STA 104 Applied Nonparametric Statistics

Chapter 4: One-Way Layout Problems: Nonparametric One-Way Analysis of Variance

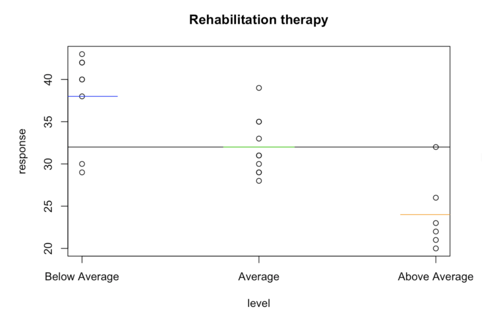
Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents

One-Way Data Layout

Treatments	Observations	Sample Sizes
1	$X_{11}, X_{12}, \ldots, X_{1n_1}$	n_1
2	$X_{11}, X_{12}, \ldots, X_{1n_2}$	n_2
$\ldots$	$\ldots$	$\ldots$
k	$X_{11}, X_{12}, \ldots, X_{1n_k}$	n_k



Setting

The data consist of $N = \sum_{i=1}^{k} n_j$ observations, with n_i observations from the ith treatment, $i = 1, \ldots, k$.

- For each treatment group $i \in \{1, \ldots, k\}$, the n_i observations are a random sample from a continuous distribution with distribution function F_i.
- The N observations are mutually independent.
- The distribution functions $F_1, \ldots, F_k$ are connected through the relationship

$$F_i(t) = F\left(t - \tau_i\right), -\infty < t < \infty,$$

for $i = 1, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ and τ_i is the unknown treatment effect for the i th population.

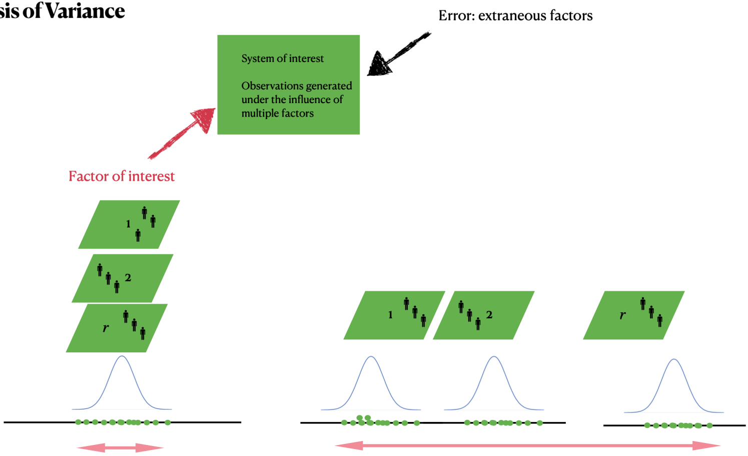
This is the usual one-way layout model: One-Way Analysis of Variance (ANOVA), commonly associated with normal assumptions:

$$X_{ij} = \theta + \tau_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,$$

where

- θ is the overall median,
- τ_i is the treatment i effect,
- the noise e_{ij}s are a random sample from a continuous distribution with median 0. (Under the additional assumption of normality, the medians θ and 0 are, of course, also the respective means.)

Analysis of Variance



Error: extraneous factors

System of interest

Observations generated under the influence of multiple factors

Factor of interest

Without factor of interest, the observations have some natural variation due to other extraneous factors, i.e. “error variance”

If the factor of interest indeed has some effects on the system, then we would expect more volatility than a system without the factor

Hypothesis

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \underbrace{\tau_1 \ldots \tau_k \text{ not all equal}}_{\text{at least two of the treatment effects are not equal}}$$

Review of One-Way ANOVA

The sum of squares for treatments is defined as

$$\mathrm{SST} = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}\right)^2$$

where $\bar{X}$ is the mean of all the observations-namely,

$$\bar{X} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_j} X_{ij}}{N}$$

The mean squares for treatment is

$$\mathrm{MST} = \frac{\mathrm{SST}}{k-1}$$

The sum of squares for error is defined as

$$\mathrm{SSE} = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

and the mean squares for error is

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-k}$$

Review of One-Way ANOVA

The F statistic is given by

$$F = \frac{\mathrm{MST}}{\mathrm{MSE}}$$

In ANOVA course, we learn that: If the observations are selected at random from normally distributed populations with equal variances, then this statistic has an F-distribution with $k - 1$ degrees of freedom for the numerator and $N - k$ degrees of freedom for the denominator. One may use this distribution to determine a p-value for the observed statistic and therefore conduct hypothesis testing.

However, if we are unwilling to assume that the population distributions are normal or the normally assumption is fundamentally wrong for the data at hand?

$\Rightarrow$ Nonparametric ANOVA

The Kruskal-Wallis Test

Hypothesis

Two-Sided Test:

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \underbrace{\tau_1 \ldots \tau_k \text{ not all equal}}_{\text{at least two of the treatment effects are not equal}}$$

Motivation

A way to obtain a nonparametric rank test for comparing k treatments is to replace the original observations with ranks and then perform the permutation F-test on these ranks.

We will obtain a statistic that is equivalent to the F statistic applied to ranks, with a permutation distribution that may be approximated by the chi-square distribution with $k - 1$ degrees of freedom.

