STATISTICAL TEST

what is the purpose statistical test?

- ✓A **statistical test** is a way to evaluate the evidence the data provides against a **null hypothesis**- $\mathbf{H_0}$
- ✓$\mathbf{H_0}$ is usually opposed to an **alternative hypothesis-** $\mathbf{H_a}$
- ✓If the data does not provide enough evidence against $\mathbf{H_0}$, $\mathbf{H_0}$ is not rejected
- ✓If instead, the data shows strong evidence against $\mathbf{H_0}$, $\mathbf{H_0}$ is rejected and $\mathbf{H_a}$ is considered as true with a quantified risk of being wrong

- ✓Whether the average size of Royal Gala apples differs from the average size of Jazz apples
  - **H_0**: average size of Royal Gala apples = average size of Jazz apples
  - **H_a**: average size of Royal Gala apples ≠ average size of Jazz apples

- ✓Whether the average accuracy of EC algorithm 1 differs from the average accuracy of EC algorithm 2
  - $\mathbf{H_0}$: average accuracy of EC algorithm 1 = average accuracy of EC algorithm 2
  - $\mathbf{H_a}$: average accuracy of EC algorithm 1 ≠ average accuracy of EC algorithm 2

Why do we need statistical tests?

- ✓In statistics, **population** refers to the total set of observations while a **sample** is a set of data collected
- ✓We’re looking at a sample rather than the entire population
- ✓A statistical test helps assess the likelihood to reject/not reject null hypothesis from the sample

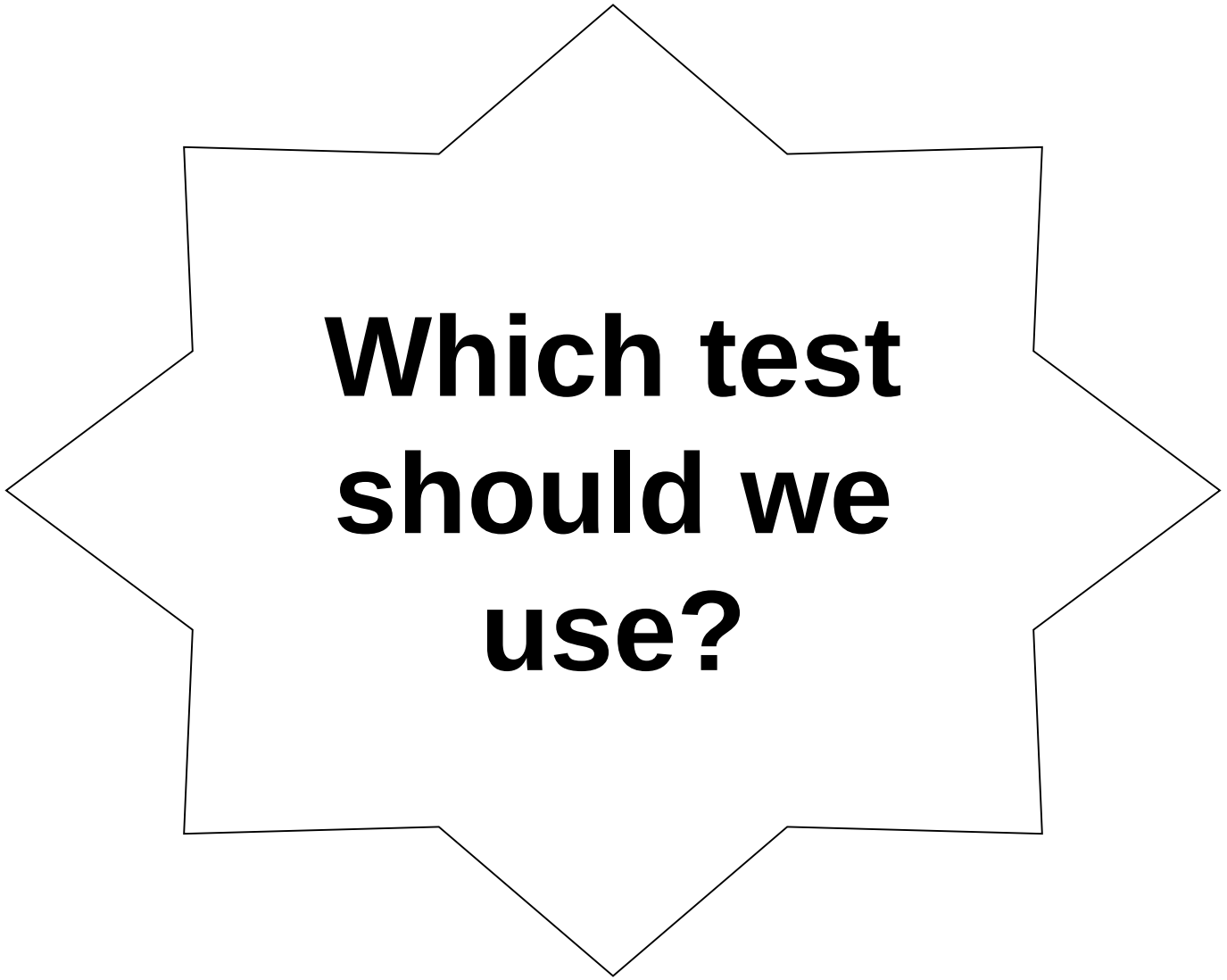
What are the steps to perform hypothesis testing?

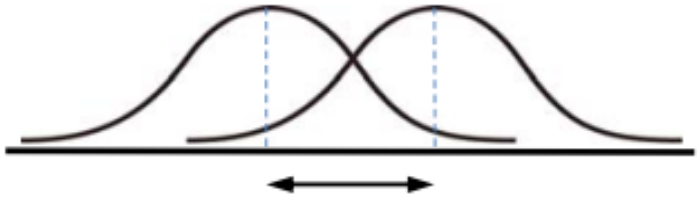
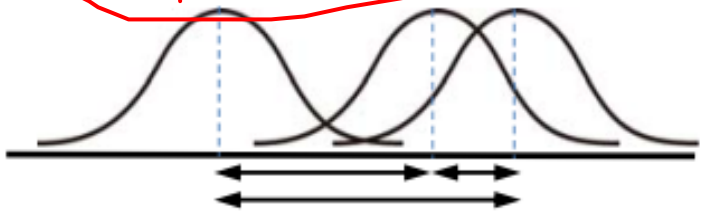
Steps to perform Hypothesis Testing

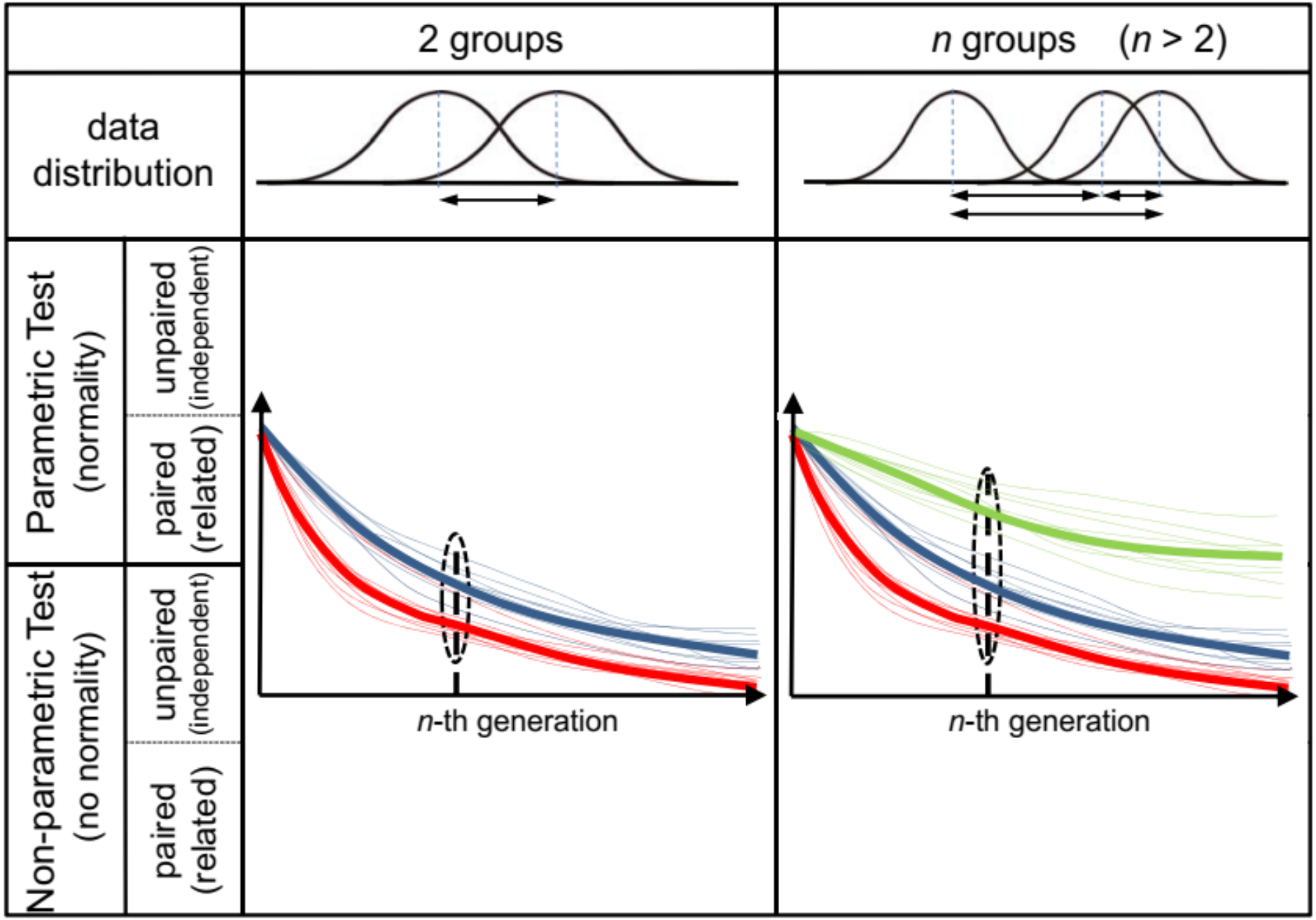
1. Set up Hypothesis (NULL and Alternate)
2. Collect data
3. Select the appropriate test statistic and level of significance
4. Compute the appropriate test statistic and make the decision.

