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

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \leq \widetilde{\mu}) = \mathbb{P}(Y \geq \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{X - \frac{n}{2}}{\sqrt{n/4}} \overset{\text{aprox.}}{\sim} N(0, 1)$$

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \leq \widetilde{\mu}) = \mathbb{P}(Y \geq \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

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

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \leq \widetilde{\mu}) = \mathbb{P}(Y \geq \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{X - \frac{n}{2}}{\sqrt{n/4}} \overset{\text{aprox.}}{\sim} N(0, 1)$$

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \le \widetilde{\mu}) = \mathbb{P}(Y \ge \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{X - \frac{n}{2}}{\sqrt{n/4}} \overset{\text{aprox.}}{\sim} N(0,1)$$

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \leq \widetilde{\mu}) = \mathbb{P}(Y \geq \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{X - \frac{n}{2}}{\sqrt{n/4}} \overset{\text{aprox.}}{\sim} N(0,1)$$

- Let $\widetilde{\mu}$ be the **median** of some unknown continuous pdf $f_Y(y)$:

$$\mathbb{P}(Y \leq \widetilde{\mu}) = \mathbb{P}(Y \geq \widetilde{\mu}) = \frac{1}{2}.$$

- For a random sample of size n is taken from $f_Y(y)$, in order to test

$$H_0 : \widetilde{\mu} = \widetilde{\mu}_0 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq \widetilde{\mu}_0,$$

let

$$X := \text{the number of observations exceeding } \widetilde{\mu}_0$$

$$\Downarrow$$

1. $X \sim \text{Binomial}(n, 1/2)$.
2. Moreover, if n is large, by CLT,

$$\frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{X - \frac{n}{2}}{\sqrt{n/4}} \overset{\text{aprox.}}{\sim} N(0, 1)$$

Sign test for median of a single sample

- When sample size n is large:

- When sample size n is small: use the exact distribution of binomial distribution.

Sign test for median of a single sample

- When sample size n is large:

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from any continuous distribution having median $\tilde{\mu}$, where $n \geq 10$. Let k denote the number of y_i's greater than $\tilde{\mu}_0$, and let $z = \frac{k-n/2}{\sqrt{n/4}}$.

a. *To test $H_0\colon \tilde{\mu} = \tilde{\mu}_0$ versus $H_1\colon \tilde{\mu} > \tilde{\mu}_0$ at the α level of significance, reject H_0 if $z \geq z_\alpha$.*
b. *To test $H_0\colon \tilde{\mu} = \tilde{\mu}_0$ versus $H_1\colon \tilde{\mu} < \tilde{\mu}_0$ at the α level of significance, reject H_0 if $z \leq -z_\alpha$.*
c. *To test $H_0\colon \tilde{\mu} = \tilde{\mu}_0$ versus $H_1\colon \tilde{\mu} \neq \tilde{\mu}_0$ at the α level of significance, reject H_0 if z is either (1) $\leq -z_{\alpha/2}$ or (2) $\geq z_{\alpha/2}$.* □

- When sample size n is small: use the exact distribution of binomial distribution.

E.g.1 In a healthy adults, the median pH for synovial fluid is 7.39.

A random sample of $n = 43$ is chosen and test

$$H_0 : \widetilde{\mu} = 7.39 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq 7.39, \qquad \text{at } \alpha = 0.10.$$

E.g.1 In a healthy adults, the median pH for synovial fluid is 7.39. A random sample of $n = 43$ is chosen and test

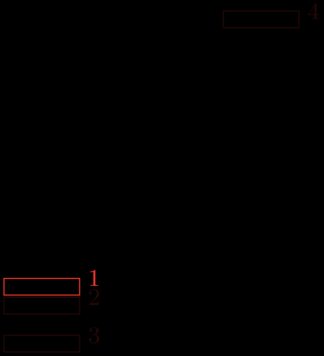
$$H_0 : \widetilde{\mu} = 7.39 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq 7.39, \qquad \text{at } \alpha = 0.10.$$

