Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on April 24, 2021

2021 Spring

Chapter 14. Nonparametric Statistics

§ 14.1 Introduction

§ 14.2 The Sign Test

§ 14.3 Wilcoxon Tests

§ 14.4 The Kruskal-Wallis Test

§ 14.5 The Friedman Test

§ 14.6 Testing for Randomness

Plan

§ 14.1 Introduction

§ 14.2 The Sign Test

§ 14.3 Wilcoxon Tests

§ 14.4 The Kruskal-Wallis Test

§ 14.5 The Friedman Test

§ 14.6 Testing for Randomness

§ 14.1 Introduction

§ 14.2 The Sign Test

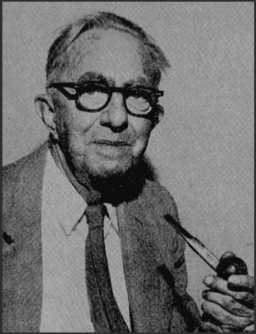
§ 14.3 Wilcoxon Tests

§ 14.4 The Kruskal-Wallis Test

§ 14.5 The Friedman Test

§ 14.6 Testing for Randomness

Frank Wilcoxon



Born	2 September 1892 County Cork, Ireland
Died	18 November 1965 (aged 73) Tallahassee, Florida, USA
Nationality	Irish American
Alma mater	Cornell University Rutgers University

Scientific career

Fields	Chemistry Statistics
Institutions	American Cyanamid Company

Testing $H_0 : \mu = \mu_0$

Setup Let $Y_1, \cdots, Y_n$ be a set of independent variables with pdfs $f_{Y_1}(y), \cdots, f_{Y_n}(y)$, respectively.

Assume that $f_{Y_i}(y)$ are continuous and symmetric.

Assume that all mean/median of f_{Y_i} are equal, denoted by μ.

Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Wilcoxon signed rank static

$$W = \sum_{k=1}^{n} R_k \, \mathbb{1}_{\{Y_k > \mu_0\}}$$

where R_i denotes the rank (increasing and starting from 1) of

$$\{|Y_1 - \mu_0|, |Y_2 - \mu_0|, \cdots, |Y_n - \mu_0|\}$$

Testing $H_0 : \mu = \mu_0$

Setup Let $Y_1, \cdots, Y_n$ be a set of independent variables with pdfs $f_{Y_1}(y), \cdots, f_{Y_n}(y)$, respectively.

Assume that $f_{Y_i}(y)$ are continuous and symmetric.

Assume that all mean/median of f_{Y_i} are equal, denoted by μ.

Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Wilcoxon signed rank static

$$W = \sum_{k=1}^{n} R_k \, \mathbb{I}_{\{Y_k > \mu_0\}}$$

where R_i denotes the rank (increasing and starting from 1) of

$$\{|Y_1 - \mu_0|, |Y_2 - \mu_0|, \cdots, |Y_n - \mu_0|\}$$

Testing $H_0 : \mu = \mu_0$

Setup Let $Y_1, \cdots, Y_n$ be a set of independent variables with pdfs $f_{Y_1}(y), \cdots, f_{Y_n}(y)$, respectively.

Assume that $f_{Y_i}(y)$ are continuous and symmetric.

Assume that all mean/median of f_{Y_i} are equal, denoted by μ.

Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Wilcoxon signed rank static

$$W = \sum_{k=1}^{n} R_k \, \mathbb{I}_{\{Y_k > \mu_0\}}$$

where R_i denotes the rank (increasing and starting from 1) of

$$\{|Y_1 - \mu_0|, |Y_2 - \mu_0|, \cdots, |Y_n - \mu_0|\}$$

Testing $H_0 : \mu = \mu_0$

Setup Let $Y_1, \cdots, Y_n$ be a set of independent variables with pdfs $f_{Y_1}(y), \cdots, f_{Y_n}(y)$, respectively.

Assume that $f_{Y_i}(y)$ are continuous and symmetric.

Assume that all mean/median of f_{Y_i} are equal, denoted by μ.

Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Wilcoxon signed rank static

$$W = \sum_{k=1}^{n} R_k \, \mathbb{I}_{\{Y_k > \mu_0\}}$$

where R_i denotes the rank (increasing and starting from 1) of

$$\{|Y_1 - \mu_0|, |Y_2 - \mu_0|, \cdots, |Y_n - \mu_0|\}$$

Testing $H_0 : \mu = \mu_0$

Setup Let $Y_1, \cdots, Y_n$ be a set of independent variables with pdfs $f_{Y_1}(y), \cdots, f_{Y_n}(y)$, respectively.

Assume that $f_{Y_i}(y)$ are continuous and symmetric.

Assume that all mean/median of f_{Y_i} are equal, denoted by μ.

Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Wilcoxon signed rank static

$$W = \sum_{k=1}^{n} R_k \, \mathbb{1}_{\{Y_k > \mu_0\}}$$

where R_i denotes the rank (increasing and starting from 1) of

$$\{|Y_1 - \mu_0|, |Y_2 - \mu_0|, \cdots, |Y_n - \mu_0|\}$$

n	1	2	3
y_n	4.2	6.1	2.0
$y_n - 3.0$	1.2	3.1	-1.0
$\lvert y_n - 3.0\rvert$	1.2	3.1	1.0
r_n	2	3	1
$\mathbb{1}_{\{y_n>3.0\}}$	1	1	0
$r_n\mathbb{1}_{\{y_n>3.0\}}$	$u_2 = 2$	$u_3 = 3$	$u_1 = 0$

$$\Downarrow$$

$$w = 2 \times 1 + 3 \times 1 + 1 \times 0 = 5.$$

Let $\{y_1, \cdots, y_n\}$ be For a sample of size n.

Some observations:

- r_i takes values in $\{1, 2, \cdots, n\}$.
- w_i takes values in $\left\{0, 1, 2, \cdots, \frac{n(n+1)}{2}\right\}$ with $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.
- W is a discrete random variable:

w	0	1	$\cdots$	$\frac{n(n+1)}{2}$
$\mathbb{P}(W = w)$				

Let $\{y_1, \cdots, y_n\}$ be For a sample of size n.

Some observations:

- r_i takes values in $\{1, 2, \cdots, n\}$.
- w_i takes values in $\left\{0, 1, 2, \cdots, \frac{n(n+1)}{2}\right\}$ with $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.
- W is a discrete random variable:

w	0	1	$\cdots$	$\frac{n(n+1)}{2}$
$\mathbb{P}(W = w)$				

Let $\{y_1, \cdots, y_n\}$ be For a sample of size n.

Some observations:

- r_i takes values in $\{1, 2, \cdots, n\}$.
- w_i takes values in $\left\{0, 1, 2, \cdots, \frac{n(n+1)}{2}\right\}$ with $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.
- W is a discrete random variable:

w	0	1	$\cdots$	$\frac{n(n+1)}{2}$
$\mathbb{P}(W = w)$				

Theorem Under the above setup and under H_0,

$$p_W(w) = \mathbb{P}(W = w) = \frac{c(w)}{2^n},$$

where $c(w)$ is the coefficient of e^{wt} in the expansion of

$$\prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

Proof Under H_0, $W = \sum_{k=1}^{n} U_k$ with follow the following distribution

$$U_k = \begin{cases} 0 & \text{with probability } 1/2 \\ k & \text{with probability } 1/2. \end{cases}$$

Then

$$M_W(t) = \prod_{k=1}^{n} M_{U_k}(t) = \prod_{k=1}^{n} \mathbb{E}\left(e^{U_k t}\right) = \prod_{k=1}^{n} \left(\frac{1}{2} + \frac{1}{2} e^{kt}\right).$$

Theorem Under the above setup and under H_0,

$$p_W(w) = \mathbb{P}(W = w) = \frac{c(w)}{2^n},$$

where $c(w)$ is the coefficient of e^{wt} in the expansion of

$$\prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

Proof Under H_0, $W = \sum_{k=1}^{n} U_k$ with follow the following distribution

$$U_k = \begin{cases} 0 & \text{with probability } 1/2 \\ k & \text{with probability } 1/2. \end{cases}$$

