MATU9D2: PRACTICAL STATISTICS

Spring 2017

PRACTICAL SESSION 4

- Chi-squared Tests (Independence & Goodness of Fit)
  CI's for Proportions

- Handout 2 of 2

Computing Science & Maths
University of Stirling

Analysis of Contingency Tables

1. Chi-Square Test of Association

This test examines ‘survey data’ i.e. categorical answer to questions often gathered by questionnaire. For example, the answers may be ‘ Vote Conservative’, ‘Vote Labour’, ‘Vote Liberal Democrat’, ‘Vote SNP’,’ Vote Green’, ‘Vote other’ in answer to a question about voting intention at the next election.

The test examines the association between two sets of data e.g. voting intentions as above and gender.

H_o : The answers are independent i.e. no association between the answers

H_1 : The answers are related i.e. an association between the answers

Enter the frequencies into a series of columns

e.g. The Voting Intention and Gender data could be entered into 6 columns , each with two rows OR two columns each with six rows i.e. a column / row for each answer .

Example

The Voting Intention & Social Class data from Lecture notes gave the following table.

		Tory	Labour	LibDem	SNP	Green	Other
Class	1	7	5	3	9	4	8
	2	13	11	5	13	13	6
	3	8	15	9	17	5	8
	4	7	8	9	7	7	3

Does this show a statistically significant relationship between Voting Intention & Social Class?

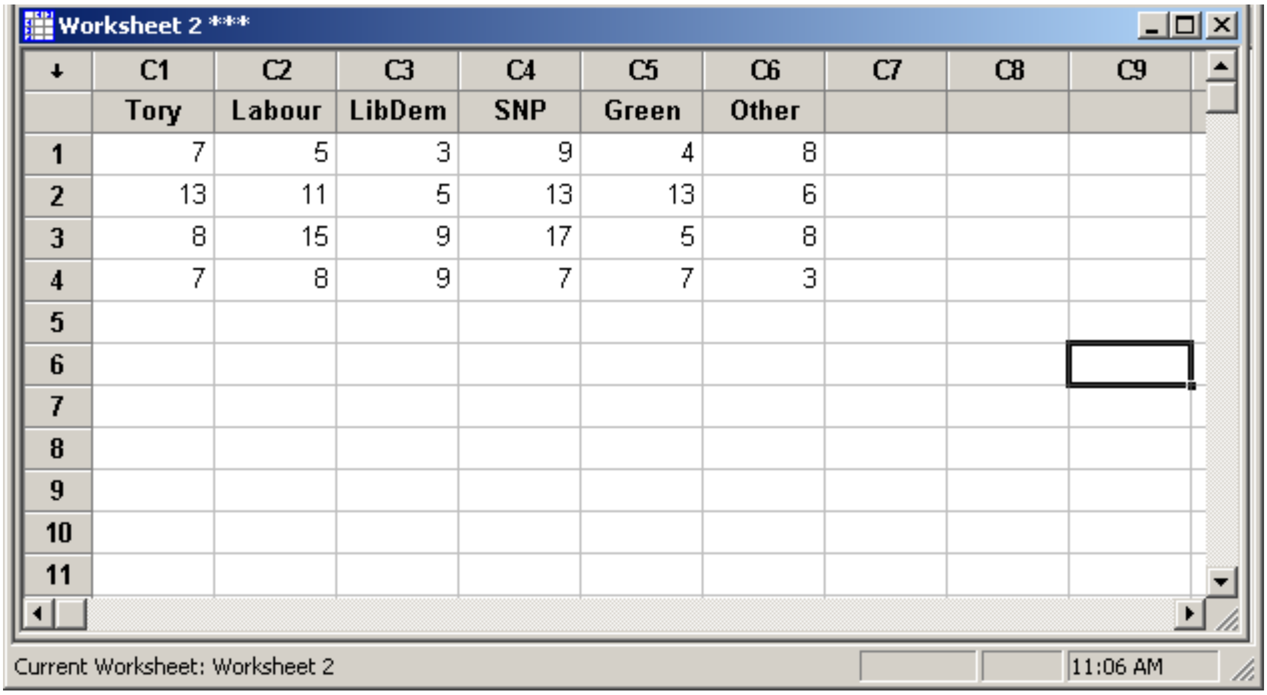
The assumptions that X^2 approximates to χ^2 is not valid if the cell frequencies are too small. A useful rule is :

df = 1, then no cell can have an expected frequency less than 5.

df > 1, then no more than 20% of the cells can have an expected frequency of less than 5, and no cell an expected frequency of less than 1.

(i) If you have <u>already calculated</u> the Observed Frequencies.

