Chi-square and bike theft patterns

Royal Geographical Society with IBG
Advancing geography and geographical learning

Specification links

AQA
A Level 3.4.2.4 Statistical skills *Inferential and relational statistical techniques to include Spearman's rank correlation and Chi-square test and the application of significance tests.*

Edexcel
A Level Appendix 1: Geographical skills. *This specification requires students to collect, analyse and interpret such information, and demonstrate the ability to understand and apply suitable analytical approaches for the different information types including, qualitative approaches such as coding and sampling and quantitative approaches such as measures of dispersion, measures of correlation and association from the following statistical tests: t-tests, Spearman's rank, Chi-square, Gini Co-efficient, Lorenz curve.*

OCR
A Level Geographical Skills 4.4 Quantitative skills *b) tests of association and significance tests, such as Chi-square, Spearman's rank, Mann-Whitney U test and T-test.*

Eduqas
A Level Appendix A Geographical Skills. *2. Number and statistical calculations: inferential statistics, including Chi-square.*

What is Chi-square?

Chi-square is a technique used to calculate the degree to which there are differences between the data you collect yourself and what you expected, in theory, *before* you went out. Chi-square calculates the statistical significance of the difference between the observed (O) and expected (E) data.

Below is the equation with each part explained:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

χ^2 Chi-square
$\sum$ the sum of
O observed
E expected

Test 1 a worked example. Chi-square: when are you most likely to have your bike stolen?

There are certain times of the week when people are more likely to suffer from bike crime than others. This is a worked example to investigate when you are most likely to have your bike stolen. The two variables are: during the week or over the weekend.

This activity is focused on bike theft from April 2009 to March 2020 across England and Wales. You will be working out statistically if what you *expect* matches what you *observe*. You will be given the

observed data, sourced from the Crime Survey for England and Wales (CSEW) www.ukdataservice.ac.uk.

Step 1
Before this statistical test is applied you must formulate a **null hypothesis** (Ho). This is a theory which says there is no statistical relationship or significance between variables. This could be:

"There will be **no** significant difference between the observed (O) timings of bike theft and the expected (E) random timings of bike theft".

Step 2
The next step is to estimate the expected (E) frequencies, using Table 1. If your null hypothesis is that there is no association between the time of week and the number of thefts then the expected frequencies would be 2/7x90 and 5/7x90 (i.e., 25.7 and 64.3).

P/hr bike theft	Observed (O)	Expected (E)	(O – E)	$(O – E)^2$	$(O – E)^2 \div E$
During the week					
At the weekend					
Total	90	90			

Table 1

Step 3
Take note: the CSEW data is given in percentages, which cannot be used in Chi-square. This crime survey data tells us that statistically that *72% of bike theft occurs at the weekend and 28% during the working week*. Using CSEW raw data from resource 1, on bike theft in the year ending March 2020 (271,000 incidents) there are 5,211.5 thefts per week. 72% of this figure (for weekend theft) is 3,752.3, whilst 28% (for theft during the week) equals 1,459.2.

This means that, during the week there are 12 bikes stolen per hour, whilst at the weekend this number rises to 78 per hour. Proceed using the figures below in Table 2 (normally, the next step is to collect the observed data through fieldwork, in this instance the Observed column has been filled out for you below). Subtract columns 2 and 3 from one another, (O – E).

P/hr bike theft	Observed (O)	Expected (E)	(O – E)	$(O – E)^2$	$(O – E)^2 \div E$
During the week	12				
At the weekend	78				
Total	90	90			

Table 2

Step 4
Now square the values for rows 2 and 3 and insert the answers into column five.

P/hr bike theft	Observed (O)	Expected (E)	(O – E)	$(O – E)^2$	$(O – E)^2 \div E$
During the week	12				
At the weekend	78				
Total	90	90			

Table 3

Step 5
Take the results from step 4 and divide by the original expected (E) value. Fill in the final column.

To complete the table, you must add up the final column. Fill in the bottom right cell of the table with the Total value. This number is the fractional part of the Chi-square equation.

Step 6

At this stage, it is advisable to rewrite the equation out again, with the $\chi^2 = \sum$ in front of the value you have calculated.

