Basics of Sampling & Data Collection

Rui Biscaia rui.biscaia@novasbe.pt



Introduction

Sampling techniques enable you to reduce the amount of data you need to collect by considering only data from a sub-group rather than all possible cases of elements (e.g., all football fans vs. fans of X number of teams)

• In selecting a sample to study, it should represent the full set of cases in a way that is meaningful, and which can be justified.

The need to sample

For same research questions it is possible to collect data from an entire population as it is a manageable size. Sampling provides a valid alternative when:

- It would be <u>impracticable</u> to survey the entire population.
- <u>Budget constraints</u> prevent you from surveying the entire population.
- <u>Time constraints prevent you from surveying the entire population.</u>

e.g., interviews, questionnaires, observation, ...

Introduction

The importance of defining the research population clearly



e.g., **Population**: All season ticket holders of the English Premier League

e.g., Target population: Season ticket holders Manchester United

e.g., Sample: a certain group of STH of MU

e.g., Case or element: a specific 'important' STH of MU

Population, target population, sample and individual cases

Introduction

An overview of sampling techniques



Saunders et al., 2020, 276

Probability sampling (or representative sampling)

It is commonly associated with <u>survey research strategies</u> where you need to make <u>inferences from your</u> sample about the population to answer your research question(s) and to meet your objective(s).

1. Identify a suitable sampling frame based on your research question(s) and objectives i.e., a complete list of all the cases in the target population from which your sample will be drawn. Possible problems when using existing databases (e.g., STH & Premier League; members of a Gym; data about sport participation – Sport England): Individual databases are often incomplete or inexistent. ٠ The information held about organisations in databases is sometimes incomplete. ٠ The information held in databases soon becomes outdated. • The way you define your **sampling** frame has implications on the **extent to which you can generalise** from your sample. By defining the sampling frame, you are defining the target population. ٠ You should not generalise beyond the sampling frame

2. Decide on a suitable <u>sample size</u>

The larger your sample's size, the lower the likely error in generalising to the target population. Your choice of sample size should be governed by:

- The level of certainty that the characteristics of the data collected will represent the characteristics of the target population
- The accuracy you require for any estimates made from your sample
- The type of analysis you are going to undertake (e.g., SEM)
- The size of the target population

e.g., t-tests and n=30

3. Be aware of the importance of a high response rate

The most important aspect of a probability sample is that it represents the target population.

- If you have many non-respondents (people who decide not to be involved in the research), your sample won't likely be representative of the target population and the data may be biased
- 4. Estimate response rates and actual sample size required

With all probability samples, it is important that your sample is <u>large enough</u> to provide you with the necessary <u>confidence in the</u> <u>data.</u> $n^a = \frac{n \times 100}{n^a}$

Response rates can vary considerably when collecting primary data







Use simple

random sampling

Use systematic

random sampling

frame contain

periodic patterns?

No

Non-probability sampling (non-random)

Probability samples may be either not be possible to obtain (i.e., not having a sample frame – do you have data about the entire fan base population?) or not be appropriate to answering the research question.

Non-probability sampling provides a range of alternatives techniques to select samples (e.g., exploratory stages of research pilot tests)



Decide on a suitable sample size

The issue of sample size is always a subject to large discussion and there are no strong rules (more based on establishing credibility).

- Generalisations are made to theory and <u>not</u> to the <u>population</u>
- e.g., interviews: Data saturation tends to be the most important criteria (i.e., collect data until little, if any, new information is obtained)

Generic guidelines about the minimum non-probability sample size (developed from Saunders, 2012)

Nature of the study	Minimum sample size
Semi-structured interviews	5-25
Ethnographic	25-36
Grounded theory	20-35
Considering a homogeneous group	4-12
Considering a heterogeneous population	12-30

Selecting the most appropriate sampling technique and the sample

Quota sampling (<u>https://www.scribbr.co.uk/research-methods/quota-sampling-method/</u>)

It is entirely non-random (selection of pre-determined number) and often used for structured interviews as part of a survey strategy.