Combine all N observations from the k samples, order them from least to greatest:

Data Layout for Ranks

Treatments	Ranks	Sample Size	Means
1	$R_{11}, R_{12}, \ldots, R_{1n_1}$	n_1	$R_{1.}$
2	$R_{11}, R_{12}, \ldots, R_{1n_2}$	n_2	$R_{2.}$
...	...	...	...
k	$R_{11}, R_{12}, \ldots, R_{1n_k}$	n_k	$R_{k.}$

$$\Rightarrow \qquad \underbrace{R_i = \sum_{j=1}^{n_i} R_{ij}}_{\text{sum of joint ranks received by treatment i observations}}$$

$$\Rightarrow \qquad \underbrace{R_i. = \frac{R_i}{n_i}}_{\text{average of joint ranks received by treatment i observations}}$$

$\Rightarrow$ Under H_0, rank vector $(R_{11}, R_{12}, \ldots, R_{1n_1}; \ldots R_{11}, R_{12}, \ldots, R_{1n_k})$ has a uniform distribution over the set of all $N!$ permutations of the rank $(1, 2, \ldots N)$

$\Rightarrow$

$$\begin{aligned} E_0(R_i) &= E_0(\sum_{j=1}^{n_i} R_{ij}) \\ &= \sum_{j=1}^{n_i} E_0(R_{ij}) \\ &= \sum_{j=1}^{n_i} \frac{\frac{N(N+1)}{2}}{N} \\ &= \sum_{j=1}^{n_i} \frac{N+1}{2} \\ &= n_i \frac{N+1}{2} \end{aligned}$$

$\Rightarrow$

$$E_0(R_{i.}) = \frac{N+1}{2}$$

we would expect average rank sum to be close to the expected value when H_0 is true.

$\Rightarrow$ Kruskal-Wallis statistics: [1] [2]

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(R_{i.} - \frac{N+1}{2} \right)^2$$
$$= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1)$$

$\Rightarrow$ The KS test statistic H is a constant times a weighted sum of squared differences between the observed treatment average ranks, $R_{i.}$, and their null expected values, $(N+1)/2$

- small values of H represent agreement with H_0
- When the treatment effects τ_i 's are not all equal, we would expect a portion of the associated treatment average ranks to differ from their common null expectation, with some tending to be larger and some smaller. The net result (after squaring the observed differences) would be a large value of H.
- This suggests rejecting H_0 in favor of H_1 for large values of H.

[1] The Kruskal-Wallis test can also be motivated by considering the usual analysis of variance $\mathcal{F}$ statistic calculated using the ranks, rather than the original observations. SSB reduces to $\sum_{j=1}^{k} n_j \left(R_{.j} - (N+1)/2 \right)^2$ when applied to the ranks rather than the original observations and SST becomes a fixed constant when calculated on the ranks. Using these facts, it can be shown that when $\mathcal{F}$ is calculated for the ranks, $\mathcal{F}$ is an increasing function of H.

[2] For the case of $k = 2$ treatments, Kruskal-Wallis test is equivalent to the two-sided Wilcoxon rank sum test.

Derivation of null distribution using permutation

Large sample approximation of null distribution

Define $T_i = R_{.j} - E_0\left(R_{.j}\right) = R_{.j} - (N+1)/2$, for $j = 1, 2, \ldots, k$. As each $R_j = \sum_{i=1}^{n_j} r_{ij}/n_j$ is an average, it is not surprising (see Kruskal and Wallis (1952), e.g., for justification) that a properly standardized version of the vector $\mathbf{T}^* = (T_1, \ldots, T_{k-1})$ has an asymptotic $(\min\left(n_1, \ldots, n_k\right)$ tending to infinity) $(k-1)$-variate normal distribution when the null hypothesis H_0 is true.

H is a quadratic form in the variables $(T_1, \ldots, T_{k-1})$, it is therefore quite natural that H has an asymptotic $(\min\left(n_1, \ldots, n_k\right)$ tending to infinity) chi-square distribution with $k-1$ degrees of freedom.

$$H \sim \chi^2_{k-1}$$

in large sample.

Procedure

Permutation

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal }],$$

at the α level of significance, Reject H_0 if $H \geq h_\alpha$; otherwise do not reject, where the constant h_α is chosen to make the type I error probability equal to α. The constant h_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of H.

Large-sample approximation

Reject H_0 if $H \geq \chi^2_{k-1,\alpha}$; otherwise do not reject, where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k - 1$ degrees of freedom.

Example:Length of YOY Gizzard Shad

To determine the number of game fish to stock in a given system and to set appropriate catch limits, it is important for fishery managers to be able to assess potential growth and survival of game fish in that system. Such growth and survival rates are closely related to the availability of appropriately sized prey. Young-of-year (YOY) gizzard shad (Dorosoma cepedianum) are the primary food source for game fish in many Ohio environments. However, because of their fast growth rate, YOY gizzard shad can quickly become too large for predators to swallow.

Thus it is useful to know both the size structure of the resident YOY shad populations. We want to assess whether there are any differences between the median lengths for the YOY gizzard shad populations in the four Kokosing Lake sites. With this in mind, Johnson (1984) sampled the YOY gizzard shad population at four different sites in Kokosing Lake (Ohio) in summer 1984.

Site I	Site II	Site III
29(5)	60(15)	33(8)
46(13)	32(7)	26(2)
37(9)	42(10)	25(1)
31(6)	45(12)	28(4)
44(11)	52(14)	27(3)
$R_1 = 44$	$R_2 = 58$	$R_3 = 18$

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(R_{i.} - \frac{N+1}{2} \right)^2 \\ &= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1) \\ &= \frac{12}{15(15+1)} \left(\frac{(44)^2}{5} + \frac{(58)^2}{5} + \frac{(18)^2}{5} \right) - 3(15+1) \\ &= 8.24 \end{aligned}$$

For the large-sample approximation:

```
> pchisq(4.85,df=2,lower.tail = F)
[1] 0.08847812
```

Hence, there is no strong evidence to reject the null hypothesis and there is no statistically significant differences between the median lengths for the YOY gizzard shad populations in the four Kokosing Lake sites.

