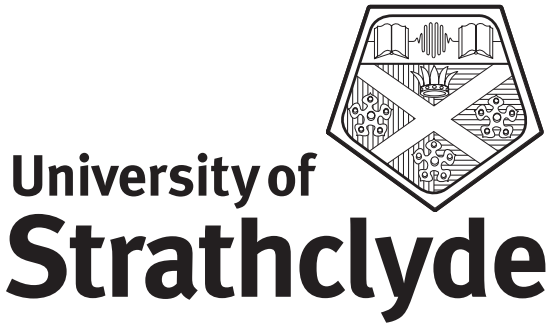


MM104 / MM107
Statistics and Data Presentation Semester 2

Project 4: Chi Squared Tests

Ainsley Miller
ainsley.miller@strath.ac.uk

Lecture Overview

In this lecture we will

- introduce the Chi Squared Test for Independence (or Association)
- learn how to carry out a Chi Square test by hand and using Minitab
- learn the assumptions of a Chi Square test.
- introduce the idea of contribution to the Chi Square test and when it is appropriate to carry it out.
- how to calculate the contribution to the Chi Square in Minitab (if appropriate).

Motivation

The Chi Squared Test for Independence (or Association) - This test is used to discover if there is a relationship between two categorical variables.

Pronunciation: We use the Greek letter chi χ pronounced as "Kai" – it rhymes with "bye".

When is this test appropriate ?

- Your two variables should be measured at an **ordinal** or **nominal level** (i.e., categorical data).
- Your two variable should consist of two or more categorical, independent groups, for example,
  - sex (Males and Females)
  - physical activity level (sedentary, low, moderate and high)
  - profession (surgeon, doctor, nurse, dentist, therapist)

If your data meets these conditions then you can continue with the test.

Hypotheses

If the categorical variables are not related then they are said to be independent.

The null (H_0) and alternative (H_1) hypothesis for a χ^2 test are as follows (you will need to write these in the context in the scenario)

- H_0: Variables are independent (no association)
- H_1: Variables are dependent (associated)

Note that the alternative hypothesis does not specify the type of association.

Chi Squared Test for Independence

We randomly sampled 100 people and ask if they smoke or not. We record their sex and smoking status.

Of the 100 people

- 52 were male of whom 13 smoked (25 %)
- 48 were female of whom 15 smoked (31 %)

We are interested in answering the research question are smoking and sex associated ?

Doing a Statistical Test

1. State the null and alternative hypothesis.
2. State the level of risk associated with the test.
3. Select the appropriate test statistic and calculate it.
4. Calculate the critical value.
5. Decide whether to reject or fail to reject the null hypothesis.

Null and Alternative Hypothesis

For our example of sex and smoking our null and alternative hypothesis are as follows:

H_0 : Sex and smoking are independent i.e. there is no association between the two variables.
H_1 : Sex and smoking are dependent i.e.there is an association between the two variables.

Observed Counts

Using the information from the study we can tabulate this into the following table. This is what we observed from the study/questionaire and is called **observed**.

	Male	Female	Total
Does not smoke	39	33	72
Smokes	13	15	28
Total	52	48	100

Expectation

The χ^2 test allows you to determine if what you observe in a distribution of frequencies is what you would expect to occur by chance. For that reason we need to calculate what we would expect to happen. The expectation is denoted by $E_{i,j}$ and is calculated as follows:

$$E_{i,j} = \frac{r_i c_j}{n}$$

where r_i is the row total for row i, c_j is the column total for row j and n is the sample size.

Expectation cont...

	Male	Female	Total
Does not smoke	$O_{1,1} = 39$ $E_{1,1} = \frac{72 \times 52}{100} = 37.44$	$O_{1,2} = 33$ $E_{1,2} = \frac{72 \times 48}{100} = 34.56$	72
Smokes	$O_{2,1} = 13$ $E_{2,1} = \frac{28 \times 52}{100} = 14.56$	$O_{2,2} = 15$ $E_{2,2} = \frac{28 \times 48}{100} = 13.44$	28
Total	52	48	100

Calculating the Test Statistic

We use the following formula to calculate the test statistic for a Chi squared test.

$$\chi^2 = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$\begin{aligned}\chi^2 &= \frac{(39-37.44)^2}{37.44} + \frac{(33-34.56)^2}{34.56} + \frac{(13-14.56)^2}{14.56} + \frac{(15-13.44)^2}{13.44} \\ &= 0.484\end{aligned}$$

So you've just calculated your χ^2 test statistic but what does it mean ?

