Categorical Data Analysis

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Dec 19, 2025

Outlines

- Introduction
- Chi-squared Test of Association
- Fisher's Exact Test
- McNemar's Test

Learning outcomes

- Understand the concepts behind each test
- Understand when to use each test
- Able to perform chi-squared, Fisher's exact and McNemar's tests using SPSS, and interpret the results

Introduction

Introduction

- Factor (IV) and Outcome (DV) variables are both categorical
- Analyses of contingency / cross-tabulation table
- Depending on # categories of each
- e.g. 2x2, 3x2, 3x3 and so on
- Analyze cell counts

Introduction

- Analyses covered:
  - Chi-squared test
  - Fisher's exact test
  - McNemar's test

Chi-squared Test of Association

About

- Non-parametric test
- TWO independent samples
- Association between TWO categorical variables

About

- Cross-tabulation between TWO variables
- The association between the variables are made by comparing the observed cell counts (from data) with the expected cell counts (i.e. the count when variables are not associated to each other)

Observed Count (O)

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	**20**	**12**	32
No	**55**	**113**	168
Column Total	75	125	200

62.5%

32.7%

Expected Count (E)

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	**32*75/200 = 12**	**32*125/200 = 20**	32
No	**168*75/200 = 63**	**168*125/200 = 105**	168
Column Total	75	125	200

E = (Row Total x Column Total) / Grand Total

Chi-square (X^2)

$$X^2 = \text{SUM}\left(\frac{[O-E]^2}{E}\right)$$

$$df = (r-1)(c-1)$$

df = degree of freedom, r = # row, c = # column

$X^2 = SUM([O - E]^2/E)$

Smoker	Lung Cancer		
	Yes	No	
Yes	**$(20-12)^2/12 = 5.33$**	**$(12-20)^2/20 = 3.20$**	
No	**$(55-63)^2/63 = 1.00$**	**$(113-105)^2/105 = 0.61$**	
	$X^2 = SUM([O-E]^2/E)$		**10.14**

These values will be used by statistical software to get P-value

$$df = (2-1) * (2-1) = 1$$

Limitation

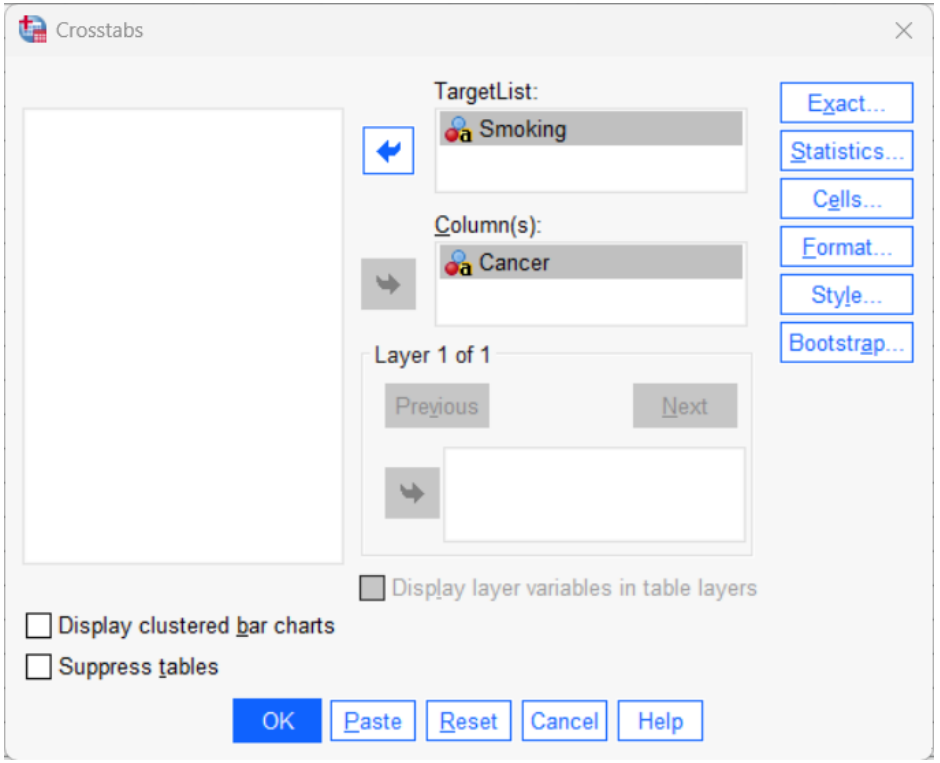
- Requirement – < 25% expected cell counts < 5
- If this assumption of X^2 is violated → Use Fisher's exact