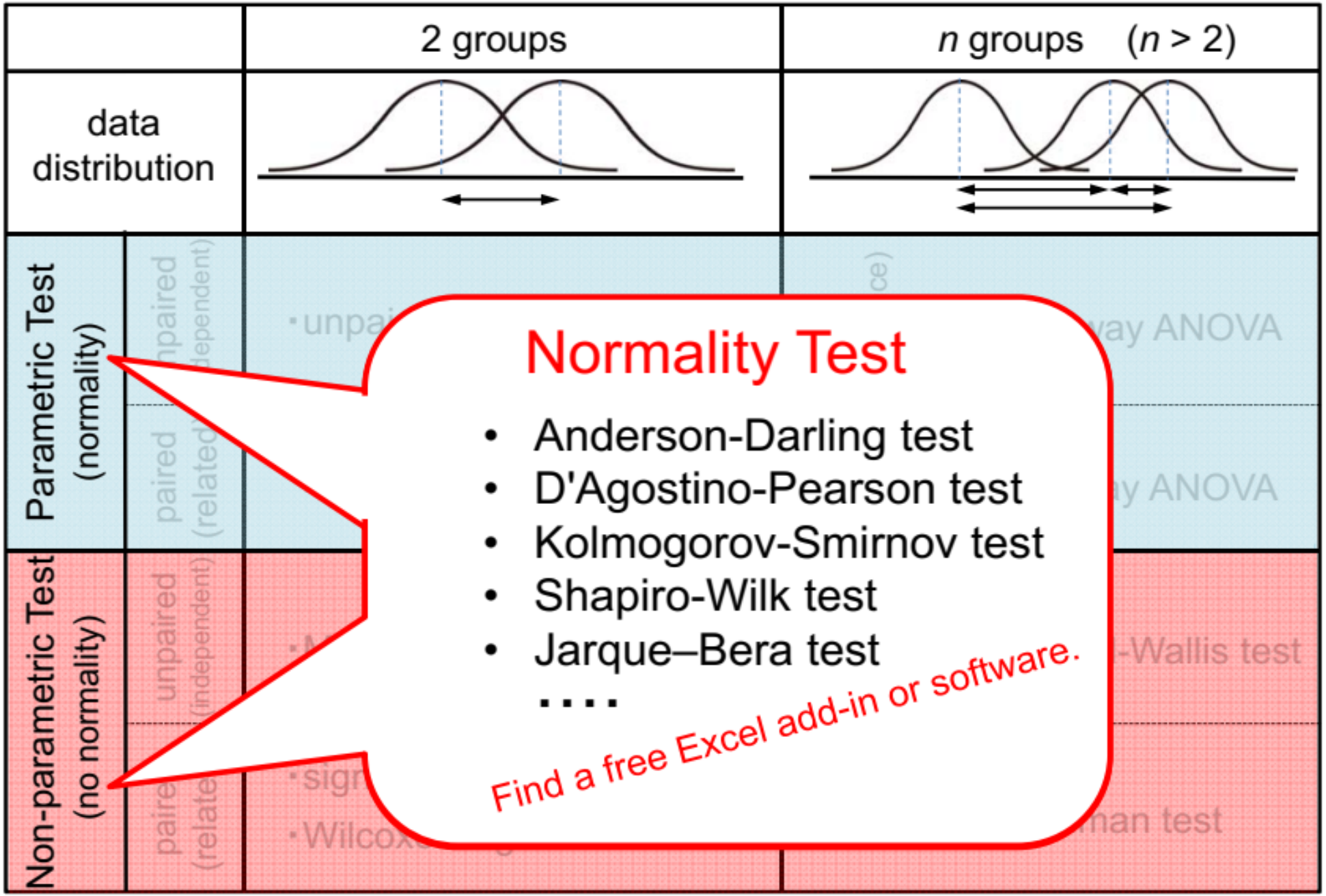
Type I Error vs Type II Error

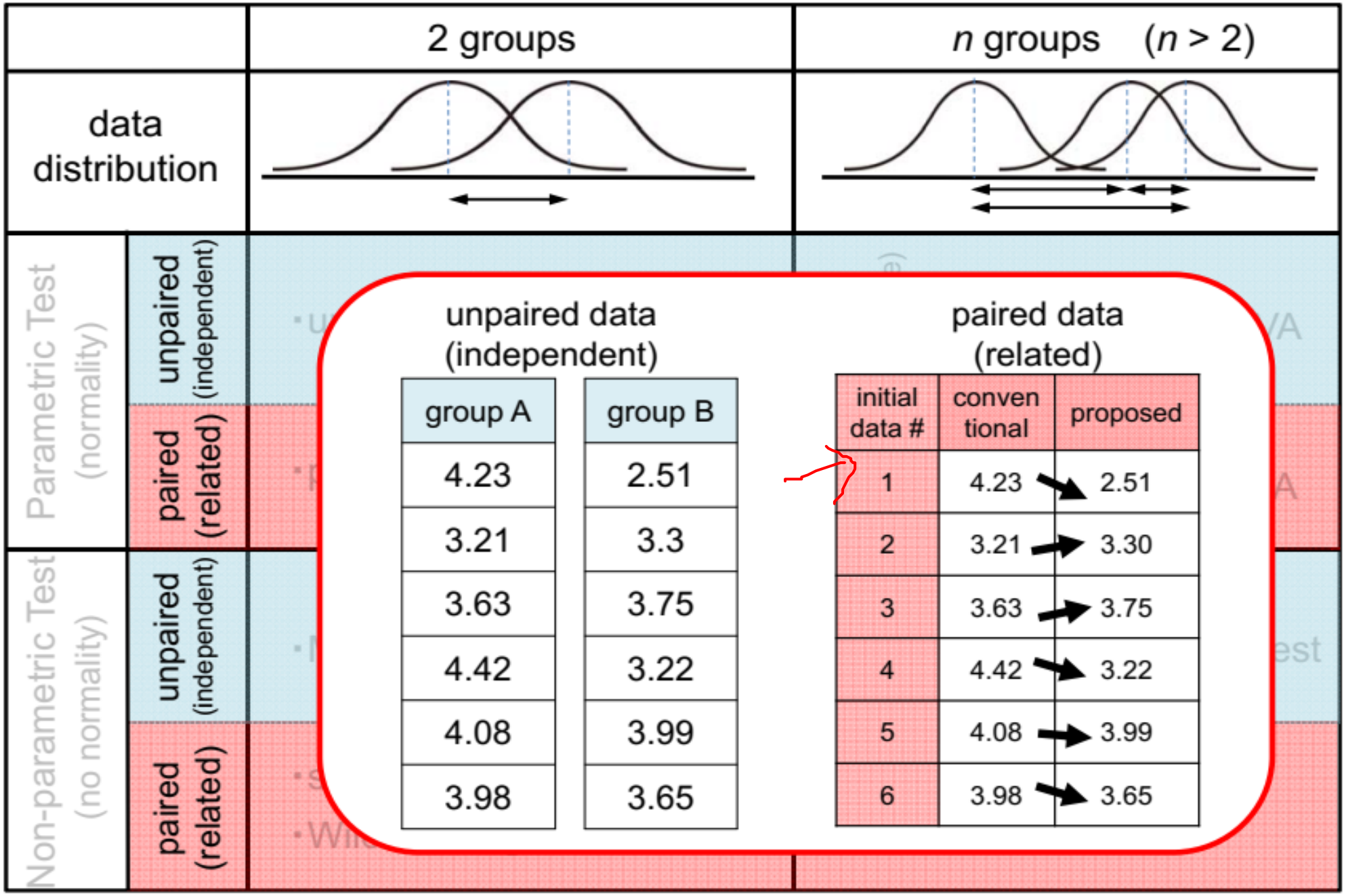
	In Reality	
Decision	H_0 is TRUE	H_0 is FALSE
Accept H_0	OK	Type II Error β = probability of Type II Error
Reject H_0	Type I Error α = probability of Type I Error	OK


Which test should we use?

		2 groups	*n* groups (*n* > 2)
data distribution			
Parametric Test (normality)	unpaired (independent)	▪ unpaired *t*-test	ANOVA (Analysis of Variance) ▪ one-way ANOVA
	paired (related)	▪ paired *t*-test	▪ two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	▪ Mann-Whitney *U*-test	one-way data ▪ Kruskal-Wallis test
	paired (related)	▪ sign test ▪ Wilcoxon signed-ranks test	two-way data ▪ Friedman test


2 groups
n groups (n > 2)
data distribution
Parametric Test (normality)
unpaired (independent)
paired (related)
Non-parametric Test (no normality)
unpaired (independent)
paired (related)
n-th generation
n-th generation

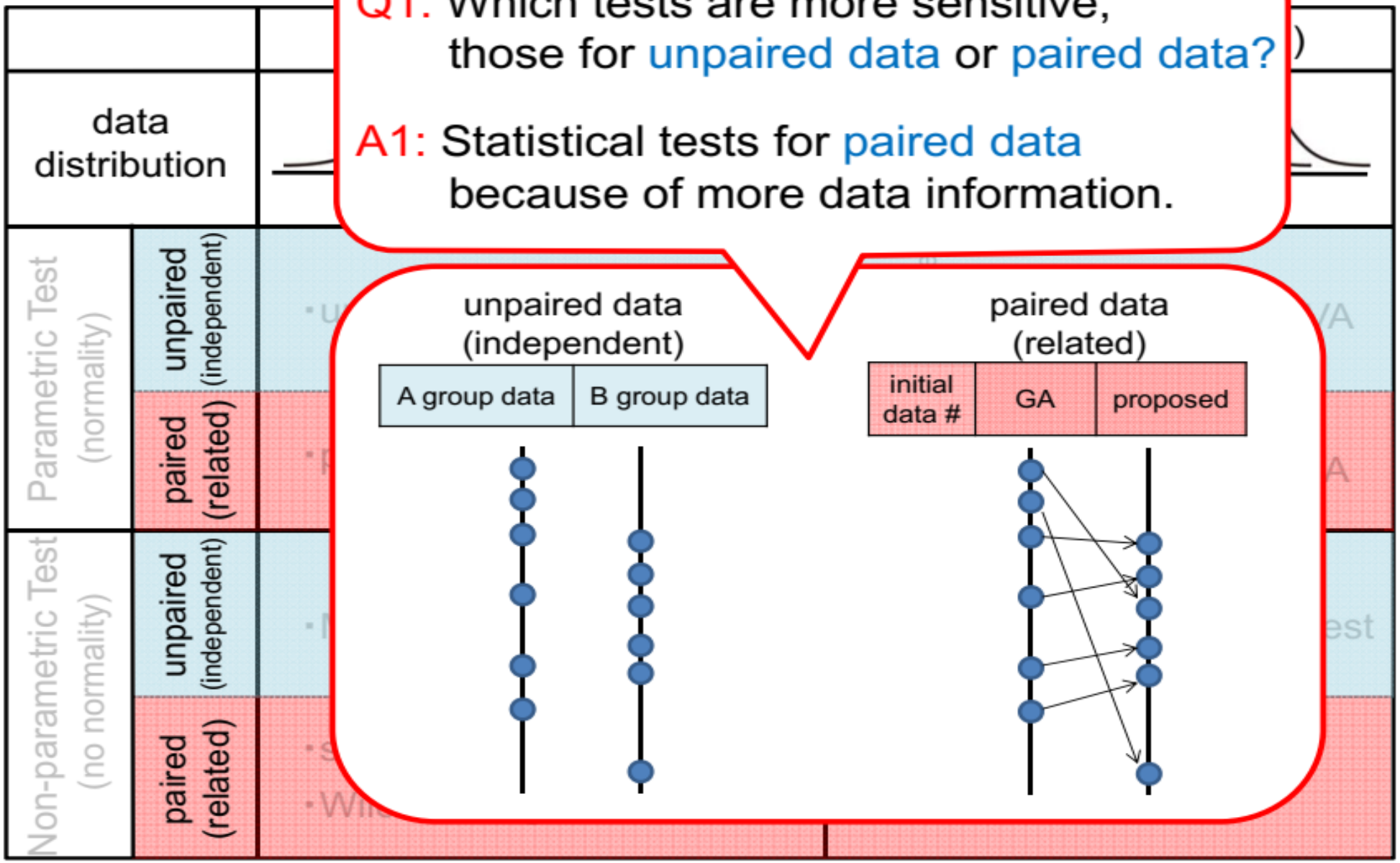

2 groups
n groups (n > 2)
data distribution
Parametric Test (normality)
Non-parametric Test (no normality)
Normality Test
• Anderson-Darling test
• D'Agostino-Pearson test
• Kolmogorov-Smirnov test
• Shapiro-Wilk test
• Jarque–Bera test
....
Find a free Excel add-in or software.


2 groups
n groups (n > 2)
data distribution
Parametric Test (normality)
Non-parametric Test (no normality)
unpaired (independent)
paired (related)
unpaired (independent)
paired (related)
unpaired data (independent)
group A
group B
4.23
2.51
3.21
3.3
3.63
3.75
4.42
3.22
4.08
3.99
3.98
3.65
paired data (related)
initial data #
conventional
proposed
1
4.23
2.51
2
3.21
3.30
3
3.63
3.75
4
4.42
3.22
5
4.08
3.99
6
3.98
3.65

Which Test Should We Use?

Q1: Which tests are more sensitive, those for unpaired data or paired data?

A1: Statistical tests for paired data because of more data information.

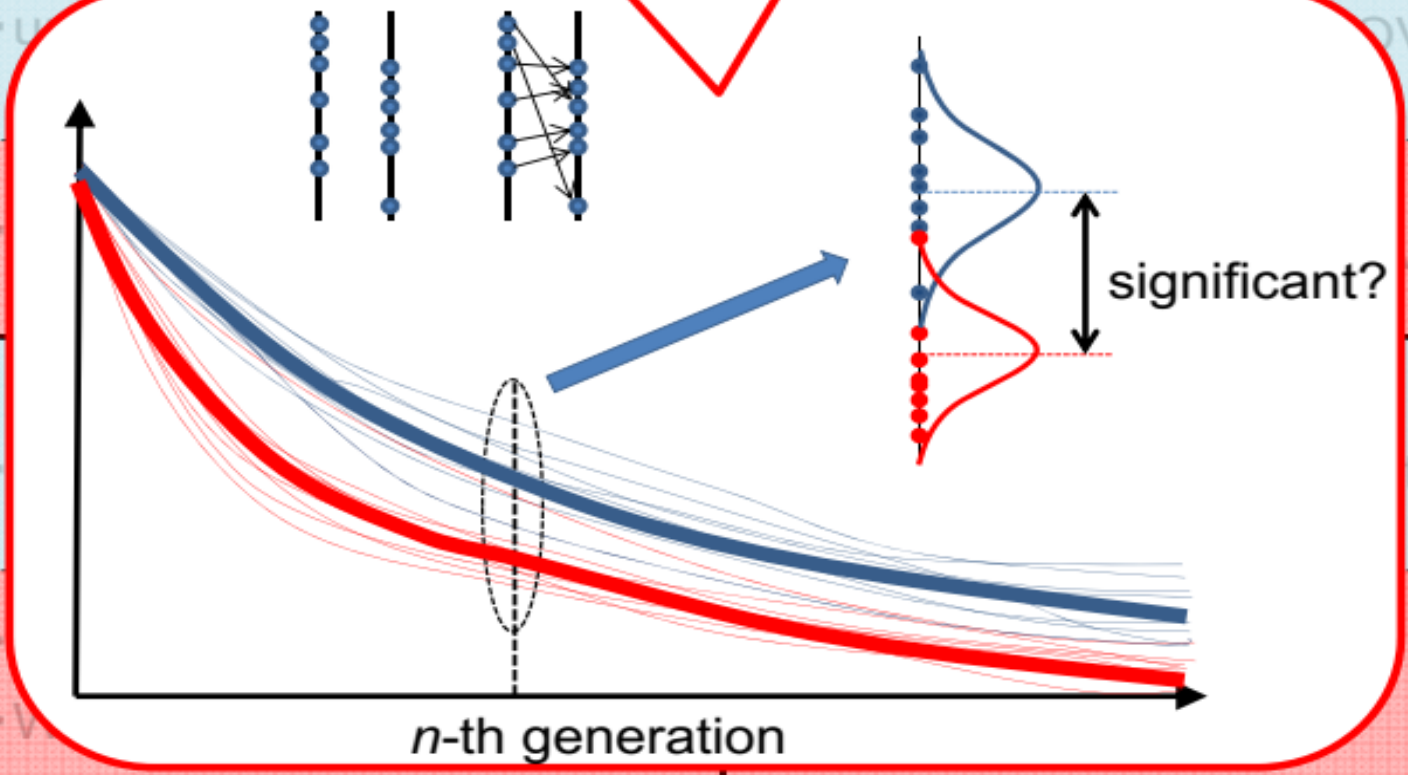


Which Test Should We Use?

Q2: How should you design your experimental conditions to use statistical tests for paired data and reduce the # of trial runs?

A2: Use the same initialized data for the set of (method A, method B) at each trial run.

