Journal of the Korean Data & Information Science Society
2009, **20**(1), 237–250

한국데이터정보과학회지

Some nonparametric test procedure for the multi-sample case[†]

Hyo-Il Park[1] · Ju Sung Kim[2]

[1]Department of Statistics, Chong-ju University
[2]Department of Informational Statistics, Chungbuk National University

Received 28 December 2008, revised 18 January 2009, accepted 23 January 2009

Abstract

We consider a nonparametric test procedure for the multi-sample problem with grouped data. We construct the test statistics based on the scores obtained from the likelihood ratio principle and derive the limiting distribution under the null hypothesis. Also we illustrate our procedure with an example and obtain the asymptotic properties under the Pitman translation alternatives. Also we discuss some concluding remarks. Finally we derive the covariance between components in the Appendix.

Keywords: Grouped data, limiting power of test, multi-sample problem, nonparametric test, permutation principle, Pitman translation alternative

1. Introduction

Suppose that we have independent K samples $X_{k1}, \ldots, X_{kn_k}$, $K \geq 3$. Also suppose that the k th population is governed by the unknown distribution function F_k, $k = 1, \ldots, K$. We assume that the unknown distribution function F_k is continuous with density f_k for each k . With these data, our interest would be to test $H_0 : F_1 = \ldots = F_K$ against the general alternative, which says that at least one equality does not hold. A considerable amount of work for the nonparametric procedure under the restricted alternatives has appeared in the literature but for the general alternative, we have hardly found any procedure except the Kruskal-Wallis test (1952), which has been widely used as a nonparametric procedure. For the right censored data, Brookmeyer and Crowley (1982) proposed a median test. However in this study, we consider the following situations. For the study of life time of light-bulb, we may decide to observe the failure time of each bulb by visiting laboratory periodically because of economic or any other reasons. Or for some specific part of a machine, we may decide to inspect the machine periodically whether the specific part fails after we run the machine for some fixed time. Therefore according to the pre-determined time schedule, we observe each object under study whether it fails or not. In these cases, the data become discretized in

[†] This work was supported by the grant of the Chungbuk National University in 2008.
[1] Professor, Department of Statistics, Chong-ju University, Chong-ju 360-764, Korea.
[2] Corresponding Author: Professor, Department of Informational Statistics, Chungbuk National University, Chong-ju 361-763, Korea. E-mail: kimjs@chungbuk.ac.kr

spite of the continuity of life time distribution. We call those as the grouped data. Heitjan (1989) reviewed extensively development of statistical inferences for the grouped data in parametric setting and indicated some of the unsolved questions in theory and application aspects. Based on these grouped data, for testing $H_0 : F_1 = \ldots = F_K$ against the general alternative, one may apply the Pearson's chi-square test as a nonparametric procedure. Or one may use the Kruskal-Wallis test by using the mid-rank among the observations which fail in the same time-interval. However in case of two-sample setting, one may apply the Puri and Sen's procedure (1985) for the grouped data. Puri and Sen proposed a class of nonparametric tests for the linear model. They derived the test statistics using the likelihood ratio principle. Therefore the procedure may be optimal in the sense of the locally most powerful test. Also Park (1993) proposed a class of nonparametric tests for the grouped and right censored data.

In this paper, we consider to propose a nonparametric test procedure for the multi sample case with the grouped observations. In the next section we begin our discussion with reviewing some results for the two sample case.

2. Review of some results for the two sample case

In this section, we review some results in case of $K = 2$. Since we are interested in the life time data, without loss of generality, we will consider the positive half real line. Suppose that the positive half real line $[0, \infty)$ is partitioned into d sub-intervals $I_j = [a_j, a_{j+1})$ for any fixed time a_j , $j = 1, \ldots, d$ with the notation that $a_1 = 0$ and $a_{d+1} = \infty$. We note that we can not observe X_{ki} directly but only have the information that X_{ki} may be contained in one of d sub-intervals. Therefore for each $k = 1, 2$ and for each $i = 1, \ldots, n_k$, each observable random variable, X^*_{ki}, can be expressed as

$$X^*_{ki} = \sum_{j=1}^{d} I_j Z_{kij},$$

where for every k , i and $j = 1, \ldots, d$

$$Z_{kij} = \begin{cases} 1, & X_{ki} \in I_j \\ 0, & \text{otherwise} \end{cases} .$$

Then for testing $H_0 : F_1 = F_2$ against $H_1 : F_1 \neq F_2$ based on the following two samples, $X^*_{11}, \ldots, X^*_{1n_1}$ and $X^*_{21}, \ldots, X^*_{2n_2}$, Puri and Sen (1985) proposed the following linear rank statistic of the form

$$T_n = \sum_{i=1}^{n_1} \sum_{j=1}^{d} \Delta_{nj} Z_{1ij} = \sum_{j=1}^{d} \Delta_{nj} n_{1j},$$

where Δ_{nj} is some score for the observations in the j th sub-interval I_j and will be explicitly defined later and n_{1j} , the number of observations of the first sample in the j th sub-interval I_j . We note that when the number of observations in each sub-interval, I_j , is at most one, this corresponds to the no tied-value case. Then one may reject $H_0 : F_1 = F_2$ in favor of $H_1 : F_1 \neq F_2$ for large values of $|T_n - E_0(T_n)|$, where $E_0(T_n)$ is the expectation of T_n under H_0 , which will be identified later also. For any given significance level, in order to