Check with built-in function:

```
> library(NSM3)
> # permutation
> pKW(x=c(29,46,37,31,44,60,32,42,45,52,33,26,25,28,27),
+     g=c(rep(1,5),rep(2,5),rep(3,5)),method='Exact')
Group sizes: 5 5 5
Kruskal-Wallis H Statistic: 8.24
Exact upper-tail probability: 0.0077
> # large sample approximation
> pKW(x=c(29,46,37,31,44,60,32,42,45,52,33,26,25,28,27),
+     g=c(rep(1,5),rep(2,5),rep(3,5)),method='Asymptotic')
Group sizes: 5 5 5
Kruskal-Wallis H Statistic: 8.24
Asymptotic upper-tail probability: 0.0162
```

The Jonckheere-Terpstra Test for Ordered Alternatives

In many practical settings, the treatments are such that the appropriate alternatives to no differences in treatment effects (H_0) are those of increasing (or decreasing) treatment effects according to some natural labeling for the treatments. Examples of such settings include "treatments" corresponding to

- degrees of knowledge of performance,
- quality or quantity of materials,
- severity of disease,
- amount of practice, drug dosage levels,
- intensity of a stimulus and temperature.

We note that the Kruskal-Wallis test does not utilize any such partial prior information regarding a postulated alternative ordering. The statistic H takes on the same value for all $k!$ possible labelings of the treatments.

In this section, we consider a procedure for testing against the a priori ordered alternatives

Hypothesis

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$$

Motivation

Number of samples in uth treatment smaller than samples in $v$the treatment, i.e. Mann-Whitney counts

$$U_{uv} = \sum_{i}^{n_u} \sum_{j}^{n_v} 1\left(X_{ui} < X_{vj}\right) \quad 1 \leqslant u < v \leqslant k$$

Jonckheere-Terpstra statistics

$$J = \sum \sum_{1 \leq u \leq v \leq k} U_{uv}$$

takes the postulated ordering into account.

Motivation

Consider the case $k = 3$.

$J = \sum_{u=1}^{v-1} \sum_{v=2}^{3} U_{uv} = U_{12} + U_{13} + U_{23}$

if $\tau_1 < \tau_2 < \tau_3$:

- U_{12} would tend to be larger than $n_1 n_2/2$ (its null expectation);
- U_{13} would tend to be larger than $n_1 n_3/2$;
- U_{23} would tend to be larger than $n_2 n_3/2$;
- consequently, $J = U_{12} + U_{13} + U_{23}$ would tend to be larger than its null expectation $(n_1 n_2 + \ n_1 n_3 + n_2 n_3)/2$.

Derivation of null distribution using permutation

When H_0 is true, all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, and $\ldots, n_k$ ranks to the treatment k observations are equally likely.

$k = 3, n_1 = n_2 = 1, n_3 = 2$

Large sample approximation of null distribution

$$
\begin{aligned}
E(J) &= E\left[\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right] \\
&= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v} P\left(X_{iu} < X_{jv}\right) \\
&= \sum_{u=1}^{v-1}\sum_{v=2}^{k} n_u n_v P\left(X_{1u} < X_{1v}\right)
\end{aligned}
$$

Under the null hypothesis H_0, $P_0\left(X_{1u} < X_{1v}\right) = \frac{1}{2}$ for every $1 \leq u < v \leq k$. It follows that

$$
\begin{aligned}
E_0(J) &= \sum_{u=1}^{v-1}\sum_{v=2}^{k} \frac{(n_u n_v)}{2} = \frac{1}{4}\sum_{\substack{u=1\\u\neq v}}^{k}\sum_{v=1}^{k} n_u n_v \\
&= \frac{1}{4}\left[\sum_{u=1}^{k}\sum_{v=1}^{k} n_u n_v - \sum_{i=1}^{k} n_i^2\right] \\
&= \frac{1}{4}\left[N^2 - \sum_{i=1}^{k} n_i^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}(J) &= \mathrm{var}\left(\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right) \\
&= \sum_{u=1}^{v-1}\sum_{v=2}^{k} \mathrm{var}\left(U_{uv}\right) + \sum_{u=1}^{v-1}\sum_{\substack{v=2\\(u,v)\neq(s,t)}}^{k-1}\sum_{t=2}^{t-1} \mathrm{cov}\left(U_{uv}, U_{st}\right)
\end{aligned}
$$

Under H_0, it can be shown that

$$\begin{aligned}
\mathrm{var}_0\,(U_{uv}) &= \frac{n_u n_v\,(n_u + n_v + 1)}{12}, \quad \text{for } 1 \le u < v \le k, \\
\mathrm{cov}_0\,(U_{uv}, U_{st}) &= 0, \quad \text{for all distinct } u, v, s, t \text{ in } \{1, \ldots, k\} \\
\mathrm{cov}_0\,(U_{uv}, U_{ut}) &= \frac{n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v, t \le k, v \ne t \\
\mathrm{cov}_0\,(U_{uv}, U_{su}) &= \frac{-n_s n_u n_v}{12}, \quad \text{for } 1 \le s < u < v \le k \\
\mathrm{cov}_0\,(U_{uv}, U_{vt}) &= \frac{-n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v < t \le k \\
\mathrm{cov}_0\,(U_{uv}, U_{sv}) &= \frac{n_u n_v n_s}{12}, \quad \text{for } 1 \le u, s < v \le k, u \ne s
\end{aligned}$$

Combining the results, it follows after significant algebraic manipulation that

$$\mathrm{var}_0(J) = \frac{N^2(2N+3) - \sum_{i=1}^{k} n_i^2\,(2n_i + 3)}{72},$$

$$J^* = \frac{J - E_0(J)}{\{\mathrm{var}_0(J)\}^{1/2}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}\right]}{\left\{\left[N^2(2N+1) - \sum_{i=1}^{k} n_i^2\,(2n_i + 3)\right]/72\right\}^{1/2}} \sim N(0, 1)$$

in large sample, follows from the fact that J can be expressed as a sum of certain mutually independent combined-samples Mann-Whitney statistics and standard theory for such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Terpstra (1952)).

