KAWASAKI MED. J. VOL. 14 : NO. 4, 187-192 (1988)

A Generalized Wilcoxon Test for Comparing Interval Data Samples

Taichi KARIYA

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

Accepted for publication on November 28, 1988

ABSTRACT. A distribution-free two-sample test is proposed that is an extension of the Wilcoxon test to interval-censored data or more generally to interval data samples. Interval-censored data are often observed in medical or biological studies and the idea of interval data is extension of interval-censored data. We define a generalized sign of difference between two observations based on their interval data under the estimated distribution of each observation. The generalized sign may be interpreted as the probability that the one is larger than the other. The test statistic is defined as the sum of the generalized signs based on all combinations of the two samples. The test is conditional on the pattern of observations.

The null hypothesis is

$H_0 : F(t) = G(t)$ against either
$H_1 : F(t) < G(t)$ or
$H_2 : F(t) < G(t)$ or $F(t) > G(t)$,

where F, G are cumulative distribution functions of the observations. The test is shown to be asymptotically normal. Working examples are presented and the tests are performed by a BASIC program which was developed for the proposed test.

Key words : Wilcoxon test — interval (censored) data — generalized sign

The statistical problem considered in this paper arises in biometrical studies comparing two treatments, where the observation for each individual is specified only by an interval. A continuous random variable X is said to be interval-censored into a non-zero interval I if the only information about a realized value of X is that the realized value lies in I. The interval-censoring of a realized value is a very common procedure in biometrics and interval-censored data are often obtained in the multistage follow-up examination about

a girl's sexual maturity age,
an eruptive age of a specified deciduous tooth,
a period of time from a surgical operation to relapse of the disease,
or a life span of a machine part and so on.

However, the situation where specifying a realized value by an interval is appropriate is not limited to interval censoring. In the case of some characteristics of a living individual, their values change every moment and especially are influenced greatly by not only physical causes but mental ones in human being. For example, maximum blood pressure of any person varies every moment;

仮谷太一

therefore observations of the same person will be different considerably every minute. Moreover, the value of some characteristics of an individual is measured by the sample material extracted from it. Another sample material from the same one will give a different observation. Consequently, it is preferable that the observation in these cases should be specified by a confidence interval.

We have engaged in statistical research in order to grant citizenship to interval data."Estimated t test and F test" is a parametric approach to interval data,[1] and "Analogous t and F test statistics based on grouped data" is an article based on grouped data which are less complicated than interval data mathematically.[2] "A generalized sign test based on paired interval data" is our first one of non-parametric approach.[3,4] And this paper is the second one of non-parametric approach. Edmond A. Gehan (1965) generalized an extension of the Wilcoxon test to samples with arbitrary censoring on the right but did not compare two interval-censored data.[5] Other some persons have considered two-sample tests with censoring, but all did not compare two interval-censored data.

The W statistic on two interval data samples

We assume that $(X_1,X_2,\ldots,X_{nx})$ and $(Y_1,Y_2,\ldots,Y_{ny})$ are samples of size n_x,n_y respectively, from populations having continuous cumulative distribution functions F(x),G(y) respectively, and that X_i and Y_j are specified by interval data (x_{iL},x_{iU}) and (y_{jL},y_{jU}) respectively, that is to say, the realized value of X_i is an element of (x_{iL},x_{iU}) and the realized value of Y_j an element of (y_{jL},y_{jU}). Any properties except continuity are not assumed on the distribution functions of variables X_i and Y_j.

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	$(y_{1L}, y_{1U}), (y_{2L}, y_{2U}), \ldots, (y_{nyL}, y_{nyU})$

The null hypothesis is H_0: $F(t)=G(t)$.
The alternative hypotheses (H_a) are either

H_1: $F(t)<G(t)$

or the two-sided version

H_2: $F(t)<G(t)$ or $F(t)>G(t)$.

We define a generalized sign U_{ij} of X_i-Y_j based on their interval data (x_{iL},x_{iU}) and (y_{jL},y_{jU}) as follows:

$$U_{ij}=E(\mathrm{sgn}(X_i-Y_j))$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{sgn}(x-y)f_i(x;x_{iL},x_{iU})g_j(y;y_{jL},y_{jU})dxdy,$$

where $f_i(x;x_{iL},x_{iU})$, $g_j(y;y_{jL},y_{jU})$ are the probability density functions of X_i,Y_j respectively, and

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

In the practical integral calculation, we assume that

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

and that

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

Then

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

Then U_{ij} may be interpreted as the probability that if $U_{ij}>0$, X_i is larger than Y_j and if $U_{ij}<0$, Y_j is larger than X_i.

Now we calculate the statistic $W=\sum_i \sum_j U_{ij}$, where the sum is over all $n_x \cdot n_y$ comparisons. The statistic W corresponds to the Wilcoxon (1945) statistic T'.[6]

It is easy to show that $W=n_y(n+1)-2T'$, if n interval data do not overlap with one another, where $n=n_x+n_y$ and T' is the sum of the ranks of the second sample in the ascending ordered combined sample.

