Statistics for Computer Science

Lecture 02

Data and Variables, Introduction to Sampling

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

Data and variables

A single observation *cannot measure change or variation, and there is the risk of the measurement being wrong.*

Therefore ***a series of observations*** *is made by recording measurements for different times or places within the study region.*

The set of measurements is known as a ***variable****. It is called a variable for good reason. Looking at the observations will reveal variations (variability) in the measurements –* ***they are not all the same****.*

There are two reasons for the variations (variability).

1. *Because* ***what is measured is not stationary*** *– it has different characteristics at different times and places.*

EXAMPLE: unemployment is greater in some places and in some years more than others; the same is true of air pollution.

2. *There are other things that degrade the quality of the measurement, creating error, that do themselves vary across time and space.*

Given both these considerations it would be surprising if all observations had exactly the same value.

Data and variables

Forming knowledge from data is central to statistical analysis.

Data are information obtained by measuring 'things' at different times or locations across a study region.

Each measurement is called an **observation**, because it is a record of what has been observed where and when the measurement was made.

That 'thing' – **the focus of the study** – will be a social, scientific or environmental feature, or process, that we want to learn more about.

EXAMPLE:

- **geographical** – elevation, surface temperature, ozone levels, soil quality, unemployment, noise, traffic congestion, access to shops and amenities, voting behaviour, water salinity, slope stability, vegetation cover, life expectancy, environmental quality, quality of life, happiness, crime rates, etc.
- **medical, biological** – laboratory variables, immunological variables, anthropological measurements (length, angles, width, height, circumference, volume, etc.)
- **economical** – GDP, exchange rates, ...

Key points

1. **Data** are measurements of something of interest. They are also called **observations** because the measurements help us to observe (and to quantify) an attribute of whatever is being studied.
2. A set of measurements is called a **variable** because the values are unlikely to have the same value at all times and locations at which the measurements were made.
3. Descriptive statistics **summarise** key information about the variable. Other types of statistics **seek to explain the causes of the variations** (variability).

Data types

It is important to distinguish between different types of numeric data because ***what you can do with data is related to the types of data they are***. This is true for more basic descriptions of data and for more advanced statistical methods.

However, thinking about data types is not only about ***matching the data to an appropriate method of analysis*** – though that is important. To consider the data type is also ***to give thought to the properties of the phenomenon under study and to the sorts of data it can generate, as well as to how the phenomenon has been conceptualised and observed by the researcher.***

The data type is a product of what is measured, how it is measured and why it is measured.

Discrete data

*Discrete data take on one from a limited (and therefore **finite**) set of possible values. Because of this it is possible to **count how many times each specific value appears in the data (countable)** and to produce a tally – a frequency table – of those counts.*

*Discrete values tend to be **whole numbers**, also known as **integer numbers**. These are values which have no fractional part and are written without a decimal point.*

EXAMPLE: 1 is an integer whereas 1.1 is not. The value -2 is also an integer but -2.0 is not: **the inclusion of the decimal point implies -2.0 is from a set of non-integer values that might include -2.1 or -2.2, and** -2.0 might itself be an approximation of -2.04 for example.

Why we are speaking about this common situation here?

Data types

*One of the most common ways of defining data types is by using the **Stevens scale** (Stevens 1946). However, it may also be too strict to apply to real-world data and, in any case,*

> scale type, as defined by Stevens, is not an attribute of the data, but rather depends upon the questions we intend to ask of the data and upon any additional information we may have. It may change due to transformation of the data, it may change with the addition of new information that helps us to interpret the data differently, or it may change simply because of the questions we choose to ask. (Velleman and Wilkinson 1993, p.69)

*In general, we distinguish between **discrete** and **continuous data**.*

Continuous data

*In contrast, continuous data are drawn from an **infinite set** and can take on any value – or, at least, any value between a lower and upper limit. They are often 'real' or 'floating-point' numbers, those with a decimal point.*

EXAMPLE: 1.001 is a real number, written to three decimal places (there are three digits after the decimal point). Another is -4.112 34, with five decimal places.

Why we are speaking about this common situation here?

