

Big Data Fundamentals and Applications

Statistical Analysis (VII)

Nonparametric Statistics

Asst. Prof. Chan, Chun-Hsiang

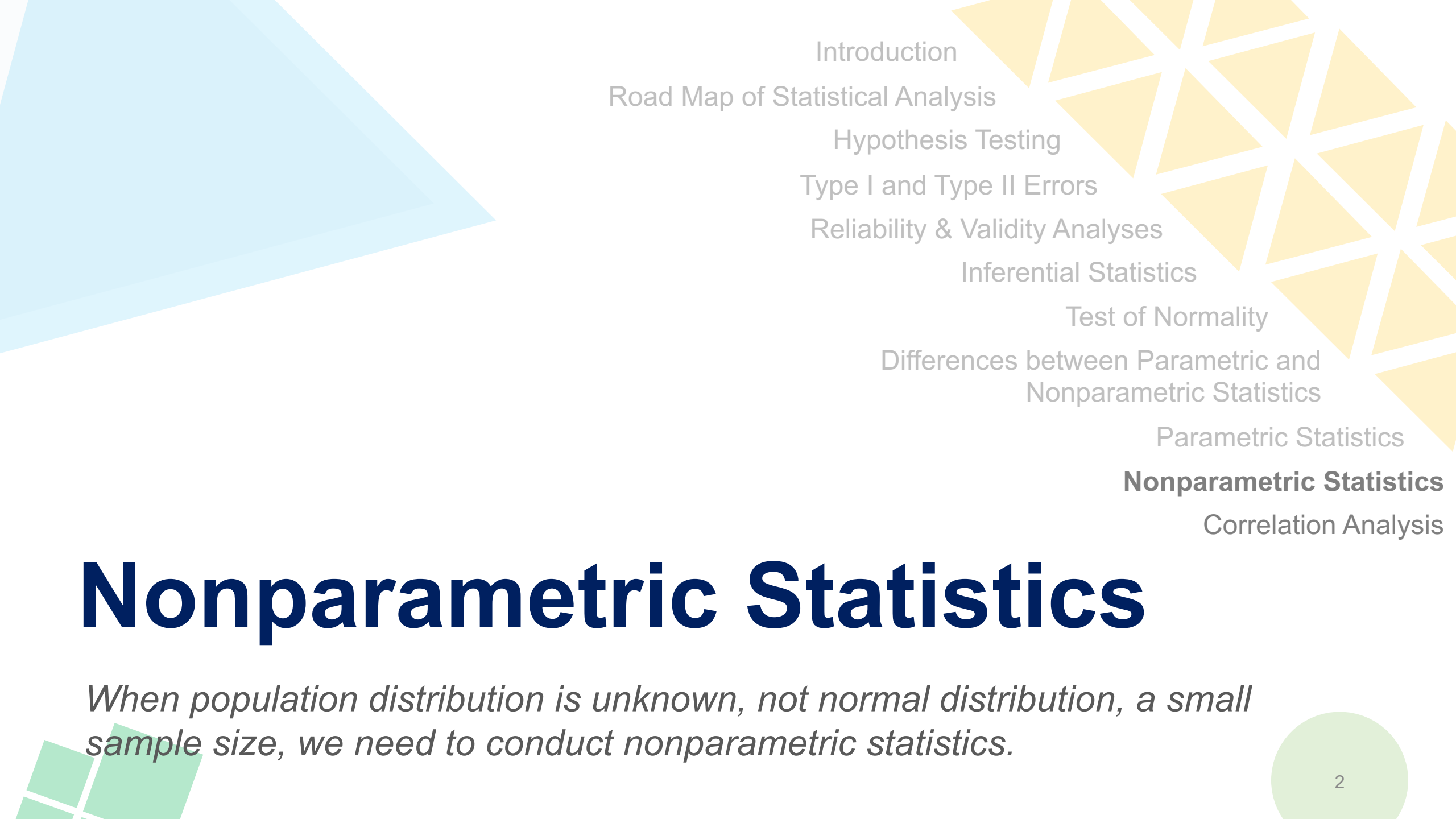
Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan

Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan

Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

1. Introduction
2. Road Map of Statistical Analysis
3. Hypothesis Testing
4. Type I and Type II Errors
5. Reliability & Validity Analyses
6. Inferential Statistics
7. Test of Normality
8. Differences between Parametric and Nonparametric Statistics
9. Parametric Statistics
10. Nonparametric Statistics
11. Correlation Analysis
12. Question Time



Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and
Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Nonparametric Statistics

When population distribution is unknown, not normal distribution, a small sample size, we need to conduct nonparametric statistics.

