Medicine & Society IIa

Analysis of Qualitative Data

Dr Azmi Mohd Tamil
Dept of Community Health
Universiti Kebangsaan Malaysia

Statistical Tests - Qualitative

Variable 1	Variable 2	Criteria	Type of Test
Qualitative	Qualitative	Sample size $\geq$ 20 dan no expected value < 5	Chi Square Test (χ^2)
Qualitative Dichotomus	Qualitative Dichotomus	Sample size > 30	Proportionate Test
Qualitative Dichotomus	Qualitative Dichotomus	Sample size > 40 but with at least one expected value < 5	χ^2 Test with Yates Correction
Qualitative Dichotomus	Qualitative Dichotomus	Sample size < 20 or (< 40 but with at least one expected value < 5)	Fisher Test

CHI-SQUARE TEST

CHI-SQUARE TEST

- The most basic and common form of statistical analysis in the medical literature.
- Data is arranged in a contingency table (*R* X *C*) comparing 2 qualitative data.
- *R* stands for number of rows and *C* stands for number of columns.

CHI-SQUARE TEST

- There are two different types of chi-square (χ^2) tests, both involve categorical data.

1. The chi-square for goodness of fit
2. The chi-square test for independence

1. The chi-square for goodness of fit

- Also referred to as one-sample chi-square.
- It explores the proportion of cases that fall into the various categories of a single variable, and compares these with hypothesized values

1. The chi-square for goodness of fit

- We test that the null hypothesis that the observed frequencies, proportion, percentage distribution for an experiment or a survey follow a certain or a given pattern theoretical distribution (hypothesized value)

2. The chi-square test for independence

- It is used to determine if two categorical variables are related.
- It compares the frequency of cases found in the various categories of one variable across the different categories of another variable.
- Each of these variables can have two or more categories.

2. The chi-square test for independence

- Example of research questions:
  - Are males more likely to be smokers than females?
  - Is the proportion of males that smoke the same as the proportion of females?
  - Is there a relationship between gender and smoking behaviour?

2. The chi-square test for independence

- Two categorical variables involved (with two or more categories in each):
  - Gender (Male / Female)
  - Smoker (Yes / No)

Assumptions for χ^2

- Random samples
- Independent observations. Each person or case can only be counted once, they cannot appear in more than one category or group, and the data from one subject cannot influence the data from another.
- Lowest expected frequency in any cell should be 5 or more.

CHI-SQUARE TEST

CRITERIA:

- Both variables are qualitative data.
- Sample size of ≥ 20.
- No cell that has expected value of < 5.

CONTINGENCY TABLE

INFECTION	YES	NO	TOTAL
MALE	a	b	e
FEMALE	c	d	f
TOTAL	g	h	n

CHI-SQUARE TEST

- *E* = expected value
- Expected value for cell a :

$$\frac{e \times g}{n}$$

- Expected value for cell b :

$$\frac{e \times h}{n}$$

FORMULA

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

FORMULA

$$x^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}$$

Only for 2 x 2 table

The steps for a χ^2 test

- Formulate the null and alternative hypotheses, and select an α -level.
- Collect a sample and compute the statistic of interest.
- Determine the degree of freedom $(R - 1)(C - 1)$

The steps for a χ^2 test

- Arrange data in contingency table.
- Calculate expected value for each cell.
- Calculate χ^2 test

The steps for a χ^2 test

- Determine the *critical values* of the test statistic as determined by the ***α*** -level.
- Compare the test statistic to the critical values. If the test statistic is :
  - more than the critical values, reject null hypothesis.
  - Equal or less than the critical values, fail to reject null hypothesis.

Example:

Jadual observasi

	+	-	
+	29	24	53
-	67	80	147
	96	104	200

Jadual jangkaan

	+	-	
+	96*53/200	104*53/200	g
-	96*147/200	104*147/200	h
	e	f	n

Example:

Jadual observasi

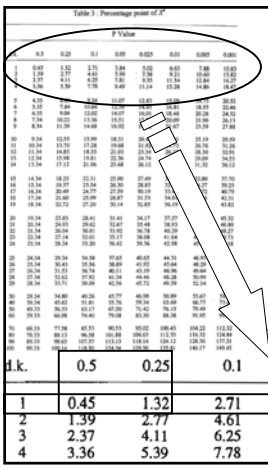
	+	-	
+	29	24	53
-	67	80	147
	96	104	200

Jadual jangkaan

	+	-	
+	25.44	27.56	g
-	70.56	76.44	h
	e	f	n

Example:

- $X^2 = \frac{(29-25.44)^2}{25.44} + \frac{(24-27.56)^2}{27.56} + \frac{(67-70.56)^2}{70.56} + \frac{(80-76.44)^2}{76.44}$
- $X^2 = 1.303$
- df = (2-1)(2-1) = 1
- Critical value for df = 1 for p=0.05 is 3.84,
- The calculated X^2 is smaller than the critical value, therefore the null hypothesis is not rejected.
- Conclusion: No association between the risk factor and the outcome.



