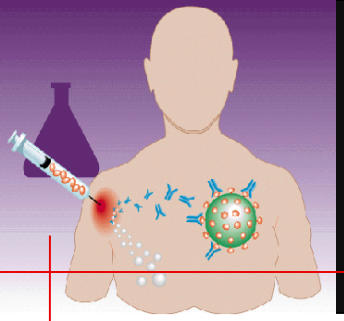
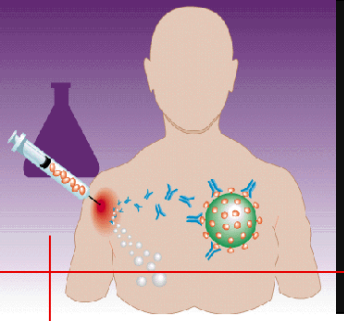


Research Week 2015

Analysis of Qualitative Data

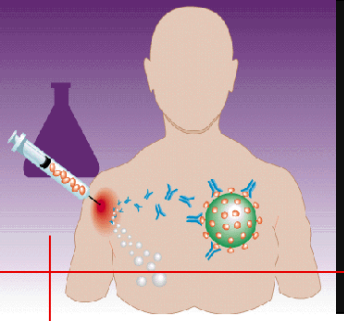
Assc. Prof. Dr Azmi Mohd Tamil
Dept of Community Health
Universiti Kebangsaan Malaysia

1970-2010
40 UKM
Mengilham Harapan
Mencipta Masa Depan



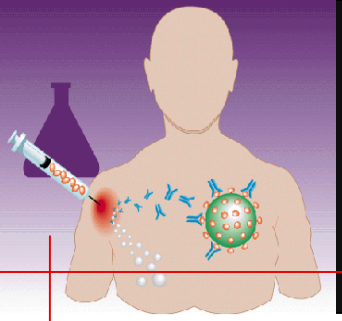
Statistical Tests - Qualitative

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

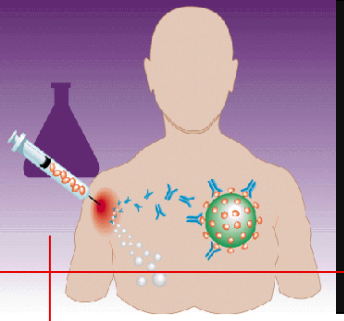


Hierarchy

- Usually we want to use Pearson Chi-Square.
- However if Pearson Chi-square test is not valid, then we opt for Yates's Correction.
- If Yates's Correction is not valid, then we opt for Fisher's Exact Test.

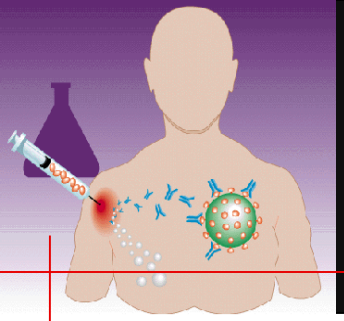


Pearson's CHI-SQUARE TEST



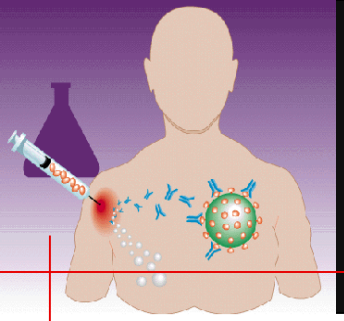
CHI-SQUARE TEST

- 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.



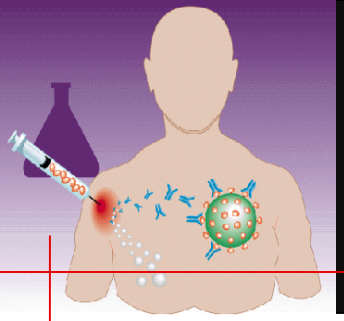
The chi-square test for independence

- 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.