Combining the results, it follows after significant algebraic manipulation that

$$\mathrm{var}_0(J) = \frac{N^2(2N+3) - \sum_{i=1}^{k} n_i^2\,(2n_i+3)}{72},$$

$$J^* = \frac{J - E_0(J)}{\{\mathrm{var}_0(J)\}^{1/2}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}\right]}{\left\{\left[N^2(2N+1) - \sum_{i=1}^{k} n_i^2\,(2n_i+3)\right]/72\right\}^{1/2}} \sim N(0,1)$$

in large sample, follows from the fact that J can be expressed as a sum of certain mutually independent combined-samples Mann-Whitney statistics and standard theory for such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Terpstra (1952)).

Procedure

First, we must label the treatments so that they are in the expected order associated with the alternative.

Calculate the $k(k-1)/2$ Mann-Whitney counts U_{uv} given by

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi\left(X_{iu}, X_{jv}\right), \quad 1 \leq u < v \leq k$$

where $\phi(a, b) = 1$ if $a < b$, 0 otherwise.

The Jonckheere-Terpstra statistic J, is then the sum of these $k(k-1)/2$ Mann-Whitney counts,

$$J = \sum_{u=1}^{v-1} \sum_{v=2}^{k} U_{uv}$$

Procedure

Permutation

To test

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k \text{with at least one strict inequality}$$

at the α level of significance.

Reject H_0 if $J \geq j_\alpha$; otherwise do not reject, where the constant j_α is chosen to make the type I error probability equal to α. The constant j_α is the upper α percentile for the null distribution of J.

Large-sample approximation

Reject H_0 if $J^* \geq z_\alpha$; otherwise do not reject.

Notes

- The Jonckheere-Terpstra test are quite superior to the Kruskal-Wallis test when the conjectured ordering of the treatment effects ($\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$) is, indeed, appropriate. In addition, small violations in the conjectured ordering for τ_i and τ_j do not seriously affect the power of the Jonckheere-Terpstra tests if i and j correspond to treatment labels near the middle of the conjectured orderings. However, if i and j are both near 1 or k, the effect of such violations can be rather substantial.
- The Jonckheere (1954a, 1954b) and Terpstra (1952) test of this section is preferred to the Kruskal-Wallis test when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from null to be in the particular direction.
- We emphasize, however, that the labeling of the treatments so that the ordered alternatives are appropriate cannot depend on the observed sample observations. This labeling must correspond completely to a factor(s) implicit in the nature of the experimental design and not the observed data.

Example:Motivational Effect of Knowledge of Performance

Hundal (1969) described a study designed to assess the purely motivational effects of knowledge of performance in a repetitive industrial task. The task was to grind a metallic piece to a specified size and shape. Eighteen male workers were divided randomly into three groups. The subjects in the control group, A, received no information about their output, subjects in group B were given a rough estimate of their output, and subjects in group C were given an accurate information about their output and could check it further by referring to a figure that was placed before them. The basic data in Table 6.6 consist of the numbers of pieces processed by each subject in the experimental period.

We apply the Jonckheere-Terpstra test with the notion that a deviation from H_0 is likely to be in the direction of increased output with increased degree of knowledge of performance. Thus, we are interested in using procedure (6.14) with the treatment labels $1 \equiv$ control (no information), $2 \equiv$ group B (rough information), and $3 \equiv$ group C (accurate information). For purpose of illustration, we take the significance level to be $\alpha = .0490$.

Control (no information)	Group B (rough information)	Group C (accurate information)
39.5	37.5	48
35	40	40.5
38	47	45
42.5	44	43
44.5	41.5	46
41	42	50

$$
\begin{aligned}
U_{12} &= 5+6+5+2+1+4 = 23 \\
U_{13} &= 6+6+6+5+4+5 = 32 \\
U_{23} &= 6+6+2+4+5+5 = 28 \\
\Rightarrow J &= 23+31+27 = 83
\end{aligned}
$$

For the large-sample approximation:

$$J^* = 2.34451$$

```
> pnorm(2.34451,lower.tail = F)
[1] 0.00952605
```

Hence, there is strong evidence in support of increased output with increase in degree of knowledge of performance.

Check with built-in function:Agreed!

```
> library(NSM3)
> motivational.effect<-list(no.Info=c(39.5,35,38,42.5,44.5,41),
+                           rough.Info=c(37.5,40,47,44,41.5,42),
+                           accurate.Info=c(48,40.5,45,43,46,50))
> pJCK(motivational.effect,method=NA)
Group sizes: 6 6 6
Jonckheere-Terpstra J Statistic: 83
Exact upper-tail probability: 0.0095
> #pJCK(motivational.effect,method="Exact")
> #pJCK(motivational.effect,method="Monte Carlo",n.mc=10000)
> pJCK(motivational.effect,method="Asymptotic")
Group sizes: 6 6 6
Jonckheere-Terpstra J* Statistic: 2.3445
Asymptotic upper-tail probability: 0.0095
```

The Fligner-Wolfe Test for Treatments versus a Control

In this section, we discuss a test procedure specifically designed for the setting where one of the treatments corresponds to a control or baseline set of conditions and we are interested in assessing which, if any, of the treatments is better than the control.

Without loss of generality, we label the treatments so that the control corresponds to treatment 1. In this setting, the null hypothesis of interest is still the same, but now it corresponds to the statement that none of the treatments $2, \ldots, k$ is different from the control (treatment 1). This is usually expressed as

$$H_0 : [\tau_i = \tau_1, i = 2, \ldots, k] .$$

Hypothesis [3]

One-Sided Upper-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

$$H_1 : [\tau_i \geq \tau_1 \ , , \text{ for } i = 2, \ldots, k, \text{ with at least one strict inequality}]$$

One-Sided Lower-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

$$H_1 : [\tau_i \leq \tau_1 \ , , \text{ for } i = 2, \ldots, k, \text{ with at least one strict inequality}]$$

[3] We do not discussed a FW test designed for a two-sided alternative. The "natural" two-sided alternative for this treatment versus control setting corresponds to [either $\tau_i \geq \tau_1$ for all $i = 2, \ldots, k$ or $\tau_i \leq \tau_1$ for all $i = 2, \ldots, k$, with at least one strict inequality]. We feel that it is rather unlikely that we would find ourselves in such a setting where either all the treatments are better than the control or all the treatments are worse than the control, but we have no idea which of the two cases pertains.