Practical in SPSS

- Dataset: lung.sav
- Variables:
  - Smoking = Yes/No
  - Cancer = Yes/No

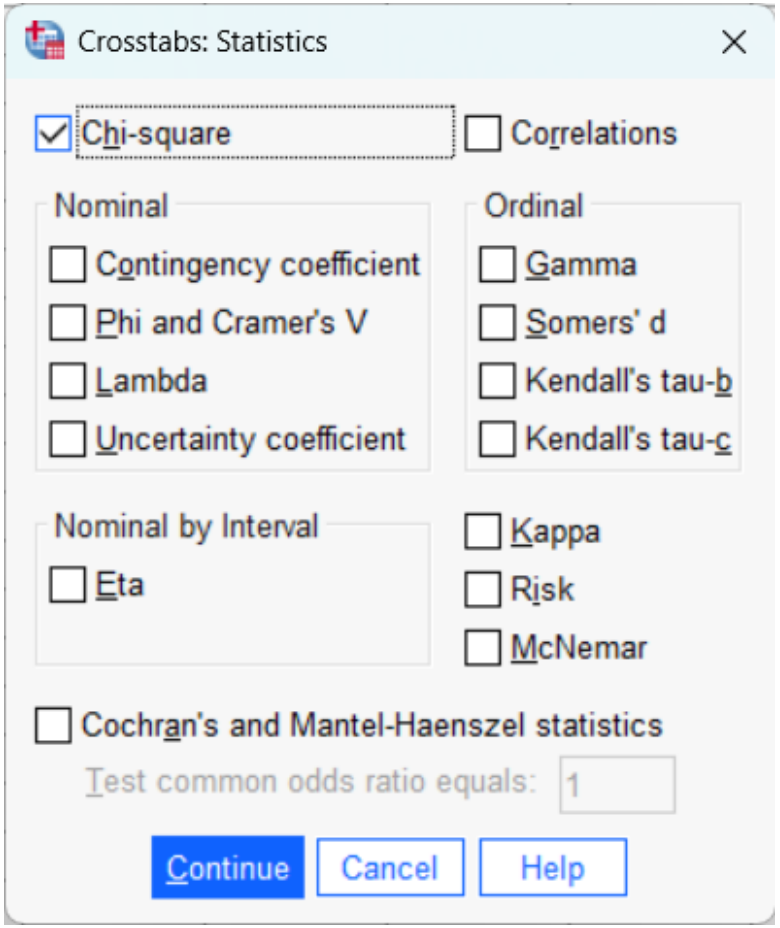
Practical in SPSS

- Open Crosstabs menu
  - Analyze → Descriptive Statistics → Crosstabs
  - TargetList = Smoking, Column(s) = Cancer