determine the critical value, we need the null distribution of T_n . Then by applying the permutation principle (cf. Good, 2000), we may obtain the null distribution of T_n for small and reasonable sample sizes. For the large sample case, in order to derive the asymptotic normality, first of all, we have to obtain the mean and variance of T_n under H_0 . From Puri and Sen (1985), we have

$$E_0(T_n) = n_1 \sum_{j=1}^{d} \Delta_{nj} \frac{n_{1j} + n_{2j}}{n_n + n_2} = n_1 \bar{\Delta}_n$$

and

$$V_0(T_n) = \frac{n_1 n_2}{n_1 + n_2 - 1} \left\{ \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{1j} + n_{2j}}{n_1 + n_2} - \bar{\Delta}_n^2 \right\},$$

where n_{2j} is the number of observations of the second sample in interval I_j . Then one can show that the standardized form

$$M_n = \frac{T_n - E_0(T_n)}{\sqrt{V_0(T_n)}}$$

converges in distribution to a standard normal random variable with the assumption about the ratio between two sample sizes by applying the central limit theorem and Slutsky's theorem. You may refer to Puri and Sen (1985) for more detailed discussion for this subject.

Now we discuss the score function Δ_{nj} in some detail. For this purpose, let $\phi(u)$, $0 < u < 1$ be any non-decreasing square-integrable function and define for each $j = 1, \ldots, d$,

$$\Delta_{nj} = \frac{1}{\hat{F}_n(a_{j+1}) - \hat{F}_n(a_j)} \int_{\hat{F}_n(a_j)}^{\hat{F}_n(a_{j+1})} \phi(u) du,$$

where $\hat{F}_n$ is the empirical distribution function of the underlying distribution function F based on the combined sample from the two samples. We note that if $\phi(u) = u$, Δ_{nj} is the Wilcoxon score. Therefore one may obtain a class of nonparametric test statistics with various choice of the score function ϕ . As a matter of fact, Puri and Sen (1985) derived the optimal score functions using the likelihood ratio principle. The optimal score functions are of the following form: for each $j = 1, \ldots, d$,

$$\Delta_j^* = \frac{1}{F(a_{j+1}) - F(a_j)} \int_{F(a_j)}^{F(a_{j+1})} \psi(u) du,$$

where $\psi(u) = -f'(F^{-1}(u))/f(F^{-1}(u))$ with the notation that $F^{-1}(u) = \inf\{t : F(t) \geq u\}$ for $0 < u < 1$ and f' is the derivative of f . Therefore if the underlying distribution function F were completely known, we might obtain a locally most powerful test using Δ_j^* . In the nonparametric case, since F and hence $\psi(u)$ as well as Δ_j^* are unknown, one may try to obtain asymptotically the optimal scores by substituting $\hat{F}_n$ for F with suitable choice of ψ . This may be achieved by using Δ_{nj} . For example, if the underlying distribution function F has a logistic density, then we may choose

$$\phi(u) = 2u - 1$$

in Δ_{nj} to produce the locally most powerful nonparametric test, which is again the Wilcoxon score for the two sample case.

3. Extension to the multi-sample problem

In this section, we consider an extension of the linear rank test procedure for the two sample case to the multi-sample ($K \geq 3$) problem for the grouped data. For this purpose, let for each k , $k = 1, \ldots , K$,

$$T_{kn} = \sum_{i=1}^{n_k} \sum_{j=1}^{d} \Delta_{nj} Z_{kij} = \sum_{j=1}^{d} \Delta_{nj} n_{kj}$$

be the linear rank statistic from the th sample, where is the number of observations in the j th sub-interval from the k th sample. Then we have that

$$E_0(T_{kn}) = n_k \sum_{j=1}^{d} \Delta_{nj} \frac{n_{\cdot j}}{n} = n_k \bar{\Delta}_n$$

and

$$V_0(T_{kn}) = n_k \frac{n - n_k}{n-1} \left\{ \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2 \right\},$$

where $n_{\cdot} = \sum_{k=1}^{K} n_{kj}$ and $n = \sum_{k=1}^{K} n_k$. Also for any $k \neq m$, the null covariance $Cov(T_{kn}, T_{mn})$ between T_{kn} and T_{mn} is as follows:

$$Cov(T_{kn}, T_{mn}) = -\frac{n_k n_m}{n-1} \left\{ \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2 \right\}.$$

All the derivations of the above moments are based on the permutation principle. The derivation of $Cov(T_{kn}, T_{mn})$ will be postponed until the appendix. Let Σ_{0n} be the null covariance matrix of $(T_{1n}, \ldots , T_{Kn})'$. Then we have the following result.

Lemma 3.1 For each n , the covariance matrix Σ_{0n} has a rank $K - 1$.