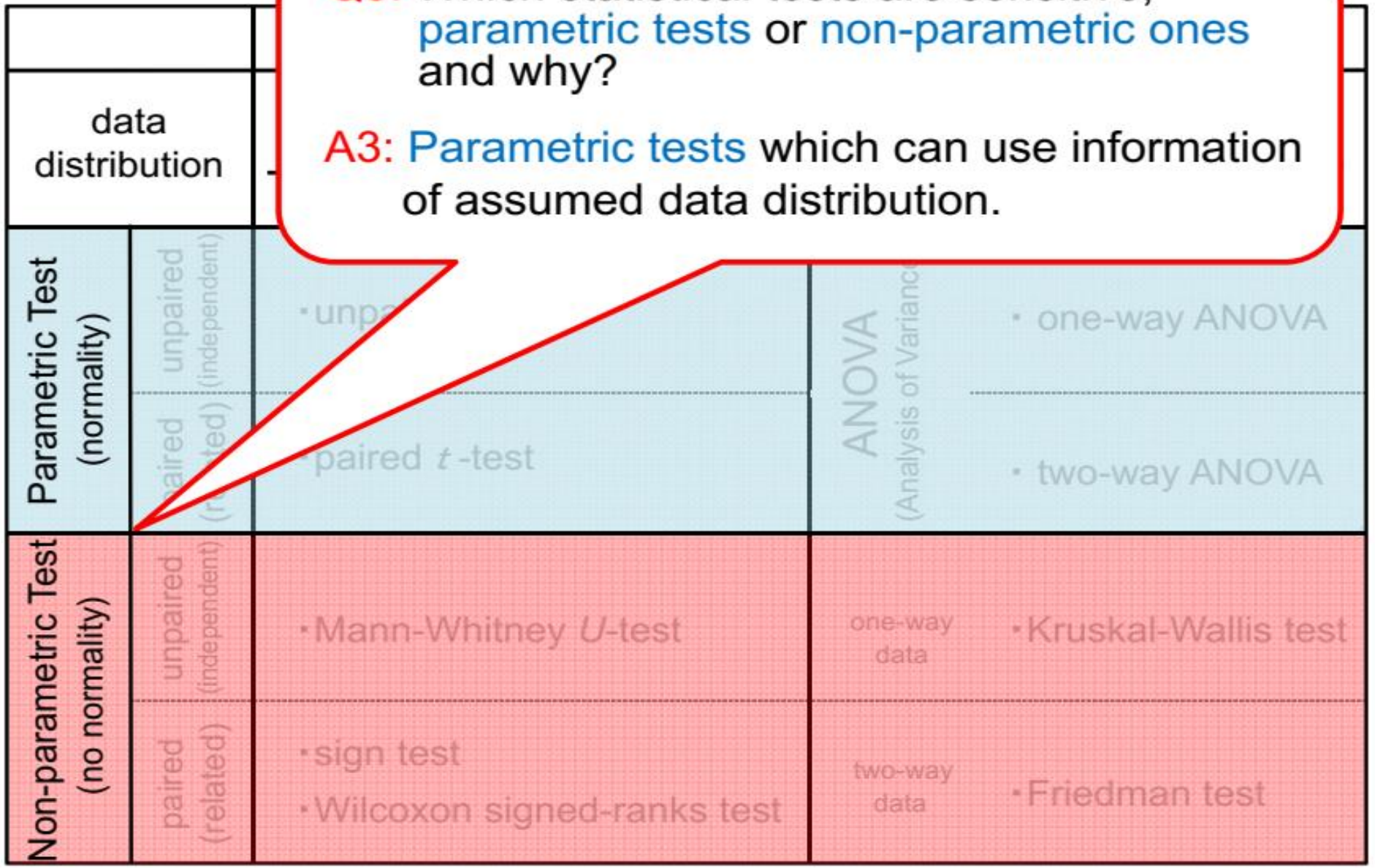
	data distribution	
Parametric Test (normality)	unpaired (independent)	
	paired (related)	
Non-parametric Test (no normality)	unpaired (independent)	
	paired (related)	



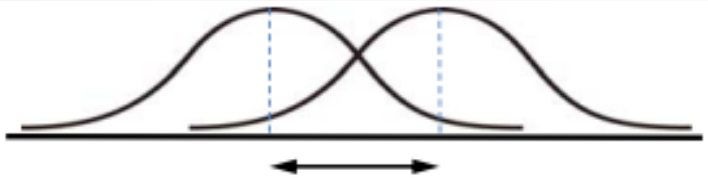
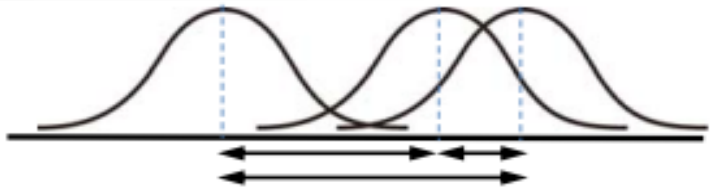
Which Test Should we Use?

Q3: Which statistical tests are sensitive, parametric tests or non-parametric ones and why?

A3: Parametric tests which can use information of assumed data distribution.



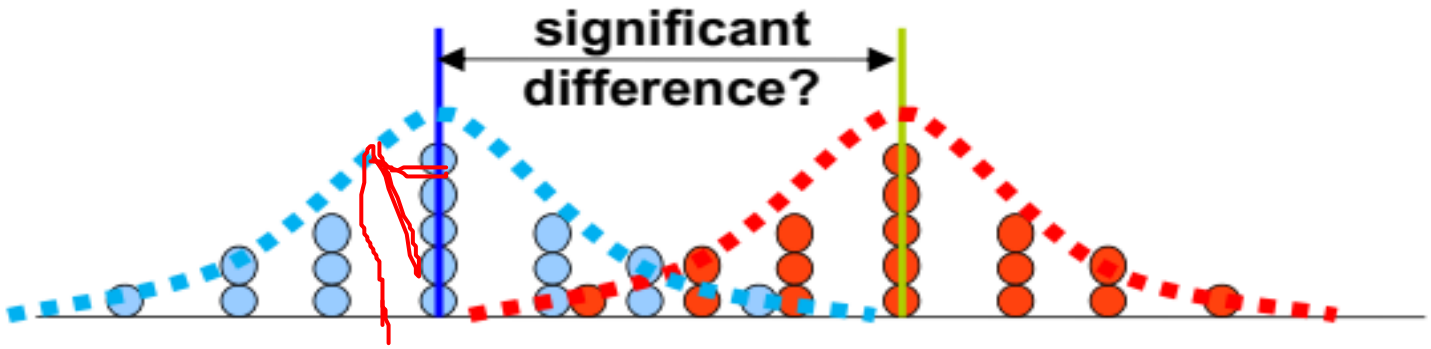
t-Test

		2 groups	*n* groups ($n > 2$)	
data distribution				
Parametric Test (normality)	unpaired (independent)	***t*-test**	ANOVA (Analysis of Variance)	・one-way ANOVA
	paired (related)			・two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	・Mann-Whitney *U*-test	one-way data	・Kruskal-Wallis test
	paired (related)	・sign test ・Wilcoxon signed-ranks test	two-way data	・Friedman test

t-Test

A	B
12	10
14	9
14	7
11	15
16	11
19	10

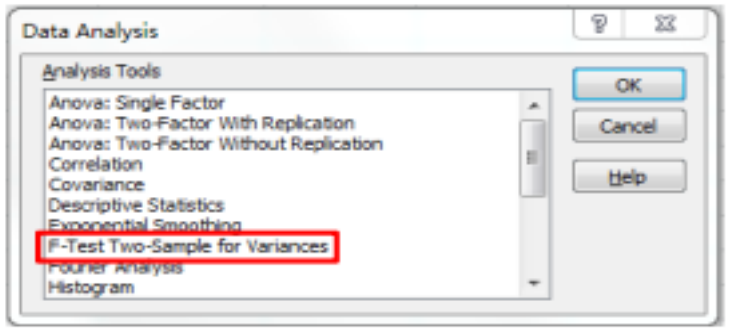
Test this difference with assuming no difference.
(null hypothesis)



Conditions to use *t*-tests:
(1) normality
(2) equal variances (not essential though)

t-Test

F-Test



When ($p > 0.05$), we assume that there is no significant difference between σ^2_A and σ^2_B.

ce with assuming no difference.
null hypothesis)

A	
12	
14	
14	
11	15
16	11
19	10

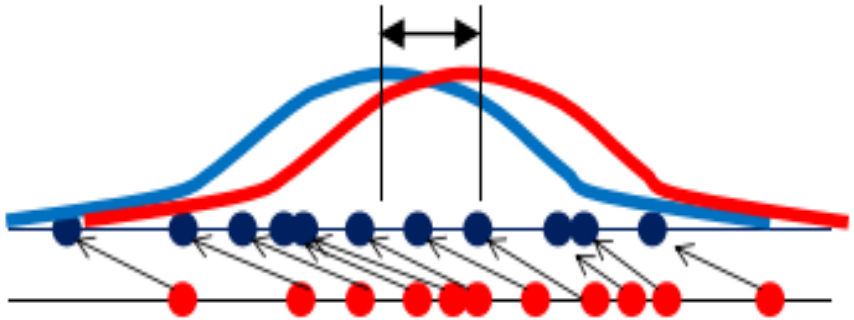
Normality Test

- Anderson-Darling test
- D'Agostino-Pearson test
- Kolmogorov-Smirnov test
- Shapiro-Wilk test
- Jarque–Bera test
- ....

Co ditions to us *t*-tests:
(1) ormality
(2) equal variances (not essential though)

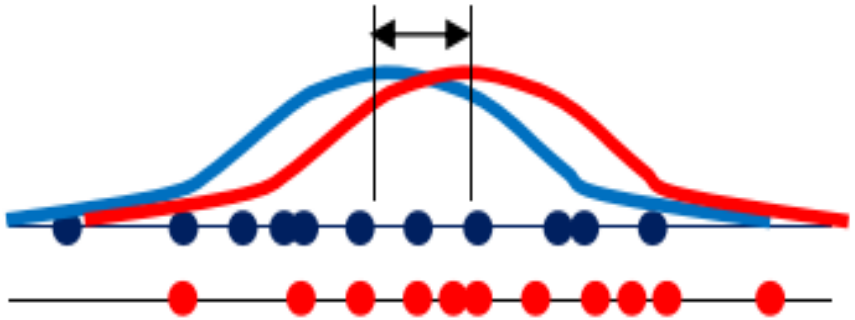
t-Test

(1) *t*-Test: Pairs two sample for means



Difference between two groups is significant ($p < 0.01$).

(2) *t*-Test: Two-sample assuming equal variances



We cannot say that there is a significant difference between two group.

paired t-test

- Step 1: $$t = \frac{m}{s/\sqrt{n}}$$

  m and **s** are the **mean** and the **standard deviation** of the difference (d), respectively. **n** is the size of d

- Step 2: read in **t test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are : df=n-1
- Step 3: If the absolute value of the **t-test statistics** (|t|) is greater than the critical value, then the difference is significant. Otherwise it isn't.

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

Calculating the mean and standard deviation of the differences gives:

$\bar{d} = 2.05$ and $s_d = 2.837$. Therefore, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$

So, we have:

$$t = \frac{2.05}{0.634} = 3.231 \qquad \text{on 19 df}$$

One-sample t-test

- Step 1: $$t = \frac{m - \mu}{s/\sqrt{n}}$$

  m and s are the mean and the standard deviation of a set of values with size n

- Step 2: read in **t test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are : df=n-1
- Step 3: If the absolute value of the **t-test statistics** (|t|) is greater than the critical value, then the difference is significant. Otherwise it isn't.

unpair t-test with equal variances

- Step 1: $t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$ $S^2 = \frac{\sum (x - m_A)^2 + \sum (y - m_B)^2}{n_A + n_B - 2}$

m_A , m_B represent the means of groups A and B

- Step 2: read in **t test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are : $df=n_A+n_B-2$
- Step 3: If the absolute value of the **t-test statistics** (|t|) is greater than the critical value, then the difference is significant. Otherwise it isn't.

unpair t-test with different variances

- Step 1:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

m_A , m_B represent the means of groups A and B

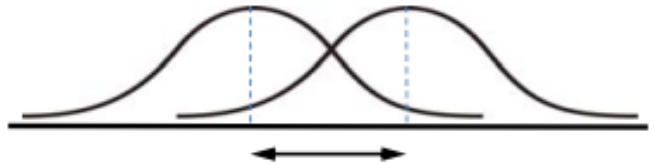
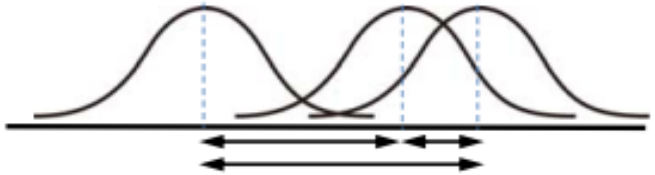
S_A , S_B are the standard deviation of the the two groups A and B

- Step 2: read in **t test table** the **critical value of Student's t distribution** corresponding to the **significance level alpha** of your choice (5%). The **degrees of freedom** (df) used in this test are :

$$df = \left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B^2}\right) / \left(\frac{S_A^4}{n_A^2(n_B - 1)} + \frac{S_B^4}{n_B^2(n_B - 1)}\right)$$

- Step 3: If the absolute value of the **t-test statistics** (|t|) is greater than the critical value, then the difference is significant. Otherwise it isn't.

ANOVA: Analysis of Variance

		2 groups	*n* groups ($n > 2$)
data distribution			
Parametric Test (normality)	unpaired (independent)	・unpaired *t*-test	ANOVA
	paired (related)	・paired *t*-test	
Non-parametric Test (no normality)	unpaired (independent)	・Mann-Whitney *U*-test	one-way data ・Kruskal-Wallis test
	paired (related)	・sign test ・Wilcoxon signed-ranks test	two-way data ・Friedman test

Why not compare groups with **multiple t-tests?**

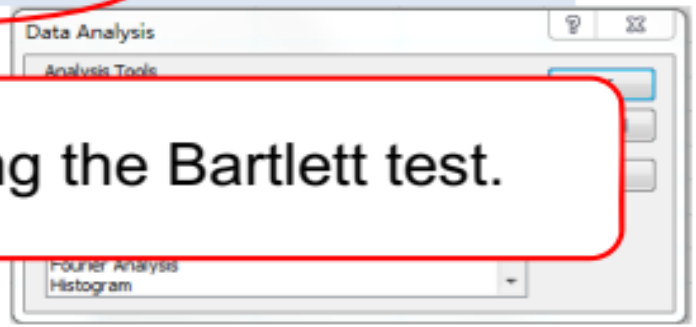
Why not compare groups with multiple t-tests?

