# Basic Nonparametric Statistical Methods

H. Wakaki

January 25, 2016

## Contents

**Reference**

"NONPARAMETRIC STATISTICAL METHODS second edition"
by Myles Hollander and Douglas A.Wolfe

# 1 Introduction

**Thema** Inference on Bernoulli trial

**Data** $X_1, \cdots, X_n$

**Assumptions**

A1. $X_1, \cdots, X_n$ : independent

A2. $\Pr\{X_i = 1\} = 1 - \Pr\{X_i = 0\} = p \quad (0 < p < 1)$

## 1.1 Binomial test

**Null hypothesis**

$$H_0 : p = p_0 \quad \text{(a specified number)}$$

**Test statistic**

$$T = \sum_{i=1}^{n} X_i$$

**Null distribution** Under $H_0$

$$T \sim B(n, p_0),$$
$$\Pr\{T = k\} = q_0(k) := \frac{n!}{k!(n-k)!} p_0^k (1-p_0)^{n-k} \quad (k = 0, 1, \cdots, n)$$

a. One–sided upper-tail test

$$H_0 : p = p_0 \text{ versus } H_1 : p > p_0$$
$$T \geq b_\alpha \Rightarrow \text{ Reject } H_0, \quad b_\alpha = \min\Big\{k : \sum_{j=k}^{n} q_0(k) \leq \alpha\Big\}$$

Randomized test

In the case of $\alpha_0 = \Pr\{T \geq b_\alpha\} < \alpha$.

Let $U$ : uniform random variable on $(0, 1)$.

$$T \geq b_\alpha \quad \Rightarrow \quad \text{Reject}$$

or

$$T = b_\alpha - 1 \text{ and } U < \frac{\alpha - \alpha_0}{q_0(b_\alpha - 1)} \quad \Rightarrow \quad \text{Reject}$$

The $P$–value

Let $t$ : observed value

$\Pr\{B \geq t\}$ is called "$P$–value", $\quad B \sim B(n, p_0)$

b. One–sided lower-tail test

$$H_0 : p = p_0 \text{ versus } H_2 : p < p_0$$

$$T \leq c_\alpha \Rightarrow \text{ Reject } H_0, \quad c_\alpha = \max\Big\{k : \sum_{j=0}^{k} q_0(k) \leq \alpha\Big\}$$

c. Two–sided test

$$H_0 : p = p_0 \text{ versus } H_3 : p \neq p_0$$
$$T \geq b_{\alpha_1} \text{ or } T \leq c_{\alpha_2} \Rightarrow \text{ Reject } H_0, \quad \alpha_1 + \alpha_2 = \alpha$$

**Large–sample approximation**

$$T^* = \frac{T - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{d} N(0,1)$$

$T^* \geq z_\alpha \quad \Rightarrow \quad$ Reject (one–sided upper–tail test)
$T^* \leq -z_\alpha \quad \Rightarrow \quad$ Reject (one–sided lower–tail test)
$|T^*| \geq z_{\alpha/2} \quad \Rightarrow \quad$ Reject (Tow–sided test)

$z_\alpha$ : upper $100\alpha$ percent point of $N(0,1)$

## 1.2 Point estimation

**Estimator**

$$\hat{p} = \frac{1}{n}T = \frac{1}{n}\#\{i : X_i = 1, 1 \leq i \leq n\}$$

**Properties**

$$\mathrm{E}[\hat{p}] = p \quad \text{(unbiased estimator)}$$
$$\mathrm{Var}[\hat{p}] = \frac{1}{n}p(1-p)$$
$$\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1) \quad (n \to \infty)$$

## 1.3 Interval estimation

Define

$$\big\{p_L(k;\alpha)\big\}_{k=0,1,\cdots,n} \quad \text{and} \quad \big\{p_U(k;\alpha)\big\}_{k=0,1,\cdots,n}$$

such that

$$p_L(k;\alpha) < p_U(k;\alpha) \quad (k = 0, 1, \cdots, n)$$
$$p_L(k;\alpha) < p_L(k+1;\alpha), \quad p_U(k;\alpha) < p_U(k+1;\alpha) \quad (k = 0, 1, \cdots, n-1)$$
$$\sum_{k:p_L(k;\alpha)\le p\le p_U(k;\alpha)} \frac{n!}{k!(n-k)!} p^k (1-p)^k \ge 1-\alpha \text{ for all } p \in (0,1).$$

$\Rightarrow$

$$\Pr\{p_L(T;\alpha) \le p \le p_U(T;\alpha)\} \ge 1-\alpha \quad \text{for all } p$$

Confidence interval

$$[p_L(T;\alpha), p_U(T;\alpha)] : 1-\alpha \text{ confidence interval for } p$$

**Large–sample approximation**

$$p_L(T;\alpha) = \hat{p} - z_{\alpha/2}\Big[\frac{\hat{p}(1-\hat{p})}{n}\Big]^{1/2}, \quad p_U(T;\alpha) = \hat{p} + z_{\alpha/2}\Big[\frac{\hat{p}(1-\hat{p})}{n}\Big]^{1/2}$$

$\Rightarrow$

$$\lim_{n\to\infty} \Pr\{p_L(T;\alpha) \le p \le p_U(T;\alpha)\} = 1-\alpha$$

# 2 The One–Sample Location Problem

**Thema** Inferences on the location

**Two types of data**

- paired replicates data — "pretreatment" and "posttreatment"
- one–sample data

## 2.1 Paired replicates analyses by way of signed ranks

**Data**

| Subject $i$ | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | $X_1$ | $Y_1$ |
| 2 | $X_2$ | $Y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $X_n$ | $Y_n$ |

**Assumptions**

A1. Let $Z_i = Y_i - X_i \quad (i = 1, \cdots, n)$.
$Z_1, \cdots, Z_n$ : independent

A2. Let $F_i(t) = \Pr\{Z_i \leq t\} \quad (i = 1, \cdots, n)$.
$F_i$ is continuous and

$$F_i(\theta + t) + F_i(\theta - t) = 1, \text{ for every } t \quad (i = 1, \cdots, n)$$

Note $\theta$ is the median which is referred as the *treatment effect*

### 2.1.1 A distribution–free signed rank test (Wilcoxon)

**Null hypothesis**

$$H_0 : \theta = 0$$

**Test statistic**

$$\psi_i = \begin{cases} 1, & (Z_i > 0) \\ 0, & (Z_i < 0) \end{cases}$$

$R_i$ : the rank of $|Z_1|, \cdots, |Z_n|$ ordered from least to greatest

$$\Rightarrow \quad T^+ = \sum_{i=1}^{n} R_i \psi_i : \text{ Wilcoxon signed rank statistic}$$

**Null distribution** Let

$V_1, \cdots, V_n$ : independent random variables

$$\Pr\{V_i = i\} = \Pr\{V_i = 0\} = \frac{1}{2} \quad (i = 1, \cdots, n)$$

$$\Rightarrow \quad T^+ \sim \sum_{i=1}^{n} V_i \quad \text{same distribution} \Rightarrow \text{ distribution free (not depend on } F_i)$$

$\because$
$\theta = 0 \Rightarrow |Z_1|, \cdots, |Z_n|, \psi_1, \cdots, \psi_n$ : independent
Define $I_i, \cdots, I_n$ by

$$I_j = i \Leftrightarrow R_i = j \quad (j = 1, \cdots, n).$$

$\Rightarrow (I_1, \cdots, I_n), \psi_1, \cdots, \psi_n$ : independent
$b_j = 0$ or $1$ $(j = 1, \cdots, n)$.

$S_n$ : set of all permutations of $(1, 2, \cdots, n)$.

$$\begin{aligned}
&\Pr\{\psi_{I_1} = b_1, \cdots, \psi_{I_n} = b_n\} \\
&\quad = \sum_{(s_1,\cdots,s_n)\in S_n} \Pr\{\psi_{I_1} = b_1, \cdots, \psi_{I_n} = b_n, (I_1, \cdots, I_n) = (s_1, \cdots, s_n)\} \\
&\quad = \sum_{(s_1,\cdots,s_n)\in S_n} \Pr\{\psi_{s_1} = b_1, \cdots, \psi_{s_n} = b_n, (I_1, \cdots, I_n) = (s_1, \cdots, s_n)\} \\
&\quad = \sum_{(s_1,\cdots,s_n)\in S_n} \Pr\{(I_1, \cdots, I_n) = (s_1, \cdots, s_n)\}\left(\prod_{i=1}^{n} \Pr\{\psi_{s_i} = b_i\}\right) \\
&\quad = \sum_{(s_1,\cdots,s_n)\in S_n} \Pr\{(I_1, \cdots, I_n) = (s_1, \cdots, s_n)\}\left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n
\end{aligned}$$

$\Rightarrow$

$\psi_{I_1}, \cdots, \psi_{I_n}$ : independent,

$\Pr\{\psi_{I_j} = 0\} = \Pr\{\psi_{I_j} = 1\} = \dfrac{1}{2} \quad (j = 1, \cdots, n)$

Define

$$V_j = \psi_{I_j} j \quad (j = 1, \cdots, n).$$

Then $V_1 \cdots, V_n$ : independent, $\Pr\{V_j = 0\} = \Pr\{V_j = j\} = \frac{1}{2}$,

$$T^+ = \sum_{j=1}^{n} V_j,$$

**Three types of test**

a. One–sided upper-tail test

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0$$

$$T^+ \geq t_\alpha \Rightarrow \text{ Reject } H_0, \quad t_\alpha = \min\left\{k : \Pr\Big\{\sum_{i=1}^{n} V_i \geq k\Big\} \leq \alpha\right\}$$

b. One–sided lower-tail test

$$H_0 : \theta = 0 \text{ versus } H_2 : \theta < 0$$

$$T^+ \leq \frac{n(n+1)}{2} - t_\alpha \Rightarrow \text{ Reject } H_0,$$

c. Two–sided test

$$H_0 : \theta = 0 \text{ versus } H_3 : \theta \neq 0$$

$$T^+ \geq t_{\alpha/2} \text{ or } T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2} \Rightarrow \text{ Reject } H_0,$$

**Large–sample approximation**

Under $H_0$

$$\mathrm{E}[T^+] = \frac{n(n+1)}{4}, \quad \mathrm{Var}[T^+] = \frac{n(n+1)(2n+1)}{24}$$
$$T^* = \frac{T^+ - \mathrm{E}[T^+]}{\sqrt{\mathrm{Var}[T^+]}} \xrightarrow{d} N(0,1) \quad (n \to \infty)$$

$$T^* \geq z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided upper–tail test)}$$
$$T^* \leq -z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided lower–tail test)}$$
$$|T^*| \geq z_{\alpha/2} \quad \Rightarrow \quad \text{Reject (Tow–sided test)}$$
$$z_\alpha : \text{upper } 100\alpha \text{ percent point of } N(0,1)$$

$\because$

Lemma (Lindeberg)
$\{X_n\}_{n=1,2,\cdots}$ : sequence of independent random variables
$\mathrm{E}[X_n] = 0, \mathrm{E}[X_n^2] = \sigma_n^2 < \infty$

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|x|>\varepsilon s_n} x^2 dF_j(x) = 0 \quad \text{for all } \varepsilon > 0$$
$$\frac{\sum_{j=1}^{n} X_n}{s_n} \xrightarrow{d} N(0,1),$$

where

$$s_n^2 = \sum_{j=1}^{n} \sigma_n^2$$

Let $X_j = V_j - \mathrm{E}[V_j]$. Then $\Pr(X_j = -\frac{j}{2}) = \Pr(X_j = \frac{j}{2}) = \frac{1}{2}$.

$$\mathrm{E}[V_j] = \frac{j}{2}, \quad \sigma_j^2 = \mathrm{Var}[V_j] = \frac{j^2}{2} - \left(\frac{j}{2}\right)^2 = \frac{j^2}{4}$$
$$s_n^2 = \mathrm{Var}[T^+] = \frac{n(n+1)(2n+1)}{24}$$
$$\left|\frac{\frac{j}{2}}{s_n}\right| \leq \frac{n}{2s_n} = \sqrt{\frac{6n}{(n+1)(2n+1)}} \to 0 \quad (n \to \infty)$$

Hence

$$\exists n_\varepsilon \ s.t. \ n > n_\varepsilon \Rightarrow \int_{|x|>\varepsilon s_n} x^2 dF_j(x) = 0,$$

which leads that the Lindeberg condition holds.