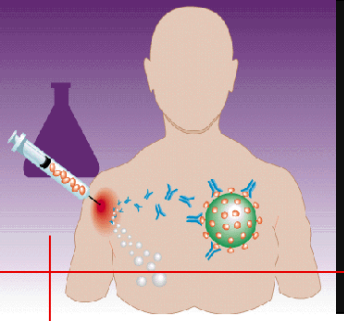
The chi-square test for independence

- 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?
    - Comparing the rate of smokers between males & females.
  - Is there a relationship between gender and smoking behaviour?



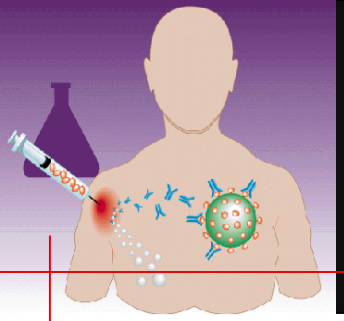
The chi-square test for independence

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



Assumptions for χ^2

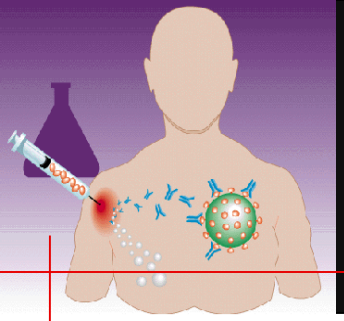
- Random samples
- Independent observations. Each person or case can only be counted once, they cannot appear in more than one category or group, and the data from one subject cannot influence the data from another.
- Lowest expected frequency in any cell should be 5 or more.



CHI-SQUARE TEST

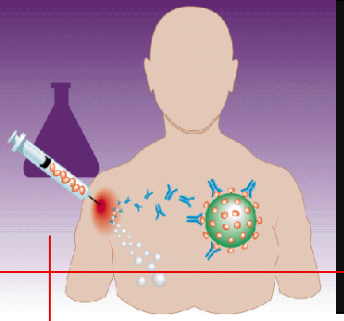
CRITERIA:

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



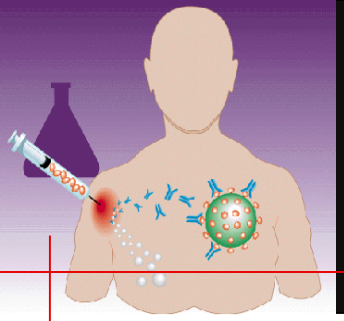
CONTINGENCY TABLE

Smoking	YES	NO	TOTAL
MALE	a (rate dis exposed)	b	e
FEMALE	c (rate dis non exposed)	d	f
TOTAL	g	h	n



FORMULA

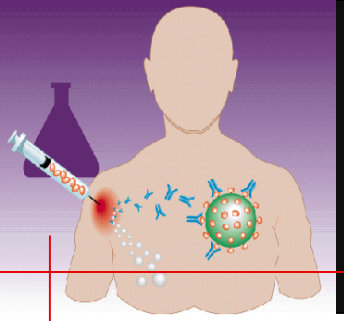
$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$



FORMULA

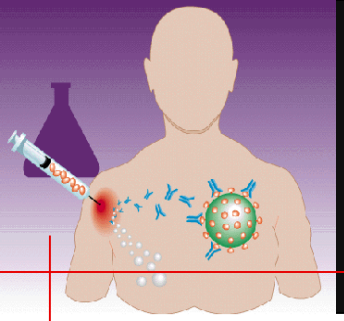
$$x^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}$$

Only for 2 x 2 table



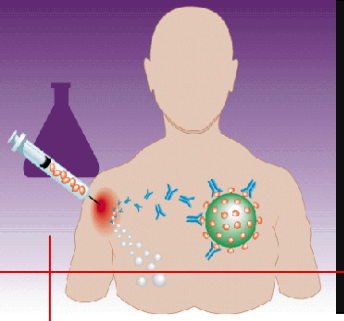
The steps for a χ^2 test

- 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 $(R - 1)\,(C - 1)$



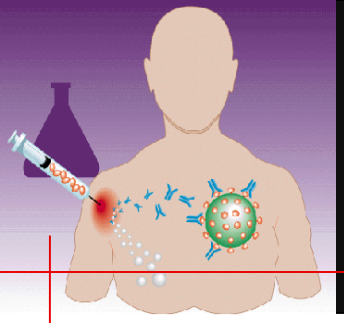
The steps for a χ^2 test

- 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:

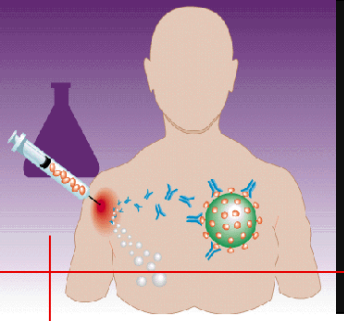
Jadual observasi

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

Male 55% vs female 45%

Jadual jangkaan

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



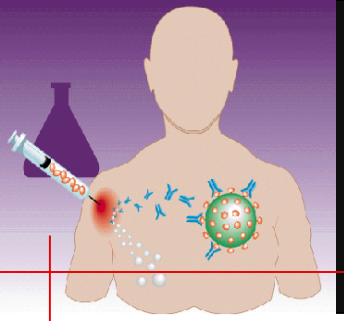
CHI-SQUARE TEST

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

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

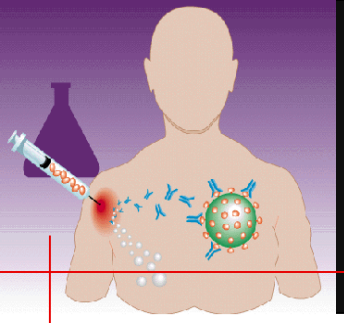
- Expected value for cell b :

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



Why need to calculate expected value?

- Because we are testing for the null hypothesis. Null hypothesis stated that there is no difference of worm infestation rate between male group and female group.
- Expected value is the value if there is no difference of worm infestation rate between the two groups.



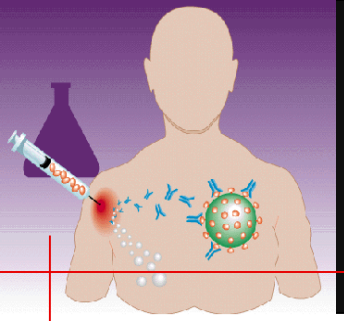
Example:

Jadual observasi

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

Jadual jangkaan

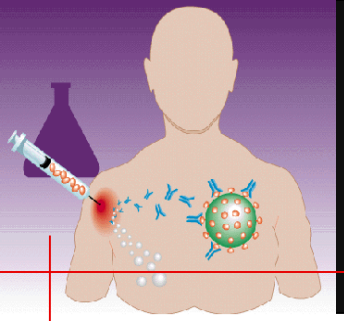
	+	-	
+	25.44	27.56	g
-	70.56	76.44	h
	e	f	n



Why need to calculate expected value?

Expected	Worm	No Worm	Total
Male	25.44 (48%)	27.56 (52%)	53
Female	70.56 (48%)	76.44 (52%)	147
Total	96 (48%)	104	200

If there is no difference of worm infestation rate between the two groups, then the worm infestation rate is similar (48% & 48%).



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 sig. diff. of worm infestation rate between male & female. No association between the risk factor and the outcome.
- Hint: Memorise critical values of 3.84 for df=1 and 5.99 for df=2.

Table 3 : Percentage point of X^2

d.k.	P Value							
	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
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.5	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	1	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.4	1	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	3	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.2	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	2	29.69	34.53
14	13.34	17.12	21.06	23.68	26.12	2	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	3	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32	4.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.	72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.8	6	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19		43.82
20	19.34	23.83	28.41	31.41	34.17	37.57		45.32
21	20.34	24.93	29.62	32.67	35.48	38.93		46.80
22	21.34	26.04	30.81	33.92	36.78	40.29		48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	4	9.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45	18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.4
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	8
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99
70	69.33	77.58	85.53	90.53	95.02	100.43	104.22	112.32
80	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
100	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

Refer to Table 3.

Look at df = 1.

