Accredited

LEVEL 3 CERTIFICATE IN

Topic Exploration Pack

H866

QUANTITATIVE REASONING (MEI)

Fallacies in Statistics

December 2015

MEI
Innovators in Mathematics Education

OCR
Oxford Cambridge and RSA

We will inform centres about any changes to the specification. We will also publish changes on our website. The latest version of our specification will always be the one on our website (www.ocr.org.uk) and this may differ from printed versions.

Oxford Cambridge and RSA Examinations is a Company Limited by Guarantee. Registered in England. Registered company number 3484466.

Registered office: 1 Hills Road
Cambridge
CB1 2EU

OCR is an exempt charity.

Contents

This Topic Exploration Pack should accompany the OCR resource 'Fallacies in Statistics' learner activities, which you can download from the OCR website.

This activity offers an opportunity for maths skills development.

Introduction

This is an important topic for student's analytical skills in dealing with probability and statistics. We deal with statistics every day when we read the newspaper, explore the Internet or in our work. Never before have people had such easy access to vast amounts of information due to the Internet and social networking. Companies and individuals that display statistics may be trying to mislead the consumer so that they can influence a person's behaviour and attitudes. Politicians often quote statistics in order to persuade people on their policies and it is the job of a diligent citizen to be able to make their own critical judgement on the validity of these statistics.

Probability is often very misunderstood and the examples of fallacies that occur in this guide are extremely common and understandable but it is important that they are understood as they are not encountered by students at GCSE level. The four concepts are discussed in detail below.

Gambler's Fallacy

This is so called because of the mistakes a gambler could make when applying incorrect statistical inference. It is best explained in an example. Imagine someone is betting on the flick of a coin. It is flicked 10 times and the first 8 times it lands heads. The Gambler decides that because heads has landed more frequently in the past then it is less likely to land in the future and bets tails. The fact is that if the coin is fair (it may be the coin is biased) then each successive trial is independent of the other and the probability of being heads on the 9th time is 50% as it was for the previous 8 flicks and the next million flicks. The Gambler has made an incorrect statistical inference because they didn't realise that successive flicks are *independent* of each other. *In other words past outcomes have no effect on future outcomes.*

Prosecutor's Fallacy

This is again best illustrated with an example. Imagine the scenario where 2 people in London witness a car being stolen by a man. They both describe the man as:

- Being in his late 30's
- Having white hair
- Having a pierced nose
- Having one finger missing on his right hand

The next day someone matching this description is found walking in the same location as the theft and is arrested. There is no forensic evidence so based on the statements of the witnesses the man is taken to court. The prosecution argues that:

- The probability of the thief being a man is 0.51
- The probability of a man being in his late 30's is 0.19
- The probability of having white hair is 0.12
- The probability of having a pierced nose is 0.04
- The probability of having a missing finger on right hand is 0.01

Therefore the probability of someone matching this description is $0.51 \times 0.19 \times 0.12 \times 0.04 \times 0.01 = 0.0000046512$ which is so incredibly small (4 in a million) that the accused must be guilty.

Alternatively, the defence argues that although this probability is correct it should have been worked out *given* the accused has these characteristics. To do this they say that 4 people in a million have this description. In the city of London with a population of 10 million it means that there are $4 \times 10 = 40$ people who match this description. Therefore the probability of the accused being guilty is $\frac{1}{40}$ or conversely the probability of innocence is $\frac{39}{40}$ = 97.5% which is quite high.

The correct statistical inference is that of the defence. The prosecutor has mistaken *conditional probability* with *unconditional probability.* The prosecutor calculated the probability someone has those characteristics whilst the defence calculated the probability they were guilty/innocent *given* someone has those characteristics.

The prosecutor's fallacy occurs when an unconditional probability is calculated instead of a conditional probability leading to vastly incorrect statistical inference.

Regression Toward the Mean

This phenomenon is incredibly common and occurs whenever a group of people are sampled twice on the same measure and non-random sampling occurs. It occurs when the measure being tested is a measure of ‘skill’ and ‘chance’. It can be very technical and is best explained with an example.

Imagine 100 students take a multiple-choice test, which is out of 50 marks. The test is completely random and tests academic aptitude and ‘skill’ and students could get high results just by having some ‘good luck’. The mean of all 100 students is measured and it comes out to be around 25 marks. However, imagine that due to some poor sampling then only the top scorers were sampled and a mean of 40 marks was observed. The next day the same students take the test again and the same students that were sampled before were again sampled. The mean of this sample will be lower than before; there is *regression toward the mean*. An incorrect conclusion would be that the students that did better on the first test couldn’t be bothered to do the next test and so did worse. This is not the case and is a fallacy.

