On a Nonparametic Two-Sample Test for Scale

by Ryôji TAMURA

(*Received Nov. 5, 1957*)

田 村 亮 二：ひろがりに関するノンパラメトリックな
二標本検定について

§ 1. Introduction

Lehmann [I] has proposed to use the following test statistic $W_{m,n}$ for testing the hypothesis H that the continuous distributions $F(x)$ and $G(x)$ are equal, or even the wider hypothesis H' that F and G differ only in location, against the alternative hypothesis that the random variable Y with the *c.d.f.* $G(y)$ is more spread out than the random variable X with the *c.d.f.* $F(x)$.

Let

$$X_1, X_2, \ldots\ldots, X_m; \quad Y_1, Y_2, \ldots\ldots, Y_n \tag{1}$$

be the two independent random samples with the continuous *c.d.f.* $F(x)$ and $G(y)$, respectivly and $W_{m,n}$ be the propotion of quadruple $X_i, X_j; Y_k, Y_l$ for which $|Y_l - Y_k| > |X_j - X_i|$ holds. Where $1 \leqq i < j \leqq m, 1 \leqq k < l \leqq n$. The null hypothesis is rejected when the sample value of $W_{m,n}$ is too large. Then it is shown that this test is unbiased against the alternative for which $F(x_1)=G(y_1)$, $F(x_2)=G(y_2)$ implies $|x_1-x_2|<|y_1-y_2|$, and consistent against the alternative that $P(|Y-Y|>|X'-X|)>\frac{1}{2}$ holds where X, X'; Y, Y' are independently distributed with the *c.d.f.* $F(x)$ and $G(y)$ respectively. Thus this Lehmann's test has the desirable properties—unbiasedness and consistency—. However $W_{m,n}$ is not completely distribution free, that is, its distribution depends upon the form of the intial distribution even under the null hypothesis. Its variance is not independent of $F(x)$, though its mean is the form $E(W_{m,n}|F=G)=\frac{1}{2}$.

From the above-mentioned we cannot ulitilise the Lehmann's statistic $W_{m,n}$ for the standpoint of the practical applications. Now we are interested in the another statistic for testing the null hypothesis H that $F=G$ against the alternative H_1 that the Y's are more spread out than the X's. We denote the event by $(Y<X, X<Y')$ that the two X's lie between the two Y's when two X's and two Y's are drawn at random from (1). And we define $Q_{m,n}$ as follows:

$$Q_{m,n}=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\{\text{the number of } (X, X'; Y, Y') \text{ for which } (Y<X, X'<Y')\} \tag{2}$$

Then our test procedure is to reject H if $Q_{m,n}$ is too large. In this paper we investigate some statistical properties of the test statistic $Q_{m,n}$ and the asymptotic efficiency of the test by $Q_{m,n}$ will be calculated in the parametric normal case.

§ 2. Some properties of $Q_{m,n}$.

In this section the statistical properties of $Q_{m,n}$—the expectation, the variance, and the asymptotic distribution—are discussed.

(a) The expressions of $Q_{m,n}$.

$Q_{m,n}$ can be expressed in a form of the ranks of one of two samples. Let $r_1 < r_2 < \ldots\ldots < r_m$ be the ranks of the m X's among the combined sample of size $m+n$. Then it may be seen that the following identity holds.

$$Q_{m,n} = \binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum_{i<j}(r_i - i)(n - r_i + j) \tag{3}$$

From this it follows by easy computation that

$$Q_{m,n} = \frac{m}{2}\binom{m}{2}^{-1}\binom{n}{2}^{-1}\left\{s_r^2 - (m-1)\left(r - \frac{m+n+1}{2}\right)^2 2(n+1)C + \text{const}\right\} \tag{3'}$$

Ignoring constant additive and multicative terms in (3′), we obtain

$$Q'_{m,n} = s_r^2 - (m-1)\left(r - \frac{m+n+1}{2}\right)^2 - 2(n+1)C \tag{3''}$$

where $$s_r^2 = \frac{1}{m}\sum_{i=1}^{m}(r_i - r)^2, \qquad r = \frac{1}{m}\sum_{i=1}^{m} r_i$$

$$C = \frac{1}{m}\sum_{i=1}^{m}(r_i - r)\left(i - \frac{m+1}{2}\right)$$

As has been given in (3″), $Q_{m,n}$ depends only on the ranks of the X's. Therefore we can consider the test by $Q_{m,n}$ as what is called the rank test. From the another point of view, $Q_{m,n}$ can be also expressed in the following form that is convenient for the use of the theory of Lehmann [1] and Hoeffiding [7]. Suppose

$$D(x_i, x_j; y_k, y_l) = \begin{cases} 1 & \text{for } y_k < x_i,\ y_j < y_l \\ 0 & \text{other wise} \end{cases} \tag{4}$$

where $i<j$, $k<l$.