AIM: to replicate the population of interest

- Less costly
- Does not require a sampling frame
- Often used for large target populations

Purposive sampling (https://www.scribbr.co.uk/research-methods/purposive-sampling-method/)

Selection of the cases that best enable to answer the research question(s)

- Often used with very small samples
- It cannot be considered statistically representative of the target population

Volunteer sampling

Snowball sampling: used when it is difficult to identify members of the desired population

- a. Contact 1-2 cases in the population
- b. Ask these cases to identify further cases
- c. Ask these new cases to identify further new cases (and so on)
- d. Stop when no new cases are given, or the sample is a large as manageable

Self-selection sampling: when you allow each case, usually individuals, to identify their desire to take part in the research.

- a. Publicise your need for cases
- b. Collect data from those who respond







Haphazard sampling

When sample cases are selected without any obvious principles of organisation in relation to the research question.

Convenience sampling

Selection of the cases simply because they are easily available (or more convenient)

- It is often used but prone to bias
- Sometimes, convenient samples may meet purposive sampling criteria that are relevant to the research aim e.g. it may
 represent a 'typical case' and show an illustrative scenario



Haphazard sampling

When sample cases are selected without any obvious principles of organisation in relation to the research question.

Convenience sampling

Selection of the cases simply because they are easily available (or more convenient)

- It is often used but prone to bias
- Sometimes, convenient samples may meet purposive sampling criteria that are relevant to the research aim e.g. it may
 represent a 'typical case' and show an illustrative scenario





Primary data vs. Secondary data

Primary data	Secondary data
Data collected by yourself and linked to your research	Data that is already available – collected by someone else for a
purpose.	different purpose.
Common sources • Surveys • Focus Groups • Observation • Experiments • Interviews •	Common sources: Government departments and official publications Industry reports National and International media Professional and Industry publications Social Media Buzz Subscription services



https://www.youtube.com/watch?v=VKR1EgfcKdQ

Secondary data

Types of secondary data and uses in research



It can include both **quantitative** (numeric) and **qualitative** (non-numeric) data and are used mainly in descriptive and explanatory research.

Secondary data

Document secondary data

Data that, unlike the spoken word, ensure physical (including digitally) evidence. It can be transposed across both time and space reanalysed for a purpose different to that which they were originally collected

• Text materials & Non-text materials

Survey-based secondary data

Data originally collected for some other purpose using a survey strategy, usually questionnaires

• Censuses; regular surveys (annual fan satisfaction surveys); ad-hoc surveys (cross sectional study)

Multiple-source secondary data

It can be compiled entirely from document or survey secondary data, or can be an amalgam of the two

- The way in which a multiple-source data set has been compiled will dictate the sort of research question(s) or objectives for which you can use it.
- It can be longitudinal data or cross-sectional data

https://www.sportengland.org/know-your-audience/data

EFL

2019

SUPPORTERS SURVEY

Searching for secondary data

Unless you have the intention to analyse one specific secondary data set, you will have to ascertain whether the data you need are available.

- 1. Establish whether the sort of data you require are likely to be available as secondary data (e.g., comments on social media platforms?)
- 2. Locate the precise data you require (e.g., Twitter; Newspaper website; Instagram?)

Advantages

- Fewer resource requirements
- Unobtrusive
- Longitudinal studies may be feasible
- Comparative/contextual data
- Unforeseen discoveries
- Permanence of data

Disadvantages

- Original purpose may not match your need
- Access may be difficult or costly
- Aggregations and definitions may be unsuitable
- No control over data quality
- Initial purpose affects how data are presented

Evaluating secondary data sources

You need to be sure that:

- 1. Available data will enable you to answer your research question(s) and to meet your objectives.
 - i.e., What may appear relevant at first may not on closer examination be appropriate to your RQ and objectives
- 2. The <u>benefits</u> associated to using secondary data will be greater than the costs.

i.e., spend time evaluating whether potential sources.