The fact is that the sample of both tests is *asymmetrical*. If the entire population was sampled on both tests then the mean would stayed roughly the same. Sampling the top half of the population in both tests meant that the second test got closer to the mean. As the test is a measure of ‘luck’ those group of top students who got the ‘best’ luck the first time will naturally not have the ‘best’ luck again and hence the mean will decrease towards the population mean. This would also happen if the bottom part of the population were sampled as their ‘worst luck’ could only improve. After the second test the mean would increase; *regressing toward the mean*. Researchers find this a very difficult effect to detect and incorrect inferences when analysing this type of data often occurs.

Regression towards the mean occurs when:

- Two measures are taken at different times
- Non-random sampling occurs
- The measure is a mixture of ‘luck’ and ‘skill’

There are lots of other subtleties and these can best be explained on this website for the interested student http://www.socialresearchmethods.net/kb/regrmean.php. The main task for the student is to identify when this effect could occur and what to do to avoid it.

Correlation/Causation

This is a common misconception that has a lot of subtleties. If two variables are measured against each other then correlation can be calculated. It could either be strong, weak or non-existent. The fallacy is to assume that just because two variables are correlated means they are dependent on each other; one of the variables may not actually *cause* and affect the other variable. If two variables A and B are found to have correlation there are different possibilities including:

1) A implies B
2) B implies A
3) A implies C which implies B
4) B implies D which implies A
5) A doesn't imply B; it is just a coincidence

It is incredibly important to fully understand the situation that is being measured to decide which of the above best reflect the correlation. Taking the following examples:

- A = A person's GCSE grade, B = The amount of sleep a person gets

  In this example it would be expected that if they are correlated then B implies A indirectly through another variable C = Concentration levels. It is not the amount of sleep that decides a person's GCSE score but having good concentration levels is a requisite for succeeding at exams and this is helped with better quality sleep. B implies C which implies A rather than B implies A.

- A = The height of a sunflower, B = The hours of sunshine

  In this case B implies A but A doesn't imply B. It would be incorrect to infer that the higher the sunflower the more sunshine occurs.

Activity 1 – The Prosecutor's Fallacy

Resources – Activity Sheet 1, Poster materials, Calculator

This is a full lesson activity. The idea of this activity is for the students to come up with their own conclusion on the guilt of the defendant if they were in the shoes of the prosecution team (incorrect reasoning) and the defence team (correct reasoning) and thus giving them an example of the Prosecutor's Fallacy in action. Hand out Activity Sheet 1 which has the details of the case:

Activity Sheet 1

"A crime was committed in the early hours of Saturday 4th April in Metropolis. Two different witnesses saw a man steal a maths protractor/calculator set from the marketplace. The thief dropped their local college timetable which was found by the police.

Day/Period	1	2	3	4
Monday	Quantitative Reasoning	ICT	Chemistry	FREE
Tuesday	Art	FREE	FREE	Chemistry
Wednesday	FREE	Quantitative Reasoning	Art	ICT
Thursday	FREE	ICT	FREE	Art
Friday	Chemistry	FREE	Quantitative Reasoning	FREE

The next morning a man was caught matching the description entirely. There was no forensic evidence available as the perpetrator had been wearing gloves and the timetable was laminated. It was noted that the timetable was one for a local college and it is unknown which college in the city it is from. A list of subjects taken at the local colleges with numbers on roll was also obtained by the police to aid with the investigation."

Your Task

Based on the theoretical probabilities of the people in the room your job is to form a statistical argument for the prosecution team and the defence team to illustrate the Prosecutor's Fallacy. The student population of the city is approximately 1 million. Present your argument in the form of a poster showing both arguments.

Teacher's notes

Activity Sheet 1 will have a fictional set of statistics that the students need to use to calculate probabilities:

Subject	No. of Students
Maths	63
Chemistry	46
Biology	52
Psychology	82
Physics	32
English Language	69
English Literature	63
History	23
Geography	21
RE	18
Government and Politics	17
Quantitative Reasoning	21
Art	25
Music	11
ICT	43
Computing	28
Photography	14
Law	18
Economics	32
Business Studies	38

From this information students will be able to make estimates for the *unconditional* probability of a student taking one of the four subjects mentioned on the timetable. First they need to find the total number of students: 716. For the subjects mentioned in the timetable we have estimated probabilities:

$$\text{P(Chemistry)} = \frac{46}{716} = 0.064$$

$$\text{P(Quantitative Reasoning)} = \frac{21}{716} = 0.029$$

$$\text{P(ICT)} = \frac{43}{716} = 0.060$$

$$\text{P(Art)} = \frac{25}{716} = 0.035$$

Therefore the prosecutor’s argument is that as the probability of someone taking all of these subjects is 0.064 × 0.029 × 0.060 × 0.035 = 0.000004 – 4 in a million and since it is tiny the person has to be guilty. Note this is the probability a student is taking these subjects.