Then $Q_{m,n}$ may be rewritten as follows

$$Q_{m,n} = \binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum_{i<j}\sum_{k<l} D(x_i, x_j; y_k, y_l) \tag{5}$$

consisting of $\binom{m}{2}\binom{n}{2}$ terms.

The expectation and the variance of $Q_{m,n}$ are obtained from the expression (5) and also the asymptotic distribution of $Q_{m,n}$ is shown by the theory of Lehmann.

(b) The expectation and variance.

We denote the expectation and variance of the statistic T by $E(T)$ and $V(T)$, respectively, and by $E(T|S)$ and $V(T|S)$ in the case that they are consider under the condition S. Then we get

$$E(Q_{m,d})=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum_{i<j}\sum_{k<l}E\{D(X_i, X_i;\ Y_k, Y_l)\}=P(Y<X, X'<Y)\ \ (=\theta)$$

Though some combinationary calculation leads to the evaluation of θ under the null hypothesis, yet we express θ in a integral form by the distribution as it is more important to reseach the behaviour of $E(Q_{m,n})$ under the alternative. Let be

$$\max(Y, Y')=Z, \qquad \min(Y, Y')=Z'$$

then

$$P(Y<X, X'<Y')=2\int\cdots\int_{\substack{x,x'<z\\ x,x'>z'}} dF(x)\, dF(x')\, dG(z)\, dG(z')$$

$$=2\iint_{z'<z}\{F(z)-F(z')\}\, dG(z')\, dG(z) \tag{6}$$

If $F=G$, we get $\theta=\frac{1}{6}$ from (6) after some computation. That is

$$E(Q_{m,n}|F=G)=2\int_{-\infty}^{\infty}\int_{-\infty}^{z}\{F(z)-F(z')\}^2\, dF(z')\, dF(z)=\frac{1}{6} \tag{7}$$

The behaviour of $E(Q_{m,n})$ under the alternative will be discussed in the section.

In order to calculate $V(Q_{m,n})$, we may accept the same method as Sundrum [4]. Then $Q^2_{m,n}$

$$Q^2_{m,n}=\binom{m}{2}^{-2}\binom{n}{2}^{-2}\sum_{i<j}\sum_{k<l}D(X_i, X_j;\ Y_k, Y_l)\}^2$$

is consisted of $\binom{m}{2}^2\binom{n}{2}^2$ terms and it's expectation can be grouped in the following nine classes of terms, involving the expectation terms shown for each class.

term	number of term	expectation
$D(i,j,k,l)\ D(i,j,k,l)$	1	θ
$D(i,j,k,l)\ D(i,m,k,l)$	$2(m-2)$	r
$D(i,j,k,l)\ D(i,j,k,f)$	$2(n-2)$	s
$D(i,j,k,l)\ D(m,n,k,l)$	$\frac{1}{2}(m-2)(n-3)$	t
$D(i,j,k,l)\ D(i,j,f,g)$	$\frac{1}{2}(m-3)(n-2)$	u
$D(i,j,k,l)\ D(i,m,k,f)$	$4(m-2)(n-2)$	v
$D(i,j,k,l)\ D(m,n,k,f)$	$(m-2)(m-3)(n-2)$	a
$D(i,j,k,l)\ D(i,m,f,g)$	$(m-2)(n-2)(n-3)$	b
$D(i,j,k,l)\ D(m,n,f,g)$	$\frac{1}{4}(m-2)(m-3)(n-3)(n-2)$	θ^2

where $D(i,j,k,l)$ stands for $D(X_i, X_j, Y_k, Y_l)$, etc.