E.g.1 In a healthy adults, the median pH for synovial fluid is 7.39. A random sample of $n = 43$ is chosen and test

$$H_0 : \widetilde{\mu} = 7.39 \quad \text{vs} \quad H_0 : \widetilde{\mu} \neq 7.39, \qquad \text{at } \alpha = 0.10.$$

Subject	Synovial Fluid pH	Subject	Synovial Fluid pH
HW	7.02	BG	7.34
AD	7.35	GL	7.22
TK	7.32	BP	7.32
EP	7.33	NK	7.40
AF	7.15	LL	6.99
LW	7.26	KC	7.10
LT	7.25	FA	7.30
DR	7.35	ML	7.21
VU	7.38	CK	7.33
SP	7.20	LW	7.28
MM	7.31	ES	7.35
DF	7.24	DD	7.24
LM	7.34	SL	7.36
AW	7.32	RM	7.09
BB	7.34	AL	7.32
TL	7.14	BV	6.95
PM	7.20	WR	7.35
JG	7.41	HT	7.36
DH	7.77	ND	6.60
ER	7.12	SJ	7.29
DP	7.45	BA	7.31
FF	7.28		

Sol 1. We first count how many samples exceeding the median (i.e., obtain the value of X)



Sol 1. We first count how many samples exceeding the median (i.e., obtain the value of X)

Subject	Synovial Fluid pH	Subject	Synovial Fluid pH
HW	7.02	BG	7.34
AD	7.35	GL	7.22
TK	7.32	BP	7.32
EP	7.33	NK	7.40
AF	7.15	LL	6.99
LW	7.26	KC	7.10
LT	7.25	FA	7.30
DR	7.35	ML	7.21
VU	7.38	CK	7.33
SP	7.20	LW	7.28
MM	7.31	ES	7.35
DF	7.24	DD	7.24
LM	7.34	SL	7.36
AW	7.32	RM	7.09
BB	7.34	AL	7.32
TL	7.14	BV	6.95
PM	7.20	WR	7.35
JG	7.41	HT	7.36
DH	7.77	ND	6.60
ER	7.12	SJ	7.29
DP	7.45	BA	7.31
FF	7.28		

Sol 1. We first count how many samples exceeding the median (i.e., obtain the value of X)

Subject	Synovial Fluid pH	Subject	Synovial Fluid pH
HW	7.02	BG	7.34
AD	7.35	GL	7.22
TK	7.32	BP	7.32
EP	7.33	NK	7.40 (4)
AF	7.15	LL	6.99
LW	7.26	KC	7.10
LT	7.25	FA	7.30
DR	7.35	ML	7.21
VU	7.38	CK	7.33
SP	7.20	LW	7.28
MM	7.31	ES	7.35
DF	7.24	DD	7.24
LM	7.34	SL	7.36
AW	7.32	RM	7.09
BB	7.34	AL	7.32
TL	7.14	BV	6.95
PM	7.20	WR	7.35
JG	7.41 (1)	HT	7.36
DH	7.77 (2)	ND	6.60
ER	7.12	SJ	7.29
DP	7.45 (3)	BA	7.31
FF	7.28		

Sol 1. We first count how many samples exceeding the median (i.e., obtain the value of X)

Subject	Synovial Fluid pH	Subject	Synovial Fluid pH
HW	7.02	BG	7.34
AD	7.35	GL	7.22
TK	7.32	BP	7.32
EP	7.33	NK	7.40 (4)
AF	7.15	LL	6.99
LW	7.26	KC	7.10
LT	7.25	FA	7.30
DR	7.35	ML	7.21
VU	7.38	CK	7.33
SP	7.20	LW	7.28
MM	7.31	ES	7.35
DF	7.24	DD	7.24
LM	7.34	SL	7.36
AW	7.32	RM	7.09
BB	7.34	AL	7.32
TL	7.14	BV	6.95
PM	7.20	WR	7.35
JG	7.41 (1)	HT	7.36
DH	7.77 (2)	ND	6.60
ER	7.12	SJ	7.29
DP	7.45 (3)	BA	7.31
FF	7.28		

