## Contents

### 3 $K$-sample Methods

3.1 Setup .................................................. 2
3.2 Classic Method Based on Normality Assumption . . . . . 3
3.3 Permutation $F$-test ...................................... 5
3.4 Kruskal-Wallis Test ..................................... 8
3.5 Multiple Comparisons .................................. 16
  3.5.1 Motivation ........................................ 16
  3.5.2 Methods for Multiple Comparisons .............. 18
  3.5.3 Multiple Comparison Permutation Tests ......... 25
3.6 Ordered Alternatives .................................. 30
3 \( K \)-sample Methods

3.1 Setup

Extend the 2-sample methods to comparison of \( K \) groups, \( K \geq 2 \).

**Data:** \( X_{ij} \): \( j \)th observation from the \( i \)th treatment group, 
\( i = 1, \cdots, K, \ j = 1, \cdots, n_i \), where \( n_1, \cdots, n_K \) may not equal. Let 
\( N = \sum_{i=1}^{K} n_i \): total number of observations.

<table>
<thead>
<tr>
<th>Trt</th>
<th>1</th>
<th>2</th>
<th>\cdots</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( x_{11} )</td>
<td>( x_{21} )</td>
<td>( \cdots )</td>
<td>( x_{K1} )</td>
</tr>
<tr>
<td></td>
<td>( x_{12} )</td>
<td>( x_{22} )</td>
<td>( \cdots )</td>
<td>( x_{K2} )</td>
</tr>
<tr>
<td></td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td></td>
<td>( x_{1n_1} )</td>
<td>( x_{2n_2} )</td>
<td>( \cdots )</td>
<td>( x_{Kn_k} )</td>
</tr>
</tbody>
</table>
Hypotheses: test whether all observations are \textit{i.i.d.}, or whether treatments differ in locations. We focus on testing the shift in locations.

\textbf{One-way ANOVA model:}

\[ X_{ij} = \mu_i + \epsilon_{ij}, \]  

(3.1)

where \( \mu_i \) is the mean for treatment \( i \), \( \epsilon_{ij} \) are \textit{i.i.d.} random variables.

Test

\[ H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \]

versus \( H_a : \mu_i, i = 1, \cdots, K, \) are not all equal (that is, at least two treatments have unequal means).

3.2 Classic Method Based on Normality Assumption

- Assume \( \epsilon_{ij} \text{ i.i.d.} \sim N(0, \sigma^2) \).
- Define **treatment i mean**: \( \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \)
- Define **grand mean**: \( \bar{X} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{n_i} X_{ij} \)
- Sum of squares for treatment:
  \[
  SST = \sum_{i=1}^{K} n_i (\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{K} n_i \bar{X}_i^2 - N \bar{X}^2
  \]
- Sum of squares for error:
  \[
  SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{K} (n_i - 1) S_i^2,
  \]
  where \( S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \) is the sample variance of treatment \( i \).
- The \( F \)-test statistic is
  \[
  F = \frac{SST/(K - 1)}{SSE/(N - K)} = \frac{MST}{MSE}.
  \]
• Under the assumption that $\epsilon_{ij} \text{ i.i.d. } \sim N(0, \sigma^2)$, $F \sim F_{K-1, N-K}$ under $H_0$.

• Then critical values and $p$-value can be obtained by using the $F$ distribution.

• If we are not willing to make the normality assumption, we can use a permutation $F$-test, i.e. use permutation to obtain the null distribution of the $F$ test statistic.

### 3.3 Permutation $F$-test

Under $H_0$, $X_{ij}$ are exchangeable. There are total 
\[ \binom{N}{n_1 n_2 \cdots n_K} = \frac{N!}{n_1! n_2! \cdots n_K!} \] ways to partition total $N$ observations into $K$ groups with sizes $n_1, \cdots, n_K$.

**Steps:**

• Calculate $F_{obs}$ using the original data.
• For each of the \( \frac{N!}{n_1!n_2!\cdots n_K!} \) permutations (or for a random sample of \( R \) permutations), calculate \( F^* \).