Step 1. Enter the frequencies into a group of columns. For Example



↓	C1	C2	C3	C4	C5	C6	C7	C8	C9
	Tory	Labour	LibDem	SNP	Green	Other			
1	7	5	3	9	4	8			
2	13	11	5	13	13	6			
3	8	15	9	17	5	8			
4	7	8	9	7	7	3			
5									
6									
7									
8									
9									
10									
11									

Choose	Stat Menu
Click on	Tables
Click on	Chisquare Test (Table in Worksheet)

Enter the column containing the frequencies into the Box e.g. C1-C6

Click on OK

(ii) If you have <u>not calculated</u> the Observed Frequencies.

i.e. For the Voting Intention data we have 200 rows of data in 2 columns with one column having Voting Intention and the other having the Social Class

Choose	Stat Menu
Click on	Tables
Click on	Cross Tabulation and Chisquared

Enter the column containing the Classification Variables as rows and columns

Click on the Box to select Chisquare Analysis and Expected Values

Click on OK

```
Example Output    Levels for 'Class' (1-4) and Party (1-6) in columns 4 and 5
                  respectively.
```

```
MTB > Table C4 C5;
SUBC>   Counts;
SUBC>   ChiSquare 3.
```

Tabulated Statistics

```
 Rows: Class     Columns: Party

           1        2        3        4        5        6      All

 1         7        5        3        9        4        8       36
        6.30     7.02     4.68     8.28     5.22     4.50    36.00
        0.28    -0.76    -0.78     0.25    -0.53     1.65      --

 2        13       11        5       13       13        6       61
       10.67    11.90     7.93    14.03     8.84     7.62    61.00
        0.71    -0.26    -1.04    -0.27     1.40    -0.59      --

 3         8       15        9       17        5        8       62
       10.85    12.09     8.06    14.26     8.99     7.75    62.00
       -0.87     0.84     0.33     0.73    -1.33     0.09      --

 4         7        8        9        7        7        3       41
        7.17     8.00     5.33     9.43     5.94     5.12    41.00
       -0.07     0.00     1.59    -0.79     0.43    -0.94      --

 All      35       39       26       46       29       25      200
       35.00    39.00    26.00    46.00    29.00    25.00   200.00
          --       --       --       --       --       --      --

Chi-Square = 16.452, DF = 15, P-Value = 0.353
2 cells with expected counts less than 5.0


  Cell Contents --
                  Count
                  Exp Freq
                  St Resid
```

2. <u>Bar Graphs for the Contingency Table Examples</u>

(i) If you have <u>already calculated</u> the Observed Frequencies.

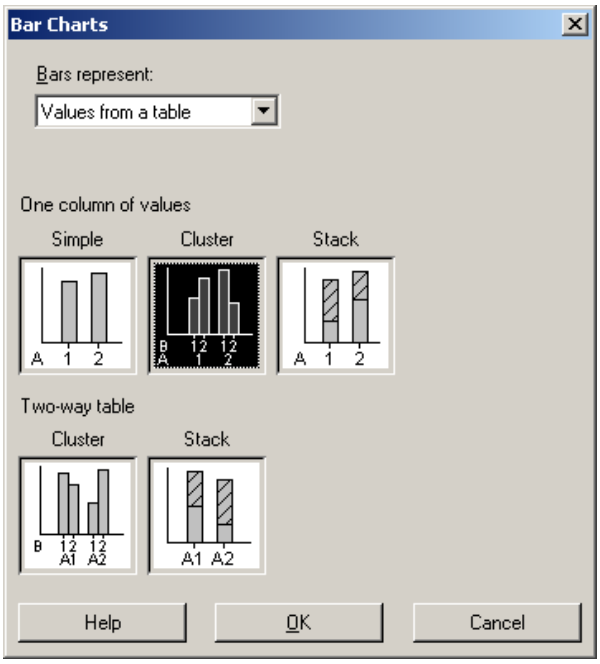
Step 1. Enter the Frequencies in a Column eg. Freq
Step 2. Enter code for 1st variable into a second column e.g. Drug
Step 3. Enter code for 2nd variable into a third column e.g. Surv

Drug	Surv	Freq
1	1	41
1	2	216
2	1	64
2	2	180

Step 4. Use the following menus:

Graph -> Bar Chart

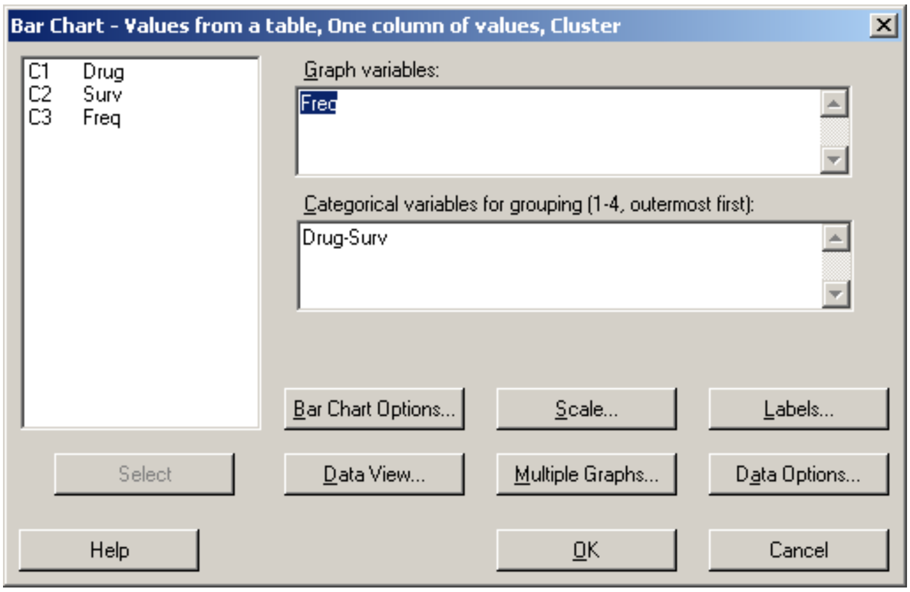
Choose Bars represent Values from a Table and a Cluster Bar Chart (see below). Click OK.



Bar Charts

Bars represent: Values from a table

The following dialogue box appears : Choose Graph variables as column with frequencies a and columns with the group codes as the Categorical Variables. Click OK.



Bar Chart - Values from a table, One column of values, Cluster

C1	Drug
C2	Surv
C3	Freq

Graph variables: Freq

Categorical variables for grouping (1-4, outermost first): Drug-Surv

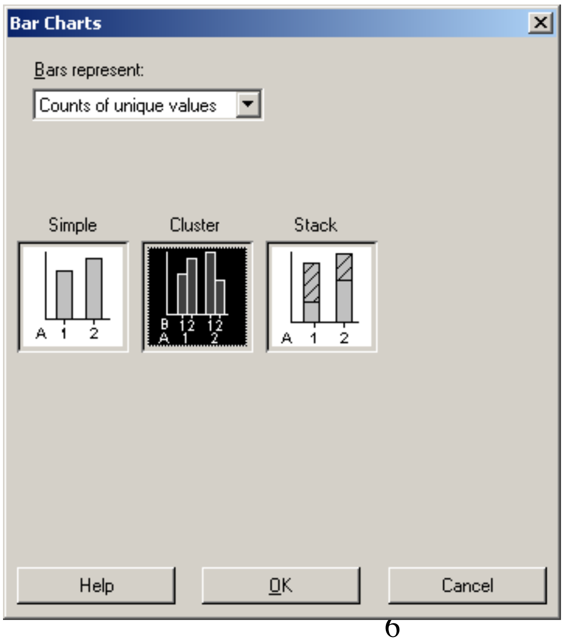
(ii) If you have <u>not calculated</u> the Observed Frequencies.

Step 1. Enter 1st variable into one column. For example, Treatment in one column
Step 2. Enter 2nd variable into another column, For example, Survival in a column

Step 3. Use the following menus:

Graph -> Bar Chart

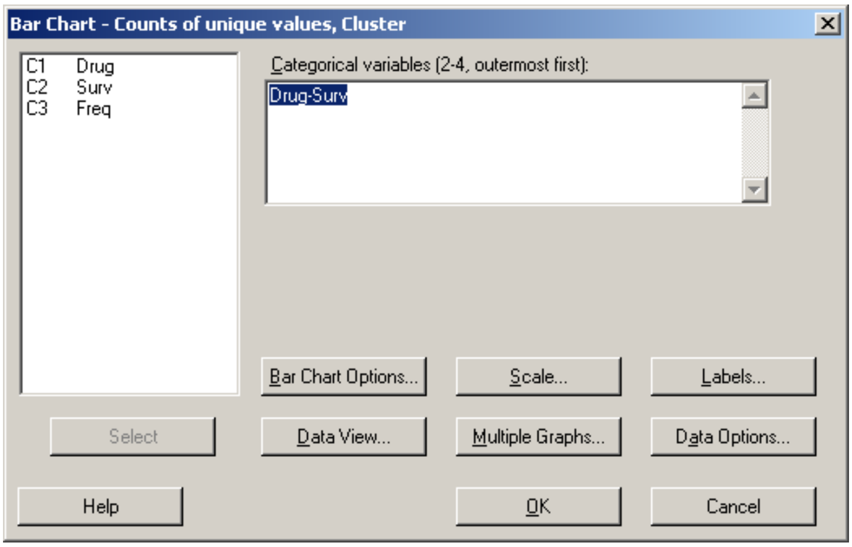
Choose Bars represent Counts of Unique Values and a Cluster Bar Chart (see below). Click OK.