Then

$$M_W(t) = \prod_{k=1}^{n} M_{U_k}(t) = \prod_{k=1}^{n} \mathbb{E}\left(e^{U_k t}\right) = \prod_{k=1}^{n} \left(\frac{1}{2} + \frac{1}{2} e^{kt}\right).$$

Theorem Under the above setup and under H_0,

$$p_W(w) = \mathbb{P}(W = w) = \frac{c(w)}{2^n},$$

where $c(w)$ is the coefficient of e^{wt} in the expansion of

$$\prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

Proof Under H_0, $W = \sum_{k=1}^{n} U_k$ with follow the following distribution

$$U_k = \begin{cases} 0 & \text{with probability } 1/2 \\ k & \text{with probability } 1/2. \end{cases}$$

Then

$$M_W(t) = \prod_{k=1}^{n} M_{U_k}(t) = \prod_{k=1}^{n} \mathbb{E}\left(e^{U_k t}\right) = \prod_{k=1}^{n} \left(\frac{1}{2} + \frac{1}{2} e^{kt}\right).$$

Hence, we have

$$M_W(t) = \frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

On the other hand,

$$M_W(t) = \mathbb{E}\left(e^{Wt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w)$$

Equating the above two expressions, namely,

$$\frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w),$$

proves the theorem. ∎

Hence, we have

$$M_W(t) = \frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

On the other hand,

$$M_W(t) = \mathbb{E}\left(e^{Wt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w)$$

Equating the above two expressions, namely,

$$\frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w),$$

proves the theorem. $\blacksquare$

Hence, we have

$$M_W(t) = \frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right).$$

On the other hand,

$$M_W(t) = \mathbb{E}\left(e^{Wt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w)$$

Equating the above two expressions, namely,

$$\frac{1}{2^n} \prod_{k=1}^{n} \left(1 + e^{kt}\right) = \sum_{w=0}^{\frac{n(n+1)}{2}} e^{wt} p_W(w),$$

proves the theorem. ∎

E.g. Find the pdf of W when $n = 2$ and 4.

Sol. When $n = 2$,

$$M_W(t) = \frac{1}{2^2}\left(1 + e^t\right)\left(1 + e^{2t}\right) = \frac{1}{2^2}(1 + e^t + e^{2t} + e^{3t}).$$

Hence,

w	0	1	2	3
$p_W(w)$	1/4	1/4	1/4	1/4

E.g. Find the pdf of W when $n = 2$ and 4.

Sol. When $n = 2$,

$$M_W(t) = \frac{1}{2^2}\left(1 + e^t\right)\left(1 + e^{2t}\right)$$
$$= \frac{1}{2^2}(1 + e^t + e^{2t} + e^{3t}).$$

Hence,

w	0	1	2	3
$p_W(w)$	1/4	1/4	1/4	1/4

When $n = 4$,

$$M_W(t) = \frac{1}{2^4}\left(1+e^t\right)\left(1+e^{2t}\right)\left(1+e^{3t}\right)\left(1+e^{4t}\right)$$
$$= \frac{1}{16}\left(e^{10t} + e^{9t} + e^{8t} + 2e^{7t} + 2e^{6t} + 2e^{5t} + 2e^{4t} + 2e^{3t} + e^{2t} + e^t + 1\right)$$

Hence,

w	0	1	2	3	4	5	6	7	8	9	10
$p_W(w)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

```
sage: var('k,t')
(k, t)
sage: product(1+e^(k*t),k,1,4)
e^(10*t) + e^(9*t) + e^(8*t) + 2*e^(7*t) + 2*e^(6*t) + 2*e^(5*t) + 2*e^(4*t) + 2*e
    ^(3*t) + e^(2*t) + e^t + 1
```

E.g. Shark studies:

Past data show that the true average TL/HDI ratio should be 14.60.
Let $Y_i = TL/HDI$.
Does the data support the above claim, namely, test

$$H_0 : \mu = 14.60 \quad \text{vs.} \quad H_1 : \mu \neq 14.60.$$

Set $\alpha = 0.05$.