Proof: This can be proved by the fact that the elementary row or column operations do not affect the rank (cf. Schott, 1997). For this, we note that

$$\Sigma_{0n} = \begin{pmatrix} \frac{n_1(n-n_1)}{n-1} & \cdots & -\frac{n_1 n_K}{n-1} \\ & \cdots & \\ -\frac{n_1 n_K}{n-1} & \cdots & \frac{n_K(n-n_K)}{n-1} \end{pmatrix} \left\{ \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2 \right\} = S_n \left\{ \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2 \right\}.$$

Then it is enough to consider the rank of S_n for that of Σ_{0n} . First by multiplying $(n-1)/\sqrt{n_k}$ for the k th row and then $1/n\sqrt{n_k}$ for the k th column of S_n and denoting

$$\boldsymbol{p}' = (\sqrt{n_1/n}, \ldots , \sqrt{n_K/n}),$$

we obtain that

$$S_n^* = \begin{pmatrix} \frac{n-n_1}{n} & \cdots & -\frac{\sqrt{n_1 n_K}}{n} \\ \cdots & & \cdots \\ -\frac{\sqrt{n_1 n_K}}{n} & \cdots & \frac{n-n_K}{n} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ & \cdots & \\ 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} \frac{n_1}{n} & \cdots & \frac{\sqrt{n_1 n_K}}{n} \\ \cdots & & \cdots \\ \frac{\sqrt{n_1 n_K}}{n} & \cdots & \frac{n_K}{n} \end{pmatrix} = \boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}',$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix. We note that the rank of S_n is the same as that of S_n^* . Therefore it is enough to obtain the rank of S_n^* for that of Σ_{0n} . For this, we note that $\boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}'$ is idempotent since

$$(\boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}')(\boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}') = \boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}'.$$

Since the rank of $\boldsymbol{I}_K - \boldsymbol{p}\boldsymbol{p}'$ is

$$\sum_{k=1}^{K}(1 - n_k/n) = K - 1,$$

we obtain the result. □

Lemma 3.2 Under $H_0 : F_1 = \ldots = F_K = F$, for each j , $j = 1, \ldots, d$, Δ_{nj} converges in probability to Δ_j , where

$$\Delta_j = \frac{1}{F(a_{j+1}) - F(a_j)} \int_{F(a_j)}^{F(a_{j+1})} \phi(u)du.$$

Proof: This result follows easily by noting that all the components in the expression of Δ_{nj} are the empirical probability and the score function ϕ is square-integrable.

In passing, we also note that under $H_0 : F_1 = \ldots = F_K = F$, $C_n = \sum_{j=1}^{d} \Delta_{nj}^2 \frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2$ converges in probability to $C_0 = \sum_{j=1}^{d} \Delta_j^2 [F(a_{j+1}) - F(a_j)] - \left[\int_0^1 \phi(u)du\right]^2$ by the same reason for Lemma 3.2. Then for any version of the generalized inverse Σ_{0n}^- of Σ_{0n} , we may propose the following test statistic for testing $H_0 : F_1 = \ldots = F_K$,

$$\boldsymbol{M}_n = \begin{pmatrix} T_{1n} - E_0(T_{1n}) \\ \cdots \\ T_{Kn} - E_0(T_{Kn}) \end{pmatrix}' \Sigma_{0n}^- \begin{pmatrix} T_{1n} - E_0(T_{1n}) \\ \cdots \\ T_{Kn} - E_0(T_{Kn}) \end{pmatrix}.$$

Then we may reject H_0 for large values of $\boldsymbol{M}_n$. For any given significance level α , in order to obtain the critical value $C_n(\alpha)$, we need the null distribution of $\boldsymbol{M}_n$. One may obtain the null distribution for $\boldsymbol{M}_n$ by applying the permutation principle for any reasonable sample sizes. For the large sample case, we consider obtaining the asymptotic distribution by applying the large sample approximation. For this purpose, we assume that for each k , $k = 1, \ldots, K$,

$$\lim_{n\to\infty} n_k/n = \lambda_k \text{ for some } \lambda_k \in (0, 1). \tag{3.1}$$

Then we obtain the asymptotic distribution with the assumption (3.1). □

Theorem 3.1 With the assumption (3.1), under H_0 , the distribution of $\boldsymbol{M}_n$ converges in distribution to a chi-square distribution with $K-1$ degrees of freedom.

Proof: From Puri and Sen (1985), for each k , we see that $(1/\sqrt{n})(T_{kn}-E_0(T_{kn}))$ converges in distribution to a normal random variable with mean 0 and variance $\lambda_k(1-\lambda_k)C_0$ with Lemma 3.2 and assumption (3.1) by applying Slutsky's theorem. Therefore from the Cramer-Wold device (cf. Billingsley, 1985) and again using Slutsky's theorem, we obtain that

$$\frac{1}{\sqrt{n}}(T_{1n}-E_0(T_{1n}),\cdots,T_{Kn}-E_0(T_{Kn}))'$$

converges in distribution to a K-variate normal random vector with 0 mean vector and covariance matrix Σ_0 , where

$$\Sigma_0=\begin{pmatrix}\lambda_1(1-\lambda_1) & \cdots & -\lambda_1\lambda_K \\ & \cdots & \\ -\lambda_1\lambda_K & \cdots & \lambda_K(1-\lambda_K)\end{pmatrix}C_0,$$

whose rank is also $K-1$. We note that for each n , Σ_{0n} is symmetric and has $K-1$ as its rank. Therefore from the Spectral Decomposition Theorem (cf. Mardia et al., 1979), Σ_{0n} can be written as

$$\Sigma_{0n}=\Gamma_n\Omega_n\Gamma_n',$$

where Ω_n is a $(K-1)\times(K-1)$ diagonal matrix of non-zero eigenvalues of Σ_{0n} and Γ_n is a $K\times(K-1)$ orthogonal matrix whose columns are standardized eigenvectors. Then $\Gamma_n\Omega_n^{-1}\Gamma_n'$ is a version of the generalized inverse of Σ_{0n} , which in turn means that the random vector

