Analysis of categorical data

Yongjua Laosiritaworn

Introductory on Field Epidemiology
6 July 2015, Thailand

Variable	Field Type	Format/Value	Special Info
ID	Number		
Ageyr	Number		age in year
case	Number	1 = case 2 = non case	Chikungunya case
Chikfam	Number	1 = yes 2 = no	Having Chikungunya case in the family
Chikvill	Number	1 = yes 2 = no	Having Chikungunya case in the village
Disfam	Number	1 = yes 2 = no	Family member had history of fever and arthralgia
Gender	Number	1 = Male 2 = Female	Gender
Workhr	Number	1 < 3 hr 2 = 3-5 hr 3 < 6 hr	working time at the rubber field (hour)
Religion	Number	1 = Buddhist 2 = others	Religion
wrubberf	Number	1 = yes 2 = no	Working in the rubber field

Questions ????

Using the Chikungunya data set: questions that you may have e.g.,

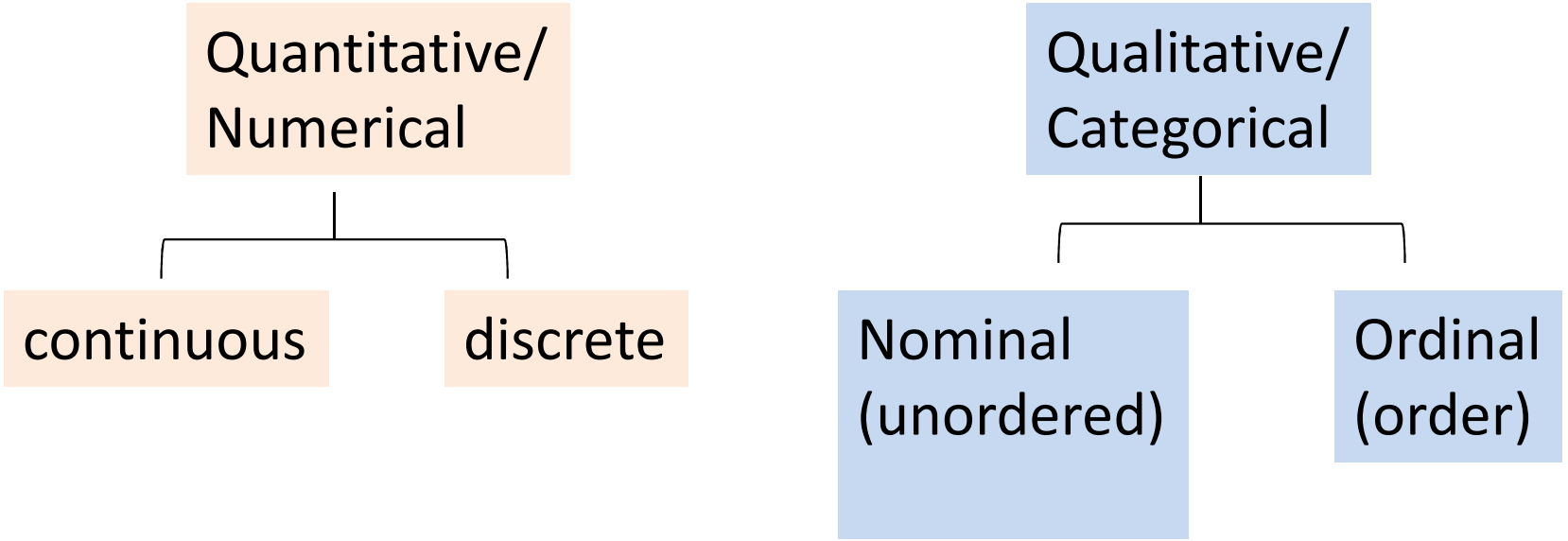
- Is the proportion of male in case and non-case similar?
- Is there evidence of relationship between having family member diagnosed as Chikungunya and the occurrence of Chikungunya fever?

Which statistical test would you use to answer these questions?