**Testing equality to the specified value**

$$H_0 : \theta = \theta_0 \ \ \text{(specified value)}$$

**Test statistic**

$$\psi_i = \begin{cases} 1, & (Z_i > \theta_0) \\ 0, & (Z_i < \theta_0) \end{cases}$$

$R_i$ : the rank of $|Z_1 - \theta_0|, \cdots, |Z_n - \theta_0|$ ordered from least to greatest

$$\Rightarrow \quad T^+ = \sum_{i=1}^{n} R_i \psi_i : \text{ Wilcoxon signed rank statistic}$$

### 2.1.2 An estimator associated with Wilcoxon's signed rank statistic (Hodges–Lehmann)

**Derivation** Choose the value of $\theta_0$ such that $H_0 : \theta = \theta_0$ is least significant, that is

$$\hat{\theta} = \underset{\theta_0}{\operatorname{argmin}} \left| T_0^+ - \frac{n(n+1)}{4} \right|,$$

$T_0^+$ : test statistic for testing "$\theta = \theta_0$"

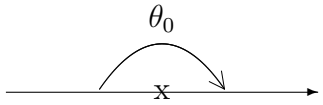
**Estimator**

$$\hat{\theta} = \text{median} \left\{ \frac{Z_i + Z_j}{2}, 1 \le i \le j \le n \right\}$$

$$W^{(1)} < \cdots < W^{(M)} : \text{ ordered values of } \frac{Z_i + Z_j}{2}$$

$$M = \frac{n(n+1)}{2} = 2k + 1 \Rightarrow \hat{\theta} = W^{(k+1)}$$

$$M = 2k \Rightarrow \hat{\theta} = \frac{W^{(k)} + W^{(k+1)}}{2}$$




$T_0^+ \Rightarrow T_0^+ - 1$

$$\theta_0 < Z_{(1)} < Z_{(2)} \cdots < Z_{(n)} \Rightarrow T_0^+ = \frac{n(n+1)}{2}$$

$$Z_{(n)} < \theta_0 \Rightarrow T_0^+ = 0$$

### 2.1.3 A distribution–free confidence interval based on Wilcoxon's signed rank test (Tukey)

**Derivation** Collect $\theta_0$ such that the two sided test for

$$H_0 : \theta = \theta_0 \text{ versus } H_3 : \theta \neq \theta_0$$

does not reject.

**Confidence interval**

$$C_\alpha = \frac{n(n+1)}{2} + 1 - t_{\alpha/2},$$

$$\theta_L = W^{(C_a)}, \quad \theta_U = W^{M+1-C_\alpha} = W^{(t_{a/2})}$$

$$\Rightarrow$$

$$\Pr{}_\theta\{\theta_L < \theta < \theta_U\} = 1 - \alpha \text{ for all } \theta$$

**Large–sample approximation**

$$C_\alpha \approx \frac{n(n+1)}{4} - z_{\alpha/2}\left\{\frac{n(n+1)(2n+1)}{24}\right\}^{1/2}$$

## 2.2 Paired replicates analyses by way of signs

**Data**

| Subject $i$ | $X_i$ | $Y_i$ |
|---|---|---|
| 1 | $X_1$ | $Y_1$ |
| 2 | $X_2$ | $Y_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $X_n$ | $Y_n$ |

**Assumptions**

B1. Let $Z_i = Y_i - X_i \quad (i = 1, \cdots, n)$.
$Z_1, \cdots, Z_n$ : independent

B2.

$$\Pr\{Z_i < \theta\} = \Pr\{Z_i > \theta\} = \frac{1}{2} \quad (n = 1, \cdots, n)$$

### 2.2.1 A distribution–free signed test (Fisher)

**Null hypothesis**

$$H_0 : \theta = 0$$

**Test statistic**

$$\psi_i = \begin{cases} 1, & (Z_i > 0) \\ 0, & (Z_i < 0) \end{cases}$$

$$\Rightarrow \quad B = \sum_{i=1}^{n} \psi_i : \text{ Fisher's signed statistic}$$

**Null distribution**

$$B \sim B\left(n, \frac{1}{2}\right)$$

**Three types of test**

a. One–sided upper-tail test

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0$$

$$B \geq b_\alpha \Rightarrow \text{ Reject } H_0, \quad b_\alpha = \min\left\{ k : \sum_{i=k}^{n} \frac{n!}{i!(n-i)!}\left(\frac{1}{2}\right)^n \leq \alpha \right\}$$

b. One–sided lower-tail test

$$H_0 : \theta = 0 \text{ versus } H_2 : \theta < 0$$
$$B \leq n - b_\alpha \Rightarrow \text{ Reject } H_0,$$

c. Two–sided test

$$H_0 : \theta = 0 \text{ versus } H_3 : \theta \neq 0$$
$$B \geq b_{\alpha/2} \text{ or } B \leq n - t_{\alpha/2} \Rightarrow \text{ Reject } H_0,$$

**Large–sample approximation**

Under $H_0$

$$\mathrm{E}[B] = \frac{n}{2}, \quad \mathrm{Var}[B] = \frac{n}{4}$$
$$B^* = \frac{B - \mathrm{E}[B]}{\sqrt{\mathrm{Var}[B]}} \xrightarrow{d} N(0,1) \quad (n \to \infty)$$

$$B^* \geq z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided upper–tail test)}$$
$$B^* \leq -z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided lower–tail test)}$$
$$|B^*| \geq z_{\alpha/2} \quad \Rightarrow \quad \text{Reject (Tow–sided test)}$$

$z_\alpha$ : upper $100\alpha$ percent point of $N(0,1)$

**Testing equality to the specified value**

$$H_0 : \theta = \theta_0$$

**Test statistic**

$$\psi_i = \begin{cases} 1, & (Z_i > \theta_0) \\ 0, & (Z_i < \theta_0) \end{cases}$$
$$\Rightarrow \quad B = \sum_{i=1}^{n} \psi_i$$

### 2.2.2 An estimator associated with the sign statistic(Hodges–Lehmann)

**Derivation** Choose the value of $\theta_0$ such that $H_0 : \theta = \theta_0$ is least significant, that is

$$\hat{\theta} = \underset{\theta_0}{\mathrm{argmin}} \left| B_0 - \frac{n}{2} \right|,$$

$B_0$ : test statistic for testing " $\theta = \theta_0$ "

**Estimator**

$$\hat{\theta} = \text{median}\{Z_i : 1 \leq i \leq n\}$$

### 2.2.3 A distribution–free confidential interval based on the sign test (Thompson, Savur)

**Derivation** Collect $\theta_0$ such that the two–sided test for

$$H_0 : \theta = \theta_0 \text{ versus } H_3 : \theta \neq \theta_0$$

does not reject.

**Confidence interval**

$$\begin{aligned} & C_\alpha = n + 1 - b_{\alpha/2} \\ & Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)} : \text{ ordered statistics of } Z_1, \cdots, Z_n \\ & \theta_L = Z_{(C_\alpha)}, \quad \theta_U = Z_{(b_{\alpha/2})} \\ \Rightarrow & \\ & \quad \Pr{}_\theta\{\theta_L < \theta < \theta_U\} = 1 - \alpha \text{ for all } \theta \end{aligned}$$

**Large–sample approximation**

$$C_\alpha \approx \frac{n}{2} - z_{\alpha/2} \left\{\frac{n}{4}\right\}^{1/2}$$

## 2.3 One–sample data

**Data**

$$X_1, \cdots, X_n$$

**Assumptions**

C1. $X_1, \cdots, X_n$ : independent

C2. Let $F_i(t) = \Pr\{X_i \leq t\} \quad (i = 1, \cdots, n)$.
$F_i$ is continuous and

$$F_i(\theta + t) + F_i(\theta - t) = 1, \text{ for every } t \quad (n = 1, \cdots, n)$$

### 2.3.1 A distribution–free signed rank test

**Null hypothesis**

$$H_0 : \theta = \theta_0$$

Define

$$Z_i = X_i - \theta_0 \quad (i = 1, \cdots, n)$$

The same procedure as §2.1.1 can be used.

### 2.3.2 An estimator based on the signed rank test

The same procedure as §2.1.2 can be used

### 2.3.3 A distribution free confidential interval based on the signed rank test

The same procedure as §2.1.3 can be used.

## 2.4 Procedures based on the sign statistic

**Data**

$$X_1, \cdots, X_n$$

**Assumptions**

D1.

$$X_1, \cdots, X_n : \text{independent}$$

D2.

$$\Pr\{X_i < \theta\} = \Pr\{X_i > \theta\} = \frac{1}{2} \quad (n = 1, \cdots, n)$$

Define

$$Z_i = X_i - \theta_0 \quad (i = 1, \cdots, n)$$

**Procedures** The same procedures in §2.2 can be used for testing problem, point estimation and interval estimation.

# 3 The Two–Sample Location Problem

**Data**

$$X_1, \cdots, X_m, Y_1, \cdots, Y_n$$

**Assumptions**

A1. $X_1, \cdots, X_m$ : a random sample from population 1
$Y_1, \cdots, Y_n$ : a random sample from population 2
that is
$X_1, \cdots, X_m$ : i.i.d. (independent and identically distributed)
$Y_1, \cdots, Y_n$ : i.i.d. (independent and identically distributed)

A2. $(X_1, \cdots, X_m), (Y_1, \cdots, Y_n)$ : independent

A3. Population distribution functions of 1 ($F(t)$) and 2 ($G(t)$)are continuous.

**Null hypothesis**

$$H_0 : F(t) = G(t) \text{ for all } t$$
$$F(t) = \Pr\{X_i \leq t\}, \quad G(t) = \Pr\{Y_i \leq t\}$$

**Alternative hypothesis(Assumption)**

$$H : G(t) = F(t - \Delta)$$

Under $H$

$$\Pr\{X_i + \Delta \leq t\} = \Pr\{X_i \leq t - \Delta\} = F(t - \Delta) = G(t)$$
$$\Rightarrow X_1 + \Delta, \cdots, X_m + \Delta, Y_1, \cdots, Y_n : i.i.d.$$

$F$ : control（対照群）
$G$ : case（処置群）
$\Delta$ : treatment effect

## 3.1 A distribution–free rank sum test (Wilcoxon, Mann and Whitney)

**Testing problems**

$$H_0 : \Delta = 0 \text{ versus } \begin{cases} H_1 : \Delta > 0 \\ H_2 : \Delta < 0 \\ H_3 : \Delta \neq 0 \end{cases}$$

**Derivation**

Under $H_0$

$$X_1, \cdots, X_m, Y_1, \cdots, Y_n : i.i.d.$$

**Test statistic**

$$W = \sum_{j=1}^{n} S_j$$
$$S_j : \text{ rank of } Y_j \text{ in } \{X_1, \cdots, X_m, Y_1, \cdots, Y_n\}$$

**Null distribution**

$\mathcal{S}_{n,N}$ : set of subsets of $\{1, 2, \cdots, N\}$ $(N = m + n)$ of size $n$

$\Rightarrow$

$\{S_1, \cdots, S_n\}$ is uniformly distributed random set on $\mathcal{S}_{n,N}$

Example $m = 4, n = 3$

$$\mathcal{S}_{4,7} = \left\{ \begin{array}{l} \{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,2,6\},\{1,2,7\},\{1,3,4\},\{1,3,5\}, \\ \{1,3,6\},\{1,3,7\},\{1,4,5\},\{1,4,6\},\{1,4,7\},\{1,5,6\},\{1,5,7\}, \\ \{1,6,7\},\{2,3,4\},\{2,3,5\},\{2,3,6\},\{2,3,7\},\{2,4,5\},\{2,4,6\}, \\ \{2,4,7\},\{2,5,6\},\{2,5,7\},\{2,6,7\},\{3,4,5\},\{3,4,6\},\{3,4,7\}, \\ \{3,5,6\},\{3,5,7\},\{3,6,7\},\{4,5,6\},\{4,5,7\},\{4,6,7\},\{5,6,7\} \end{array} \right\}$$

| $w$ | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Pr\{W=w\}$ | $\frac{1}{35}$ | $\frac{1}{35}$ | $\frac{2}{35}$ | $\frac{3}{35}$ | $\frac{4}{35}$ | $\frac{4}{35}$ | $\frac{5}{35}$ | $\frac{4}{35}$ | $\frac{4}{35}$ | $\frac{3}{35}$ | $\frac{2}{35}$ | $\frac{1}{35}$ | $\frac{1}{35}$ |

**Three types of test**

a. One–sided upper-tail test

$$H_0 : \Delta = 0 \text{ versus } H_1 : \Delta > 0$$
$$W \geq w_\alpha \Rightarrow \text{ Reject } H_0,$$

b. One–sided lower-tail test

$$H_0 : \Delta = 0 \text{ versus } H_2 : \Delta < 0$$
$$W \leq n(m+n+1) - w_\alpha \Rightarrow \text{ Reject } H_0,$$

c. Two–sided test

$$H_0 : \Delta = 0 \text{ versus } H_3 : \Delta \neq 0$$
$$W \geq w_{\alpha/2} \text{ or } W \leq n(m+n+1) - w_{\alpha/2} \Rightarrow \text{ Reject } H_0,$$