$$(T_{1n}-E_0(T_{1n}),\ldots,T_{Kn}-E_0(T_{Kn}))\Gamma_n\Omega_n^{-1/2}$$

converges in distribution to a normal random vector with 0 mean vector and covariance matrix $\boldsymbol{I}_{K-1}$, where $\boldsymbol{I}_{K-1}$ is the $(K-1)\times(K-1)$ identity matrix. Therefore

$$\begin{pmatrix}T_{1n}-E_0(T_{1n})\\ \cdots \\ T_{Kn}-E_0(T_{Kn})\end{pmatrix}'\Gamma_n\Omega_n^{-1}\Gamma_n'\begin{pmatrix}T_{1n}-E_0(T_{1n})\\ \cdots \\ T_{Kn}-E_0(T_{Kn})\end{pmatrix}$$

converges in distribution to a chi-square random variable with $K-1$ degrees of freedom. Now we note that for each n , $(T_{1n}-E_0(T_{1n}),\ldots,T_{Kn}-E_0(T_{Kn}))'$ lies in the space which is spanned by Σ_{0n} since

$$(1,\ldots,1)\begin{pmatrix}T_{1n}-E_0(T_{1n})\\ \cdots \\ T_{Kn}-E_0(T_{Kn})\end{pmatrix}=0,$$

where $(1,\ldots,1)'$ consists of the null space of Σ_{0n} . This means that $\boldsymbol{M}_n$ is G-inverse invariant for each n . Thus we obtain the result. □

Remark 3.1 As another statistic for testing $H_0: F_1=\ldots=F_K=F$, one may consider using the following form:

$$KW_n=(n-1)\frac{\sum_{k=1}^K n_k\left(T_{kn}/n_k-\bar{\Delta}_n\right)^2}{n\left\{\sum_{j=1}^d\Delta_{nj}^2\frac{n_{\bullet j}}{n}-\bar{\Delta}_n^2\right\}}=(n-1)\frac{\sum_{k=1}^K\left(T_{kn}-n_k\bar{\Delta}_n\right)^2}{nn_k\left\{\sum_{j=1}^d\Delta_{nj}^2\frac{n_{\bullet j}}{n}-\bar{\Delta}_n^2\right\}} \qquad (3.2)$$

which is a modified version of Kruskal-Wallis (1952) statistic for the grouped data. The derivation of this form and corresponding asymptotic normality will be discussed briefly in the sequel. Since no generalized inverse of the covariance matrix is involved in the expression of KW_n , it may be useful for practical purpose. However since KW_n does not reveal the covariance structure explicitly, we used the form $\boldsymbol{M}_n$ for the purposes of discussion of the asymptotic properties of our proposed test in the sequel.

4. An example

In order to illustrate our procedure, we consider the blood lead data, which were analyzed by Hasselblad et al. (1980) under the log-normal assumption. The data consist of year, ethnic group, age and lead level from 1970 to 1976. The blood lead levels were recorded with some interval. In this study, suppose that we are interested in detecting any difference among the three ethnic groups, white, black and Puerto Rican. For this purpose, we only consider only the data of 1970. In the following table, we summarized the frequencies between the blood lead levels and ethnic groups. We chose the Wilcoxon score, $\phi(u) = u$ and obtained

표 **4.1** Data for blood levels and ethnic groups

	lead level							
races	0-14	15-24	25-34	35-44	45-54	55-64	65+	total
Black	317	2245	3424	1870	651	220	125	8852
Puero Rican	559	3148	2996	1074	306	109	65	8259
White	111	522	424	157	41	16	14	1285
total	987	5915	6844	3101	998	345	206	18396

the following statistics which are necessary for the analysis of our procedure:

$$T_{1n} = 1184.39, \quad T_{2n} = 1178.63, \quad T_{3n} = 180.34$$

$$E_0(T_{1n}) = 1223.84, E_0(T_{2n}) = 1223.84, E_0(T_{3n}) = 177.66$$

$$\Sigma_{0n} = \begin{pmatrix} 15.37 & -13.30 & -2.07 \\ -13.30 & 15.24 & -1.93 \\ -2.07 & -1.93 & 4.00 \end{pmatrix}$$

and a generalized inverse Σ_{0n}^{-} of Σ_{0n} is as follows:

$$\Sigma_{0n}^{-} = \begin{pmatrix} 0.04 & 0.01 & -0.06 \\ 0.01 & 0.05 & -0.06 \\ -0.06 & -0.06 & 0.11 \end{pmatrix}.$$

Then we obtain that

$$\boldsymbol{M}_n = 103.00,$$

which shows the strongly significant difference among the ethnic groups. Also if we consider the Pearson's chi-square test, then we obtain 848.54 for the chi-square statistic, whose p -value is less than 0.0001. Therefore one may draw the same conclusion with our test. All the calculations were carried out using the IML/SAS on PC.

