ON A CLASS MULTISAMPLE RANK TESTS BASED ON TRIMMED SAMPLES

Tamura, Ryoji
Department of Mathematics, Kumamoto University

https://doi.org/10.5109/13057

出版情報：統計数理研究. 15 (1/2), pp.1-6, 1972-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON A CLASS OF MULTISAMPLE RANK TESTS BASED ON TRIMMED SAMPLES

By

Ryoji TAMURA*

(Received May 5, 1971)

1. Introduction.

Hettmansperger [5] has shown that the Mann-Whitney test has larger asymptotic relative efficiencies in Pitman's sense for some distributions with heavy tails by using the trimmed samples instead of the complete samples. Recently Tamura [8] has generalized Hettmansperger's results to Bhapkar's test [3] for the c-sample problem. Along the same line as [8], we shall propose a class of rank tests based on the trimmed samples and discuss in detail about Kruskal-Wallis test [6] as its special case.

Let $X_{i1}<\cdots<X_{in_i}$ be order statistics from absolutely continuous cdf $F_i(x)=F(x-\theta_i)$, $i=1,\cdots,c$ where $F(x)$ has symmetric density $f(x)$ of unknown functional form. We further assume for $0<\alpha<\frac{1}{2}$ that $f(x)$ is continuously differentiable in some neighborhood of the unique population quantiles b_α and $b_{1-\alpha}$ of order α and $1-\alpha$, respectively. The hypothesis H_0, to be tested, is specified by $\theta_1=\cdots=\theta_c$ against the alternatives that not all θ's are equal. For this problem, a class of test statistics will be proposed on the basis of only the middle n_i-2k_i random variables $X_{ik_i+1}<\cdots<X_{in_i-k_i}$, $i=1,\cdots,c$, where $k_i=[n_i\alpha]$ denotes the largest integer not exceeding $n_i\alpha$. Throughout this paper, we assume that the sample size n_i, $i=1,\cdots,c$, increases in such a way that $\lim_{N\to\infty} n_i/N=\lambda_i$, $0<\lambda_i<1$ where $N=\sum_{i=1}^{c} n_i$. Some definitions and assumptions are given in Section 2. In Section 3, we derive the asymptotic distributions of the proposed statistics. Section 4 is concerned with the test of Kruskal-Wallis type.

2. Definitions and assumptions.

Let us define for $i=1,\cdots,c$

$$T_i=(n_i-2k_i)^{-1}\sum_{\beta=1}^{N-2k} E_\beta^{(i)} Z_\beta^{(i)}, \qquad k=\sum_{i=1}^{c} k_i \tag{2.1}$$

where

* Department of Mathematics, Kumamoto University, Kumamoto.

$$Z_\beta^{(i)} = \begin{cases} 1 & \text{if the } \beta\text{th smallest among the combined trimmed samples is from } F_i(x) \\ 0 & \text{otherwise} \end{cases}$$

and the $E_\beta^{(i)}$'s are some given constants. Then we can represent T_i by

$$T_i = \int_{-\infty}^{\infty} J_N(H_N(x))dG_i^*(x, n_i-2k_i) \tag{2.2}$$

where we set $J_N(\beta/(N-2k+1)) = E_\beta^{(i)}$ and $G_i^*(x, n_i-2k_i)$ is the empirical cdf based on the i-th trimmed sample and

$$H_N(x) = \sum_{i=1}^{c} \lambda_i G_i^*(x, n_i-2k_i).$$

ASSUMPTION (A).

(i) $\lim_{N\to\infty} J_N(t) = J(t)$ exists for $0<t<1$ and is not constant.

(ii) $\int_{I_N} [J_N(H_N)-J(H_N)]dG_i^*(x, n_i-2k_i) = o_p(N^{-\frac{1}{2}})$, $I_N = \{x\,;\, o<H_N(x)<1\}$

(iii) $J_N(1) = o(N^{\frac{1}{2}})$

(iv) $|d^jJ(t)/dt^j| \leqq M(t(1-t))^{-j-\frac{1}{2}+\delta}$, $j=0, 1, 2$

and some $\delta>0$ where M is a generic constant.