Motivation

First combine all N observations from the k samples and order them from least to greatest.

$\Rightarrow$ Letting R_{ij} denote the rank of X_{ij} in this joint ranking, the Fligner-Wolfe statistic FW is then the sum of these joint ranks for the non-control treatments,

$$\mathrm{FW} = \sum_{j=2}^{k} \sum_{i=1}^{n_j} R_{ij}$$

$\Rightarrow$ When some of the τ_i 's are strictly greater than the control effect τ_1, we would expect the joint ranks for the observations from those treatments to be larger than the joint ranks for the control observations. The net result would be a larger value of FW. This suggests rejecting H_0 in favor of H_1 for large values of FW .

Derivation of null distribution using permutation

FW can be viewed as a two-sample Wilcoxon rank sum statistic computed for the $m = n_1$ control treatment observations (playing the role of the X 's in the two sample setting) and the $n = \sum_{j=2}^{k} n_j$ combined observations from treatments $2, \ldots, k$ (playing the role of the Y 's in the two-sample setting).

As a result, the null distribution of FW is the same as that of the Wilcoxon rank sum statistic with sample sizes m, n.

Thus, the critical value f_α is just the upper α th percentile w_α for the null distribution of the Wilcoxon rank sum statistic with sample sizes m, n.

Large sample approximation of null distribution

The statistic FW has the same probability distribution as the null distribution of the two-sample Wilcoxon rank sum statistic W with sample sizes m, n. Hence, it follows directly from the Large-Sample Approximation for two-sample Wilcoxon rank sum statistic

$$\mathrm{FW}^* = \frac{\mathrm{FW} - \frac{n(N+1)}{2}}{\sqrt{mn(N+1)/12}} \sim N(0,1)$$

has, as $\min(n_1, N^*)$ tends to infinity, an asymptotic $N(0,1)$ distribution when H_0 is true.

Procedure

Permutation

To test

One-Sided Upper-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

$$H_1 : [\tau_i \geq \tau_1 \text{ , , for } i = 2, \ldots, k, \text{ with at least one strict inequality}]$$

at the α level of significance, Reject H_0 if $\mathrm{FW} \geq f_\alpha$; otherwise do not reject.

Large-sample approximation

Reject H_0 if $\mathrm{FW}^* \geq z_\alpha$; otherwise do not reject.

Procedure

Permutation

To test

One-Sided Lower-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

$$H_1 : [\tau_i \leq \tau_1 \;,, \text{ for } i = 2, \ldots, k, \text{ with at least one strict inequality}]$$

at the α level of significance, Reject H_0 if $\mathrm{FW} \leq f_{1-\alpha}$; otherwise do not reject.

Large-sample approximation

Reject H_0 if $\mathrm{FW}^* \leq -z_\alpha$; otherwise do not reject.

Remarks

The test deal with very restricted alternatives where all the treatments are either at least as good as the control or all the treatments are no better than the control, respectively. They are not appropriate tests when the possibility exists that some of the treatments might be better and some might be worse than the control. For such mixed alternatives, one would need to use the general alternatives Kruskal-Wallis test.

In many settings where we are interested in comparing a number of treatments with a control, we will have additional a priori information regarding the relative magnitude of the treatment effects. In a drug development, for instance, increasing dosage levels may be compared with a zero-dose control. If the treatment effects are not identical to that of the control, then it is often reasonable to assume that the higher the dose of the drug applied, the better (say, higher) will be the resulting effect on a patient, corresponding to monotonically ordered treatment effects.

However, it may also be the case that a subject might potentially succumb to toxic effects at high doses, thereby actually decreasing the associated treatment effects. Such a setting would correspond to an ordering in the treatment effects that is monotonically increasing up to a point, followed by a monotonic decrease; that is, an umbrella pattern on the treatment. This is a research topic very important for pharmaceutical industry – how to find the best dosage?

Example:Motivational Effect of Knowledge of Performance

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, the no information category clearly serves as a control population, and it is very natural to ask if additional performance information of either type (rough or accurate) leads to improved performance as measured by an increase in the number of pieces processed.

Control (no information)	Group B (rough information)	Group C (accurate information)
39.5	37.5	48
35	40	40.5
38	47	45
42.5	44	43
44.5	41.5	46
41	42	50

```
> sum(rank(c(39.5,35,38,42.5,44.5,41,37.5,40,47,44,41.5,42,48,40.5,45,43,46,50)
[1] 133
```

$$
\begin{aligned}
FW &= 133 \\
\mathrm{FW}^* &= \frac{\mathrm{FW} - \frac{n(N+1)}{2}}{\sqrt{mn(N+1)/12}} \\
&= \frac{133 - 114}{\{112.12\}^{1/2}} = 1.79437
\end{aligned}
$$

```
> pnorm(1.79437,lower.tail = F)
[1] 0.03637707
```

Thus, we have sufficient evidence from the Fligner-Wolfe treatments-versus-control test that additional performance knowledge (either rough or accurate) leads to an increase in the number of pieces produced.

Multiple Comparisons

- We have discussed procedures designed to test the null hypothesis of no difference in treatment groups against a variety of alternative hypotheses. Upon rejection of H_0 with one of these test procedures for a given set of data, our conclusions range from the general statement that there are some unspecified differences among the treatment effects (associated with the Kruskal-Wallis test) to the more informative relationships between the treatment effects associated with test procedures designed for the ordered alternatives or the treatments-versus-control setting.
- However, in none of these test procedures are our conclusions specific or pair-specific; that is, the tests are not designed to enable us to reach conclusions about specific pairs of treatment effects, such as which specific treatments are better than the control.

Rationale for Multiple Comparison Procedures.