The conditional exact distribution and the conditional mean and variance of W

We consider the conditional exact distribution of W under the null hypothesis H_0. Since the n individuals are labeled differently, there are $n!/(n_x!n_y!)$ possible allocations of the individuals to two samples with n_x, n_y observations respectively, where $n=n_x+n_y$. This is independent of whether there are tied observations in the n combined observations or not. Under null hypothesis H_0, these $n!/(n_x!n_y!)$ possible allocations occur with same probability and so we can construct the conditional exact distribution of W by all values of W which are calculated for these allocations. Then the significance probability (also called P-value) of the observed two samples is as follows:

P-value= (the number of allocations which W values are not less than W_0)/n,

where W_0 is the value of W for the observed two samples. By the P-value, we can decide to reject or accept the null hypothesis H_0 against the alternative hypothesis H_1. The test procedure against the alternative hypothesis H_2 is almost the same as the one against H_1.

Now let the conditional mean and variance of W be denoted by $E(W \mid P, H_0)$ and $V(W \mid P, H_0)$ respectively, where P is the pattern of observed interval data. The expectations are over all the $n!/(n_x!n_y!)$ equally likely samples leading to the same observed pattern P.

It is easy to see

$$E(W \mid P,H_0)=0 \qquad \cdots\cdots(2)$$

by symmetry of the allocation. Then the variance is given as follows:

$$V(W \mid P,H_0)= E(W-E(W \mid P,H_0))^2=E(W^2)= \sum_{s=1}^{\binom{n}{nx}} W_{(s)}{}^2 / \binom{n}{nx}$$

$$=\binom{n-2}{nx-1} \sum_{i=1}^{n} U_{i\cdot}{}^2 / \binom{n}{nx}$$

$$=n_x n_y \Big(\sum_{i=1}^{n} U_{i\cdot}{}^2 \Big) / \{n(n-1)\}, \qquad \cdots\cdots(3)$$

where $W_{(s)}$ is the value of W for the s-th sample allocation

and $U_{i\cdot}= \sum_{j=1}^{n} (j \neq i) U_{ij} = \sum_{j=1}^{n} U_{ij}$ $(\because U_{ii}=0)$.

Asymptotic normality and the calculation of W in large sample

It is easy to show that

$$U_{ij}= - U_{ji},\ -1 \leqq U_{ij} \leqq +1 \text{ and } U_{ii}=0 \qquad \cdots\cdots(4)$$

by the definition of U_{ij}.

Now we calculate generalized signs U_{ij} $(i \neq j)$, $i,j=1,2,\ldots,n$ for all combinations of two interval data from the n combined observed interval data. From these n(n-1) generalized signs, $n_x \cdot n_y$ signs corresponding to each allocation are selected, and summed up into the W. Accordingly the total of all W's contains $n_x \cdot n_y (n!/(n_x! n_y!))$ signs. The allocations are symmetric and so generalized signs are equally likely summed up into W's. Therefore, each sign equally appears in W's $n_x \cdot n_y (n!/(n_x! n_y!))/\{n(n-1)\}$ times.

Then if we set a W against an elementary error in measurement, we can easily arrive at asymptotic normality of W in the same way as Gauss' law of errors. On the basis of the property, the test is expected to be consistent against one-sided alternatives $F(t)< G(t)$ and against two-sided alternatives where either $F(t)< G(t)$ or $F(t)> G(t)$. According to the results of our simulation study, the normal approximation is very good if both sample sizes are reasonably large (say both sizes seven or more). Of course, we have to calculate exactly when both sample sizes are small.

The calculation of W statistic and its variance $V(W \mid P,H_0)$ are simple when n_x,n_y are small. However if n_x,n_y are large, then both the W and its variance calculation are lengthy even by a computer. When n_x,n_y are considerably large, we can calculate the asymptotic variance of W by the formula (3). Since the mean of W is zero by the symmetry of W distribution, a value of

$$Z= W/\sqrt{V(W \mid P,H_0)} \qquad \cdots\cdots(5)$$

is taken as asymptotically normal with zero mean and unit variance. Consequently, to test H_0 against either H_1 or H_2, we can use the value of Z.

Working examples

In this section, we apply the W test to samples on the maximum blood pressures of young men and women. The measurement of blood pressure of each person was carried out five times successively by an automatic sphygmomanometer. The subjects are classmates of about twenty years. The class is composed of thirty male students and twenty-two female students.

We are interested in the difference of sex in maximum blood pressures. Each interval datum and median are derived from his/her five measured values. We calculated U_{ij} under the assumption that X_i and Y_j distribute uniformly in (x_{iL},x_{iU}) and (y_{jL},y_{jU}) respectively.