E.g. Shark studies:

Table 14.3.2 Measurements Made on Ten Sharks Caught Near Santa Catalina

Total Length (mm)	Height of First Dorsal Fin (mm)	*TL/HDI*
906	68	13.32
875	67	13.06
771	55	14.02
700	59	11.86
869	64	13.58
895	65	13.77
662	49	13.51
750	52	14.42
794	55	14.44
787	51	15.43

Past data show that the true average TL/HDI ratio should be 14.60.

Let $Y_i = TL/HDI$.

Does the data support the above claim, namely, test

$$H_0 : \mu = 14.60 \quad \text{vs.} \quad H_1 : \mu \neq 14.60.$$

Set $\alpha = 0.05$.

E.g. Shark studies:

Table 14.3.2 Measurements Made on Ten Sharks Caught Near Santa Catalina

Total Length (mm)	Height of First Dorsal Fin (mm)	*TL/HDI*
906	68	13.32
875	67	13.06
771	55	14.02
700	59	11.86
869	64	13.58
895	65	13.77
662	49	13.51
750	52	14.42
794	55	14.44
787	51	15.43

Past data show that the true average TL/HDI ratio should be 14.60.

Let $Y_i = TL/HDI$.

Does the data support the above claim, namely, test

$$H_0 : \mu = 14.60 \quad \text{vs.} \quad H_1 : \mu \neq 14.60.$$

Set $\alpha = 0.05$.

E.g. Shark studies:

Table 14.3.2 Measurements Made on Ten Sharks Caught Near Santa Catalina

Total Length (mm)	Height of First Dorsal Fin (mm)	*TL/HDI*
906	68	13.32
875	67	13.06
771	55	14.02
700	59	11.86
869	64	13.58
895	65	13.77
662	49	13.51
750	52	14.42
794	55	14.44
787	51	15.43

Past data show that the true average TL/HDI ratio should be 14.60.

Let $Y_i = TL/HDI$.

Does the data support the above claim, namely, test

$$H_0 : \mu = 14.60 \quad \text{vs.} \quad H_1 : \mu \neq 14.60.$$

Set $\alpha = 0.05$.

E.g. Shark studies:

Table 14.3.2 Measurements Made on Ten Sharks Caught Near Santa Catalina

Total Length (mm)	Height of First Dorsal Fin (mm)	*TL/HDI*
906	68	13.32
875	67	13.06
771	55	14.02
700	59	11.86
869	64	13.58
895	65	13.77
662	49	13.51
750	52	14.42
794	55	14.44
787	51	15.43

Past data show that the true average TL/HDI ratio should be 14.60.

Let $Y_i = TL/HDI$.

Does the data support the above claim, namely, test

$$H_0 : \mu = 14.60 \quad \text{vs.} \quad H_1 : \mu \neq 14.60.$$

Set $\alpha = 0.05$.

Sol. Computing the Wilcoxon signed rank statistics:

Hence, $w = 4.5$.

Now check the table to find the critical region:

$$C = \{w : w \leq 8 \quad \text{or} \quad w \geq 47\}.$$

Conclusion: Rejection! ∎

Sol. Computing the Wilcoxon signed rank statistics:

Table 14.3.3 Computations for Wilcoxon Signed Rank Test

$TL/HDI\ (= y_i)$	$y_i - 14.60$	$\lvert y_i - 14.60 \rvert$	r_i	z_i	$r_i z_i$
13.32	−1.28	1.28	8	0	0
13.06	−1.54	1.54	9	0	0
14.02	−0.58	0.58	3	0	0
11.86	−2.74	2.74	10	0	0
13.58	−1.02	1.02	6	0	0
13.77	−0.83	0.83	4.5	0	0
13.51	−1.09	1.09	7	0	0
14.42	−0.18	0.18	2	0	0
14.44	−0.16	0.16	1	0	0
15.43	+0.83	0.83	4.5	1	4.5

Hence, $w = 4.5$.

