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

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,

  grp1 : no knowledge
  grp2 : fair knowledge
  grp3 : good knowledge

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

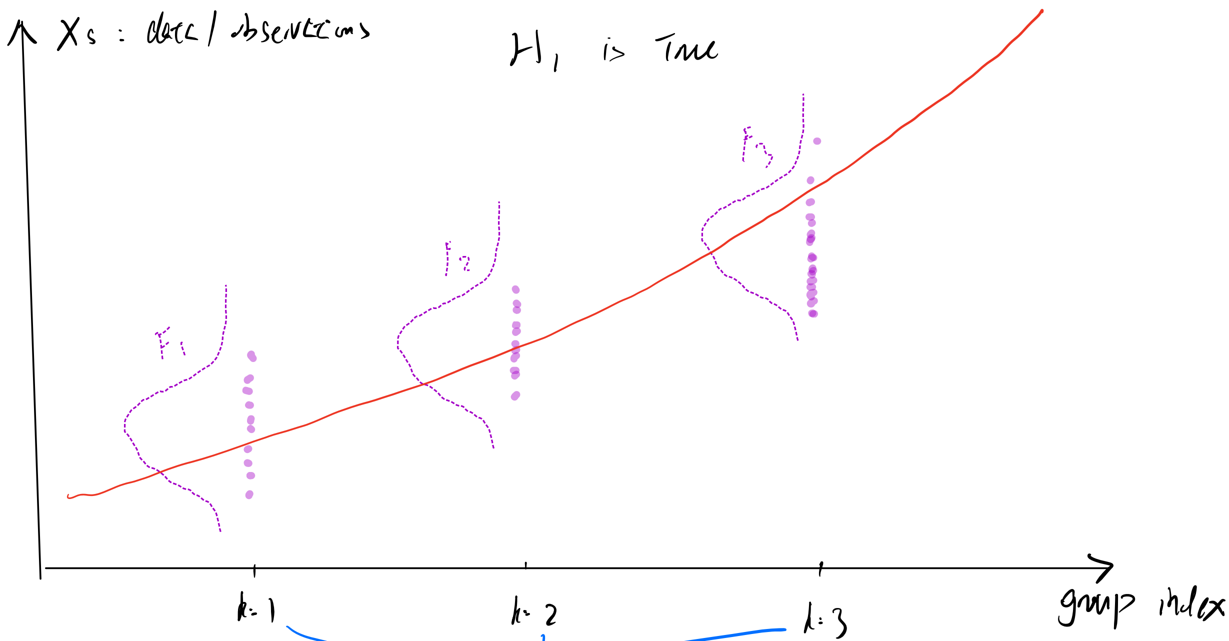
We note that the Kruskal-Wallis test does not utilize any such partial prior information regarding a postulated alternative ordering. The statistic H takes on the same value for all $k!$ possible labelings of the treatments.

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

Hypothesis

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

$$\underbrace{H_1 : \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k}_{\text{Ordered alternative}}$$



grp labels determined by applications

linear regression : precursor to LR

study the "trend"