5. Asymptotic properties of the test

In this section, in order to deal with the asymptotic properties of our test, we consider the location translation model. This means that for each , there is a location translation parameter $\delta_k \in \boldsymbol{R}^1$ such that for all $x \in \boldsymbol{R}^1$,

$$F_k(x) = F(x - \delta_k).$$

First of all, we derive the limiting power of our test under the following Pitman translation alternatives: For each k and n , let

$$H_{1n} : \delta_{kn} = c_k/\sqrt{n},$$

where c_k 's are some fixed real numbers. Also for each n , let

$$G_n = \sum_{k=1}^{K} \frac{n_k}{n} F_k \text{ and } \hat{G}_n = \sum_{k=1}^{K} \frac{n_k}{n} \hat{F}_{kn_k},$$

where for each k , $\hat{F}_{kn_k}$ is the empirical distribution function of F_k based on $X^*_{k1}, \ldots, X^*_{kn_k}$. Also we assume that

$$\int_{-\infty}^{\infty} f'(x)dx = 0, \tag{5.1}$$

where f' is the derivative of f and f is the corresponding density function of F . The assumption (5.1) is known as a regularity assumption on density (cf. Bickel and Doksum, 1977). From now on, we use Δ_j's instead of Δ_{nj}'s since we consider obtaining the limiting power of test for each fixed score function. Then we note that for each k ,

$$E_{H_{1n}}(T_{kn}) = n_k \sum_{j=1}^{d} \Delta_j \left(\hat{G}_n(a_{j+1}) - \hat{G}_n(a_j) \right)$$

or

$$E_{H_{1n}}(T_{kn}/n_k) = \sum_{j=1}^{d} \Delta_j \left(\hat{G}_n(a_{j+1}) - \hat{G}_n(a_j) \right).$$

From Glivenko-Cantelli lemma, we note with probability one that as $n \to \infty$

$$\sup_{j \in \{1, \ldots, d\}} \left| \hat{G}_n(a_j) - G_n(a_j) \right| \to 0.$$

Also let

$$\mu_n(\boldsymbol{\delta}_n) = \sum_{j=1}^{d} \Delta_j \left(G_n(a_{j+1}) - G_n(a_j) \right),$$

where $\boldsymbol{\delta}_n = (\delta_{1n}, \ldots, \delta_{Kn})'$. Then we note that for every k , we have with probability one that as $n \to \infty$

$$|E_{H_{1n}}(T_{kn}/n_k) - \mu_n(\boldsymbol{\delta}_n)| \to 0.$$

Therefore for each k , we may use $\mu_n(\boldsymbol{\delta}_n)$ instead of $E_{H_{1n}}(T_{kn}/n_k)$ for the asymptotic properties of our test. We note that

$$\begin{aligned} G_n(a_{j+1}) - G_n(a_j) &= \sum_{k=1}^{K} \frac{n_k}{n} (F_k(a_{j+1}) - F_k(a_j)) \\ &= \sum_{k=1}^{K} \frac{n_k}{n} (F_k(a_{j+1} - \delta_k) - F_k(a_j - \delta_k)) . \end{aligned}$$

Therefore we have that for each n ,

$$\begin{aligned} \frac{\partial \mu_n(\delta)}{\partial \delta_k} |_{\delta_k=0} &= \frac{n_k}{n} \sum_{j=1}^{d} \Delta_j \left(f(a_j) - f(a_{j+1})\right) \\ &= \frac{n_k}{n} \sum_{j=1}^{d} \Delta_j \Delta_j^* \left(F(a_{j+1}) - F(a_j)\right) \end{aligned}$$

since

$$\Delta_j^* = \frac{1}{F(a_{j+1}) - F(a_j)} \int_{F(a_j)}^{F(a_{j+1})} \psi(u) du = \frac{f(a_j) - f(a_{j+1})}{F(a_{j+1}) - F(a_j)}.$$

Also with the same arguments used for the derivation of the null variance, one can easily obtain the variance of T_{kn} under the Pitman translation alternatives as follows: for each k ,

$$V_{H_{1n}}(T_{kn}) = n_k \frac{n - n_k}{n - 1} \left\{ \sum_{j=1}^{d} \Delta_j^2 \left(\hat{G}_n(a_{j+1}) - \hat{G}_n(a_j) \right) - \left[\sum_{j=1}^{d} \Delta_j \left(\hat{G}_n(a_{j+1}) - \hat{G}_n(a_j) \right) \right]^2 \right\}.$$

Then it is easy to see that for each k , as $n \to \infty$,

$$V_{H_{1n}}(T_{kn}) / V_0(T_{kn}) \to 1$$

from the fact that under the Pitman translation alternatives, with probability one,

$$\hat{G}_n(a_{j+1}) - \hat{G}_n(a_j) \to F(a_{j+1}) - F(a_j).$$

Also with the same arguments used for the variances, for any pair $k \neq m$,

$$Cov_{H_{1n}}(T_{kn}, T_{mn}) / Cov_0(T_{kn}, T_{mn}) \to 1.$$

Finally, we have that for each k ,

$$\begin{aligned} \lim_{n \to \infty} \frac{n_k (\partial \mu_n(\boldsymbol{\delta}) / \partial \delta_k)|_{\delta_k=0}}{\sqrt{n V_0(T_{kn})}} &= \lambda_k \sqrt{\frac{\lambda_k}{1 - \lambda_k}} \frac{\sum_{j=1}^{d} \Delta_j \Delta_j^* \left(F(a_{j+1}) - F(a_j)\right)}{\sqrt{\sum_{j=1}^{d} \Delta_j^2 \left(F(a_{j+1}) - F(a_j)\right) - \bar{\Delta}^2}} \\ &= \zeta_k \end{aligned}$$

, say.