Now check the table to find the critical region:

$$C = \{w : w \leq 8 \quad \text{or} \quad w \geq 47\}.$$

Conclusion: Rejection! ∎

Sol. Computing the Wilcoxon signed rank statistics:

Table 14.3.3 Computations for Wilcoxon Signed Rank Test

$TL/HDI\ (=y_i)$	$y_i - 14.60$	$\lvert y_i - 14.60\rvert$	r_i	z_i	$r_i z_i$
13.32	−1.28	1.28	8	0	0
13.06	−1.54	1.54	9	0	0
14.02	−0.58	0.58	3	0	0
11.86	−2.74	2.74	10	0	0
13.58	−1.02	1.02	6	0	0
13.77	−0.83	0.83	4.5	0	0
13.51	−1.09	1.09	7	0	0
14.42	−0.18	0.18	2	0	0
14.44	−0.16	0.16	1	0	0
15.43	+0.83	0.83	4.5	1	4.5

Hence, $w = 4.5$.

Now check the table to find the critical region:

$$C = \{w : w \leq 8 \quad \text{or} \quad w \geq 47\}.$$

Conclusion: Rejection!

Sol. Computing the Wilcoxon signed rank statistics:

Table 14.3.3 Computations for Wilcoxon Signed Rank Test

$TL/HDI\ (= y_i)$	$y_i - 14.60$	$\lvert y_i - 14.60\rvert$	r_i	z_i	$r_i z_i$
13.32	−1.28	1.28	8	0	0
13.06	−1.54	1.54	9	0	0
14.02	−0.58	0.58	3	0	0
11.86	−2.74	2.74	10	0	0
13.58	−1.02	1.02	6	0	0
13.77	−0.83	0.83	4.5	0	0
13.51	−1.09	1.09	7	0	0
14.42	−0.18	0.18	2	0	0
14.44	−0.16	0.16	1	0	0
15.43	+0.83	0.83	4.5	1	4.5

Hence, $w = 4.5$.

Now check the table to find the critical region:

$$C = \{w : w \leq 8 \quad \text{or} \quad w \geq 47\}.$$

Conclusion: Rejection! ■

```
> x <- c(13.32, 13.06, 14.02, 11.86, 13.58, 13,77, 13.51, 14.42, 14.44, 15.43)
> wilcox.test(x, mu = 14.60, alternative = "two.sided")

        Wilcoxon signed rank exact test

data:  x
V = 15, p-value = 0.123
alternative hypothesis: true location is not equal to 14.6
```

Large-sample Wilcoxon Signed Rank Test

Theorem Under the same setup and H_0, we have

$$\mathbb{E}(W) = \frac{n(n+1)}{4} \quad \text{and} \quad \mathrm{Var}(W) = \frac{n(n+1)(2n+1)}{24}.$$

Proof.

$$\mathbb{E}(W) = \mathbb{E}\left(\sum_{k=1}^{n} U_k\right) = \sum_{k=1}^{n}\left(0 \cdot \frac{1}{2} + k \cdot \frac{1}{2}\right)$$
$$= \sum_{k=1}^{n} \frac{k}{2} = \frac{n(n+1)}{4}.$$

$$\mathrm{Var}(W) = \mathrm{Var}\left(\sum_{k=1}^{n} U_k\right) = \sum_{k=1}^{n} \mathrm{Var}(U_k) = \sum_{k=1}^{n}\left[\mathbb{E}(U_k^2) - \mathbb{E}(U_k)^2\right]$$
$$= \sum_{k=1}^{n}\left[\frac{k^2}{2} - \left(\frac{k}{2}\right)^2\right] = \sum_{k=1}^{n} \frac{k^2}{4} = \frac{1}{4}\frac{n(n+1)(2n+1)}{6}$$

∎

Large-sample Wilcoxon Signed Rank Test

Theorem Under the same setup and H_0, we have

$$\mathbb{E}(W) = \frac{n(n+1)}{4} \quad \text{and} \quad \mathrm{Var}(W) = \frac{n(n+1)(2n+1)}{24}.$$

Proof.