Collecting terms together and simplyfing, we get

$$V(Q_{m,n})=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\Big\{(a-\theta^2)m^2n+(b-\theta^2)mn^2+(4v+6\theta^2-5a-5b)mn$$
$$+\Big(\frac{1}{2}t+\frac{2}{3}\theta^2-2a\Big)m^2+\Big(\frac{1}{2}u+\frac{3}{2}\theta^2-2b\Big)n^2+\Big(2r-\frac{5}{2}t+10a+6b-8v-\frac{15}{2}\theta^2\Big)m$$
$$+\Big(2s-\frac{5}{2}u+6a-8v-\frac{15}{2}\theta^2\Big)n+(\theta+3t+3u+16v+9\theta^2-4r-4s-12a-12b)\Big\} \quad (8)$$

For evaluating the parameters occurring in the above expression, it is convenient to express them in terms of the probabilities of a certain ordered arrangement of a gived number of X's and Y's drawn at random from the respective population. In the following, we express, for example, the event by $(xyxy)$ that when two X's and Y's are drawn at random from (1) and arranged in order of magnitude, they have the indicated arrangement. Then we can get

$$\theta=P(yxxy)$$
$$r=P(yxxxy)$$
$$s=\frac{1}{3}\{P(yxxyy)+P(yyxxy)\}$$
$$t=P(yxxxxy)$$
$$u=\frac{2}{3}P(yyxxyy)$$
$$v=\frac{1}{3}\{P(yxxxyy)+P(yyxxxy)\}+\frac{1}{9}\{P(yxxyxy)+P(yxyxxy)\}$$
$$a=\frac{1}{3}\{P(yxxxxyy)+P(yyxxxxy)\}+\frac{1}{6}\{P(yxxyxxy)+P(yxxxyxy)+P(yxyxxxy)\}$$
$$b=\frac{2}{3}P(yyxxxyy)+\frac{2}{9}\{P(yyxxyxy)+P(yxyxxyy)\}+\frac{1}{18}P(yxyxyxy)$$

In the null case these probabilities can evaluated easily as follows.

$$r=\frac{1}{10},\qquad s=t=\frac{1}{15},\qquad u=x=\frac{2}{45},\qquad a=b=\frac{1}{30}$$

Substituting these values into (8), we find the variance of $Q_{m,n}$ for the null case.

$$V(Q_{m,n}\mid F=G)=\frac{(m+n+1)(2mn+3m-n-2)}{90mn(m-1)(n-1)} \quad (9)$$

(c) The asymptotic distribution of $Q_{m,n}$.

Considering the expression of $Q_{m,n}$ in the form (5), we notice the following two facts: (i) $D(x_i, x_j;\ y_k, y_l)$ is a real valuted symmetric function of (x_i, x_j) and (y_k, y_l) (ii) $E(D)=\theta$, $V(D)<\infty$. Thus $Q_{m,n}$ is U-statistic that have been studied by Hoeffiding [7] and Lehmann [1]. The asympotic distribution of the extended U-statistic is researched by Lehmann [1], Weger [5], and Fraser [6] and it is shown that they are asymptotically normally distributed. Stating by the form of Weger in our case, we get the following theorem.

THEOREM.

Let $X_1, X_2, \ldots\ldots, X_m$ and $Y_1, Y_2, \ldots\ldots, Y_n$ be the independently distributed random variables from the distributions $F(x)$ and $G(y)$ respectively. A function $D(x_i, x_j; y_k, y_l)$ is symmetric in the x's alone and in the y's alone as defined already. Futher

$$E\{D(X_i, X_j;\ Y_k, Y_l)\}=\theta$$
$$V\{D(X_i, Y_j;\ Y_k, X_l)\}<\infty$$
$$Q_{m,n}=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum D(X_i, X_j;\ Y_k, Y_l)$$

where the summation is extended over all subscripts $1\leqq i<j\leqq m$, $1\leqq k<l\leqq n$. Then as $n\to\infty$ where $m/n=c$, $\left(\frac{mn}{m+n}\right)^{\frac{1}{2}}(Q_{m,n}-\theta)$ is asymptotically normally distributed.

§ 3. The symmetrical case.

In § 2 we have investigated the general expression of $E(Q_{m,n})$ and $V(Q_{m,n})$ and futher under the null hypothesis obtained their numerical value. In this section it is our purpose to get the more detailed properties of $Q_{m,n}$. Now we assume that the *c. d. f.* F and G are symmetric and have the same median zero without the loss of generality. Then the expression of θ is as follows from (6).

$$\theta=2\iint_{z'<z}\{F(z)-F(z')\}^2 dG(z)\, dG(z')$$

by nothing the identity

$$F(x)-F(z')=(F(z)-G(z))+(G(z)-G(z'))+(G(z')-F(z'))$$

and (7), we get

$$\theta=\frac{1}{6}+2\int_{-\infty}^{\infty}\{F(z)-G(z)\}^2 dG(z)$$
$$+4\iint_{z'<z}\{F(z)-G(z)\}\{G(z)-G(z')\}dG(z')dG(z)$$
$$+4\iint_{z'<z}\{F(z)-G(z)\}\{G(z')-F(z')\}dG(z')dG(z)$$
$$+4\iint_{z'<z}\{G(z)-G(z')\}\{G(z')-F(z')\}dG(z')dG(z)$$

