九州大学学術情報リポジトリ
Kyushu University Institutional Repository

On the Nonparametric Tests based on Certain U-Statistics

Tamura, Ryoji
Shimane University

https://doi.org/10.5109/12995

出版情報：統計数理研究. 9 (2/3), pp.61-67, 1960-10. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

ON THE NONPARAMETRIC TESTS BASED ON CERTAIN U-STATISTICS

By

Ryoji TAMURA

(*Received*, May 20, 1959
Revised, January 30, 1960)

§ 1. **Introduction** The object of this paper is to discuss the two sample tests for scale based on some of the certain *generalized U-statistics* which have been investigated by Hoeffding [1], Lehmann [2], Sukhatme [3], [4], [5] and Fraser [6].

Let now $X_1, X_2, \cdots, X_m$ and $Y_1, Y_2, \cdots, Y_n$ be two independent random samples from populations with comulative distribution functions $F(x-\xi)$ and $G(y-\eta)=F[(y-\eta)/\delta]$ respectively, ξ and η being the population medians of X and Y, and δ a scale parameter. Moreover assume that $F(x)$ and $G(y)$ be absolutely continuous. The problem considered in this paper is to test the hypothesis that two populations have the same scale parameter against the alternative that the Y's are more spread out than the X's and vice versa, or in symbols

$$H:\ \delta=1,\quad A:\ \delta\neq 1\quad (\text{or } A':\ \delta>1).$$

This is narrower than the alternative $F(x)\neq G(x)$. Let ξ and η be known, say $\xi=\eta=0$, without loss of generality, so that the distribution functions of X and Y differ only in the scale parameter. In this case several nonparametric tests have been proposed, particularly by Mood [7], Sukhatme [3], [5], and Tamura [8]. These tests are based on some of the statistics which are generally called as *the generalized U-statistics*. On the other hand, Sukhatme has proposed to use the modified statistics which are obtained by substituting their estimates into *the U-statistics* instead of ξ and η, in case where the informations about ξ and η are unknown. We shall also propose the new statistic $\hat{Q}_N$ as the test criterion which is essentially *the modified generalized U-statistic*, and investigate some of properties of $\hat{Q}_N$.

§ 2. **The test based on the statistic Q_N —*Q test*—.** When the location parameters ξ and η are known, *say* $\xi=\eta=0$, the auther [8] has proposed the statistic Q_N as the test criterion

$$Q_N=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum_{\alpha,\beta}\varphi(X_{\alpha_1}, X_{\alpha_2}, Y_{\beta_1}, Y_{\beta_2}), \tag{1}$$

where

$$\varphi(u_1, u_2, v_1, v_2) = \begin{cases} 1 & \text{for } v_1 < u_1 u_2 < v_2 \text{ or } v_2 < u_1 u_2 < v_1 \\ 0 & \text{otherwise} \end{cases},$$

the notation $v_1 < u_1 u_2 < v_2$ *means that two* u_1, u_2 *lie between* v_1 *and* v_2 ($v_1 < v_2$), *and the summation runs over all subscripts* α, β *such that* $1 \leqq \alpha_1 < \alpha_2 \leqq m$, $1 \leqq \beta_1 < \beta_2 \leqq n$. Let ρ_1, ρ_2 be fixed non-negative numbers such that $m = N\rho_1$, $n = N\rho_2$ and $\rho_1 + \rho_2 = 1$. Then it is easily shown that Q_N is *the so-called generalized U-statistic* investigated by Lehmann and other authers. The mean value $\theta = EQ_N$ and the variance $\sigma^2 = \operatorname{var} Q_N$ of Q_N may be found after some computations as follows,

$$\theta = P(Y_1 < X_1 X_2 < Y_2 \text{ or } Y_2 < X_1 X_2 < Y_1) \tag{2}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{F(y) - F(y')\}^2 dG(y) dG(y')$$