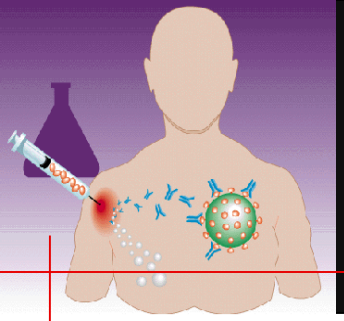
$X^2 = 1.303$, larger than 0.45 (p=0.5) but smaller than 1.32 (p=0.25).

0.45 (p=0.5) < 1.303 < 1.32 (p=0.25)

Therefore if $X^2 = 1.303$, 0.5>p>0.25.

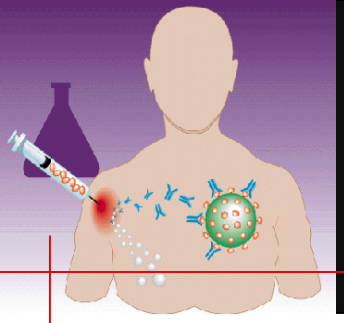
Since the calculated p > 0.05, null hypothesis not rejected.

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.

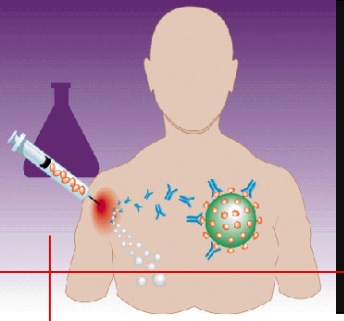


Examples for Validity of X^2

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

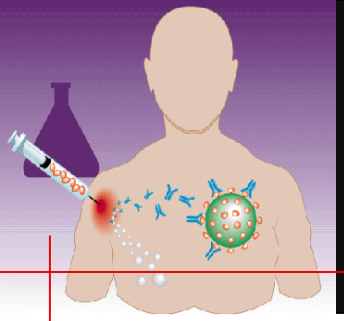
Expected	Stressed	Not Stressed	Total
Underweight	13	7	20
Normal	46	25	70
Overweight	6.5	3.5	10
Total	65	35	100

- 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.



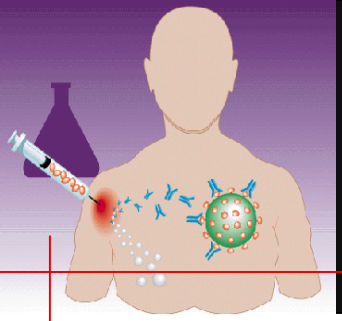
YATES' CORRECTION

- 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)



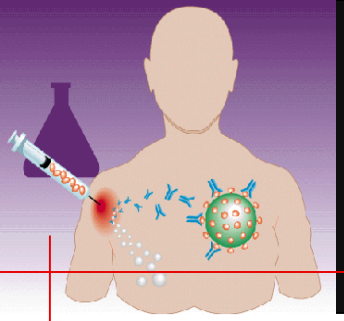
CRITERIA FOR YATES CORRECTION

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



YATES CORRECTION FORMULA

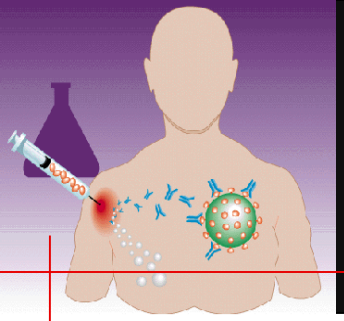
$$\chi^2 = \sum \left[\frac{(|O - E| - 0.5)^2}{E} \right]$$



FISHER'S EXACT TEST

CRITERIA:

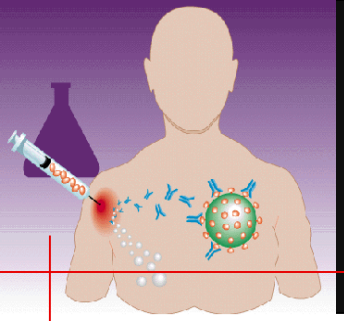
- Both variables are dichotomous qualitative (2 X 2 table).
- Sample size of < 20.
- Sample size of 20 to < 40 but one of the cell has expected value of < 5.