The form (2.2) and the assumption (A) have been dealt by Chernoff-Savage [4] and Puri [7]. Further we define for $i=1, \cdots, c$

$$\begin{aligned} &Y_{i1} = n_i^{\frac{1}{2}}(X_{ik_i+1}-\theta_i-b_\alpha), \qquad Y_{i2} = n_i^{\frac{1}{2}}(X_{in_i-k_i}-\theta_i-b_{1-\alpha}) \\ &\boldsymbol{Y}' = (Y_{11}, Y_{12}, \cdots, Y_{c1}, Y_{c2}). \end{aligned} \tag{2.3}$$

We here notice that the statistics T_i, given $\boldsymbol{Y}$, are the the rank statistics of Chernoff-Savage type [4] based on the trimmed samples from the cdf $G_i(x)$ with density $g_i(x)$,

$$g_i(x) = \begin{cases} f(x-\theta_i)/[F(b_{1-\alpha}+Y_{i2}/n_i^{\frac{1}{2}})-F(b_\alpha+Y_{i1}/n_i^{\frac{1}{2}})] \\ \qquad \text{for } b_\alpha+\theta_i+Y_{i1}/n_i^{\frac{1}{2}} \leqq x \leqq b_{1-\alpha}+\theta_i+Y_{i2}/n_i^{\frac{1}{2}} \\ 0 \qquad \text{otherwise}. \end{cases} \tag{2.4}$$

Finally, we define for $i=1, \cdots, c$

$$\begin{aligned} &R_i = (N-2k)^{\frac{1}{2}}(T_i - E(T_i\,|\,\boldsymbol{Y})), \qquad \boldsymbol{R}' = (R_1, \cdots, R_{c-1}) \\ &W_i = (N-2k)^{\frac{1}{2}}\left(T_i - \int_0^1 J(t)dt\right), \qquad \boldsymbol{W}' = (W_1, \cdots, W_{c-1}) \end{aligned} \tag{2.5}$$

where $E(*\,|\,\boldsymbol{Y})$ is the expected value of the statistic $*$, given $\boldsymbol{Y}$.

3. Asymptotic distributions.

Now we shall consider the asymptotic distributions of the proposed statistics under the hypothesis H_0 and the following sequence of alternatives

$$H_N:\ F_i(x)=F(x-\nu_i/N^{\frac{1}{2}}),\qquad i=1,\cdots,c \tag{3.1}$$

where not all ν's are equal.

LEMMA 3.1. *The random vector* $(\boldsymbol{R}, \boldsymbol{Y})$ *is asymptotically normally distributed if the assumption* (*A*) *holds.*

PROOF. It has been shown by Puri [7] that the random vector $\boldsymbol{R}$ has the joint normal distribution $N(\boldsymbol{0}, \Sigma^{(1)})$, given $\boldsymbol{Y}$, if the assumption (A) holds where $\Sigma^{(1)}=\|\sigma_{ij}^{(1)}\|$ $i, j=1,\cdots,c-1$

$$\sigma_{ij}^{(1)}=\lambda_i^{-1}(\delta_{ij}-\lambda_i)\Big[\int_0^1 J^2(t)dt-\Big(\int_0^1 J(t)dt\Big)^2\Big] \tag{3.2}$$

where $\delta_{ij}=1$ or 0 for $i=j$ or $i\neq j$. It is also well known that $\boldsymbol{Y}$ has the asymptotic normal cdf $N(\boldsymbol{0}, \boldsymbol{\Omega})$ where

$$\boldsymbol{\Omega}=\left\|\begin{matrix}\boldsymbol{\Omega}_1\cdots\boldsymbol{0}\\ \cdots\cdots\cdots\\ \boldsymbol{0}\cdots\boldsymbol{\Omega}_1\end{matrix}\right\|\qquad \boldsymbol{\Omega}_1=f^{-2}(b_\alpha)\left\|\begin{matrix}\alpha(1-\alpha) & \alpha^2\\ \alpha^2 & \alpha(1-\alpha)\end{matrix}\right\| \tag{3.3}$$

These facts establish the asymptotic normality of $(\boldsymbol{R}, \boldsymbol{Y})$.

