parameters from the sample itself. The

method of obtaining unbiased estimates is illustrated by a few examples below. Let U_i be the number of Y's in the sample greater than $X_{(i)}$, where $X_{(i)}$ is the i-th ordered value of the X's amongst themselves. Then, using primes for estimators,

$$P'(XY) = \frac{1}{mn} \sum_{i=1}^{m} U_i$$

$$P'(XXXXY) = \frac{4}{mn(m-1)(m-2)(m-3)} \sum_{i=1}^{m} i(i-1)(i-2)(i-3)\, U_i$$

$$P'(XYYYY) = \frac{1}{mn(n-1)(n-2)(n-3)} \sum_{i=1}^{m} U_i(U_i-1)(U_i-2)(U_i-3)$$

$$P'(XYXY) = \frac{4}{mn(m-1)(n-1)} \sum_{\substack{i \\ i<j}} \sum_{j} (U_i-U_j)\, U_j$$

$$P'(XYYXY) = \frac{6}{mn(m-1)(n-1)(n-2)} \sum_{\substack{i \\ i<j}} \sum_{j} (U_i-U_j)(U_i-U_j-1)\, U_j \qquad (17)$$

etc. This however is too tedious a procedure for common use.

Haldane, J. B. S. and Smith, C. A. B. (1948), Ann. of Eugenics, 14, 117.

Mann, H. B. and Whitney, D. R. (1947), Ann. of Math. Stats. 18, 50.

Pitman, E. J. G., (1948), Lectures on Non-parametric Inference (mimeographed) University of North Carolina.

Salzer, H.E., Zucker, Ruth, and Capuano, Ruth (1952), Journal of Research of National Bureau of Standards, 48, 111.

Sundrum, R. M., (1953), J.R. Statist. Soc., B, 15...

Wilcoxon, F., (1945), Biometrics, 1, 80.