Statistics for Computer Science

Lecture 03

Sampling Bias, Sampling Methods

doc. PaedDr RNDr. Stanislav Katina, PhD.
Institute of Mathematics and Statistics, Masaryk University
Honorary Research Fellow, The University of Glasgow

Sampling bias

Sampling bias most often occurs through a ***lack of forethought rather than the deliberate manipulation of data***, and can arise for a variety of reasons:

A. Bias in regard to **geographical site or situation and the overall size and controlling factors behind your research question**; these could relate to water supply, elevation or geology in a physical geography setting or cultural and economic factors in a human geography context.

B. Bias introduced when **sampling at an unrepresentative time/or period**.

C. Bias introduced through **the method you use to collect your data**.

D. **Too small a data set**, even when important controlling factors have been weighed up as part of the sample design, can lead to bias where **the underlying variability of the population is relatively high compared with the sample size**.

Sampling bias

Bias in the *Oxford English Dictionary* is defined as follows:

'***An opinion, feeling or influence that strongly favours one side in an argument or one item in a group or series***.'

In a scientific sense, your intuition or opinion may be correct, but it is vital that you do not **invalidate your research** by rigging the data in a way that **favours the outcome** you desire, intentionally or otherwise.

Sampling bias (**situational bias** and **situation specifity**) occurs when ***the sample is, in fact, unrepresentative of the target population***; it favours some elements over others.

In geographical terms ***participants should reflect the salient characteristics of the population about which inference is being drawn***.

Sampling bias and Disraeli's argument

At the root of the phrase:

'***Lies, damned lies, and statistics***'

(attributed to Benjamin Disraeli) is this issue of bias; Disraeli's argument may have related to the manipulation of state figures rather than scientific statistics, but the underlying issue is identical.

If the sample is biased the results that you might subsequently claim would lack credibility, be unpublishable and have a high chance of being misleading.

Examples of sampling bias

EXAMPLE: An article was written for a newspaper in the late 1980s concerning the **drinking habits of students**. This story did actually run, but is reported here anecdotally and from memory. The gist of the piece was that, on the basis of evidence from approximately 40 students from one particular college of one particular university, all students spent a high proportion of their government grant on gin and tonic; hence, the government grant to students should be abolished. Prima facie this not a geographical topic, but if we look carefully there are a number of space-time issues that emerge that are common to many more overtly geographical research projects.

Examples of sampling bias

- **Were the data collected in a bar at a time when gin and tonic was on special offer**, or in a week of celebratory significance? In other words, were the data temporally representative of the 'norm'?
- In a British context, and as a generalisation, gin and tonic is a combination of drinks more commonly (if not necessarily) **served in the south of the country**. The volume of all alcoholic drinks consumed, or even all spirits, would have been more representative of the nation's student population as a whole.
- Forty students was a **desperately small proportion of the national student community** at that time; the volume of data in comparison with the overall scene you wish to represent matters. ***Sample size is not everything, but low sample sizes can certainly contribute to bias where the overall population is variable as in this case***.
- The article itself had a tendency to **support stories that favour a right-of-centre political viewpoint.** That is, the results appeared to have been used to support '***An opinion, feeling or influence that strongly favours one side in an argument***.

Examples of sampling bias

Intuitively we will realise the following flaws in the sampling plan that led to an unlikely conclusion:

- **We do not know whether the proportion of males to females in this study matched that of the overall student population in the UK** at that time, but given that in the 1980s female drinking was less prevalent than it is now, this proportion could have been material even to the outcome of the small study.
- **We do not know the religious background of these students**. The consumption of alcohol is not a predominant feature in Muslim society, for example, so in this particular case the religion of the students is a salient issue.
- Both the gin and tonic association and the particular name of the college and indeed university at the time in question were, in a British context, suggestive of a **particularly privileged economic background amongst this student group**. To suggest that the results were indicative of the behaviour of all students was to extrapolate the findings beyond the supporting evidence.
- **Not all students drink gin and tonic as a matter of preference**, wherever they study. Were the data collected in a bar particularly known for its range of spirits? In other words, were the data collected at a representative range of sites?

Examples of sampling bias

SUMMARY:

This all adds up to a case of biased statistics: too few students who in all likelihood were from similar economic and religious backgrounds, of unknown gender, compounded by geographical bias at the **micro scale** (one bar) and **macro scale** (southern England being used to represent the UK).

Examples of sampling bias

EXAMPLE: In 1948, the *Chicago Daily Tribune* was so confident of its opinion polls that it went to print with the title '**Dewey beats Truman**'; but Truman won. The **public opinion polls** prior to the British 1992 election are a more recent example of statistical disaster as a result of bias (Smith 1996).

