




On the Analysis of Paired Quality of Life Data: Was there a Change?

*Conference on Quality of Life Research in Asia
May 2005*

Daniel YT Fong

Department of Nursing Studies
Li Ka Shing Faculty of Medicine
The University of Hong Kong



Road Map



- ❖ The Situations
- ❖ The Problem
- ❖ A Solution
- ❖ The Conclusions



When Do Paired QoL Data Arise : Situation 1

- ❖ A clinical trial to examine the effects of an intervention in 100 patients with Type 2 diabetes mellitus
- ❖ Each patient had Beck Depression Inventory (BDI) measured at baseline and 18 weeks after treatment.
- ❖ BDI has 21 items with total score ranging from 0 (none) to 63 (severe).

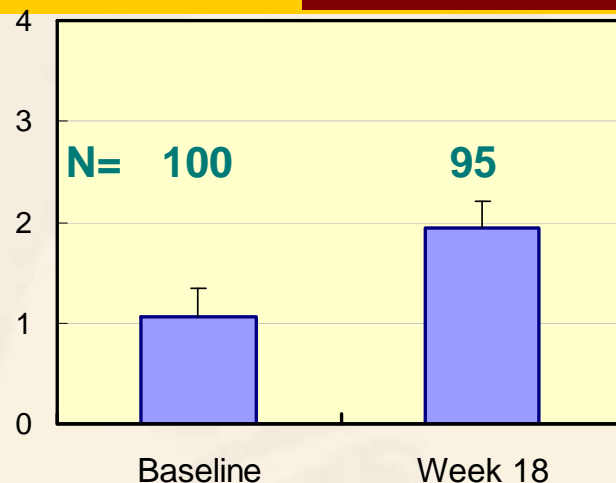
Was there a change of BDI score?



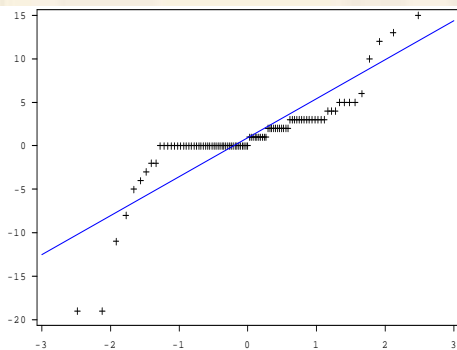
Was there a change of BDI score?

Seems yes !

18 weeks after treatment

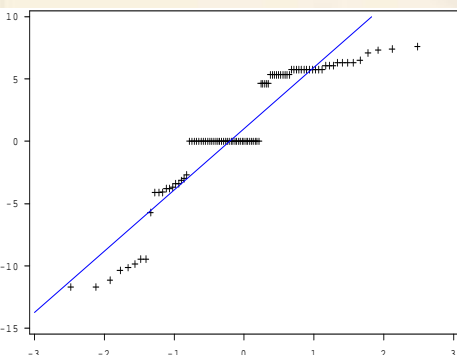


QQ plots



	n	Mean (SD)
Week 18 - Baseline	95	0.9 (4.5)

	p-value
Paired t-test	0.044
Signed rank test	<0.001
Sign test	<0.001



With logit transformation	n	Mean (SD)
Week 18 - Baseline	95	1.0 (4.9)

With logit transformation	p-value
Paired t-test	<0.001
Signed rank test	<0.001
Sign test	<0.001



When Do Paired QoL Data Arise : Situation 2

- ❖ A clinical trial to examine the effects of a single daily dose of fexcofenadine 120mg on the relief of symptoms in patients suffering from allergic rhinitis for at least 2 years
- ❖ **Symptom scores**, each rated on a 5-point Likert scale
(0=Absent; 1=Mild; 2=Moderate; 3=Severe; 4=Very severe)
, were measured at baseline as well as 1 week and 2 weeks after treatment

Was there a change of symptom score?

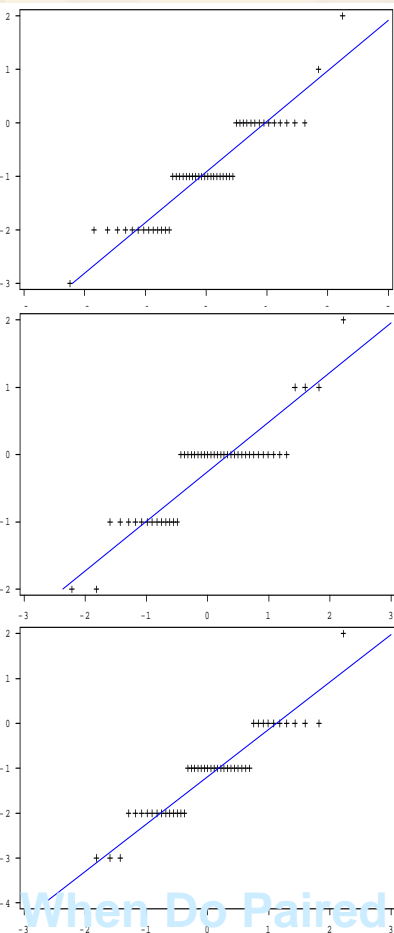


Was there a change of symptom score?