- To elicit such pairwise specific information, we turn to the class of multiple comparison procedures.
- The aim of applying such procedures goes beyond the point of deciding whether the treatments are equivalent to the (often more important) problem of selecting which, if any, treatments differ from one another. Thus, the user makes $k(k-1)/2$ decisions, one for each pair of treatments.

Rationale for Multiple Comparison Procedures.

- The multiple comparison procedure is designed so that the Experimentwise Error Rate is controlled to be equal to α; that is, the probability of falsely declaring any pair of treatment effects to be different, when in fact all of the treatment effects are the same, is equal to α.
- The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1 - \alpha$ when the hypothesis H_0) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply such techniques when we have evidence (rejecting the Kruskal-Wallis test) that H_0 is not true.
- This protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when in fact H_0 is false, and this difficulty becomes more severe as k increases.

Two-Sided All-Treatments Multiple Comparisons for General Alternative

After rejection of

$$H_0 : \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F}$$

$$H_1 : \qquad \underbrace{\tau_1 \ldots \tau_k \text{ not all equal}}_{\text{at least two of the treatment effects are not equal}}$$

with the Kruskal-Wallis test, it is important to reach conclusions about exactly which treatment is different from which treatment, that is, all $\left(\begin{array}{c} k \\ 2 \end{array}\right) = k(k-1)/2$ individual differences between pairs of treatment effects (τ_i, τ_j), for $i < j$, and these conclusions are naturally two-sided in nature.

Hypothesis

$$\left\{\begin{array}{ll} H_0 : \tau_1 = \tau_2 & H_1 : \tau_1 \neq \tau_2 \\ H_0 : \tau_1 = \tau_3 & H_1 : \tau_1 \neq \tau_3 \\ \ldots & \\ H_0 : \tau_{k-1} = \tau_k & H_1 : \tau_{k-1} \neq \tau_k \end{array}\right\} \frac{k(k-1)}{2} \text{ simultaneous tests/multiple}$$

comparisons

Motivation

⇒ For each pair of treatments (i, j), for $1 \leq i < j \leq k$, let

$$W_{ij} = \sum_{b=1}^{n_j} R_{jb}$$

where R_{jb} are the ranks of X_{jb} among the combined i th and j th samples; that is, W_{ij} is the Wilcoxon rank sum of the j th sample ranks in the joint two-sample ranking of the i th and j th sample observations.

⇒ standardized (under H_0) version of W_{ij} multiplied by $\sqrt{2}$

$$W_{ij}^{*} = \sqrt{2}\left[\frac{W_{ij} - E_0\left(W_{ij}\right)}{\left\{\operatorname{var}_0\left(W_{ij}\right)\right\}^{1/2}}\right] = \frac{W_{ij} - \frac{n_i\left(n_i+n_j+1\right)}{2}}{\left\{n_i n_j\left(n_i+n_j+1\right)/24\right\}^{1/2}}, \quad \text{for } 1 \leq i < j \leq k.$$

$\Rightarrow$

- When H_0 is true, the $[k(k-1)/2]$-component vector $\left(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k}\right)$ has, as $\min\left(n_1, \ldots, n_k\right)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$.
- and when $n_1 = n_2 = \cdots = n_k$,

$$\left(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k}\right) \sim (\ldots Z_i - Z_j \ldots)$$

  the $[k(k-1)/2]$-component vector of differences where $Z_1 \ldots Z_k$ are indpendent $N(0,1)$

$\Rightarrow$

$$\max_{1 \leq i < j \leq k} \left|W^*_{ij}\right| \sim range(Z_1 \ldots Z_k)$$

$\Rightarrow$ To get the null distribution for the simultaneous tests, it is equivalent to know the null distribution of $\max_{1 \leq i < j \leq k} \left|W^*_{ij}\right|$

$\Rightarrow$ It is then equivalent to the distribution of the range when we draw k independent $N(0,1)$

Procedure

For each pair of treatments (i, j), for $1 \leq i < j \leq k$,

Decide $\tau_i \neq \tau_j$ if $\left| W^*_{ij} \right| \geq q_\alpha$; otherwise decide $\tau_u = \tau_v$.

q_α is the upper α quantile of the range of k normal variates.

Example: Length of YOY Gizzard Shad

Site I	Site II	Site III
29(5)	60(15)	33(8)
46(13)	32(7)	26(2)
37(9)	42(10)	25(1)
31(6)	45(12)	28(4)
44(11)	52(14)	27(3)

```
> library(NSM3)
> cRangeNor(0.1,k=3)
[1] 2.903
```

Decide $\tau_u \neq \tau_v$ if $|W^*_{uv}| \geq 2.903$.

$$
\begin{aligned}
W^*_{12} &= \frac{[34 - 5(11)/2]}{\sqrt{5 \times 5 \times 11/24}} = 1.92 \\
W^*_{13} &= \frac{[17 - 5(11)/2]}{\sqrt{5 \times 5 \times 11/24}} = -3.10 \\
W^*_{23} &= \frac{[16 - 5(11)/2]}{\sqrt{5 \times 5 \times 11/24}} = -3.397
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow |W^*_{12}| = 1.92 < 2.903 \quad &\Longrightarrow \quad \text{decide } \tau_1 = \tau_2 \\
|W^*_{13}| = 3.10 > 2.903 \quad &\Longrightarrow \quad \text{decide } \tau_1 \neq \tau_3 \\
|W^*_{23}| = 3.397 > 2.903 \quad &\Longrightarrow \quad \text{decide } \tau_2 \neq \tau_3
\end{aligned}
$$

Thus, at an experimentwise error rate of .05, the multiple comparison decisions can be summarized by the statement $(\tau_1 = \tau_2) \neq (\tau_3)$.

This multiple comparison procedure provides more detailed information about the lengths of the YOY gizzard shad population in Kokosing Lake. We now know that sites I and II may be viewed as providing similar living environments for gizzard shad. However, we also know that the common living environment at sites I and II is significantly different from the common living environment at sites III.