Nonparametric Statistics

	One Sample	Two Independent Sample	Paired Sample	Multiple Independent Sample
Binary		Chi-squared Fisher's exact	McNemar's test	Chi-squared
Normal and continuous variable	One-Sample t-test	Independent Sample t-test	Paired t-test	One-way ANOVA
Non-normal and continuous variable	Wilconxon signed rank test Signed test	Mann-Whitney U test (Wilconxon rank-sum test)	Wilconxon signed rank test Signed test	Kruskal-Wallis test
Rank variable	Wilconxon signed rank test Signed test	Mann-Whitney U test (Wilconxon rank-sum test)	Wilconxon signed rank test Signed test	Kruskal-Wallis test

Chi-square Test

Edu.	Men	Women	Total
BS	30	40	70
MS	80	50	130
Total	110	90	200

- As abovementioned, chi-square test is to measure the independence, normality, and goodness of fit. Here, we discuss about the independence relationship between two variables.
- **Null hypothesis (H_0):** two variables are independent.
- **Alternative hypothesis (H_1):** two variables are dependent.

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - m_i)^2}{m_i}$$

$$\chi^2 = \frac{\left(30 - \frac{70 \times 110}{200}\right)^2}{\frac{70 \times 110}{200}} + \frac{\left(40 - \frac{70 \times 90}{200}\right)^2}{\frac{70 \times 90}{200}} + \frac{\left(80 - \frac{130 \times 110}{200}\right)^2}{\frac{130 \times 110}{200}} + \frac{\left(90 - \frac{130 \times 90}{200}\right)^2}{\frac{130 \times 90}{200}}$$

$$\chi^2 = \frac{(30 - 38.5)^2}{38.5} + \frac{(40 - 31.5)^2}{31.5} + \frac{(80 - 71.5)^2}{71.5} + \frac{(90 - 58.5)^2}{58.5} = 6.416 \Rightarrow P \text{ value} = 0.0113$$

Edu.	Men	Women
BS	38.5	31.5
MS	71.5	58.5

Fisher's Exact Test

	Sample	Sample	Total
Var 1	A	B	A+B
Var 2	C	D	C+D
Total	A+C	B+D	A+B+C+D=N

- If your sample size for each category (A, B, C, D) is smaller than 5 or total sample size (N) is smaller than 40, then we prefer to use Fisher's exact test.
- **Null hypothesis (H_0):** the distribution of two groups is equal.
- **Alternative hypothesis (H_1):** the distribution of two groups is unequal.

$$P = \frac{\binom{A+B}{A} \binom{C+D}{C}}{\binom{N}{A+C}} = \frac{(A+B)! (C+D)! (A+C)! (B+D)!}{N! A! B! C! D!}$$

Fisher's Exact Test

$$P = \frac{\binom{A+B}{A} \binom{C+D}{C}}{\binom{N}{A+C}} = \frac{(A+B)! (C+D)! (A+C)! (B+D)!}{N! A! B! C! D!}$$

$$P = \frac{\binom{7}{3} \binom{13}{8}}{\binom{20}{8}} = \frac{7! 13! 11! 9!}{20! 3! 4! 8! 5!} = 0.26819$$

	Sample	Sample	Total
Var 1	A	B	A+B
Var 2	C	D	C+D
Total	A+C	B+D	A+B+C+D=N

Edu.	Men	Women	Total
BS	3	4	7
MS	8	5	13
Total	11	9	20

McNemar's Test

- **McNemar's test** is a statistical test used on paired nominal data, like a paired-sample χ^2 test.
- It is applied to 2×2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (that is, whether there is "marginal homogeneity").
- **Goal:** To investigate whether the number of changes from "yes" to "no" and from "no" to "yes" before and after the experiment is equal, it is called "significant change test".