Bar Charts

Bars represent: Counts of unique values

The following dialogue box appears : Choose Graph variables as column with frequencies and columns with the group codes as the Categorical Variables. Click OK.


Bar Chart - Counts of unique values, Cluster

C1 Drug
C2 Surv
C3 Freq

Categorical variables (2-4, outermost first): Drug-Surv

Confidence Intervals & Tests for Proportions

1. One Sample

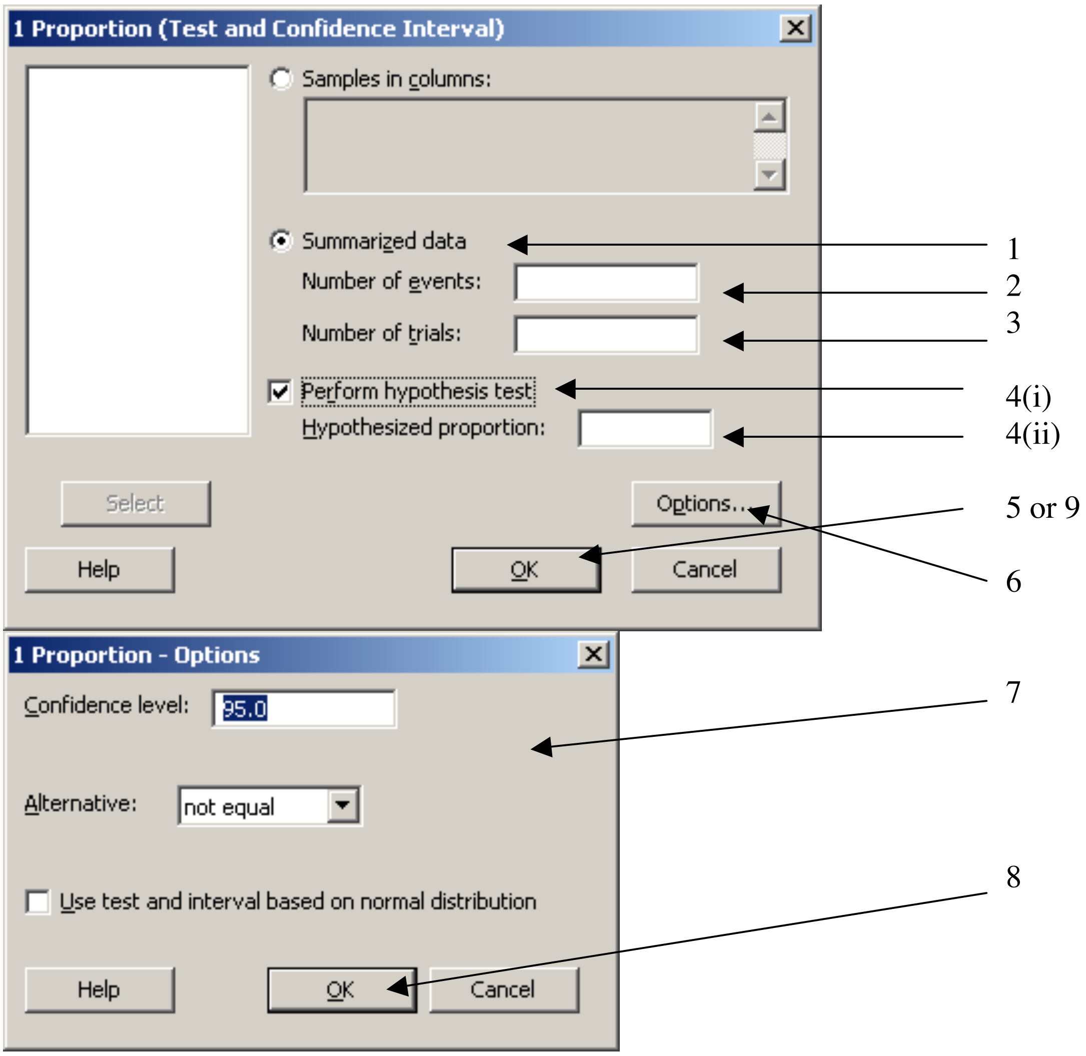
Null hypothesis H_0 : $\theta = \theta_o$ against

Alternative H_1 : $\theta \neq \theta_o$ or $\theta > \theta_o$ or $\theta < \theta_o$