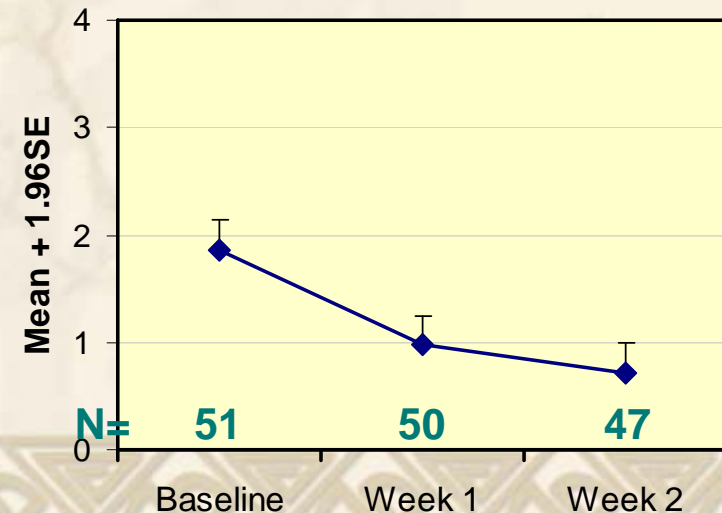
Seems yes, at each week !

Symptom Score (Eye)

QQ plots




	n	Mean (SD)	p-value		
			Paired t-test	Signed rank test	Sign test
Week 1 – Week 0	50	-0.9 (0.9)	<0.001	<0.001	<0.001
Week 2 – Week 1	47	-0.3 (0.7)	0.022	0.034	0.019
Week 2 – Week 0	47	-1.2 (1.1)	<0.001	<0.001	<0.001



When Do Paired QoL Data Arise : Situation 2

When Do Paired QoL Data Arise : Other Situations

- 
1. Test-retest reliability
 2. Criterion/Concurrent validity
 3. Inter-rater difference
e.g. the questionnaire is not self-administered
 4. Difference between different modes of administration
e.g. self-completion vs telephone interview
 5. Any studies with matching



**So,
What is the Problem?**

A Common Characteristic of Virtually ALL QoL Measures

PHYSICAL WELL-BEING

	Not at all	A little bit	Somewhat	Quite a bit	Very much
GP1 I have a lack of energy.....	0	1	2	3	4
GP2 I have nausea.....	0	1	2	3	4
GP3 Because of my physical condition, I have trouble meeting the needs of my family.....	0	1	2	3	4
GP4 I have pain.....	0	1	2	3	4
GP5 I am bothered by side effects of treatment.....	0	1	2	3	4
GP6 I feel ill.....	0	1	2	3	4
GP7 I am forced to spend time in bed.....	0	1	2	3	4

SOCIAL/FAMILY WELL-BEING

	Not at all	A little bit	Somewhat	Quite a bit	Very much
GS1 I feel close to my friends.....	0	1	2	3	4
GS2 I get emotional support from my family.....	0	1	2	3	4
GS3 I get support from my friends.....	0	1	2	3	4
GS4 My family has accepted my illness.....	0	1	2	3	4
GS5 I am satisfied with family communication about my illness.....	0	1	2	3	4
GS6 I feel close to my partner (or the person who is my main support).....	0	1	2	3	4
Q1 <i>Regardless of your current level of sexual activity, please answer the following question. If you prefer not to answer it, please check this box <input type="checkbox"/> and go to the next section.</i>					
GS7 I am satisfied with my sex life.....	0	1	2	3	4

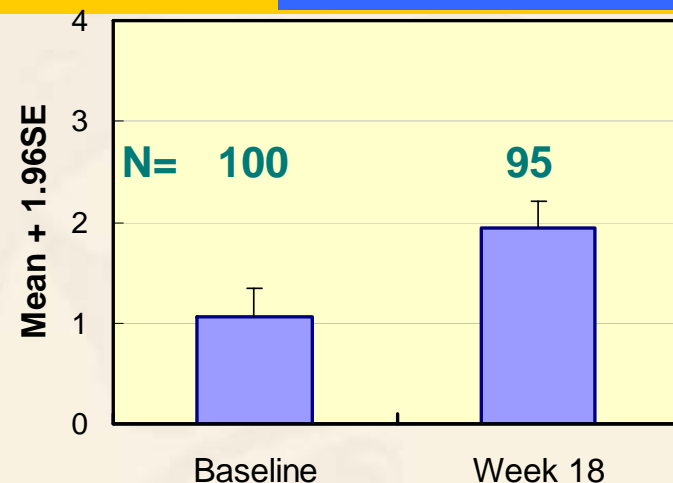
- ❖ Most QoL scores are calculated as a total or a weighted total of a set of item responses on a categorical scale
- ❖ Therefore, virtually all QoL measures are **discrete**

Was there a change of BDI score?

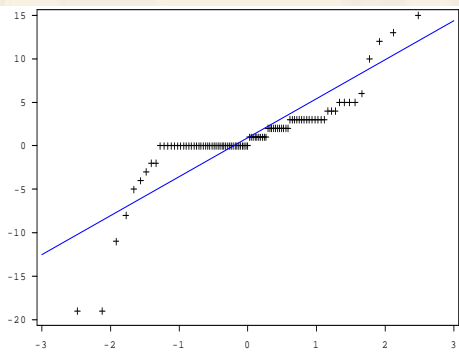
No !