Reminder

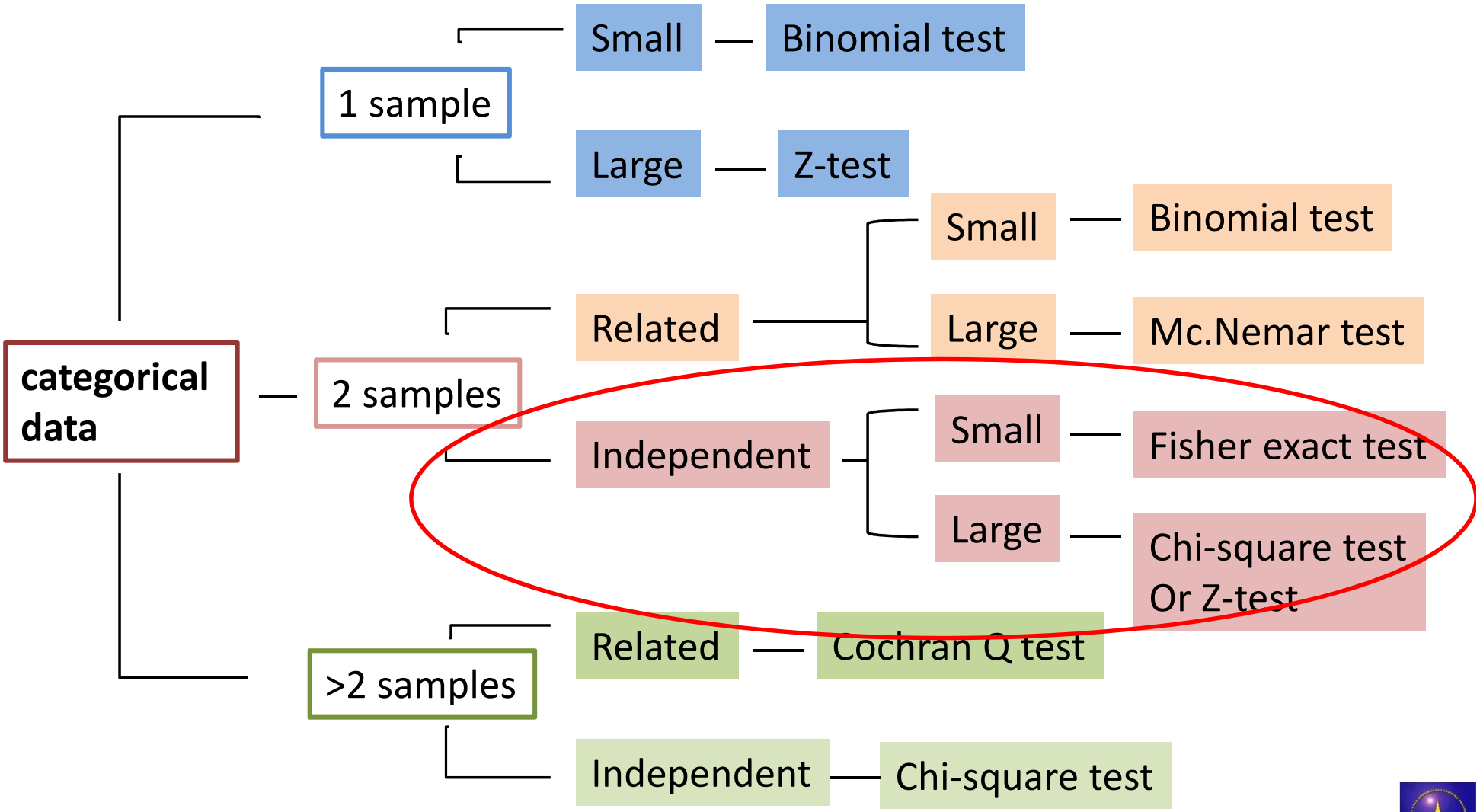
- What type of data these variables are?
  - Gender
  - Having family member diagnosed as Chikungunya

Reminder: data type



- Sometimes we may use numbers to label categories. For example, male=1 and female =2; white=1, black=2, other=3.
- This does not change the nature of the variable being categorical:
    * *they are still not computable*
    * *for unordered variables, they are still not ordered (consider* race)

Hypothesis Tests for Categorical Data



Introduction (1)

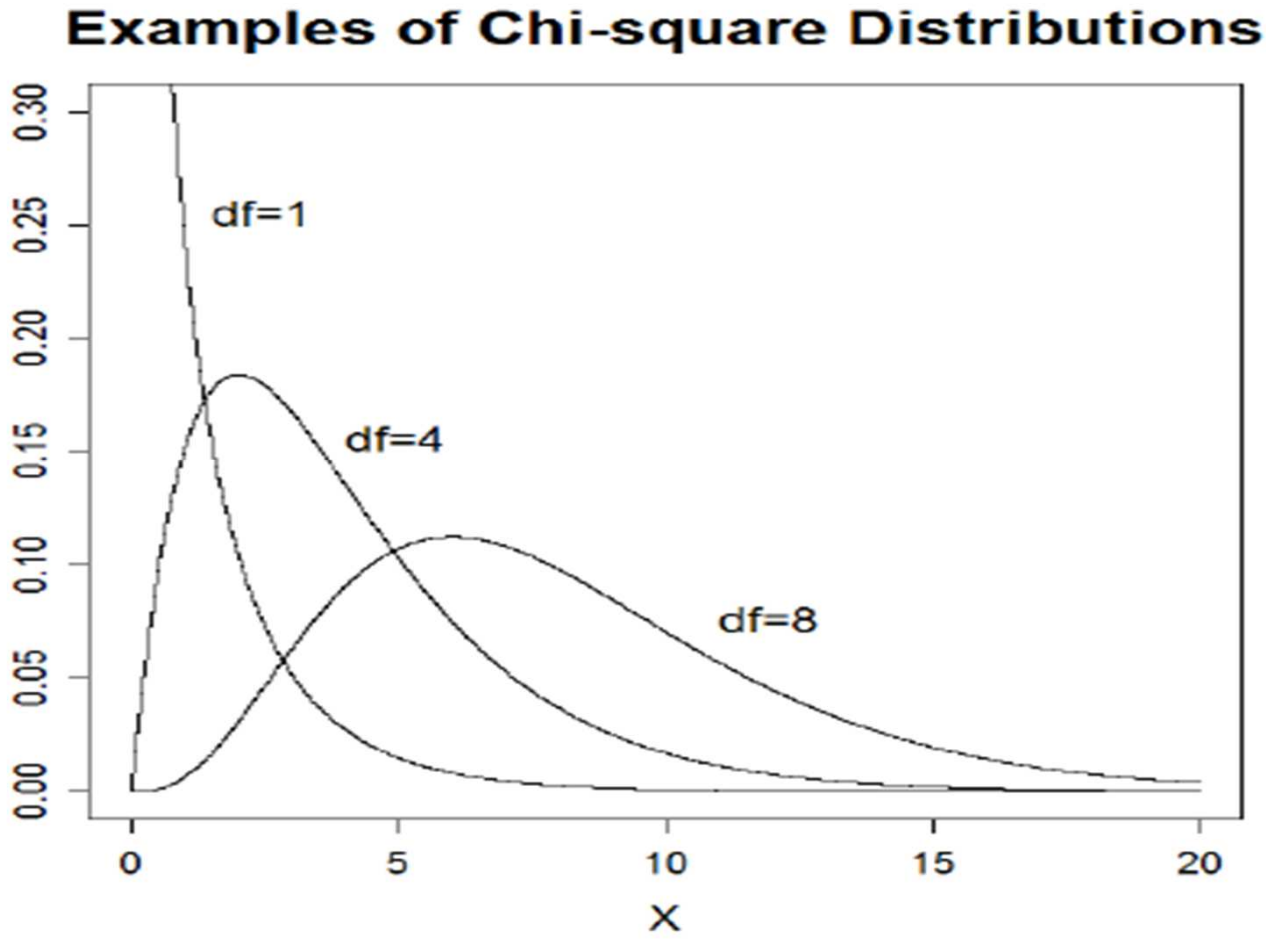
- We often have occasions to make comparisons between two characteristics of something to see if they are linked or related to each other.