Let Σ_{1n} be the covariance matrix of $(T_{1n}, \ldots, T_{Kn})'$ under the Pitman translation alternatives for each n . Then from the above arguments, we see that the asymptotic distribution of

$$\boldsymbol{M}_n = \begin{pmatrix} T_{1n} - E_{H_{1n}}(T_{1n}) \\ \cdots \\ T_{Kn} - E_{H_{1n}}(T_{Kn}) \end{pmatrix}' \Sigma_{1n}^{-} \begin{pmatrix} T_{1n} - E_{H_{1n}}(T_{1n}) \\ \cdots \\ T_{Kn} - E_{H_{1n}}(T_{Kn}) \end{pmatrix}$$

coincides with that of

$$\boldsymbol{M}_n^* = \begin{pmatrix} T_{1n} - E_0(T_{1n}) + \zeta_1 c_1 \\ \cdots \\ T_{Kn} - E_0(T_{Kn}) + \zeta_K c_K \end{pmatrix}' \Sigma_{0n}^{-} \begin{pmatrix} T_{1n} - E_0(T_{1n}) + \zeta_1 c_1 \\ \cdots \\ T_{Kn} - E_0(T_{Kn}) + \zeta_K c_K \end{pmatrix}$$

whose asymptotic distribution is a non-central chi-square with $K-1$ degrees of freedom and non-centrality parameter Θ (cf. Johnson and Katz, 1970), where

$$\Theta = \frac{1}{2} \begin{pmatrix} \zeta_1 c_1 \\ \cdots \\ \zeta_K c_K \end{pmatrix}' \Sigma_0^{-} \begin{pmatrix} \zeta_1 c_1 \\ \cdots \\ \zeta_K c_K \end{pmatrix}. \tag{5.2}$$

Let $q_\alpha(K-1)$ be the upper α -percentile point of the (central) chi-square distribution with $K-1$ degrees of freedom. Also let $Q(\Theta)$ be a chi-square random variable with $K-1$ degrees of freedom and non-centrality parameter Θ . Then the asymptotic power of our proposed test under the Pitman translation alternatives is

$$\lim_{n\to\infty} \Pr_{H_{1n}} \{\boldsymbol{M}_n^* \geq q_\alpha(K-1)\} = \Pr\{Q(\Theta) \geq q_\alpha(K-1)\} > \alpha,$$

since $\Theta > 0$ and $\Pr\{Q(\Theta) \geq q\}$ is strictly increasing in Θ for each fixed q .

From now on, we consider the intrinsic loss on efficiency due to grouping (cf. Puri and Sen, 1985). For this matter, we consider the two-sample case. If we use Δ_j^* instead of Δ_j , then we obtain that

$$\begin{aligned} \lim_{n\to\infty} \frac{n_1 \mu_n'(0)}{\sqrt{n V_0(T_n)}} &= \sqrt{\lambda_1 \lambda_2} \frac{\sum_{j=1}^{d} \Delta_j^{*2} \left(F(a_{j+1}) - F(a_j)\right)}{\sqrt{\sum_{j=1}^{n} \Delta_j^{*2} \left(F(a_{j+1}) - F(a_j)\right)}} \\ &= \sqrt{\lambda_1 \lambda_2} \sqrt{\sum_{j=1}^{n} \Delta_j^{*2} \left(F(a_{j+1}) - F(a_j)\right)} \end{aligned}$$

since

$$\begin{aligned} \bar{\Delta}^* &= \sum_{j=1}^{d} \frac{1}{F(a_{j+1}) - F(a_j)} \int_{F(a_j)}^{F(a_{j+1})} \psi(u) du \left(F(a_{j+1}) - F(a_j)\right) \\ &= \sum_{j=1}^{d} \int_{F(a_j)}^{F(a_{j+1})} \left[-\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \right] du \\ &= -\int_{-\infty}^{\infty} f'(x) dx \\ &= 0 \end{aligned}$$

using the variable-transformation-technique and from the regularity assumption (5.1). Furthermore if we observe X_{ki} directly not X_{ki}^* , then we obtain that

$$\lim_{n\to\infty} \frac{n_1\mu_n'(0)}{\sqrt{nV_0(T_n)}} = \sqrt{\lambda_1\lambda_2}\frac{\int_0^1 \psi^2(u)du}{\sqrt{\int_0^1 \psi^2(u)du}} = \sqrt{\lambda_1\lambda_2}\sqrt{\int_0^1 \psi^2(u)du}.$$

The intrinsic loss on efficiency due to grouping, $L(G)$ is defined as

$$L(G) = \frac{\sum_{j=1}^d \Delta_j^{*2}\left(F(a_{j+1}) - F(a_j)\right)}{\int_0^1 \psi^2(u)du}.$$

We note that $\sqrt{\lambda_1\lambda_2}\sqrt{\int_0^1 \psi^2(u)du}$ and $\sqrt{\lambda_1\lambda_2}\sqrt{\sum_{j=1}^n \Delta_j^{*2}\left(F(a_{j+1}) - F(a_j)\right)}$ are the efficacies of the tests based on the statistic $\boldsymbol{M}_n$ for the grouped and non-grouped data, respectively. Therefore $L(G)$ is the square of the ratio of the efficacy for the grouped data relative to that of non-grouped data and can not exceed 1. To see this, we note that with Cauchy-Schwarz inequality (cf. Chung, 1974)

$$\begin{aligned}
0 &< \sum_{j=1}^n \Delta_j^{*2}\left(F(a_{j+1}) - F(a_j)\right) \\
&= \sum_{j=1}^d \left[\frac{1}{F(a_{j+1}) - F(a_j)}\int_{F(a_j)}^{F(a_{j+1})} \psi(u)du\right]^2 \left(F(a_{j+1}) - F(a_j)\right) \\
&\le \sum_{j=1}^d \left[\frac{1}{F(a_{j+1}) - F(a_j)}\int_{F(a_j)}^{F(a_{j+1})} \psi^2(u)du\right] \left(F(a_{j+1}) - F(a_j)\right) \\
&= \sum_{j=1}^d \int_{F(a_j)}^{F(a_{j+1})} \psi^2(u)du \\
&= \int_0^1 \psi^2(u)du.
\end{aligned}$$