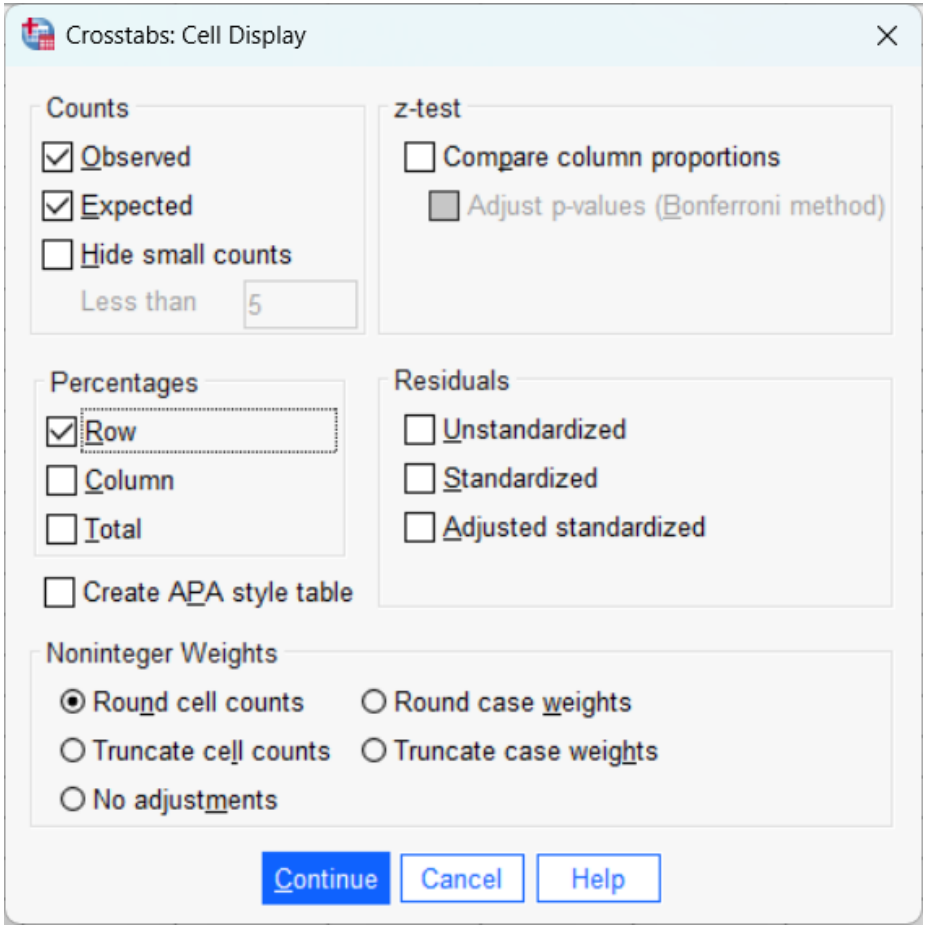
Practical in SPSS

- Statistics button
  - Check Chi-square
  - Continue



Practical in SPSS

- Cells button
  - Check Observed, Expected and Row
  - Continue, then OK in Crosstabs main window



Practical in SPSS

- Results

P < 0.05, Sig. Association between Smoking & Cancer (2 sided)

Smoking * Cancer Crosstabulation

			Cancer		
			cancer	no cancer	Total
Smoking	smoking	Count	20	12	32
		Expected Count	12.0	20.0	32.0
		% within Smoking	62.5%	37.5%	100.0%
	no smoking	Count	55	113	168
		Expected Count	63.0	105.0	168.0
		% within Smoking	32.7%	67.3%	100.0%
Total		Count	75	125	200
		Expected Count	75.0	125.0	200.0
		% within Smoking	37.5%	62.5%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	10.159[a]	1	.001		
Continuity Correction[b]	8.929	1	.003		
Likelihood Ratio	9.830	1	.002		
Fisher's Exact Test				.002	.002
Linear-by-Linear Association	10.108	1	.001		
N of Valid Cases	200				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.00.

b. Computed only for a 2x2 table

Assumption fulfilled for X^2 (E < 5 less < 25%)

Fisher's Exact Test

About

- Alternative of chi-squared test – when its requirement is not fulfilled
- For cross-tabulation with small cell counts (rare disease) – small expected cell counts
- Gives exact *P*-value, no statistical distribution involved (unlike chi-squared distribution)

Observed Count (O)

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	**5**	**10**	15
No	**2**	**28**	30
Column Total	7	38	45

33.3%

6.7%

Expected Count (E)

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	**15*7/45 = 2.33**	**15*38/45 = 12.67**	15
No	**30*7/45 = 4.67**	**30*38/45 = 25.33**	30
Column Total	7	38	45

2/4 cells < 5 = 50%
Cannot use X^2!

Fisher's exact

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	a	b	a + b
No	c	d	c + d
Column Total	a + c	b + d	n

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Fisher's exact

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	a = 5	b = 10	a + b = 15
No	c = 2	d = 28	c + d = 30
Column Total	a + c = 7	b + d = 38	n = 45

Calculate these by statistical software to get P-value

$$p = \frac{15!\,30!\,7!\,38!}{5!\,10!\,2!\,28!\,45!}$$

Fisher's exact

Smoker	Lung Cancer		Row Total
	Yes	No	
Yes	a = 5	b = 10	a + b = 15
No	c = 2	d = 28	c + d = 30
Column Total	a + c = 7	b + d = 38	n = 45

$$p = 0.028$$

Practical in SPSS

- Dataset: lung_small.sav
- Variables:
  - Smoking = Yes/No
  - Cancer = Yes/No

Practical in SPSS

- Same steps as chi-squared test, read Fisher's exact result

Practical in SPSS

- Results

P < 0.05, Sig. Association between Smoking & Cancer (exact 2 sided)