3. You have access to the data.

- i.e., typically, not a problem with data available on the Internet and University Library but will imply time investment.
- i.e., For other data sources, you may need permission prior to get access and must consider ethical implications.



https://www.sportengland.org/know-your-audience/data



https://www.uksport.gov.uk/resources/transparency-and-open-data

Evaluating secondary data sources



The measures need to match those that you need

e.g., fan satisfaction: what you need vs. fan engagement: what you have access to

The data cover the population about which you need the data, for the time you need, and contain variables that enable to answer your research question(s)

e.g., exclude what is not relevant for your research

 It depends on the method in which the data were collected and the source (see week 6 slides)

e.g., Sport participation in the UK: Sport England vs. personal blog?

Potential deliberate distortion of data

e.g., satisfaction surveys not showing an overall picture Data that do not represent the topic of interest

Evaluating secondary data sources



Evaluating your secondary data sources

Overall suitability

- Does the data set contain the information you require to answer your research question(s) and meet your objectives?
- ✓ Do the measures used match those you require?
- ✓ Is the data set a proxy for the data you really need?
- Does the data set cover the population that is the subject of your research?
- Does the data set cover the geographical area that is the subject of your research?
- Can data about the population that is the subject of your research be separated from unwanted data?
- Are the data for the right time period or sufficiently up to date?
- Are data available for all the variables you require to answer your research question(s) and meet your objectives?
- ✓ Are the variables defined clearly?

Precise suitability

- How reliable is the data set you are thinking of using?
- ✓ How credible is the data source?
- ✓ Is it clear what the source of the data is?
- ✓ Do the credentials of the source of the data
- (author, institution or organisation sponsoring the data) suggest it is likely to be reliable?
- Do the data have an associated copyright statement?
- Do associated published documents exist?
- Does the source contain contact details for obtaining further information about the data?

- ✓ Is the method described clearly?
- If sampling was used, what was the procedure and what were the associated sampling errors and response rates?
- Who was responsible for collecting or recording the data?
- (For surveys) Is a copy of the questionnaire or interview checklist included?
- (For compiled data) Are you clear how the data were analysed and compiled?
- Are the data likely to contain measurement bias?
- What was the original purpose for which the data were collected?
- Who was the target audience and what was its relationship to the data collector or compiler (were there any vested interests)?
- Have there been any documented changes in the way the data are measured or recorded including definition changes?
- How consistent are the data obtained from this source when compared with data from other sources?
- ✓ Have the data been recorded accurately?
- Are there any ethical concerns with using the data?

Costs and benefits

- What are the financial and time costs of obtaining these data?
- Can the data be downloaded into a spreadsheet, statistical analysis software or word processor?
- ✓ Do the overall benefits of using these secondary data sources outweigh the associated costs?

And finally

 Is permission required to use these data and, if 'yes', can you obtain it?

Source: Authors' experience; Dale et al. (1988); Dochartaigh (2007); Hair et al. (2011); Smith (2006); Stewart and Kamins (1993); Vartanian (2011)

Primary data: Observation

Observation involves the systematic viewing, recording, description, analysis and interpretation of people's behaviours.

- **Participant observation** is qualitative and often focused on discovering the meanings people attach to their actions.
- **Structured observation** is quantitative and more concerned with the frequency of actions.

Technology facilitates new ways to conduct observation.

- Internet-mediated observation involves the collection of data from online communities.
- **Observation using videography** involves recording moving images onto electronic media to collect observation data.



While participant observation and structured observation have been traditionally linked to primary data collection, <u>technology makes it possible</u> to apply observational techniques to <u>both primary and secondary data</u>.

