MSCI152: Introduction to Business Intelligence and Analytics

Lecture 2: Sampling Methods

Lancaster University Management School

Overview

- Collecting sample data
- Bias and uncertainty
- Sampling Process
- Sampling Methods

Collecting sample data

Sample: A **subset** of members selected from a population

- Exhibits **characteristics** typical of those possessed by the population of interest
- Data is collected **from the sample** with the objective to analyse and make inferences **about the population**
- Sample must be **well selected** to **well represent** the population

Examples of data from samples

- Annual inflation rate
- Annual GDP (Gross Domestic Product)
- Immigration and emigration figures
- Results of medical experiments
- Weight of infants at birth by country
- Satisfaction of customers purchasing at Amazon online
- Conversion rate of British Pound against Euro
- Interest rate on car loans

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

Difference between uncertainty and bias

Uncertainty:

- Limits of knowledge due to using a sample rather than the whole population
- Can be measured ("sampling variation"): see textbook, not covered this year
- Can be reduced (at a cost) by taking a larger sample and structuring the sample (e.g., stratified sampling)

Bias:

- Nature of the method means that the sample results are likely to be systematically different to the population
- To get rid of bias need to change the method

Difference between uncertainty and bias

Suppose we want to estimate accurately the average height of Lancaster undergraduates

Method 1: High uncertainty (but not biased)

- List of all undergraduate students
- Choose 2 people at random as the sample
- Measure the average height of the sample

Method 2: Biased (but low uncertainty)

- Reduce the list of students by including only male students
- Choose 100 at random as the sample
- Measure the average height of the sample

Stages of the sampling process

Defining the population of interest

Planning stage

- Specify a set of elements that are possible to measure
- Specify a **sampling method** for selecting the elements
- Determining the sample size

Implementing the sampling plan

Conducting sampling (i.e., collecting data!)

Sampling Methods

We want the sample to be a fair representation of the whole population (**no bias**)

We also want the process to be **efficient**

Useful sampling methods:

- **Simple random** sampling
- **Systematic** sampling
- **Stratified** sampling
- **Cluster** sampling
- **Convenience** sampling
- **Voluntary response** sampling
- **Quota** sampling

Simple Random Sampling

Selection so that each individual member of the population has an equal chance of being selected

Hence, every subset of size n ($n \geq 1$) elements has an equal chance of selection from population of size N ($n \leq N$)

Simple Random Sampling

Examples of **how to** achieve it:

- Flip a coin
- Throw a die
- Pull names from a hat
- Use random numbers on a computerised list

Examples of **use**:

- Select staff members from a company for a detailed interview
- Jury service: random selection from electoral register

Simple Random Sampling

Advantages:

- Pure form of sampling, conceptually simple
- **No inherent bias**
- Can **analyse well** mathematically
- The textbooks like this method best !!

Disadvantages:

- Need to be able to list the whole population
  - often impractical, time-consuming or impossible
- Subject to **sampling variation**
  - may get an unusual sample by chance
  - some other methods can make this less likely
  - particularly an issue if relatively small sample

Systematic Sampling

How to achieve it:

- Choose an **integer positive number** k
- Select some starting point (often at random)
- Then select **every** $k^{\textbf{th}}$ **element** in the population
- e.g., $k = 3$ (a 1 in 3 sample)

Systematic Sampling

How to get a **"1 in k"** sample

- Find a random number r between 1 and k
- Include the r^{th}, the $(k+r)^{\text{th}}$, $(2k+r)^{\text{th}}$, etc.
- e.g., if $k = 100$ let $r = 57$, then take the 57^{th}, 157^{th}, 257^{th}, ...

For example, every k^{th} person arriving at a shop, every k^{th} item manufactured, etc.

Every item still has an equal chance of being selected

But **not every combination** has equal chance

- e.g.: if you are chosen then the person sitting next to you cannot be

Use of Systematic Sampling

Quality control:

- examine every 100^{th} car produced
- Not suitable for smaller items (e.g., every 100^{th} nail)

Local council checking up on loft insulation grants:

- Every 4^{th} recipient checked if they already had loft insulation
- Every 9^{th} recipient checked afterwards to see if they had actually installed it
- So every 36^{th} recipient got checked both ways

Museum wanting to know views of customers:

- Interview every 50^{th} visitor (e.g., with a financial incentive)
- e.g., every customer whose ticket number ends in 33 or 83

Systematic Sampling

Advantages:

- You can do it as you go along
- You do not have to have the complete population available
- Conceptually simple
- Easy to do it and easy to explain how to do it
- May be very convenient
  - e.g., for the museum interviewees do not accumulate
- May make variety more likely than simple random sampling
  - e.g., for the museum we get visitors all day long

May have to be **careful** to avoid fixed patterns

- e.g., if you check on the typesetting of a newspaper every 28 days you always get the same day of the week

Stratified Sampling

Stratum $\sim$ level, layer, region, etc

Population is heterogeneous and composed of strata
(e.g., gender, income, religion, education).