A quick note on the χ^2 test statistic

The chi-squared test statistic is not an index of the strength of the association.

For example, doubling the observed frequencies in our table will double the chi-squared test statistic value but the strength of the association is unchanged.

Note that the chi-squared test can only be used when the numbers in the table are frequencies, not when they are percentages, proportions or measurements.

Critical Value

It is very difficult to calculate a p-value for a χ^2 test by hand, so we need to use statistical tables to find a **critical value**. To find the critical value we need to calculate the **degrees of freedom**.
We use the Greek letter ν pronounced "nu" when referring to degrees of freedom.

> The degrees of freedom in a χ^2 test is $(r-1)(c-1)$ where r is the number of rows and c is the number of columns. Think of this as how many categories there are for each variable.

In this example the degrees of freedom are $(2-1)(2-1) = 1$.

Statistical Tables

Don't worry too much about the statistical tables for the Chi squared test now as we will cover those in our written statistics tutorials in Week 9 and 10.

Using the statistical tables we find that our critical value is 3.84.

- Test Statistic $\leq$ Critical Value: Fail to reject the null hypothesis of the statistical test.
- Test Statistic $>$ Critical Value: Reject the null hypothesis of the statistical test.

Conclusion

Since our test Statistic $\leq$ Critical Value we fail to reject the null hypothesis, concluding that there is no association between sex and smoking.

Assumptions for the χ^2 test

The χ^2 test for association is only valid if:

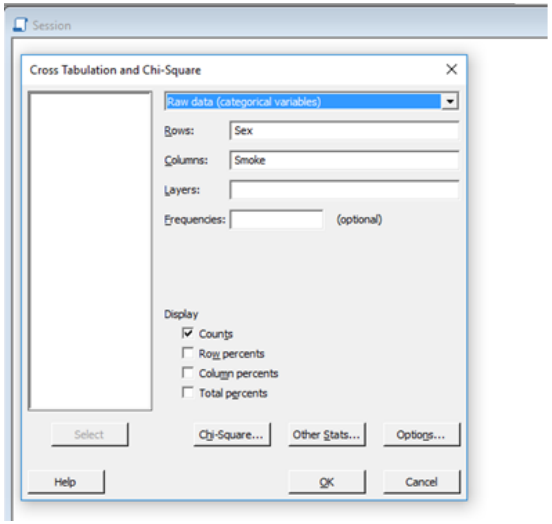
- All expected values $E_{i,j}$ are greater than the value of 1.
- No more than 20 % of the expected values $E_{i,j}$ are less than 5.
- Cell values are independent.

In this example all these assumptions were met. The cells are independent as your smoking preference does not impact someone else.

Minitab - Observed Counts?

Let's firstly make our table of observed counts.

Stat > Tables > Cross Tabulation and Chi Square. It does not matter which way round you have the rows and columns.



Minitab - Observed Counts cont...

Tabulated Statistics: Sex, Smoke

Rows: Sex Columns: Smoke

	0	1	All
0	33	15	48
1	39	13	52
All	72	28	100

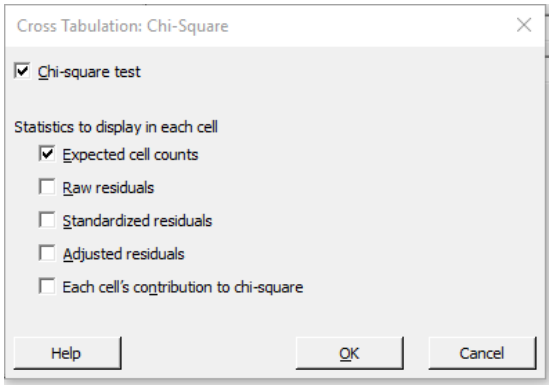
Cell Contents
Count

This table would need to be altered prior to going in the report and presentation, as it is not clear what the 0s and 1s represent.
Note that this table has rows and columns in a different order than in the example, but that's okay.