Example of participant observation

Staff member:				
Behaviour	Was the behaviour observed?	Comments		
Smiles and makes eye contact with the customer				
Greets the customer in a friendly manner				
Gives the customer undivided attention throughout the transaction				
Suggests extra items that have not been ordered by the customer				
Places items on clean tray with tray liner facing customer				
Ensures that customer is told where all relevant extras (e.g. cream, sugar) are located				
Explains to customer reasons for any delays and indicates likely duration of delay				
Neatly double-folds bags containing items with the Fastfoodchain logo facing the customer				
Price of order is stated and customer thanked for payment				
Lays all money notes across till drawer until change is given and clearly states the appropriate amount of change				
Customer is finally thanked for transaction, hope expressed that the meal will be enjoyed, and an invitation to return to the restaurant issued				

Primary data: Interviews

The research interview is a purposeful conversation between two or more people.

• It is about asking purposeful questions and carefully listen the answers to be able to explore these further.

Types of Interviews

- a. Structured interviews: questionnaires based on pre-determined and standardised questions (interviewer completed forms)
- b. Semi-structured interviews: the researcher has a list of themes and key questions to be covered but its use may vary from interview to interview
- c. Unstructured (or in-depth) interviews: more informal (no pre-determined list of questions, but important to have a clear idea of what to explore). The interviewee talks freely about events, beliefs and behaviours in relation to the topic area

	Exploratory	Descriptive	Explanatory	Evaluative			
Structured		(D) (D)	\mathbb{C}				
Semi-structured	(\mathbf{P})		$\mathbb{P}\mathbb{P}$	ÐÐ			
Unstructured	\mathbf{P}			\bigcirc			
P = more frequent, P = less frequent							

Semi-structured or in-depth interviews

Preparation stage:

- You need to be knowledgeable about the research topic and context.
- Develop interview themes (and supply information to the interviewee) before the interview.
- Select the appropriate location for the interview.

During the interview:

- Appropriateness of your appearance at the interview i.e., perceived credibility
- Nature of your comments to open the interview

i.e., the first few minutes will likely affect the rest of the interview

• Approach to questioning

i.e., reduce the scope for bias with questions phrased clearly; avoid theoretical jargons; try to link the questions with real-life experiences of the participants

• Appropriate use of different types of questions

i.e., open questions; probing questions; Specific and closed questions; other means to further your questioning – "That's interesting"," Will you please tell me more"; questions to avoid



To help you conduct your semi-structured or in-depth interview

Appearance at the interview

 How will your appearance at the interview affect the willingness of the interviewee to share data?

Opening the interview

- How will you open the interview to gain the confidence of your interviewee?
- What will you tell your interviewee about yourself, the purpose of your research, its funding and your progress?
- What concerns, or need for clarification, may your interviewee have?
- How will you seek to overcome these concerns or provide this clarification?
- In particular, how do you intend to use the data to which you are given access, ensuring, where appropriate, its confidentiality and your interviewee's anonymity?
- What will you tell your interviewee about their right not to answer particular questions and to end the interview should they wish?
- How will you explain the structure of the interview?

Asking questions and behaviour during the interview

- How will you use appropriate language and tone of voice, and avoid jargon when asking questions or discussing themes?
- How will you word open questions appropriately to obtain relevant data?
- ✓ How will you word probing questions to build on,
- clarify or explain your interviewee's responses?
- How will you avoid asking leading questions that may introduce forms of bias?
- ✓ Have you devised an appropriate order for your questions to avoid asking sensitive questions too early where this may introduce interviewee bias?
- How will you maintain a check on the interview themes that you intend to cover and to steer the discussion where appropriate to raise and explore these aspects?
- How will you avoid overzealously asking questions and pressing your interviewee for a response where it should be clear that they do not wish to provide one?
- How will you avoid projecting your own views or feelings through your actions or comments?
- How might you identify actions and comments made by your interviewee that indicate an aspect of the discussion that should be explored in order to reveal the reason for the response?
- How will you listen attentively and demonstrate this to your interviewee?



To help you conduct your semi-structured or in-depth interview

- ✓ How will you summarise and test your understanding of the data that are shared with you in order to ensure accuracy in your interpretation?
- Where appropriate, how will you deal with difficult participants while remaining polite?

Recording data during the interview

0

✓ How will you record the data that are revealed to you during the interview? Where this involves

Semi-structured or in-depth interviews

using an audio-recorder, have you raised this as a request and provided a reason why it would help you to use this technique?