**Large–sample approximation**

Under $H_0$

$$\mathrm{E}[W] = \frac{n(m+n+1)}{2}, \quad \mathrm{Var}_0[W] = \frac{mn(m+n+1)}{12}$$
$$W^* = \frac{W - \mathrm{E}[W]}{\sqrt{\mathrm{Var}_0[W]}} \xrightarrow{d} N(0,1) \quad (n \to \infty)$$

$$W^* \geq z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided upper–tail test)}$$
$$W^* \leq -z_\alpha \quad \Rightarrow \quad \text{Reject (one–sided lower–tail test)}$$
$$|W^*| \geq z_{\alpha/2} \quad \Rightarrow \quad \text{Reject (Tow–sided test)}$$
$$z_\alpha : \text{upper } 100\alpha \text{ percent point of } N(0,1)$$

Proof

$$\phi(x,y) = \begin{cases} 1 & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$
$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j)$$
$$\Rightarrow \quad W = U + \frac{n(n+1)}{2}$$

$\because$

$$R(Y_j): \text{ rank of } Y_j$$

$$\Rightarrow \quad R(Y_j) = \sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + 1,$$

$$W = \sum_{j=1}^{n}\sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j=1}^{n}\sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + n$$

$$\sum_{j=1}^{n}\sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) = 0 + 1 + \cdots + n - 1 = \frac{n(n-1)}{2}$$

$$\mathrm{E}[U] = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{E}[\phi(X, Y)], \quad X, Y \overset{i.i.d.}{\sim} F,$$

$$\mathrm{E}[\phi(X, Y)] = \mathrm{E}_Y[\phi(X, Y)|Y] = \mathrm{E}_Y[F(Y)] = \frac{1}{2}$$

$\because \quad F(Y) \sim U(0, 1)$ Uniform distribution

$$\mathrm{E}[W] = \frac{mn}{2} + \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2}$$

$$\mathrm{Var}[U] = \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n} \mathrm{Cov}[\phi(X_i, Y_j), \phi(X_{i'}, Y_{j'})]$$

If $i \neq i', j \neq j'$

$\mathrm{Cov}[\phi(X_i, Y_j), \phi(X_{i'}, Y_{j'})] = 0$

If $i \neq i', j = j'$

$$\mathrm{Cov}[\phi(X_i, Y_j), \phi(X_{i'}, Y_j)] = \mathrm{E}[\phi(X_i, Y_j)\phi(X_{i'}, Y_j)] - \mathrm{E}[\phi(X_i, Y_j)] \cdot \mathrm{E}[\phi(X_{i'}, Y_j)]$$

$$= \mathrm{E}_{X_i, X_j}\mathrm{E}\big[\phi(X_i, Y_j)\phi(X_{i'}, Y_j)|X_i, X_{i'}\big] - \frac{1}{4}$$

$$= \mathrm{E}_{X_i, X_{i'}}\big[\Pr\{\max\{X_i, X_{i'}\} < Y|X_i, X_{i'}\}\big] - \frac{1}{4}$$

$$= \mathrm{E}_{X_i, X_{i'}}\big[\min\big\{1 - F(X_i), 1 - F(X_{i'})\big\}\big] - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

If $i = i', j \neq j'$

$\mathrm{Cov}[\phi(X_i, Y_j), \phi(X_i, Y_{j'})] = \dfrac{1}{12}$ similarly as above

If $i = i', j = j'$

$$\mathrm{Cov}[\phi(X_i, Y_j), \phi(X_i, Y_j)] = \mathrm{E}[\phi(X_i, Y_j)^2] - \frac{1}{4} = \mathrm{E}[\phi(X_i, Y_j)] - \frac{1}{4} = \frac{1}{4}$$

$$\mathrm{Var}[U] = \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\left\{ \frac{(1 - \delta_{ii'})\delta_{jj'}}{12} + \frac{\delta_{ii'}(1 - \delta_{jj'})}{12} + \frac{\delta_{ii'}\delta_{jj'}}{4} \right\}$$

$$= \frac{1}{12}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\big\{\delta_{jj'} + \delta_{ii'} + \delta_{ii'}\delta_{jj'}\big\} = \frac{1}{12}(m^2n + mn^2 + mn) = \frac{mn(m+n+1)}{12}$$

Theorem (CLT for a general Mann–Whittney statistic)

Assume

$$
\begin{aligned}
&X_1, \cdots, X_m : i.i.d., \quad Y_1, \cdots, Y_n : i.i.d. \\
&(X_1, \cdots, X_m), (Y_1, \cdots, Y_n) : \text{ independent} \\
&\mathrm{E}[\varphi(X_i, Y_j)] = 0, \quad \mathrm{Var}[\varphi(X_i, Y_j)] = \sigma^2 < \infty.
\end{aligned}
$$

Let

$$
U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(X_i, Y_j).
$$

If

$$
\lim_{N \to \infty} \frac{N}{m} = \exists \rho_{10}, \quad \lim_{N \to \infty} \frac{N}{n} = \exists \rho_{01}
$$

then

$$
\sqrt{N} U \xrightarrow{d} N(0, \sigma_*^2), \quad \sigma_*^2 = \rho_{10} \sigma_{10}^2 + \rho_{01} \sigma_{01}^2
$$

where $N = m + n$, and

$$
\sigma_{10}^2 = \mathrm{Var}_X[\mathrm{E}[\varphi(X, Y)|X]], \quad \sigma_{01}^2 = \mathrm{Var}_Y[\mathrm{E}[\varphi(X, Y)|Y]]
$$

Proof

Let

$$
\begin{aligned}
U^* &= \frac{1}{m} \sum_{i=1}^{m} \varphi_{10}(X_i) + \frac{1}{n} \sum_{j=1}^{n} \varphi_{01}(Y_j), \\
\varphi_{10}(x) &= \mathrm{E}[\varphi(x, Y)], \quad \varphi_{01}(y) = \mathrm{E}[\varphi(X, y)].
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathrm{Var}[U^*] &= \frac{1}{m} \sigma_{10}^2 + \frac{1}{n} \sigma_{01}^2, \\
\sqrt{N} U^* &\xrightarrow{d} N(0, \sigma_*^2).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[U] &= \frac{1}{m} \sigma_{10}^2 + \frac{1}{n} \sigma_{01}^2 + \frac{1}{mn} (\sigma^2 - \sigma_{10}^2 - \sigma_{01}^2), \\
\mathrm{Cov}[U, U^*] &= \frac{1}{m} \sigma_{10}^2 + \frac{1}{n} \sigma_{01}^2.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathrm{E}[N(U - U^*)^2] = N\big\{\mathrm{Var}[U] - 2\mathrm{Cov}[U, U^*] + \mathrm{Var}[U^*]\big\} \\
&\quad = \frac{N}{mn} (\sigma^2 - \sigma_{10}^2 - \sigma_{01}^2) \to 0 \quad \Rightarrow \quad \sqrt{N}(U - U^*) \xrightarrow{P} 0. \\
&\sqrt{N} U = \sqrt{N} U^* + \big\{\sqrt{N}(U - U^*)\big\} \xrightarrow{d} N(0, \sigma_*^2)
\end{aligned}
$$

## 3.2 An estimator associated with Wilcoxon's rank sum statistic (Hodges–Lehmann)

**Testing equality to the specified value**

$$H_0 : \Delta = \Delta_0 \text{ (specified value)}$$

Let

$$S_j(\Delta_0) : \text{ rank of } Y_j - \Delta_0 \text{ in } \{X_1, \cdots, X_m, Y_1 - \Delta_0, \cdots, Y_n - \Delta_0\}$$

**Two–sided test**

$$H_0 \quad \text{versus} \quad H_3 : \Delta \neq \Delta_0$$

$$W(\Delta_0) = \sum_{j=1}^{n} S_j(\Delta_0) \geq w_{\alpha/2} \text{ or } W \leq n(m+n+1) - w_{\alpha/2} \Rightarrow \text{ Reject } H_0$$

**Derivation** Choose the value of $\Delta_0$ such that $H_0 : \Delta = \Delta_0$ is least significant, that is

$$\hat{\Delta} = \text{argmin}_{\Delta_0} \left| W(\Delta_0) - \frac{n(m+n+1)}{2} \right|$$

**Estimator**

$$\hat{\Delta} = \text{median}\{(Y_j - X_i); i = 1, \cdots, m; j = 1, \cdots, n\}$$

$\because$

$$W(\Delta_0) = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j - \Delta_0) + \frac{n(n+1)}{2}.$$

## 3.3 A distribution-free confidence interval based on Wilcoxon's rank sum test (Moses)

**Derivation** Collect $\Delta_0$ such that the two–sided test for

$$H_0 : \Delta = \Delta_0 \text{ versus } H_3 : \Delta \neq \Delta_0$$

does not reject.

**Confidence interval**

$$C_\alpha = \frac{n(2m+n+1)}{2} + 1 - w_{\alpha/2}$$

$$U_{(1)} < U_{(2)} < \cdots < U_{mn} : \text{ ordered statistic of } \{(Y_j - X_i); i = 1, \cdots, m; j = 1, \cdots, n\}$$

$$\Delta_L = U_{(C_\alpha)}, \quad \Delta_U = U_{mn+1-(C_\alpha)}$$

$\Rightarrow$

$$\Pr_{\Delta}\{\Delta_L < \Delta < \Delta_U\} = 1 - \alpha \text{ for all } \Delta.$$

**Large–sample approximation**

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\left\{\frac{mn(m+n+1)}{12}\right\}^{1/2}.$$

# 4 The Two–Sample Dispersion Problem and Other Two–Sample Problems

**Data**

$$X_1, \cdots, X_m, Y_1, \cdots, Y_n$$

**Assumptions**

A1. $X_1, \cdots, X_m$ : a random sample from continuous population 1

$Y_1, \cdots, Y_n$ : a random sample from continuous population 2

that is

$X_1, \cdots, X_m$ : i.i.d. (independent and identically distributed)

$Y_1, \cdots, Y_n$ : i.i.d. (independent and identically distributed)

A2. $(X_1, \cdots, X_m), (Y_1, \cdots, Y_n)$ : independent

## 4.1 A distribution–free rank test for dispersion–case of equal medians (Ansari–Bradley)

**Null hypothesis**

$F, G$ : the distribution functions corresponding to population 1 and 2.

$$H_0 : F(t) = G(t), \text{ for every } t$$

**Alternative hypothesis (Assumption)**

$$H : F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right) \text{ and } G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right)$$

**Additional assumption**

A3. $\theta_1 = \theta_2$

**Testing problems**

$$H_0 : \gamma = 1 \text{ versus } \begin{cases} H_1 : \gamma^2 > 1 \\ H_2 : \gamma^2 < 1 \\ H_3 : \gamma^2 \neq 1 \end{cases},$$

where

$$\gamma = \frac{\eta_1}{\eta_2}$$

**Test statistic**

$$C = \sum_{j=1}^{n} S_j,$$
$$S_j = \frac{N+1}{2} - \left|\frac{N+1}{2} - R_j\right|$$
$$R_j : \text{ rank of } Y_j \text{ in } \{X_1, \cdots, X_m, Y_1, \cdots Y_n\},$$
$$N = m + n$$

Note

$$R_j \le \frac{N+1}{2} \quad \Rightarrow \quad S_j = R_j$$
$$R_j > \frac{N+1}{2} \quad \Rightarrow \quad S_j = N + 1 - R_j$$

**Three types of test**

a. One–sided upper-tail test

$$H_0 : \gamma^2 = 1 \text{ versus } H_1 : \gamma^2 > 1$$
$$C \ge c_\alpha \Rightarrow \text{ Reject } H_0,$$

b. One–sided lower-tail test

$$H_0 : \gamma^2 = 1 \text{ versus } H_2 : \gamma^2 < 1$$
$$C \le [c_{1-\alpha} - 1] \Rightarrow \text{ Reject } H_0,$$

c. Two–sided test

$$H_0 : \gamma^2 = 1 \text{ versus } H_3 : \gamma^2 \ne 1$$
$$C \ge c_{\alpha/2} \text{ or } C \le [c_{1-\alpha/2} - 1] \Rightarrow \text{ Reject } H_0,$$

**Null distribution**

$$(S_1, \cdots, S_n) : \text{ random sample of } n \text{ integers from}$$
$$\begin{cases} \{1, 2, \cdots, \frac{N}{2}, \frac{N}{2}, \frac{N}{2} - 1, \cdots, 1\} & (N \text{ :even }) \\ \{1, 2, \cdots, \frac{N+1}{2}, \frac{N+1}{2} - 1, \cdots, 1\} & (N \text{ :odd }) \end{cases}$$
without replacement

Lemma

$(X_1, \cdots, X_n)$ : random sample of size $n$ from finite population

$$\Pi = (a_1, a_2, \cdots, a_N)$$

without replacement.