$$\sigma^2 = \binom{m}{2}^{-1} \binom{n}{2}^{-1} \Big\{ (a-\theta^2) mn^2 + (b-\theta^2) mn^2 + (4v + 6\theta - 5a - 5b) mn + (t/2 + 2\theta^2/3 - 2a) m^2 + (u/2 + 2\theta^2/3 - 2b) n^2 + (2r - 5t/2 + 10a + 6b - 8v - 15\theta^2/2) m + (2s - 5u/2 + 10b6a - 8v - 15\theta^2/2) n + (\theta + 3t + 3u + 16v + 9\theta^2 - 4r - 4s - 12a - 12b) \Big\}, \tag{3}$$

where setting $F_i = F(y_i)$, $G_i = G(y_i)$,

$$r = 2 \int_{-\infty}^{\infty} \int_{y_2}^{\infty} (F_1 - F_2)^3 \, dG_1 dG_2$$

$$s = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} (F_1 - F_2)^2 G_2 dG_2 dG_1$$

$$t = 2 \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} (F_2 - F_1)^4 dG_1 dG_2$$

$$u = 16 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} (F_1 - F_2)^2 G_2 (1 - G_1) \, dG_1 \, dG_2$$

$$v = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} (F_1 - F_2)^3 G_2 dG_2 dG_1 + 4 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} (F_1 - F_2)^2 (F_2 - F_3) dG_3 dG_2 dG_1$$

$$a = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} (F_1 - F_2)^4 G_2 dG_2 dG_1 + 6 \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \int_{y_2}^{\infty} (F_1 - F_2)^2 (F_2 - F_3)^2 dG_1 dG_3 dG_2$$

$$+ 8 \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \int_{y_2}^{\infty} (F_1 - F_2)^3 (F_2 - F_3) dG_1 dG_3 dG_2$$

$$b = 16 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} (F_1 - F_2)^3 G_2 (1 - G_1) \, dG_2 dG_1 + 32 \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} (F_1 - F_2)(F_2 - F_3)^2 G_3 dG_3$$

$$dG_2 dG_1$$

$$+2\int_{-\infty}^{\infty}\int_{-\infty}^{y_1}\int_{-\infty}^{y_2}\int_{-\infty}^{y_3}(F_1-F_2)(F_2-F_3)(F_3-F_4)dG_4dG_3dG_2dG_1.$$

Under the null-hypothesis $F=G$, we can evaluate θ and σ^2 as follows

(4) $\theta_0=E(Q_N\,|\,H)=1/6$

(5) $\sigma_0^2=\mathrm{var}\,(Q_N H)=(m+n)/45mn+o(1/N).$

Futhermore, if we set

(6)
$$\psi_1(x_1)=E\varphi(x_1, X_2, Y_1, Y_2)-\theta$$
$$\psi_2(y_1)=E\varphi(X_1, X_2, y_1, Y_2)-\theta,$$

where x_1, y_1 are arbitrary fixed numbers and the expectations is taken with respect to the random variable X_1, X_2, Y_1, Y_2, then the asymptotic distribution of $\sqrt{N}(Q_N-EQ_N)$ is the nondegenerate normal distribution since

(7)
$$E\psi_1^2(X_1)+E\psi_2^2(Y_1)>0.$$

In a matter of fact,

$$\psi_2(y_1)=E\varphi(X_1, X_2, y_1, Y_2)-\theta$$
$$=\int_{-\infty}^{\infty}\{F(y)-F(y_1)\}^2dG(y)-\theta.$$

Setting $\int_{-\infty}^{\infty}\{F(y)-F(y_1)\}^2dG(y)=I(y_1)$, then $E\psi_2^2(Y_1)=0$ implies that

$$I(Y_1)=EI(Y_1).$$

However this identity cannot hold with probability one. This fact shows that the variance σ^2 of Q_N is of order N^{-1}.