- Every time you conduct a t-test there is a chance that you will make a Type I error. . This error is usually 5%
- When we try to compare means of three groups, A, B, and C, using the *t* test, we need to implement 3 pairwise tests, i.e., A vs B, A vs C, and B vs C.
- By running k times t-tests on the same data you will have increased your chance of "making a mistake" to $1-(0.95)^k$
- These are unacceptable errors

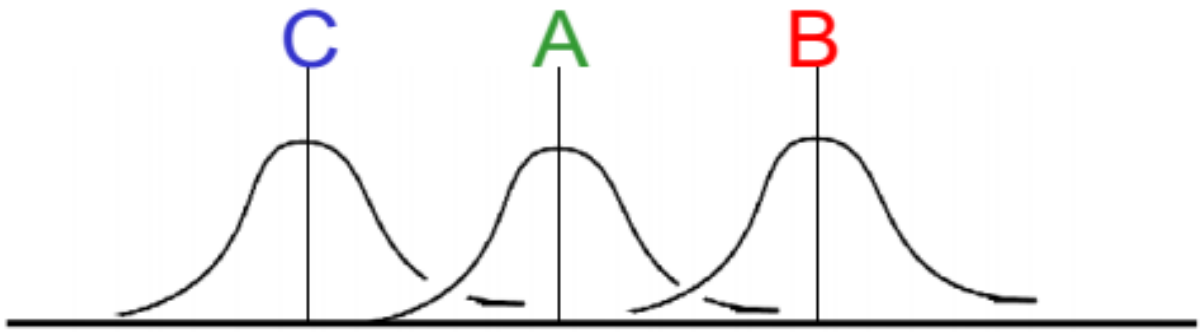
ANOVA: Analysis of Variance

1. Analysis of more than two data groups.
2. Normality and equal variance are required.

Check it using the Bartlett test.



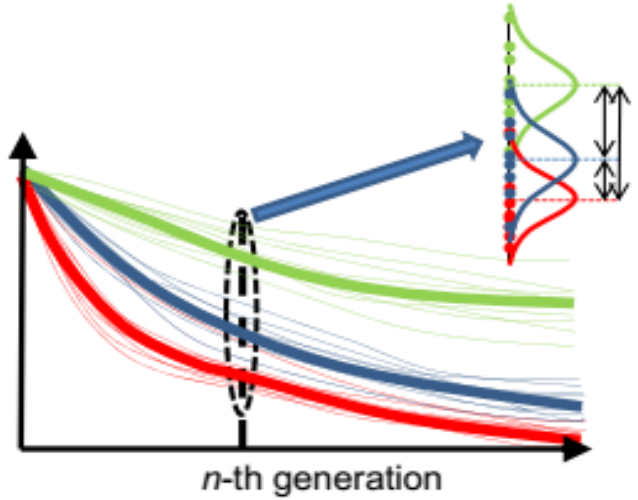
A	B	C
11.0	12.8	9.4
9.3	11.3	12.4
11.5	9.5	16.8
16.4	14.0	14.3
16.0	15.2	17.0
15.0	13.0	14.6
12.8	12.4	17.0
13.6	15.0	14.3
13.0	12.4	15.6
12.0	17.8	15.0
13.4	12.6	18.6
10.0	13.4	12.4
10.8	16.8	15.4



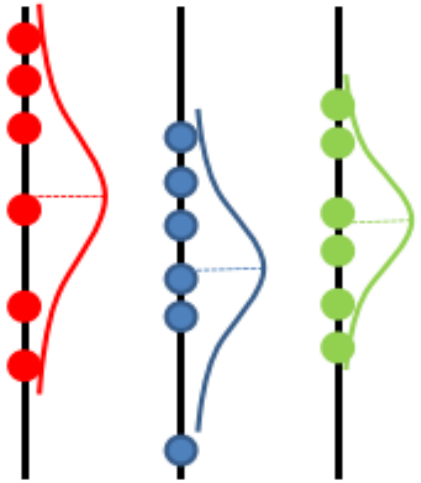
three *t*-tests $\neq$ one ANOVA

Three times of t-test with (p<0.05) equivalent one ANOVA (p<0.14). $1-(1-0.05)^3 = 0.14$

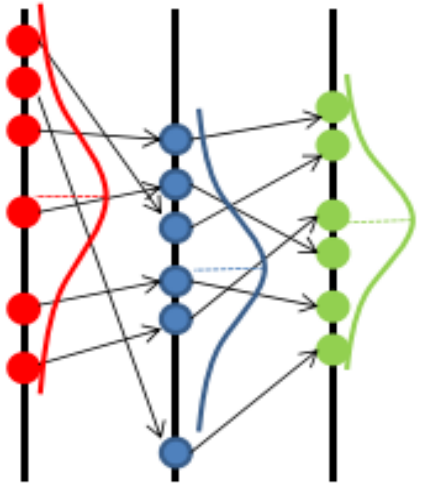
ANOVA: Analysis of Variance



When data are independent, use one-way ANOVA (single factor ANOVA).



When data correspond each other, use two-way ANOVA (two-factor ANOVA).



ANOVA: Analysis of Variance

one-factor (one-way) ANOVA

column factor

group A	group B	group C
4.23	2.51	3.04
3.21	3.3	2.89
3.63	3.75	3.55
4.42	3.22	4.39
4.08	3.99	3.86
3.98	3.65	3.5
3.75	2.62	3.6
3.22	2.93	3.21

We cannot say that three groups are significantly different. (p=0.089)

two-factor (two-way) ANOVA

column factor

sample factor

initial condition	group A	group B	group C
#1	4.23	2.51	3.04
#2	3.21	3.3	2.89
#3	3.63	3.75	3.55
#4	4.42	3.22	4.39
#5	4.08	3.99	3.86
#6	3.98	3.65	3.5
#7	3.75	2.62	3.6
#8	3.22	2.93	3.21

There are significant difference *somewhere* among three groups. (p<0.05)

ANOVA: Analysis of Variance

Output of the one-way ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.11342	2	3.05671	15.30677	3.6E-05	3.354131
Within Groups	5.39181	27	0.199697			
Total	11.50523	29				

When (p-value < 0.01 or 0.05), there is(are) significant difference *somewhere* among data groups.

Output of the two-way ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	0.755233	2	0.377617	2.755097	0.103596	3.885294
Columns	3.582272	1	3.582272	26.13631	0.000256	4.747225
Interaction	0.139411	2	0.069706	0.508573	0.613752	3.885294
Within	1.644733	12	0.137061			
Total	6.12165	17				

- Significant difference among Sample (e.g. initial conditions) cannot be found ($p > 0.05$).
- Significant difference can be found *somewhere* among Columns (e.g. three methods) ($p < 0.01$).
- We need not care an interaction effect between two factors (e.g. initial condition vs. methods) ($p > 0.05$).

Column factor

Sample factor

A	B	C
11.0	12.8	9.4
9.3	11.3	12.4
11.5	9.5	16.8
16.4	14.0	14.3
16.0	15.2	17.0
15.0	13.0	14.6
12.8	12.4	17.0
13.6	15.0	14.3
13.0	12.4	15.6
12.0	17.8	15.0
13.4	12.6	18.6
10.0	13.4	12.4
10.8	16.8	15.4

ANOVA: Analysis of Variance

Q1: Where is *significant* among A, B, and C?

A1: Apply multiple comparisons between all pairs among columns.

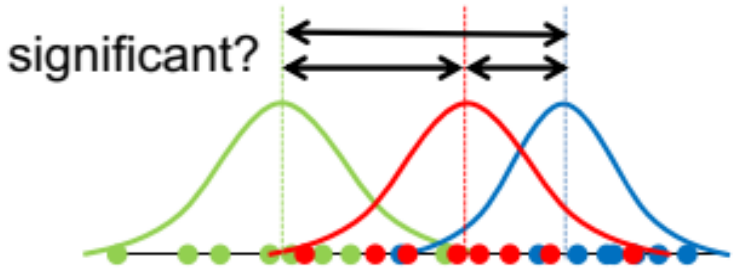
(Fisher's PLSD method, Scheffé method, Bonferroni-Dunn test, Dunnett method, Williams method, Tukey method, Nemenyi test, Tukey-Kramer method, Games/Howell method, Duncan's new multiple range test, Student-Newman-Keuls method, etc. Each has different characteristics.)

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	0.755233	2	0.377617	2.755097	0.103596	3.885294
Columns	3.582272	1	3.582272	26.13631	0.000256	4.747225
Interaction	0.139411	2	0.069706	0.508573	0.613752	3.885294
Within	1.644733	12	0.137061			
Total	6.12165	17				

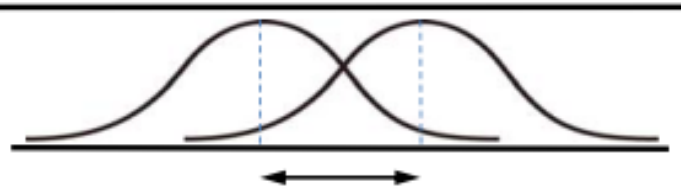
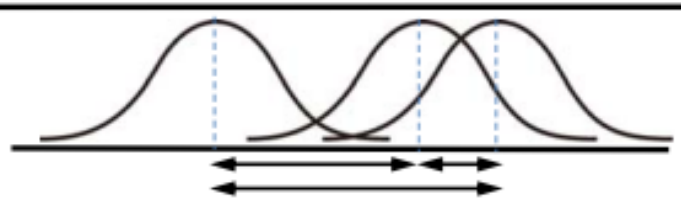
Column factor

Sample factor

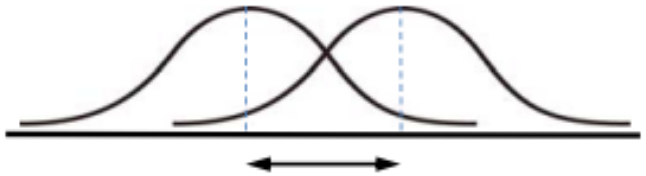
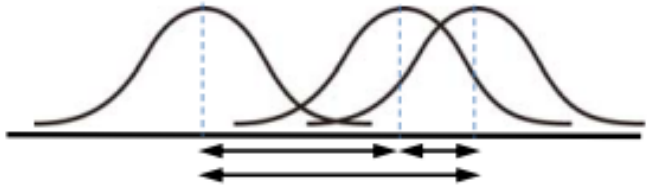
A	B	C
11.0	12.8	9.4
9.3	11.3	12.4
11.5	9.5	16.8
16.4	14.0	14.3
16.0	15.2	17.0
15.0	13.0	14.6
12.8	12.4	17.0
13.6	15.0	14.3
13.0	12.4	15.6
12.0	17.8	15.0
13.4	12.6	18.6
10.0	13.4	12.4
10.8	16.8	15.4



Non-Parametric Tests

		2 groups	*n* groups ($n > 2$)
data distribution			
Parametric Test (normality)		If normality and equal variances are not guaranteed, use non-parametric tests.	
Non-parametric Test (no normality)	unpaired (independent)	·Mann-Whitney *U*-test	one-way data: ·Kruskal-Wallis test
	paired (related)	·sign test ·Wilcoxon signed-ranks test	two-way data: ·Friedman test

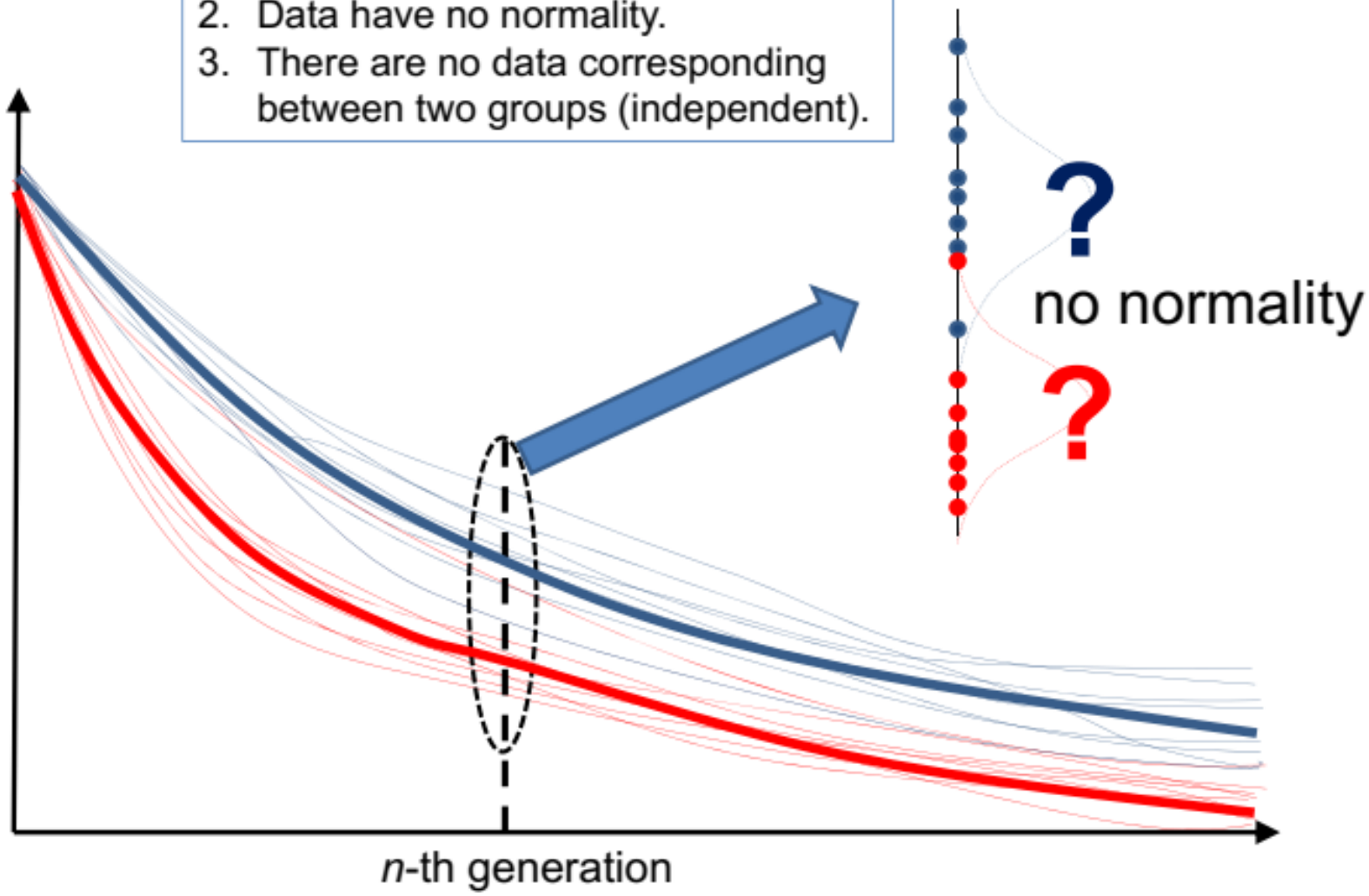
Mann-Whitney *U*-test

		2 groups	*n* groups ($n > 2$)	
data distribution				
Parametric Test (normality)	unpaired (independent)	·unpaired *t*-test	ANOVA (Analysis of Variance)	· one-way ANOVA
	paired (related)	·paired *t*-test		· two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	·Mann-Whitney *U*-test	one-way data	·Kruskal-Wallis test
	paired (related)	·sign test ·Wilcoxon signed-ranks test	two-way data	·Friedman test