Considering the third and the fifth menbers in the right side of the above identity, we have

$$2\int_{-\infty}^{\infty}(F-G)G^2dG+2\int_{-\infty}^{\infty}(G-F)dG+2\int_{-\infty}^{\infty}(G-F)G^2dG-4\int_{-\infty}^{\infty}G(G-F)dG$$
$$=4\int_{-\infty}^{\infty}(G-F)(1-2G)dG$$

Next let

$$D_1=\{(z',z)\,|\,-z\leqq z'\leqq z,\ 0\leqq z<\infty\},\qquad D_2=\{(z,z)\,|\,z'\leqq z\leqq z',\ -\infty<z'\leq 0\}$$

and

$$R(t)=\int_{-\infty}^{\infty}(G-F)dG$$

, then

$$\text{the fourth menber }=4\iint_{D_1}+4\iint_{D_2}$$

$$=4\int_0^{\infty}\{F(z)-G(z)\}\,\{R(z)-R(-z)\}\,dG(z)+4\int_{-\infty}^{0}\{G(z)-F(z)\}\,\{R(z)-(-z)\}\,dG(z)$$

However, F and G are symmetric and have median 0 from the assumption, so that the value of the above integral vanishes. Therefor we have

$$\theta=\frac{1}{6}+2\int_{-\infty}^{\infty}(F-G)^2dG+4\int_{-\infty}^{\infty}(G-F)\,(1-2G)dG$$

THEOREM.

Assume that the continuous c. d. f. F and G are symmetric and have the same median, then we have

$$\theta=\frac{1}{6}+2\int_{-\infty}^{\infty}\{F(z)-G(z)\}+dG(z)+4\int_{-\infty}^{\infty}\{G(z)-F(z)\}\{1-2G(z)\}\,dG$$

and under the alternative that Y is more spread out than X—this means that $G(z)\geqq F(z)$ for all $z<0$ and $G(z)\leqq F(z)$ for all $z>0$, strictly inequality for some interval of z—, we have

$$\theta>\frac{1}{6}$$

As the first part of this theorem has been proved already, we now prove the second part. Under the alternative we have from the assumption that

$$G(z)-F(z)\geqq 0,\qquad 1-2G(z)\geqq 0\quad\text{for}\quad 0\leqq z<\infty$$

, so that we can get

$$\int_0^{\infty}(G-F)\,(1-2G)\,dG>0$$

Similarly we have

$$\int_{-\infty}^{0}(G-F)\,(1-2G)\,dG>0\quad\text{for}\quad -\infty<z\leq 0$$

Combining the both results we get $\theta>\frac{1}{6}$.

It follows the next theorem by Lehmann [1] that our test is consistent.

THEOREM (Lehmann).

If $t_n(x_1,\dots,x_n)$ is a test statistic for testing $\theta=\theta_0$ against $\theta>\theta_0$ for each n, and

$$E\{t_n(X_1, \dots, X_n)\}=\theta$$
$$V\{t_n(X_1, \dots, X_n)\}\to 0 \qquad (\text{as } n\to\infty)$$

, then the test with $t_n(X_1, \dots, X_n)>\theta_0+C_n$ as the critical region is consistent for the alternative $\theta>\theta_0$.

§ 4. Asymptotic efficiency.

In general, let T_n (and T_n^*) be a one-sided test statistic for testing $\theta=\theta_0$ which is a function of n sample X from the c. d. f. $F(X; \theta)$. Let the mean of T_n (and T_n^*) be $\mu_n(\theta)$ ($\mu_n^*(\theta)$) and the variance $\sigma_n(\theta)$ ($\sigma_n^*(\theta)$). Moreover suppose that T_n (and T_n^*) is asympotically normally distributed. Then the asymptotic efficiency of T_n against T_n^* is difined by Mood [3] to be

$$\lim_{n\to\infty}\left\{\frac{1}{\sigma_n(\theta_0)}\left(\frac{d\mu_n(\theta)}{d\theta}\right)_{\theta_0}\Big/\frac{1}{\sigma_n^*(\theta_0)}\left(\frac{d\mu_n^*(\theta)}{d\theta}\right)_{\theta_0}\right\} \tag{10}$$