- How will you allow your interviewee to maintain control over the use of an audio-recorder, where used, if they wish to do this?
- ✓ Have you practised to ensure you can carry out a number of tasks at the same time, including listening, note taking and identifying where you need to probe further?

Closing the interview

How will you draw the interview to a close within the agreed time limit and thank the interviewee for their time and the data they have shared with you?

Primary data: Questionnaires

Written/printed or computer-based schedule of questions and a pro forma to record answers to the questions.



The choice of the questionnaire is influenced by:

- Characteristics of respondents (postal questionnaires for Millennials?)
- Importance of reaching a particular person as respondent (e.g., face-to-face vs. Internet)
- Importance of respondents' answers not being contaminated or distorted (e.g., postal questionnaires may be influenced by other people in the household)
- Sample size needed for the analysis (internet vs. face-to-face; self-completed vs. participant-completed)
- Type of questions needed (e.g., close questions; Likert-type scale)
- Number of questions needed (e.g., 40 questions: self-completed vs. participant-completed?).

e.g.

- <u>Postal (mail) questionnaires</u> do not necessarily ensure the respondent is the person who is being targeted.
- Internet questionnaires with a hyperlink may offer greater control, BUT do not ensure the 'quality of responses'.
- Interviewer-completed questionnaires ensures the participant is whom you want BUT have less anonymity.
- <u>Respondent-completed questionnaires</u> are quicker and relatively anonymous BUT may lead to incomplete and/or frivolous responses.

Primary data: Questionnaires

Individual questions

- Open questions (e.g., list three things you like about...)
- List questions (e.g., select your favourite brand from the list below)
- Category questions (e.g., how often do attend live sport events? 1 per week vs. 1 per month vs. 1 per year; etc.)
- Ranking questions (e.g., indicate the factors listed below by order of importance when making the decision to attend a live sport event)
- Rating questions (e.g., Indicate your level of agreement with the following sentences; 1= totally disagree and 7=totally agree)
- Quantity questions (e.g., what is your year of birth?)



<u>https://www.onlinesurveys.ac.uk/</u> <u>https://www.qualtrics.com/uk/?rid=ip&prevsite=pt-br&newsite=uk&geo=GB&geomatch=uk</u> https://www.surveymonkey.com/

Questionnaires: wording of the questions and order

Box 11.10 Checklist

Your question wording

- Does your question collect data at the right level of detail to answer your investigative question as specified in your data requirements table?
- Will respondents have the necessary knowledge to answer your question? A question on the implications of a piece of European Union legislation would yield meaningless answers from those who were unaware of that legislation.
 Does your question appear to talk down to
- Does your question appear to take or the respondents? It should not!
- Does your question challenge respondents' mental or technical abilities? Questions that do this are less likely to be answered.
- Are the words used in your question familiar to all respondents, and will all respondents understand them in the same way? In particular, you should use simple words and avoid jargon, abbreviations and colloquialisms.
- Are there any words that sound similar and might be confused with those used in your question? This is a particular problem with interviewercompleted questionnaires.
- Are there any words that look similar and might be confused if your question is read quickly? This is particularly important for self-completed questionnaires.
- Are there any words in your question that might cause offence? These might result in biased responses or a lower response rate.
- Can your question be shortened? Long questions are often difficult to understand, especially in interviewer-completed questionnaires, as the respondent needs to remember the whole question. Consequently, they often result in no response at all.
- Are you asking more than one question at the same time? The question 'How often do you visit your mother and father?' contains two separate questions, one about each parent, so responses would probably be impossible to interpret.
- Does your question include a negative or double negative? Questions that include the word 'not' are sometimes difficult to understand. The question 'Would you rather not use a

non-medicated shampoo?' is far easier to understand when rephrased as: 'Would you rather use a medicated shampoo?' Is your question unambiguous? This can arise from