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

$\Rightarrow$

$$\mathrm{E}[\bar{X}] = \mu = \frac{1}{N}\sum_{j=1}^{N} a_j,$$

$$\mathrm{Var}[\bar{X}] = \frac{N-n}{n(N-1)}\sigma^2, \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(a_j-\mu)^2$$

Proof.

$$\Pr\{X_j = a_k\} = \frac{(N-1)(N-2)\cdots(N-n+1)}{N(N-1)\cdots(N-n+1)} = \frac{1}{N}$$

$$\Rightarrow \mathrm{E}[X_j] = \mu, \quad \mathrm{Var}[X_j] = \mathrm{E}[X_j^2] - \mu^2 = \frac{1}{N}\sum_{j=1}^{N} a_j^2 - \mu^2 = \sigma^2$$

$$\Pr\{X_i = a_k, X_j = a_m\} = \frac{(N-2)\cdots(N-n+1)}{N(N-1)\cdots(N-n+1)} = \frac{1}{N(N-1)}$$

$$\frac{1}{N}\Big(\sum_{k=1}^{N} a_k\Big)^2 = \frac{1}{N}\sum_{k=1}^{N} a_k^2 + \frac{1}{N}\sum_{k\neq m} a_k a_m$$

$(i \neq j) \Rightarrow$

$$\mathrm{E}[X_i X_j] = \frac{1}{N(N-1)}\sum_{k\neq m} a_k a_m = \frac{1}{N-1}\left\{N\mu^2 - (\sigma^2+\mu^2)\right\} = \mu^2 - \frac{1}{N-1}\sigma^2$$

$$\mathrm{E}[\bar{X}] = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[X_j] = \mu$$

$$\mathrm{Var}[\bar{X}] = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}[X_j] + \frac{1}{n^2}\sum_{i\neq j}\mathrm{Cov}[X_i, X_j] = \frac{1}{n}\sigma^2 - \frac{n(n-1)}{n^2(N-1)}\sigma^2$$

Case of $N$ : even

$$\mathrm{E}_0[C] = \frac{2}{N}\sum_{k=1}^{N/2} k = \frac{n(N+2)}{4}$$

$$\mathrm{Var}_0[C] = \frac{n(N-n)}{N-1}\left\{\frac{2}{N}\sum_{k=1}^{N/2} k^2 - \Big(\frac{N+2}{4}\Big)^2\right\} = \frac{mn(N-2)(N+2)}{48(N-1)}$$

Case of $N$ : odd

$$\mathrm{E}_0[C] = \frac{n(N+1)^2}{4N}$$

$$\mathrm{Var}_0[C] = \frac{mn(N+1)(3+N^2)}{48N^2}$$

**Large–sample approximation**

$$C^* = \frac{C - \mathrm{E}_0[C]}{\sqrt{\mathrm{Var}_0[C]}} \to N(0,1) \quad (m, n \to \infty)$$

$\because$ Under $H_0$

$$F^{-1}(Y_1), \cdots, F^{-1}(Y_n), F^{-1}(X_1), \cdots, F^{-1}(X_m) \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$$

$\big($ uniform distribution on $(0,1)\big)$

The following theorem can be applied since

$$C = \frac{n(N+1)}{2} - \frac{N+2}{2} \sum_{i=1}^{n} \Big| 1 - \frac{2R_i}{N+1} \Big|.$$

Theorem (CLT for a general rank statistic) Let

$$U_1, \cdots, U_n, U_{n+1}, \cdots, U_{m+n} \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$$

$R_i$ : rank of $U_i$

$$S_N = \sum_{i=1}^{N} c_i \phi\Big(\frac{R_i}{N+1}\Big),$$

$N = m + n$,

$\phi$ : Lipschitz continuous function on $[0,1]$ which is not constant

If

$$\lim_{N\to\infty} \frac{\max_i (c_i^{(N)} - \bar{c}^{(N)})^2}{\sum_{i=1}^{N} (c_i^{(N)} - \bar{c}^{(N)})^2} = 0 \quad \Big(\bar{c}^{(N)} = \frac{1}{N} \sum_{i=1}^{N} c_i^{(N)}\Big),$$

Then

$$\frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[S_N]}} \to N(0,1) \quad (N \to \infty)$$

Proof

Define

$$T_N = \sum_{i=1}^{N} d_i \phi(U_i), \quad d_i = c_i^{(N)} - \bar{c}^{(N)}$$

Then

(1) $$\frac{T_N - \mathrm{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}} \to N(0,1)$$

(2) $$\frac{\mathrm{Var}[S_N - T_N]}{\mathrm{Var}[T_N]} \to 0$$

(3) $$\frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[S_N]}} - \frac{T_N - \mathrm{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}} \xrightarrow{P} 0$$

when $(N \to \infty)$.

Proof of (1)

$$g(t) := \log \mathrm{E}[\exp(it\phi(U_i))] = i\mu t - \frac{t^2}{2}\sigma^2 + \varepsilon(t)t^2$$

$$\varepsilon(t) = \frac{1}{2}\{g''(\theta t) - g''(0)\} \quad (0 < \theta < 1)$$

$$\mu = \mathrm{E}[\phi(U_i)] = \frac{1}{i}g'(0), \quad \sigma^2 = \mathrm{Var}[\phi(U_i)] = -g''(0)$$

$$\left|\log \mathrm{E}\left[\exp\left(\frac{T_N - \mathrm{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}}\right)\right] + \frac{t^2}{2}\right| = \left|\sum_{j=1}^{N} \varepsilon\left(\frac{d_j t}{\sqrt{\sum_{j=1}^N d_j^2}}\right)\frac{d_j^2 t^2}{\sum_{j=1}^N d_j^2}\right|$$

$$\leq \max_j \left|\varepsilon\left(\frac{d_j t}{\sqrt{\sum_{j=1}^N d_j^2}}\right)\right| \to 0$$

Proof of (3)

$$\mathcal{L} = \{X : \text{ random variable such that } \mathrm{E}[X] = 0, \mathrm{Var}[X] < \infty\}$$

$$\Longrightarrow \quad \mathrm{Cov}[X, Y] : \text{ an inner product on the linear space } \mathcal{L}$$

$$\sqrt{\mathrm{E}\left[\left(\frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[S_N]}} - \frac{T_N - \mathrm{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}}\right)^2\right]}$$

$$= \left\{\mathrm{E}\left[\left(\frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[S_N]}} - \frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[T_N]}} + \frac{S_N - \mathrm{E}[S_N]}{\sqrt{\mathrm{Var}[T_N]}} - \frac{T_N - \mathrm{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}}\right)^2\right]\right\}^{1/2}$$

$$\leq \frac{\left|\sqrt{\mathrm{Var}[T_N]} - \sqrt{\mathrm{Var}[S_N]}\right|}{\sqrt{\mathrm{Var}[T_N]}} + \sqrt{\frac{\mathrm{Var}[S_N - T_N]}{\mathrm{Var}[T_N]}} \leq 2\sqrt{\frac{\mathrm{Var}[S_N - T_N]}{\mathrm{Var}[T_N]}}$$

Proof of (2)

(2-1) $U_1|_{R_1=i,R_2=j} \sim U_1|_{R_1=i} \quad (i \neq j)$

(2-2) $$\mathrm{Cov}\left[\phi\left(\frac{R_1}{N+1}\right), \phi(U_2)\right] = -\frac{1}{N-1}\mathrm{Cov}\left[\phi\left(\frac{R_1}{N+1}\right), \phi(U_1)\right]$$

(2-3) $$\mathrm{Cov}\left[\phi\left(\frac{R_1}{N+1}\right), \phi\left(\frac{R_2}{N+1}\right)\right] = -\frac{1}{N-1}\mathrm{Var}\left[\left(\frac{R_1}{N+1}\right)\right]$$

(2-4) $$\frac{\mathrm{Var}[S_N - T_N]}{\mathrm{Var}[T_N]} = \left\{\frac{N}{N-1}\frac{\mathrm{Var}\left[\phi\left(\frac{R_1}{N+1}\right) - \phi(U_1)\right]}{\mathrm{Var}[\phi(U_1)]} - \frac{1}{N-1}\right\}$$

(2-5) $$\mathrm{Var}\left[\phi\left(\frac{R_1}{N+1}\right) - \phi(U_1)\right] \to 0$$

Proof of (2-1)

Formula : $\displaystyle\int_a^b (u-a)^j (b-u)^k = (b-a)^{j+k+1}\frac{j!k!}{(j+k+1)!}$

If $i<j$

$$\Pr\{R_1=i,R_2=j\}=\frac{1}{N(N-1)}$$

$$\Pr\{U_1\le t\,|\,R_1=i,R_2=j\}$$

$$=\frac{1}{\mathrm{P}(R_1=i,R_2=j)}\frac{(N-2)!}{(i-1)!(j-i-1)!(N-j)!}$$
$$\cdot\Pr\big\{U_3<U_1,\cdots,U_{i+1}<U_1,U_1<U_{i+2},\cdots,U_{j-1}<U_2,U_2<U_{j+1},\cdots,U_n,U_1\le t\big\}$$
$$=\frac{N!}{(i-1)!(j-i-1)!(N-j)!}\int_0^t\left\{\int_{u_1}^1 u_1^{i-1}(u_2-u_1)^{j-i-1}(1-u_2)^{N-j}du_2\right\}du_1$$
$$=\frac{N!}{(i-1)!(N-i)!}\int_0^t u_1^{i-1}(1-u_1)^{N-i}du_1,$$

$$\Pr\{U_1\le t\,|\,R_1=i\}$$
$$=N\cdot\frac{(N-1)!}{(i-1)!(N-i)!}\Pr\{U_2<U_1,\ldots,U_i<U_1,U_1<U_{i+1},\ldots,U_1<U_N,U_1\le t\}$$
$$=\frac{N!}{(i-1)!(N-i)!}\int_0^t u_1^{i-1}(1-u_1)^{N-i}du_1,$$

Similarly,

$$\Pr\{U_2\le t|R_1=i,R_2=j\}$$
$$=\frac{1}{\Pr\{R_1=i,R_2=j\}}\frac{(N-2)!}{(i-1)!(j-i-1)!(N-j)!}$$
$$\cdot\Pr\big\{U_3<U_1,\cdots,U_{i+1}<U_1,U_1<U_{i+2},\cdots,U_{j-1}<U_2,U_2<U_j,\cdots,U_n,U_2\le t\big\}$$
$$=\frac{N!}{(i-1)!(j-i-1)!(N-j)!}\mathrm{E}[1_{U_2\le t,U_1<U_2}U_1^{i-1}(U_2-U_1)^{j-i-1}(1-U_2)^{N-j}]$$
$$=\frac{N!}{(i-1)!(j-i-1)!(N-j)!}\int_0^t\left\{\int_0^{u_2} u_1^{i-1}(u_2-u_1)^{j-i-1}(1-u_2)^{N-j}du_1\right\}du_2$$
$$=\frac{N!}{(j-1)!(N-j)!}\int_0^t u_2^{j-1}(1-u_2)^{N-j}du_2=\Pr\{U_2\le t|R_2=j\}$$

Proof of (2-2)

$$\psi(j) = \phi\Big(\frac{j}{N+1}\Big) - \frac{1}{N}\sum_{i=1}^{N}\phi\Big(\frac{i}{N+1}\Big)$$

$$\begin{aligned}
\mathrm{Cov}\Big[\phi\Big(\frac{R_1}{N+1}\Big), \phi(U_2)\Big] &= \mathrm{E}\big[\psi(R_1)\phi(U_2)\big] \\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\psi(i)\mathrm{E}\big[\phi(U_2)\big|R_1 = i, R_2 = j\big] \\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\psi(i)\mathrm{E}\big[\phi(U_2)\big|R_2 = j\big] \\
&= \frac{1}{N(N-1)}\Big\{\sum_{i=1}^{N}\psi(i)\Big\}\Big\{\sum_{j=1}^{N}\mathrm{E}\big[\phi(U_2)\big|R_2 = j\big]\Big\} \\
&\quad - \frac{1}{N(N-1)}\sum_{i=1}^{N}\psi(i)\mathrm{E}\big[\phi(U_2)\big|R_2 = i\big] \\
&= -\frac{1}{N-1}\mathrm{E}\big[\psi(R_2)\phi(U_2)\big] = -\frac{1}{N-1}\mathrm{Cov}\big[\psi(R_2), \phi(U_2)\big]
\end{aligned}$$

Proof of (2-3)

$$\begin{aligned}
\mathrm{Cov}\Big[\phi\Big(\frac{R_1}{N+1}\Big), \phi\Big(\frac{R_2}{N+1}\Big)\Big] &= \mathrm{E}[\psi(R_1), \psi(R_2)] \\
&= \frac{1}{N(N-1)}\sum_{i\neq j}\psi(i)\psi(j) = \frac{1}{N(N-1)}\Big\{\sum_{i=1}^{N}\psi(i)\Big\}^2 - \frac{1}{N(N-1)}\sum_{i=1}^{N}\{\psi(i)\}^2 \\
&= -\frac{1}{N-1}\mathrm{Var}\Big[\phi\Big(\frac{R_1}{N+1}\Big)\Big]
\end{aligned}$$