The first sex-wise sample data are as follows:

male : $(115,124)_{124}$, $(106,114)_{111}$, $(108,112)_{110}$, $(117,128)_{127}$
$(100,119)_{117}$, $(118,134)_{123}$, $(124,130)_{128}$, $(116,131)_{126}$

female : $(105,112)_{107}$, $(110,113)_{110}$, $(110,116)_{112}$, $(\ 90,108)_{98}$
$(\ 99,108)_{101}$, $(112,127)_{123}$, $(108,142)_{113}$, $(103,106)_{104}$

These are random samples of size eight from male and female data respectively. Interval data are shown by number pairs in parentheses and medians by suffices.

The second sex-wise sample data are as follows:

male : $(111,118)_{115}$, $(107,110)_{109}$, $(117,130)_{120}$, $(114,120)_{118}$
$(121,130)_{123}$, $(116,122)_{121}$, $(106,114)_{111}$, $(\ 99,108)_{107}$

female : $(\ 92,101)_{96}$, $(\ 84,\ 95)_{91}$, $(\ 98,102)_{99}$, $(\ 99,109)_{101}$
$(110,113)_{110}$, $(\ 90,101)_{98}$, $(104,118)_{108}$, $(105,112)_{107}$

These are random samples of size eight from male and female data which are measured at about same hour in the afternoon.

When we applied the W test to these samples, the P-values under the null hypothesis H_0 were as follows:

TABLE 1. P-values on interval data and medians

sample no.	sample size	P-values on interval data		P-values on medians	
		exact	approximate	exact	approximate
1	8, 8	0.0340*	0.0334*	0.0050**	0.0058**
2	8, 8	0.0033**	0.0047**	0.0006***	0.0014***
all	30, 22		0.0003***		0.0001***

* : 5% significant, ** : 1% significant, *** : 0.1% significant

The overlapping pattern of the interval data of the first sample and the distribution of W based on every sample allocation under the null hypothesis are shown in the following Fig. 1.

From these P-values we can reject the null hypothesis H_0 that there are no difference between male and female in maximum blood pressures. In the examples we could get interval data and medians of all persons respectively since the subjects were healthy young men and women, but in practical cases we can not get such reliable data as these. The P-value on the interval data are considerably different from the P-value on the medians. Generally speaking, P-values on interval data of random samples from a population are more stable

than P-value on medians of the same samples. The P-value on ordinary observations will be more irregular. The figures in Table 1 will give information about the difference between both P-values.

The second sample was more homogeneous than the first one and then the P-value of the second one was smaller than the P-value of the first one.

The exact P-value was nearly equal to the approximate P-value because the both sample sizes were eight. However, when sample sizes are less than five, both P-values will be reasonably different. Especially, samples from clinical trials are small in size and then exact calculation of P-value is in great request.

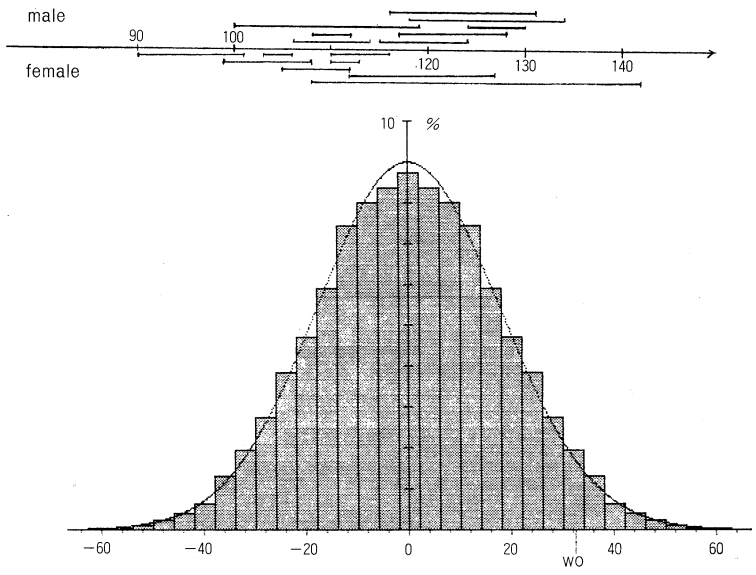

Fig. 1. Overlapping pattern and W's distribution the first sample

REFERENCES

1) Kariya, T.: Estimated t test and F test with interval-censored normal data. Kawasaki Med. J. **10** : 197–206, 1984
2) Kariya, T.: Analogous t and F test statistics based on grouped data. Proceedings of the Pacific Statistical Congress-1985. North-Holland, Elsevier Science Publishers B.V. 1986, pp. 275–279
3) Kariya, T.: A generalized sign test based on paired interval data. The second Japan–China symposium on statistics : 125–128, 1986
4) Kariya, T.: A generalized sign test for paired interval data samples. Jpn. J. of Applied Statistics **16** : 77–88, 1987 (in Japanese)
5) Gehan, E.A.: A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika **52** : 203–223, 1965
6) Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1** : 80–83, 1945