- Is your question unantiguous: wind several poor sentence structure, using words with several different meanings or having an unclear investigative question. If you ask 'When did you leave school?' some respondents might state the year, others might give their age, while those still in education might give the time of day! Ambiguity can also occur in category questions. If you ask employers how many employees they have on their payroll and categorise their answers into three groups (up to 100, 100–250, 250 plus), they will not be clear which group to choose if they have 100 or 250 employees.
- Does your question imply that a certain answer is correct? If it does, the question is biased and will need to be reworded, such as with the question 'Many people believe that too little money is spent on our public Health Service. Do you believe this to be the case?' For this question, respondents are more likely to answer 'yes' to agree with and please the interviewer.
 Does your question prevent certain answers from
- being given? If it does, the question is biased and will need to be reworded. The question 'Is this the first time you have pretended to be sick?' implies that the respondent has pretended to be sick whether they answer yes or no!
- ✓ Is your question likely to embarrass the respondent? If it is, then you need either to reword it or to place it towards the end of the survey when you will, it is to be hoped, have gained the respondent's confidence. Questions on income can be asked as either precise amounts (more embarrassing), using a quantity question, or income bands (less embarrassing), using a category question. Questions on self-perceived shortcomings are unlikely to be answered.
- ✓ Have you incorporated advice appropriate for your type of questionnaire (such as the maximum number of categories) outlined in the earlier discussion of question types?
- Are answers to closed questions written so that at least one will apply to every respondent and so that each of the responses listed is mutually exclusive?
 Are the instructions on how to record each answer clear?



Your question order

- ✓ Are questions at the beginning of your questionnaire more straightforward and ones the respondent will enjoy answering? Questions about attributes and behaviours are usually more straightforward to answer than those collecting data on opinions.
- ✓ Are questions at the beginning of your questionnaire obviously relevant to the stated purpose of your questionnaire? For example, questions requesting contextual information may appear irrelevant.
- ✓ Are questions and topics that are more complex placed towards the middle of your questionnaire? By this stage most respondents should

be undertaking the survey with confidence but should not yet be bored or tired.

- Are personal and sensitive questions towards the end of your questionnaire, and is their purpose clearly explained? On being asked these a respondent may refuse to answer; however, if they are at the end of an interviewer-completed questionnaire you will still have the rest of the data!
- Are filter questions and routing instructions easy to follow so that there is a clear route through the questionnaire?
- ✓ (For interviewer-completed questionnaires) Are instructions to the interviewer easy to follow?
- Are questions grouped into obvious sections that will make sense to the respondent?
- ✓ Have you re-examined the wording of each question and ensured it is consistent with its position in the questionnaire as well as with the data you require?

Review Questions

- 1. Why are sampling techniques commonly used in research?
- 2. Explain the four stages in the process of probability sampling.
- 3. What non-probability sampling techniques can be used?
- 4. What is the difference between probability and non-probability sampling?
- 5. Explain the difference between primary and secondary data.
- 6. Give three examples of situations where you might use secondary data as part of your research.
- 7. Explain the potential advantages and disadvantages of using secondary data.
- 8. What should one consider when evaluating secondary data sources?
- 9. What is participant observation and what different types may be used? Does it fall under primary or secondary data? Why?
- 10. What types of interviews can one consider? Elaborate on the different types.
- 11. What types of questionnaires can be used in research? What aspects should guide the decision to select a certain type of questionnaire?

Saunders, M., Lewis, P. and Thornhill, A., 2020. *Research methods for business students.* 8th ed. Harlow: Pearson Education. (Chapters 7-11)

Veal, A. J. and Darcy, S., 2014. *Research methods in sport studies and sport management*. A practical guide. New York: Routledge. (Chapters 7, 9 and 13)

Gratton, C. and Jones, I., 2015. *Research methods for sports studies*. 3rd ed. New York: Routledge. (Chapters 8-12)

Collis, J. and Hussey, R., 2014. *Business Research: A Practical Guide for Undergraduate and Postgraduate Students.* 4th ed. New York: Palgrave Macmillan. (Chapters 6 and 10)