Mann-Whitney *U*-test

(Wilcoxon-Mann-Whitney test, two sample Wilcoxon test)

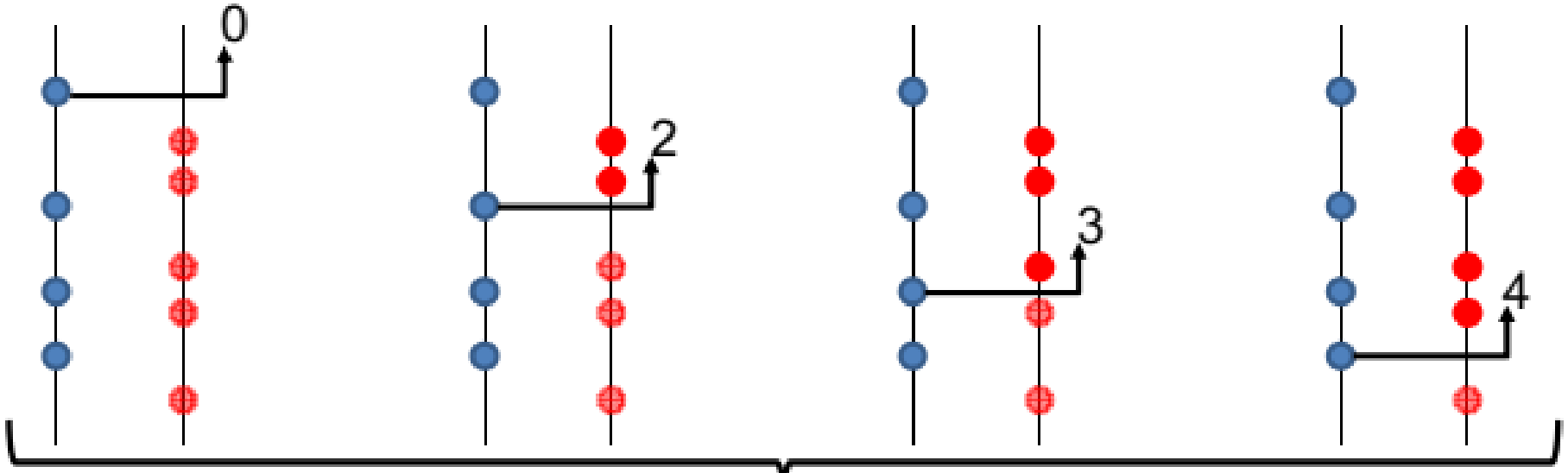
1. Comparison of two groups.
2. Data have no normality.
3. There are no data corresponding between two groups (independent).



Mann-Whitney *U*-test

(Wilcoxon-Mann-Whitney test, two sample Wilcoxon test)

1. Calculate a *U* value.



$U = 0 + 2 + 3 + 4 = 9$

$U' = 11$ $\quad (U + U' = n_1 n_2)$

(when two values are the same, count as 0.5.)

Mann-Whitney *U*-test (cont.)

(Wilcoxon-Mann-Whitney test, two sample Wilcoxon test)

2. See a *U*-test table.

- Use the smaller value of *U* or *U'*.
- When $n_1 \le 20$ and $n_2 \le 20$, see a Mann-Whitney test table. (where n_1 and n_2 are the # of data of two groups.)
- Otherwise, since *U* follows the below normal distribution roughly,

$$N\left(\mu_U, \sigma_U^2\right) = N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right)$$

normalize *U* as $z = \dfrac{U - \mu_U}{\sigma_U}$ and check a standard normal distribution table

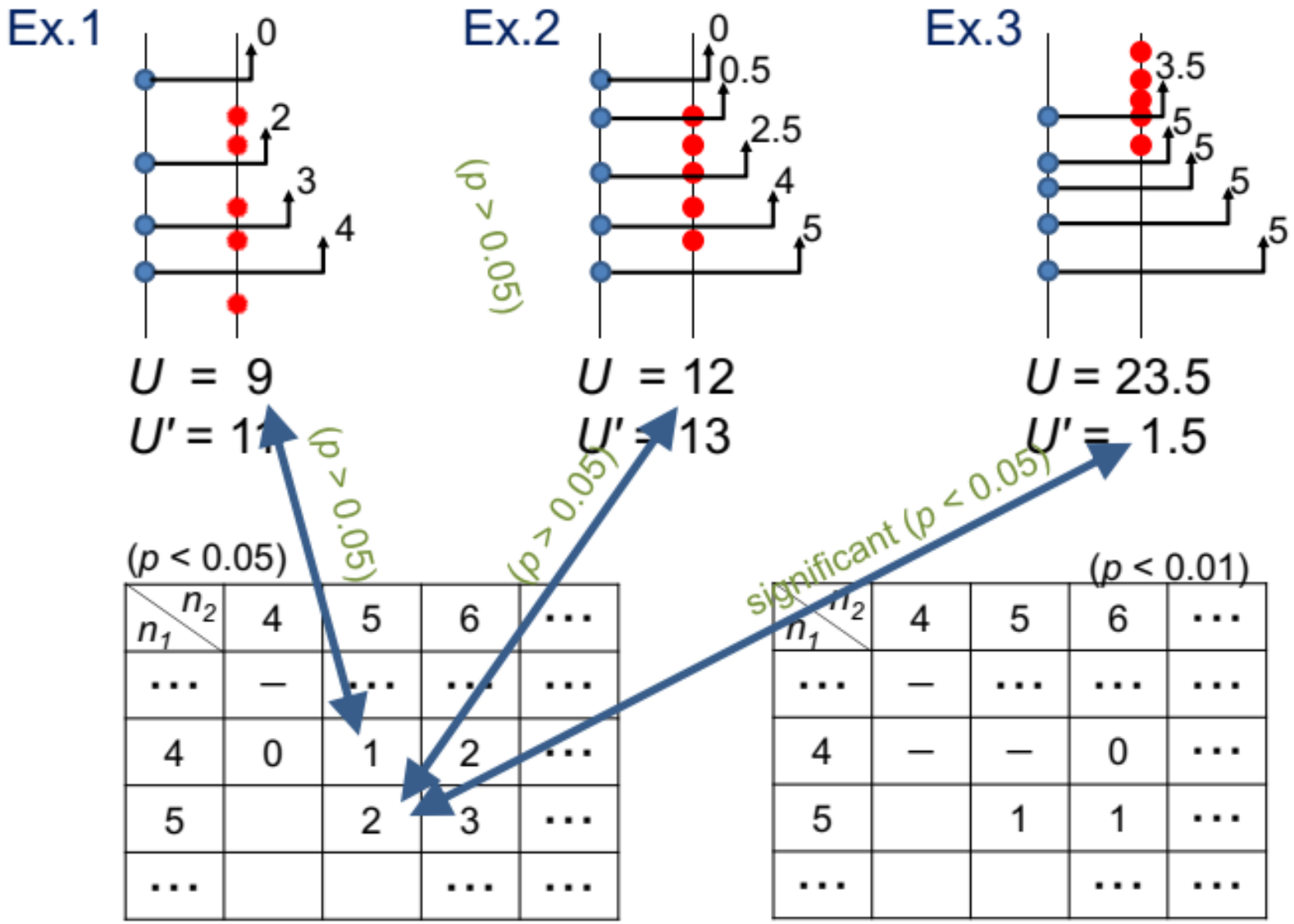
with the z, where $\mu_U = \dfrac{n_1 n_2}{2}$ and $\sigma_U = \sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

Use an Excel function to calculate the p-value for the *z*-value:

p-value = 1 - NORM.S.DIST(z)

Examples: Mann-Whitney *U*-test

(Wilcoxon-Mann-Whitney test, two sample Wilcoxon test)

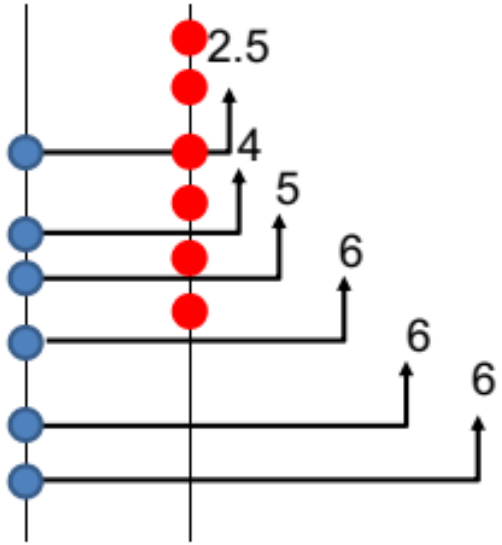


n_1 \ n_2	4	5	6	⋯
⋯	–	⋯	⋯	⋯
4	0	1	2	⋯
5		2	3	⋯
⋯			⋯	⋯

n_1 \ n_2	4	5	6	⋯
⋯	–	⋯	⋯	⋯
4	–	–	0	⋯
5		1	1	⋯
⋯			⋯	⋯

Exercise: Mann-Whitney *U*-test

(Wilcoxon-Mann-Whitney test, two sample Wilcoxon test)



$U = 29.5$

$U' = 6.5$ (Since $U' > 5$, ($p > 0.05$): significance is not found)

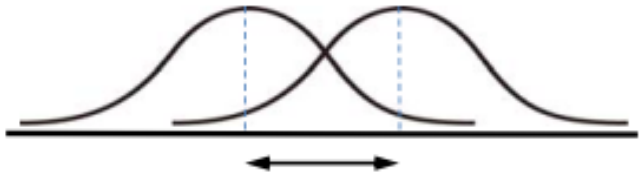
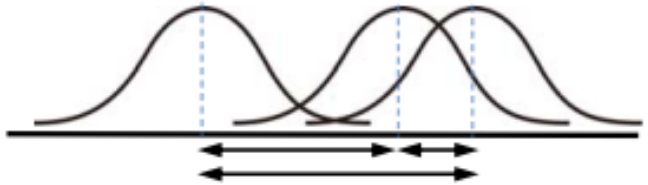
($p < 0.05$)

n_1 \ n_2	4	5	6	7
3	–	0	1	1
4	0	1	2	3
5		2	3	5
6			5	6

($p < 0.01$)

n_1 \ n_2	4	5	6	7
3	–	–	–	–
4	–	–	0	0
5		1	1	1
6			2	3

Sign Test

		2 groups	*n* groups ($n > 2$)	
data distribution				
Parametric Test (normality)	unpaired (independent)	・unpaired *t*-test	ANOVA (Analysis of Variance)	・one-way ANOVA
	paired (related)	・paired *t*-test		・two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	・Mann-Whitney *U*-test	one-way data	・Kruskal-Wallis test
	paired (related)	・sign test ・Wilcoxon signed-ranks test	two-way data	・Friedman test