• Calculate

\[
p\text{-value} = \frac{\# \text{ of } F^*\text{'s } \geq F_{obs}}{\# \text{ of permutations}}
\]

**Note:**

• The Sum of Squares of Total

\[
SSTotal = SST + SSE = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} X_{ij}^2 - N \bar{X}^2,
\]

which does not change across the permutations.

• Equivalently, we can base our permutation test on

\[
SST = \sum_{i=1}^{K} n_i (\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{K} n_i \bar{X}_{i}^2 - N \bar{X}^2 \text{ or}
\]

\[
SSX = \sum_{i=1}^{K} n_i \bar{X}_{i}^2
\]
instead of \( F = \frac{MST}{MSE} \) since \( F \) is an increasing function of both \( SST \) and \( SSX \).

**Example 3.3.1** Compare permutation \( F \)-test and one-way ANOVA \( F \)-test. The observations for three treatments are randomly sampled from \( N(15, 9^2) \), \( N(25, 9^2) \) and \( N(30, 9^2) \).

\[
\begin{array}{lcccccc}
  & j & 1 & 2 & 3 & 4 & 5 \\
trt1 & 25.07 & -1.45 & 22.61 & 28.58 & 15.51 \\
trt2 & 20.80 & 25.29 & 27.52 & 13.48 & 16.70 \\
trt3 & 34.13 & 31.70 & 30.78 & 24.22 & 29.86 \\
\end{array}
\]

**See the R code**

3.4 **Kruskal-Wallis Test**

Nonparametric rank test for comparing $K$ treatments:

- Replace the original observations with ranks
- Carry out the permutation test on the ranks
- This leads to a test equivalent to the Kruskal-Wallis test

The Kruskal-Wallis statistic is

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{K} n_i (\bar{R}_i - \frac{N+1}{2})^2,$$

where

- $\bar{R}_i$: the average rank for the $i$th treatment
- $(N + 1)/2$ is the average of all the ranks $1, 2, \cdots, N$
- $n_i(\bar{R}_i - \frac{N+1}{2})^2$: Sum of Squares for Treatment (SST) on ranks
• KW critical values for a limited number of scenarios can be found in Table A6

• For large samples,

\[ KW \sim \chi^2_{K-1} \] approximately.

R used the \( \chi^2_{K-1} \) approximation for \( p \)-value calculation

• The utility of approximate \( p \)-values based on the \( \chi^2_{K-1} \) distribution is questionable

• We can obtain the \( p \)-value directly by using the permutation test based on the \( KW \) statistic

• For large \( N \), the total number of permutations is large, so we may use a random sample of the permutations as an approximation. For example, \( n = 5 \) for each of 3 treatments, \( N = 15 \), \( \binom{15}{5 \ 5 \ 5} = 756, 756 \)
Adjustment for Ties:

- When there are ties, we adjust the ranks using midranks (average ranks) for the tied data.
- Use the adjusted ranks, calculate the KW test statistic for tied data

\[
KW_{ties} = \frac{1}{S_R^2} \sum_{i=1}^{K} n_i \left( \bar{R}_i - \frac{N + 1}{2} \right)^2,
\]

where \( S_R^2 \) is the sample variance of the combined adjusted ranks.

**Example 3.4.1** Refer to Example 3.3.1.

\( \bar{R}_i = 5.8, 6, 12.2, n_i = 5, i = 1, 2, 3, N = 15, (N + 1)/2 = 8 \). So

\[
KW = \frac{12}{15 \times 16} \sum_{i=1}^{3} 5(\bar{R}_i - 8)^2
\]

\[
= \frac{12 \times 5}{15 \times 16} \left\{ (5.8 - 8)^2 + (6 - 8)^2 + (12.2 - 8)^2 \right\} = 6.62.
\]
### 3 K-sample Methods