- One way to do this is to work out what we would expect to find if there was **no relationship** between them (the usual null hypothesis) and what we actually **observe**.

Introduction (2)

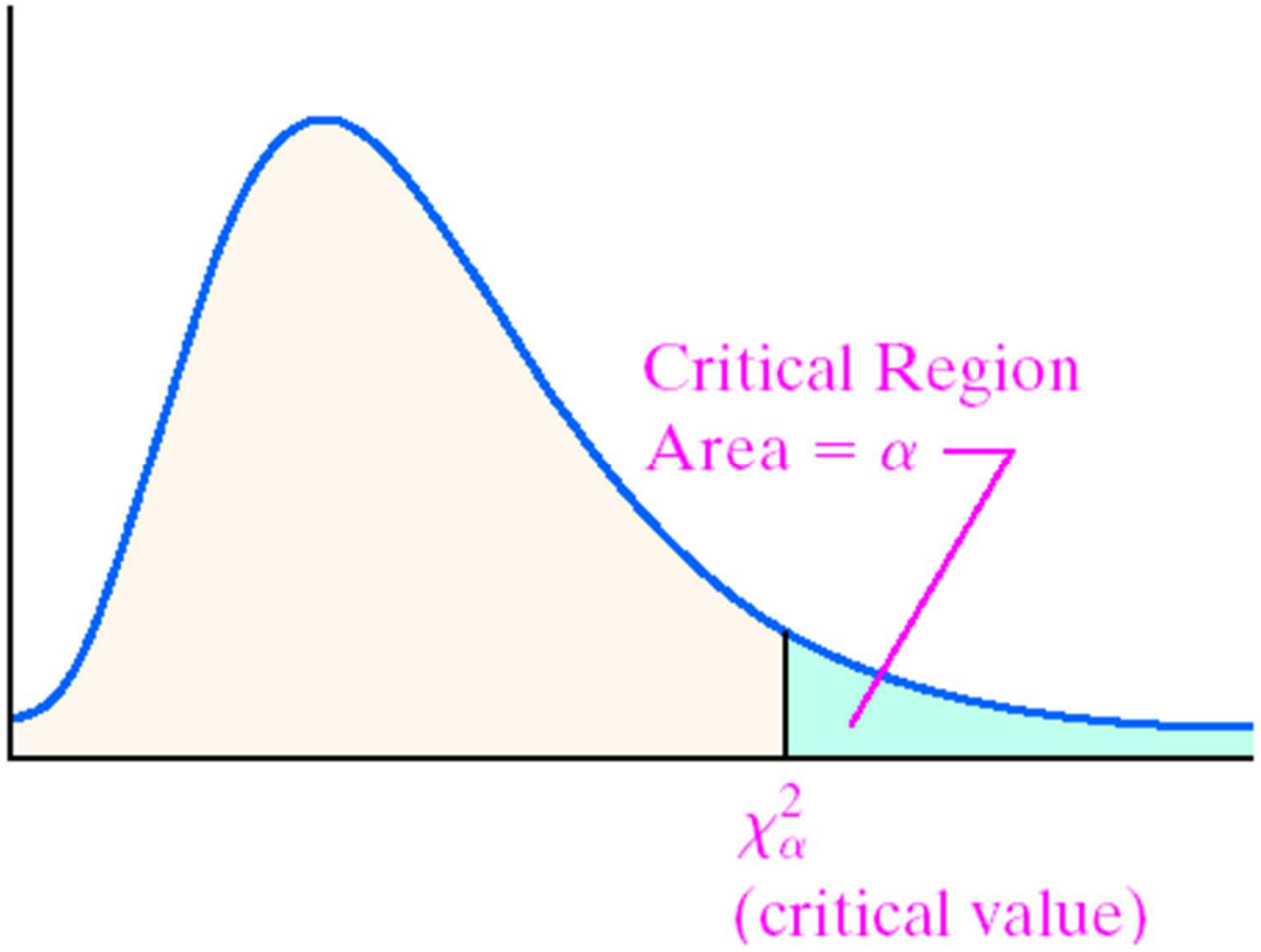
- The test we use to measure the differences between what is observed and what is expected according to an assumed hypothesis is called the **chi-square test (χ^2)**
- Instead of looking at differences between means, **chi-square** -> examine differences between frequencies
- **Chi-square test:** a hypothesis test in which the sampling distribution of the test statistic has a chi-square distribution when the null hypothesis is true

Chi-square distribution



Characteristic of Chi-Square distribution is asymmetry, have only positive value, skewed to the right, when df increase the distribution looks more normal

Chi-square distribution



Introduction (3)