Minitab - Expected Counts

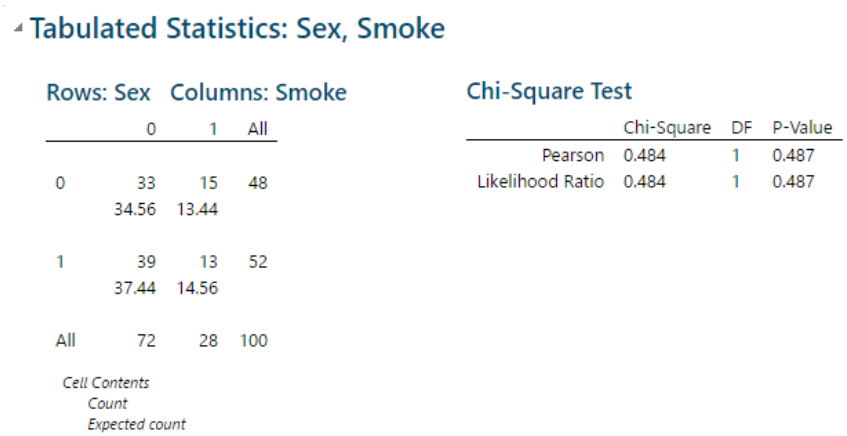
Let's now make our table of observed and expected counts.

Stat > Tables > Cross Tabulation and Chi Square and then click *Chi-Square*



Minitab - Expected Counts cont...

We get table and also the results of the Chi Square test.



Tabulated Statistics: Sex, Smoke

Rows: Sex Columns: Smoke

	0	1	All
0	33	15	48
	34.56	13.44	
1	39	13	52
	37.44	14.56	
All	72	28	100

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	0.484	1	0.487
Likelihood Ratio	0.484	1	0.487

We get two Chi Square test statistics out, we are interested in the Pearson one (they don't always give out the same result). We can see we get the same test statistic that we calculated by hand and we can see our p value is 0.487. Since this is greater than 0.05 we fail to reject the null hypothesis and conclude there is no association between sex and smoking (same conclusion that we obtained previously).

Further Analysis - Contribution to the Chi Square

Minitab displays each cell's contribution to the chi-square statistic, which tells us how much of the total chi-square statistic is attributable to each cell's divergence. **This test is only appropriate if we reject the null hypothesis.**

Stat > Tables > Cross Tabulation and Chi Square and then click *Chi-Square* and tick each *Each cell's contribution to chi Square*

Contribution to the Chi Square Example

Consider a sample of mothers and the relationship between housing tenure and whether they had a pre-term delivery. Initial analysis tested the hypothesis

H_0 : Housing tenure and pre-term delivery are independent i.e. there is no association between the two variables.
H_1 : Housing tenure and pre-term delivery are dependent i.e. there is an association between the two variables.

and found that the p-value was 0.033, concluding that there is an association between housing tenure and pre-term delivery.

Hence, calculate the contribution to the Chi Square and interpret the result.

Contribution to the Chi Square Example cont..

Rows: Housing tenure	Columns: Delivery		
	Preterm	Term	All
Council tenant	29	229	258
	17.70	240.30	
	7.21307	0.53132	
Lives with parents	6	66	72
	4.94	67.06	
	0.22759	0.01676	
Other	3	36	39
	2.68	36.32	
	0.03931	0.00290	
Owner-occupied	50	849	899
	61.68	837.32	
	2.21101	0.16286	
Private tenant	11	164	175
	12.01	162.99	
	0.08433	0.00621	
All	99	1344	1443

Cell Contents
Count
Expected count
Contribution to Chi-square

Interpretation: The largest contributor to the chi-squared test statistic is pre-terms birth for council tenants (7.21307) - more of these tenants had a pre-term delivery than expected if the null hypothesis was true.