<table>
<thead>
<tr>
<th>$j$</th>
<th>Raw</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>trt1</td>
<td>trt2</td>
</tr>
<tr>
<td>1</td>
<td>25.1</td>
<td>20.8</td>
</tr>
<tr>
<td>2</td>
<td>-1.5</td>
<td>25.3</td>
</tr>
<tr>
<td>3</td>
<td>22.6</td>
<td>27.5</td>
</tr>
<tr>
<td>4</td>
<td>28.6</td>
<td>13.5</td>
</tr>
<tr>
<td>5</td>
<td>15.5</td>
<td>16.7</td>
</tr>
<tr>
<td>trt mean</td>
<td>18.1</td>
<td>20.8</td>
</tr>
</tbody>
</table>

```r
# Kruskal-Wallis test with chi-square approximation
> kruskal.test(x,grps)

Kruskal-Wallis rank sum test
data: x and grps
Kruskal-Wallis chi-squared = 6.62, df = 2, p-value = 0.03652
```
> summary(aov(rank.x ~ factor(grps)))

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F-value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(grps)</td>
<td>2</td>
<td>132.4</td>
<td>66.2</td>
<td>5.3821</td>
<td>0.02146 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>147.6</td>
<td>12.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

> SST = summary(aov(rank.x ~ factor(grps)))[[1]][1,2]
> 132.4*12/(N*(N+1))
[1] 6.62

# K-W test with permutation-based approx p-value
> Fobs <- getF(rank.x, grps)
> permFs <- perm.approx.F(rank.x, grps, R=1000)
> mean(permFs >= Fobs)
[1] 0.032
Example 3.4.2 Motivational Effect of Knowledge of Performance: Table 6.6 of Hollander and Wolfe (page 205). 18 male workers were divided randomly into 3 groups: control (no information is given), Group B(some rough information), Group C (accurate information of outputs). The number of pieces processed by each person in the experimental period.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Group B</th>
<th>Group C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40 (5.5)</td>
<td>38 (2.5)</td>
<td>48 (18)</td>
</tr>
<tr>
<td>2</td>
<td>35 (1)</td>
<td>40 (5.5)</td>
<td>40 (5.5)</td>
</tr>
<tr>
<td>3</td>
<td>38 (2.5)</td>
<td>47 (17)</td>
<td>45 (15)</td>
</tr>
<tr>
<td>4</td>
<td>43 (10.5)</td>
<td>44 (13)</td>
<td>43 (10.5)</td>
</tr>
<tr>
<td>5</td>
<td>44 (13)</td>
<td>40 (5.5)</td>
<td>46 (16)</td>
</tr>
<tr>
<td>6</td>
<td>41 (8)</td>
<td>42 (9)</td>
<td>44 (13)</td>
</tr>
<tr>
<td>trt mean</td>
<td>40.2 (6.75)</td>
<td>41.8 (8.75)</td>
<td>44.3 (13)</td>
</tr>
</tbody>
</table>
Solution:
Sorted data: 35 38 38 40 40 40 40 41 42 43 43 44 44 44 45 46 47 48
The values in the parentheses are the ranks. Midranks are used when there are ties.

> # Kruskal-Wallis test with chi-square approximation
> kruskal.test(x, grps)
Kruskal-Wallis rank sum test
data: x and grps
Kruskal-Wallis chi-squared = 4.3615, df = 2, p-value = 0.1130

> # K-W test with permutation-based approx p-value
> summary(aov(rank.x ~ factor(grps)))

Df     Sum Sq Mean Sq    F value Pr(>F)
factor(grps) 2   122.25   61.12   2.5882 0.1082
Residuals    15   354.25   23.62

> (SST = summary(aov(rank.x ~ factor(grps)))[[1]][1,2])
[1] 122.25
K-sample Methods

```r
> (SR2 = var(rank.x))
[1] 28.02941
> (SST/SR2)
[1] 4.36149

> #Fobs <- summary(aov(rank(x)~factor(grps)))[1,4]
> Fobs <- getF(rank.x, grps)
> set.seed(122356)
> permFs <- perm.approx.F(rank.x, grps, R=1000)
> mean(permFs >= Fobs)
[1] 0.124
```
3.5 Multiple Comparisons