THEOREM 3.1. *The random vector* $\boldsymbol{W}$ *has the asymptotic normal distribution under the assumption* (*A*) *and* $J(1)<\infty$, $J(0)<\infty$.

PROOF. By expanding the Puri's results [7]

$$E(T_i|\boldsymbol{Y})=\int_{-\infty}^{\infty}J\Big[\sum_{k=1}^{c}\lambda_k G_k(x)\Big]dG_i(x)+O_p(N^{-1})$$

in a Taylor series, we get the following (3.4).

$$\begin{aligned}E(T_i|\boldsymbol{Y})=&\int_0^1 J(t)dt+(\nu_i-\bar{\nu})\int_{b_\alpha}^{b_{1-\alpha}}J'[(F(t)-\alpha)/(1-2\alpha)]f(t)dF(t)\\ &\div N^{\frac{1}{2}}(1-2\alpha)^2+\sum_{k=1}^{c}\lambda_k n_k^{-\frac{1}{2}}(r_1Y_{k1}-r_2Y_{k2})+n_i^{-\frac{1}{2}}(s_1Y_{i1}-s_2Y_{i2})+O_q(N^{-1})\end{aligned} \tag{3.4}$$

where

$$r_1=(1-2\alpha)^{-3}f(b_\alpha)\int_{b_\alpha}^{b_{1-\alpha}}J'[(F(t)-\alpha/(1-2\alpha)](F(t)-1+\alpha)dF(t)$$

$$r_2=(1-2\alpha)^{-3}f(b_{1-\alpha})\int_{b_\alpha}^{b_{1-\alpha}}J'[(F(t)-\alpha)/(1-2\alpha)](F(t)-\alpha)dF(t)$$

$$s_1=(1-2\alpha)^{-2}f(b_\alpha)\Big[\int_{b_\alpha}^{b_{1-\alpha}}J[(F(t)-\alpha)/(1-2\alpha)]dF(t)-(1-2\alpha)J(0)\Big]$$

$$s_2=(1-2\alpha)^{-2}f(b_{1-\alpha})\Big[\int_{b_\alpha}^{b_{1-\alpha}}J[(F(t)-\alpha)/(1-2\alpha)]dF(t)-(1-2\alpha)J(1)\Big]$$

$$\bar{\nu}=\sum_{i=1}^{c}\lambda_i\nu_i\,.$$

For any constant vector $\boldsymbol{a}'=(a_1, \cdots, a_{c-1})$, the scalor product $\boldsymbol{a}'\boldsymbol{W}$, where

$$\boldsymbol{a}'\boldsymbol{W}=\boldsymbol{a}'\boldsymbol{R}+(N-2k)^{\frac{1}{2}}\sum_{i=1}^{c-1} a_i\Big[E(T_i\,|\,\boldsymbol{Y})-\int_0^1 J(t)dt\Big], \tag{3.5}$$

becomes to a linear function of $\boldsymbol{R}$ and $\boldsymbol{Y}$ from (3.4). Thus lemma 3.1 and the theorem by Anderson [1], p. 76, lead the asymptotic normality of $\boldsymbol{a}'\boldsymbol{W}$. The asymptotic normality of the random vector $\boldsymbol{W}$ may be established from that of $\boldsymbol{a}'\boldsymbol{W}$ for any $\boldsymbol{a}$. We can easily see from (3.4) that $EW_i\sim 0$ under H_0.

Denoting the covariance matrix of $\boldsymbol{W}$ by Σ, the statistic

$$V_\alpha=\boldsymbol{W}'\Sigma^{-1}\boldsymbol{W} \tag{3.6}$$

may be used as the test statistic for H_0. As a special case of V_α, we shall, in Section 4, discuss about the test of Kruskal-Wallis type.