McNemar's Test

		Test 2		Row Total
		Positive	Negative	
Test 1	Positive	a	b	a+b
	Negative	c	d	c+d
Column Total		a+c	b+d	N

- The test is applied to a 2×2 contingency table, which tabulates the outcomes of two tests on a sample of N subjects, as follows.
- The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, i.e., $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$.
- **Null hypothesis (H_0):** $p_b = p_c$
- **Alternative hypothesis (H_1):** $p_b \neq p_c$
- The **McNemar test statistic** is: $\chi^2 = \frac{(b-c)^2}{b+c}, df = 1$

McNemar's Test

- If either b or c is small ($b + c < 25$) then χ^2 is not well-approximated by the chi-squared distribution.
- An exact binomial test can then be used, where b is compared to a binomial distribution with size parameter $n = b + c$ and $p = 0.5$. Effectively, the exact binomial test evaluates the imbalance in the discordant b and c .
- To achieve a two-sided P value, the P value of the extreme tail should be multiplied by 2. For $b \geq \frac{n}{2}$:

$$\text{Exact } - P \text{ value} = 2 \sum_{i=b}^n \binom{n}{i} (0.5)^i (1 - 0.5)^{n-i}$$

McNemar's Test

		After :: OP		Row Total
		Yes	No	
Before :: DM	Yes	205	30	235
	No	60	65	125
Column Total		265	95	360

$$\chi^2 = \frac{(b - c)^2}{b + c} = \frac{(60 - 30)^2}{60 + 30} = \frac{900}{90} = 10$$

P value = 0.001565.

The null hypothesis is rejected because $0.001565 < 0.05$.

Signed Test

- **Assumption:** categorical data with non-normality or rank data
- **Goal:** test whether the median (M_0) of a single population is a certain value or whether the distribution of paired populations is the same.
- Given a series of data as $x_1, x_2, x_3, x_4, \dots, x_n$

$$z_i = \begin{cases} \text{when } x_i - M_0 < 0 \Rightarrow -1 \\ \text{when } x_i - M_0 = 0 \Rightarrow 0 \\ \text{when } x_i - M_0 > 0 \Rightarrow +1 \end{cases}$$

$S^+, S^-, S^0 \in \text{binomial distribution}$
 S^+, S^-, S^0 are the numbers of $M > M_0, M < M_0,$
and $M = M_0,$ respectively.

	Left-tailed Signed Test	Right-tailed Signed Test	Two-tailed Signed Test
H_0 vs H_1	$\begin{cases} H_0: M = M_0 \\ H_1: M > M_0 \end{cases}$	$\begin{cases} H_0: M = M_0 \\ H_1: M < M_0 \end{cases}$	$\begin{cases} H_0: M = M_0 \\ H_1: M \neq M_0 \end{cases}$
Rejection region	$S^+ \leq c$	$S^- \leq c$	$S^0 \leq c$
P value	$\sum_{k=0}^{S^+} \binom{n'}{k} (0.5)^{n'}$	$\sum_{k=0}^{S^-} \binom{n'}{k} (0.5)^{n'}$	$2 \sum_{k=0}^{S^0} \binom{n'}{k} (0.5)^{n'}$

Signed Test

- If the number of non-zero observations is larger than 20 ($n \geq 20$), then signed test (Y) could be transferred to Z test.

$$Z = \frac{Y - \mu}{\sigma} = \frac{Y - 0.5n}{0.5\sqrt{n}}$$

where $\mu = np = 0.5n$ and $\sigma^2 = npq = 0.5^2n$ only if $H_0: P = 0.5$.

- **Null hypothesis (H_0):** $M = M_0$ || **Alternative hypothesis (H_1):** $M >, <, \neq M_0$
- **Null hypothesis (H_0):** $P = 0.5$ || **Alternative hypothesis (H_1):** $P >, <, \neq 0.5$
- Adopt Z test as abovementioned.

Signed Test (One-Sample) [category]

- 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.
- The **null hypothesis** is that consumers do **not** prefer product B **over** product A.
- The **alternative hypothesis** is that consumers **prefer** product B **over** product A.
- This is a **one-sided (directional) test**.

Signed Test (One-Sample) [category]

- At the end of the study, ...

Consumer ID	Preference	Signed
1	B	+
2	A	-
3	B	+
4	B	+
5	B	+
6	No reported	
7	B	+
8	B	+
9	B	+
10	B	+

The tie (no reported) is excluded.

→ $n = \text{number of } + \& - = 8 + 1 = 9.$

$$\sum_{k=8}^9 \binom{9}{k} 0.5^k 0.5^{9-k} = \left[\binom{9}{8} + \binom{9}{9} \right] \times 0.5^9$$
$$= 10 \times 0.001953125 \sim 0.0195$$

$P(8 \text{ or } 9 \text{ heads in } 9 \text{ flips of a fair coin}) = 0.0195.$
The null hypothesis is rejected, and the manufacturer concludes that consumers prefer product B over product A.