One-Sided All-Treatments Multiple Comparisons for Ordered Treatment Effects Alternatives

After rejection of

$$
\begin{aligned}
H_0 :& \underbrace{\tau_1 = \ldots = \tau_k}_{F_1 = F_2 \ldots = F_k \equiv F} \\
H_1 :& \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k
\end{aligned}
$$

with the Jonckheere-Terpstra test, it is important to reach conclusions about exactly which $\leq$ is $<$ as opposed to $=$, that is, whether there is strict ordering among all $\left(\begin{array}{c} k \\ 2 \end{array} \right) = k(k-1)/2$ individual differences between pairs of treatment effects $\left(\tau_i, \tau_j\right)$, for $i < j$, and these conclusions are naturally one-sided, in accordance with the ordered alternatives setting.

Hypothesis

$$\left\{\begin{array}{ll} H_0 : \tau_1 = \tau_2 & H_1 : \tau_1 < \tau_2 \\ H_0 : \tau_1 = \tau_3 & H_1 : \tau_1 < \tau_3 \\ \dots & \\ H_0 : \tau_{k-1} = \tau_k & H_1 : \tau_{k-1} < \tau_k \end{array}\right\} \tfrac{k(k-1)}{2} \text{ simultaneous tests/multiple}$$

comparisons

Motivation

$\Rightarrow$ For each pair of treatments (i, j), for $1 \leq i < j \leq k$, let

$$W_{ij} = \sum_{b=1}^{n_j} R_{jb}$$

where R_{jb} are the ranks of X_{jb} among the combined i th and j th samples; that is, W_{ij} is the Wilcoxon rank sum of the j th sample ranks in the joint two-sample ranking of the i th and j th sample observations.

$\Rightarrow$ standardized (under H_0) version of W_{ij} multiplied by $\sqrt{2}$

$$W_{ij}^{*} = \sqrt{2}\left[\frac{W_{ij} - E_0\left(W_{ij}\right)}{\left\{\operatorname{var}_0\left(W_{ij}\right)\right\}^{1/2}}\right] = \frac{W_{ij} - \frac{n_i\left(n_i+n_j+1\right)}{2}}{\left\{n_i n_j\left(n_i+n_j+1\right)/24\right\}^{1/2}}, \quad \text{for } 1 \leq i < j \leq k.$$

Motivation

$\Rightarrow$

- When H_0 is true, the $[k(k-1)/2]$-component vector $\left(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*\right)$ has, as $\min\left(n_1, \ldots, n_k\right)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$, and

$$\left(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*\right) \sim (\ldots \frac{Z_j - Z_i}{\sqrt{\frac{n_i+n_j}{2n_i n_j}}} \ldots)$$

  the $[k(k-1)/2]$-component vector of differences where $Z_1, \ldots, Z_k$ are mutually independent and Z_i has an $N\left(0, 1/n_i\right)$ distribution

$\Rightarrow$

$$\max_{1\le i<j\le k} W_{ij}^* \sim \max_{1\le i<j\le k} \frac{Z_j - Z_i}{\sqrt{\frac{n_i+n_j}{2n_i n_j}}}$$

$\Rightarrow$ To get the null distribution for the simultaneous tests, it is equivalent to know the null distribution of $\max_{1\le i<j\le k} W_{ij}^*$

$\Rightarrow$ It is then equivalent to the distribution of the maximum difference when we draw k independent normal random variables with $N(0, 1/n_i)$

Procedure

For each pair of treatments (i, j), for $1 \leq i < j \leq k$,

Decide $\tau_i < \tau_j$ if $W_{ij}^* \geq d_\alpha$; otherwise decide $\tau_i = \tau_j$.

d_α is the upper α percentile of the maximum range of k normal variates with $N(0, 1/n_i)$.

Example: Motivational Effect of Knowledge of Performance

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, we found using the Jonckheere-Terpstra test that there was sufficient evidence in the sample data to conclude that $\tau_1 \leq \tau_2 \leq \tau_3$ with at least one strict inequality.

To examine which of the types of information (none, rough, or accurate) lead to differences in median numbers of pieces processed.

Control (no information)	Group B (rough information)	Group C (accurate information)
39.5	37.5	48
35	40	40.5
38	47	45
42.5	44	43
44.5	41.5	46
41	42	50

```
> library(NSM3)
> cHayStonLSA(alpha=0.05,k=3)
[1] 2.94
> cHaySton(.05,rep(6,3),method=NA)

Monte Carlo Approximation (with  10000  Iterations) used:

Group sizes: 6 6 6
For the given experimentwise alpha=0.05, the upper cutoff value is
Hayter-Stone W*=3.1703756956, with true experimentwise alpha level=0.0367
```

```
> sum(rank(c(39.5,35,38,42.5,44.5,41,37.5,40,47,44,41.5,42))[7:12])
[1] 44
> sum(rank(c(39.5,35,38,42.5,44.5,41,48,40.5,45,43,46,50))[7:12])
[1] 53
> sum(rank(c(37.5,40,47,44,41.5,42,48,40.5,45,43,46,50))[7:12])
[1] 49
```

$$
\begin{aligned}
W_{12}^* &= \frac{[44 - 6(13)/2]}{\sqrt{6 \times 6 \times 13/24}} = 1.132277 \\
W_{13}^* &= \frac{[53 - 6(13)/2]}{\sqrt{6 \times 6 \times 13/24}} = 3.170376 \\
W_{23}^* &= \frac{[49 - 6(13)/2]}{\sqrt{6 \times 6 \times 13/24}} = 2.264554
\end{aligned}
$$

$$
\begin{array}{rcl}
\Rightarrow |W_{12}^*| < 2.94 & \Longrightarrow & \text{decide } \tau_1 = \tau_2 \\
|W_{13}^*| > 2.94 & \Longrightarrow & \text{decide } \tau_1 < \tau_3 \\
|W_{23}^*| < 2.94 & \Longrightarrow & \text{decide } \tau_2 = \tau_3
\end{array}
$$

Thus, at an experimentwise error rate of .05, we have reached the conclusion that $\tau_1 < \tau_3$ but$\tau_1 = \tau_2$ and $\tau_2 = \tau_3$.