- Chi-square test was first proposed by Karl Pearson -> most of the time known as (Pearson) Chi-square
- Different types of chi-square test have been applied:
  - only focus on Pearson's chi-square test for independence, Yate's corrected chi-square, Mantel-Haenzel chi-square, and Fisher exact

Assumption for chi-square test

- Random sample
- Independence: observations are always assumed to be independent of each other
- Sample size per cell: expected value is ≥ 5 in 80% of cells in larger tables, or no cells with zero≥ 5 in all cells of 2x2
- Sample size (as a whole) is large enough, otherwise will lead to an unacceptable type II error

Pearson's Chi–Square test for Independence/homogeniety

- Test hypothesis concerning whether there is association between a row variable and column variable in a contingency table **or**

  test whether different populations have the same proportion of individuals with some characteristic

  H_0 : $p_0 = p_1$

  H_a : $p_0 \neq p_1$

- If the two risks are the same, then the risk difference = 0 and both relative risk and odds ratios = 1

  H_0 : RR/ OR = 1

  H_a : RR/ OR $\neq$ 1

The idea behind chi-square test

- To compare actual counts to the counts we would expect if the null hypothesis were true (if the variables are independent)
- If a significant difference between the actual counts and expected counts exists -> reject null hypothesis

Expected Frequencies in a Chi-Square Independence Test

- Expected frequency in each cell =

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

2x2 contingency table

	D(+)	D(-)	Total
E(+)	a	b	a+b
E(-)	c	d	c+d
total	a+c	b+d	N

Expected value

	D(+)	D(-)	Total
E(+)	a	b	a+b
Expected	**(a+b)(a+c)/N**	**(a+b)(b+d)/N**	
E(-)	c	d	c+d
Expected	**(c+d)(a+c)/N**	**(c+d)(b+d)/N**	
total	a+c	b+d	N

Test Statistic for the Test of Independence

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where
-O_i represent the observed number of counts in the *i*th cell
- E_i represent the expected number of counts in the *i*th cell

follows the chi-square distribution with
$(r - 1)(c - 1)$ degrees of freedom (df),

where r is the number of rows and c is the number of columns in the contingency table provided

Exercise 1

- Is Gender associated with the occurrence of the disease?

Gender	1	2	TOTAL
1	54	312	366
Col %	49.5	53	52.4
2	55	277	332
Col %	50.5	47	47.6
TOTAL	109	589	698

Ex1. Calculate expected value

Gender	1	2	TOTAL
1	54	312	366
expected	57.15	308.85	
	= (366* 109)/698	= (366*589)/698	
2	55	277	332
expected	51.85	280.15	
	= (332*109)/698	= (332*589)/698	
TOTAL	109	589	698

Ex.1 Calculate the X^2 statistic and interpretation

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = $(54 - 57.15)^2 / (57.15)$ +

$(312 - 308.85)^2 / (308.85)$ +

$(55 - 51.85)^2 / (51.85)$ +

$(277 - 280.15)^2 / (280.15)$

χ^2 = 0.43

Example

CASE			
Gender	**1**	**2**	**TOTAL**
1	54	312	366
Row %	14.8	85.2	100.0
Col %	49.5	53.0	52.4
2	55	277	332
Row %	16.6	83.4	100.0
Col %	50.5	47.0	47.6
TOTAL	109	589	698
Row %	15.6	84.4	100.0
Col %	100.0	100.0	100.0

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	0.8717	0.5791	1.3120	(T)
Odds Ratio (MLE)	0.8719	0.5781	1.3144	(M)
		0.5668	1.3404	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	0.8906	0.6308	1.2573	(T)
Risk Difference (RD%)	-1.8122	-7.2154	3.5910	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	0.4338		0.5101115263
Chi-square - Mantel-Haenszel	0.4332		0.5104147302
Chi-square - corrected (Yates)	0.3072		0.5793934551
Mid-p exact		0.2560568964	
Fisher exact		0.2894921579	

Chi-square test for an RxC contingency table

- The same concept as 2x2 table
- For the case where we have *c different groups (columns), and we're checking each* group for different levels of the row factor - >
  - each column will have P(level 1), P(level 2), …,P(level r)
- Test whether the distributions are the same for each group (each column)

Example

CASE

Gender	1	2	TOTAL
1	54	312	366
Row %	14.8	85.2	100.0
Col %	49.5	53.0	52.4
2	55	277	332
Row %	16.6	83.4	100.0
Col %	50.5	47.0	47.6
TOTAL	109	589	698
Row %	15.6	84.4	100.0
Col %	100.0	100.0	100.0

Chi-square test for an RxC contingency table

- Test statistic

Test Statistics:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(\{r\text{-}1\}\{c\text{-}1\})$$

- Use this test when
  - No more than 20% of the cells have expected value < 5
  - No cell has expected value < 1

Exercise 2

- Is there an evidence that the distribution of working hour in the rubber field different between case and non case?
- Is there an evidence of association between working hour in the field and the disease?
- What is the null hypothesis?

H_0: The working hour in case and non case are equal or
there is no association between working hour and the disease

Ha: The working hour in case and non case are not equal or
there is an association between working hour and the disease

Exercise 2

- How would you interpret this result?