Signed Test (One-Sample) [numeric]

- In a clinical trial, survival time (weeks) is collected for 10 subjects with non-Hodgkin's lymphoma. The exact survival time was not known for one subject who was still alive after 362 weeks, when the study ended.
- The subjects' survival times were 49, 58, 75, 110, 112, 132, 151, 276, 281, 362+
- The plus sign indicates the subject still alive at the end of the study.
- **The researcher wished to determine if the median survival time was less than or greater than 200 weeks.**

Signed Test (One-Sample) [numeric]

- **Null hypothesis** is that median survival is 200 weeks.
- **Alternative hypothesis** is that median survival is not 200 weeks.
- **A two-sided test:** the alternative median may be greater than or less than 200 weeks.

Patient ID	Survival Week	Signed
1	49	-
2	58	-
3	75	-
4	110	-
5	112	-
6	132	-
7	151	-
8	276	+
9	281	+
10	362+	+

Signed Test (One-Sample) [numeric]

- The probability for each value of k is given in the table below.

$$\sum_{k=0}^{10} \binom{10}{k} 0.5^k 0.5^{10-k} = \sum_{k=0}^{10} \binom{10}{k} 0.5^{10} = 0.0009765625 \times \sum_{k=0}^{10} \binom{10}{k}$$

k	0	1	2	3	4	5	6	7	8	9	10
Prob	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

- The probability of **0, 1, 2, 3, 7, 8, 9, or 10** heads in 10 tosses is the sum of their individual probabilities: **0.0010 + 0.0098 + 0.0439 + 0.1172 + 0.1172 + 0.0439 + 0.0098 + 0.0010** = 0.3438, where is not rejected at a significance level of P value = 0.05.

Signed Test (Two-Sample)

Entry	Bus A	Bus B	Difference
1	42	41	+
2	50	40	+
3	45	30	+
4	31	12	+
5	26	15	+
6	80	35	+
7	50	65	-
8	60	45	+
9	40	52	-
10	30	25	+

- A two-tailed test
- A result as extreme or extreme more than 6 positive differences includes the results of 8,9,10 positive differences, and the results of 0,1,2 positive differences.

Signed Test (Two-Sample)

- The probabilities can be calculated using the binomial test, with the probability of heads = probability of tails = 0.5.

<i>k</i>	0	1	2	3	4	5	6	7	8	9	10
Prob	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

- The two-sided probability of a result as extreme as 8 of 10 positive difference is the sum of these probabilities:
 $0.0010 + 0.0098 + 0.0439 + 0.0439 + 0.0098 + 0.0010 = 0.1094$, where is not rejected at a significance level of P value = 0.05.

Wilcoxon Signed Rank Test

- **Goal:** test the location of a population based on a sample of data, or to compare the locations of two populations using two matched samples.
- **Characteristics:** The one-sample version serves a purpose similar to that of the one-sample Student's t -test. For two matched samples, it is a paired difference test like the paired Student's t -test.
- **Reasons:** The Wilcoxon test can be a good alternative to the t -test when population means are not of interest; for example, when one wishes to test whether a population's median is nonzero, or whether there is a better than 50% chance that a sample from one population is greater than a sample from another population.

Wilcoxon Signed Rank Test

- Wilcoxon signed rank test considers both **the signs of differences and the difference values**; hence, the statistical power of Wilcoxon signed rank test is higher than signed test.
- Given a series of data as $x_1, x_2, x_3, x_4, \dots, x_n$

$$z_i = |x_i - M_0|, i = 0, 1, 2, \dots, n$$

	Left-tailed Signed Test	Right-tailed Signed Test	Two-tailed Signed Test
H_0 vs H_1	$\begin{cases} H_0: M = M_0 \\ H_1: M > M_0 \end{cases}$	$\begin{cases} H_0: M = M_0 \\ H_1: M < M_0 \end{cases}$	$\begin{cases} H_0: M = M_0 \\ H_1: M \neq M_0 \end{cases}$
Rejection region	$T^+ \leq c$	$T^- \leq c$	$T^0 = \min(T^+, T^-) \leq c$