Therefore we obtained that

$$L(G) = \frac{\sum_{j=1}^d \Delta_j^{*2}\left(F(a_{j+1}) - F(a_j)\right)}{\int_0^1 \psi^2(u)du} \le 1.$$

We note that as $\max_j \left(F(a_{j+1}) - F(a_j)\right) \to 0$,

$$\sum_{j=1}^d \Delta_j^{*2}\left(F(a_{j+1}) - F(a_j)\right) \to \int_0^1 \psi^2(u)du.$$

Therefore as $\max_j \left(F(a_{j+1}) - F(a_j)\right) \to 0$, we see that

$$L(G) \to 1.$$

We note that the non-centrality parameter (5.2) can be used to obtain the asymptotic relative efficiency (ARE) of the test based on $\boldsymbol{M}_n$ with other tests, which we want to compare their performance. Also we note that the ARE will be determined by the patterns of the partition of the interval as well as the distribution functions.

6. Some concluding remarks

First of all, we discuss briefly to obtain the statistic, KW_n and derive the asymptotic normality for KW_n . For the simplicity of our arguments, we only consider the case that $K = 3$. For the extension to the case that $K > 3$, nothing new is involved except the notational complexity. If we express the exponent of a bivariate normal distribution with any two T_{kn} and T_{mn} among three components, T_{1n} , T_{2n} and T_{3n} , we have

$$-\frac{1}{2(1-\rho^2)}\left[\frac{(T_{kn}/n_k-\bar{\Delta}_n)^2}{V_0(T_{kn}/n_k)}-2\rho\frac{T_{kn}/n_k-\bar{\Delta}_n}{\sqrt{V_0(T_{kn}/n_k)}}\frac{T_{mn}/n_m-\bar{\Delta}_n}{\sqrt{V_0(T_{mn}/n_m)}}+\frac{(T_{mn}/n_m-\bar{\Delta}_n)^2}{V_0(T_{mn}/n_m)}\right],$$

where

$$\rho=-\sqrt{\frac{n_k}{n-n_k}\frac{n_m}{n-n_m}}.$$

Then by multiplying the above exponent by -2 and manipulating the algebraic structure with the following facts

$$T_{1n}+T_{2n}+T_{3n}=n\bar{\Delta}_n \text{ and } n=n_1+n_2+n_3,$$

we obtain (3.2) with 3 in place of K . It is well-known that the exponent of any bivariate normal distribution multiplied by -2 has a chi-square distribution with 2 degrees of freedom. Therefore it follows immediately that the asymptotic distribution of KW_n is a chi-square distribution with 2 degrees of freedom when $K = 3$. Also we note that we have obtained the same value for $\boldsymbol{M}_n$ and KW_n for the example in the previous section. Therefore one may conjecture that $\boldsymbol{M}_n$ and KW_n are equivalent. From this, our procedure can be considered as generalization of the Kruskal-Wallis test in the aspects to improve the power of test as well as to be applied to the grouped data.

We applied the Pearson's chi-square test as a nonparametric procedure to the example in order to test $H_0 : F_1 = F_2 = F_3$ and used the table of chi-square distribution with 12 degrees of freedom. Therefore the asymptotic distribution of the Pearson's chi- square statistic depends on the number of the sub-intervals as well as the number of samples. However we note that the asymptotic distribution of our test statistic is completely independent of the number of sub-intervals. This may be an advantage of our procedure.

7. Appendix

In this appendix, we derive the expression of $Cov_0(T_{kn}, T_{mn})$, the covariance between the two components, T_{kn} and T_{mn} . For this, we use the permutational arguments. Since

$$T_{kn}=\sum_{i=1}^{n_k}\sum_{j=1}^{d}\Delta_{nj}Z_{kij}=\sum_{j=1}^{d}\Delta_{nj}\sum_{i=1}^{n_k}Z_{kij}$$

