A Bivariate Permutation Test for Analysis of Three Interval Data Samples

Taichi KARIYA

Emeritus Professor, Kawasaki Medical School,
Kurashiki 701-01, Japan

Accepted for publication on March 6, 1992

ABSTRACT. Let X, Y, and Z be three random variables with unknown continuous cumulative distribution functions F, G, and H. D.R. Whitney[1] proposed the (U,V) statistics as a bivariate extension of Wilcoxon's U statistic in 1951. It is well known that Whitney's (U,V) tests based on these statistics are particularly powerful against the following two types of alternatives, respectively :

(1) $F(t) > G(t)$ and $F(t) > H(t)$ (for all t),
(2) $G(t) > F(t) > H(t)$ (for all t).

A bivariate permutation (U,V) test that is an extension of Whitney's (U,V) test is proposed for samples in which observations are specified only by intervals with known probability density functions. Here, the two statistics, U and V, are based on generalized signs instead of ranks. The proposed bivariate permutation (U,V) test is not rough even in small samples, because the value of the generalized sign can take on real value densely. In the same way, we can construct a multivariate permutation test for many samples in which observations are specified only by intervals. Computer programs were developed for determining the critical region of (U,V) in a given sample of n_x x's, n_y y's and n_z z's.

Key words : bivariate permutation test — interval data — generalized sign

The statistical problem considered in this paper arises in biomedical studies comparing several treatments, in which the observation for each individual is specified only by an interval with a known probability density function or by a membership function. This mainly occurs because there are many biomedical fuzzy variates ; for example, a blood pressure, the size of a tumor, the incubation period and the period from a surgical operation to a relapse of the disease.

We assume that $(X_1, X_2, \ldots, X_{nx})$, $(Y_1, Y_2, \ldots, Y_{ny})$, $(Z_1, Z_2, \ldots, Z_{nz})$ are random samples from populations having the unknown continuous cumulative distribution functions F(x), G(y), and H(z) respectively. In spite of the assumption of continuity of the distribution function, sometimes tied observa-

c.d.f.	sample size	interval data
F(x)	n_x	$(x_{1L}, x_{1U}), (x_{2L}, x_{2U}), \ldots, (x_{nxL}, x_{nxU})$
G(y)	n_y	$(y_{1L}, y_{1U}), (y_{2L}, y_{2U}), \ldots, (y_{nyL}, y_{nyU})$
H(z)	n_z	$(z_{1L}, z_{1U}), (z_{2L}, z_{2U}), \ldots, (z_{nzL}, z_{nzU})$

仮谷太一

tions will appear in practice, so they should not be excluded from the test.

We wish to test the hypothesis Ho: $F=G=H$ against the alternative that

A1: $F(t)>G(t)$ and $F(t)>H(t)$ (for all t), or

A2: $G(t)>F(t)>H(t)$ (for all t).

The test is conditional on the given overlapping pattern of observations. As noted in our previous papers,[2-4] we define the generalized sign of X_i-Y_j to be U_{ij} based on their interval data, (x_{iL},x_{iU}) and (y_{jL},y_{jU}), and the generalized sign of X_i-Z_k to be V_{ik} based on their interval data, (x_{iL},x_{iU}) and (z_{kL},z_{kU}), in the following manner:

$$U_{ij}=E(\mathrm{sgn}(X_i-Y_j))=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{sgn}(x-y)s_i(x;\ x_{iL},\ x_{iU})t_j(y;\ y_{jL},\ y_{jU})dxdy,$$

$$V_{ik}=E(\mathrm{sgn}(X_i-Z_k))=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{sgn}(x-z)s_i(x;\ x_{iL},\ x_{iU})w_k(z;\ z_{kL},\ z_{kU})dxdz,$$

where $s_i(x;\ x_{iL},\ x_{iU})$, $t_j(y;\ y_{jL},\ y_{jU})$, and $w_k(z;\ z_{kL},\ z_{kU})$ are known probability density functions of X_i, Y_j, and Z_k, respectively, and are not related to F, G, and H at all, and

$$\mathrm{sgn}(u-v)=\begin{cases}1 & \text{if } u>v,\\ 0 & \text{if } u=v,\\ -1 & \text{if } u<v.\end{cases}$$

In the practical integration, we assume that

$$s_i(x;x_{iL},x_{iU})=0 \quad \text{if } x<x_{iL} \text{ or } x>x_{iU}, \text{ and } \int_{x_{iL}}^{x_{iU}} s_i(x;x_{iL},x_{iU})dx=1,$$

$$t_j(y;y_{jL},y_{jU})=0 \quad \text{if } y<y_{jL} \text{ or } y>y_{jU}, \text{ and } \int_{y_{jL}}^{y_{jU}} t_j(y;y_{jL},y_{jU})dy=1,$$

and $w_k(z;z_{kL},z_{kU})=0$ if $z<z_{kL}$ or $z>z_{kU}$, and $\int_{z_{kL}}^{z_{kU}} w_k(z;z_{kL},z_{kU})dz=1.$