Access the Stat Menu -> Basic Statistics -> 1 Proportion

In the dialogue box,

1. If you only have sample size and number of successes – click Summarised data
2. Enter the Number of events
3. Enter Number of Trials
4. If you want to perform a test – click to get a tick in this box
   Then : Enter the claimed proportion
5. If you do not want to change the Confidence Level or H_1 – Click OK
6. If you do want to change the Confidence Level or H_1 – Click Options - the second box appears
7. Change Confidence Level or Alternative Hypothesis
8. Click OK
9. Click OK



1 Proportion (Test and Confidence Interval)

○ Samples in columns:

● Summarized data

Number of events:

Number of trials:

☑ Perform hypothesis test

Hypothesized proportion:

1 Proportion - Options

Confidence level: 95.0

Alternative: not equal

□ Use test and interval based on normal distribution

1. **<u>One Sample Proportion</u>** **<u>Example Output</u>**

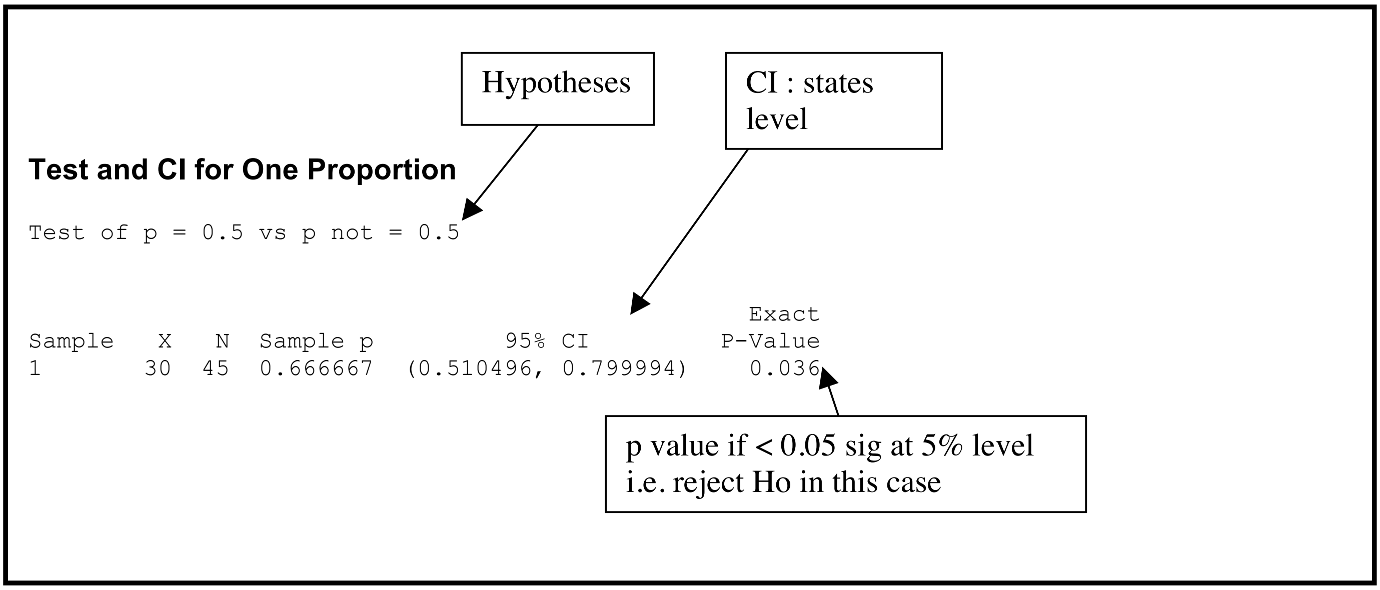
$H_o: \quad \theta = 0.5$

$H_1: \quad \theta \neq 0.5$

From the output below:

$p<0.05$ and 95% CI for θ is (0.51, 0.80)

<u>Conclusion</u> : Sufficient evidence to reject Ho in favour of H1 at 5% level since $p<0.05$. Same conclusion using the 95% CI for θ i.e. since 95% CI for θ does not include 0.5, sufficient at 5% significance level to conclude that the population proportion is significantly different to 0.5.