Now complete the statistical equation by working out this final part of the sum.

Your answer will be χ2 (Chi-square).

Step 7
The aim of the Chi-square test is to find out whether the observed pattern agrees with or differs from the expected theoretical predictions. This can be achieved by comparing the calculated result of the test with its level of significance.

On its own the Chi-square result (χ^2) is meaningless — it needs to be validated against 'critical values'. A critical values table is created by statistical experts to allow you to work out if your expected values were statistically *significant*.

In order to proceed geographers next need to decide on the confidence level. The confidence level is normally one of two options, either: 95% or 99% explaining that you expect there to be either a 5% likelihood of the result being calculated by chance i.e., you are *highly* confident, or a 1% likelihood i.e., you are *very* confident. (There is sometimes a 99.9% confidence level, which describes *extreme* confidence).

This is summarised below in table 4.

	Highly confident	Very confident	Extremely confident
Confidence level	95%	99%	99.9%
Uncertainty level	5%	1%	0.1%
Probability level	0.05	0.01	0.001

Table 4

Step 8

Finally, use Table 5 to work out the Degrees of Freedom (df). This is normally n-1 with n representing the number of observations. In this example there are 2 (*During the week* and *At the weekend*).

Work out the Degrees of Freedom (df) for this Chi-square statistical test, (the relevant row from Table 5 has been highlighted) and identify the significance value.

If the result **is larger** than the critical value (in this case df 1) then it is a valid result in our data, and the **null hypothesis is rejected** (and you accept *the* hypothesis — that there *is* a statistical link between what you expected (E) and what you observed (O).

If the result is smaller than the critical value then the null hypothesis is accepted, concluding there is no significant difference between the observed (O) timings of bike theft and the expected (E) random timings of bike theft.

Critical values of the χ^2 distribution		
df	0.05	0.01
1	3.841	6.6355
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.59	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578

Table 5

Test 2 Chi-square: what proportion of stolen bikes were locked?

This is a second Chi-square task using the same ONS 2020 Nature of crime: bike theft data report.

This data has already been converted from percent to raw data (using the Unweighted base – number of incidents).

The survey data asks whether victims of bike theft locked their bike or left it unlocked. In this statistical test you will be working out if the level of locked bikes you expect (E) matches how many were left securely (O).

Over a 10-year period from March 2010 to March 2020 5,376 incidents of bike theft were surveyed in the CSEW. On average, that is 45 per month over this time period.

Step 1
Write out a null hypothesis (Ho).

Estimate the number of stolen bikes per month that were (and were not) locked up for your expected frequency (E), from 2010 to 2020. If your null hypothesis is that there is very strong association between the whether the bike was locked or not and the number of thefts, then the expected frequencies would be 1/10x45 and 9/10x45 (i.e., 4.5 and 40.5).

Av p/m	Observed (O)	Expected (E)	(O – E)	$(O – E)^2$	$(O – E)^2 \div E$
Bike was locked					
Bike was unlocked					
Total	45	45			

Table 6

Step 2
Create a null hypothesis for this second Chi-square test.

Step 3
Use table 7 on the next page to calculate this second Chi-square statistical test.

If you required any further help watch some of the suggested videos at the end of this resource. From now on you will use the equation and earlier steps from the first worked example.

Av p/m	Observed (O)	Expected (E)	(O – E)	$(O – E)^2$	$(O – E)^2 \div E$
Bike was locked	18.0				
Bike was unlocked	27				
Total	45	45			

Table 7

Patterns of bike theft

From resources 2 and 3 you have learnt that students and the young are most likely to have their bike stolen, probably over the weekend. Bike crime tends to be highly patterned with particular places at particular times having high rates of theft, common examples include in and around the victim's home, at university campuses and at transport hubs such as railway and bus stations.

Within these patterns there is likely to be variation in the levels of cycle theft. In Oxford, for example, cycle theft is likely to be higher at some schools than others. This is seen in a study from Washington DC (in 2018) which explored the levels of bicycle theft across different metro stations there. The study found that cycle thefts were higher at stations where there were more cycles parked, few businesses nearby (and hence fewer potential guardians) and higher levels of crime more generally (perhaps indicated a greater number of active offenders operating in the area).