Therefore,

$$U_{ij}=\begin{cases}-1 & \text{if } x_{iU}\leqq y_{jL},\\ +1 & \text{if } y_{jU}\leqq x_{iL},\\ \text{a real number which is larger than } -1 & \\ \qquad \text{and less than } +1 & \text{otherwise,}\end{cases}$$

and

$$V_{ik}=\begin{cases}-1 & \text{if } x_{iU}\leqq z_{kL},\\ +1 & \text{if } z_{kU}\leqq x_{iL},\\ \text{a real number which is larger than } -1 & \\ \qquad \text{and less than } +1 & \text{otherwise.}\end{cases}$$

Then $|U_{ij}|$ may be interpreted as the probability that X_i is larger than Y_j if $U_{ij}>0$, and as the probability that Y_j is larger than X_i if $U_{ij}<0$. The same may be said of V_{ik}.

Now we calculate the statistic $U=\sum_i\sum_j U_{ij}$, where the sum is over all $n_x\cdot n_y$ comparisons of two samples, **X** and **Y**, and the statistic $V=\sum_i\sum_k V_{ik}$, where the sum is over all $n_x\cdot n_z$ comparisons of two samples, **X** and **Z**. If there is no overlapping in the interval data, it is easy to show that $U=2\cdot WU-n_x\cdot n_y$ and $V=2\cdot WV-n_x\cdot n_z$ where WU and WV are Whitney's (or Wilcoxon's) U and V, that is more concretely to say, the number of times a y precedes an x and the number of times a z precedes an x in ascending order of the sample values. Hence the proposed test is equivalent to Whitney's test if there is no overlapping in the interval data.

As a critical region for the hypothesis Ho: F=G=H against the alternative A_1: F(t)>G(t), F(t)>H(t) we propose to use $U \leqq K_1$, $V \leqq K_2$; or against the alternative A_2: G(t)>F(t)>H(t), $U \geqq K_3$, $V \leqq K_4$, where the constants K_i are chosen to give the correct significance level. Even if the significance level is fixed, the constants K_i are not uniquely determined. A reasonable principle to follow in this case would be to choose K_i so that

$$P(U \leqq K_1) \doteqdot P(V \leqq K_2) \quad \text{or} \quad P(U \geqq K_3) \doteqdot P(V \leqq K_4)$$

according to which alternative is chosen.

Since the test statistics U and V are conditional on the given pattern of observations and are not always integers, a convenient recurrence relation does not hold, as is the case with Whitney's (U,V) statistics regarding the number of sequences of n_x x's, n_y y's and n_z z's in which a y precedes an x U times and a z precedes an x V times in their ascending order. We must compute, therefore, the probability value of the given data samples each time.

Moments of joint distribution of U and V

Now let's consider the exact joint distribution of (U,V) under the null hypothesis Ho. In spite of the assumption of continuity of the distribution function, tied observations will appear in practice. It is true that these tied observations are not exactly equal but the differences are very small. Therefore, they should be handled separately. In accordance with the way of thinking, we deal with any observation by its datum number. Thus, as the n individual observations are labeled differently, there are $n!/(n_x! \cdot n_y! \cdot n_z!)$ possible allocations of the n observations to three samples with n_x, n_y, and n_z observations, respectively, where $n = n_x + n_y + n_z$. Under the null hypothesis Ho, these $n!/(n_x! \cdot n_y! \cdot n_z!)$ possible allocations occur with the same probability and so we can derive the conditional exact distribution of (U,V) by all the values of (U, V) calculated for each of these allocations. Now we determine a critical region for the hypothesis Ho so that $U \leqq K_1$, $V \leqq K_2$ against the alternative A_1: F>G, F>H; or $U \geqq K_3$, $V \leqq K_4$ against the alternative A_2: G>F>H, where the constants K_i are chosen to give the correct significance level. In this case we have to choose

$$P(U \leqq K_1) \doteqdot P(V \leqq K_2) \quad \text{or} \quad P(U \geqq K_3) \doteqdot P(V \leqq K_4)$$

according to which alternative is chosen. This is because even though the value of $P(U \leqq Uo, V \leqq Vo)$ is very small, both the values of $P(U \leqq Uo)$ and $P(V \leqq Vo)$ are not always small, where Uo and Vo are the the value of U and V for the observed data samples. And even if the value of $P(U \geqq Uo, V \leqq Vo)$ is very small, both the values of $P(U \geqq Uo)$ and $P(V \leqq Vo)$ are not always small.