we have that for any $k \neq m$,

$$
\begin{aligned}
&E_0\left[\left(\sum_{j=1}^{d}\Delta_{nj}\sum_{i=1}^{n_k}Z_{kij}\right)\left(\sum_{g=1}^{d}\Delta_{ng}\sum_{h=1}^{n_m}Z_{mhg}\right)\right]\\
&=\sum_{i=1}^{n_k}\sum_{h=1}^{n_m}\sum_{j=1}^{d}\sum_{g=1}^{d}\Delta_{nj}\Delta_{ng}E_0(Z_{kij}Z_{mhg})\\
&=\sum_{i=1}^{n_k}\sum_{h=1}^{n_m}\sum_{j=1}^{d}\Delta_{nj}^2E_0(Z_{kij}Z_{mhj})+\sum_{i=1}^{n_k}\sum_{h=1}^{n_m}\sum\sum\nolimits_{j\neq g}\Delta_{nj}\Delta_{ng}E_0(Z_{kij}Z_{mhg})\\
&=\sum_{i=1}^{n_k}\sum_{h=1}^{n_m}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}\frac{n_{\bullet j}-1}{n-1}\right\}+\sum_{i=1}^{n_k}\sum_{h=1}^{n_m}\left\{\sum\sum\nolimits_{j\neq g}\Delta_{nj}\Delta_{ng}\frac{n_{\bullet j}}{n}\frac{n_{\bullet m}}{n-1}\right\}\\
&=n_kn_m\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}\frac{n_{\bullet j}-1}{n-1}\right\}+n_kn_m\left\{\sum\sum\nolimits_{j\neq g}\Delta_{nj}\Delta_{ng}\frac{n_{\bullet j}}{n}\frac{n_{\bullet m}}{n-1}\right\}\\
&=n_kn_m\frac{n}{n-1}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}\frac{n_{\bullet j}-1}{n}\right\}+n_kn_m\frac{n}{n-1}\left\{\sum\sum\nolimits_{j\neq g}\Delta_{nj}\Delta_{ng}\frac{n_{\bullet j}}{n}\frac{n_{\bullet m}}{n}\right\}\\
&=n_kn_m\frac{n}{n-1}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}\frac{n_{\bullet j}-1}{n}\right\}+n_kn_m\frac{n}{n-1}\left[\left\{\sum_{j=1}^{d}\Delta_{nj}\frac{n_{\bullet j}}{n}\right\}^2-\sum_{j=1}^{d}\Delta_{nj}^2\left(\frac{n_{\bullet j}}{n}\right)^2\right]\\
&=n_kn_m\frac{n}{n-1}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}\frac{n_{\bullet j}-1}{n}\right\}+n_kn_m\frac{n}{n-1}\left[\bar{\Delta}_n^2-\sum_{j=1}^{d}\Delta_{nj}^2\left(\frac{n_{\bullet j}}{n}\right)^2\right]\\
&=n_kn_m\frac{n}{n-1}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\left(\frac{n_{\bullet j}}{n}\right)^2-\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n^2}\right\}+n_kn_m\frac{n}{n-1}\left[\bar{\Delta}_n^2-\sum_{j=1}^{d}\Delta_{nj}^2\left(\frac{n_{\bullet j}}{n}\right)^2\right]\\
&=-\frac{n_kn_m}{n-1}\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\bullet j}}{n}+n_kn_m\frac{n}{n-1}\bar{\Delta}_n^2.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
Cov_0(T_{kn}, T_{mn}) &= E_0(T_{kn}T_{mn}) - E_0(T_{kn})E_0(T_{mn}) \\
&= E_0\left[\left(\sum_{j=1}^{d}\Delta_{nj}\sum_{i=1}^{n_k}Z_{kij}\right)\left(\sum_{g=1}^{d}\Delta_{ng}\sum_{h=1}^{n_m}Z_{mhg}\right)\right] \\
&\qquad - E_0\left[\sum_{j=1}^{d}\Delta_{nj}\sum_{i=1}^{n_k}Z_{kij}\right]E_0\left[\sum_{g=1}^{d}\Delta_{ng}\sum_{h=1}^{n_m}Z_{mhg}\right] \\
&= -\frac{n_k n_m}{n-1}\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\cdot j}}{n} + n_k n_m\frac{n}{n-1}\bar{\Delta}_n^2 - n_k n_m\bar{\Delta}_n^2 \\
&= -\frac{n_k n_m}{n-1}\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\cdot j}}{n} + \frac{n_k n_m}{n-1}\bar{\Delta}_n^2 \\
&= -\frac{n_k n_m}{n-1}\left\{\sum_{j=1}^{d}\Delta_{nj}^2\frac{n_{\cdot j}}{n} - \bar{\Delta}_n^2\right\}.
\end{aligned}
$$

References

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical statistics, basic ideas and selected topics,* Holden-Day, Inc., San Francisco.

Billingsley, P. (1985). *Probability and measure,* 2nd Ed., John Wiley and Sons, Inc., New York.

Brookmeyer, R and Crowley, J. (1982). A k-sample median test for censored data. *Journal of American Statistical Association*, **77**, 433-440.

Chung, K. L. (1974). *A course in probability theory*, 2nd Ed., Academic Press, New York.

Good, P. (2000). *Permutation tests - A practical guide to resampling methods for testing hypotheses*, Springer, New York.

Hasselblad, V., Stead, A. G. and Galke, W. (1980). Analysis of coarsely grouped data from the lognormal distribution. *Journal of American Statistical Association*, **75**, 771-778.

Heitjan, D. F. (1989). Inference from grouped continuous data: a review. *Statistical Science*, **4**, 164-183.

Johnson, N. L. and Katz, S. (1970). *Distribution in Statistics-Continuous univariate distributions-2,* Houghton Mifflin Co., Boston.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of American Statistical Association*, **47**, 583-621.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate analysis*, Academic Press, New York.

Park, H. I. (1993). Nonparametric rank-order tests for right censored and grouped data in linear model. *Communications in Statistics*, **22**, 3143-3158.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric methods in multivariate analysis,* John Wiley and Sons, Inc., New York.

Puri, M. L. and Sen, P. K. (1985). *Nonparametric methods in general linear model,* John Wiley and Sons, Inc., New York.

Schott, J. R. (1997). *Matrix analysis for statistics,* Wiley and Sons, Inc., New York.