$$\begin{aligned}\mathbb{E}(W) &= \mathbb{E}\left(\sum_{k=1}^{n} U_k\right) = \sum_{k=1}^{n}\left(0 \cdot \frac{1}{2} + k \cdot \frac{1}{2}\right) \\ &= \sum_{k=1}^{n} \frac{k}{2} = \frac{n(n+1)}{4}.\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(W) &= \mathrm{Var}\left(\sum_{k=1}^{n} U_k\right) = \sum_{k=1}^{n} \mathrm{Var}(U_k) = \sum_{k=1}^{n}\left[\mathbb{E}(U_k^2) - \mathbb{E}(U_k)^2\right] \\ &= \sum_{k=1}^{n}\left[\frac{k^2}{2} - \left(\frac{k}{2}\right)^2\right] = \sum_{k=1}^{n} \frac{k^2}{4} = \frac{1}{4}\frac{n(n+1)(2n+1)}{6}\end{aligned}$$

$\blacksquare$

Hence when n is large (usually $n \geq 12$),

$$\frac{W - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{W - [n(n+1)]/4}{\sqrt{[n(n+1)(2n+1)]/24}} \overset{approx}{\sim} N(0,1).$$

$$\Downarrow$$

Hence when n is large (usually $n \geq 12$),

$$\frac{W - \mathbb{E}(W)}{\sqrt{\operatorname{Var}(W)}} = \frac{W - [n(n+1)]/4}{\sqrt{[n(n+1)(2n+1)]/24}} \overset{approx}{\sim} N(0,1).$$

$$\Downarrow$$

Let w be the signed rank statistic based on n independent observations, each drawn from a continuous and symmetric pdf, where $n > 12$. Let

$$z = \frac{w - [n(n+1)]/4}{\sqrt{[n(n+1)(2n+1)]/24}}$$

a. *To test $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu > \mu_0$ at the α level of significance, reject H_0 if $z \geq z_\alpha$.*

b. *To test $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu < \mu_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.*

c. *To test $H_0\colon \mu = \mu_0$ versus $H_1\colon \mu \neq \mu_0$ at the α level of significance, reject H_0 if z is either (1) $\leq -z_{\alpha/2}$ or (2) $\geq z_{\alpha/2}$.* $\square$

The Wilcoxon Rank Sum Test

– Nonparametric counterpart of the pooled two-sample t-test

Setup Let $x_1, \cdots, x_n$ and $y_{n+1}, \cdots, y_{n+m}$ be two independent random samples from $f_X(x)$ and $f_Y(y)$, respectively.

Assume that $f_X(x)$ and $f_Y(y)$ are the same except for a possible shift in location.

Test $H_0 : \mu_x = \mu_y$ vs. ...

Test statistic

$$W = \sum_{k=1}^{n+m} R_i Z_i$$

where R_i is the rank (starting from the lowest with rank 1) and

$$Z_i = \begin{cases} 1 & \text{the ith entry comes from } f_X(x) \\ 0 & \text{the ith entry comes from } f_Y(y). \end{cases}$$

The Wilcoxon Rank Sum Test

– Nonparametric counterpart of the pooled two-sample t-test

Setup Let $x_1, \cdots, x_n$ and $y_{n+1}, \cdots, y_{n+m}$ be two independent random samples from $f_X(x)$ and $f_Y(y)$, respectively.

Assume that $f_X(x)$ and $f_Y(y)$ are the same except for a possible shift in location.

Test $H_0 : \mu_x = \mu_y$ vs. ...

Test statistic

$$W = \sum_{k=1}^{n+m} R_i Z_i$$

where R_i is the rank (starting from the lowest with rank 1) and

$$Z_i = \begin{cases} 1 & \text{the ith entry comes from } f_X(x) \\ 0 & \text{the ith entry comes from } f_Y(y). \end{cases}$$

The Wilcoxon Rank Sum Test

– Nonparametric counterpart of the pooled two-sample t-test

Setup Let $x_1, \cdots, x_n$ and $y_{n+1}, \cdots, y_{n+m}$ be two independent random samples from $f_X(x)$ and $f_Y(y)$, respectively.