Even when great care is taken designing samples to avoid causes of bias, the unexpected can happen as shown in the second example, above. The cause of the **highly embarrassing mistake** in the Tribune case was that the opinion polls were based on telephone surveys, but in 1948 many people with lower incomes did not own telephones (McAfee 2002, p.226). George W. Bush too was predicted to lose the 2005 US election with a heavy defeat according to the exit polls. In fact, he won convincingly.

Primary and secondary data

In bother examples, the ***data were collected for the direct purpose of the studies reported***. Such data are referred to as **primary data**.

Dates gathered by a researcher as part of a study, whether by interview survey, experiment or observation, fall into this category.

In contrast, **secondary data** are data that have been collected as part of a separate study and potentially a different purpose, but that offer potential value to your work. These data might for example include published government statistics, meteorological records, elevation surfaces and remotely sensed land cover classifications.

Because you have not controlled the collection of these data in order to make sure that the sample is representative in regard to the process that you yourself are researching, the probability of sample bias creeping into a study using secondary data is strong.

Whether the data set is **fit for use** in your particular study must be carefully weighed.

Examples of sampling bias

Bringing this scenario into the present, if you are undertaking a **geographical project** today, for example looking at the effectiveness of web media for communicating local issues, or evaluating the effectiveness of public participatory GIS (PPGIS) using visual media for community decision making, you should consider whether you need to seek out those without easy access to the Internet as well as those who can connect easily.

Broadband access is not uniform geographically, economically or across age-cohorts universal.

Examples of sampling bias

EXAMPLE: *UK Meteorological Office land surface temperature archive* as an example of a **secondary data** set that we wish to consider for the purpose of creating raster maps at 1 km resolution for climate variables over England and Wales. We also wish to compute associated summary descriptive statistics for England and Wales as part of our projec.

Like other **national meteorological networks**, great care is taken by the UK Meteorological Office that **measurements adopted within the archive use instruments calibrated to a certain accuracy, based on equipment at sites conforming to carefully specified local characteristics and using well-documented observing and recording protocols** (Met Office 2010). A very positive aspect to this particular secondary data collection is that the **metadata** associated with it are strong; **we know what we have, and how and when it was recorded**. This is to ensure that users can evaluate whether the data, and their consistent protocols, are fit for the purpose they need them for.

Key points: Metadata

Metadata are formally described as '***data about data***'. The term encompasses information such as ***the date of data collection***, the age or address of the informant, spatial resolution, attribute precision, sample size, method by which data were acquired through processing in the lab and the type of instrument used to make the recording.

Examples: Fitness for purpose (use)

EXAMPLE: **The locations for which daily temperature data are available for a period we wish to study against the national picture for England and Wales** (Jarvis 2000).

The **mean elevation** of 174 sample meteorological recording stations, scattered across England and Wales, was **83 m**. *Based on a GIS analysis using Ordnance Survey 50 m raster elevation data, and using the central point of each 1 km cell across the landscape of England and Wales, the mean elevation for England and Wales was computed at **125 m***. Further, the standard deviation in elevation within a 50 km2 area around each 1 km cell across the country was 47 m based on the sample points and 60 m for the overall landscape at 1 km resolution. These figures show that the Met Office collection is **biased towards lower and flatter elevations relative to the overall situation for England and Wales**. An **average temperature** for the total area taken from the subset of 174 archival records, is likely to be higher than reality. Why might this be so?

Key points: Fitness for purpose (use)

Fitness for purpose (use) is the **effectiveness** and **appropriateness** of the data **for the particular task for which they are to be used**. Factors affecting fitness include **where and when the data were collected and by whom** (expert or amateur; scale and place of collection; numerical precision of record). Key here is the interplay between defined purpose and characteristics of the data.

Examples: Fitness for purpose (use)

The primary purpose behind these national meteorological records is **the forecasting of daily weather**. Accurate weather forecasts are particularly pertinent for the building and agricultural industries and for air travel. Thus, many meteorological stations are based at **air airports** or **airfields**, ***generally not sited in mountainous regions***; local sources of meteorological data are more commonly available in **lowland agricultural regions**, again which tend to be at lower elevations. ***We might conclude that the data archive is fit for its main purpose, but is it fit for our purpose?***

The answer to this question depends to a large degree on the **intended purpose for the temperature surfaces themselves**; ***the data are biased***, but if we determine that the temperature surfaces are to be used for agriculture-related modelling, the outputs for which are less relevant in mountainous areas, then we can still conclude that the data set is ***sufficiently representative for our task***. The geographical coverage, longitudinal nature of the data set (data collected over a number of time periods, in this example multiple years) and the quality assurance applied to the archive also contribute to this assessment.