Test and CI for One Proportion

```
Test of p = 0.5 vs p not = 0.5


                                                          Exact
Sample   X   N  Sample p        95% CI          P-Value
1       30  45  0.666667  (0.510496, 0.799994)    0.036
```

Hypotheses

CI : states level

p value if < 0.05 sig at 5% level
i.e. reject Ho in this case

2. **Two Independent Samples**

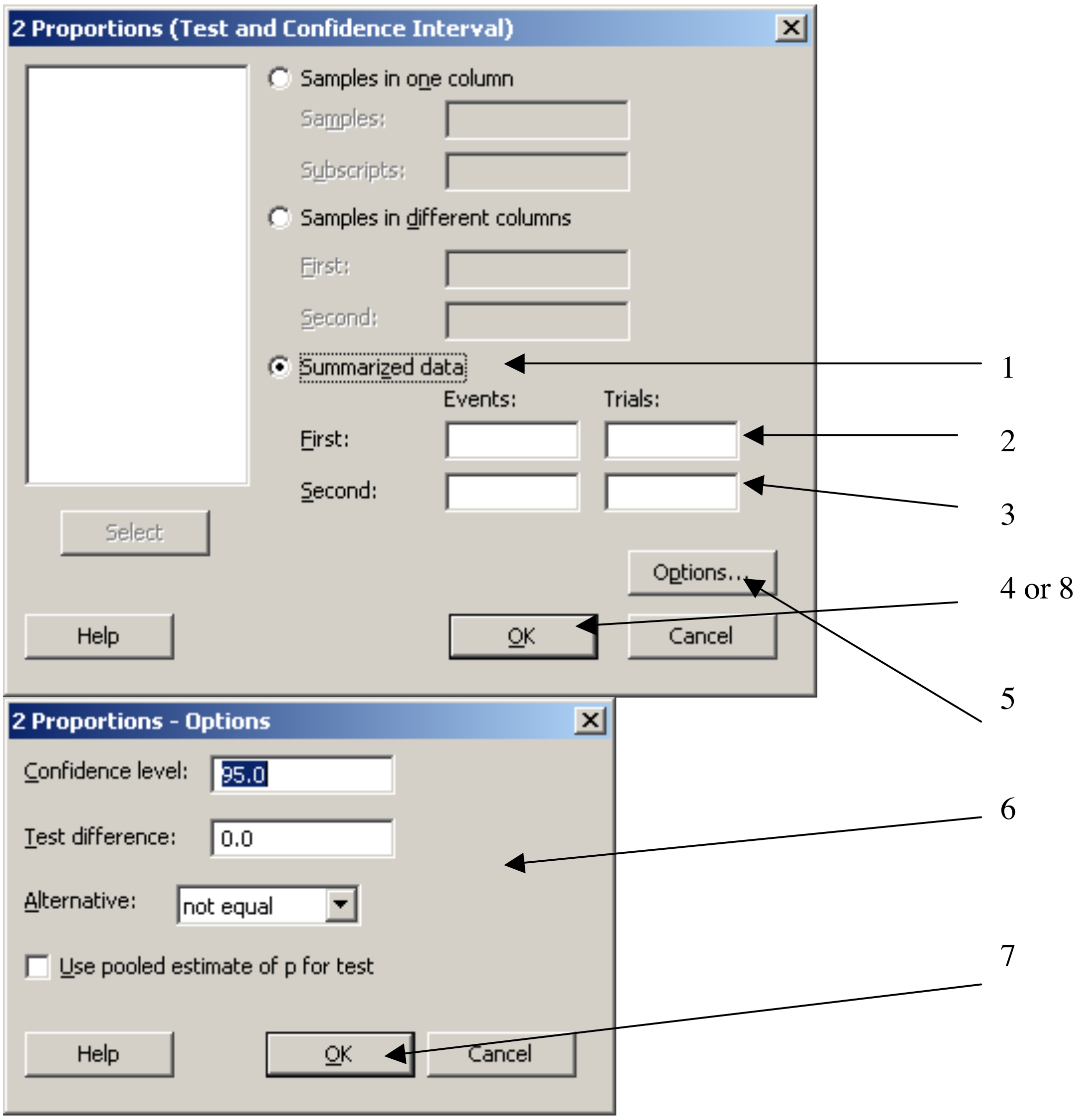
Null hypothesis H$_0$: $\theta_1 = \theta_2$ against

Alternative H$_1$: $\theta_1 \neq \theta_2$ or $\theta_1 > \theta_2$ or $\theta_1 < \theta_2$

Access the Stat Menu -> Basic Statistics -> 2 Proportions

In the dialogue box,

1. If you only have sample sizes and number of successes – click Summarised data
2. Enter the Number of events & Trials for Sample 1
3. Enter Number of Events & Trials for Sample 2
4. If you do not want to change the Confidence Level or H$_1$ – Click OK
5. If you do want to change the Confidence Level or H$_1$ – Click Options - the second box appears
6. Change Confidence Level or Alternative Hypothesis
7. Click OK
8. Click OK



2 Proportions (Test and Confidence Interval)

○ Samples in one column

○ Samples in different columns

● Summarized data

2 Proportions - Options

Confidence level: 95.0

Test difference: 0.0

Alternative: not equal

□ Use pooled estimate of p for test

3. Goodness of Fit Test

This test compares observed frequencies with the frequencies expected if the data follow a specified distribution.