Sign Test

(1) Sign Test

significance test between the # of winnings and losses

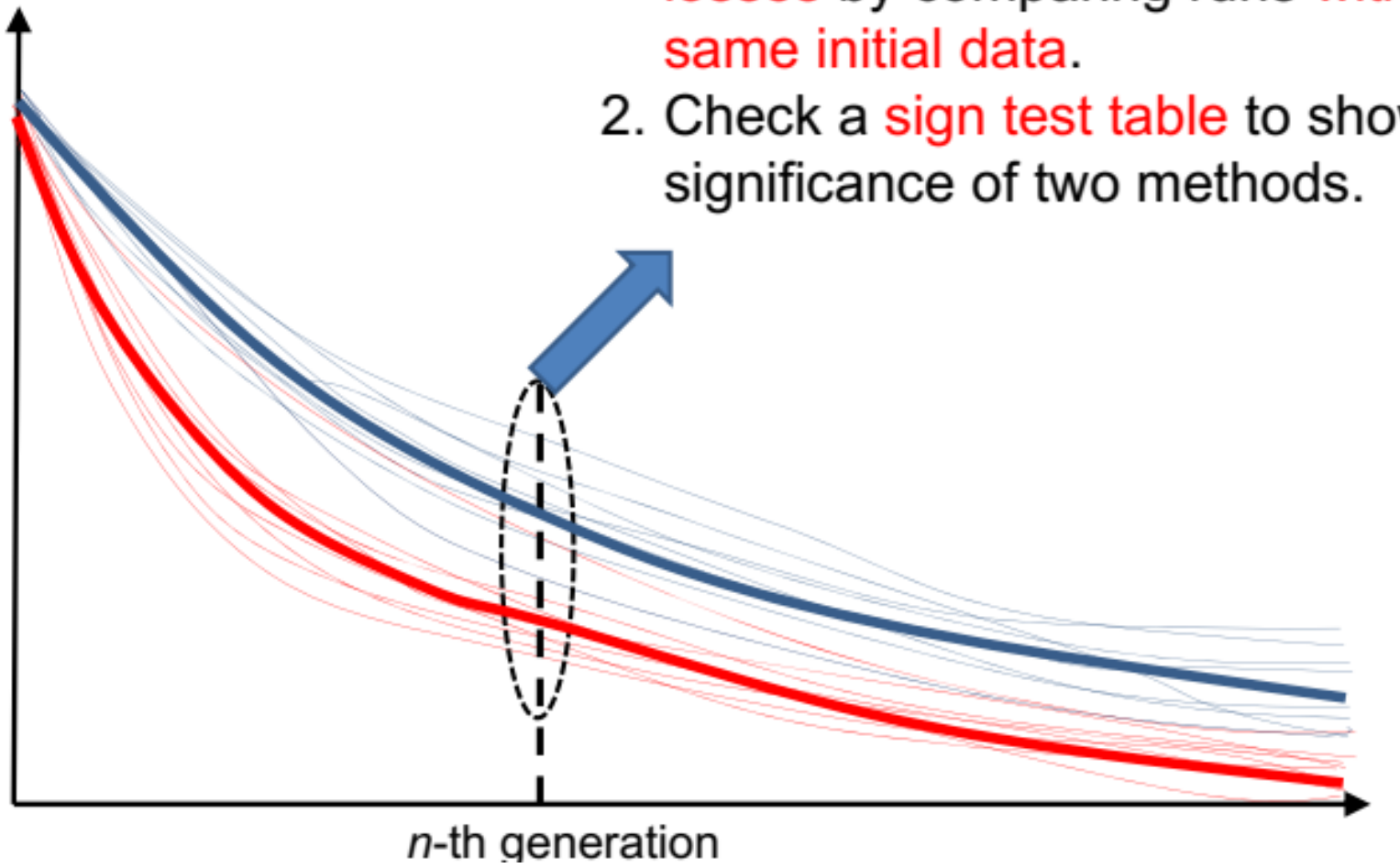
(2) Wilcoxon's Signed Ranks Test

significance test using both the # of winnings and losses and the level of winnings/losses

data of 2 groups		# of winnings and losses		the level of winnings/losses
173	174	-	+	-1
143	137	+	-	+6
158	151	+	-	+7
156	143	+	-	+13
176	180	-	+	-4
165	162	+	-	+3

Sign Test

1. Calculate the # of winnings and losses by comparing runs with the same initial data.
2. Check a sign test table to show significance of two methods.



Sign Test

Let's think about the case of $N = 17$.

To say that n_1 and n_2 are significantly different,

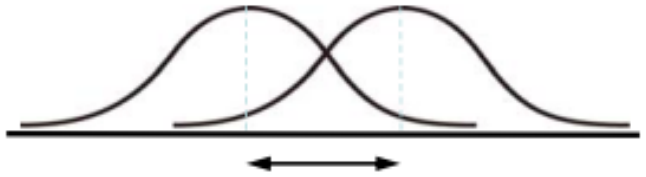
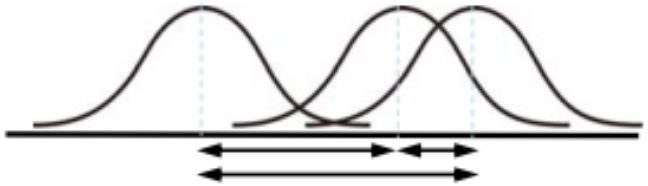
(n_1 vs. n_2) = (17 vs. 0), (16 vs. 1), or (15 vs. 2) ($p < 0.01$)

or

(n_1 vs. n_2) = (14 vs. 3) or (13 vs. 4) ($p < 0.05$)

	level of significance	
N	1%	5%
1		
2		
3		
4		
5		
6		0
7		0
8	0	0
9	0	1
10	0	1
11	0	1
12	1	2
13	1	2
14	1	2
15	2	3
16	2	3
17	2	4
18	3	4
19	3	4
20	3	5
21	4	5
22	4	5
23	4	6
24	5	6
25	5	7

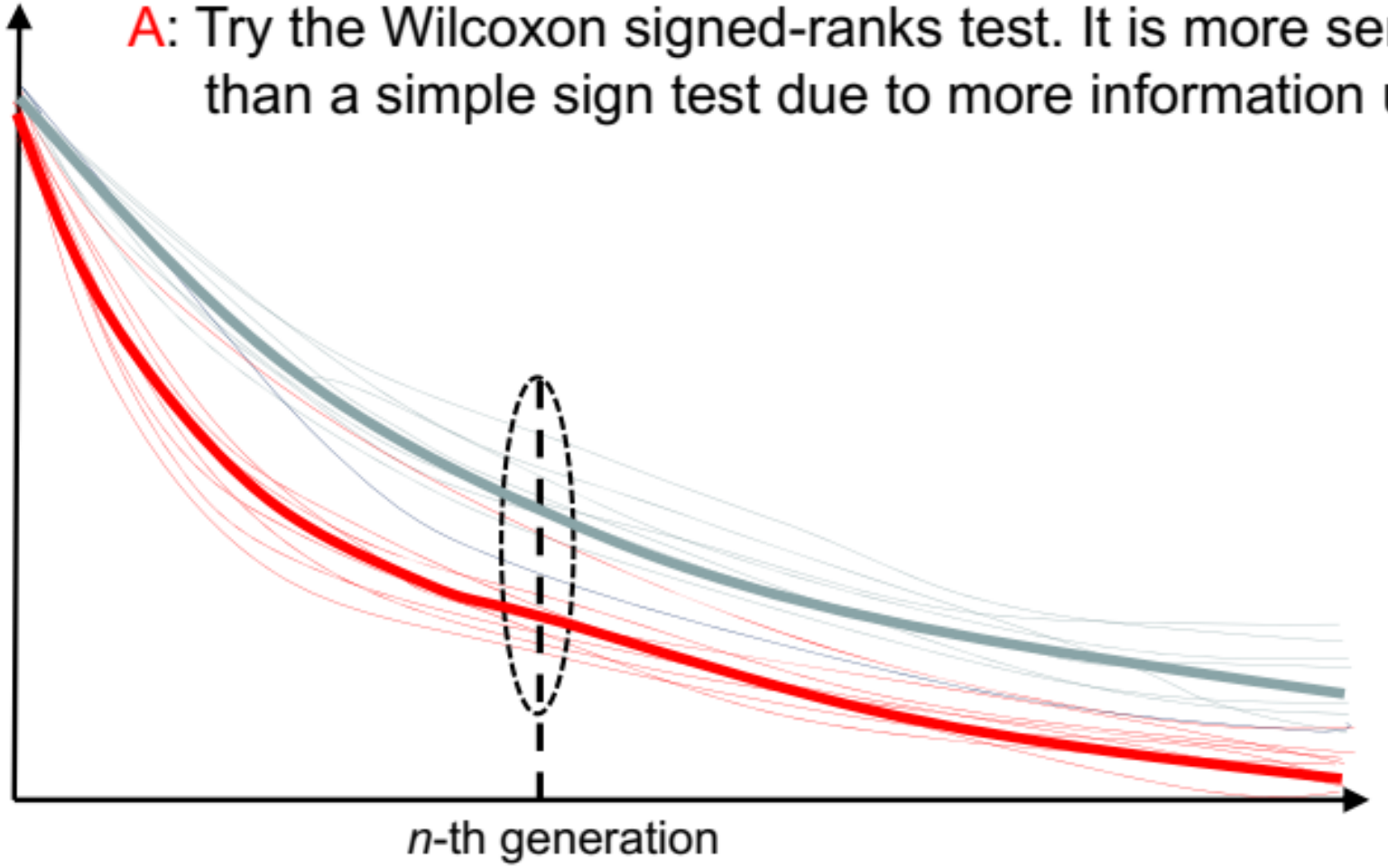
Wilcoxon Signed-Ranks Test

		2 groups	*n* groups ($n > 2$)	
data distribution				
Parametric Test (normality)	unpaired (independent)	·unpaired *t*-test	ANOVA (Analysis of Variance)	· one-way ANOVA
	paired (related)	·paired *t*-test		· two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	·Mann-Whitney *U*-test	one-way data	·Kruskal-Wallis test
	paired (related)	·sign test ·Wilcoxon signed-ranks test	two-way data	·Friedman test

Wilcoxon Signed-Ranks Test

Q: When a sign test could not show significance, how to do?

A: Try the Wilcoxon signed-ranks test. It is more sensitive than a simple sign test due to more information use.



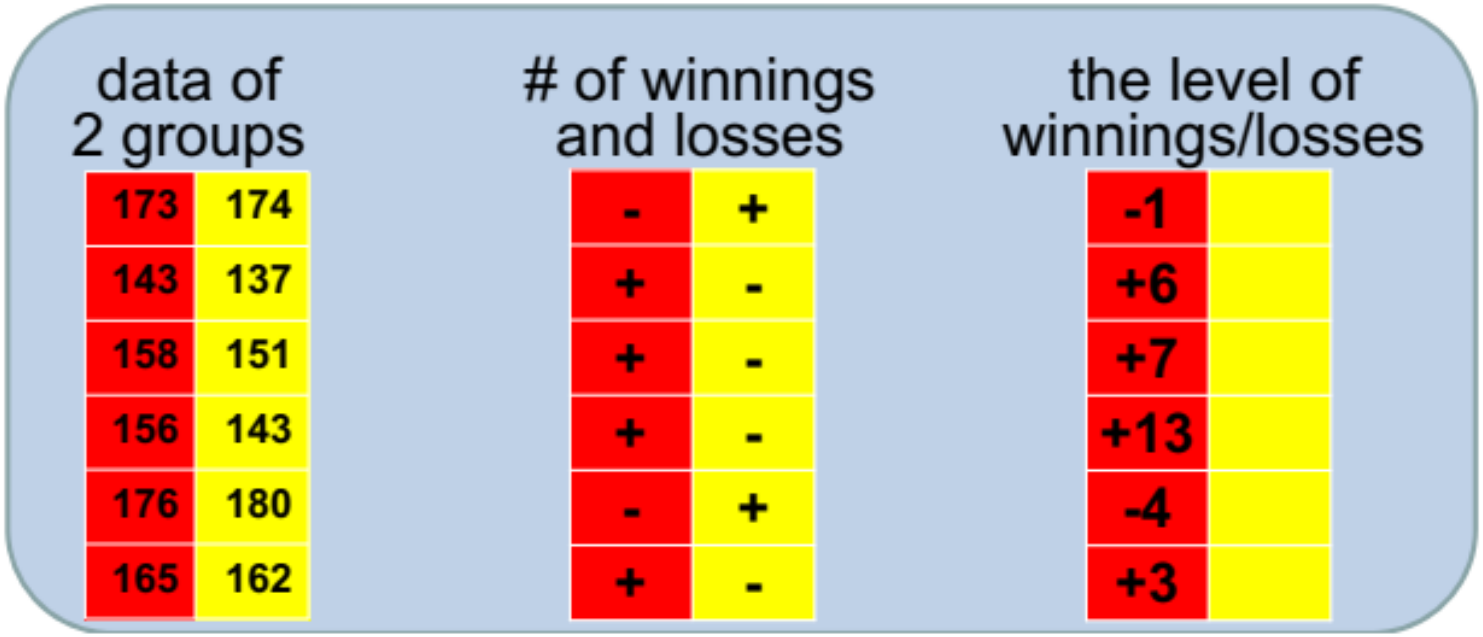
Wilcoxon Signed-Ranks Test

(1) Sign Test

significance test between the # of winnings and losses

(2) Wilcoxon's Signed Ranks Test

significance test using both the # of winnings and losses and the level of winnings/losses



Wilcoxon Signed-Ranks Test

Example:

		(step 1)	(step 2)	(step 3)	(step 4)
v (system A)	*v* (system B)	difference *d*	rank of $\|d\|$	add sign to the ranks	rank of fewer # of signs
182	163	19	7	7	
169	142	27	8	8	
172	173	−1	1	−1	1
143	137	6	4	4	
158	151	7	5	5	
156	143	13	6	6	
176	172	4	3	3	
165	168	−3	2	−2	2

$n = 8$

(step 5) $T = \sum \#of(Step4)$

$= 3$

(step 6)

Wilcoxon test table

(step 6)

$\begin{cases} n = 8 \\ T = 3 \end{cases}$

T=3 ≤ 3 (n=8, p<0.05),
then difference between systems A and B is significant.

T=3 > 0 (n=8, p<0.01),
then we cannot say there is a significant difference.