Fitness for purpose (use)

Increasingly, data collected as part of large collaborative scientific experiments are archived for **access from the Internet**, yet, as you may discover for yourself, **supporting metadata are often sparse**. ***This makes bias more difficult to spot.***

Overall, however, **the same questions you need to consider in regard to sample design for primary data should be asked of a secondary data source**.

If critical questions cannot be answered directly or by analysis as above, the data should not be used in your study.

Sampling methods

Sampling methods can be divided into two groups:

1. **probabilistic methods** –
   A. systematic,
   B. simple random,
   C. stratified random,
   D. multi-stage random and
   E. clustered random method
2. **non-probabilistic methods** –
   F. judgemental,
   G. snowball,
   H. quota and
   I. convenience

Where you wish to deepen your study to go beyond the descriptive statistical techniques and use one of the analytical, inferential statistical techniques, you will need to adopt a probabilistic sampling method in your work that allows sampling error and representativeness to be modelled.

Key points: Sample design

Sample design refers to the choice of method and the scale and detail of how you plan to implement your approach.

Absolutely key to this process is having **a clear research question**.

Before this phase, you also need to have developed a broad understanding of the factors that are likely to affect the human or physical process you are investigating through a thorough search of the literature, **the scientific scope of your study**, and weigh up the scale(s) at which you consider the processes are operating.

Non-probabilistic sampling

It is not advisable to use these methods when you wish to conduct inferential statistical analyses.

Why? This is because models (such as probability theory) cannot be used to determine sampling error when the researcher has selected the sampling elements, making it difficult to quantify how representative the sample might be.

However, where your **purpose is exploratory** and the **statistics** are intended simply to be **descriptive**, for difficult or rare populations, or where the development of rapport with participants is vital to your study, this class of sampling can be valuable.

It is commonly of greater relevance in **qualitative research**, where data can only be represented on nominal scales and cannot be quantified in numeric terms.

Judgemental sampling

Judgemental sampling involves the researcher making sampling selections based on their ***prior experience to judge which samples to leave in or out to construct a representative sample overall***.

Where variety as opposed to representativeness is sought, this type of purposive sampling methodology is useful.

EXAMPLE: If you were to record the **opinions expressed in a public planning meeting** of some type as part of your qualitative research evidence, this would **gather the extent of views but with an emphasis on those most vocally expressed**.

EXAMPLE: One might equally make contact with a **prominent local councillor** to speak about **the issues of a small town** on the basis that that person encounters the views of many residents as part of their role.

Snowball sampling

Snowball sampling involves ***sample members identifying other suitable candidates with particular desired characteristics, a 'chain letter' type of approach***.

Geographically speaking, and perhaps counter-intuitively, the effect may be to achieve a highly dispersed sample group in a geographical sense but **it is very unlikely to be representative**.

Note that sending a questionnaire into work with a friend or parent is in reality a form of snowball sampling and such data should not subsequently form the basis for inferential tests.

For ***populations about whom data is difficult to gather and where rapport is important***, for example when considering **homeless populations** or **activist groups**, this **selection-by-recommendation approach** is likely to be highly valuable when used as a basis for further qualitative analysis.

A quota approach could also be used ***to select a sample from those initially recommended by the snowball method***.

Quota sampling

Quota sampling is a sampling method, where ***data are added to a sample until a predefined quota is reached***.

EXAMPLE: If you were making a **study of unemployment in relation to use or access to recreational facilities**, two groups might be selected where the overall proportions of employed to unemployed between the two groups matches that of the government statistics for the area. Critically, **there is no randomised element within this selection**.

One might target the first 200 people to arrive at the nearest railway station for a train between 8 and 8:30 in the morning, who meet important criteria relevant to the study (for example, travelling to work in London for a study on city commuting) or the first 30 people to arrive at a government job-seeking unit on a particular morning.

This method is the most likely of the non-probabilistic group to result in a representative sample being gathered.

Convenience sampling

Convenience sampling is a method, where ***accessibility is the key to selection within a trial***.

EXAMPLE: **An academic developing a new data visualisation technique designed to improve learning** might for example invite their Bc or MSc class to take part in trials regarding its effectiveness for certain tasks on a voluntary basis.

The outcomes would provide initial information and feedback to the researcher, but may prove to be **unrepresentative of the overall student cohort of Bc or MSc students** in the subject nationally, since the convenience approach in this particular case would have ***a tendency to interest students particularly attracted to a visual learning style***.