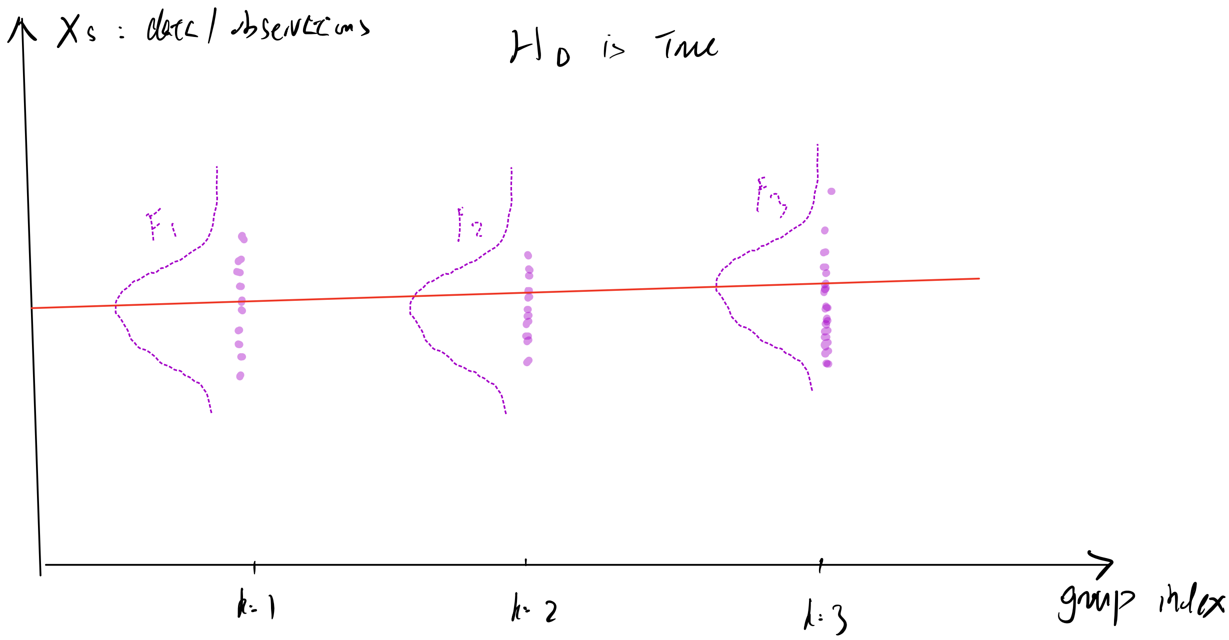
$$U_{12} \triangleq \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}(X_{1i} < X_{2j}) \quad \underline{\underline{\text{large}}}$$

Mann - Whitney count for subgrp 1 and 2

$$U_{23} \triangleq \cdots \cdots \quad \underline{\underline{\text{large}}}$$

$$U_{13} \triangleq \cdots \cdots \quad \underline{\underline{\text{large}}}$$

$$\boxed{U_{12} + U_{23} + U_{13}} \quad \text{large}$$



$$\underbrace{U_{12}}_{\text{Mann-Whitney count for subgroup 1 and 2}} \triangleq \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}(X_{1i} < X_{2j}) \approx \frac{n_1 \times n_2}{2}$$

$$U_{23} \triangleq \cdots\cdots \approx \frac{n_2 \times n_3}{2}$$

$$U_{13} \triangleq \cdots\cdots \approx \frac{n_1 \times n_3}{2}$$

$$\boxed{U_{12} + U_{23} + U_{13}} \approx \frac{1}{2}\left(n_1 n_2 + n_1 n_3 + n_2 n_3 \right)$$

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-T

$$\sum_{u} \sum_{v}$$

$$1 \leq u < v \leq k$$

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

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 $N!/\left(\prod_{j=1}^{k} n_j!\right)$ 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.