Forward

CASE

workhr	1	2	TOTAL
1	16	138	154
Row %	10.4	89.6	100.0
Col %	15.5	24.5	23.1
2	25	185	210
Row %	11.9	88.1	100.0
Col %	24.3	32.8	31.5
3	62	241	303
Row %	20.5	79.5	100.0
Col %	60.2	42.7	45.4
TOTAL	103	564	667
Row %	15.4	84.6	100.0
Col %	100.0	100.0	100.0

Single Table Analysis

Chi-square	df	Probability
10.8706	2	0.0044

Issues with chi-square test

- Sufficiently valid if no cell (2x2 table) has an expected value of less than 5
- The chi-square distribution only approximate estimates of the discrete probabilities associated with frequency data -> p-value based on Pearson's Chi-square will underestimate the true p-value

Yates-Corrected Chi-Square Test

- Corrected chi-square by subtract 0.5 from the absolute difference. The statistic is

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

- Yield more conservative p-value than the uncorrected version
- Continuity correction is not used for contingency tables larger than 2x2 (does not help in the approximation of the test statistic)

Fisher’s exact Test

- A statistical test used in the analysis of contingency table where sample size is small (particularly expected value <5)
- Fisher's exact test assumes that the row and column totals are known in advance (fix margin)
- A significance of the deviation from a null hypothesis can be calculated exactly

Fisher exact test

- From 2x2 table, Fisher Exact Test can be calculated by

	D(+)	D(-)	Total
E(+)	a	b	a+b
E(-)	c	d	c+d
total	a+c	b+d	N

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

- Fisher exact test give more conservative p-value -> requires more evidence than is necessary to reject null hypothesis

Mid p-values

- Use for small sample size
- Add half the probability of the observed result to the probability of the more extreme results
- Less conservative than Fisher exact
- For larger samples,
  - the p-value obtained from a test with Yates' correction will correspond to the Fisher approach,
  - the P value from the uncorrected test will correspond to the mid P value

Exercise 2

- Is there evidence of association between religion and the occurrence of Chikungunya ?
  - To consider which test would be used -> is any of the cell have an expected value <5?

	CASE		
Religion	**1**	**2**	**TOTAL**
1	105	570	675
Row %	15.6	84.4	100.0
Col %	96.3	96.8	96.7
2	4	19	23
Row %	17.4	82.6	100.0
Col %	3.7	3.2	3.3
TOTAL	109	589	698
Row %	15.6	84.4	100.0
Col %	100.0	100.0	100.0

Which statistical test would you use and how to interpret it

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	0.8750	0.2918	2.6238	(T)
Odds Ratio (MLE)	0.8752	0.3090	3.0558	(M)
		0.2830	3.6083	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	0.8944	0.3608	2.2174	(T)
Risk Difference (RD%)	-1.8357	-17.5659	13.8944	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	0.0569		0.8114946059
Chi-square - Mantel-Haenszel	0.0568		0.8116271477
Chi-square - corrected (Yates)	0.0029		0.9572877003
Mid-p exact		0.3889797608	
Fisher exact		0.4953752151	

Warning: The expected value of a cell is <5. Fisher Exact Test should be used.

Exercise 3

- Is there evidence of relationship between having family member diagnosed as Chikungunya and the occurrence of Chikungunya fever?

CASE			
Chikfam	**1**	**2**	**TOTAL**
1	21	72	93
Row %	22.6	77.4	100.0
Col %	20.4	12.9	14.0
2	82	487	569
Row %	14.4	85.6	100.0
Col %	79.6	87.1	86.0
TOTAL	103	559	662
Row %	15.6	84.4	100.0
Col %	100.0	100.0	100.0

Which statistical test would you use and how to interpret it?

Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	1.7322	1.0100	2.9709	(T)
Odds Ratio (MLE)	1.7306	0.9919	2.9440	(M)
		0.9566	3.0345	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	1.5669	1.0231	2.3998	(T)
Risk Difference (RD%)	8.1694	-0.8050	17.1438	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	4.0605		0.0438974760
Chi-square - Mantel-Haenszel	4.0544		0.0440572250
Chi-square - corrected (Yates)	3.4625		0.0627749100
Mid-p exact		0.0266430440	
Fisher exact		0.0351613476	

Mantel-Haenzel Method

- It is preferred when testing the significance of linear relationship between ordinal variables (more powerful than Pearson chi-square)
- Both variables must lie on an ordinal scale
- The test statistic can be calculated by

$$Q_{MH} = (n\text{-}1)r^2$$

where r^2 is the Pearson correlation between the row variable and the column variable

- should not be used with tables with small cell counts

http://esra.univ-paris1.fr/sashtml/stat/chap28/sect19.htm
http://faculty.chass.ncsu.edu/garson/PA765/chisq.htm

Mantel-Haenzel Method

- Apply when interest is comparing two groups in terms of a dichotomous outcome over several levels of a third variable (stratify analysis for confounding control)
- Only valid if the relative risks/ odds ratios are homogenous across strata
- The Mantel-Haenszel chi-square coefficient tests whether the common odds ratio across the k strata is 1.0, indicating no effect of the stratification variable

Question?

- If you are interested to see whether there is an association between age greater than 15 and age less than or equal to 15 with the occurrence of Chikungunya disease.
- Which statistic would you use to test this hypothesis?
- How would you operate to test this hypothesis?

Z – test for categorical data analysis

- When comparing proportion from the two samples:
  - p_1 from sample n1 & p_2 from sample n2
- Z – test can be applied to categorical data **but *only when the normal approximation to the binomial distribution is valid*** for each of the two samples
  - $n_1\hat{p}\hat{q} \geq 5$ and $n_2\hat{p}\hat{q} \geq 5$
  - where $\hat{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2) = (x_1+x_2)/(n_1 + n_2)$ and $\hat{q} = 1 - \hat{p}$