Sol 1. We first count how many samples exceeding the median (i.e., obtain the value of X)

Subject	Synovial Fluid pH	Subject	Synovial Fluid pH
HW	7.02	BG	7.34
AD	7.35	GL	7.22
TK	7.32	BP	7.32
EP	7.33	NK	7.40 (4)
AF	7.15	LL	6.99
LW	7.26	KC	7.10
LT	7.25	FA	7.30
DR	7.35	ML	7.21
VU	7.38	CK	7.33
SP	7.20	LW	7.28
MM	7.31	ES	7.35
DF	7.24	DD	7.24
LM	7.34	SL	7.36
AW	7.32	RM	7.09
BB	7.34	AL	7.32
TL	7.14	BV	6.95
PM	7.20	WR	7.35
JG	7.41 (1)	HT	7.36
DH	7.77 (2)	ND	6.60
ER	7.12	SJ	7.29
DP	7.45 (3)	BA	7.31
FF	7.28		

Hence, we have $k = 4$, $n = 43$, and since n is large, we use the z test:

$$z = \frac{4 - 43/2}{\sqrt{43/4}} = -5.34.$$

Since the critical regions (two-sided test here) are

$$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$$
$$||$$
$$(-\infty, -2.58) \cup (2.58, \infty),$$

we reject the hypothesis.

Or equivalently, the p-value is

$$2 \times \mathbb{P}(Z < -5.34) = 9.294658 \times 10^{-8}.$$

■

```
> pnorm(-5.34) *2
[1] 9.294658e-08
```

Hence, we have $k = 4$, $n = 43$, and since n is large, we use the z test:

$$z = \frac{4 - 43/2}{\sqrt{43/4}} = -5.34.$$

Since the critical regions (two-sided test here) are

$$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$$
$$||$$
$$(-\infty, -2.58) \cup (2.58, \infty),$$

we reject the hypothesis.

Or equivalently, the p-value is

$$2 \times \mathbb{P}(Z < -5.34) = 9.294658 \times 10^{-8}.$$

■

```
> pnorm(-5.34) *2
[1] 9.294658e-08
```

Hence, we have $k = 4$, $n = 43$, and since n is large, we use the z test:

$$z = \frac{4 - 43/2}{\sqrt{43/4}} = -5.34.$$

Since the critical regions (two-sided test here) are

$$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$$
$$||$$
$$(-\infty, -2.58) \cup (2.58, \infty),$$

we reject the hypothesis.

Or equivalently, the p-value is

$$2 \times \mathbb{P}(Z < -5.34) = 9.294658 \times 10^{-8}.$$

```
> pnorm(-5.34) *2
[1] 9.294658e-08
```

Hence, we have $k = 4$, $n = 43$, and since n is large, we use the z test:

$$z = \frac{4 - 43/2}{\sqrt{43/4}} = -5.34.$$

Since the critical regions (two-sided test here) are

$$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$$
$$||$$
$$(-\infty, -2.58) \cup (2.58, \infty),$$

we reject the hypothesis.

Or equivalently, the p-value is

$$2 \times \mathbb{P}(Z < -5.34) = 9.294658 \times 10^{-8}.$$

```
> pnorm(-5.34) *2
[1] 9.294658e-08
```

Sol 2. We can also carry out the exact computation thanks to computer:

The exact p-value should be

$$2 \times \mathbb{P}(X \leq 5) = 2 \sum_{k=0}^{5} \binom{43}{k} \left(\frac{1}{2}\right)^{43} = 2.49951 \times 10^{-7},$$

which is smaller than $\alpha = 0.10$.