Assume that $f_X(x)$ and $f_Y(y)$ are the same except for a possible shift in location.

Test $H_0 : \mu_x = \mu_y$ vs. ...

Test statistic

$$W = \sum_{k=1}^{n+m} R_i Z_i$$

where R_i is the rank (starting from the lowest with rank 1) and

$$Z_i = \begin{cases} 1 & \text{the ith entry comes from } f_X(x) \\ 0 & \text{the ith entry comes from } f_Y(y). \end{cases}$$

The Wilcoxon Rank Sum Test

– Nonparametric counterpart of the pooled two-sample t-test

Setup Let $x_1, \cdots, x_n$ and $y_{n+1}, \cdots, y_{n+m}$ be two independent random samples from $f_X(x)$ and $f_Y(y)$, respectively.

Assume that $f_X(x)$ and $f_Y(y)$ are the same except for a possible shift in location.

Test $H_0 : \mu_x = \mu_y$ vs. ...

Test statistic

$$W = \sum_{k=1}^{n+m} R_i Z_i$$

where R_i is the rank (starting from the lowest with rank 1) and

$$Z_i = \begin{cases} 1 & \text{the ith entry comes from } f_X(x) \\ 0 & \text{the ith entry comes from } f_Y(y). \end{cases}$$

Theorem Under the above setup and under H_0,

$$\mathbb{E}[W] = \frac{n(n+m+1)}{2} \quad \text{and} \quad \mathrm{Var}(W) = \frac{nm(n+m+1)}{12}.$$

Hence when sample sizes are large, namely, $n, m > 10$,

$$\frac{W - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{W - [n(n+m+1)]/2}{\sqrt{[nm(n+m+1)]/12}} \quad \overset{approx}{\sim} \quad N(0,1).$$

Theorem Under the above setup and under H_0,

$$\mathbb{E}[W] = \frac{n(n+m+1)}{2} \quad \text{and} \quad \mathrm{Var}(W) = \frac{nm(n+m+1)}{12}.$$

Hence when sample sizes are large, namely, $n, m > 10$,

$$\frac{W - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{W - [n(n+m+1)]/2}{\sqrt{[nm(n+m+1)]/12}} \quad \overset{approx}{\sim} \quad N(0,1).$$

E.g. Baseball ...

Test if $H_0 : \mu_X = \mu_Y$ vs. $H_0 : \mu_X \neq \mu_Y$

Group X

Group Y

E.g. Baseball ...

Test if $H_0 : \mu_X = \mu_Y$ vs. $H_0 : \mu_X \neq \mu_Y$

Obs. #	Team	Time (min)	r_i	z_i	$r_i z_i$
1	Baltimore	177	21	1	21
2	Boston	177	21	1	21
3	California	165	7.5	1	7.5
4	Chicago (AL)	172	14.5	1	14.5
5	Cleveland	172	14.5	1	14.5
6	Detroit	179	24.5	1	24.5
7	Kansas City	163	5	1	5
8	Milwaukee	175	18	1	18
9	Minnesota	166	9.5	1	9.5
10	New York (AL)	182	26	1	26
11	Oakland	177	21	1	21
12	Seattle	168	12.5	1	12.5
13	Texas	179	24.5	1	24.5
14	Toronto	177	21	1	21
15	Atlanta	166	9.5	0	0
16	Chicago (NL)	154	1	0	0
17	Cincinnati	159	2	0	0
18	Houston	168	12.5	0	0
19	Los Angeles	174	16.5	0	0
20	Montreal	174	16.5	0	0
21	New York (NL)	177	21	0	0
22	Philadelphia	167	11	0	0
23	Pittsburgh	165	7.5	0	0
24	San Diego	161	3.5	0	0
25	San Francisco	164	6	0	0
26	St. Louis	161	3.5	0	0
					$w' = 240.5$

Group X

Group Y

E.g. Baseball ...