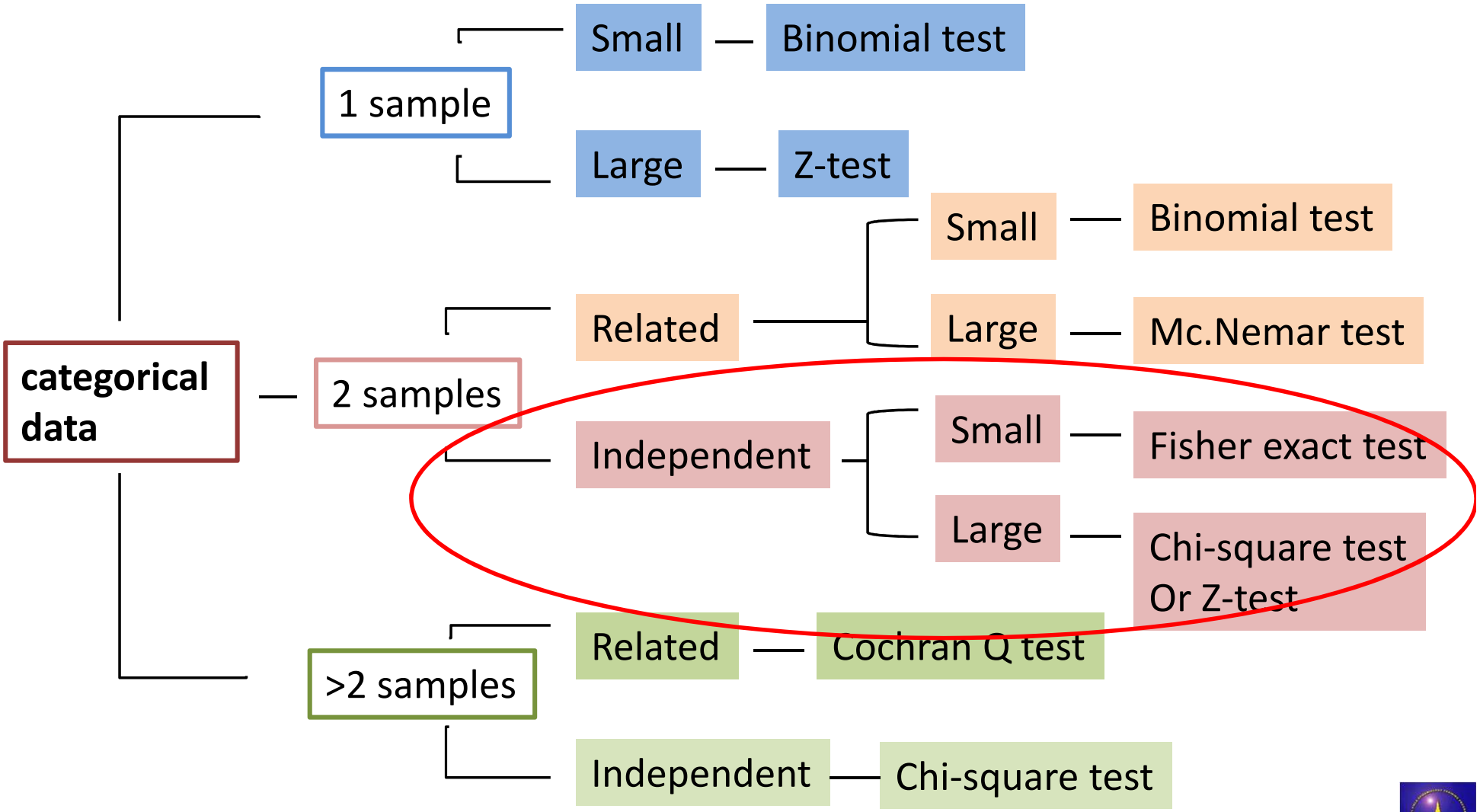
Z – test for categorical data analysis

- The test statistic can be calculated as:

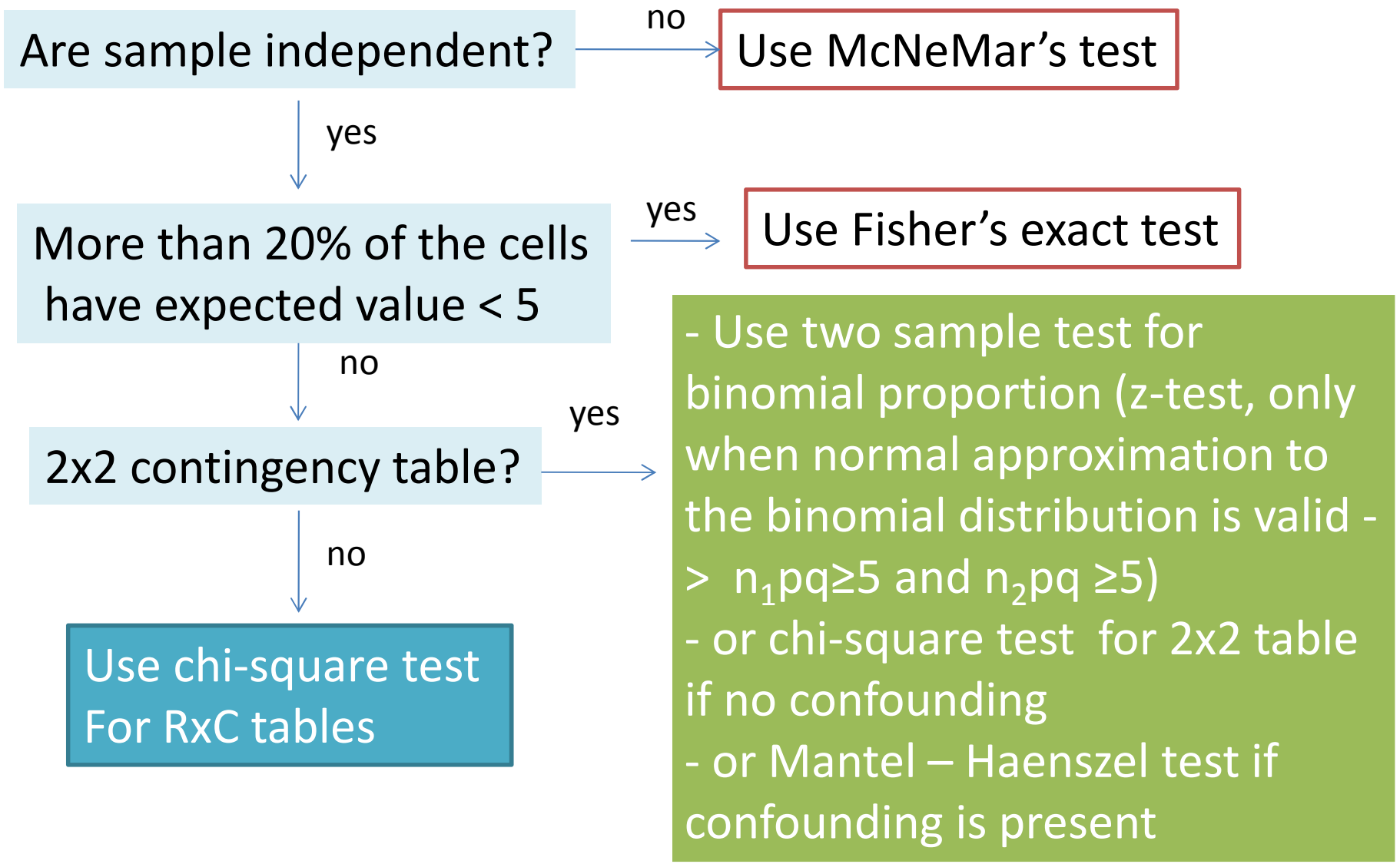
$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})}\sqrt{(1/n_1)+(1/n_2)}} \text{ where } \widehat{p} = \frac{X+Y}{n_1+n_2}.$$