Hence, rejection! ∎

```
> pbinom(5,43,0.5) * 2
[1] 2.49951e-07
```

Sol 2. We can also carry out the exact computation thanks to computer: The exact p-value should be

$$2 \times \mathbb{P}(X \leq 5) = 2 \sum_{k=0}^{5} \binom{43}{k} \left(\frac{1}{2}\right)^{43} = 2.49951 \times 10^{-7},$$

which is smaller than $\alpha = 0.10$.

Hence, rejection! ■

```
> pbinom(5,43,0.5) * 2
[1] 2.49951e-07
```

Sol 2. We can also carry out the exact computation thanks to computer: The exact p-value should be

$$2 \times \mathbb{P}(X \leq 5) = 2 \sum_{k=0}^{5} \binom{43}{k} \left(\frac{1}{2}\right)^{43} = 2.49951 \times 10^{-7},$$

which is smaller than $\alpha = 0.10$.

Hence, rejection! ■

```
> pbinom(5,43,0.5) * 2
[1] 2.49951e-07
```

Sign test for paired data

E.g. A manufacturer produces two products, A and B. The manufacturer wishes to know if consumers prefer product B over product A.

A sample of 10 consumers are each given product A and product B, and asked which product they prefer:

Preferences	Number
B	8
A	1
No preference	1

Test at $\alpha = 0.10$ that

H_0: consumers do not prefer B over A

vs.

H_1: consumers do prefer B over A.

Sign test for paired data

E.g. A manufacturer produces two products, A and B. The manufacturer wishes to know if consumers prefer product B over product A.

A sample of 10 consumers are each given product A and product B, and asked which product they prefer:

Preferences	Number
B	8
A	1
No preference	1

Test at $\alpha = 0.10$ that

$$H_0\text{: consumers do not prefer B over A}$$

vs.

$$H_1\text{: consumers do prefer B over A.}$$

Sign test for paired data

E.g. A manufacturer produces two products, A and B. The manufacturer wishes to know if consumers prefer product B over product A.

A sample of 10 consumers are each given product A and product B, and asked which product they prefer:

Preferences	Number
B	8
A	1
No preference	1

Test at $\alpha = 0.10$ that

H_0: consumers do not prefer B over A

vs.

H_1: consumers do prefer B over A.

Sol. We first remove the ties. So that we have a random (paired-data) sample of size $n = 9$.

Under H_0, the consumers have no preference for B over A. Hence, we may believe that consumers will choose A or B with probability $\frac{1}{2}$.

Hence, to get more extreme values in this setting would give the p-value:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{9} \binom{9}{k} \left(\frac{1}{2}\right)^9 = 0.0195.$$

Conclusion, Rejection! ■

Sol. We first remove the ties. So that we have a random (paired-data) sample of size $n = 9$.

Under H_0, the consumers have no preference for B over A. Hence, we may believe that consumers will choose A or B with probability $\frac{1}{2}$.

Hence, to get more extreme values in this setting would give the p-value:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{9} \binom{9}{k} \left(\frac{1}{2}\right)^9 = 0.0195.$$

Conclusion, Rejection! ■

Sol. We first remove the ties. So that we have a random (paired-data) sample of size $n = 9$.

Under H_0, the consumers have no preference for B over A. Hence, we may believe that consumers will choose A or B with probability $\frac{1}{2}$.

Hence, to get more extreme values in this setting would give the p-value:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{9} \binom{9}{k} \left(\frac{1}{2}\right)^9 = 0.0195.$$

Conclusion, Rejection! ■

Sol. We first remove the ties. So that we have a random (paired-data) sample of size $n = 9$.

Under H_0, the consumers have no preference for B over A. Hence, we may believe that consumers will choose A or B with probability $\frac{1}{2}$.

Hence, to get more extreme values in this setting would give the p-value:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{9} \binom{9}{k} \left(\frac{1}{2}\right)^9 = 0.0195.$$

Conclusion, Rejection! ■