3.5.1 Motivation

- The $F$-test and K-W test can only test if there is any difference among $K$ treatments.
- When $H_0 : \mu_1 = \cdots = \mu_K$ is rejected, we want to know which treatment differs from the others, i.e. to identify where the difference is.
- One way: pairwise comparison. E.g $K = 3$
  - Perform two-sample test to test $H_{01} : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ at significance level 5%. That is, the chance of incorrectly rejecting $H_{01}$ when two means are the same is 0.05.
  - Perform 2-sample test to test $H_{02} : \mu_1 = \mu_3$ at level 0.05.
  - Perform 2-sample test to test $H_{03} : \mu_2 = \mu_3$ at level 0.05.
  - If three tests are independent, then when $\mu_1 = \mu_2 = \mu_3$, the
probability of incorrectly rejecting at least one of 
$H_{0i}, i = 1, 2, 3$ is

which is greater than 0.05.

- And this Type I error rate increases quickly with $K$. When 
  $\alpha = 0.05$, $K = 6$, total $\binom{6}{2} = 15$ pairwise comparisons, the 
  overall Type I error rate is $1 - (1 - \alpha)^{15} = 0.54$.

- **Multiple comparison**: to determine which treatments differ 
  from others and meanwhile control the overall false rejection rate 
  under the desired level $\alpha$. 


3.5.2 Methods for Multiple Comparisons

Method 1: Bonferroni Adjustment

- \( K \) treatments, total \( n = \binom{K}{2} = \frac{K(K-1)}{2} \) number of pairwise comparisons

- Compare \( p \)-value for each pairwise comparison with significance level

\[
\frac{\alpha}{n} = \frac{\alpha}{K(K-1)/2}
\]

- This guarantees protection against inflation but tends to be too conservative.

For example, \( K = 6 \), \( n = \binom{6}{2} = 15 \). Let overall desired rejection rate \( \alpha = 0.05 \). We reject each pairwise comparison if the \( p \)-value \( \leq 0.05/15 = 0.0033 \). Then the probability of observing at least one
significant result is

\[ P(\geq 1\text{rejection}) = 1 - P(\text{no rejections}) = 1 - (1 - 0.0033)^{15} = 0.048. \]

Here, we’re just a bit under our desired 0.05 level. We benefit here from assuming that all tests are independent of each other. In practical applications, that is often not the case. Depending on the correlation structure of the tests, the Bonferroni correction could be extremely conservative, leading to a high rate of false negatives.

**Method 2: Fisher’s Protected Least Significant Difference (LSD)**

- The first multiple comparison invented.
- First check overall test for equality of multiple treatments for example \( F \)-test of KW test etc.
  - If the omnibus test is significant, then conduct pairwise comparison test with significance level \( \alpha \)
If not significant, then do not proceed and declare no significant different among all treatments

- Problem: if only one (or some) of the treatments is/are different from others, Method 2 can lead to many false conclusions of statistical significance, so the overall type I error is not well controlled.

- This method is not recommended.

**Method 3: Tukey’s Honest Significant Difference (HSD)**

- Define the normalized mean difference between groups $i$ and $j$ as:

$$T_{ij} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}},$$

where $MSE = SSE/df$ is the mean squared error in the one-way ANOVA, and $df = N - K$. 
• Define the largest difference statistic as

\[ Q = \max_{i<j} T_{ij}, \quad 1 \leq i < j \leq K. \]

• All pairwise (normalized) differences \( T_{ij} \) are compared with the critical values of \( Q \)—the largest difference. This makes Tukey’s HSD approach conservative.