When n > 25

As T follows the below normal distribution roughly,

$$N\left(\mu_T, \sigma_T^2\right) = N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

normalize T as the below and check a standard normal distribution table with the z; see μ_T and σ_T in the above equation.

$$z = \frac{T - \mu_T}{\sigma_T}$$

Wilcoxon Test Table: significance point of T

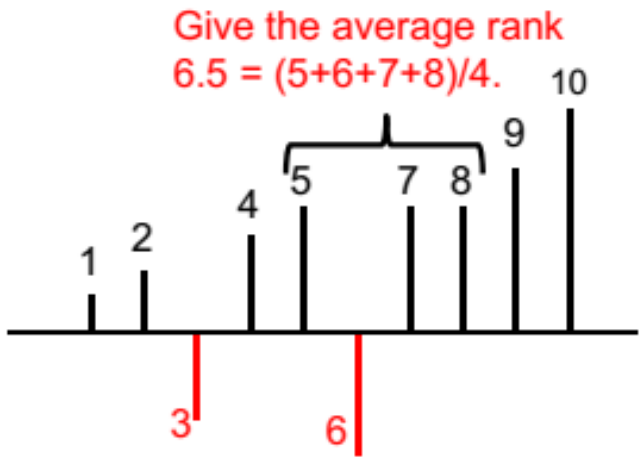
one-tail	p < 0.025	p < 0.005
two-tail	p < 0.05	p < 0.01
n = 6	0	
7	2	
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7
13	17	9
14	21	12
15	25	15
16	29	19
17	34	23
18	40	27
19	46	32
20	52	37
21	58	42
22	65	48
23	73	54
24	81	61
25	89	68

Wilcoxon Signed-Ranks Test

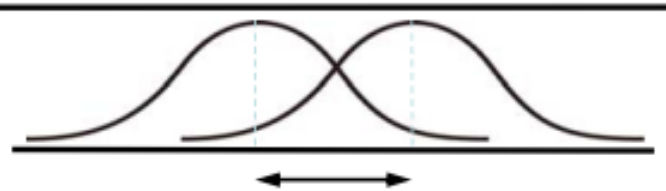
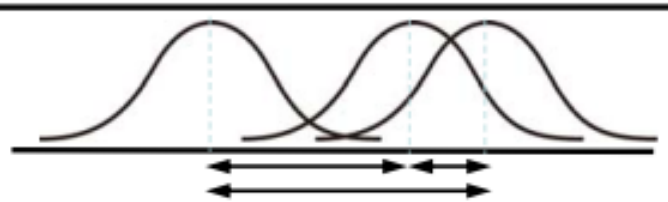
		(step 1)	(step 2)	(step 3)	(step 4)
v (system A)	*v* (system B)	difference *d*	rank of $\|d\|$	add sign to the ranks	rank of fewer # of signs
176	163	13	7 → 6.5 Tip #2	6.5	
~~142~~	~~142~~	~~0~~ Tip #1			
172	173	−1	1	−1	1
143	137	6	4	4	
158	151	7	5	5	
156	143	13	6 → 6.5 Tip #2	6.5	
176	172	4	3	3	
165	168	−3	2	−2	2

Tips:

1. When $d = 0$, ignore the data.
2. When there are the same ranks of $|d|$, give average ranks.

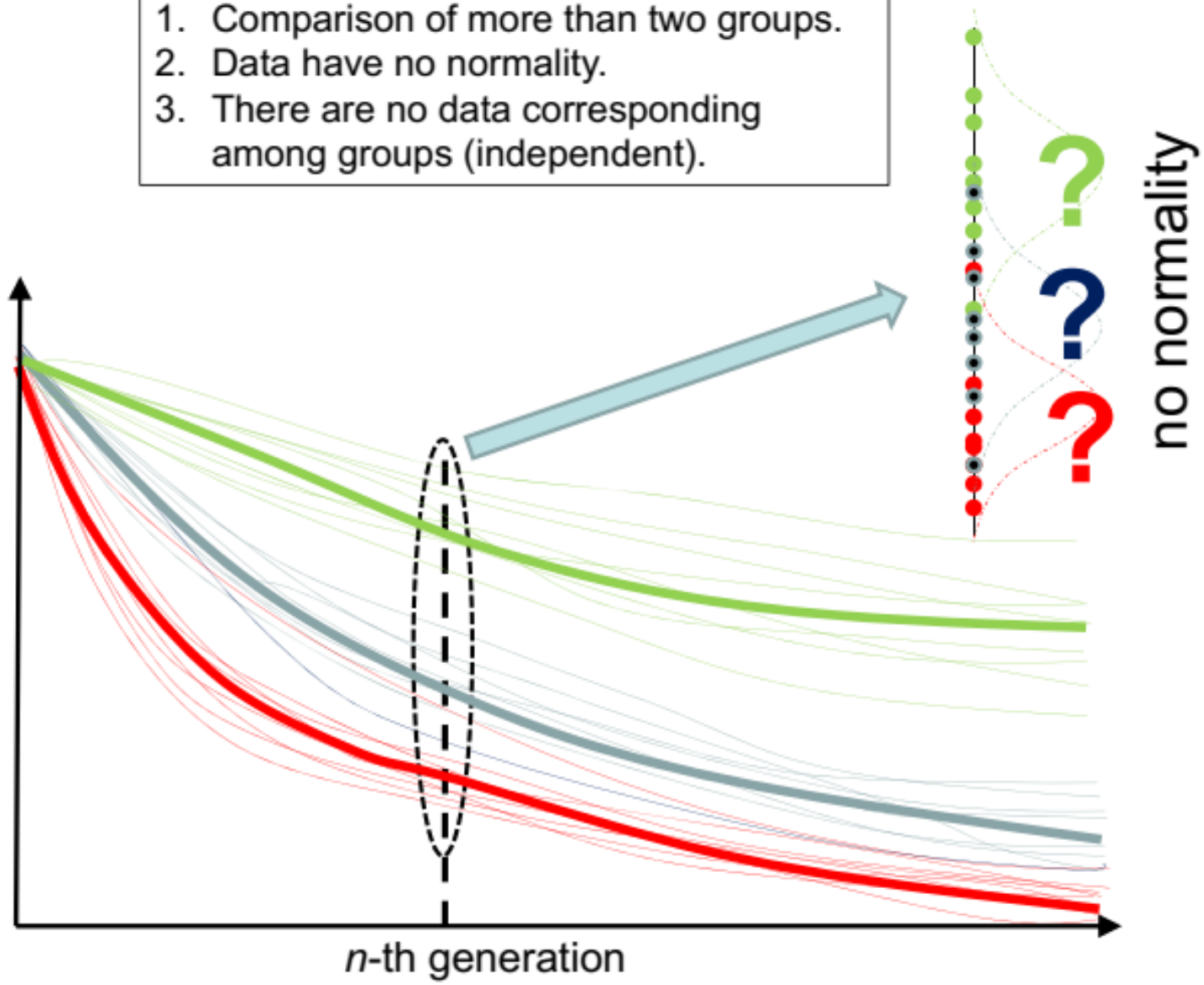


Kruskal-Wallis Test

		2 groups	*n* groups ($n > 2$)
data distribution			
Parametric Test (normality)	unpaired (independent)	·unpaired *t*-test	ANOVA (Analysis of Variance) · one-way ANOVA
	paired (related)	·paired *t*-test	· two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	·Mann-Whitney *U*-test	·Kruskal-Wallis test
	paired (related)	·sign test ·Wilcoxon signed-ranks test	two-way data ·Friedman test

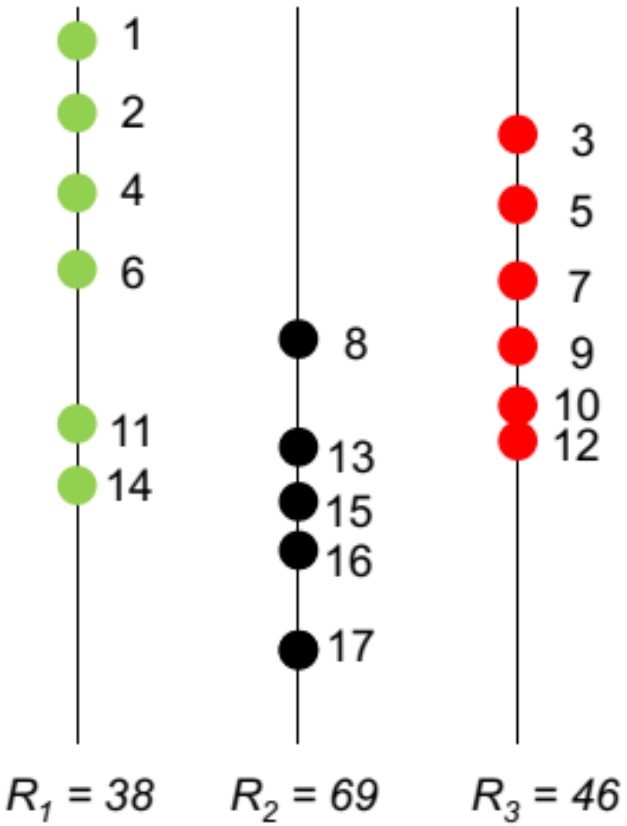
Kruskal-Wallis Test

1. Comparison of more than two groups.
2. Data have no normality.
3. There are no data corresponding among groups (independent).



Kruskal-Wallis Test

N: total # of data
k: # of groups
n_i: # of data of group *i*
R_i : sum of ranks of group *i*



How to Test

1. Rank all data.
2. Calculate *N*, *k*, n_i and R_i .
3. Calculate statistical value *H*.

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

4. If $k = 3$ and $N \le 17$, compare the *H* with a significant point in a Kruskal-Wallis test table. Otherwise, assume that *H* follows the χ^2 distribution and test the *H* using a χ^2 distribution table of (k-1) degrees of freedom

Example: Kruskal-Wallis Test

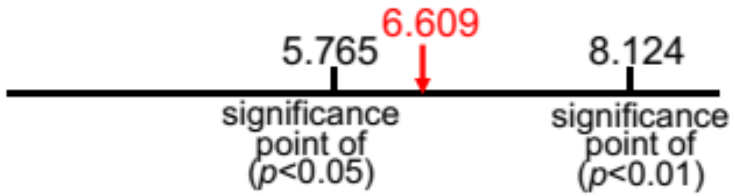
$N = n_1+n_2+n_3$ = 17 data
k = 3 groups
(n_1, n_2, n_3) = (6, 5, 6)
(R_1, R_2, R_3) = (38, 69, 46)

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

$$= \frac{12}{17(17+1)}\left(\frac{38*38}{6}+\frac{69*69}{5}+\frac{46*46}{6}\right) - 3(17+1)$$

= 6.609

Since significant points of (p<0.05) and (p<0.01) for (n_1, n_2, n_3) = (6, 5, 6) are 5.765 and 8.124, respectively, there are significant difference(s) *somewhere* among three groups (p<0.05).



Kruskal-Wallis Test Table
(for k = 3 and $N \le 17$)

n_1	n_2	n_3	p < 0.05	p < 0.01
2	2	2	-	-
2	2	3	4.714	-
2	2	4	5.333	-
2	2	5	5.160	6.533
2	2	6	5.346	6.655
2	2	7	5.143	7.000
2	2	8	5.356	6.664
2	2	9	5.260	6.897
2	2	10	5.120	6.537
2	2	11	5.164	6.766
2	2	12	5.173	6.761
2	2	13	5.199	6.792
2	3	3	5.361	-
2	3	4	5.444	6.444
2	3	5	5.251	6.909
2	3	6	5.349	6.970
2	3	7	5.357	6.839
2	3	8	5.316	7.022
2	3	9	5.340	7.006
2	3	10	5.362	7.042
2	3	11	5.374	7.094
2	3	12	5.350	7.134
2	4	4	5.455	7.036
2	4	5	5.273	7.205
2	4	6	5.340	7.340
2	4	7	5.376	7.321
2	4	8	5.393	7.350
2	4	9	5.400	7.364
2	4	10	5.345	7.357
2	4	11	5.365	7.396
2	5	5	5.339	7.339
2	5	6	5.339	7.376
2	5	7	5.393	7.450
2	5	8	5.415	7.440
2	5	9	5.396	7.447
2	5	10	5.420	7.514
2	6	6	5.410	7.467
2	6	7	5.357	7.491
2	6	8	5.404	7.522
2	6	9	5.392	7.566
2	7	7	5.398	7.491
2	7	8	5.403	7.571

n_1	n_2	n_3	p < 0.05	p < 0.01
3	3	3	5.606	7.200
3	3	4	5.791	6.746
3	3	5	6.649	7.079
3	3	6	5.615	7.410
3	3	7	5.620	7.228
3	3	8	5.617	7.350
3	3	9	5.589	7.422
3	3	10	5.588	7.372
3	3	11	5.583	7.418
3	4	4	5.599	7.144
3	4	5	5.656	7.445
3	4	6	5.610	7.500
3	4	7	5.623	7.550
3	4	8	5.623	7.585
3	4	9	5.652	7.614
3	4	10	5.661	7.617
3	5	5	5.706	7.578
3	5	6	5.602	7.591
3	5	7	5.607	7.697
3	5	8	5.614	7.706
3	5	9	5.670	7.733
3	6	6	5.625	7.725
3	6	7	5.689	7.756
3	6	8	5.678	7.796
3	7	7	5.688	7.810
4	4	4	5.692	7.654
4	4	5	5.657	7.760
4	4	6	6.681	7.795
4	4	7	5.650	7.814
4	4	8	5.779	7.853
4	4	9	5.704	7.910
4	5	5	5.666	7.823
4	5	6	5.661	7.936
4	5	7	5.733	7.931
4	5	8	5.718	7.992
4	6	6	5.724	8.000
4	6	7	5.706	8.039
5	5	5	5.780	8.000
5	5	6	5.729	8.028
5	5	7	5.708	8.108
5	6	6	5.765	8.124

Example: Kruskal-Wallis Test

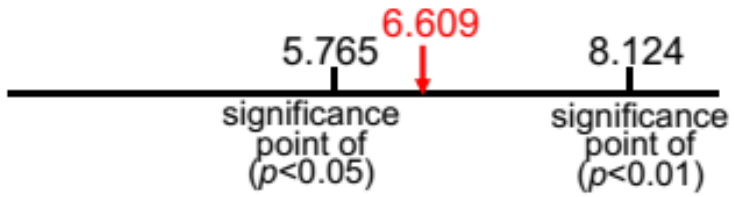
$N = n_1+n_2+n_3$ = 17 data
k = 3 groups
(n_1, n_2, n_3) = (6, 5, 6)

Q1: Where is *significant* among A, B, and C?

A1: Apply multiple comparisons between all pairs among columns.

(Fisher's PLSD method, Scheffé method, Bonferroni-Dunn test, Dunnett method, Williams method, Tukey method, Nemenyi test, Tukey-Kramer method, Games/Howell method, Duncan's new multiple range test, Student-Newman-Keuls method, etc. Each has different characteristics.)

Since sign[illegible] of (p<0.05) and (p<0.01) for (n_1, n_2[illegible] (6, 5, 6) are 5.765 and 8.124, respectiv[illegible], there are significant difference(s) *somewhere* among three groups (p<0.05).