Proof of (2-4)

$$\begin{aligned}
&\mathrm{Cov}[\psi(R_1) - \phi(U_1), \psi(R_2) - \phi(U_2)] = \mathrm{Cov}[\psi(R_1), \psi(R_2)] - 2\mathrm{Cov}[\psi(R_1), \phi(U_2)] \\
&\quad = -\frac{1}{N-1}\Big\{\mathrm{Var}[\psi(R_1)] - 2\mathrm{Cov}[\psi(R_1), \phi(U_1)]\Big\} \\
&0 = \Big\{\sum_{i=1}^{N} d_i\Big\}^2 = \sum_{i=1}^{N} d_i^2 + \sum_{i\neq j} d_i d_j \\
&\mathrm{Var}[S_N - T_N] = \sum_{i=1}^{N} d_i^2 \mathrm{Var}[\psi(R_i) - \phi(U_i)] + \sum_{i\neq j} d_i d_j \mathrm{Cov}[\psi(R_i) - \phi(U_i), \psi(R_j) - \phi(U_j)] \\
&\quad = \Big(\sum_{i=1}^{N} d_i^2\Big)\Big\{\frac{N}{N-1}\Big\{\mathrm{Var}[\psi(R_1)] - 2\mathrm{Cov}[\psi(R_1), \phi(U_1)]\Big\} + \mathrm{Var}[\phi(U_1)]\Big\} \\
&\mathrm{Var}[T_N] = \Big(\sum_{i=1}^{N} d_i^2\Big)\mathrm{Var}[\phi(U_1)]
\end{aligned}$$

Proof of (2-5)

$$f_{U_1|R_1=j}(u) = \frac{N!}{(j-1)!(N-j)!}u^{j-1}(1-u)^{N-j}$$

$$\mathrm{Var}[U_1|R_1=j] = \mathrm{E}[U_1^2|R_1=j] - \mathrm{E}[U_1|R_1=j]^2$$

$$= \frac{N!}{(j-1)!(N-j)!}\frac{(j+1)!(N-j)!}{(N+2)!} - \left(\frac{N!}{(j-1)!(N-j)!}\frac{j!(N-j)!}{(N+1)!}\right)^2$$

$$= \frac{j(N+1-j)}{(N+1)^2(N+2)} \qquad \text{(Formula in page 22)}$$

$$\mathrm{E}\Big[\Big\{\phi\Big(\frac{R_1}{N+1}\Big) - \phi(U_1)\Big\}^2\Big] = \frac{1}{N}\sum_{j=1}^{N}\mathrm{E}\Big[\Big\{\phi\Big(\frac{j}{N+1}\Big) - \phi(U_1)\Big\}^2\Big|R_1=j\Big]$$

$$= \frac{1}{N}\sum_{j=1}^{N}\Big[\int_{|u-\frac{j}{N+1}|<\varepsilon}\Big\{\phi\Big(\frac{j}{N+1}\Big) - \phi(u)\Big\}^2 f_{U_1|R_1=j}(u)du$$

$$\int_{|u-\frac{j}{N+1}|>\varepsilon}\Big\{\phi\Big(\frac{j}{N+1}\Big) - \phi(u)\Big\}^2 f_{U_1|R_1=j}(u)du\Big]$$

$$< \frac{1}{N}\sum_{j=1}^{N}\Big[D\varepsilon^2\Pr\Big\{\Big|U_1 - \frac{j}{N+1}\Big| < \varepsilon \,|\, R_1=j\Big\} + M\Pr\Big\{\Big|U_1 - \frac{j}{N+1}\Big| > \varepsilon \,|\, R_1=j\Big\}\Big]$$

$$\Big(D = \max_{x<y}\Big\{\frac{\phi(x)-\phi(y)}{x-y}\Big\}^2, M = \max_{x<y}\{\phi(x)-\phi(y)\}^2\Big)$$

$$< \frac{1}{N}\sum_{j=1}^{N}\Big\{D\varepsilon^2 + \frac{M}{\varepsilon^2}\mathrm{Var}[U_1|R_1=j]\Big\} = \frac{1}{N}\sum_{j=1}^{N}\Big\{D\varepsilon^2 + \frac{M}{\varepsilon^2}\frac{j(N+1-j)}{(N+1)^2(N+2)}\Big\}$$

$$= D\varepsilon^2 + \frac{M}{\varepsilon^2}\Big\{\frac{1}{2(N+2)} - \frac{2N+1}{6(N+1)(N+2)}\Big\} = D\varepsilon^2 + \frac{M}{6\varepsilon^2(N+1)}$$

$$\mathrm{E}\Big[\Big\{\phi\Big(\frac{R_1}{N+1}\Big) - \phi(U_1)\Big\}^2\Big] < 2\sqrt{\frac{DM}{N+1}} \qquad \Big(\varepsilon^2 = \sqrt{\frac{M}{D(N+1)}}\Big)$$

## 4.2 An asymptotically distribution–free test for dispersion based on the Jackknife–medians not necessarily equal (Miller)

**Assumptions**

A1. $X_1, \cdots, X_m$ : a random sample from continuous population 1

$Y_1, \cdots, Y_n$ : a random sample from continuous population 2

that is

$X_1, \cdots, X_m$ : i.i.d. (independent and identically distributed)

$Y_1, \cdots, Y_n$ : i.i.d. (independent and identically distributed)

A2. $(X_1, \cdots, X_m), (Y_1, \cdots, Y_n)$ : independent

**Further hypothesis (Assumption)**

$$H : F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right) \text{ and } G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right)$$

**Additional assumption**

A4. $H$ has finite fourth moment.

$$\mathrm{E}\Big[\Big(\frac{X_1-\theta_1}{\eta_1}\Big)^4\Big] < \infty$$

**Procedure**

$$\bar{X}_i = \sum_{s\neq i}^{m} \frac{X_s}{m-1}, \quad D_i^2 = \sum_{s\neq i} \frac{(X_s-\bar{X}_i)^2}{m-2} \quad (i = 1, 2, \cdots, m)$$

$$\bar{Y}_j = \sum_{t\neq i}^{n} \frac{Y_t}{n-1}, \quad E_j^2 = \sum_{t\neq i} \frac{(Y_t-\bar{Y}_j)^2}{n-2} \quad (i = 1, 2, \cdots, n)$$

$$S_i = \log D_i^2 \quad (i = 1, \cdots, m)$$

$$T_j = \log D_j^2 \quad (j = 1, \cdots, n)$$

$$\bar{X}_0 = \frac{1}{m}\sum_{s=1}^{m} X_s, \quad S_0 = \log\left[\sum_{s=1}^{m} \frac{(X_s-\bar{X}_0)^2}{m-1}\right],$$

$$\bar{Y}_0 = \frac{1}{n}\sum_{t=1}^{n} Y_t, \quad T_0 = \log\left[\sum_{t=1}^{n} \frac{(Y_t-\bar{Y}_0)^2}{n-1}\right]$$

$$A_i = mS_0 - (m-1)S_i \quad (i = 1, \cdots, m)$$

$$B_j = nT_0 - (n-1)T_j \quad (j = 1, \cdots, n)$$

$$\bar{A} = \sum_{i=1}^{m} \frac{A_i}{m}, \quad \bar{B} = \sum_{j=1}^{n} \frac{B_j}{n}$$

$$V_1 = \sum_{i=1}^{m} \frac{(A_i-\bar{A})^2}{m(m-1)}, \quad V_2 = \sum_{j=1}^{n} \frac{(B_j-\bar{B})^2}{n(n-1)}$$

Test statistic

$$Q = \frac{\bar{A}-\bar{B}}{\sqrt{V_1+V_2}}$$

a. One–sided upper-tail test

$$H_0 : \gamma^2 = 1 \text{ versus } H_1 : \gamma^2 > 1$$
$$Q \geq z_\alpha \Rightarrow \text{ Reject } H_0,$$

b. One–sided lower-tail test

$$H_0 : \gamma^2 = 1 \text{ versus } H_2 : \gamma^2 < 1$$
$$Q \leq -z_\alpha \Rightarrow \text{ Reject } H_0,$$

c. Two–sided test

$$H_0 : \gamma^2 = 1 \text{ versus } H_3 : \gamma^2 \neq 1$$
$$|Q| \geq z_{\alpha/2} \Rightarrow \text{ Reject } H_0,$$

**Jackknife technique**

$$X_1, \cdots, X_n \overset{i.i.d.}{\sim} F, \quad \mathrm{E}[X_1] = \mu, \mathrm{Var}[X_1] = \sigma^2$$
$$\hat{\theta}_n : \text{an estimator based on } X_1, \cdots, X_n$$
$$\hat{\theta}_{n-1}^{(i)} : \text{the estimator based on } X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n$$
$$\tilde{\theta}_{n,i} = n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1}^{(i)} : \text{jackknife pseudo–value}$$

Tukey's conjecture

(A1) $\tilde{\theta}_{n,1}, \cdots, \tilde{\theta}_{n,n}$ can be vied as an *i.i.d.* sample

(A2) $\mathrm{Var}[\tilde{\theta}_{n,i}] \approx \mathrm{Var}[\sqrt{n}\hat{\theta}_n]$

Jackknife estimator

$$\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_{n,i} : \text{Bias modified Jackknife estimator}$$
$$V = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{n-1}^{(i)} - \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{n-1}^{(j)}\right)^2$$
$$: \text{Jackknife variance estimator for } \mathrm{Var}[\hat{\theta}_n]$$

Case of a smooth function of the sample mean

$$X_1, \cdots, X_n \overset{i.i.d.}{\sim} F, \quad \mathrm{E}[X_1] = \mu, \mathrm{Var}[X_1] = \sigma^2$$
$$\hat{\theta}_n = h(\bar{X}_n), \quad h : C^{(3)}\text{–class}, \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\theta}_n^{(i)} = h(\bar{X}_{n-1}^{(i)}), \quad \bar{X}_{n-1}^{(i)} = \frac{1}{n-1}\sum_{j\neq i}^{n} X_j$$

Asymptotic normality of $\hat{\theta}_n$

$$Z_n = \sqrt{n}(\bar{X}_n - \mu) \longrightarrow N(0, \sigma^2) \quad \text{central limit theorem}$$
$$h(\bar{X}_n) = h\Big(\mu + \frac{1}{\sqrt{n}}Z_n\Big) = h(\mu) + \frac{1}{\sqrt{n}}h'(\mu)Z_n + \frac{1}{2n}h''(\mu)Z_n^2 + \cdots$$
$$\sqrt{n}\{h(\bar{X}_n) - h(\mu)\} = h'(\mu)Z_n + \frac{1}{2\sqrt{n}}h''(\mu)Z_n^2 + \cdots \longrightarrow N\big[0, \{h'(\mu)\}^2\sigma^2\big]$$

Bias of the jackknife estimator

$$\bar{X}_n = \frac{n-1}{n}\bar{X}^{(i)}_{n-1} + \frac{1}{n}X_i = \bar{X}^{(i)}_{n-1} + \frac{1}{n}(X_i - \bar{X}^{(i)}_{n-1})$$
$$= \mu + \frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1} + \frac{1}{n}(X_i - \mu) - \frac{1}{n\sqrt{n-1}}Z^{(i)}_{n-1}, \quad Z^{(i)}_{n-1} = \sqrt{n-1}(\bar{X}^{(i)}_{n-1} - \mu)$$
$$h(\bar{X}_n) = h(\mu) + h'(\mu)\Big\{\frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1} + \frac{1}{n}(X_i - \mu) - \frac{1}{n\sqrt{n-1}}Z^{(i)}_{n-1}\Big\}$$
$$+ \frac{1}{2}h''(\mu)\Big\{\frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1} + \frac{1}{n}(X_i - \mu) - \frac{1}{n\sqrt{n-1}}Z^{(i)}_{n-1}\Big\}^2$$
$$+ \frac{1}{6}h^{(3)}(\mu)\Big\{\frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1} + \frac{1}{n}(X_i - \mu) - \frac{1}{n\sqrt{n-1}}Z^{(i)}_{n-1}\Big\}^3 + O_p(n^{-2})$$
$$h(\bar{X}^{(i)}_{n-1}) = h(\mu) + h'(\mu)\frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1}$$
$$+ \frac{1}{2}h''(\mu)\frac{1}{n-1}\{Z^{(i)}_{n-1}\}^2 + \frac{1}{6}h^{(3)}(\mu)\frac{1}{(n-1)^{3/2}}\{Z^{(i)}_{n-1}\}^3 + O_p(n^{-2})$$
$$\tilde{\theta}_{n,i} = nh(\bar{X}_n) - (n-1)h(\bar{X}^{(i)}_{n-1})$$
$$= h(\mu) + \frac{h'(\mu)}{\sqrt{n-1}}Z^{(i)}_{n-1} + \frac{h''(\mu)}{2(n-1)}\{Z^{(i)}_{n-1}\}^2 + \frac{h^{(3)}(\mu)}{6(n-1)^{3/2}}\{Z^{(i)}_{n-1}\}^3$$
$$+ h'(\mu)\Big\{(X_i - \mu) - \frac{1}{\sqrt{n-1}}Z^{(i)}_{n-1}\Big\}$$
$$+ \frac{h''(\mu)}{2}\Big\{\frac{2}{\sqrt{n-1}}Z^{(i)}_{n-1}(X_i - \mu) + \frac{1}{n}(X_i - \mu)^2 - \frac{2}{n-1}\{Z^{(i)}_{n-1}\}^2\Big\}$$
$$+ \frac{h^{(3)}(\mu)}{6}\Big\{\frac{3}{n-1}\{Z^{(i)}_{n-1}\}^2(X_i - \mu)\Big\} + O_p(n^{-3/2})$$
$$= h(\mu) + h'(\mu)(X_i - \mu) + \frac{h''(\mu)}{\sqrt{n-1}}Z^{(i)}_{n-1}(X_i - \mu)$$
$$+ \frac{h''(\mu)}{2}\Big\{\frac{1}{n}(X_i - \mu)^2 - \frac{1}{(n-1)}\{Z^{(i)}_{n-1}\}^2\Big\}$$
$$+ \frac{h^{(3)}(\mu)}{2(n-1)}\{Z^{(i)}_{n-1}\}^2(X_i - \mu) + O_p(n^{-3/2})$$