18 weeks after treatment



QQ plots



	n	Mean (SD)
Week 18 - Baseline	95	0.9 (4.5)

?? ? !!!

- ❖ Only took 18 different values
- ❖ Indeed, **Median = 0 !**

Week 18 - Baseline	n	%
-19	2	2.11
-11	1	1.05
-8	1	1.05
-5	1	1.05
-4	1	1.05
-3	1	1.05
-2	2	2.11
0	39	41.05
1	10	10.53
2	11	11.58
3	14	14.74
4	3	3.16
5	4	4.21
6	1	1.05
10	1	1.05
12	1	1.05
13	1	1.05
15	1	1.05

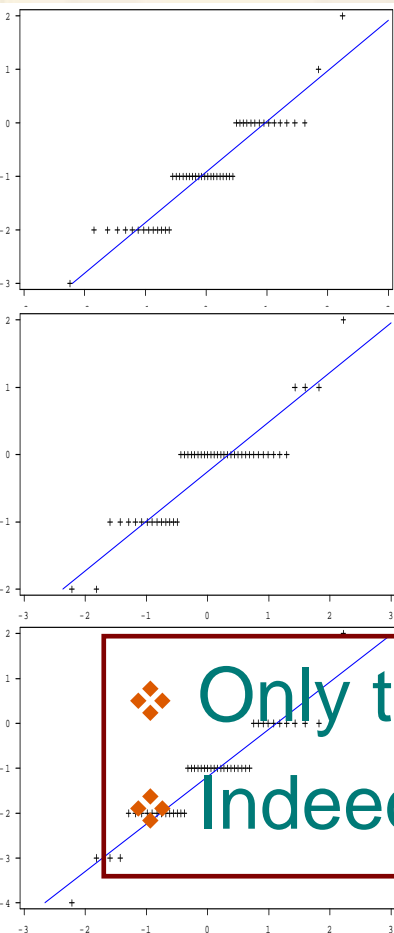
Was there a change of symptom score?

Not in the 2nd week !

Symptom Score (Eye)

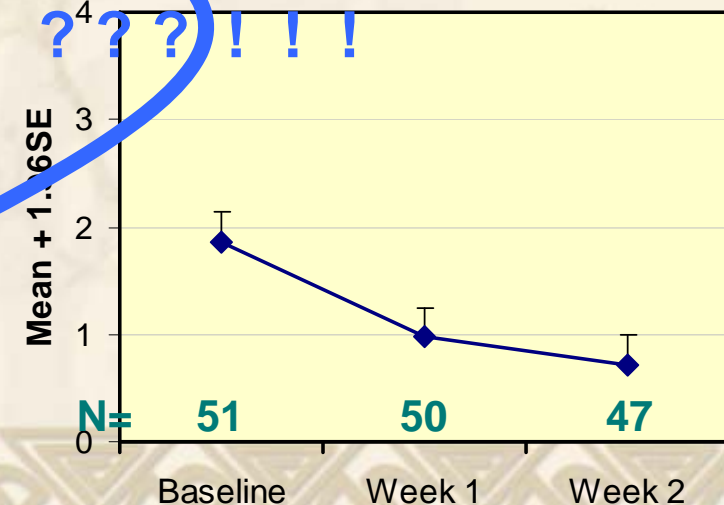


QQ plots



			p-value		
	n	Mean (SD)	Paired t-test	Signed rank test	Sign test
Week 1 – Week 0	50	-0.9 (0.9)	<0.001	<0.001	<0.001
Week 2 – Week 1	47	-0.3 (0.7)	0.022	0.034	0.019

Week 2 – Week 1	n	%
-2	2	4.26
-1	13	27.66
0	28	59.57
1	3	6.38
2	1	2.13

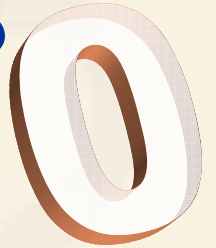


❖ Only took 5 different values

❖ Indeed, **Median = 0 !**

QoL Data Arise : Situation 2

So, What is the Problem?



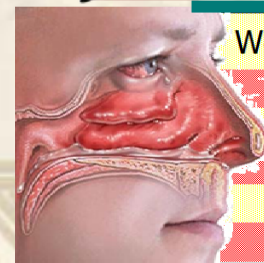
Ties

- ❖ The Problem is due to the zero differences (ties)
- ❖ All the 3 tests: the paired t-test, the signed rank test and the sign test, when first developed, were not prepared for ties

Week 18 - Baseline	n	%
-19	2	2.11
-11	1	1.05
-8	1	1.05
-5	1	1.05
-4	1	1.05
-3	1	1.05
-2	2	2.11
0	39	41.05
1	10	10.53
2	11	11.58
3	14	14.74
4	3	3.16
5	4	4.21
6	1	1.05
10	1	1.05
12	1	1.05
13	1	1.05
15	1	1.05



Week 2 - Week 1	n	%
-2	2	4.26
-1	13	27.66
0	28	59.57
1	3	6.38
2	1	2.13



Anything Wrong with Our Methods?