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.



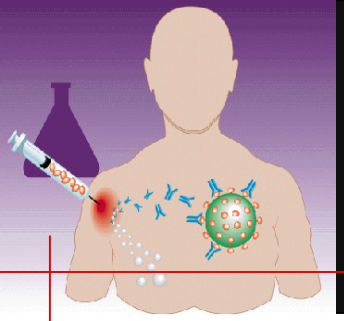
Example

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 an expected value < 5 (4.67) and small sample size, therefore the best type of analysis possible is **Fisher's Exact Test**.

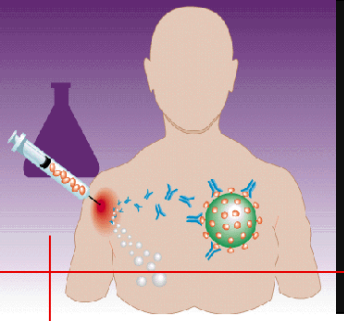


Step 1

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

- $p1 = \frac{19!14!11!22!}{33!8!11!3!11!}$

$= \frac{4.758 \times 10^{56}}{3.3471 \times 10^{57}} = 0.142$



Step 2

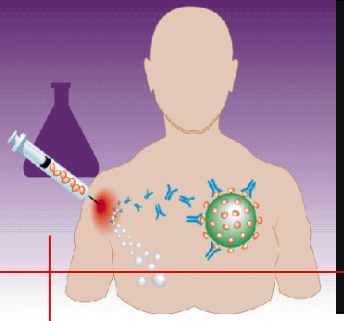
- Create 3 more extreme tables by deducting 1 from the smallest value. Continue to do so till the cell becomes zero;

KB	N	
9	10	19
2	12	14
11	22	33

KB	N	
10	9	19
1	13	14
11	22	33

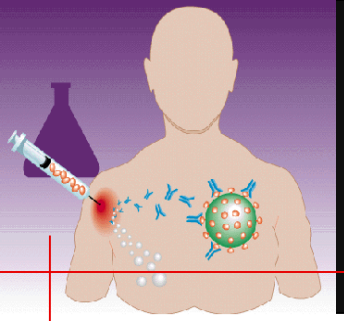
KB	N	
11	8	19
0	14	14
11	22	33

- 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 larger than 0.05, therefore the null hypothesis is not rejected.
- There is no association between occupation and UW ;-)



SPSS Output

KERJA * OBESITI Crosstabulation

Count

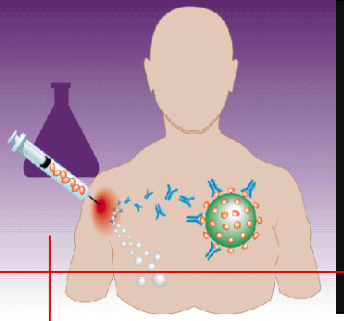
		OBESITI		
		Underweight	Normal	Total
KERJA	Bus Driver	8	11	19
	Taxi Driver	3	11	14
Total		11	22	33

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				

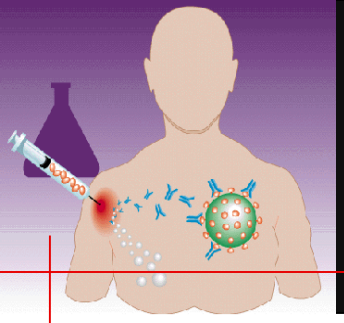
a. Computed only for a 2x2 table

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



McNemar Test

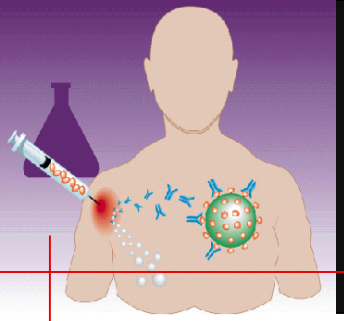
- It is a test to compare before and after findings in the same individual or to compare findings in a matched analysis (for dichotomous variables)
- Example: a researcher wanted to compare the attitudes of medical students toward confidence in statistics analysis before and after the intensive statistics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