It will now be shown that the test based on Q_N —*Q test*— is consistent, that is, the probability tends to one as N tends to infinity that *the Q_N test* will reject the hypothesis H tested when the given alternative is true. In fact, the power function for the alternative A is indicated as follows

(8)
$$P(|Q_N-\theta_0|>\lambda_\alpha\sigma_0\,|\,A),$$

where λ_α is a constant that is defined by

$$(2\pi)^{-\frac{1}{2}}\int_{-\lambda\alpha}^{\lambda\alpha}\exp(-x^2/2)dx=1-\alpha.$$

Then the above probability (8) is transformed as follows

(9)
$$1-\frac{1}{\sqrt{2\pi}\,\sigma}\int_{\theta_0-\lambda_\alpha\sigma_0}^{\theta_0+\lambda_\alpha\sigma_0}\exp\{-(y-\theta)^2/2\sigma^2\}dy=1-\frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2}\exp(-x^2/2)dx,$$

where

(10) $$t_1=(\theta_0-\lambda_\alpha\sigma_0-\theta)/\sigma,\ t_2=(\theta_0+\lambda_\alpha\sigma_0-\theta)/\sigma.$$

However σ_0 and σ is of order $N^{-\frac{1}{2}}$ so that the above power tends to one as $N\to\infty$.

To close this section, we calculate the asymptotic efficiency of this test *in the Mood's sense* for $G(x)=F(x/6)$. Then from (2) we can directly get

(11) $$\left.\frac{\partial EQ_N}{\partial\sigma}\right|_{\sigma=1}=4\int_{-\infty}^{\infty}xf(x)^2F(x)\,dx-2\int_{-\infty}^{\infty}xf(x)^2\,dx,$$

therefore the efficacy of *the Q test* is equal to

(12) $$180mn(m+n)^{-1}\{2\int_{-\infty}^{\infty}xf^2F\,dx-\int_{-\infty}^{\infty}xf^2\,dx\}^2.$$

Also, the efficacy of *the variance ratio F-test* is equal to

$$4mn(m+n)^{-1}(\beta_2-1).$$

Hence the asymptotic relative efficiency of our *Q test* with respect to the *variance ratio F-test* is given by

(13) $$e=45(\beta_2-1)\{2\int xf^2Fdx-\int xf^2dx\}^2,$$

where

$$\beta_2=\int_{-\infty}^{\infty}\{x-E(X)\}^4\,dF/[\int_{-\infty}^{\infty}\{x-E(X)\}^2dF]^2.$$

We can obtain after some computations the following table about the asymptotic relative efficiency with respect to *F-test* for some alternatives.

Table

density function	asymptotic efficiency
1. normal $f(x)=\frac{1}{\sqrt{2\pi}}\exp(-x^2/2),-\infty<x<\infty$	$15/2\pi^2\doteqdot 0.76$
2. uniform $f(x)=\frac{1}{2},\ -1\leqq x\leqq 1$	1
3. exponential $f(x)=\frac{1}{2}\exp(-\lvert x\rvert),\ -\infty<x<\infty$	1.08

On the other hand, the asymptotic relative efficiency of Sukhatme's *T test* with respect to *F-test* is about 0.61 for the normal alternative and about 0.94 for the exponential, futhermore that of Sukhatme's *S test* is equal to 0.69 for the normal and 0.80 for the uniform. Hence it follows that our *Q test* is more efficient than not only Sukhatme's *T* and *S tests*, but *the variance ratio F-test* for some non-normal alternatives.

§ 3. **The test based on $\hat{Q}^N$.** The test in the above section presupposes the knowledge about the location of the two populations which are not always available. Then our concern is with the case that there are no know-

ledge about ξ and η. In the latter case, we estimate ξ and η by the sample medians $\tilde{X}$ and $\tilde{Y}$, and use the deviations of the observations from the sample medians rather than the observations themselves. And we define the following new statistic $\hat{Q}_N$ which is known as *the modified generalized U-statistic* in general.