1. Paired t-test

- ∞ Tests about the mean
- ∞ Requires normality though may survive in moderate departure (Swinscor & Campbell, 2002)


2. Signed rank test

- ∞ Tests about the median
- ∞ Requires symmetry
- ∞ Ties handling procedures rely on asymptotic assumption (Lehmann, 1975) but are implemented in software like SAS

3. Sign test

- ∞ Tests about directional symmetry, i.e. same proportion of +ve differences and -ve differences (Randles, 2001)
- ∞ No distributional requirements
- ∞ Ties are discarded

Do we have a Solution?

- 
- ❖ Yes !
 - ❖ Two modifications of the sign test for testing about the median
 1. The modified sign test
 - ❖ Simple to implement
 2. The likelihood ratio sign test
 - ❖ Computationally more intensive but more powerful
 - ❖ They are not implemented in any marketed software!

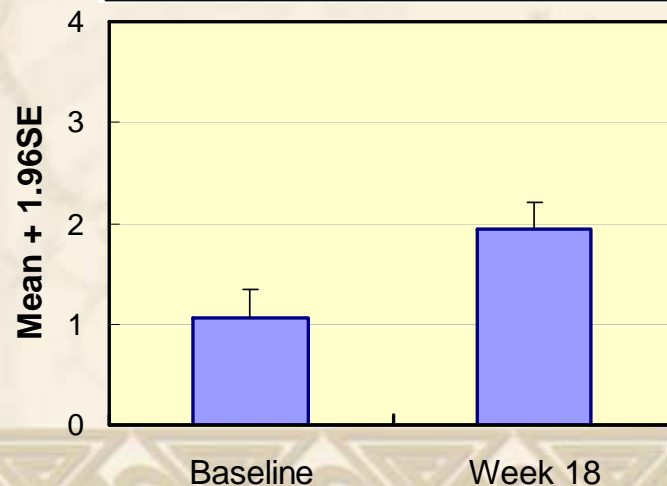
So, Was there a Change?

	n	Mean (SD)
Week 18 - Baseline	95	0.9 (4.5)

Median = 0

	p-value
Paired t-test	0.044
Signed rank test	<0.001
Sign test	<0.001
Modified sign test	0.581
Likelihood ratio sign test	1.000

Week 18 - Baseline	n	%
-19	2	2.11
-11	1	1.05
-8	1	1.05
-5	1	1.05
-4	1	1.05
-3	1	1.05
-2	2	2.11
0	39	41.05
1	10	10.53
2	11	11.58
3	14	14.74
4	3	3.16
5	4	4.21
6	1	1.05
10	1	1.05
12	1	1.05
13	1	1.05
15	1	1.05



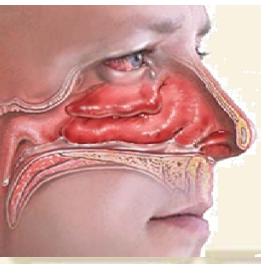
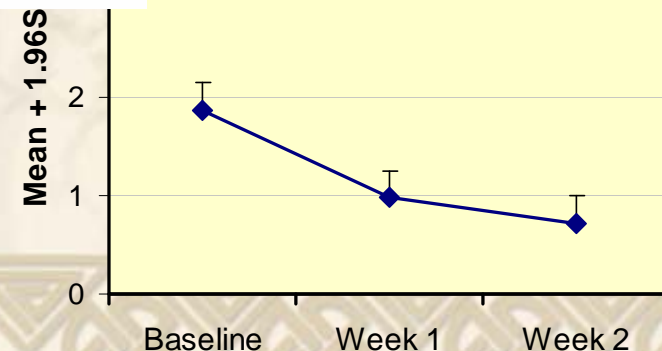
Was there a change of symptom score?

So, Was there a Change?