Data Collected

Concordant
discordant

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



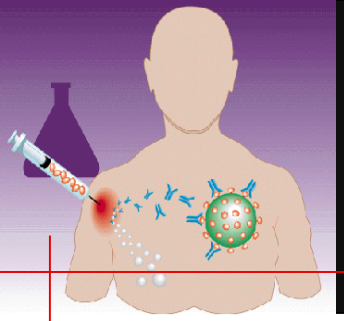
McNemar

- **SIGNIFICANCE OF DIFFERENCE IN EXPOSURE**

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

Odds ratio = c / b

$$b + c \geq 25$$

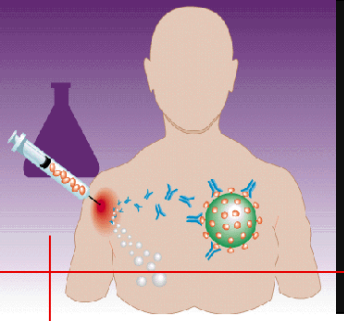


Calculation

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

- OR = c/b = 22/8 = 2.75

1970-2010 40 UKM
Mengilham Harapan
Mencipta Masa Depan



Calculation of the degree of freedom

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

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

$$= 1$$

Table 3 : Percentage point of X^2

d.k.	P Value							
	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
5	4.35		9.24	11.07	12.83	15.09		20.52
6	5.35	7.84	10.64			16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51		20.09	21.96	26.13
9	8.34	11.39	14.68	16.92		21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.4		25.19	29.59
11	10.34	13.70	17.28	19.68	21.92		26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.2	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74		29.69	34.53
14	13.34	17.12	21.06	23.68	26.12		31.32	36.12
15	14.34	18.25	22.31	25.00	27.49		32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32	4.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.	72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.8	6	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19		43.82
20	19.34	23.83	28.41	31.41	34.17	37.57		45.32
21	20.34	24.93	29.62	32.67	35.48	38.93		46.80
22	21.34	26.04	30.81	33.92	36.78	40.29		48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	4	9.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45	18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.4
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	8
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99
70	69.33	77.58	85.53	90.53	95.02	100.43	104.22	112.32
80	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
100	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

Refer to Table 3.

Look at df = 1.

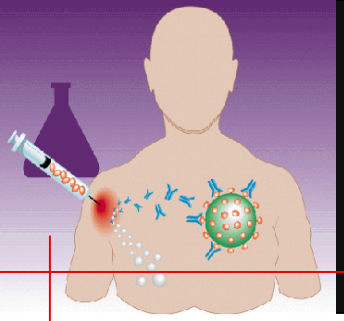
$X^2 = 6.53$, larger than 5.02 (p=0.025) but smaller than 6.63 (p=0.01).

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.

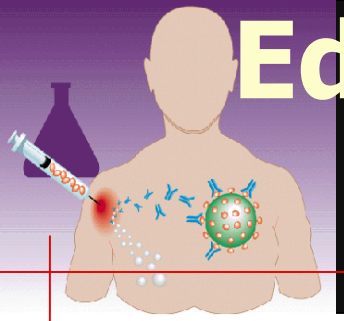
d.k.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47



Determination of the *p* value

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

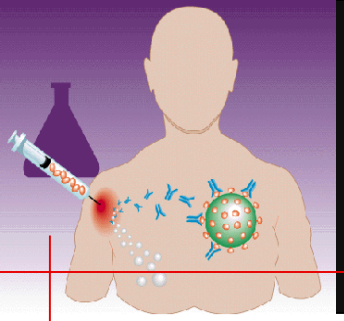
Interpretation: there is a significant change in the attitudes of medical students toward confidence in statistics analysis before and after the intensive statistics course.