The defence argument works out the probability:

$$\text{P(Innocence given they take the subject)}$$

So in this case there are 1 million people in Metropolis and hence 1 × 4 = 4 people will take this combination of subjects. Therefore:

$$\text{P(Innocence given they take the subject)} = \frac{3}{4} = 0.75$$

Which is a large probability and hence they should be declared innocent given the lack of forensic evidence.

Students can present their answers in the form of a poster or a fake mini-trial with students playing different roles could be played out for the more adventurous!

Activity 2 – Statistical Diagrams

Resources – *http://nrich.maths.org/7759*, *http://www.suffolkmaths.co.uk/pages/1TeachingResources.htm*

The NRICH website has an excellent resource on statistical diagrams and their relative advantages and disadvantages called ‘Charting More Success’ found at the website above. There is a full set of teacher notes, worksheets and extension activities.

The second website Suffolk Maths has a vast number of resources on misleading statistics. Go to the page and click on the ‘Handling Data Resource’ tab. A Dropbox folder will appear, select ‘5 Misleading Statistics’ which contains a number of different resources that have been tried and tested on misleading statistics and diagrams.

Activity 3 – Correlation/Causation

Resources – Activity Sheet 3, mini-whiteboards

The purpose of this activity is to demonstrate the idea of correlation versus causation. This activity is to be used as a starter activity and as stimulus for a debate about what inferences can be made about correlation. There are a number of ways you can proceed with the activity. Hand out Activity Sheet 3 which has 54 different variables on it. In a spreadsheet create a random number generator by writing the formula =randbetween (1,54).

Generate two sets of random numbers. The first set of numbers corresponds to Variable A and the second to Variable B. Students match the numbers to a variable on the sheet. On the mini-whiteboards students then decide to write what kind of correlation there is (if any). The choices are:

Variable A implies Variable B
Variable B implies Variable A
Variable A implies Variable C implies Variable B
Variable B implies Variable D implies Variable A
No correlation

For the cases of indirect causation students should be able to come up with Variables C and D which may cause the correlation. For example:

Variable A = Test score at English
Variable B = Amount of alcohol drunk

Here a student could argue that B implies C implies A by saying that the more alcohol drunk means less sleep and hence lower concentration and hence a lower test score at English. Variable C in this case is amount of sleep.

Students show their mini-whiteboard with their answer on. Ask individual students why they have made their choice. There is no correct answer to this and different students may have different ideas. The aim of this activity is to give students a grasp of the issues correlation may have and for them to think beyond the 'numbers' and think about the actual variables that are being measured. An interesting and entertaining debate will ensue! Awards could be made for the most imaginative causation between two incredibly obscurely linked variables!

Activity 4 – Regression toward the mean – Famous People Ages

Resources – Activity Sheet 4, Laptops/Access to internet with Flash capabilities
http://www.bbc.co.uk/science/humanbody/sleep/sheep/reaction_version5.swf

This short activity demonstrates the effect of regression towards the mean.

What do you notice about the change in averages from the 1st to the 2nd test?

Ask students to go onto the website above to play the sheep dash game. The idea is to test reaction times which will have elements of 'skill' and elements of 'luck'. A sheep runs across the field and the idea is to click before they reach the other side. If you click prematurely then there is a 1 second penalty. The average of the five sheep is given at the end.

Ask students to do this 3 times and get an average of their 3 averages. They should record this on Activity Sheet 4. These statistics should be collected as a class and be displayed on a spreadsheet for the whole class to see. Sort the data from largest to smallest using the 'Sort and Filter' function. Students then calculate the averages of the worst 3 times, the best 3 times and the average as a whole. These are recorded on their sheet. Then they repeat the experiment and the process is repeated. The mean of the whole class should be roughly the same whilst the mean of the worst should have improved whilst the mean of the best should have got worse. The data has *regressed toward the mean*. Whilst not going into too much detail this is an excellent illustration of this effect and a discussion can be had based on the Teacher Guide at the beginning of the document.

OCR customer contact centre
General qualifications
Telephone 01223 553998
Facsimile 01223 552627
Email general.qualifications@ocr.org.uk

A DIVISION OF
CAMBRIDGE ASSESSMENT