- When the normal approximation is met, Z-test yield the same p-value as chi-square test
- Giving the gender and disease example, can we also use Z-test to test the hypothesis?
- Giving the religion and disease example, can we also use Z-test to test the hypothesis?

Hypothesis Tests for Categorical Data



Conclusion: flow chart for appropriate methods of statistical inference for categorical data



Chi-square test for trend

- Tests whether there is a *linear* trend between row (or column) number and the fraction of subjects in the left column (or top row)
- Provides a more powerful test than the unordered independence test above
- Use this test only if $npq \geq 5$,
  - where p = number of cases/total number of the population and q = 1-p

Chi-square test for trend

- Is the longer we stay at home could prevent the occurrence of Chikungunya?
- To answer this question, should answer the question “whether the proportion of home stay difference between case and control or not”

Forward

CASE

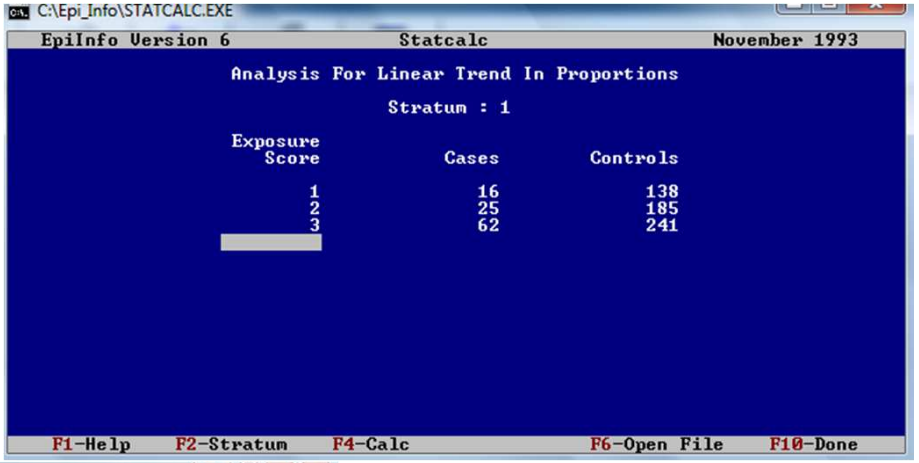
workhr	1	2	TOTAL
1	16	138	154
Row %	10.4	89.6	100.0
Col %	15.5	24.5	23.1
2	25	185	210
Row %	11.9	88.1	100.0
Col %	24.3	32.8	31.5
3	62	241	303
Row %	20.5	79.5	100.0
Col %	60.2	42.7	45.4
TOTAL	103	564	667
Row %	15.4	84.6	100.0
Col %	100.0	100.0	100.0