In some cases i.e. the Binomial or Uniform ‘discrete’ distributions the calculation of expected frequencies is relatively simple.

Example

A random sample of 500 units is taken from each day’s production and inspected for defective units. The number of defectives in the last working week were as follows :

Day	Number of defectives
Monday	15
Tuesday	6
Wednesday	3
Thursday	5

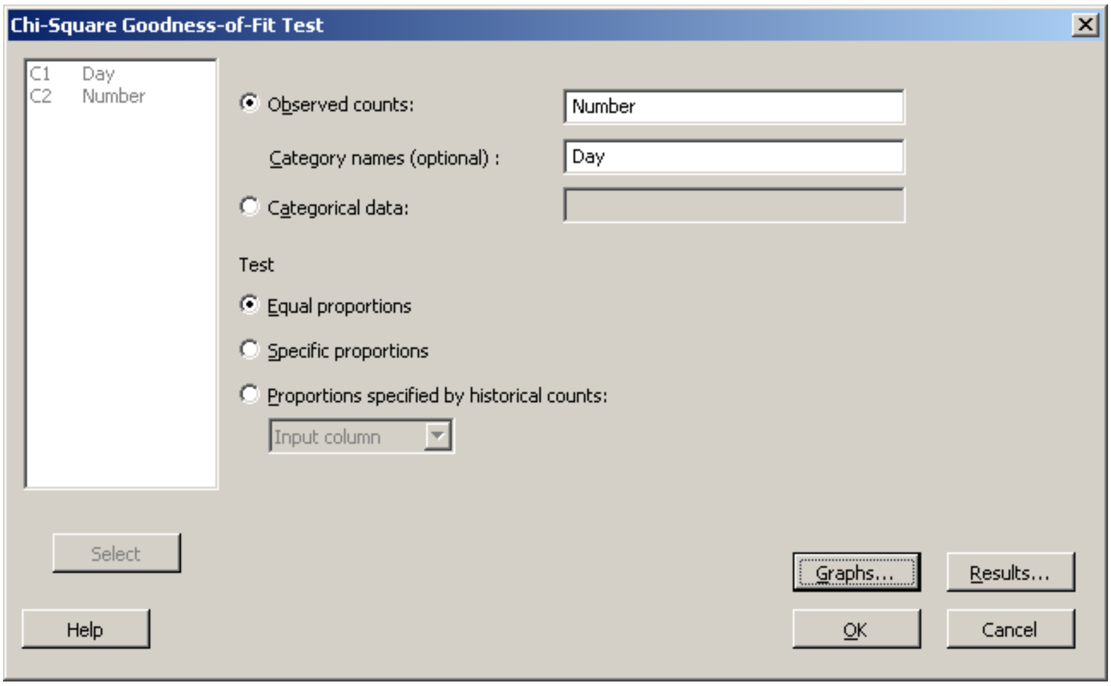
Friday	15

Test the hypothesis that the difference between the days is due to chance.

Comments

This translates into

Ho : Data comes from a Uniform Distribution
H1 : Data comes from some other Distribution



Chi-Square Goodness-of-Fit Test

C1 Day
C2 Number

● Observed counts: Number

Category names (optional) : Day

○ Categorical data:

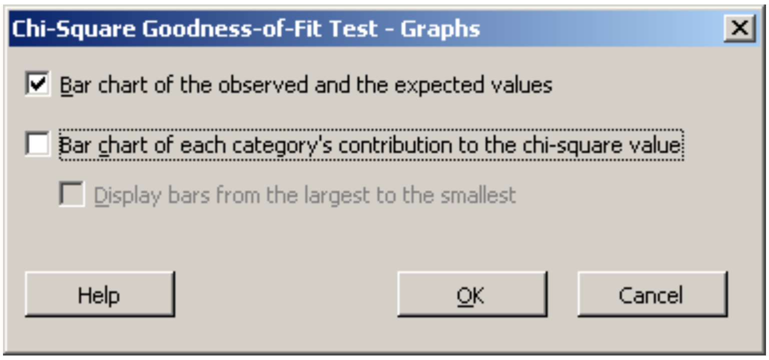
Test

● Equal proportions

○ Specific proportions

○ Proportions specified by historical counts:

Input column



Chi-Square Goodness-of-Fit Test - Graphs

☑ Bar chart of the observed and the expected values

□ Bar chart of each category's contribution to the chi-square value

□ Display bars from the largest to the smallest

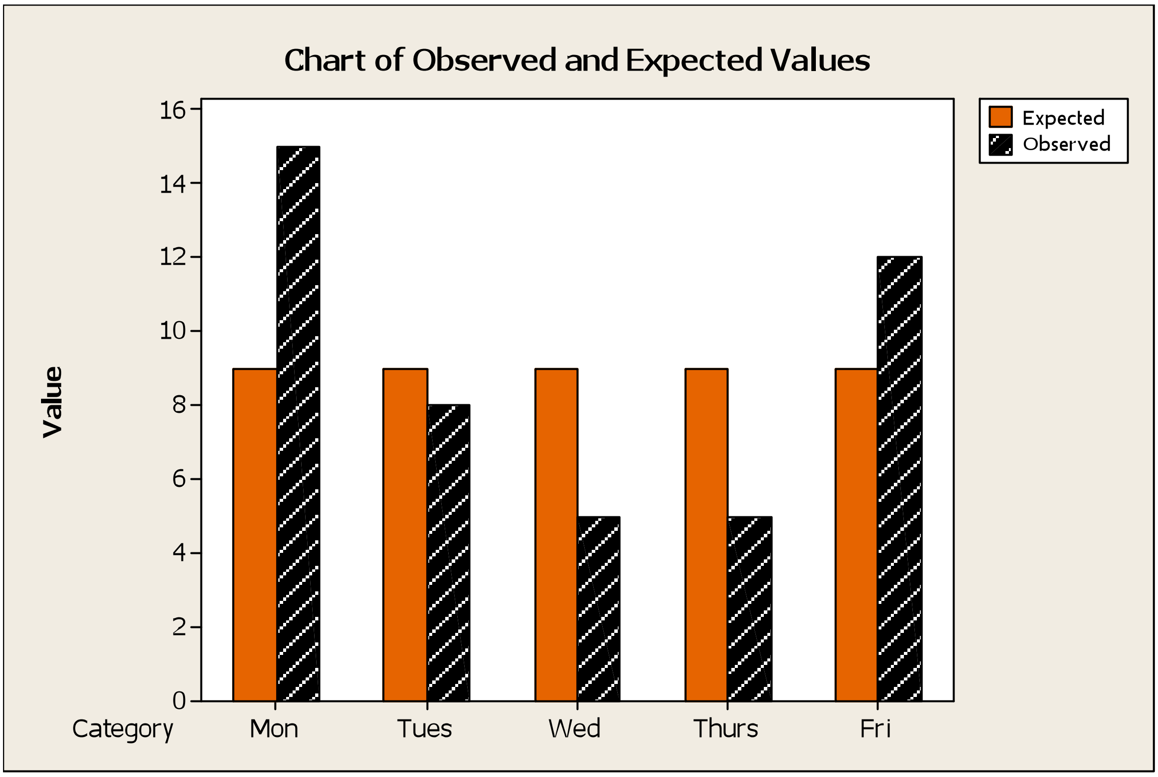
3. Goodness of Fit : Example Output

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Number

```
Using category names in Day


                              Test             Contribution
Category  Observed  Proportion  Expected    to Chi-Sq
Mon             15         0.2         9      4.00000
Tues             8         0.2         9      0.11111
Wed              5         0.2         9      1.77778
Thurs            5         0.2         9      1.77778
Fri             12         0.2         9      1.00000


 N  DF    Chi-Sq  P-Value
45   4  8.66667    0.070
```



EXERCISES

1. The following results were obtained from a study to investigate the viewing habits of people living in Scotland. The responses examined below is the number of hours of television watched each week by an individual and their age and gender. Number of hours viewing is coded as either less than 21 hours or 21 greater than 21 hours per week and Age is coded as less than 18, between 18 and 50, and over 50.

MALE		Age		
		< 18	18 - 50	> 50
	< 21hours	10	20	9
Viewing	> 21 hours	10	4	11

FEMALE			Age	
		< 18	18 - 50	> 50
	< 21hours	18	18	12
Viewing	> 21 hours	12	4	8

(i) Are Age and Number of hours of TV watched associated for either gender? Prior to using the formal test, draw an appropriate plot to form a subjective impression.

(ii) Is there an association between Age and Number of hours of TV watched in the general population as estimated by this sample ?

(iii) Compare the above results.

(iv) Calculate a 95% CI for Proportion of the population who watch more than 21 hours.

2. The table below gives the number of claims made in the last year by the motorists insured with a particular insurance company.

Number of Claims	Insurance Groups			
	I	II	III	IV
0	900	2000	1500	30
1	200	700	500	15
2 or more	30	40	40	4

Is there an association between the number of claims and the insurance groups?

3. A random sample of 500 units is taken from each day's production and inspected for defective units. The number of defectives recorded in the last working week were as follows :

Day	Number of defectives
Monday	15
Tuesday	8
Wednesday	5
Thursday	5
Friday	12

Test the hypothesis that the difference between the days is due to chance.