$Z^{(i)}_{n-1}, (X_i)$ : independent

$$\Rightarrow \quad \mathrm{E}[\tilde{\theta}_{n,i}] = h(\mu) + \frac{h''(\mu)}{2}\Big\{\frac{1}{n} - \frac{1}{n-1}\Big\}\sigma^2 + O(n^{-3/2}) = h(\mu) + O(n^{-3/2})$$

$$(i \neq j) \Rightarrow \mathrm{E}[\{\tilde{\theta}_{n,i} - h(\mu)\}\{\tilde{\theta}_{n,j} - h(\mu)\}]$$
$$= \frac{\{h''(\mu)\}^2}{n-1}\mathrm{E}[Z^{(i)}_{n-1}Z^{(j)}_{n-1}(X_i - \mu)(X_j - \mu)] + O(n^{-3/2}) = O(n^{-3/2})$$

$$
\begin{aligned}
\mathrm{E}[\{\tilde{\theta}_{n,i} - h(\mu)\}^2] &= \{h'(\mu)\}^2 \mathrm{E}[(X_i-\mu)^2] + \frac{h'(\mu)h''(\mu)}{n}\mathrm{E}[(X_i-\mu)^3] \\
&\quad + \frac{h'(\mu)h^{(3)}(\mu)}{n-1}\mathrm{E}[\{Z_{n-1}^{(i)}\}^2(X_i-\mu)^2] + \frac{\{h''(\mu)\}^2}{n-1}\mathrm{E}[\{Z_{n-1}^{(i)}\}^2(X_i-\mu)^2] \\
&\quad + O(n^{-3/2}) \qquad (\kappa_3 = \mathrm{E}[(X_i-\mu)^3]) \\
&= \{h'(\mu)\}^2\sigma^2 + \frac{h'(\mu)h''(\mu)}{n}\kappa_3 + \frac{1}{n-1}\big[h'(\mu)h^{(3)}(\mu) + \{h''(\mu)\}^2\big]\sigma^4 + O(n^{-3/2})
\end{aligned}
$$

Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ then

$$
\begin{aligned}
\hat{\theta}_n = h(\bar{X}_n) &= h(\mu) + \frac{h'(\mu)}{\sqrt{n}}Z_n + \frac{h''(\mu)}{2n}Z_n^2 + \frac{h^{(3)}(\mu)}{6n}Z_n^3 + O_p(n^{-3/2}) \\
\mathrm{Var}[\sqrt{n}\hat{\theta}_n] &= \{h'(\mu)\}^2\sigma^2 + \frac{h'(\mu)h''(\mu)}{n}\kappa_3 \\
&\quad + \frac{1}{n}\big[h'(\mu)h^{(3)}(\mu) + \frac{1}{2}\{h''(\mu)\}^2\big]\sigma^4 + O(n^{-3/2})
\end{aligned}
$$

So

$$
\mathrm{Var}[\sqrt{n}\hat{\theta}_n] - \mathrm{Var}[\tilde{\theta}_n] = O(n^{-1})
$$

Asymptotic normality of $\tilde{\theta}_n$

$$
\begin{aligned}
&\bar{X}_{n-1}^{(i)} = \bar{X}_n - \frac{1}{n-1}(\bar{X}_n - X_i) \\
&nh(\bar{X}_n) - (n-1)h(\bar{X}_{n-1}^{(i)}) \\
&\qquad = h(\bar{X}_n) - h'(\bar{X}_n)(\bar{X}_n - X_i) - \frac{h''(\bar{X}_n)}{2(n-1)}(\bar{X}_n - X_i)^2 + O_p(n^{-2}) \\
&\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_{n,i} = h(\bar{X}_n) - \frac{h''(\bar{X}_n)}{2n}S_n + O_p(n^{-2}), \qquad S_n = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}_n - X_i)^2 \\
&\sqrt{n}\big\{\hat{\theta}_n - h(\mu)\big\} - \sqrt{n}\big\{\tilde{\theta}_n - h(\mu)\big\} = O_p(n^{-1/2}) \quad (\because \bar{X}_n \xrightarrow{P} \mu,\ S_n \xrightarrow{P} \sigma^2)
\end{aligned}
$$

## 4.3 A distribution–free rank test for either location or dispersion (Lepage)

**Assumptions**

A1. $X_1, \cdots, X_m$ : a random sample from continuous population 1

$Y_1, \cdots, Y_n$ : a random sample from continuous population 2

that is

$X_1, \cdots, X_m$ : i.i.d. (independent and identically distributed)

$Y_1, \cdots, Y_n$ : i.i.d. (independent and identically distributed)

A2. $(X_1, \cdots, X_m), (Y_1, \cdots, Y_n)$ : independent

**Further hypothesis (Assumption)**

$$H : F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right) \text{ and } G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right)$$

**Testing problem**

$$H_0 : \theta_1 = \theta_2 \text{ and } \eta_1 = \eta_2 \text{ versus } H_1 : \theta_1 \neq \theta_2 \text{ and/or } \eta_1 \neq \eta_2$$

**Test statistic**

$R_i$ : rank of $Y_i$ in $X_1, \cdots, X_m, Y_1, \cdots, Y_n$

$W = \sum_{i=1}^{n} R_i$ : Wilcoxon rank sum statistic ,

$C = \sum_{i=1}^{n} \left\{ \frac{N+1}{2} - \left| \frac{N+1}{2} - R_i \right| \right\}$ : Ansari–Bradley scale statistic

$D = (W^*)^2 + (C^*)^2$, where $W^* = \dfrac{W - \mathrm{E}_0[W]}{\sqrt{\mathrm{Var}_0[W]}}$ and $C^* = \dfrac{C - \mathrm{E}_0[C]}{\sqrt{\mathrm{Var}_0[C]}}$

**Procedure**

$D \geq d_\alpha \Rightarrow$ reject $H_0$

$d_\alpha$ is obtained based on the fact that

$(R_1, \cdots, R_n)$ is a random sample from $\{1, 2, \cdots, N\}$ without replacement

**Large–sample approximation**

$D \longrightarrow \chi^2_2, \quad (\min\{m, n\} \to \infty)$

$D \geq \chi^2_{2,\alpha} \Rightarrow$ reject

$\chi^2_2$ : chi–square distribution with degree of freedom 2

$\chi^2_{2,\alpha}$ : upper $100\alpha$ percent point of $\chi^2_2$

$\because$

Let

$$T_W = \frac{m}{N}\sum_{i=1}^{n} U_i - \frac{n}{N}\sum_{i=1}^{m} U_{n+i}$$

$$T_C = \frac{m}{N}\sum_{i=1}^{n} \phi(U_i) - \frac{n}{N}\sum_{i=1}^{m} \phi(U_{n+i}) \quad \phi(u) = \left|1 - 2u\right|.$$

where $U_i = F^{-1}(X_i) \quad i = 1, \cdots, m; \quad U_i = F^{-1}(Y_i) \quad i = 1, \cdots, n$

Then from the proof of Theorem in section 4.1,

$$W^* - \frac{T_W}{\sqrt{\mathrm{Var}(T_W)}} \xrightarrow{P} 0, \quad C^* - \frac{T_C}{\sqrt{\mathrm{Var}(T_C)}} \xrightarrow{P} 0$$

Let $a_1, a_2$ : any constants, and

$$T_a = a_1 T_W + a_2 T_C = \frac{m}{N}\sum_{i=1}^{n}\{a_1 U_i + a_2\phi(U_i)\} - \frac{n}{N}\sum_{i=1}^{m}\{a_1 U_{i+n} + a_2\phi(U_{i+n})\}$$

Then

$$\mathrm{E}[T_a] = 0 \text{ and } \mathrm{Var}[T_a] = \frac{mn}{12N}(a_1^2 + a_2^2)$$
$$\frac{T_a}{\sqrt{\mathrm{Var}[T_a]}} = \frac{a_1 T_W^* + a_2 T_C^*}{\sqrt{a_1^2 + a_2^2}} \to N(0,1) \quad (\min\{m,n\} \to \infty)$$
$$T_W^* = \frac{T_W}{\sqrt{\mathrm{Var}(T_W)}}, \quad T_C^* = \frac{T_C}{\sqrt{\mathrm{Var}(T_C)}},$$

Hence

$$\begin{pmatrix} T_W^* \\ T_C^* \end{pmatrix} \to N_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

## 4.4 A distribution–free test for general differences in two populations (Kolmogorov–Smirnov)

**Assumptions**

A1. $X_1, \cdots, X_m \overset{i.i.d.}{\sim} F, \quad Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} G$

A2. $(X_1, \cdots, X_m), (Y_1, \cdots, Y_n)$ : independent

**Testing problem**

$$H_0 : F(t) = G(t) \text{ for any } t \quad \text{versus} \quad H_1 : F(t) \neq G(t) \text{ for at least one } t$$

**Test statistic**

$$J = \frac{mn}{d} \max_{-\infty<t<\infty}\{|F_m(t) - F_n(t)|\},$$

where

$d$ : greatest common divisor of $m$ and $n$

$$F_m(t) = \frac{1}{m}\#\{X_i | X_i \leq t, i = 1, \cdots, m\}$$
(the empirical distribution function for $X$)

$$G_n(t) = \frac{1}{m}\#\{Y_i | Y_i \leq t, i = 1, \cdots, n\}$$
(the empirical distribution function for $Y$)

**Procedure**

$$J \geq j_\alpha \Rightarrow \text{ reject } H_0$$

Under $H_0$ the conditional distribution of $(X_1, \cdots, X_m)$ can be viewed as the one of random sample from $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(N)}$ without replacement.

**Deriving $J$**

data

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 3 | 3 | 5 | 7 | 9 | | | | | |
| $Y$ | 3 | 4 | 4 | 6 | 7 | 8 | 10 | 10 | 11 | 12 |

$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_N$ : ordered statistic of $X_1, \cdots, X_m, Y_1, \cdots, Y_n$

| $i$ | $Z_{(i)}$ | $F_5(Z_{(i)})$ | $G_{10}(Z_{(i)})$ | $\lvert F_5(Z_{(5)}) - G_{10}(Z_{(i)})\rvert$ |
|---|---|---|---|---|
| 1 | 3 | 2/5 | 1/10 | 3/10 |
| 2 | 3 | 2/5 | 1/10 | 3/10 |
| 3 | 3 | 2/5 | 1/10 | 3/10 |
| 4 | 4 | 2/5 | 3/10 | 1/10 |
| 5 | 4 | 2/5 | 3/10 | 1/10 |
| 6 | 5 | 3/5 | 3/10 | 3/10 |
| 7 | 6 | 3/5 | 4/10 | 2/10 |
| 8 | 7 | 4/5 | 5/10 | 3/10 |
| 9 | 7 | 4/5 | 5/10 | 3/10 |
| 10 | 8 | 4/5 | 6/10 | 2/10 |
| 11 | 9 | 5/5 | 6/10 | 4/10 |
| 12 | 10 | 5/5 | 8/10 | 2/10 |
| 13 | 10 | 5/5 | 8/10 | 2/10 |
| 14 | 11 | 5/5 | 9/10 | 1/10 |
| 15 | 12 | 5/5 | 10/10 | 0 |

$$\Rightarrow J = \frac{10\times 5}{5}\frac{4}{10} = 4$$

**Large–sample approximation**

$$J^* = \sqrt{\frac{mn}{N}} \max_{-\infty<t\infty} \{|F_m(t) - F_n(t)|\}$$

$$\mathrm{P}_0(J^* < s) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, 0 \text{ for } s >, \leq 0 \quad (\min\{m, n\} \to \infty)$$

# 5 The One-Way Layout

**Data**

| Treatments | | | |
|---|---|---|---|
| 1 | 2 | $\cdots$ | k |
| $X_{11}$ | $X_{12}$ | $\cdots$ | $X_{1k}$ |
| $X_{21}$ | $X_{22}$ | $\cdots$ | $X_{2k}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ |
| $X_{n_1 1}$ | $X_{n_2 2}$ | $\cdots$ | $X_{n_k k}$ |

**Assumptions**

A1. $X_{11}, X_{21}, \cdots, X_{n_1 1}, \cdots, X_{1k}, \cdots, X_{n_k k}$ : independent

A2. $X_{1j}, X_{2j}, \cdots, X_{n_j j} \overset{i.i.d.}{\sim} F_j$

A3. $F_1, \cdots, F_k$ belong to the same location family of distributions:

$$F_j(t) = F(t - \tau_j), \quad -\infty < t < \infty, \quad j = 1, \cdots, k$$

$\tau_j$ : unknown parameter

$F$ : continuous distribution function with unknown median $\theta$

One–way layout model

$$X_{ij} = \theta + \tau_j + e_{ij} \quad i = 1, \cdots, n_j, \quad j = 1, \cdots, k$$

$\tau_j$ : treatment $j$ effect

$e_{ij}$ : i.i.d. , median$[e_{ij}] = 0$

**Hypothesis**

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

## 5.1 A distribution–free test for general alternatives (Kruskal–Wallis)

**Testing problem**

$$H_0 \quad \text{versus } H_1 : \tau_1, \cdots \tau_k \text{ are not all equal}$$

**Test statistic**

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

$$N = \sum_{j=1}^{k} n_j, \quad R_j = \sum_{i=1}^{n_j} r_{ij}, \quad R_{\cdot j} = \frac{R_j}{n_j}, \quad j = 1, \cdots, k$$

$r_{ij}$ : rank of $X_{ij}$ in all observations

**Procedure**

$$H \geq h_\alpha \Rightarrow \text{ reject } H_0$$

$h_\alpha$ is derived based on the fact that

> Under $H_0$, all $N!/(\prod_{j=1}^{k} n_j!)$ assignments of $n_1$ ranks to the treatment 1 observations, $n_2$ ranks to the treatment 2 observations, $\cdots, n_k$ ranks to the treatment k observations, are equally likely.