• Under \( H_0 : \) no difference among \( K \) treatments,

\[ Q \sim \text{studentized range Q-distribution} \ q(K, df). \]

• Let \( q(\alpha, K, df) \) be the \( \alpha \)th percentage point of the null distribution of \( Q \).

• Limited values of \( q(\alpha, K, df) \) are given in Table A8.
• For comparing treatments \( i \) and \( j \), if the statistic

\[
T_{ij} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} > q(\alpha, K, df),
\]

we declare treatments \( i \) and \( j \) different. That is, we declare a significant difference between treatments \( i \) and \( j \) if

\[
|\bar{X}_i - \bar{X}_j| > q(\alpha, K, df) \sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} = HSD.
\]

• The null distribution of \( Q \) can be obtained by permutation.

• R functions
  – Tukey.HSD (Judy’s implementation based on permutation)
  – TukeyHSD (existing R function based on Q-distribution)
Example 3.5.1 The following table shows simulated data from $N(1, 1)$, $N(10, 1)$ and $N(1, 1)$. Use the given summary statistics to carry out multiple comparisons using Bonferroni, Fisher’s LSD and Tukey’s HSD methods. Use significance level 0.05.

<table>
<thead>
<tr>
<th>i \ j</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$\bar{X}_i$</th>
<th>$S^2_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>group1</td>
<td>-0.01</td>
<td>0.15</td>
<td>0.25</td>
<td>-0.90</td>
<td>0.86</td>
<td>0.07</td>
<td>0.40</td>
</tr>
<tr>
<td>group2</td>
<td>11.14</td>
<td>9.56</td>
<td>10.33</td>
<td>8.45</td>
<td>9.59</td>
<td>9.81</td>
<td>1.00</td>
</tr>
<tr>
<td>group3</td>
<td>0.77</td>
<td>0.03</td>
<td>1.96</td>
<td>3.24</td>
<td>2.40</td>
<td>1.68</td>
<td>1.65</td>
</tr>
</tbody>
</table>

ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(grps)</td>
<td>2</td>
<td>272.83</td>
<td>136.42</td>
<td>134.3</td>
<td>6.12e-09 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>12</td>
<td>12.19</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pairwise t-test results:

> t.test(x1, x2, var.equal=TRUE)
\begin{verbatim}
t = -18.3981, df = 8, p-value = 7.843e-08
> t.test(x1, x3, var.equal=TRUE)
t = -2.5157, df = 8, p-value = 0.03605
> t.test(x2, x3, var.equal=TRUE)
t = 11.1831, df = 8, p-value = 3.662e-06
\end{verbatim}
3.5.3 Multiple Comparison Permutation Tests

We can avoid messing with tables by adopting the permutation.

**Bonferroni Permutation Tests**

- Perform 2-sample permutation tests on all pairs of treatments and compare the permutation $p$-value for each pair with the adjusted significance level $\alpha' = \frac{\alpha}{K(K - 1)/2}$
- If the $p$-value for one pair is less than $\alpha'$, declare significance difference between this pair of treatments
- If the permutation test is based on ranks, need be careful: in $K$-sample comparison test such as KW test, the ranks range from 1 to $N = n_1 + \cdots + n_K$, but when comparing treatments $i$ and $j$, the ranks range from 1 to $n_i + n_j$
- This procedure controls the overall error rate $\leq \alpha$
Fisher’s Protected LSD Permutation Tests

- Carry out permutation tests to test if there is any difference among $K$ treatments, e.g. permutation based on F statistic or KS statistic.
- If and only if we reject the above omnibus test, we then consider pairwise comparisons.
- If comparing treatments $i$ and $j$, randomly permute the labels of $n_i + n_j$ observations associated with treatment $i$ and $j$ to form the permuted data for groups $i$ and $j$. For each permutation (or a random sample of $R$ permutations), compute $T_{ij}^*$. One logical statistic is

$$T_{ij}^* = \bar{X}_i^* - \bar{X}_j^*,$$

where $\bar{X}_i^*$ and $\bar{X}_j^*$ are the sample means of the $i$th and the $j$th treatment groups based on the permutation data.
• Compute permutation \( p \)-value

\[
p-value = \frac{\text{\# of } |T_{ij}^*|'s \geq |T_{ij}|}{R},
\]

where \( T_{ij} \) is the statistic \( \bar{X}_i - \bar{X}_j \) based on the observed data, and \( R \) is the number of permutations.