Wilcoxon Signed Rank Test (Two-Sample)

i	$x_{2,i}$	$x_{1,i}$	$x_{2,i} - x_{1,i}$	
			sgn	abs
1	125	110	+1	15
2	115	122	-1	7
3	130	125	+1	5
4	140	120	+1	20
5	140	140	0	0
6	115	124	-1	9
7	140	123	+1	17
8	125	137	-1	12
9	140	135	+1	5
10	135	145	-1	10

i	$x_{2,i}$	$x_{1,i}$	$x_{2,i} - x_{1,i}$			
			sgn	abs	R_i	$sgn \cdot R_i$
5	140	140		0		
3	130	125	+1	5	1.5	1.5
9	140	135	+1	5	1.5	1.5
2	115	122	-1	7	3	-3
6	115	124	-1	9	4	-4
10	135	145	-1	10	5	-5
8	125	137	-1	12	6	-6
1	125	110	+1	15	7	7
7	140	123	+1	17	8	8
4	140	120	+1	20	9	9

order by absolute difference

Wilcoxon Signed Rank Test (Two-Sample)

- sgn is the sign function, abs is the absolute value, R_i and is the rank. Notice that pairs 3 and 9 are tied in absolute value. They would be ranked 1 and 2, so each gets the average of those ranks, 1.5.
- $W = 1.5 + 1.5 - 3 - 4 - 5 - 6 + 7 + 8 + 9 = 9$
- $|W| < W_{crit}(\alpha=0.05, 9, two-sided) = 15$
- \therefore failed to reject H_0 that the median of pairwise differences is different from zero.
- The P value for this result is 0.6113.

Wilcoxon Signed Rank Test

- When n is larger than 20, T could be approximated as normal distribution.
- The T value could be standardized as $Z = (T - \mu_T) / \sigma_T$.
- $\mu_T = \frac{n(n+1)}{4}$
- $\sigma_T = \sqrt{n(n+1)(2n+1)/24}$
- Usually, we have sample sizes of 20, 40, or even higher, then we may simply adopt paired t-test instead of Wilcoxon signed rank test.

Mann-Whitney U Test

- **Mann-Whitney U Test** is used instead of **two-independent t-test**, where test if the mean values from two groups are equal.
- The assumption of normality is not required.
- **Assumption:** two populations are continuous distribution with the same variations; two samples are random, where the sample sizes are n_1, n_2 .
- **Null hypothesis (H_0):** $\eta_1 - \eta_2 = 0$
- **Alternative hypothesis (H_1):** $\eta_1 - \eta_2 \neq 0$

$$U = \frac{n_1 n_2 + n_1(n_1 + 1)}{2} - R_1, \text{ where } R_1 \text{ is the sum of rank of first group.}$$

- When n_1 & $n_2 \geq 10$, then U approximates to normal distribution, where $\mu_u = \frac{n_1 n_2}{2}$, $\sigma_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$, $Z = \frac{U - \mu_u}{\sigma_u}$.

Mann-Whitney U Test

Class A	Rank	Class B	Rank
30	2	20	1
40	4	35	3
45	6	41	5
60	8	50	7
80	9	81	10
90	11	92	13
91	12	93	14
95	15	96	16
97	17	98	18
99	19	100	20
$n_1 = 10$	$R_1 = 103$	$n_2 = 10$	

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U = 10 \times 10 + \frac{10(10 + 1)}{2} - 103 = 52$$

since n_1 & n_2 are larger than 10.

U can be standardized as Z

$$\mu_u = \frac{n_1 n_2}{2} = \frac{10 \times 10}{2} = 50$$

$$\sigma_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{100 \times 21}{12} = 175 \Rightarrow \sigma_u = 13.229$$

$$Z = \frac{U - \mu_u}{\sigma_u} = \frac{52 - 50}{13.229} = 0.15118 < 1.96 (Z_{0.025})$$

Wilcoxon Rank Sum Test

- Null hypothesis (H_0): $\eta_1 - \eta_2 = 0$
- Alternative hypothesis (H_1): $\eta_1 - \eta_2 \neq 0$

$T = R_1$, where R_1 is the sum of rank of first group.

- When n_1 & $n_2 \geq 10$, then T approximates to normal distribution,

where $\mu_T = \frac{n_1(n_1+n_2+1)}{2}$, $\sigma_T^2 = \frac{n_1n_2(n_1+n_2+1)}{12}$, $Z = \frac{U-\mu_T}{\sigma_T}$.