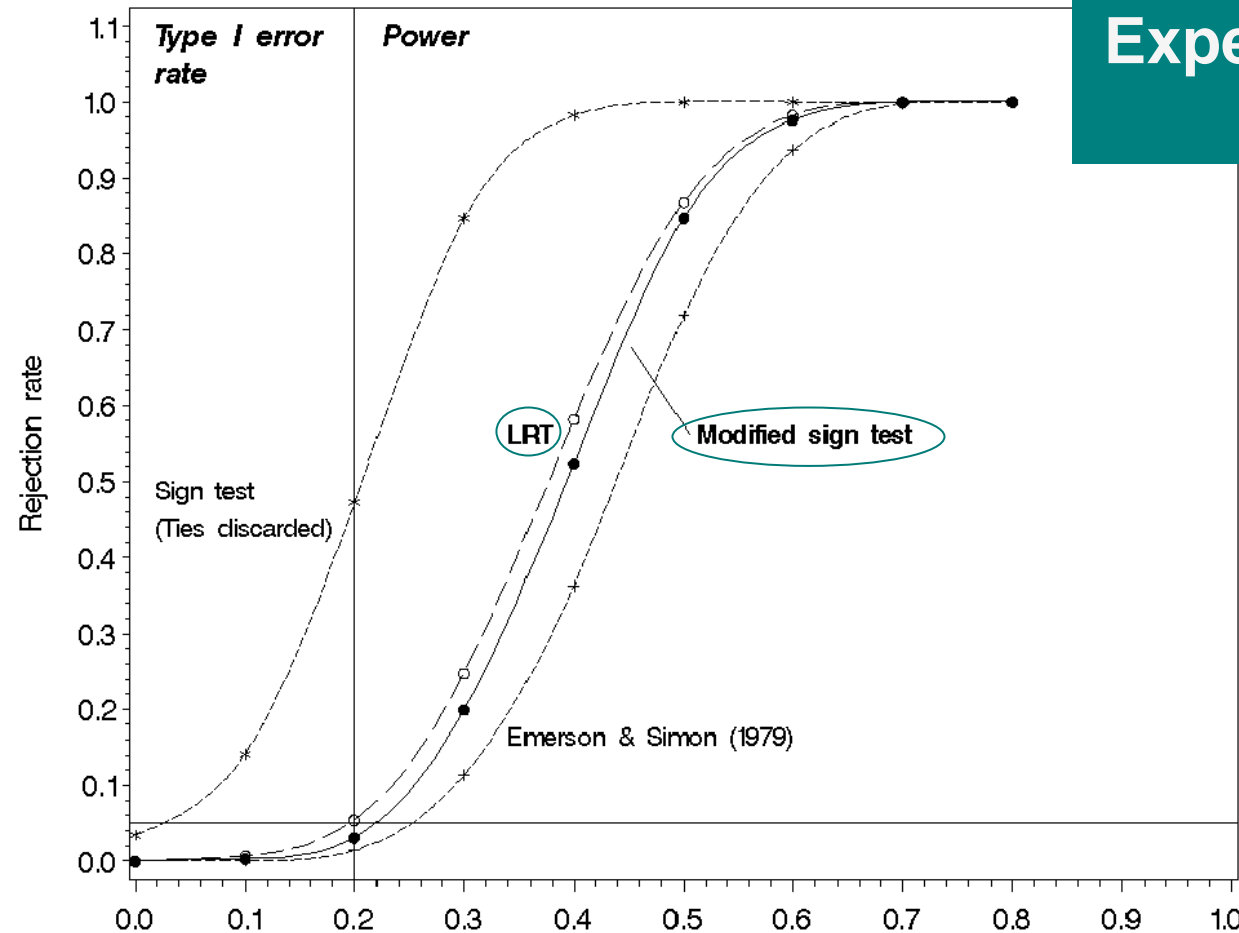
Not in the 2nd week !

	n	Mean (SD)	p-value		
			Paired t-test	Signed rank test	Sign test
Week 1 – Week 0	50	-0.9 (0.9)	<0.001	<0.001	<0.001
Week 2 – Week 1	47	-0.3 (0.7)	0.022	0.034	0.019
Week 2 – Week 0	47	-1.2 (1.1)	<0.001	<0.001	<0.001

	n	Median	p-value	
			Modified sign test	Likelihood ratio sign test
Week 1 – Week 0	50	-1	0.008	0.002
Week 2 – Week 1	47	0	0.996	1.000
Week 2 – Week 0	47	-1	<0.001	0.002



How Better was the likelihood ratio sign test (LRT) ?