• Note that the permutation distribution of \( T_{ij}^* \) is a valid reference distribution for comparing any two treatments with the same sample sizes as \( n_i \) and \( n_j \).

**Tukey’s HSD Permutation Tests**

• Using the observed raw data (or ranks), calculate the mean squared error \( MSE = \frac{SSE}{(N - K)} \) and calculate

\[
T_{ij} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}
\]

for all \( K(K - 1)/2 \) pairs of \((i, j)\).
For each partition of $N$ observations into $K$ groups with sizes $n_1, \cdots, n_K$ (or for a sample of $R$ such permutations), calculate

$$Q^* = \max_{ij} T_{ij}^* = \max_{ij} \frac{|\bar{X}_i^* - \bar{X}_j^*|}{\sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} = \frac{\max_i (\bar{X}_i^*) - \min_i (\bar{X}_i^*)}{\sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

- Calculate $p$-value for each $T_{ij}$ (comparison of treatments $i$ and $j$):
  $$p\text{-value} = \frac{\# \text{of } Q\text{'s } \geq T_{ij}}{R}.$$

- If $p\text{-value} \leq \alpha$, declare significant difference between treatments $i$ and $j$. 


**Example 3.5.2 (Clay Percentage) (Table 3.3.1 in Higgins)** Six samples of soil were selected from 4 locations and the percentage of clay was determined in each sample.

<table>
<thead>
<tr>
<th>location (j)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>group mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>location 1</td>
<td>26.5</td>
<td>15.0</td>
<td>18.2</td>
<td>19.5</td>
<td>23.1</td>
<td>17.3</td>
<td>19.93</td>
</tr>
<tr>
<td>location 2</td>
<td>16.5</td>
<td>15.8</td>
<td>14.1</td>
<td>30.2</td>
<td>25.1</td>
<td>17.4</td>
<td>19.85</td>
</tr>
<tr>
<td>location 3</td>
<td>19.2</td>
<td>21.4</td>
<td>26.0</td>
<td>21.6</td>
<td>35.0</td>
<td>28.9</td>
<td>25.35</td>
</tr>
<tr>
<td>location 4</td>
<td>26.7</td>
<td>37.3</td>
<td>28.0</td>
<td>30.1</td>
<td>33.5</td>
<td>26.3</td>
<td>30.32</td>
</tr>
</tbody>
</table>

See R code
3.6 Ordered Alternatives

In some situations, it may be reasonable to suspect that the responses from different treatments follow some order.

- For example, researchers may anticipate the degree of pain relief will be greater for treatments with larger doses of the pain relief drug. Let $\mu_d$ denote the mean of pain reduction with dose $d$. We may want to test

$$H_a : \mu_0 \leq \mu_1 < \cdots < \mu_5.$$

- Argonomists may believe that the average yield of corn obtained with different levels of fertilizers (none, low, medium, high) will be

$$\mu_{\text{none}} \leq \mu_{\text{low}} \leq \mu_{\text{medium}} \leq \mu_{\text{high}}.$$
Let $F_i(x)$ be the CDF of the treatment $i$ group, $i = 1, \cdots, K$. We are interested in assessing the hypotheses:

$$H_0 : F_1(x) = F_2(x) = \cdots = F_K(x)$$

against

$$H_a : F_1(x) \geq F_2(x) \geq \cdots \geq F_K(x) \quad \text{(at least one strict inequality)},$$

(i.e. group 1 has the smallest values, \ldots, group $k$ has the largest values)

If we have location shift alternatives, the above $H_a$ can be expressed as

$$H_a : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K \quad \text{(at least one strict inequality)}.$$

**Jonckheere-Terpstra Test**

- Let $T_{ij}$ be any reasonable test statistic for testing

$$H_0 : F_i(x) = F_j(x), \quad v.s. \quad H_a : F_i(x) \geq F_j(x), \quad i < j$$

i.e. for one-sided alternative where the $j$th group gives larger
values.