Clearly, assumption violated for X^2 (E < 5 less < 25%)

Smoker * Cancer Crosstabulation

			Cancer		
			No	Yes	Total
Smoker	No	Count	28	2	30
		Expected Count	25.3	4.7	30.0
		% within Smoker	93.3%	6.7%	100.0%
	Yes	Count	10	5	15
		Expected Count	12.7	2.3	15.0
		% within Smoker	66.7%	33.3%	100.0%
Total		Count	38	7	45
		Expected Count	38.0	7.0	45.0
		% within Smoker	84.4%	15.6%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.414[a]	1	.020		
Continuity Correction[b]	3.574	1	.059		
Likelihood Ratio	5.109	1	.024		
Fisher's Exact Test				.032	.032
Linear-by-Linear Association	5.293	1	.021		
N of Valid Cases	45				

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 2.33.

b. Computed only for a 2x2 table

McNemar’s Test

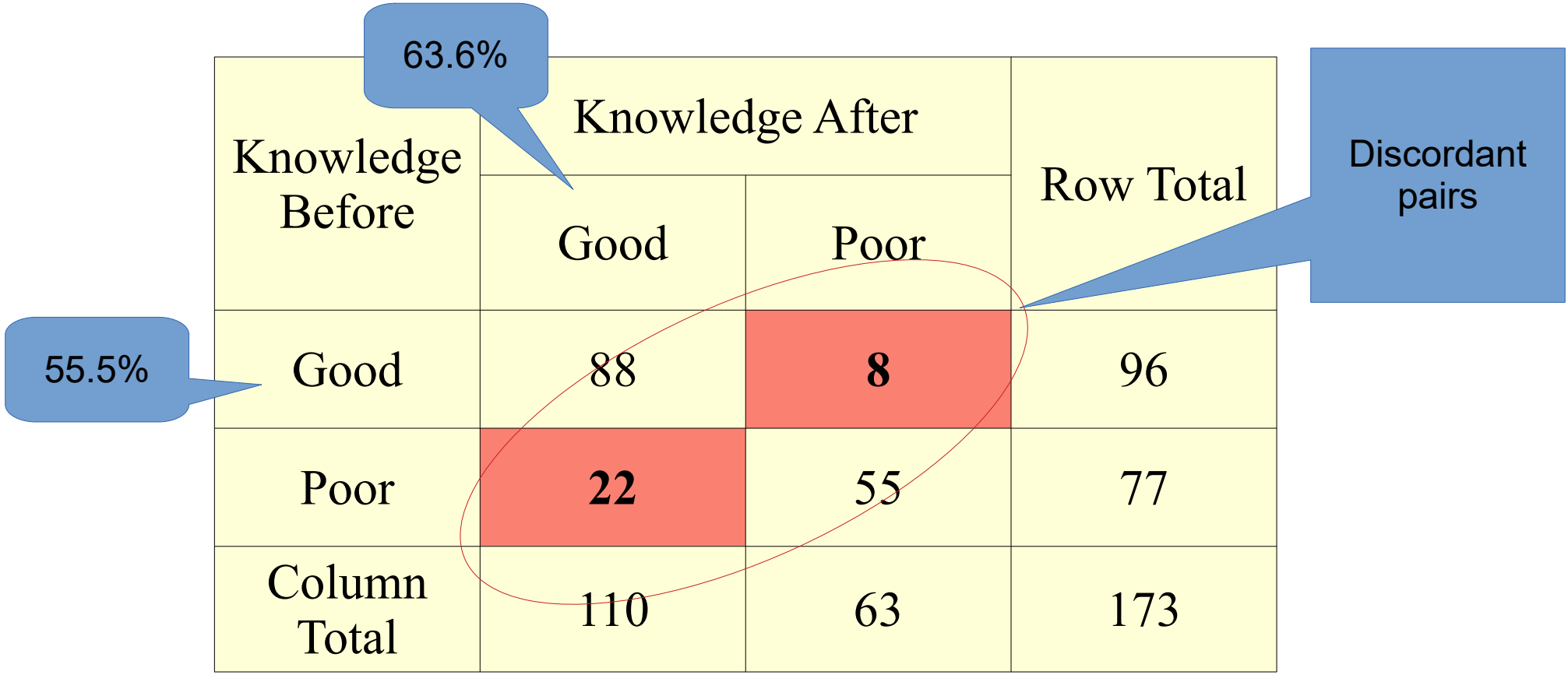
About

- Non-parametric test
- TWO dependent samples
- Association between TWO repeated categorical outcomes
- Any change?