$$\hat{Q}_N=\binom{m}{2}^{-1}\binom{n}{2}^{-1}\sum_{\alpha\ \beta}\varphi\,(X_{\alpha_1}-\tilde{X}, X_{\alpha_2}-\tilde{X}, Y_{\beta_1}-\tilde{Y}, Y_{\beta_2}-\tilde{Y}), \tag{14}$$

where the summation runs over all subscripts α, β *such that* $1\leqq\alpha_1<\alpha_2\leqq m$, $1\leqq\beta_1<\beta_2\leqq n$ *and the function* φ *means the same as* (1). In this section we shall now prove that $\hat{Q}_N$ has the same asymptotic normal distribution as the statistic Q_N, that is, in symbol

$$\lim_{N\to\infty} L\sqrt{N}(\hat{Q}_N-EQ_N)=\lim_{N\to\infty} L\sqrt{N}(Q_N-EQ_N)=N(0,\sigma_1^2) \tag{15}$$

under the some assumptions about F and G, where $\lim_{N\to\infty} L\ T_N$ means the limiting distribution of T and σ_1^2 is the asymptotic variance of $\sqrt{N}\,Q_N$. In order to prove the fact that $\hat{Q}_N$ has the same asymptotic distribution as the statistic Q_N, we must check up the regular conditions that have been given by Sukhatme. It is well known that $\sqrt{m}(\tilde{X}-\xi)$ and $\sqrt{n}(\tilde{Y}-\eta)$ are bounded in probability and each has a limiting distribution. Setting $A(t_1, t_2)$ as follows

$$A(t_1, t_2)=E\varphi(X_1-t_1, X_2-t_1, Y_1-t_2, Y_2-t_2|\xi=\eta=0), \tag{16}$$

where the exectation being taken with respect to all the X's and Y's, then it is necessary to show the following identities

$$\left.\frac{\partial A(t_1, t_2)}{\partial t_1}\right|_{t_1=t_2=0}=\left.\frac{\partial A(t_1, t_2)}{\partial t_2}\right|_{t_1=t_2=0}=0. \tag{17}$$

Since $A(t_1, t_2)$ is equal to the next probability from (16)

$$P(Y_1-t_2<X_1-t_1\ \ X_2-t_1<Y_2-t_2\ \text{ or }\ Y_2-t_2<X_1-t_1\ \ X_2-t_1<Y_1-t_2),$$

we can transform $A(t_1, t_2)$ to the following integral representation

$$A(t_1, t_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{F(y'+t_1-t_2)-F(y+t_1-t_2)\}^2dG(y)dG(y'). \tag{18}$$

Hence it follows by differentiating (18) that

$$\left.\frac{\partial A(t_1, t_2)}{\partial t_1}\right|_{t_1=t_2=0}=4\{\int_{-\infty}^{\infty}F(x)f(x)dG(x)-\int_{-\infty}^{\infty}F(x)dG(x)\int_{-\infty}^{\infty}f(x)dG(x)\}.$$

Assume that $f(x)$ and $g(x)$ are symmetric about the origin, then we can prove after some computations that the value of the first partial derivative

to t_1 vanishes in $t_1=t_2=0$. Similarly we can get the identity

$$\left.\frac{\partial A(t_1, t_2)}{\partial t_2}\right|_{t_1=t_2=0}=0.$$

Next set

(19) $W(x_1, x_2, y_1, y_2, t_1, t_2)=\varphi(x_1-t_1, x_2-t_1, y_1-t_2, y_2-t_2)-A(t_1, t_2)$,

then it must be proved that the following inequality follows

(20) $E\,|W(X_1, X_2, Y_1, Y_2, t_1, 0)-W(X_1, X_2, Y_1, Y_2, 0, 0)|\leqq M_1 t_1$
$E\,|W(X_1, X_2, Y_1, Y_2, 0, t_2)-W(X_1, X_2, Y_1, Y_2, 0, 0)|\leqq M_2 t_2$,

where M_1 and M_2 are certain constants. In fact, for $t_1>0$, we get

$$E\,|W(X_1, X_2, Y_1, Y_2, t_1, 0)-W(X_1, X_2, Y_1, Y_2, 0, 0)|$$
$$\leqq E\,|\varphi(X_1-t_1, X_2-t_1, Y_1, Y_2)-\varphi(X_1, X_2, Y_1, Y_2)+A(t_1, 0)-A(0, 0)|.$$