One-Sided Treatments-versus-Control Multiple Comparisons for Treatment-versus-Control Alternatives

When the main interest is on treatment-versus-control comparisons, we do not compare all treatments, but only each noncontrol treatment with the control on a directional bias. This situation arises, for example, in drug screening in the examination of many new treatments in hopes of improving on a standard, and there is no initial reason to perform between treatment comparisons. Of course, comparisons could be carried out later between treatments that were selected as being better, if there is intention to pick the optimal one.

After rejection of

One-Sided Upper-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \dots, k]$$

$$H_1 : [\tau_i \geq \tau_1 \ , , \text{ for } i = 2, \dots, k, \text{ with at least one strict inequality}]$$

One-Sided Lower-Tail Test:

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \dots, k]$$

$$H_1 : [\tau_i \leq \tau_1 \ , , \text{ for } i = 2, \dots, k, \text{ with at least one strict inequality}]$$

with the Fligner-Wolf test, it is important to reach conclusions about exactly which treatment is better than control, and these conclusions are naturally one-sided, in accordance with the directional alternatives setting.

Hypothesis

One-Sided Lower-Tail Test:

$$\left\{\begin{array}{ll} H_0 : \tau_1 = \tau_2 & H_1 : \tau_2 > \tau_1 \\ H_0 : \tau_1 = \tau_3 & H_1 : \tau_3 > \tau_1 \\ \dots & \\ H_0 : \tau_{k-1} = \tau_k & H_1 : \tau_k > \tau_1 \end{array}\right\} k-1 \text{ simultaneous tests/multiple comparisons}$$

Motivation

$\Rightarrow$ Jointly rank all N of the sample observations and let $R_{1.}, \ldots, R_{k.}$ be the averages of these joint ranks associated with treatments $1, \ldots, k$, respectively. (as in Kruskal-Wallis statistic.)

$\Rightarrow$ For each of the $k-1$ noncontrol treatments, calculate the difference $R_{i.} - R_{1.}, i = 2, \ldots, k$.

$\Rightarrow$ When H_0 is true, and $n_1 = b$ and $n_2 = \cdots = n_k = n$, with both n and b large: the $k-1$-component vector

$$(R_{2.} - R_{1.}, R_{3.} - R_{1.}, \ldots, R_{k.} - R_{1.}) \sim (Z_1 ... Z_{k-1})$$

$(k-1)N(0,1)$ variables with common correlation $\rho = n/(b+n)$

$\Rightarrow$

$$\max_{2 \leq i \leq k} R_{i.} - R_{1.} \sim \max_{1 \leq i \leq k-1} Z_i$$

$\Rightarrow$ To get the null distribution for the simultaneous tests, it is equivalent to know the null distribution of $\max_{1 \leq i \leq k-1} Z_i$, the maximum when we draw $k-1$ normal random variables with common correlation $\rho = n/(b+n)$.

Procedure

When $n_1 = b$ and $n_2 = \cdots = n_k = n$

For each treatments i,

$$\text{Decide } \tau_i > \tau_1 \text{ if } (R_{i.} - R_{1.}) \geq m^*_\alpha \left[\frac{N(N+1)}{12}\right]^{1/2} \left(\frac{1}{b} + \frac{1}{n}\right)^{1/2}$$

otherwise decide $\tau_u = \tau_1, u = 2, \ldots, k$.

- m^*_α is the α upper percentile of the $\max_{1 \leq i \leq k-1} Z_i$, the maximum when we draw $k-1$ normal random variables with common correlation $\rho = n/(b+n)$.

General setting: arbitrary sample sizes (Bonferroni's Inequality) [4]

For each treatments i,

$$\text{Decide } \tau_i > \tau_1 \text{ if } (R_{i.} - R_{1.}) \geq z_{\alpha^*} \left[\frac{N(N+1)}{12}\right]^{1/2} \left(\frac{1}{n_1} + \frac{1}{n_u}\right)^{1/2}$$

otherwise decide $\tau_u = \tau_1, u = 2, \ldots, k$.

- $\alpha^* = \alpha/(k-1)$

[4] Bonferroni's general approximate procedure can often be quite conservative in practice, as a direct result of the conservative nature of the Bonferroni Inequality.

Example: Motivational Effect of Knowledge of Performance

To further investigate which (if either) of the two types of additional information (rough or accurate) lead to improvement or increase in median numbers of pieces processed relative to the no information control (treatment 1).

Control (no information)	Group B (rough information)	Group C (accurate information)
39.5	37.5	48
35	40	40.5
38	47	45
42.5	44	43
44.5	41.5	46
41	42	50

```
# large-sample approximation
> cMaxCorrNor(alpha=0.05,k=2,rho=6/12)
[1] 1.91
> sqrt(18*19/12)*sqrt(1/6+1/6)*1.91
[1] 5.887015
```

Decide $\tau_u > \tau_i$ if $(R_u - R_1) \geq 5.88$.

```
> ranks=rank(c(39.5,35,38,42.5,44.5,41,37.5,40,47,44,41.5,42,48,40.5,45,43,46,5
> R1=mean(ranks[1:6])
> R2=mean(ranks[7:12])
> R3=mean(ranks[13:18])
> R2-R1
[1] 2.333333
> R3-R1
[1] 7.166667
```

$$
\begin{aligned}
(R_{2.} - R_{1.}) &= 2.3 < 5.88 \Rightarrow \text{ decide } \tau_2 = \tau_1, \\
(R_{3.} - R_{1.}) &= 7.1 \geq 5.88 \Rightarrow \text{ decide } \tau_3 > \tau_1.
\end{aligned}
$$

Thus at an experimentwise error rate of .05, we have reached the conclusion that accurate information leads to significantly more pieces processed than the no information control, while rough information do not lead to significant improvement compared to no information control.