About

- Cross-tabulation between TWO variables limited to 2x2 only
- It is concerned with whether the subjects still have the same outcomes (concordant) or different outcomes (discordant) upon repetition (pre-post)
- The association is determined by looking at the discordant cells

Observed Count (O)



Knowledge Before	Knowledge After		Row Total
	Good	Poor	
Good	88	**8**	96
Poor	**22**	55	77
Column Total	110	63	173

Chi-square (X^2) for McNemar

Knowledge Before	Knowledge After		Row Total
	Good	Poor	
Good	a	b	a + b
Poor	c	d	c + d
Column Total	a + c	b + d	n

$$X^2 = \frac{(b-c)^2}{b+c} \text{ with } df = 1$$

Chi-square (X^2) for McNemar

Knowledge Before	Knowledge After		Row Total
	Good	Poor	
Good	88	**8**	15
Poor	**22**	55	30
Column Total	7	38	45

These values will be used by statistical software to get P-value

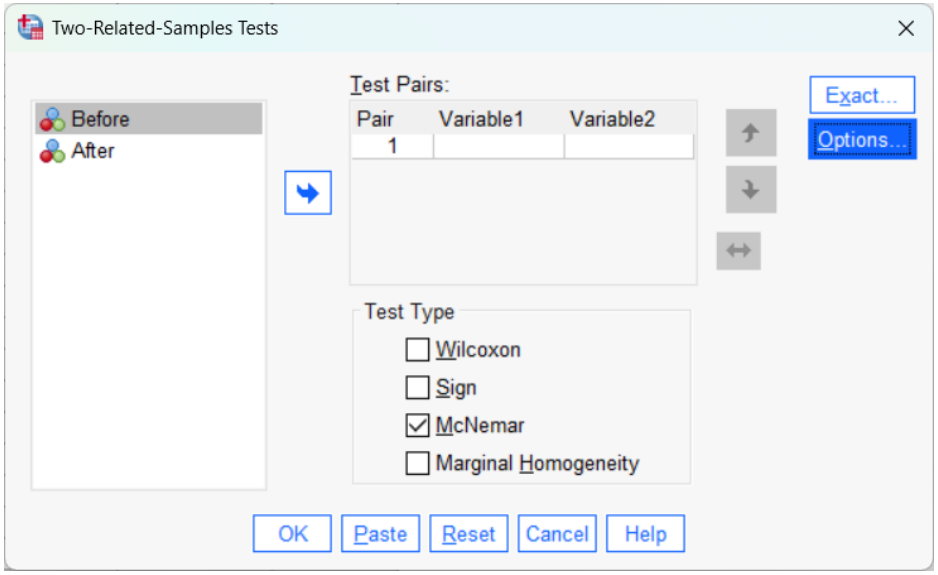
$$X^2 = \frac{(8-22)^2}{8+22} = 6.53, df = 1$$

Practical in SPSS

- Dataset: knowledge.sav
- Variables:
  - Before = Good/Poor
  - After = Good/Poor

Practical in SPSS

- Open 2 Related Samples menu
  - Add Before and After in Test Pairs
  - Uncheck Wilcoxon, Check McNemar
  - OK



Practical in SPSS

- Results

Smoker & Cancer

Smoker	Cancer: No	Cancer: Yes
No	28	2
Yes	10	5

SPSS uses exact P-value instead of X^2

P < 0.05, Sig. change pre-post

Test Statistics[a]

	Smoker & Cancer
N	45
Exact Sig. (2-tailed)	.039[b]

a. McNemar Test

b. Binomial distribution used.

Tutorial

Tutorial

- Datasets:
  - X^2: alzheimer.sav
  - Fisher: eofad.sav
  - McNemar: mmse.sav