MSCI152: Introduction to Business Intelligence and Analytics

Lecture 3: Sampling Issues

Lancaster University Management School

Overview

- Collecting data
- Sampling
- Bias

8 out of 10 cats prefer Whiskas

“8 out of 10 cat owners said their cat prefers it”

8 out of 10 cats prefer Whiskas

Who says so? **Why** are they saying this?

What does this mean?

- 8 out of 10 particular cat owners?
- 80% of cat owners in the country (in which one)?
- 80% of cat owners worldwide?

What do the cats prefer it to?

- The cheapest cat food?
- All other cat foods?
- Cat foods of about the same price?
- Liver? Fish? Mice?

8 out of 10 cats prefer Whiskas

How large is the effect?

- Is it a strong preference or a weak one?
- Is it enough to justify the higher price?

What was done to determine this “fact”?

- Have they checked all the existing cats?
- Observation of cats in controlled trials to assure the same conditions for all cats?
- Survey of cat owners:
  - what exact questions were asked?
  - could easily have been loaded questions

8 out of 10 cats prefer Whiskas

Have they tested only a **sample** of cat owners?

How **large** was the sample?

- 10? 100? 1000? 10000?
- What percentage of all the cat owners?

How was the sample **chosen**?

- Survey attached to tins of Whiskas?
- Survey in cat-lovers' magazine?
- Survey of selected cat breeders?
- Random telephone interviews?
- The Whiskas producer owner's cats?

Slogan became:

"8 out of 10 cat owners who expressed a preference said their cat prefers it"

Inquiry into the UK Advertising Industry

The Royal Statistical Society made a submission to a House of Lords inquiry.

Particular concerns raised were:

- **Sample size**
  - One product had a recorded sample size of 25
  - A L'Oreal Men Expert product advert has a sample size of 54
- **Sample suitability**
- **Stretched evidence**
  - A L'Oreal Glam Shine Stain advert asserted that '73%' of the 60 people surveyed agreed
- **Substantiation**

Collecting data from a population

Population (or “sampling frame”) is the complete **set** of the objects we are interested in

- e.g., all people in the UK, all BMW cars, all patients going through a medical treatment,. . .

Census: Collection of data from every population member

- Expensive and lengthy process
- In the UK: it takes place every 10 years; every household is contacted

Referendum: Collection of data from every population member on a voluntary basis

- Expensive
- e.g., 2016: “Should the United Kingdom remain a member of the European Union or leave the European Union?”
- 2014: “Should Scotland be an independent country?”

Sampling: Data collection from samples

Types of sampling

- Activity we conduct to **access** a sample (or samples) within the population.

Methods of sampling

- Action we take to **construct** a sample (or samples) within the population

Common types of sampling

Surveys

- What is the pattern of traffic on the M6?
- How many sparrows are there in England?

Quality control

- What proportion of products are damaged in transportation?
- How long will my light bulbs last?

Interviews (email, phone, face to face) about customer satisfaction:

- food or service at restaurant, hotel, shop
- services by travel agency

Market research

- What sort of new magazine would sell well?
- What do cats think of Whiskas?

Often using an Internet **questionnaire**

Properties of good sampling

Needs to be **practical**

- Is it feasible?
- How long will it take?
- How much will it cost?
- Will I be able to get a big enough sample?

Needs to lead to **reliable** data for **conclusive** analysis

- Must minimize errors, missing values, etc.
- Must offer sufficient amount of data
- Must **avoid** all possible types of **bias**
  - which can be very difficult

Questionnaires

Need careful planning

- What information do you need?
- Type of questions (select one answer, select several answers, rank answers, open questions etc.)
- Clear and unambiguous
- Neutral phrasing of questions
- Length of survey and order of questions
- Make it clear, easy and short!

Think about it from the point of view of a respondent

Pilot it first: Small exercise to see what sort of responses you might get

Questionnaire Example

Skein Airlines has a website where people can find out about their flights and other information. The person responsible for designing and maintaining the site is planning to add a questionnaire as follows. What are the problems and how can the questionnaire be improved?

1. How did you first hear about the Skein Airlines site?

   TV advertising / Print advertising / Link from another website

2. Did you find the information you needed? Yes / No

3. On a scale of 1 to 10, how useful was the information and how clearly was it set out? ___

4. What is the speed of your Internet connection? ___

5. How frequently do you fly?

   Every day / Once a week / Once a month / Once a year

Comments on Questionnaire

- The goal of questionnaire is not clear
- In 3: 2 questions in one
- People may not know the speed of their Internet connection
- What is the relevance of the Internet connection to this company?
- Amongst others . . .

Improving a questionnaire

Imagine you are answering it and be critical:

- Is it possible to answer in any circumstances?
- Can we answer it exactly (or close enough)?
- Do we know what each part of it means?
- Does the answer depend on something?
- Is everyone familiar with the terminology?
- Could the question irritate/offend someone?