4. The test of Kruskal-Wallis type.

The test statistic of Kruskal-Wallis type can be obtained by setting $J(t)=t$ in (3.6). In this case we rewrite the statistics such as T_i, W_i and etc. in the previous sections by T_i^*, W_i^* and etc.

LEMMA 4.1. *The asymptotic mean vector and covariance matrix of $\boldsymbol{W}^*$ is given by $\mu'=(\mu_1, \cdots, \mu_{c-1})$ and $\Sigma=\|\sigma_{ij}\|$ where*

$$\mu_i=(\nu_i-\bar{\nu})\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx/(1-2\alpha)^{3/2}, \tag{4.1}$$

$$\sigma_{ij}=(\delta_{ij}\lambda_i^{-1}-1)(1+4\alpha)/12(1-2\alpha) \tag{4.2}$$

PROOF. From the identity

$$E(T_i\,|\,\boldsymbol{Y})=\frac{1}{2}+(\nu_i-\bar{\nu})\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx/N^{\frac{1}{2}}(1-2\alpha)^2 \tag{4.3}$$

$$+f(b_\alpha)(n_i^{-\frac{1}{2}}Z_i-\sum_{k=1}^{c}\lambda_k n_k^{-\frac{1}{2}}Z_k)/2(1-2\alpha)+O_p(N^{-1})$$

$$Z_j=Y_{j1}+Y_{j2}$$

which is obtained from (3.4), we first get

$$EW_i^*=(\nu_i-\bar{\nu})\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx/(1-2\alpha)^{3/2}.$$

Next from

$$W_i^*=R_i^*+(N-2k)^{\frac{1}{2}}\Big[E(T_i^*\,|\,\boldsymbol{Y})-\frac{1}{2}\Big],$$

we get

$$\sigma_{ij}=\mathrm{cov}\,(R_i^*, R_j^*)+(N-2k)\,\mathrm{cov}\,(E(T_i^*\,|\,\boldsymbol{Y}),\ E(T_j^*\,|\,\boldsymbol{Y}))$$

$$+(N-2k)^{\frac{1}{2}}[\mathrm{cov}\,(R_i^*, E(T_j^*\,|\,\boldsymbol{Y}))+\mathrm{cov}\,(R_j^*, E(T_i^*\,|\,\boldsymbol{Y})].$$

Noticing the relations (3.2), (3.3), (4.3) and the following

$$\mathrm{cov}\,(R_i^*,\ E(T_j^*\mid Y)) = E_Y[\{E(T_j^*\mid Y) - E(T_j^*)\}E(R_i^*\mid Y)]$$
$$= 0\,,$$

we can get

$$\sigma_{ij} = (\delta_{ij} - \lambda_i)/12\lambda_i + f^2(b_\alpha)E[(\lambda_i^{-\frac{1}{2}}Z_i - \sum_{k=1}^{c}\lambda_k^{-\frac{1}{2}}Z_k)(\lambda_j^{-\frac{1}{2}}Z_j\sum_{k=1}^{c}\lambda_k^{-\frac{1}{2}}Z_k)]$$
$$\div 4(1-2\alpha)$$
$$= (\delta_{ij} - \lambda_i)(1+4\alpha)/12\lambda_i(1-2\alpha)\,.$$

Under H_0 we get $\mu = \mathbf{0}$ by setting all $\nu_i = 0$.

THEOREM 4.1. *The test statistic*

$$V_\alpha^* = 12(1-2\alpha)(1+4\alpha)^{-1}N^{-1}\sum_{i=1}^{c} n_i W_i^{*2} \tag{4.4}$$

is asymptotically distributed as x_{c-1}^2 *with* $c-1$ *degree of freedom under* H_0 *and as noncentral* $x_{c-1}^2(\delta_\alpha)$ *with* $c-1$ *degree of freedom and the noncentrality parameter* δ_α *under* H_N *where*

$$\delta_\alpha = 12\sum_{i=1}^{c}\lambda_i(\nu_i - \bar{\nu})^2\Big(\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx\Big)^2\Big/(1-2\alpha)^2(1+4\alpha)\,. \tag{4.5}$$