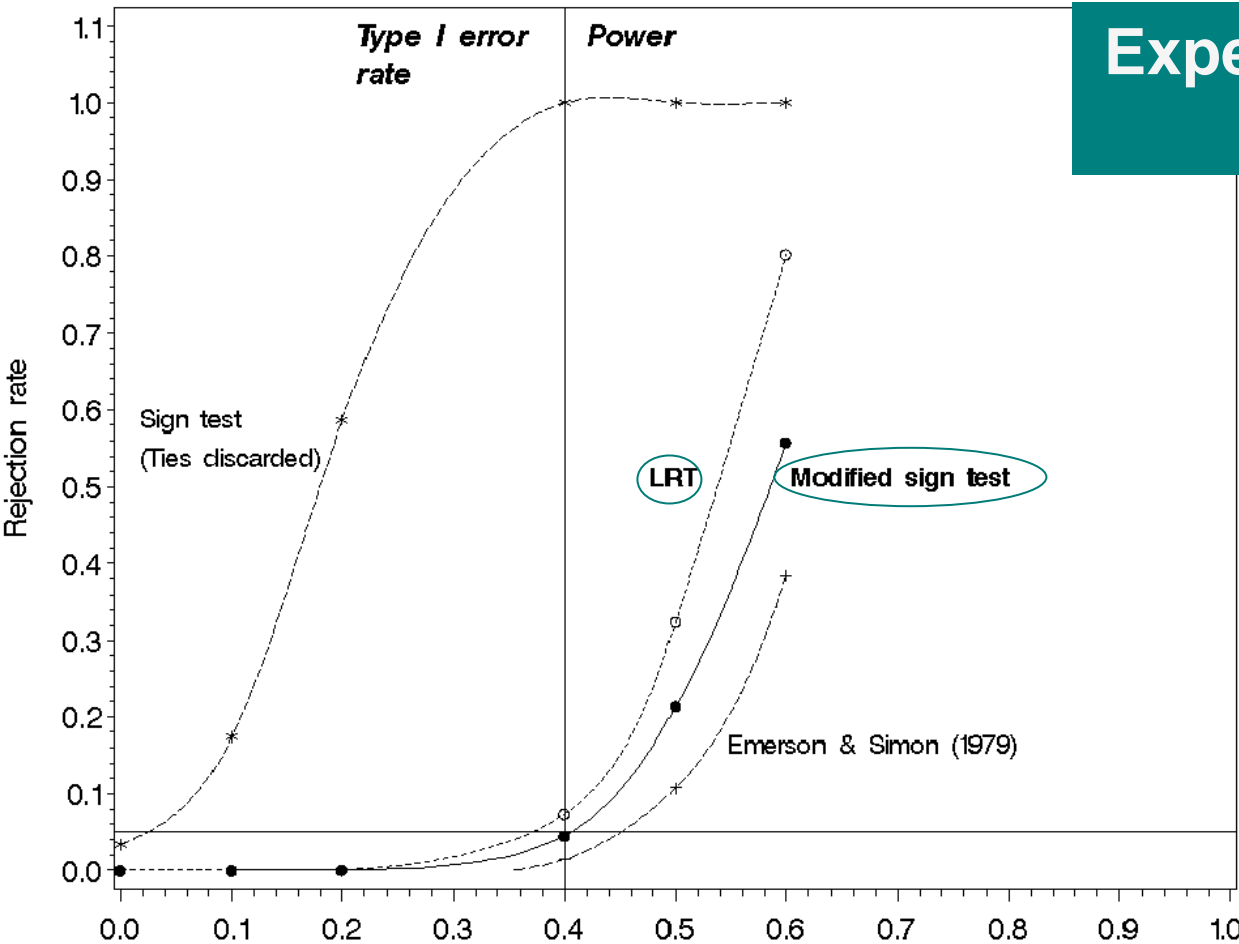
Expected ties proportion = 20%

❖ Not much better !

Proportion of +ve differences - Proportion of -ve differences

N = 80

How Better was the likelihood ratio sign test (LRT) ?