Refer to Table 3.
Look at df = 1.

X^2 = 1.303, larger than 0.45 (p=0.5) but smaller than 1.32 (p=0.25).
0.45 (p=0.5) < 1.303 < 1.32 (p=0.25)
Therefore if X^2 = 1.303, 0.5<p<0.25.
Since the calculated p > 0.05, null hypothesis not rejected.

d.k.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47

Validity of X^2

- For contingency tables larger than 2x2 (i.e. 3x2 or 3x3), X^2 is valid if less than 20% of the expected values are less than 5 and none are less than one.
- If there are many expected values of less than 5, you can try merging the cells with small values to overcome this problem.

Examples for Validity of X^2

Observed	Stressed	Not Stressed	Total
Underweight	6	14	20
Normal	50	20	70
Overweight	9	1	10
Total	65	35	100

Expected	Stressed	Not Stressed	Total
Underweight	13	7	20
Normal	46	25	70
Overweight	6.5	3.5	10
Total	65	35	100

- Only one cell out of six cells have expected values of less than 5, which is 3.5.
- 1/6 = 16.67%, less than 20%, so X^2 still valid.

YATES CORRECTION

- When sample sizes are small, the use of χ^2 will introduces some bias into the calculation, so that the χ^2 value tends to be a little too large.
- To remove the bias, we use continuity correction (Yates Correction)

CRITERIA FOR YATES CORRECTION

- Both variables are dichotomous qualitative (2 X 2 table).
- Sample size of ≥ 40.
- One of the cell has expected value of < 5.

YATES CORRECTION FORMULA

$$\chi^2 = \sum \left[\frac{(|O - E| - 0.5)^2}{E} \right]$$

FISHER'S EXACT TEST

CRITERIA:

- Both variables are dichotomous qualitative (2 X 2 table).
- Sample size of < 20.
- Sample size of 20 to < 40 but one of the cell has expected value of < 5.

FORMULA FOR FISHER'S EXACT TEST

$$\frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{N!\,a!\,b!\,c!\,d!}$$

Weird since you have to calculate for many tables, until one of the cell becomes 0, then total up all the p values.

Example

Distribution of Underweight and Normal Weight for Taxi Drivers and Bus Drivers

	Underweight	Normal	Total
Bus Drivers	8	11	19
Taxi Drivers	3	11	14
Total	11	22	33

There is an association between the prevalence of underweight and the type of vehicle driven by the public vehicle drivers.

In this analysis, it is a 2 X 2 table, cells with expected value < 5 and small sample size, therefore the best type of analysis possible is **Fisher's Exact Test**.

Step 1

- $\frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{N!\,a!\,b!\,c!\,d!}$
- p1 = $\frac{19!14!11!22!}{33!8!11!3!11!}$ = 0.142

Step 2

- Create 3 more extreme tables by deducting 1 from the smallest value. Continue to do so till the cell becomes zero;

KB	N	
9	10	19
2	12	14
11	22	33

KB	N	
10	9	19
1	13	14
11	22	33

KB	N	
11	8	19
0	14	14
11	22	33

- p2 = 0.0434
  p3 = 0.00668
  p4 = 0.00039

Step 3

- Total p = 0.142+0.0434+0.00668+0.00039
  = 0.19247
- This is the p value for single-tailed test. To make it the p value for 2 tailed, times the value with 2; p = 0.385.
- p is bigger than 0.05, therefore the null hypothesis is not rejected.
- There is no association between occupation and UW ;-)

SPSS Output

KERJA * OBESITI Crosstabulation

Count

		OBESITI		
		Underweight	Normal	Total
KERJA	Bus Driver	8	11	19
	Taxi Driver	3	11	14
Total		11	22	33

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.551[b]	1	.213		
Continuity Correction[a]	.760	1	.383		
Likelihood Ratio	1.598	1	.206		
Fisher's Exact Test				.278	.193
Linear-by-Linear Association	1.504	1	.220		
N of Valid Cases	33				

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.67.