PROOF. It follows from Theorem 3.1 and Lemma 4.1 that $\boldsymbol{W}^*$ is asymptotically normal $N(\mathbf{0}, \Sigma)$ under H_0 and $N(\mu, \Sigma)$ under H_N. Therefore $\boldsymbol{W}^{*\prime}\Sigma^{-1}\boldsymbol{W}^*$ is asymtotically distributed as x_{c-1}^2 under H_0 and $x_{c-1}^2(\delta_\alpha)$ under H_N where

$$\Sigma^{-1} = 12(1-2\alpha)(1+4\alpha)^{-1}\lambda_c^{-1}\left\|\begin{matrix} \lambda_1(\lambda_1+\lambda_c) & \lambda_1\lambda_2 \cdots \lambda_1\lambda_{c-1} \\ \lambda_2\lambda_1 & \lambda_2(\lambda_2+\lambda_c) \cdots \lambda_2\lambda_{c-1} \\ \cdots\cdots & \cdots\cdots \\ \lambda_{c-1}\lambda & \lambda_{c-1}\lambda_2 \cdots \lambda_{c-1}(\lambda_{c-1}+\lambda_c) \end{matrix}\right\|$$

$$\delta_\alpha = \mu'\Sigma^{-1}\mu\,.$$

Some calculations show that

$$\boldsymbol{W}^{*\prime}\Sigma^{-1}\boldsymbol{W}^* = 12(1-2\alpha)(1+4\alpha)^{-1}\sum_{i=1}^{c}\lambda_i W_i^{*2}\,.$$

and

$$\mu'\Sigma^{-1}\mu = 12\sum_{i=1}^{c}\lambda_i(\nu_i-\bar{\nu})^2\Big(\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx\Big)^2\Big/(1-2\alpha)^2(1+4\alpha)\,.$$

We here notice that Kruskal-Wallis's test may be denoted by V_0^*.

It has been shown by Andrews [2] that the Pitman efficiency e_α of the V_α^* test respective to the V_0^* test is given by the ratio of the noncentrality parameters in the asymptotic x_{c-1}^2 distributions of their test statistics. From (4.5), we have

$$e_\alpha = \Big(\int_{b_\alpha}^{b_{1-\alpha}} f^2(x)dx\Big)^2\Big/(1+4\alpha)(1-2\alpha)^2\Big(\int_{-\infty}^{\infty} f^2(x)dx\Big)^2 \tag{4.6}$$

Lastly we give the numerical values of e_α for some distributions with heavier tails than the normal distribution.

Table of e_α.

α	.1	.2	.25	.3	.35	.4
Logistic	.99	.97	.95	.92	.88	.84
D. Exp.	1.03	1.09	1.13	1.16	1.20	1.25
Cauchy	1.09	1.26	1.34	1.40	1.43	1.44

References

[1] ANDERSON, T. W. (1958) Introduction of multivariate statistical analysis. Wiley, New York.

[2] ANDREWS, F. C. (1954) *Asymptotic behavior of some rank tests for analysis of variance*, Ann. Math. Statist., **25** 724–736.

[3] BHAPKAR, V. P. (1961) *A nonparametric test for the problem of several samples*, Ann. Math. Statist., **32** 1108–1117.

[4] CHERNOFF, H. and SAVAGE, I. R. (1958) *Asymptotic normality and efficiency of certain non-parametric test statistics*, Ann. Math. Statist., **29** 972–994.

[5] HETTMANSPERGER, T. P. (1968) *On the trimmed Mann-Whitney statistic*, Ann. Math. Statist. **39** 1610–1614.

[6] KRUSKAL, W. H. and WALLIS, W. A. (1952) *Use of ranks in one-criterion varance analysis*, J. A. S. A., **47** 583–621.

[7] PURI, M. L. (1964) *Asymptotic efficiency of a class of c-sample tests*, Ann. Math. Statist., **35** 102–121.

[8] TAMURA, R. (1971) *On a c-sample test based on trimmed samples*, Ann. Math. Statist., **42** 1455–1460.