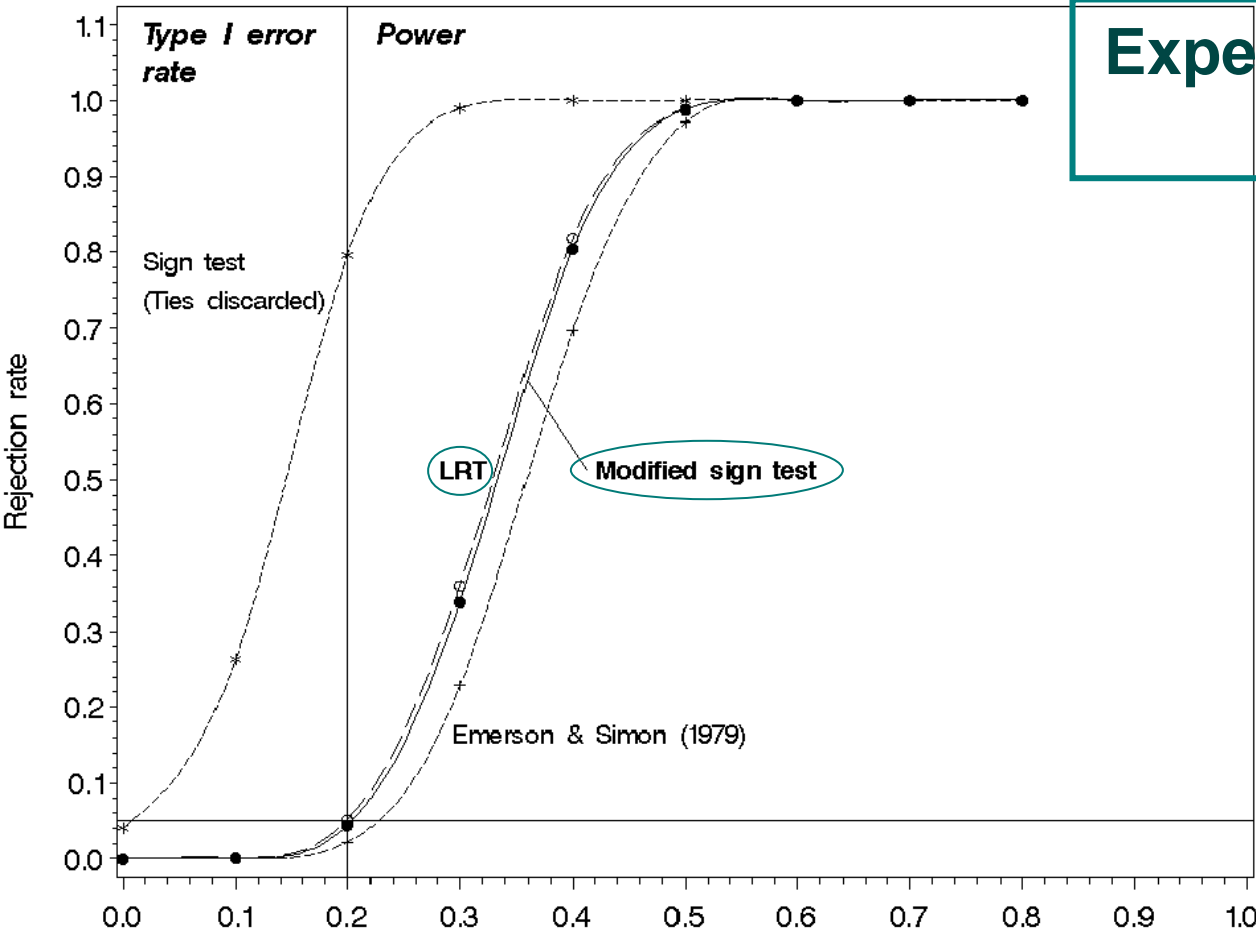
Expected ties proportion
= 40%

- ❖ Better !
- ❖ Difference increases with the expected ties proportion

Proportion of +ve differences – Proportion of –ve differences

N = 80

How Better was the likelihood ratio sign test (LRT) ?



Expected ties proportion = 40%

- ❖ Not much better !
- ❖ Difference decreases with the sample size

Ideas on Sample Size?

- ❖ No sample size calculation is currently available
- ❖ Monte Carlo simulation was used
- ❖ Need to specify

1. Ties proportion (p_0)

2. Either

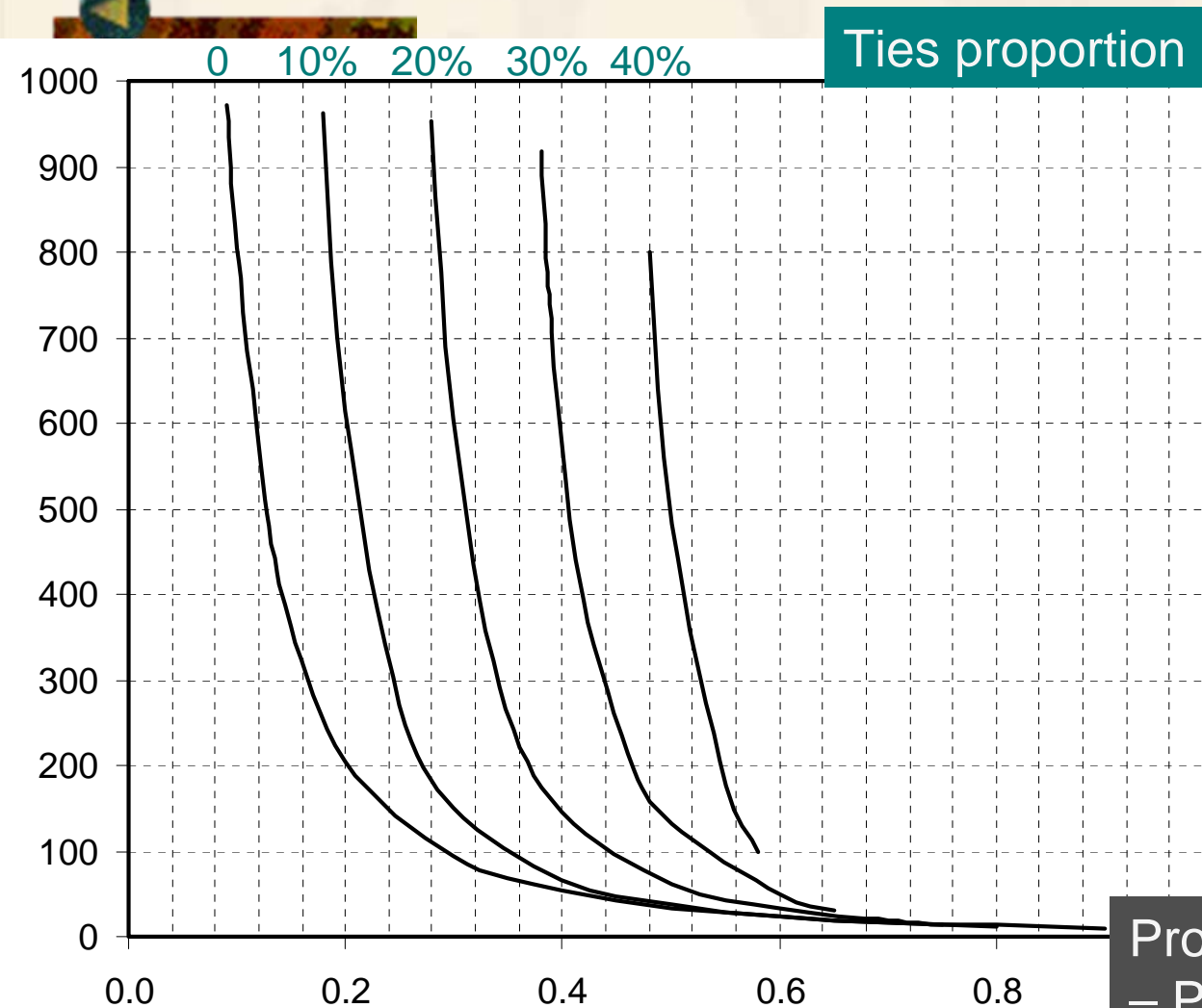
- a. $d =$ proportion of +ve differences (p_+)
– proportion of –ve differences (p_-); or