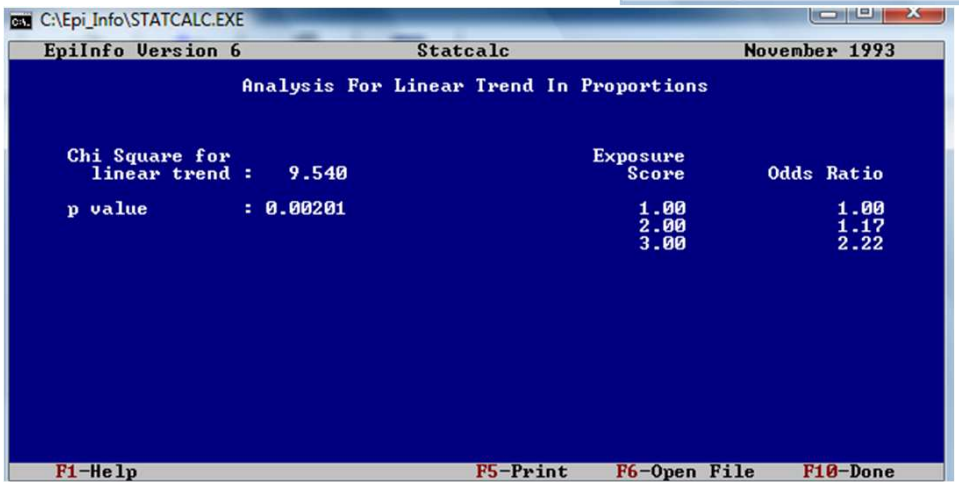
Single Table Analysis

Chi-square	df	Probability
10.8706	2	0.0044

In Epi-info

- Chi-square for trend can be calculated in statcalc





Example: Chi-square test for trend

	Successes	Failures	Total	Per cent
Observed	19	497	516	3.68
Expected	26.58	489.42		
Observed	29	560	589	4.92
Expected	30.33	558.67		
Observed	24	269	293	8.19
Expected	15.09	277.91		
Total	72	1326	1398	5.15

Total Chi^2 = 7.884843, (2 DF), P = .0194

Chi^2 for linear trend = 7.19275, (1 DF), P = .0073

Remaining Chi^2 (non-linearity) = 0.692093, (1 DF), P = .4055

Exercise 4: Mantel Haenzel stratify chi-square

- We are wondering whether gender is a confounder, so the question is *" is there evidence of association between having family member diagnosed as Chikungunya and the occurrence of disease after controlling for gender?"*

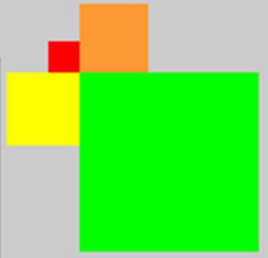
Stratify by gender (Gender=1)

Chikfam : case, Gender=1

Forward

CASE

Chikfam	1	2	TOTAL
1	8	37	45
Row %	17.8	82.2	100.0
Col %	15.7	12.4	12.9
2	43	261	304
Row %	14.1	85.9	100.0
Col %	84.3	87.6	87.1
TOTAL	51	298	349
Row %	14.6	85.4	100.0
Col %	100.0	100.0	100.0



Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	1.3124	0.5725	3.0082	(T)
Odds Ratio (MLE)	1.3113	0.5390	2.9338	(M)
		0.4937	3.1179	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	1.2568	0.6325	2.4975	(T)
Risk Difference (RD%)	3.6330	-8.2047	15.4708	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	0.4146		0.5196269899
Chi-square - Mantel-Haenszel	0.4134		0.5202258435
Chi-square - corrected (Yates)	0.1746		0.6760671897
Mid-p exact		0.2577813538	
Fisher exact		0.3263236233	

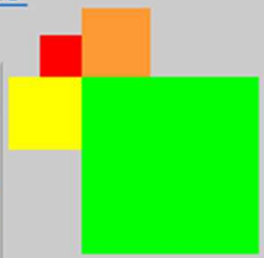
Stratify by gender (Gender=2)

Chikfam : case, Gender=2

Back Forward Current Procedure

CASE

Chikfam	1	2	TOTAL
1	13	35	48
Row %	27.1	72.9	100.0
Col %	25.0	13.4	15.3
2	39	226	265
Row %	14.7	85.3	100.0
Col %	75.0	86.6	84.7
TOTAL	52	261	313
Row %	16.6	83.4	100.0
Col %	100.0	100.0	100.0



Single Table Analysis

	Point Estimate	95% Confidence Interval Lower	Upper	
PARAMETERS: Odds-based				
Odds Ratio (cross product)	2.1524	1.0460	4.4292	(T)
Odds Ratio (MLE)	2.1462	1.0150	4.3860	(M)
		0.9542	4.6193	(F)
PARAMETERS: Risk-based				
Risk Ratio (RR)	1.8403	1.0647	3.1809	(T)
Risk Difference (RD%)	12.3664	-0.9094	25.6421	(T)

(T=Taylor series; C=Cornfield; M=Mid-P; F=Fisher Exact)

STATISTICAL TESTS	Chi-square	1-tailed p	2-tailed p
Chi-square - uncorrected	4.4861		0.0341722923
Chi-square - Mantel-Haenszel	4.4718		0.0344600699
Chi-square - corrected (Yates)	3.6379		0.0564804335
Mid-p exact		0.0229071317	
Fisher exact		0.0327204930	

Summary

Parameters	Point Estimate	95%Confidence Interval Lower	Upper	
Odds Ratio Estimates				
Crude OR (cross product)	1.7322	1.0100,	2.9709	(T)
Crude (MLE)	1.7306	0.9919,	2.9440	(M)
		0.9566,	3.0345	(F)
Adjusted OR (MH)	1.7231	1.0037,	2.9582	(R)
Adjusted OR (MLE)	1.7177	0.9844,	2.9220	(M)
		0.9494,	3.0117	(F)
Risk Ratios (RR)				
Crude Risk Ratio (RR)	1.5669	1.0231,	2.3998	
Adjusted RR (MH)	1.5596	1.0167,	2.3925	

(T=Taylor series; R=RGB; M=Exact mid-P; F=Fisher exact)

STATISTICAL TESTS (overall association)	Chi-square	1-tailed p	2-tailed p
MH Chi-square - uncorrected	3.9418		0.0471
MH Chi-square - corrected	3.3543		0.0670
Mid-p exact		0.0282	
Fisher exact		0.0372	
In the following two tests, low p values suggest that ratios differ by stratum			
Chi-square for differing Odds Ratios by stratum (interaction)	0.7778		0.3778