$k = 3, n_1 = n_2 = 1, n_3 = 2$ $\qquad \dfrac{4!}{2!\,1!\,1!} = 12$

I	II	III
3	4	1, 2

$U_{12} = 1$
$U_{23} = 0$ $\Rightarrow J = 1$
$U_{13} = 0$

I	II	III
2	4	1, 3

$U_{12} = 1$
$U_{23} = 0$ $\Rightarrow J = 2$
$U_{13} = 1$

I	II	III
2	3	1, 4

$U_{12} = 1$
$U_{23} = 1$ $\Rightarrow J = 3$
$U_{13} = 1$

I	II	III
4	3	1, 2

$U_{12} = 0$
$U_{23} = 0$ $\Rightarrow J = 0$
$U_{13} = 0$

I	II	III
4	2	1, 3

$U_{12} = 0$
$U_{23} = 1$ $\Rightarrow J = 1$
$U_{13} = 0$

I	II	III
3	2	1, 4

$U_{12} = 0$
$U_{23} = 1$ $\Rightarrow J = 2$
$U_{13} = 1$

I	II	III
1	4	2
		3

$\omega_{12}: 1$

$\omega_{23}: 0 \quad \Rightarrow J: 3$

$\omega_{13}: 2$

I	II	III
1	3	2
		4

$\omega_{12}: 1$

$\omega_{23}: 1 \quad \Rightarrow J: 4$

$\omega_{13}: 2$

I	II	III
1	2	3
		4

$\omega_{12}: 1$

$\omega_{23}: 2 \quad \Rightarrow J: 5$

$\omega_{13}: 2$

I	II	III
4	1	2
		3

$\omega_{12}: 0$

$\omega_{23}: 2 \quad \Rightarrow J: 2$

$\omega_{13}: 0$

I	II	III
3	1	2
		4

$\omega_{12}: 0$

$\omega_{23}: 2 \quad \Rightarrow J: 3$

$\omega_{13}: 1$

I	II	III
2	1	3
		4

$\omega_{12}: 0$

$\omega_{23}: 2 \quad \Rightarrow J: 4$

$\omega_{13}: 2$

null dsn of J :

J	prob
0	1/12
1	2/12
2	3/12
3	3/12
4	2/12
5	1/12

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

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

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

Under H_0, it can be shown that

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

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)).

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

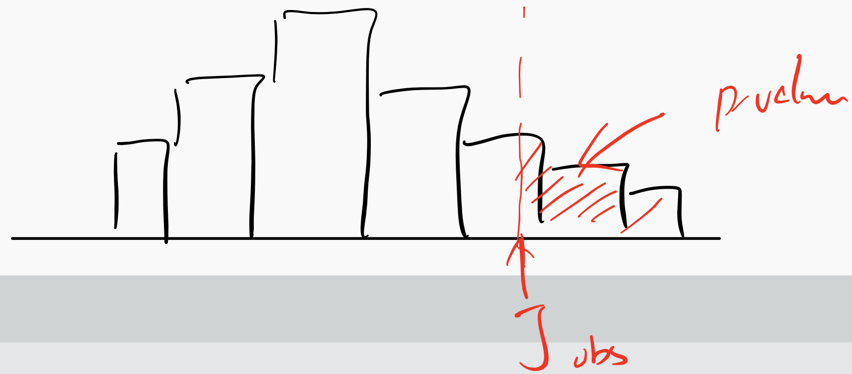
$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi\left(X_{iu}, X_{jv}\right), \quad 1 \leq u < v \leq k$$

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

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

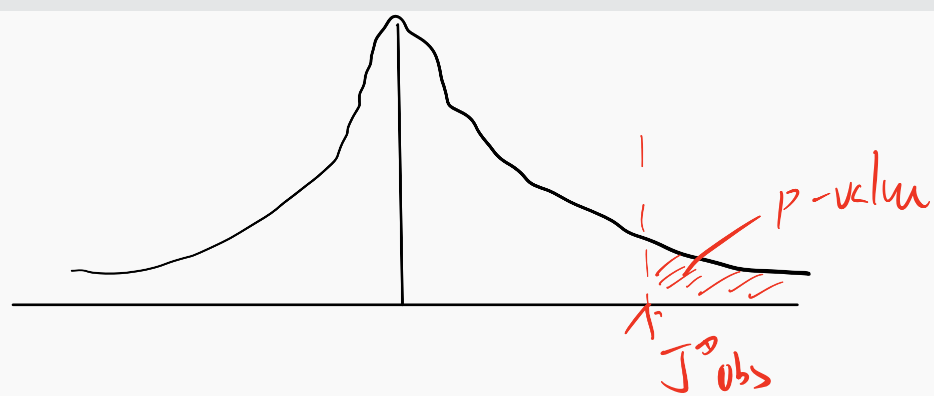
$$H_1: \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k \text{with at least one strict inequality}$$

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 [redacted] 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

$H_0: \tau_1 = \tau_2 = \tau_3$

vs

$H_1: \tau_1 \le \tau_2 \le \tau_3$

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

$\frac{18!}{6!\,6!\,6!}$ = huge

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
```