Median = 0 if and only if $|d| \leq p_0$

- b. $r = p_+ / p_-$

Median = 0 if and only if $1-2p_0 \leq r \leq 1/(1-2p_0)$

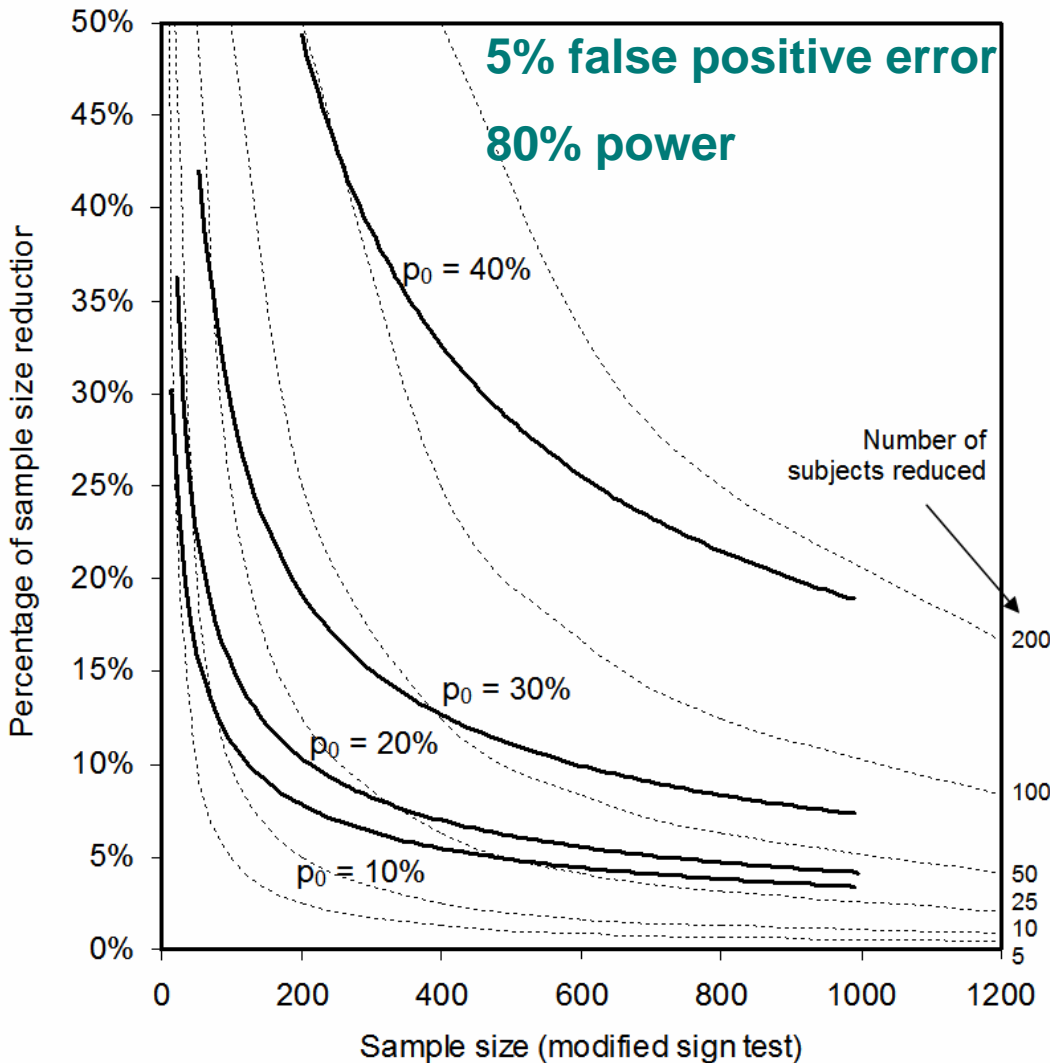
Sample Size for the Likelihood Ratio Sign Test



❖ Sample size may be severely under-determined if ties when foreseeable are not accounted

Proportion of +ve differences
– Proportion of -ve differences

Advantage of using the LRT over the modified sign test



- ❖ % sample size reduction increases exponentially when the sample size becomes small
- ❖ Not too much general difference when false positive error rate was 1%, 2% or 10%, or power was 90%

Conclusions and Recommendations

- ❖ Ignorance of ties may lead to false positive error
- ❖ Be sensitive to the presence of ties
- ❖ Check both the mean and the median for potential counter-intuitive results
- ❖ The likelihood ratio sign test is the preferable
- ❖ The modified sign test works well in large sample size

Acknowledgements



Collaborators:

- ❖ WK Ho, Department of Surgery, HKU
- ❖ CW Kwan, Clinical Trials Centre, HKU
- ❖ KF Lam, Department of Statistics & Actuarial Science, HKU
- ❖ KSL Lam, Department of Medicine, HKU
- ❖ YW Lee, Department of Statistics & Actuarial Science, HKU
- ❖ JST Sham, Department of Clinical Oncology, HKU

Research Grant:

- ❖ CRCG, HKU

*Thank
You*