Example for deriving the exact null distribution

$$k = 3, n_1 = n_2 = n_3 = 2, \quad \text{Number of rank assignments } = \frac{6!}{2!2!2!} = 90$$

$$H = \frac{12}{6(6+1)} \frac{R_1^2 + R_2^2 + R_3{}^2}{2} - 3(6+1) = \frac{A}{7} - 21, \quad A = R_1^2 + R_2^2 + R_3{}^2$$

| | I | | II | | III | | |
|---|---|---|---|---|---|---|---|
| (a) | 1 | 2 | 3 | 4 | 5 | 6 | $A = 179, H = 4.57$ |
| (b) | 1 | 2 | 3 | 5 | 4 | 6 | $A = 173, H = 3.71$ |
| (c) | 1 | 2 | 3 | 6 | 4 | 5 | $A = 171, H = 3.43$ |
| (d) | 1 | 3 | 2 | 4 | 5 | 6 | $A = 173, H = 3.71$ |
| (e) | 1 | 3 | 2 | 5 | 4 | 6 | $A = 165, H = 2.57$ |
| (f) | 1 | 3 | 2 | 6 | 4 | 5 | $A = 161, H = 2.00$ |
| (g) | 1 | 4 | 2 | 3 | 5 | 6 | $A = 171, H = 3.43$ |
| (h) | 1 | 4 | 2 | 5 | 3 | 6 | $A = 155, H = 1.14$ |
| (i) | 1 | 4 | 2 | 6 | 3 | 5 | $A = 153, H = 0.857$ |
| (j) | 1 | 5 | 2 | 3 | 4 | 6 | $A = 161, H = 2.00$ |
| (k) | 1 | 5 | 2 | 4 | 3 | 6 | $A = 153, H = 0.857$ |
| (l) | 1 | 5 | 2 | 6 | 3 | 4 | $A = 149, H = 0.286$ |
| (m) | 1 | 6 | 2 | 3 | 4 | 5 | $A = 155, H = 1.14$ |
| (n) | 1 | 6 | 2 | 4 | 3 | 5 | $A = 149, H = 0.286$ |
| (o) | 1 | 6 | 2 | 5 | 3 | 4 | $A = 147, H = 0$ |

(I,II,III) $\Rightarrow$ (I,III,II), (II,I,III),(II,III,I),(III,I,II),(III,II,I)

| $h$ | 0 | 0.286 | 0.857 | 1.14 | 2.00 | 2.57 | 3.43 | 3.71 | 4.57 |
|---|---|---|---|---|---|---|---|---|---|
| $\Pr\{H = h\}$ | $\frac{1}{15}$ | $\frac{2}{15}$ | $\frac{2}{15}$ | $\frac{2}{15}$ | $\frac{2}{15}$ | $\frac{1}{15}$ | $\frac{2}{15}$ | $\frac{2}{15}$ | $\frac{1}{15}$ |

**Large–sample approximation**

$$H \longrightarrow \chi^2_{k-1}, \quad (\min\{n_1, \cdots, n_k\} \to \infty)$$

$$H \geq \chi^2_{k-1,\alpha} \Rightarrow \text{ reject } H_0$$

Proof

$$S_n := \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{a_j}{n_j} r_{ij}, \quad a_1, \cdots, a_k : \text{ constants}$$

Then

$$\frac{S_N - E[S_N]}{\sqrt{\mathrm{Var}[S_N]}} \to N(0,1)$$

$\because$ Theorem in section 4.1 can be applied since

$$c_{ij}^{(N)} := \frac{N+1}{n_j} a_j \quad (i = 1, \cdots, n_j; j = 1, \cdots, k)$$

$$\bar{c}^{(N)} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n_j} c_{ij}^{(N)} = \frac{N+1}{N} \sum_{j=1}^{k} a_j = \frac{N+1}{N} A \text{ (say)}$$

$$\frac{\max_{i,j}(c_{ij}^{(N)} - \bar{c}^{(N)})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (c_{ij}^{(N)} - \bar{c}^{(N)})^2} = \frac{\max_j (\frac{a_j}{n_j} - \frac{A}{N})^2}{\sum_{j=1}^{k} n_j (\frac{a_j}{n_j} - \frac{A}{N})^2} < \frac{1}{n_{j_*}} \to 0, \quad j_* = \mathrm{argmax}_j (\frac{a_j}{n_j} - \frac{A}{N})^2$$

$$\mathrm{E}[r_{ij}] = \frac{1}{N} \sum_{l=1}^{N} l = \frac{N+1}{2} = \mu \text{ (say)} ,$$

$$\mathrm{Var}[r_{ij}] = \frac{1}{N} \sum_{l=1}^{N} l^2 - \mu^2 = \frac{(N+1)(N-1)}{12} = \sigma^2 \text{ (say)} ,$$

$$\begin{aligned} \mathrm{Cov}[r_{ij}, r_{i'j'}] &= \frac{1}{N(N-1)} \sum_{l \neq l'} l l' - \mu^2 \\ &= \frac{N}{N-1} \left\{ \left( \frac{1}{N} \sum_{l=1}^{N} l \right)^2 - \frac{1}{N^2} \sum_{l=1}^{N} l^2 \right\} - \mu^2 \\ &= \frac{N}{N-1} \left\{ \mu^2 - \frac{1}{N} (\sigma^2 + \mu^2) \right\} - \mu^2 = -\frac{1}{N-1} \sigma^2 \end{aligned}$$

$$\mathrm{E}[S_N] = \left(\sum_{j=1}^{k} a_j\right)\mu$$

$$\mathrm{Var}[S_N] = \sum_{j=1}^{k} \mathrm{Var}\left[\sum_{i=1}^{n_j} \frac{a_j}{n_j} r_{ij}\right] + \sum_{j\neq j'}^{k} \mathrm{Cov}\left[\sum_{i=1}^{n_j} \frac{a_j}{n_j} r_{ij}, \sum_{i=1}^{n_{j'}} \frac{a_{j'}}{n_{j'}} r_{ij'}\right]$$

$$\mathrm{Var}\left[\sum_{i=1}^{n_j} \frac{a_j}{n_j} r_{ij}\right] = \left(\frac{a_j}{n_j}\right)^2 \left\{\sum_{i=1}^{n_j} \mathrm{Var}[r_{ij}] + \sum_{i\neq i'}^{n_j} \mathrm{Cov}[r_{ij}, r_{i'j}]\right\}$$

$$= \left(\frac{a_j}{n_j}\right)^2 \left\{n_j - \frac{n_j(n_{j-1})}{N-1}\right\}\sigma^2$$

$$\mathrm{Cov}\left[\sum_{i=1}^{n_j} \frac{a_j}{n_j} r_{ij}, \sum_{i=1}^{n_{j'}} \frac{a_{j'}}{n_{j'}} r_{ij'}\right] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_{j'}} \frac{a_j a_{j'}}{n_j n_{j'}}\left(-\frac{\sigma^2}{N-1}\right)$$

$$\mathrm{Var}[S_N] = (a_1, \cdots, a_k)' B \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} \sigma^2,$$

$$B = (b_{jl}), \quad b_{jl} = \begin{cases} \frac{1}{n_j} - \frac{1}{N-1}\frac{n_j-1}{n_j} & (j = l) \\ -\frac{1}{N-1} & (j \neq l) \end{cases}$$

$$B = \frac{N}{N-1} \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_k} \end{pmatrix} - \frac{1}{N-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (1, \cdots, 1)$$

$$= \frac{N}{N-1} \begin{pmatrix} \frac{1}{\sqrt{n_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{n_k}} \end{pmatrix} \left\{ I_k - \frac{1}{N} \begin{pmatrix} \sqrt{n_1} \\ \vdots \\ \sqrt{n_k} \end{pmatrix} (\sqrt{n_1}, \cdots, \sqrt{n_k}) \right\} \begin{pmatrix} \frac{1}{\sqrt{n_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{n_k}} \end{pmatrix}$$

Define

$$\boldsymbol{Z}_N = \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} r_{i1} \\ \vdots \\ \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} r_{ik} \end{pmatrix}, \; D_N = \begin{pmatrix} \frac{1}{\sqrt{n_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{n_k}} \end{pmatrix}, \; \boldsymbol{n} = \begin{pmatrix} \sqrt{n_1} \\ \vdots \\ \sqrt{n_k} \end{pmatrix} \text{ and } \Pi_N = I_k - \frac{1}{N}\boldsymbol{n}\boldsymbol{n}'$$

then

$$S_N = \boldsymbol{a}' D_N \boldsymbol{Z}_N,$$
$$\mathrm{E}[S_N] = \frac{N+1}{2} \boldsymbol{a}' D_N \boldsymbol{n},$$
$$\mathrm{Var}[S_N] = \frac{N(N+1)}{12} \boldsymbol{a}' D_N \Pi_N D_N \boldsymbol{a}.$$

Define

$$\Pi_N = I_k - \frac{1}{N}\boldsymbol{n}\boldsymbol{n}' = \underset{k\times(k-1)}{H_N} H_N'$$

$$\boldsymbol{U}_N = \sqrt{\frac{12}{N(N+1)}} H_N' \Big(\boldsymbol{Z}_N - \frac{N+1}{2}\boldsymbol{n}\Big) = \sqrt{\frac{12}{N(N+1)}} H_N' \boldsymbol{Z}_N$$

$$\Rightarrow \quad \boldsymbol{U}_N \to N_{k-1}(\boldsymbol{0}, I_{k-1})$$

$$(\because\ \boldsymbol{s}'\boldsymbol{U}_N = S_N - \mathrm{E}[S_N] \text{ with } \boldsymbol{a} = \sqrt{\tfrac{12}{N(N+1)}} D_N^{-1} H_N \boldsymbol{s},\ \text{and } \mathrm{Var}[\boldsymbol{s}'\boldsymbol{U}_N] = \boldsymbol{s}'\boldsymbol{s})$$

$$\Pi_N \boldsymbol{Z}_N = \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} r_{i1} \\ \vdots \\ \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} r_{ik} \end{pmatrix} - \frac{1}{N}\boldsymbol{n} \sum_{j=1}^{k}\sum_{i=1}^{n_j} r_{ij} = \begin{pmatrix} \sqrt{n_1}(R_{\cdot 1} - \frac{N+1}{2}) \\ \vdots \\ \sqrt{n_k}(R_{\cdot k} - \frac{N+1}{2}) \end{pmatrix}$$

$$\boldsymbol{U}_N'\boldsymbol{U}_N = \frac{12}{N(N+1)} \boldsymbol{Z}_N' \Pi_N \boldsymbol{Z}_N = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left( R_{\cdot j} - \frac{N+1}{2} \right)^2 = H$$

Hence

$$H = \boldsymbol{U}_N \boldsymbol{U}_N \to \chi^2_{k-1}$$

Note

$$\boldsymbol{Z}_N' \Pi_N \boldsymbol{Z}_N = \boldsymbol{Z}_N' \boldsymbol{Z}_N - \frac{1}{N}(\boldsymbol{Z}_N'\boldsymbol{n})^2 = \sum_{j=1}^{k} \frac{1}{n_j} \Big( \sum_{i=1}^{n_j} r_{ij} \Big)^2 - \frac{1}{N} \Big( \frac{N(N+1)}{2} \Big)^2$$