How to achieve it:

- Divide your population into "strata"
- Sample within each stratum (random or systematic)

Divide your population into "strata"

- Results from each stratum are expected to be different
- More variation is expected between strata than within strata
- Each member of the population in one stratum only
- E.g., for sampling buying preferences we could have 6 strata:

$$\{female \leq 25\},\ \{25 < female \leq 50\},\ \{female > 50\},$$
$$\{male \leq 25\},\ \{25 < male \leq 50\}, \{male > 50\}$$

Stratified Sampling

Defining your strata: Think carefully

- stratum for people whose names begin with J: no sense unless studying names
- Geographical strata may make sense when sampling housing costs, but may not when sampling favourite films
- It needs to be practical, i.e., you need to know into which stratum each individual falls

Sampling within each stratum

- Either: **sample proportionately** – get a sample of the same size from every stratum
- Or: **sample disproportionately** – extrapolate strata separately (get larger samples from strata that might be expected to vary more; this will give a more reliable final outcome)

Stratified Sampling: Example

Quality control in a pie company

- It wants to assess the quality of pies it produces
- **Different types of pie** may be different
- Pies produced in **different factories** may be different
- Pies produced on **different days of the week** may be different

So, define three strata

- By (1) pie type, (2) factory, and (3) day of production
- E.g., {Pork pie, Lancaster, Tuesday}
- Take a sample within each stratum

Stratified Sampling

Advantages:

- Good at reducing sampling variation
- Sample is more representative and so we can be more confident about extrapolating results
- Avoids the problem that a simple random sample could be unusual by chance
- Covers variety in population
- May give much more useful information
- whether about pies or about shopping habits

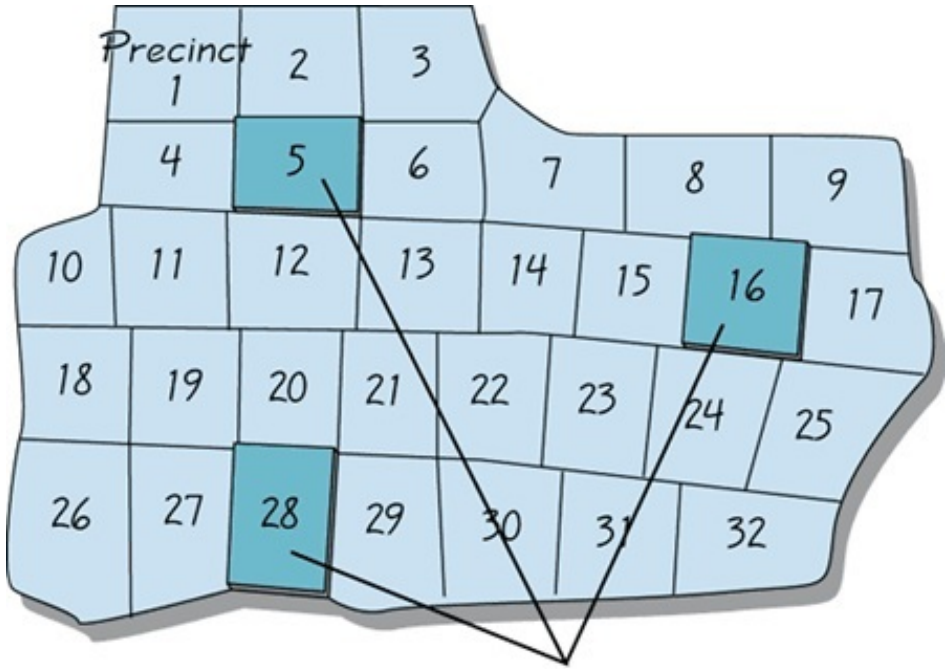
Disadvantages:

- Need to identify relevant strata
- Need stratified information about the population

Cluster Sampling

How to achieve it:

- Divide the population into clusters
- (Randomly) select some of those clusters
- Sample or take all from selected clusters



Assumes clusters are "mini populations"

Cluster Sampling

Single-stage:

- Divide population into clusters
- Select clusters to sample from at random
- Choose (or sample from) all the members of selected clusters

Multi-stage:

- Several stages of selecting clusters at random
- Clusters within clusters
- Final stage may be random sample of individuals

Cluster Sampling

Examples of clusters:

- **Geographical:** province, county, town, district, street, etc.
- **Organisation:** company, school, university, etc.
- **Products:** batch, carton, box, etc.

Advantage:

- Often cheaper and quicker

Disadvantage:

- Clusters may not be truly representative
- depends on differences between the clusters

Stratified vs. Cluster Sampling

Both divide population into groups, but

Stratified sampling:

- Groups have different characteristics
- Choose members from each group

Cluster sampling:

- Each group is a mini version of the whole population
- Choose some groups only and ignore the others

Convenience Sampling

Basic: Choose anyone/anything

Market research

- Choose people who look friendly?
- Ignore the fact that some people do not want to talk to you, or return your questionnaire
- often inevitable; called a "voluntary response" sample

Quality control:

- Choose items that are easy to access (those on top of a pile)
- Companies wanting to know about new products
- Easiest to ask current customers
- but really want to know about potential new customers

Voluntary Response Sampling

Invite a group to respond

- e.g. Internet surveys, customer satisfaction surveys, TV/radio phone-in polls

Advantage:

- Cheap and quick

Disadvantage:

- Problem is low response rate, e.g. may only get 5%
- Those with strong opinions are more likely to respond
- Those with negative opinions may be more likely to respond

Quota Sampling

How to achieve it:

- Stratified sample, select strata
- Within each stratum do **convenience sampling** until a given **quota** (number) is reached

E.g., surveys in the high street: six strata

- $\{females < 25\}, \{25 \leq females < 50\}, \{females \geq 50\}, \{males < 25\}, \{25 \leq males < 50\}, \{males \geq 50\}$
- get **100** in each stratum

E.g., surveys in Lancaster University: two strata

- $\{female\ students\ 60\%\}$, $\{male\ students\ 40\%\}$
- get **60 females** and **40 males** into a sample of 100

Very hard to avoid serious bias

- but a lot better than basic convenience sampling

Other Sampling Issues

How to collect the data?

- post / phone / Internet / interviews / questionnaires

How much data do you need?

- The more the better (as long as the computer can handle it)

Non-response

- For some methods, may only get 5% or lower response rate
- Need to consider if that introduces bias
- How can it be minimised?

Questionnaire design

- Much harder than most people think

Wrap up

Here we:

- Discussed Sampling methods

Next time:

- Sampling issues