Hence it follows from (18)

(21) $$A(t_1, 0)-A(0, 0)\leqq 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{|F(y'+t_1)-F(y')|+|F(y+t_1)-F(y)|\}dG(y')dG(y)$$
$$\leqq 4at_1,$$

since $|F(y+t)-F(y)|\leqq at$ if the distribution function $F(x)$ has a derivative $f(x)$ bounded in absolute value by a. On the other hand, the expected value $E\,|\varphi(X_1-t_1, X_2-t_1, Y_1, Y_2)-\varphi(X_1, X_2, Y_1, Y_2)|$ is equal to the probability

(22) $P\{\varphi(X_1-t_1, X_2-t_1, Y_1, Y_2)=1$ and $\varphi(X_1, X_2, Y_1, Y_2)=0\}$
$+P\{\varphi(X_1-t_1, X_2-t_1, Y_1, Y_2)=0$ and $\varphi(X_1, X_2, Y_1, Y_2)=1\}$.

This probability is transformed to the following integral form

$$4\Big[\iint_{x<x'}\{G(x')-G(x'-t_1)\}G(x-t_1)dF(x)dF(x')$$
$$+\iint_{x<x'}\{G(x)-G(x-t_1)\}\{1-G(x')\}dF(x)\,dF(x)\Big]$$

and this value is bounded by

$$8\int_{-\infty}^{\infty}\{G(x)-G(x-t_1)\}dF(x)\leqq 8bt_1,$$

provided the distribution function $G(x)$ has a derivative $g(x)$ bounded in absolute value by b. It can be therefore shown that

(23) $E\,|W(X_1, X_2, Y_1, Y_2, t_1, 0)-W(X_1, X_2, Y_1, Y_2, 0, 0)|\leqq(4a+8b)t_1$.

Similarly it can be shown that

$$(24)\quad EW(X_1, X_2, Y_1, Y_2, 0, t_2) - W(X_1, X_2, Y_1, Y_2, 0, 0) \leqq (8a+4b)t_2.$$

Thus the statistic $\hat{Q}_N$ has the same asymptotic normal distribution as the statistic Q_N from the above results, so that it has been shown that the test based on the statistic $\hat{Q}_N$ is asymptotically distribution free and has the same asymptotic properties as the test on Q_N. This consequence is stated in the following theorem.

Theorem If the X's and Y's are distributed symmetrically about the respective medians and have bounded density functions, the test of the hypothesis H based on the statistic $\hat{Q}_N$ is asymptotically distribution free.

SHIMANE UNIVERSITY.

References

[1] W. HOEFFDING,: *A class of statistics with asymptotically normal distribution.* Ann. Math. Stat. Vol. **19** (1948) 293-325.

[2] E. L. LEHMANN,: *Consistency and unbiasedness of certain nonparametric tests.* Ann. Math. Stat. Vol. **22** (1951) 165-179.

[3] B. V. SUKHATME,: *On certain two sample nonparametric tests for comparing variance.* Ann. Math. Stat. Vol. **28** (1957) 188-194.

[4] B. V. SUKHATME,: *Testing the hypothesis that two populations differ only in location.* Ann. Math. Stat. Vol. **29** (1958) 60-78.

[5] B. V. SUKHATME,: *A two sample distribution free test for comparing variance.* Biometrika. Vol. **45** (1958) 544-548.

[6] D. A. S. FRASER,: *NONPARAMETRIC METHODS OF STATISTIC.* John Wily & Son. 1957.

[7] A. M. MOOD,: *On the asymptotic efficiency of certain nonparametric two sample tests.* Ann. Math. Stat. Vol. **25** (1954) 514-522.

[8] R. TAMURA,: *On a nonparametric two sample test for scale.* Report Shimane Univ. No. 8 (1958) 1-8.