• For instance, $T_{ij}$ can be chosen as the Wilcoxon’s rank-sum test statistic or Mann-Whitney’s test statistic for comparing group $i$ versus group $j$.

**Note:** for testing $H_a : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$ using $T_{ij}$,

• if we choose $T_{ij}$ as the rank-sum test statistic, define $T_{ij}$ as the sum of ranks of the $j$th treatment for $i < j$, i.e. the treatment with larger values;

• equivalently, if we choose $T_{ij}$ as the Mann-Whitney test statistic, define $T_{ij}$ as the number of pairs such that the observation from treatment $i \leq$ the observation from treatment $j$.

Therefore, a larger value of $T_{ij}$ supports the alternative that $\mu_i \leq \mu_j$. 
• Define the $JT$ test statistic as

$$JT = \sum_{i<j} T_{ij}.$$ 

• The null distribution of $JT$ can be obtained by using permutation
  – Compute $JT$ from the raw observed data.
  – Calculate $JT^*$ for each of (or a sample $R$ of) the possible allocations of the $N$ observations into $K$ groups of sizes $n_1, \cdots, n_K$.
  – Define $p$-value as

$$p\text{-value} = \frac{\# \text{of } JT^*\text{'s } \geq JT_{obs}}{R}.$$ 

  – Reject $H_0$ if $p$-value $\leq \alpha$.

• We can use F-test or KW test to test if there is any difference among $K$ treatments, but they would lose power since we have a specific alternative in mind.
Example 3.6.1  The basal area increment (BAI) for 16 stands of mixed species of oak trees in southern Ohio was measured. The BAI is related to yearly growth increment in a tree. The 16 stands were grouped according to the growing site index. As growing site index increases, the growing environment becomes more favorable for a stand of trees. The BAI data were grouped into 5 distinct categories according to the associated growing site index values.

<table>
<thead>
<tr>
<th>Growing site index interval</th>
<th>66-68</th>
<th>69-71</th>
<th>72-74</th>
<th>75-77</th>
<th>78-80</th>
</tr>
</thead>
<tbody>
<tr>
<td>stand idx1</td>
<td>1.91</td>
<td>2.44</td>
<td>2.45</td>
<td>2.52</td>
<td>2.78</td>
</tr>
<tr>
<td>stand idx2</td>
<td>1.53</td>
<td>2.04</td>
<td>2.36</td>
<td>2.88</td>
<td></td>
</tr>
<tr>
<td>stand idx3</td>
<td>2.08</td>
<td>1.60</td>
<td>2.73</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>stand idx4</td>
<td></td>
<td></td>
<td></td>
<td>1.60</td>
<td>1.66</td>
</tr>
<tr>
<td>stand idx5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.66</td>
</tr>
</tbody>
</table>
Pairwise rank-sum statistics $T_{ij}$:

\[
\begin{array}{cccc}
  i \backslash j & 2 & 3 & 4 & 5 \\
 1 & 5 & 22 & 18 & 23 \\
 2 & 11 & 8 & 12 & \\
 3 & 16 & 21 & & \\
 4 & & 16 & & \\
\end{array}
\]

Therefore, $JT = 5 + 22 + \cdots + 16 = 152$. By using permutation, we obtain the $p$-value=0.019. So we reject $H_0$.

Compare to KW test:

Kruskal-Wallis rank sum test
data:  x and grps
Kruskal-Wallis chi-squared = 5.9669, df = 4, p-value = 0.2016

So by incorporating the direction of treatment effects, we are able to detect a significant difference among the treatments.

See R code for implementation and other examples