When there is overlapping in interval data, we must calculate the means, the variances of U, V and the correlation coefficient between U and V each time. The conditional means and variances of U, V under Ho are denoted by $\boldsymbol{E}(U|P, Ho)$, $\boldsymbol{E}(V|P, Ho)$, and $\boldsymbol{V}(U|P, Ho)$, $\boldsymbol{V}(V|P, Ho)$, respectively, where P is the overlapping pattern of the observed interval data. The expectations are obtained by summing over all the $n!/(n_x! \cdot n_y! \cdot n_z!)$ equally likely samples leading to the same observed pattern P.

It is easy to see

$$\boldsymbol{E}(U|P, Ho)=0 \quad \text{and} \quad \boldsymbol{E}(V|P, Ho)=0$$

by the symmetry of the allocation.

Then the variances and the correlation coefficient between U and V under

Ho are given as follows:

$$\boldsymbol{V}(U|P,\ Ho)=\boldsymbol{E}(U|P,\ Ho)^2=\sum_{l=1}^{m} U_{(l)}/m,$$

$$\boldsymbol{V}(V|P,\ Ho)=\boldsymbol{E}(V|P,\ Ho)^2=\sum_{l=1}^{m} V_{(l)}/m,$$

and

$$\rho=\frac{\mathrm{Cov}\ (U,\ V)}{\sqrt{\boldsymbol{V}(U|P,Ho)\cdot \boldsymbol{V}(V|P,Ho)}}=\sum_{l=1}^{m} U_{(l)}\cdot V_{(l)}/m/\boldsymbol{V}(U|P,Ho)/\boldsymbol{V}(V|P,Ho),$$

where $U_{(l)}$ ($V_{(l)}$) is the value of U (V) for the l-th allocation sample, $m=n!/(n_x!\cdot n_y!\cdot n_z!)$ and Cov(U,V) is the covariance between U and V.

SOME NUMERICAL STUDIES

We carried out the following numerical studies to confirm the relationship between our proposed test and Whitney's (U,V) test and to observe how the depth of overlapping among interval data affects the constants K_i.

The i-th interval datum is indicated by its midpoint and one half of its length as follows: $(t_i \pm d_i)$. For example, (35 ± 2) implies the interval datum $(35-2,\ 35+2)$. Generally, a $(t_i\pm d_i)$ implies an interval datum $(t_i-d_i,\ t_i+d_i)$.

And $\{t_1,\ t_2, \ldots,\ t_n\}$ was composed of $n=n_x+n_y+n_z$ random numbers sampled from a normal population with mean 50 and variance 10^2.

When $d_i=0$ or when d_i $(i=1,\ 2, \ldots,\ n)$ are relatively small and so $(t_i-d_i,\ t_i+d_i)$ $(i=1,\ 2, \ldots,\ n)$ do not overlap with one another, our proposed test will coincide with Whitney's (U,V) test.

We are then interested in the hypothesis F=G=H under the alternative G>F>H. In order not to make things unduly complicated, we will consider that $d_i=d$ $(i=1,\ 2, \ldots,\ n)$ and assume that s(x), t(y), and w(z) are all distributed uniformly.

(1) Study 1

Twelve random numbers sampled from a normal population $N(50,\ 10^2)$ were divided into 6 x's, 3 y's, and 3 z's, so that the sequence of x, y, and z in magnitude would coincide with the sequence yyxxxyxzxzzx in the Whitney's example, as shown in Table 1.

TABLE 1. Numerical data sample 1

sample size	interval data
$n_x=6$	$(35\pm d)$, $(44\pm d)$, $(49\pm d)$, $(51\pm d)$, $(56\pm d)$, $(65\pm d)$
$n_y=3$	$(27\pm d)$, $(32\pm d)$, $(50\pm d)$
$n_z=3$	$(52\pm d)$, $(58\pm d)$, $(60\pm d)$

where d takes the value 0, 0.5, 1.0, 1.5, 2.0, 2.5, and 5.0 in turn.

The results of our proposed exact test based on numerical data of sample 1 are shown in Table 2, where the significance level α is held at almost 0.05.

(2) Study 2

Now thirteen random numbers sampled from a normal population $N(50,\ 10^2)$ were divided into 6 x's, 4 y's, and 3 z's, so that the sequence of x, y, and z in magnitude would coincide with the sequence yyxyxzxyxxzzx, as shown in Table 3.

The results of our proposed exact test based on the numerical data of

TABLE 2. The results of the test based on the numerical data of sample 1.