$$\text{Test if } H_0 : \mu_X = \mu_Y \text{ vs. } H_0 : \mu_X \neq \mu_Y$$

Obs. #	Team	Time (min)	r_i	z_i	$r_i z_i$
1	Baltimore	177	21	1	21
2	Boston	177	21	1	21
3	California	165	7.5	1	7.5
4	Chicago (AL)	172	14.5	1	14.5
5	Cleveland	172	14.5	1	14.5
6	Detroit	179	24.5	1	24.5
7	Kansas City	163	5	1	5
8	Milwaukee	175	18	1	18
9	Minnesota	166	9.5	1	9.5
10	New York (AL)	182	26	1	26
11	Oakland	177	21	1	21
12	Seattle	168	12.5	1	12.5
13	Texas	179	24.5	1	24.5
14	Toronto	177	21	1	21
15	Atlanta	166	9.5	0	0
16	Chicago (NL)	154	1	0	0
17	Cincinnati	159	2	0	0
18	Houston	168	12.5	0	0
19	Los Angeles	174	16.5	0	0
20	Montreal	174	16.5	0	0
21	New York (NL)	177	21	0	0
22	Philadelphia	167	11	0	0
23	Pittsburgh	165	7.5	0	0
24	San Diego	161	3.5	0	0
25	San Francisco	164	6	0	0
26	St. Louis	161	3.5	0	0
					$w' = 240.5$

Group X

Group Y

E.g. Baseball ...

Test if $H_0 : \mu_X = \mu_Y$ vs. $H_0 : \mu_X \neq \mu_Y$

Obs. #	Team	Time (min)	r_i	z_i	$r_i z_i$
1	Baltimore	177	21	1	21
2	Boston	177	21	1	21
3	California	165	7.5	1	7.5
4	Chicago (AL)	172	14.5	1	14.5
5	Cleveland	172	14.5	1	14.5
6	Detroit	179	24.5	1	24.5
7	Kansas City	163	5	1	5
8	Milwaukee	175	18	1	18
9	Minnesota	166	9.5	1	9.5
10	New York (AL)	182	26	1	26
11	Oakland	177	21	1	21
12	Seattle	168	12.5	1	12.5
13	Texas	179	24.5	1	24.5
14	Toronto	177	21	1	21
15	Atlanta	166	9.5	0	0
16	Chicago (NL)	154	1	0	0
17	Cincinnati	159	2	0	0
18	Houston	168	12.5	0	0
19	Los Angeles	174	16.5	0	0
20	Montreal	174	16.5	0	0
21	New York (NL)	177	21	0	0
22	Philadelphia	167	11	0	0
23	Pittsburgh	165	7.5	0	0
24	San Diego	161	3.5	0	0
25	San Francisco	164	6	0	0
26	St. Louis	161	3.5	0	0
					$w' = 240.5$

Group X

Group Y

In this case, $n = 14$, $m = 12$, $w = 240.5$.

$$\mathbb{E}(W) = \frac{14(14+12+1)}{2} = 189,$$
$$\mathrm{Var}(W) = \frac{14 \times 12 \times (14+12+1)}{12} = 378.$$

Hence, the approximate z-score is

$$z = \frac{w - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{240.5 - 189}{\sqrt{378}} = 2.65.$$

…

In this case, $n = 14$, $m = 12$, $w = 240.5$.

$$\mathbb{E}(W) = \frac{14(14+12+1)}{2} = 189,$$
$$\mathrm{Var}(W) = \frac{14 \times 12 \times (14+12+1)}{12} = 378.$$

Hence, the approximate z-score is

$$z = \frac{w - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{240.5 - 189}{\sqrt{378}} = 2.65.$$

...

In this case, $n = 14$, $m = 12$, $w = 240.5$.

$$\mathbb{E}(W) = \frac{14(14+12+1)}{2} = 189,$$
$$\mathrm{Var}(W) = \frac{14 \times 12 \times (14+12+1)}{12} = 378.$$

Hence, the approximate z-score is

$$z = \frac{w - \mathbb{E}(W)}{\sqrt{\mathrm{Var}(W)}} = \frac{240.5 - 189}{\sqrt{378}} = 2.65.$$

...