Now we compute the efficiency of the test by $Q_{m,n}$ relative to the standard F test by using

$$f(x)=\frac{1}{\sqrt{2\pi}}exp\left\{-\frac{x^2}{2}\right\}, \qquad g(x)=\frac{1}{\sigma}f\left(\frac{x}{\sigma}\right) \qquad \sigma>1$$

On differentiating $E(Q_{m,n})$ with regard to σ and introducing $\sigma=1$, we obtain

$$\left(\frac{dE(Q)}{d\sigma}\right)_{\sigma=1}=2\iint_{z'<z}\{F(z)-F(z')\}^2(z'^2-1)\,dF(z')\,dF(z)$$
$$+2\iint_{z'<z}\{F(z)-F(z')\}(z^2-1)\,dF(z')\,dF(z)$$
$$=\frac{2}{3}\int_{-\infty}^{\infty}(z^2-2)\{(1-F(z)^3+F(z)^3\}\,dF(z)=\frac{1}{2\sqrt{3\pi}}$$

On the other hand,

$$V(Q_{m,n}|\sigma=1)=\frac{m+n}{45mn}+0\left(\frac{1}{m^2}\right)$$

For the F test we have $\mu^*(\sigma)=\frac{1}{\sigma^2}$ and the variance under the null hypothesis is $\frac{2(m+n)}{mn}$. Substituting these results into (10), we find the asympotic efficiency to be

$$\lim\left\{\sqrt{\frac{45mn}{m+n}}\frac{1}{2\sqrt{\pi}}\Big/\sqrt{\frac{2mn}{m+n}}\right\}=\frac{1}{\pi}\sqrt{\frac{15}{2}}$$

Therefore it is shown that our test has the asymptotic efficiency 87% against F test in the normal case.

§ 5. Acknowledgment.

The rough results of this problem have been obtained in May, 1957. On the other hand the auther has seen in the Ann. Math. Stat. Vol. 28 No. 1 (1957) which he has

recieved in June that the problem of the same category has been discussed on the next titled paper by B. V. Sukhatme [8] "On certain two-sample nonparametric tests for variance."

He proposed in [8] the following test statistic

$$T=\frac{1}{mn}\sum_{i}^{m}\sum_{j}^{n}\psi(X_i, X_j)$$

where $$\psi(X, Y)=\begin{cases} 1 & \text{if either } 0<X<Y \text{ or } Y<X<0 \\ 0 & \text{otherwise} \end{cases}$$

To use this statistic for testing variance is similar to use Mann-Whitney's statistic with regard to the nonparametric two-sample test for location, where Mann-Whitney's statistic is the proportion of the number of (X_i, X_j) with $X_i<Y_j$. It may safely be said that our statistic is the extension of B. V. Sukhatme's T-statistic for testing scale parameter as well as Lehmann's statistic is the extension of Mann-Whitney's statistic for testing the location parameter, where Lehmann's statistic is the praportion of the number of $(X_i, X_j; Y_k, Y_l)$ for which both X lie the same side of both Y.

The auther wishes to express his thanks to Mr. A. Kudō of Kyushu University and Mr. M. Ikeda of Wakayama University for giving some remarks for this paper.

Reference

〔1〕 E. L. Lehmann: Consistency and unbiasedness of certain nonparametric tests. Ann. Math. Stat. Vol. 22 (1951)

〔2〕 H. B. Mann & D. R. Whitney: On a test of whether one of two random variables is stohastically larger than the other. Ann. Maht. Stat. Vol. 18 (1947)

〔3〕 A. M. Mood: On a osymptotic efficiency of certain nonparametric two sample tests. Ann. Math. Stat. Vol. 25 (1954)

〔4〕 R. M. Sundrum: On Lehmann's two sample test. Ann. Math. Stat. Vol. 25 (1954)

〔5〕 L. H. Weger: Properties of some two sample tests based on a particular measure of discrepancy. Ann. Math. Stat. Vol. 27 (1956)

〔6〕 D. A. D. Fraser: NONPARAMETRIC METHODS IN STATISTICS. J. Wiley & Son (1957)

〔7〕 W. Hoeffding: A class of statistics with asymptotically normal distribution. Ann. Math. Stal. Vol. 19 (1947)

〔8〕 B. V. Sukhatme: On certain two-sample nonparametric tests for variance. Ann. Math. Stal. Vol. 27 (1957)