No.	d	ρ	K_3	K_4	α	$P(U\geqq K_3)$	$P(V\leqq K_4)$	Uo	Vo	β
0	0	.3	6	−6	.0438	.274	.274	12	−10	.0044
1	0.5	.3	6	−6	.0438	.274	.274	12	−10	.0044
2	1.0	.302	5	−5	.0438	.274	.274	12	−9.75	.0037
3	1.5	.304	4.3	−4.3	.0485	.287	.287	12	−9.44	.0031
4	2.0	.306	4.3	−4.3	.0499	.290	.290	11.94	−9.13	.0031
5	2.5	.307	4.4	−4.4	.0498	.291	.290	11.84	−8.84	.0032
6	5.0	.314	4.4	−4.4	.0488	.286	.286	11.47	−8.25	.0036

where $\alpha=P(U\geqq K_3, V\leqq K_4)$ and $\beta=P(U\geqq Uo, V\leqq Vo)$, and ρ is the correlation coefficient between U and V.

TABLE 3. Numerical data sample 2

sample size	interval data
$n_x=6$	(35±d), (44±d), (49±d), (51±d), (56±d), (65±d)
$n_y=4$	(27±d), (32±d), (42±d), (50±d)
$n_z=3$	(48±d), (58±d), (60±d)

where d takes the value 0, 0.5, 1.0, 1.5, 2.0, 2.5, and 5.0 in turn.

TABLE 4. The results of the test based on the numerical data of sample 2.

No.	d	ρ	K_3	K_4	α	$P(U\geqq K_3)$	$P(V\leqq K_4)$	Uo	Vo	β
0	0	.330	6	−6	.0468	.305	.274	16	−6	.0063
1	0.5	.330	6	−6	.0468	.305	.274	16	−6	.0063
2	1.0	.332	5.5	−5.0	.0468	.305	.274	16	−6.25	.0035
3	1.5	.334	5.4	−4.3	.0481	.301	.287	15.89	−6.33	.0034
4	2.0	.336	5.2	−4.3	.0478	.296	.291	15.69	−6.38	.0035
5	2.5	.337	5.2	−4.3	.0499	.297	.299	15.48	−6.36	.0035
6	5.0	.345	5.1	−4.2	.0491	.297	.298	14.82	−6.21	.0035

where $\alpha=P(U\geqq K_3, V\leqq K_4)$ and $\beta=P(U\geqq Uo, V\leqq Vo)$, and ρ is the correlation coefficient between U and V.

sample 2 are shown in Table 4.

THE RESULTS OF SOME NUMERICAL STUDIES

From Tables 2 and 4 the following conclusions can be made. It is clear that if the value of d is zero or relatively small and the interval data do not overlap with one another, the proposed bivariate permutation (U,V) test perfectly coincides with Whitney's (U,V) test, in accordance with our expectation. As the value of d increases gradually and overlapping of interval data becomes greater, the critical region widens little by little and soon becomes stable.

When the value of d is equal to or greater than 1.5, the critical region is almost stable. Then the length of the interval data is equal to or greater than 0.3 times of the standard deviation of F, G, or H where its value is 10. For example, when the value of d is 2.0; that is, the length of the interval data is equal to 0.4 times of the standard deviation of F, G, or H, the critical region of (U,V) is $U\geqq 4.3$ and $V\leqq -4.3$ under the significance level $\alpha\doteqdot 0.05$, from Table 2. On the other hand, the critical region of (U,V) on the Whitney's test is $U \geqq 6$ and $V \leqq -6$ under about same significance level.

Therefore, the proposed bivariate permutation (U,V) test based on interval data samples is more powerful than Whitney's test based on medians samples, when the length of the interval data is equal to or greater than 0.3 times of the standard deviation of F, G, or H. About the same results can also be obtained from Table 4.

Moreover, similar results were also obtained in the case where s(x), t(y), and w(z) were all distributed normally in each interval.

Consequently, when an observation is fuzzy and not specified by an exact value, we should specify it by an interval datum and not by its mean or median based on its several observed values for each sample individual. The proposed bivariate permutation (U,V) test should be applied to interval data samples with enthusiasm.

REFERENCES

1) Whitney, D.R.: A bivariate extension of the U statistic. A.M.S. **22**: 274–282, 1951
2) Kariya, T.: Estimated T test and F test with interval-censored normal data. Kawasaki Med. J. **10**: 197–206, 1984
3) Kariya, T.: Analogous t and F test statistics based on grouped data. Proceedings of the Pacific Statistical Congress-1985. Elsevier Science Publishers B.V. (North-Holland): 275–279, 1986
4) Kariya, T.: A generalized sign test based on paired interval data. The Second Japan-China Symposium on Statistics: 125–128, 1986