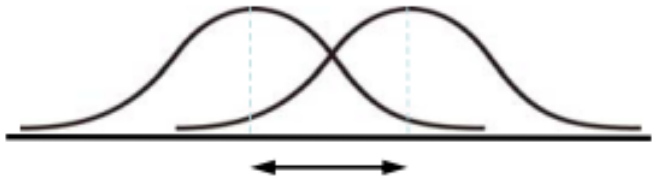
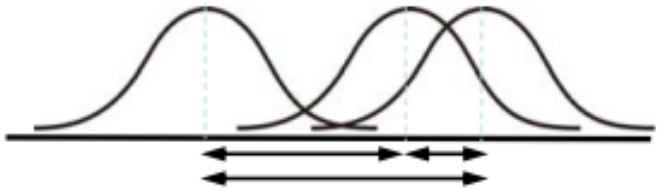
Kruskal-Wallis Test Table

(for k = 3 and $N \leq 17$)

n_1	n_2	n_3	p < 0.05	p < 0.01
2	2	2	-	-
2	2	3	4.714	-
2	2	4	5.333	-
2	2	5	5.160	6.533
2	2	6	5.346	6.655
2	2	7	5.143	7.000
2	2	8	5.356	6.664
2	4	7	5.376	7.321
2	4	8	5.393	7.350
2	4	9	5.400	7.364
2	4	10	5.345	7.357
2	4	11	5.365	7.396
2	5	5	5.339	7.339
2	5	6	5.339	7.376
2	5	7	5.393	7.450
2	5	8	5.415	7.440
2	5	9	5.396	7.447
2	5	10	5.420	7.514
2	6	6	5.410	7.467
2	6	7	5.357	7.491
2	6	8	5.404	7.522
2	6	9	5.392	7.566
2	7	7	5.398	7.491
2	7	8	5.403	7.571

n_1	n_2	n_3	p < 0.05	p < 0.01
3	3	3	5.606	7.200
3	3	4	5.791	6.746
3	3	5	6.649	7.079
3	3	6	5.615	7.410
3	3	7	5.620	7.228
3	3	8	5.617	7.350
3	3	9	5.589	7.422
3	3	10	5.588	7.372
3	3	11	5.583	7.418
3	4	4	5.599	7.144
3	4	5	5.656	7.445
3	4	6	5.610	7.500
3	4	7	5.623	7.550
3	4	8	5.623	7.585
3	4	9	5.652	7.614
3	4	10	5.661	7.617
3	5	5	5.706	7.578
3	5	6	5.602	7.591
3	5	7	5.607	7.697
3	5	8	5.614	7.706
3	5	9	5.670	7.733
3	6	6	5.625	7.725
3	6	7	5.689	7.756
3	6	8	5.678	7.796
3	7	7	5.688	7.810
4	4	4	5.692	7.654
4	4	5	5.657	7.760
4	4	6	6.681	7.795
4	4	7	5.650	7.814
4	4	8	5.779	7.853
4	4	9	5.704	7.910
4	5	5	5.666	7.823
4	5	6	5.661	7.936
4	5	7	5.733	7.931
4	5	8	5.718	7.992
4	6	6	5.724	8.000
4	6	7	5.706	8.039
5	5	5	5.780	8.000
5	5	6	5.729	8.028
5	5	7	5.708	8.108
5	6	6	5.765	8.124

Friedman Test

		2 groups	*n* groups ($n > 2$)
data distribution			
Parametric Test (normality)	unpaired (independent)	▪ unpaired *t*-test	ANOVA (Analysis of Variance) ▪ one-way ANOVA
	paired (related)	▪ paired *t*-test	▪ two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	▪ Mann-Whitney *U*-test	one-way data ▪ Kruskal-Wallis test
	paired (related)	▪ sign test ▪ Wilcoxon signed-ranks test	▪ Friedman test

Friedman Test

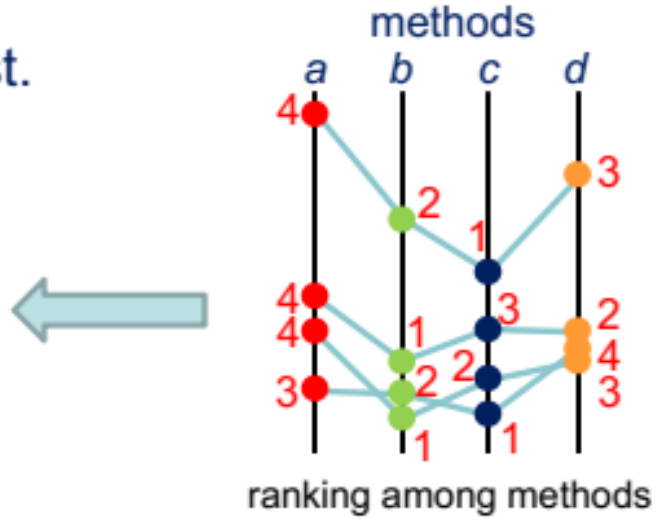
Step 1: Make a ranking table.

Step 2: Sum ranks of the factor that you want to test.

benchmark tasks	method			
	a	b	c	d
A	4	2	1	3
B	3	2	1	4
C	4	1	2	3
D	4	1	3	2
Σ	15	6	7	12

# of data (n = 4)

# of methods (k = 4)



Step 3: Calculate the Friedman test value, χ^2_r.

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

where (k, n) are the # of levels of factors 1 and 2.

Step 4: If k =3 or 4, compare χ^2_r with a significant point in a Friedman test table.
Otherwise, use a χ^2 table of (k-1) degrees of freedom.

Example: Friedman Test

Step 1: Make a ranking table.

Step 2: Sum ranks of the factor that you want to test.

benchmark tasks	method			
	a	b	c	d
A	4	2	1	3
B	3	2	1	4
C	4	1	2	3
D	4	1	3	2
Σ	15	6	7	12

# of data (n = 4)

# of methods (k = 4)

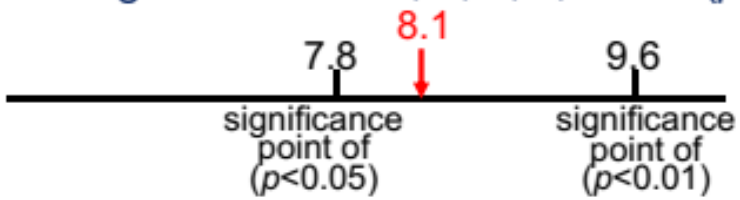
Step 3: Calculate the Friedman test value, χ^2_r.

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

$$= \frac{12}{4*4*5}\left(15^2 + 6^2 + 7^2 + 12^2 + \right) - 3*4*5$$

$$= 8.1$$

Step 4: Since significant point for (k,n) = (4,4) is7.80, there is/are significant difference(s) *somewhere* among four methods, *a*, *b*, *c*, and *d* (p<0.05).



Friedman test table.

k	n	p<0.05	p<0.01
3	3	6.00	–
	4	6.50	8.00
	5	6.40	8.40
	6	7.00	9.00
	7	7.14	8.86
	8	6.25	9.00
	9	6.22	9.56
	∞	5.99	9.21
4	3	7.40	9.00
	4	7.80	9.60
	5	7.80	9.96
	∞	7.81	11.34

Example: Friedman Test

Step 1: Make a ranking table.

Step 2: Sum ranks of the factor that you want to test.

Q1: Where is *significant* among *a*, *b*, *c*, or *d*?

A1: Apply multiple comparisons between all pairs among columns.

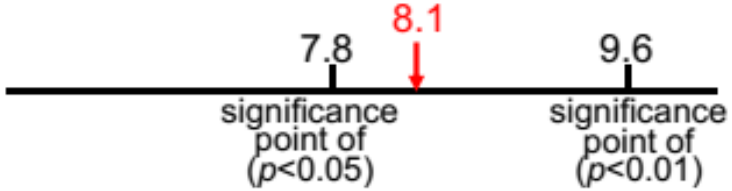
(Fisher's PLSD method, Scheffé method, Bonferroni-Dunn test, Dunnett method, Williams method, Tukey method, Nemenyi test, Tukey-Kramer method, Games/Howell method, Duncan's new multiple range test, Student-Newman-Keuls method, etc. Each has different characteristics.)

Ste

$$= \frac{12}{4*4*5}\left(15^2 + 6^2 + 7^2 + 12^2 + \right) - 3*$$

$$= 8.1$$

Step 4: Since significant point for (*k*,*n*) = (4,4) is7 0, there is/are significant difference(s) *somewhere* among four methods, *a*, *b*, *c*, and *d* (*p*<0.05).

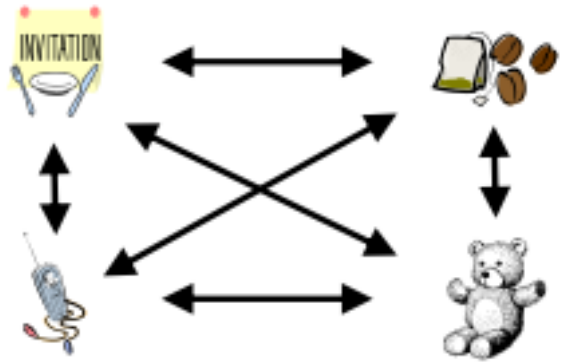


est table.

		0.05	*p*<0.01
		.00	—
		6.50	8.00
3	5	6.40	8.40
	6	7.00	9.00
	7	7.14	8.86
	8	6.25	9.00
	9	6.22	9.56
	∞	5.99	9.21
4	3	7.40	9.00
	4	7.80	9.60
	5	7.80	9.96
	∞	7.81	11.34

Multiple Comparisons

When there is significant difference among groups, multiple comparison is used to know which group is significantly difference from others.



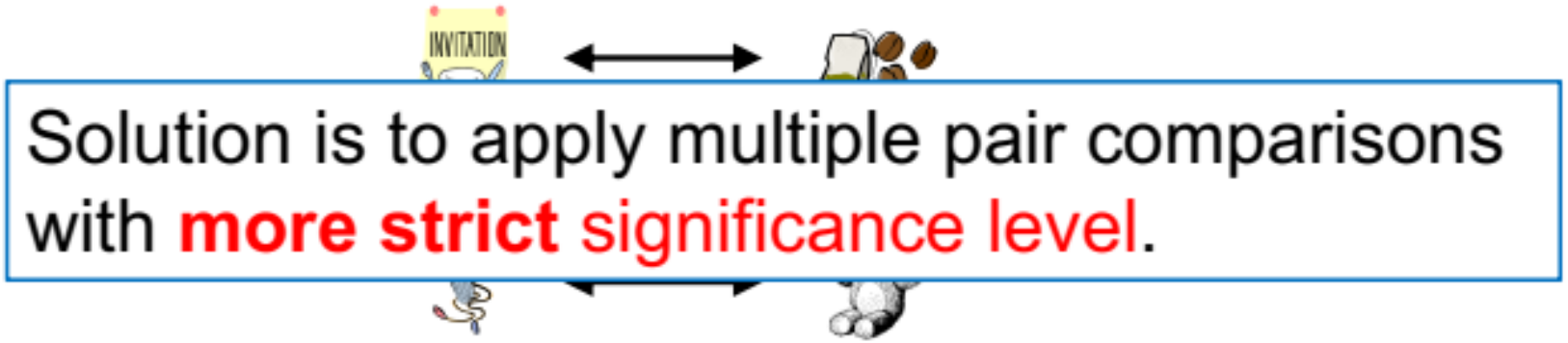
Example $_4C_2 = 6$ times of pair comparisons with ($p < 0.05$) simply.

Don't apply multiple pair comparisons

$1 - (1 - 0.05)^6$ = significance level 26.5%!

Multiple Comparisons

When there is significant difference among groups, multiple comparison is used to know which group is significantly difference from others.



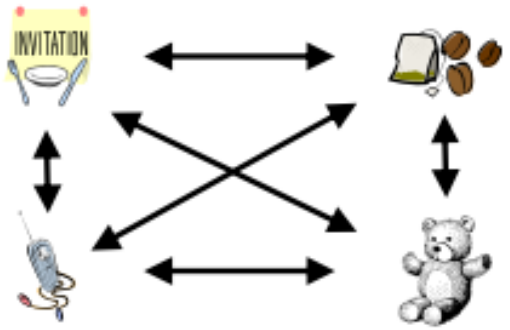
Solution is to apply multiple pair comparisons with **more strict** significance level.

Example $_4C_2 = 6$ times of pair comparisons with ($p < 0.05$)

$1 - (1 - 0.05)^6$ = significance level 26.5%!

Multiple Comparisons
-- Bobferroni method --

When pair comparisons are applied m times, let's use a significance level of p / m.



${}_4C_2 = 6$ times of pair comparisons with $(p < \frac{0.05}{6})$

Features:
(1) Simple.
(2) Rather strict, i.e. showing significances is rather hard.

Multiple Comparisons
-- Holm method --

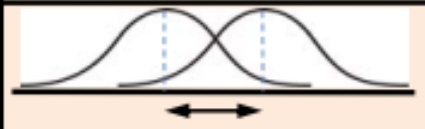
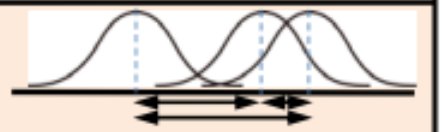
Corrected Bonferroni method to detect significances easily.

Example

pair comparisons	*p*-value	corrected *p*-value eqn.	corrected *p*-value
 vs. 	0.0076	= *p*-value* 6	0.0456
 vs. 	0.0095	= *p*-value* 5	0.0475
 vs. 	0.0280	= *p*-value* 4	0.1120
 vs. 	0.0320	= *p*-value* 3	0.0960
 vs. 	0.0380	= *p*-value* 2	0.0760
 vs. 	0.0410	= *p*-value* 1	0.0410

SUMMARY

1. We overview which statistical test we should use for which case.

		2 groups	*n* groups ($n > 2$)
data distribution			
Parametric Test (normality)	unpaired (independent)	・unpaired *t*-test	ANOVA (Analysis of Variance) ・one-way ANOVA
	paired (related)	・paired *t*-test	・two-way ANOVA
Non-parametric Test (no normality)	unpaired (independent)	・Mann-Whitney *U*-test	one-way data ・Kruskal-Wallis test
	paired (related)	・sign test ・Wilcoxon signed-ranks test	two-way data ・Friedman test

+

Scheffé's method of paired comparison for *Human Subjective Tests*

2. We can appeal the effectiveness of our experiments with correct use of statistical tests.


Thank
you