Edward's Continuity Correction for McNemar Test

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

When b+c is less than 25



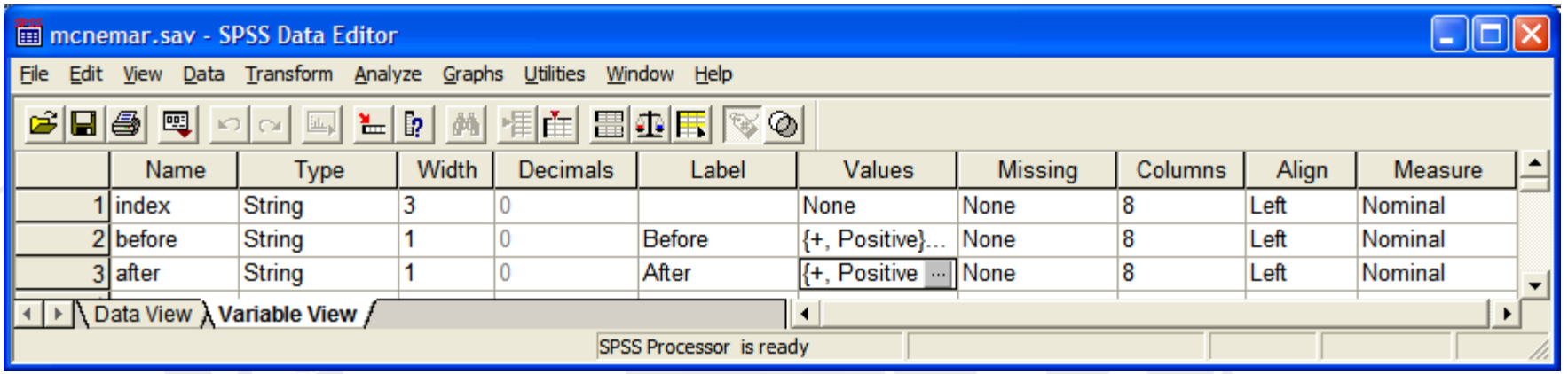
Also use in matched pair case-control studies

Example;

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

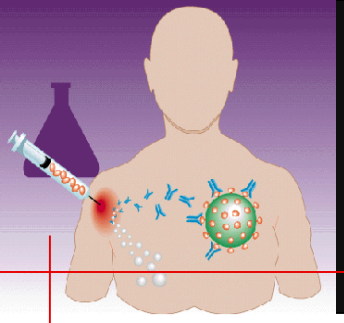
1970-2010 40 UKM Mengilham Harapan Mencipta Masa Depan

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



mcnemar.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs
Utilities Window Help