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

## 5.2 A distribution–free test for ordered alternatives (Jonckheere–Terpstra)

**Testing problem**

$$H_0 \text{ versus } H_2 : \tau_1 \le \tau_2 \le \cdots \le \tau_k, \quad \text{with at least one strict inequality}$$

**Test statistic**

$$J = \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} U_{ab}$$

$$U_{ab} = \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \phi(X_{ia}, X_{jb}),\ 1 \le a < b \le k$$

$$\phi(x,y) = \begin{cases} 1 & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

**Procedure**

$$J \geq j_\alpha \Rightarrow \text{ reject } H_0$$

Example for deriving the exact null distribution

$$R_{ij}: \text{ rank of } X_{ij} \quad \Rightarrow \quad U_{ab} = \sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\phi(X_{ia}, X_{jb}) = \sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\phi(R_{ia}, R_{jb})$$

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

| | I | II | III | | $J$ |
|---|---|---|---|---|---|
| (a) | 1 | 2 | 3 | 4 | 5 |
| (b) | 2 | 1 | 3 | 4 | 4 |
| (c) | 1 | 3 | 2 | 4 | 4 |
| (d) | 3 | 1 | 2 | 4 | 3 |
| (e) | 1 | 4 | 2 | 3 | 3 |
| (f) | 4 | 1 | 3 | 2 | 2 |
| (g) | 2 | 3 | 1 | 4 | 3 |
| (h) | 3 | 2 | 1 | 4 | 2 |
| (i) | 2 | 4 | 1 | 3 | 2 |
| (j) | 4 | 2 | 1 | 3 | 1 |
| (k) | 3 | 4 | 1 | 2 | 1 |
| (l) | 4 | 3 | 1 | 2 | 0 |

| $j$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $\Pr\{J=j\}$ | $\frac{1}{12}$ | $\frac{2}{12}$ | $\frac{3}{12}$ | $\frac{3}{12}$ | $\frac{2}{12}$ | $\frac{1}{12}$ |

**Large–sample approximation**

Under $H_0$

$$\mathrm{E}[J] = \frac{N^2 - \sum_{j=1}^k n_j}{4},$$
$$\mathrm{Var}[J] = \frac{N^2(2N+3) - \sum_{j=1}^k n_j^2(2n_j+3)}{72}$$
$$\frac{J - \mathrm{E}[J]}{\sqrt{\mathrm{Var}[J]}} \to N(0,1) \quad (\min\{n_1, \cdots, n_k\} \to \infty)$$

Proof

$$\mathrm{E}[J] = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{E}[U_{ab}] = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a n_b \mathrm{E}[\phi(X,Y)] = \frac{N^2 - \sum_{a=1}^k n_a^2}{2}\mathrm{E}[\phi(X,Y)]$$

$$J^* = \sum_{a=1}^{k-1}\sum_{b=a+1}^{k} U_{ab}^*$$

$$U_{ab}^* = n_b\sum_{i=1}^{n_a}\phi_{10}(X_{ia}) + n_b\sum_{j=1}^{n_b}\phi_{01}(X_{jb})$$

$$\phi_{10}(x) = \mathrm{E}[\phi(X,Y)|X=x], \ \phi_{01}(y) = \mathrm{E}[\phi(X,Y)|Y=y]$$

Formulas

$$
\begin{aligned}
&X, Y, Z \overset{i.i.d.}{\sim} F \\
&\sigma_{10}^2 = \mathrm{Var}[\phi_{10}(X)],\ \sigma_{01}^2 = \mathrm{Var}[\phi_{01}(X)],\ \sigma_{11}^2 = \mathrm{Cov}[\phi_{10}(X), \phi_{01}(X)] \\
&\Rightarrow\ \mathrm{Cov}[\phi(X,Y), \phi(X,Z)] = \sigma_{10}^2, \quad \mathrm{Cov}[\phi(X,Y), \phi(Z,Y)] = \sigma_{01}^2, \\
&\quad\ \mathrm{Cov}[\phi(X,Y), \phi(Y,Z)] = \sigma_{11}, \quad \mathrm{Cov}[\phi(X,Y), \phi(Z,X)] = \sigma_{11}, \\
&\quad\ \mathrm{Cov}[\phi(X,Y), \phi_{10}(X)] = \sigma_{10}^2, \quad \mathrm{Cov}[\phi(X,Y), \phi_{10}(Y)] = \sigma_{11}, \\
&\quad\ \mathrm{Cov}[\phi(X,Y), \phi_{01}(X)] = \sigma_{11}, \quad \mathrm{Cov}[\phi(X,Y), \phi_{01}(Y)] = \sigma_{01}^2,
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{Var}[U_{ab}] = n_a n_b(\sigma^2 - \sigma_{10}^2 - \sigma_{01}^2) + n_a n_b^2 \sigma_{10}^2 + n_a^2 n_b \sigma_{01}^2 \\
&\mathrm{Cov}[U_{ab}, U_{ac}] = n_a n_b n_c \sigma_{10}^2,\ (b \neq c) \\
&\mathrm{Cov}[U_{ab}, U_{bc}] = n_a n_b n_c \sigma_{11}, \\
&\mathrm{Cov}[U_{ab}, U_{cb}] = n_a n_b n_c \sigma_{01}^2,\ (a \neq c)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{Cov}[U_{ab}, U_{ab}^*] = n_a n_b^2 \sigma_{10}^2 + n_a^2 n_b \sigma_{01}^2, \\
&\mathrm{Cov}[U_{ab}, U_{ac}^*] = n_a n_b n_c \sigma_{10}^2,\ (b \neq c) \\
&\mathrm{Cov}[U_{ab}, U_{bc}^*] = \mathrm{Cov}[U_{ab}, U_{ca}^*] = n_a n_b n_c \sigma_{11} \\
&\mathrm{Cov}[U_{ab}, U_{cb}^*] = n_a n_b n_c \sigma_{01}^2,\ (a \neq c)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{Var}[U_{ab}^*] = n_a n_b^2 \sigma_{10}^2 + n_a^2 n_b \sigma_{01}^2 \\
&\mathrm{Cov}[U_{ab}^*, U_{ac}^*] = n_a n_b n_c \sigma_{10}^2,\ (b \neq c) \\
&\mathrm{Cov}[U_{ab}^*, U_{bc}^*] = n_a n_b n_c \sigma_{11}, \\
&\mathrm{Cov}[U_{ab}^*, U_{cb}^*] = n_a n_b n_c \sigma_{01}^2,\ (a \neq c)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[J - J^*] &= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} \mathrm{Var}[U_{ab} - U_{ab}^*] \\
&\quad + \sum_{a=1}^{k-1} \sum_{\substack{b=a+1 \\ (a,b) \neq (a',b')}}^{k} \sum_{a'=1}^{k-1} \sum_{b'=a'+1}^{k} \mathrm{Cov}[(U_{ab} - U_{ab}^*), (U_{a'b'} - U_{a'b'}^*)] \\
&= \left\{ \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_a n_b \right\} (\sigma^2 - \sigma_{10}^2 - \sigma_{01}^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[J] &= \sum_{a=1}^{k-1}\sum_{b=a+1}^{k}\mathrm{Var}[U_{ab}] + \sum_{a=1}^{k-1}\sum_{\substack{b=a+1\\(a,b)\neq(a',b')}}^{k}\sum_{a'=1}^{k-1}\sum_{b'=a'+1}^{k}\mathrm{Cov}[U_{ab},U_{a'b'}] \\
&= \left\{\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a n_b\right\}(\sigma^2-\sigma_{10}^2-\sigma_{01}^2) + 2\left\{\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{11} \\
&\quad + \left\{\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a n_b^2 + 2\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{10}^2 \\
&\quad + \left\{\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a^2 n_b + 2\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{01}^2 \\
\mathrm{Var}[J^*] &= 2\left\{\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{11} \\
&\quad + \left\{\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a n_b^2 + 2\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{10}^2 \\
&\quad + \left\{\sum_{a=1}^{k-1}\sum_{b=a+1}^{k} n_a^2 n_b + 2\sum_{a=1}^{k-2}\sum_{b=a+1}^{k-1}\sum_{c=b+1}^{k} n_a n_b n_c\right\}\sigma_{01}^2
\end{aligned}
$$

$\phi_{10}(x) = \mathrm{E}[\phi(x,Y)] = 1-F(x),\ \phi_{01}(y) = \mathrm{E}[\phi(X,y)] = F(y)$

$\mathrm{E}[\phi(X,Y)] = \mathrm{E}[\phi_{10}(X)] = \mathrm{E}[\phi_{01}(X)] = \dfrac{1}{2} \quad \because F(X) \sim \mathcal{U}(0,1)$

$\mathrm{E}[J] = \dfrac{N^2 - \sum_{a=1}^{k} n_a^2}{4}$

$\sigma_{10}^2 = \sigma_{01}^2 = \dfrac{1}{12}$

$\sigma_{11} = \mathrm{E}[\{1-F(X)\}F(X)] - \dfrac{1}{4} = -\dfrac{1}{12}$

$\sigma^2 = \mathrm{E}[\phi(X,Y)^2] - \dfrac{1}{4} = \mathrm{E}[\phi(X,Y)] - \dfrac{1}{4} = \dfrac{1}{4}$

$$
N^3 = \sum_{a=1}^{k}\sum_{b=1}^{k}\sum_{c=1}^{k} n_a n_b n_c = 6\sum_{\substack{a=1\\a<b<c}}^{k}\sum_{b=1}^{k}\sum_{c=1}^{k} + 3\sum_{\substack{a=1\\a<b}}^{k}\sum_{b=1}^{k}(n_a^2 n_b + n_a n_b^2) + \sum_{a=1}^{k} n_a^3
$$

$$
\mathrm{Var}[J] = \frac{1}{12}\left\{\frac{N^2 - \sum_{a=1}^{k} n_a^2}{2} + \frac{N^3 - \sum_{a=1}^{k} n_a^3}{3}\right\} = \frac{N^2(2N+3) - \sum_{a=1}^{k} n_a^2(2n_a+3)}{72}
$$

$$
\mathrm{Var}[J^*] = \frac{N^3 - \sum_{a=1}^{k} n_a^3}{36}, \quad \mathrm{Var}[J-J^*] = \frac{N^2 - \sum_{a=1}^{k} n_a^2}{24}
$$

$$
\Rightarrow \quad \frac{J-\mathrm{E}[J]}{\sqrt{\mathrm{Var}[J]}} - \frac{J^*-\mathrm{E}[J^*]}{\sqrt{\mathrm{Var}[J^*]}} \xrightarrow{P} 0
$$

$$U^*_{ab} = n_b \sum_{i=1}^{n_a} \phi_{10}(X_{ia}) + n_a \sum_{j=1}^{n_b} \phi_{01}(X_{ib})$$

$$\begin{aligned} J^* = \sum_{a<b} U^*_{ab} &= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_b \sum_{i=1}^{n_a} \phi_{10}(X_{ia}) + \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_a \sum_{i=1}^{n_b} \phi_{01}(X_{ib}) \\ &= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_b \sum_{i=1}^{n_a} \phi_{10}(X_{ia}) + \sum_{a=2}^{k} \sum_{b=1}^{a-1} n_b \sum_{i=1}^{n_a} \phi_{01}(X_{ia}) \\ &= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_b \sum_{i=1}^{n_a} \{1 - F(X_{ia})\} + \sum_{a=2}^{k} \sum_{b=1}^{a-1} n_b \sum_{i=1}^{n_a} F(X_{ia}) \\ &= \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_a n_b + \sum_{a=1}^{k} \left\{ - \sum_{b=a+1}^{k} n_b + \sum_{b=1}^{a-1} n_b \right\} \sum_{i=1}^{n_a} F(X_{ia}) \end{aligned}$$

$$c_{ia} = \left\{ - \sum_{b=a+1}^{k} n_b + \sum_{b=1}^{a-1} n_b \right\}, \ (i = 1, \cdots, n_a; a = 1, \cdots, k)$$

$$\bar{c}_{ia} = \frac{1}{N} \sum_{a=1}^{k} \sum_{i=1}^{n_a} c_{ia} = \frac{1}{N} \left\{ - \sum_{a=1}^{k-1} \sum_{b=a+1}^{k} n_a n_b + \sum_{a=2}^{k} \sum_{b=1}^{a-1} n_a n_b \right\} = 0$$

$$\sum_{a=1}^{k} \sum_{i=1}^{n_a} c_{ia}^2 = \frac{N^3 - \sum_{a=1}^{k} n_a^3}{3}, \ |c_{ia}| < N$$

$$\frac{\max_{i,a} c_{ia}^2}{\sum_{a=1}^{k} \sum_{i=1}^{n_a} c_{ia}^2} \leq \frac{3N^2}{N^3 - \sum_{a=1}^{k} n_a^3} \to 0$$