McNemar Test

- It is a test to compare before and after findings in the same individual or to compare findings in a matched analysis (for dichotomous variables)
- Example: a researcher wanted to compare the attitudes of medical students toward confidence in statistics analysis before and after the intensive statistics course.

Data Collected

		Post-course -ve	Post-course +ve	Total
Pre-course	-ve	150 (a)	22 (b)	172
	+ve	8 (c)	20 (d)	28
Total		158	42	200

McNemar

- **SIGNIFICANCE OF DIFFERENCE IN EXPOSURE**

$X^2 = (b-c)^2 / b + c$ (1 df)

Odds ratio = c / b

Calculation

- McNemar $\chi 2 = \frac{(b - c)^2}{b + c}$
- $= \frac{(22 - 8)^2}{22 + 8}$
- $= 6.5333$

Yates Correction for McNemar Test

McNemar $\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$

When an expected value is less than 5

Calculation of the degree of freedom

$df = (R - 1)(C - 1)$

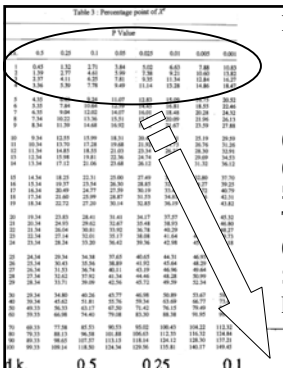
$= (2 - 1)(2 - 1)$

$= 1$

Determination of the *p* value

Value from the chi-square table for 6.53 on df=1, $p < 0.02$ (statistically significant)

Interpretation: there is a significant change in the attitudes of medical students toward confidence in statistics analysis before and after the intensive statistics course.



Refer to Table 3.
Look at df = 1.

$X^2 = 6.53$, larger than 5.02 (p=0.025) but smaller than 6.63 (p=0.01).
5.02 (p=0.025) < 6.53 < 6.63 (p=0.01)
Therefore if $X^2 = 6.53$, $0.01<p<0.025$.
Since the calculated $p > 0.05$, null hypothesis rejected.

d.k.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47

Also use in matched pair case-control studies

Example;

MATCHED PAIR C-C STUDY

CASES

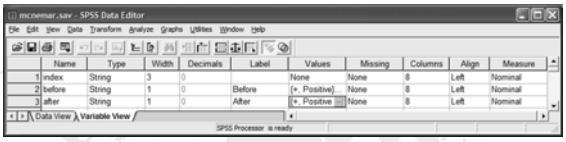
CONTROL		Exposed	Not exposed
	Exposed	**a** (both pairs exposed)	**b** (pairs of controls exposed)
	Not exposed	**c** (pairs of cases exposed)	**d** (both pairs not exposed)

MATCHED PAIRS C-C STUDY

- **SIGNIFICANCE OF DIFFERENCE IN EXPOSURE**

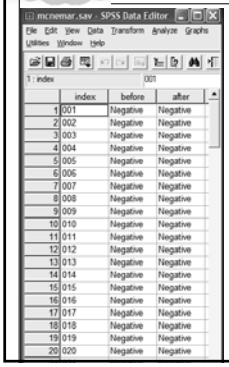
$X^2 = (b-c)^2 / b + c$ (1 df)
Odds ratio = c / b

McNemar in SPSS



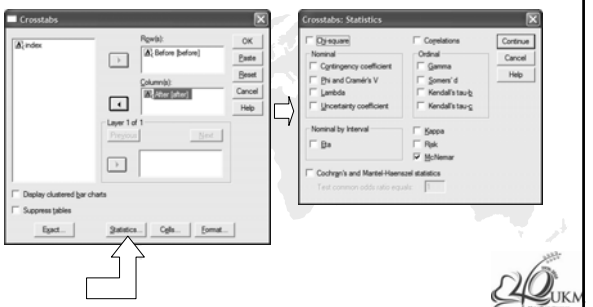
- The code for before and after (or case & control) must be similar.
- i.e. if one uses 1 for “Present” for before, 1 should also means the same for after.

McNemar in SPSS



- Data should be entered in pairs as illustrated on the left here.

SPSS – Crosstabs Command



SPSS Output

Before * After Crosstabulation

Count

		After		
		Positive	Negative	Total
Before	Positive	20	8	28
	Negative	22	150	172
Total		42	158	200

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.016[a]
N of Valid Cases	200	

a. Binomial distribution used.