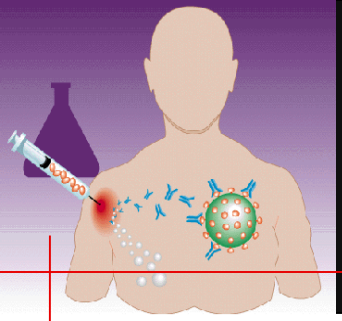
1 : index 001

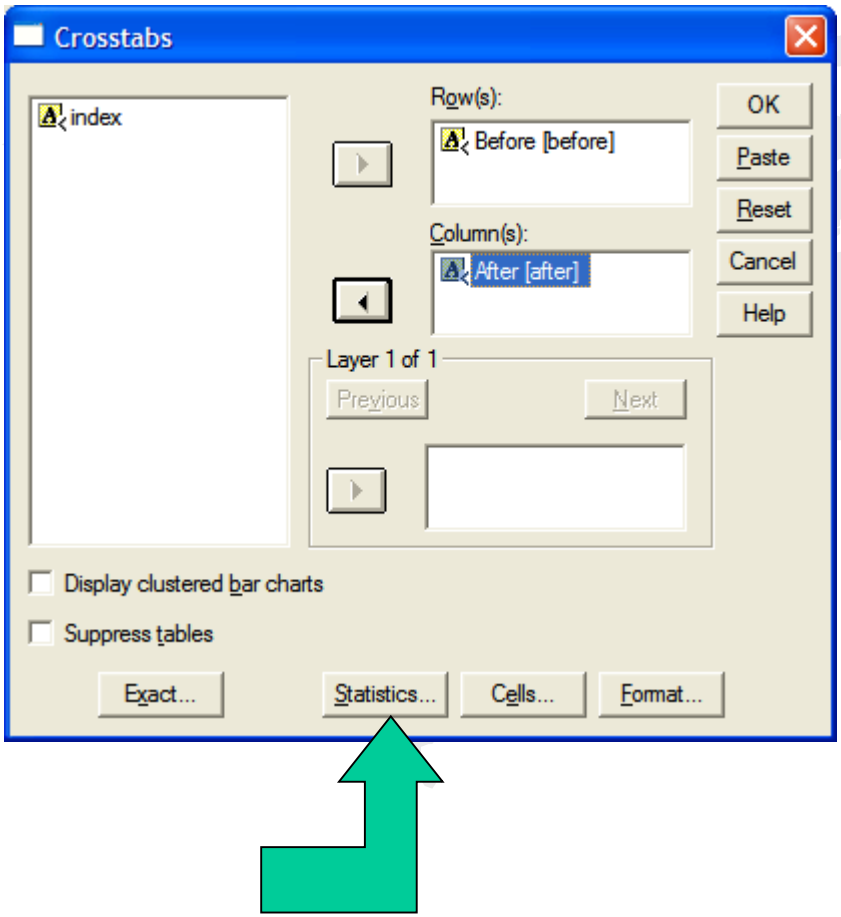
	index	before	after
1	001	Negative	Negative
2	002	Negative	Negative
3	003	Negative	Negative
4	004	Negative	Negative
5	005	Negative	Negative
6	006	Negative	Negative
7	007	Negative	Negative
8	008	Negative	Negative
9	009	Negative	Negative
10	010	Negative	Negative
11	011	Negative	Negative
12	012	Negative	Negative
13	013	Negative	Negative
14	014	Negative	Negative
15	015	Negative	Negative
16	016	Negative	Negative
17	017	Negative	Negative
18	018	Negative	Negative
19	019	Negative	Negative
20	020	Negative	Negative

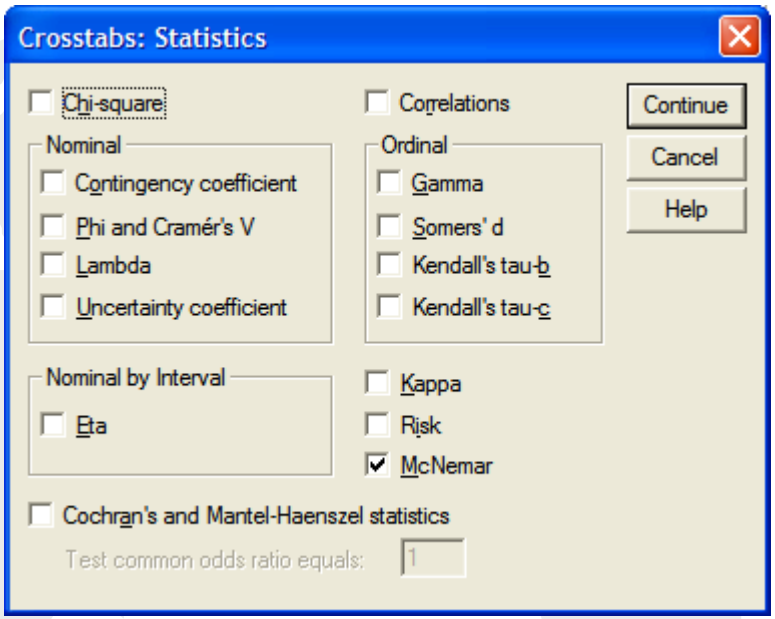
- Data should be entered in pairs as illustrated on the left here.

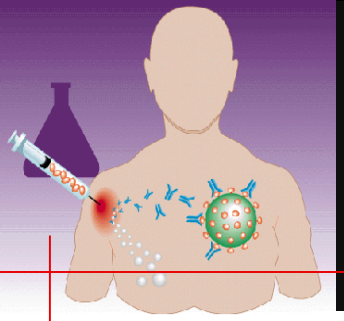
1970-2010 40 UKM
Mengilham Harapan
Mencipta Masa Depan

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.