Think outside of the box:

- Should some kind of people be disallowed to participate?
- E.g.: kids, foreigners, random visitors,. . .
- Check non-standard circumstances

Bias

Bias $\sim$ tendency, prejudice, unfairness

- either positive or negative

Biased sample

- measurements, observations or responses are likely to be **unrepresentative of the population** as a whole, because of the way the sample is chosen
- e.g., test Whiskas only on cats in rich families

Analysis of a biased sample

- **results may be very misleading**
- a severely biased sample is usually worthless

Example of a biased sample

Traffic on the M6 motorway

- It opened in the 1960s
- A study was undertaken to estimate the proportion of vehicles of different types on the M6

A student went to a bridge and began counting

- Too many vehicles to record them all
- So they took a sample
- For convenience just one lane was chosen
- The outside (fast) lane

Thus coming to the following conclusion

- There are no lorries on the M6 !!!

Example of a biased election poll

1948 US presidential election

- Poll on election day said Thomas Dewey (Republican) would win easily
- Newspapers were so confident of the result that they announced it in their headlines the next morning, but
- Harry Truman (Democrat) won!

Why did they get it so wrong?

- Poll was conducted by telephone
- Back in 1948 a lot of people did not have phones
- Richer people were more likely to have a phone than poorer people
- Richer people were more likely to vote Republican, poorer people were more likely to vote Democrat

Why might the polls be wrong

- It is very difficult!
- Late changes of opinion
- People lying
- Turnout
- Postal voting
- Phrasing of the question
- Overseas voters
- Sampling procedure

2016 EU Referendum
Polls: Remain to win
Actual: Leave won

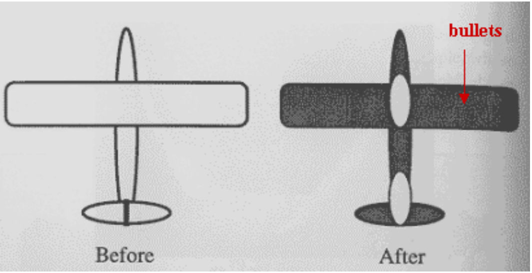
Another biased survey example

- You may well have been surveyed and asked why you chose to come to Lancaster University
- The recruitment people want to know so that they can try to attract even more of you next year (!)
- But are they asking the wrong people?
- Should they really be asking the people who chose not to come here "Why did you choose not to come here?"
  - These are the types of people that need to be persuaded what a wonderful place this is
- The same in the restaurants, pubs,. . .

Survivor bias

A bias caused by considering only objects that pass some selection critera e.g., companies surviving a financial crash or customers who are successful on obtaining a loan.

Real example from World War II in the 1940s: You inspect a sample of war planes that returned home and plot the pattern of bullet holes. The dark regions have high density of bullet holes. Your task is to recommend where to put extra armour on the new planes. *What would you recommend?*



Size of sample

A small sample means **we do not have much information**

- Uncertainty when extrapolating to the whole population

Choosing the minimum size of sample is **important**

- But it requires a good understanding of probability

Sample size is generally **less important than bias**

- We can get surprisingly accurate results from relatively small samples if great care is taken to avoid bias, but not too small
- e.g., political opinion polls generally use samples of about 1000, for **any** population of potential voters and they usually turn out about right (!)

Exercise

A shopkeeper is considering stocking FairTrade coffee. It costs 20% more than the current brand stocked in the shop.

The shopkeeper wants to know the likely demand for the new product, so she conducts a survey on one day by asking each customer who buys the coffee currently stocked:

"**Would you buy FairTrade coffee for 20% more than you have just paid?**"

and records the number of people who say

"**Yes**"

Think of reasons why the survey may be biased, i.e.,

- why the results from her survey may not reflect the demand for the new brand.

Exercise: Possible sources of bias

- Only the customers shopping in her shop are asked
- The customers may say “yes” because they are coffee drinkers
- In face-to-face contact, the customers feel obligated to say “yes” because they want to be nice to the shopkeeper; but in reality they will not spend more money on coffee

What can cause bias?

Sampling procedure:

- Sampling out of the population
    - taken the wrong way
- Particular groups underrepresented, overrepresented or missed out altogether
- Sampling one way because it is easier
    - e.g., kids, students, unemployed and retired people are typically happier to respond on the street; employed are in a rush (lunch break, etc.)
- Sampling according to our personal preferences
    - e.g., asking attractive people

What can cause bias?

Changes in circumstances:

For example, UK immigration and emigration statistics:

- in 2003 almost all flights to or from Poland were from or to Gatwick, Heathrow or Manchester, hence surveys of arriving passengers were carried out there
- by 2008 there were a lot of new flights from and to Poland but mainly at other airports
- however, the 2008 survey just looked at the original three airports hence underestimated the increase in immigrant and emigrant numbers from and to Poland

What can cause bias?

Systematic inaccuracy:

- lying, trying to please, misinterpreting the questions
- inaccurate data recording, e.g.:
  - a computerised survey of doctors found an amazingly high proportion were born on 11th Nov 1911
  - poor lab equipment when taking scientific measurements
  - in surveys: ambiguous or misleading questions
- difficulty in measuring **fuzzy concepts**
  - fuzzy $\sim$ imprecise, ambiguous, vague
  - emotions, feelings, quality

What can cause bias?

Psychological factors:

During interviews – the interviewer's effect on the respondent

- distortion of response to a personal or telephone interview which results from differential reactions to the social style and personality of interviewers or to their presentation of particular questions.

Non-response

- People who do not return questionnaires; e.g., on customer satisfaction – who is likely to return the questionnaire?
- They might answer some questions but not others

Wrap up

We covered:

- The challenges of data collection
- Sampling and bias

Next:

- Presenting Data