Mburu and Helbich (2016) found that pawnshops, universities, vacant houses, bicycle renting places, parking stands and repair shops were all equally as influential as train stations in increasing the risk of bike theft.

Evidence also suggests that the risk of bike theft is transmissible. Following the occurrence of bike theft at one location, the risk of bike theft occurring nearby in the short term is elevated. For example, using police recorded crime data for the county of Dorset, England, Johnson and colleagues (2008) found that when a bike is stolen from one location, further incidents are more likely to occur nearby and up to about 450 yards for a period of time (around 3 to 5 weeks).

Bike use and bike theft is highly seasonal. In England and Wales levels of cycle theft are higher in the summer months compared to the winter months.

Finally, numerous studies point out that less secure ways of parking a bike are more common and that less than half of bikes which are stolen were locked.

1. Do the above statements all match your Chi-square result for tests 1 and 2?

2. Why are insecure locking practices so common: is it a lack of secure locks? A lack of secure parking facilities? Or perhaps an assumption because bike theft is rare, "it won't happen to me"?

Further work

- A Jamie Cox video from AQA on a 6-mark data response question involving Chi-square https://www.youtube.com/watch?v=1okYBWJ0unU
- Chi-square help https://www.mathsisfun.com/data/chi-square-test.html
- A medical example from the BMJ 'the psychiatrist wants to investigate whether the distribution of the patients by social class differed in these two units' https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-square-tests
- Chi-square from a physiology angle, a good explanation of the Critical Values for Chi-square distribution https://www.youtube.com/watch?v=LZjrI_UmAko
- Levy, J. M., Irvin-Erickson, Y., & La Vigne, N. (2018). A case study of bicycle theft on the Washington DC Metrorail system using a Routine Activities and Crime Pattern theory framework. Security Journal, 31(1), 226-246.
- Mburu, L. W., & Helbich, M. (2016). Environmental risk factors influencing bicycle theft: A spatial analysis in London, UK. PLoS one, 11(9), e0163354.
- Johnson, S. D., Sidebottom, A., & Thorpe, A. (2008). Bicycle theft. Washington, DC: US Department of Justice, Office of Community Oriented Policing Services.

Answers

1. Below is the complete Chi-square table for: when are you most likely to have your bike stolen? The table is filled in with an expected (E) value of 64.3 for *During the week*, and 25.7 for *At the weekend*.

P/hr bike theft	Observed (O)	Expected (E)	(O – E)	(O – E)²	(O – E)² ÷ E
During the week	12	64.3	-52.3	2735.2	42.5
At the weekend	78	25.7	52.3	2735.2	106.4
Total	90	90			$\chi 2 = 148.9$

$\chi 2$ (Chi-square) is 148.9 in this statistical test which is higher than the critical value of 6.6355 (using the 0.01 column) so, in this example, the test is over 99% significant — it is a valid result in our data and the null hypothesis is rejected. This means there **is** a significant difference between the observed (O) and the expected (E).

2. Below is the completed Chi-square table for: what proportion of stolen bikes were locked? The table has been filled in with an expected (E) value of 4.5 for *Bike was locked*, and 40.5 for *Bike was unlocked*.

Av p/m	Observed (O)	Expected (E)	(O – E)	(O – E)²	(O – E)² ÷ E
Bike was locked	18.0	4.5			
Bike was unlocked	27	40.5			
Total	45	45			$\chi 2 = 328.5$

$\chi 2$ (Chi-square) is 328.5 in this second statistical test. This is again much higher than the critical value of 6.6355 (using the 0.01 column) so, in this example, the test is over 99% significant — it is a valid result in our data and the null hypothesis is rejected. For a second time the test has confirmed that there is a difference between the observed (O) and the expected (E).

In both equations the exceptionally high χ2 result is the test telling us that the Ho might be flawed i.e., in test 1 for example, it was immediately clear that during the week there are very low levels of bike theft (compared to the weekend). Generally speaking, if the “O-E” is large then the χ2 result will be high, and we can assume that the initial thinking (E) is unlikely to be true.