Wilcoxon Rank Sum Test

Class A	Rank	Class B	Rank
30	2	20	1
40	4	35	3
45	6	41	5
60	8	50	7
80	9	81	10
90	11	92	13
91	12	93	14
95	15	96	16
97	17	98	18
99	19	100	20
$n_1 = 10$	$R_1 = 103$	$n_2 = 10$	

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{10 \times 21}{2} = 105$$

$$\sigma_T^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{100 \times 21}{12}$$

$$= 175 \Rightarrow \sigma_u = 13.229$$

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{103 - 105}{13.229} = -0.15118$$

where is smaller than -1.96 ($Z_{0.025}$)

Kruskal-Wallis Test

- **Kruskal Wallis Test for K Independent Random Samples**
- The assumption of normality is not required.
- **Goal:** test whether the medians of K independent groups are different.
- **Usage:** K-W test can be instead of one-way ANOVA F-test for non-normal distributed dataset.
- **Assumption:** K populations are continuous distribution with the same variations; K samples are random, where the sample sizes are $n_1, n_2, \dots, n_k \geq 5$
- **Null hypothesis (H_0):** $\eta_1 = \eta_2 = \dots = \eta_k$
- **Alternative hypothesis (H_1):** *at least one $\eta_i \neq 0$*

Kruskal-Wallis Test

Calculation Kruskal-Wallis Test

$$W = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$$

when $n \geq 5$, it approximates to χ^2 distribution ($df = k - 1$)

Kruskal-Wallis Test (Example)

The following table shows the midterm exam scores of statistics of the three classes. Please perform whether the midterm exam scores are different among the three classes. $\alpha = 0.05$.

Class A	Rank	Class B	Rank	Class C	Rank
80	13.5	45	4	53	8
60	9.5	85	15	51	7
87	16	62	12	35	2
60	9.5	10	1	47	5
50	6	80	13.5	91	17
92	18	93	19	95	20
44	3	61	11	Rank Sum =	59
Rank Sum =	75.5	Rank Sum =	75.5		

Kruskal-Wallis Test (Example)

$$W = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$$

$$W = \frac{12}{20(20+1)} \left(\frac{75.5^2}{7} + \frac{75.5^2}{7} + \frac{59^2}{6} \right) - 3(20+1) = 11.17952$$

This is a two-tailed test.

$$\chi_{\alpha=0.025, df=2}^2 = 7.378, P \text{ value} = 0.003736$$

$\therefore 11.17952 > 7.378$ ($0.004 < 0.025$),

\therefore null hypothesis is rejected.

Reading

Nonparametric Correlation Techniques: Techniques for Correlating Nominal & Ordinal Variables

<https://staff.blog.ui.ac.id/r-suti/files/2010/05/noparcoringrelationtechniq.pdf>

Parametric and Non-parametric tests for comparing two or more groups

<https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests>

魏克生符號檢定(Wilcoxon sign rank)與符號檢定(sign rank)-SPSS無母數統計

<https://www.yongxi->

[stat.com/%E9%AD%8F%E5%85%8B%E7%94%9F%E7%AC%A6%E8%99%9F%E6%AA%A2%E5%AE%9A_%E7%84%A1%E6%AF%8D%E6%95%B8%E7%B5%B1%E8%A8%88/](https://www.yongxi-stat.com/%E9%AD%8F%E5%85%8B%E7%94%9F%E7%AC%A6%E8%99%9F%E6%AA%A2%E5%AE%9A_%E7%84%A1%E6%AF%8D%E6%95%B8%E7%B5%B1%E8%A8%88/)

第十四章 無母數統計檢定

[https://itunesu-assets.itunes.apple.com/apple-assets-us-std-](https://itunesu-assets.itunes.apple.com/apple-assets-us-std-000001/CobaltPublic3/v4/de/a6/56/dea65603-072f-0678-b2dc-e07d67536737/304-7970202547785007611-14.4.pdf)

[000001/CobaltPublic3/v4/de/a6/56/dea65603-072f-0678-b2dc-e07d67536737/304-7970202547785007611-14.4.pdf](https://itunesu-assets.itunes.apple.com/apple-assets-us-std-000001/CobaltPublic3/v4/de/a6/56/dea65603-072f-0678-b2dc-e07d67536737/304-7970202547785007611-14.4.pdf)

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))