Systematic sampling

The sampling frame is ***sampled regularly*** (***systematically***), **starting from a randomly selected first point or element**.

EXAMPLE: Lay out **a quadrat** to record biogeographical data **in a grid pattern every 5 m across your area.**

EXAMPLE: select **every 100th person in the electoral roll in sequence** to receive your questionnaire.

Your **sampling frequency** will be determined by the number of samples you wish to collect from the sample frame itself.

Typically, the **total sample frame** is divided by the **sample size** to determine the **sample interval**. The sampling interval could be expressed in ***distance terms***, or by the ***number of intervening records in a list***.

EXAMPLE: A small *systematic sample* being used for **bulk density** and **soil moisture** of a **60 × 60 m2 patch** of the *Karoo National Park in South Africa*. The **sample locations** are **regularly placed**, and in this case an identical sampling structure has been used for both variables.

Simple random sampling

There is an equal chance of selecting one element of the population from another.

Selection is made via a **random number algorithm**, and **without replacement**.

EXAMPLE: A small *simple random sample* being used for **bulk density** and **soil moisture** of a **60 × 60 m2 patch** of the *Karoo National Park in South Africa*. The **sample locations** are **regularly placed**, and in this case an identical sampling structure has been used for both variables.

Simple random sampling carries inefficiencies relative to systematic samples with or without a random element.

EXAMPLE: As Pearson and Rose (2001) demonstrate in their study **estimating the magnitude and distribution of radioactive caesium in a Tennessee reservoir**, the **relative inefficiency** of simple random sampling as a method is not an exception for spatial studies.

In his book *Statistical Rules of Thumb*, van Belle (2002) generalises the situation and suggests that ***you should 'always consider alternatives to simple random sampling for a potential increase in efficiency, lower costs and validity***.

Systematic sampling: problems

The major difficulty when using systematic sampling arises ***in relation to periodicity within the underlying phenomenon***.

EXAMPLE: It is a common cultural phenomenon in the UK for houses on one side of a street to receive odd numbers and the other, even numbers.

Problem: An extreme outcome might be that if you sampled every other household in a street systematically, you might find that your sample covers every house on one side of the street at the expense of the other.

EXAMPLE: In a physical geography setting, **terrain might vary systematically** also; consider the drumlin fields in western Scotland mapped by Smith et al. (2006) in which terrain undulates approximately every 300 m in a classic '***basket of eggs' topography***.

Systematic sampling will cause **bias** where the ***underlying structure of the phenomenon being measured has the same spatial structure as that of the sample interval***, or indeed the ***phenomenon measured varies at a smaller interval than that chosen for measurement***.

Simple random sampling: problems

Simple random sampling meets the probabilistic requirements for a representative sample.

In **reality**, you should suspect that ***sampling units are highly heterogeneous in an unevenly distributed manner***, and you should consider an alternative method such as stratified random sampling.

There is also a **danger** that ***random sampling can miss capturing small spatial structures*** unless the sample is particularly large, owing to the ***uneven distribution of the measured values***.

Stratified random sampling

To achieve representative data in a geographical setting, the **stratified random sample** is particularly important.

This is especially so where the population units are more similar within each stratum (for example, land cover, elevation class, geology) ***than they are across the strata***.

We distinguish to types of this sampling:

1. **proportionate stratified sampling** and
2. **disproportionate stratified sampling**

Stratified random sampling

EXAMPLE: In a *human geography* setting, you might use **statistics derived from the census *to select similar but mutually exclusive wards within which to conduct interviews to randomly selected households***. This allows sub-groups within the population, important to the study, to gain **fair representation**. Stratification also might usefully be undertaken by **gender** or **age categories**.

Be aware too that **stratification** might usefully be carried out ***over time rather than space*** in a longer running study.

EXAMPLE: Exposure to particulate matter such as PM4 has a strongly **seasonal component** (Fanshawe et al. 2008).

Stratified random sampling

Under a ***proportionate stratified sampling scheme***, **the number of units drawn from each stratum is proportionate to the population of the underlying strata**. This provides a sample representative of the overall population.

Wider availability of digital coverages for soil, geology, elevation and census units amongst others make the use of GIS as part of the sampling design process something worth considering.

Where strata form the basis for the comparative study, for example the variation of a process by geology, you require data that has been selected randomly in all respects apart from underlying geology to provide an appropriate comparison.

The sampling frame, or area, would be divided into areas according to geology and, under a ***disproportionate stratified sampling scheme***, **identical numbers of samples would be sampled randomly from each area to provide data that were not biased to one** geological type **over another**.

Stratified random sampling: problems

There are similarities between **quota sampling** and stratified sampling.

However, the **critical difference** is that the ***selection of individuals*** (be these points or people) ***is not truly random for quota sampling and researcher bias can therefore potentially affect selections***.

For this reason, ***stratified approaches carry less risk of overgeneralisation and potential administrative benefits***.

EXAMPLE: In an environmental study, the number of **agencies** or **rangers** that you need to contact to discuss access to your target population is potentially reduced.

Overall, stratified random sampling is a commonly used and effective method for sampling geographical phenomena and is regularly used within academic research.

It does, however, rely on accurate information about the processes affecting the phenomenon you are researching and the availability of adequate information concerning appropriate strata.

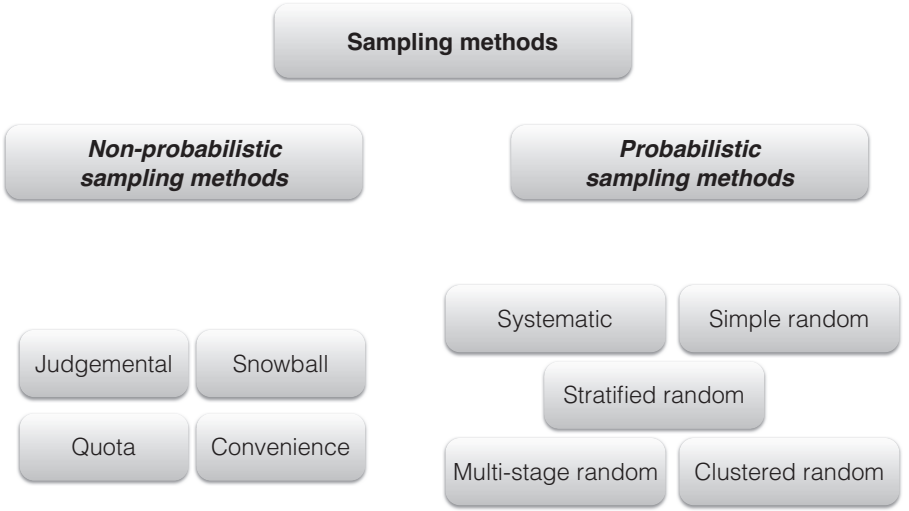
Multi-stage random sampling

In **multi-stage random sampling**, ***the sampling frame is divided into multiple hierarchical levels. Each level is sampled randomly in order to select the elements to be included in the next level.***

When is this method effective? ***The target population is very large in a temporal or spatial sense.***

EXAMPLE: If we were to conduct the undergraduate student **alcohol consumption** worldwide for example, we might select a **subset of countries**, then a **subset of universities** and then, finally, a **subset of students** from this set to interview.

The multi-stage random approach is also useful for ensuring that a ***full range of scale separations is present in the sampling***.



Clustered random sampling

This method is identical to the multi-stage random method, except for the fact that all elements are included at the final (most detailed) level.

The method is most commonly used **when access to the complete sampling frame is unavailable**.

EXAMPLE: We might for example **randomly pick six areas from a gridded geographical sample frame** and then **comprehensively survey each of these** for evidence of a rare plant or animal.

The critical issue here is that the sample members are chosen as a group (or cluster) rather than as individuals.

Sampling error and sample size

As sample size increases, sampling error decreases. More **homogeneous populations** will have a **lower sample error**, given the same sample size, than a **more variable population**.

If you are able to measure your data directly within the field, an **empirical rule of thumb** is to stop sampling when the standard deviation of your measured values achieves a degree of stability. For indirect measurements, a two-stage sampling process is recommended.

Another **common-sense approach** to determining how many measurements you may need is to draw on the experiences of a published study using a similar population to yours.

Measurement accuracy

It is essential that you consider ***sample measurement accuracy*** as a component of ***sample design*** and ensure that your data collection strategy is ***fit for purpose*** as regards precision and scale.

Measurement error can be caused by both instrument and observer.

EXAMPLE: As a general rule, **studies using GPS** to record elevation across local areas for use in numerical analysis work should be avoided unless you have access to differential GPS equipment and training in the postprocessing of data. It might seem tempting to use handheld single unit GPS to estimate **cliff height**, for example. If, however, you consider the **strong likelihood of a 10 m error** in both upper and lower measurements, the overall error as a portion of cliff height will be high. It is very often the case that the nearest contour will provide a better estimation of elevation than single GPS.

StKa, Oct 20 2015