*Because there are an infinite number of values continuous data could take, it is futile to produce a frequency table for them. Many or all of the values that do appear will do so only once. Instead, we can arrange them into groups (for example, 0–4.99, 5–9.99, 10–14.99, etc.) and then count the number of members in each group. **The result will not be independent of the groupings**. We will see that the way we group continuous data affects our portrayals of them. These data are call **interval data**.*

Continuous data

It could be argued that no data are perfectly continuous since there are always limits to the ***precision*** *(****the number of digits****) by which events can be measured and recorded.*

No measurements are drawn from a truly infinite set. However, the difference between discrete and continuous data is better understood ***as a property of what is being measured than of the data themselves****.*

EXAMPLE: most light switches have two discrete states, either on or off, whilst the luminance of energy-saving light bulbs increases, on a continuous scale, from when they receive an electric current to when they are fully lit.

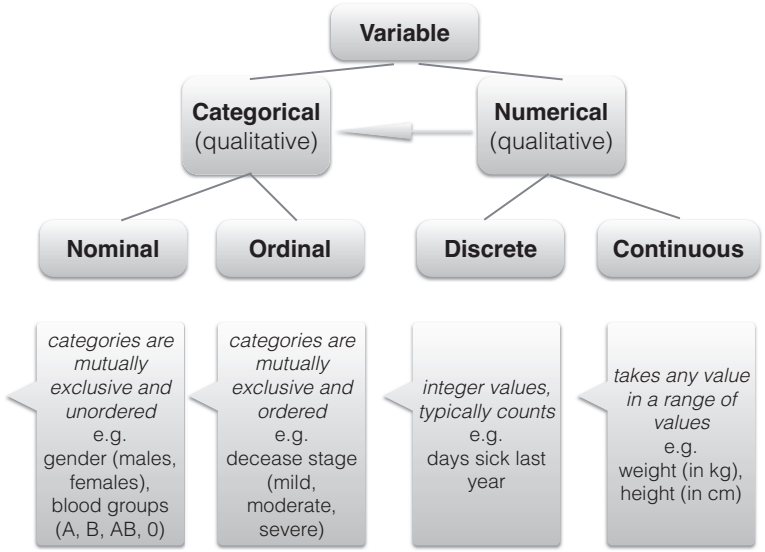
It is also better to understand the words discrete and continuous as ***two ends of a continuum along which different sets of data are placed****. It then becomes a matter of asking*

which is the better model for the data I have,

what I want to do with it and

for what I am trying to study?*'*

Integer data are never really continuous, but unless the set contains only a few (discrete) values then it is neither unusual nor necessarily mistaken to treat it as if it were continuous.



StKa, Oct 11 2015

Categorical data

An exception to our focus on numeric data will be ***categorical variables*** *– those that are labels distinguishing one category from others.*

EXAMPLE:

M and F, for males and females

1, 2 and 3 for different types of land use (1 = arable; 2 = forest; etc.)

'left' and 'right' for political allegiance

low, moderate and high for levels of risk

'good' and 'bad' as qualitative judgements; and so forth

We will use categorical variables to split a set of measurements into groups that are then compared. *We distinguish (Agresti 2007)*

1. ***ordinal data*** *– the categories are ordered in some way*
2. ***nominal data*** *– categories are not ordered but simply have names*
3. ***binary data***

StKa, Oct 11 2015

Key points

1. **Discrete data** are those that take on one from a restricted set of possible values. They are usually whole or integer data.
2. **Continuous data** could take on any value, or any value within a lower and upper limit and to a certain level of precision. They are 'real' or 'floating-point' numbers.

In practice, the difference between these data types is not fixed or immutable. Discrete data are often analysed as though they are continuous, and continuous data can be grouped into discrete categories (for example, by sorting height data into 'short', 'average' and 'tall').

StKa, Oct 11 2015

Derived data

Percentages

Ratios and quotients

Rates

Scores

All of these data can be treated as numerical variables for most analyses. Where the variable is derived using more than one value, it is important to record all of the values used.

Censored data *– either laboratory data below or above detection limit or failures in time where we are following subject for a certain period of time and recording a failure*

Sampling – learning objectives

- **Summarise** the following probabilistic sampling methods – **systematic, simple random, stratified random, multi-stage random and cluster sampling**.
- **Highlight** the effectiveness of different sampling methods and be able to **construct** a sample design that is effective and efficient for a given problem.
- **Appreciate** the relationship between **sample size**, **precision** and **confidence** in general terms.
- **Weigh up the practicalities** of a particular sample design, taking into account cost, safety, access and environmental/human ethics.

Sampling – learning objectives

- Know the **distinction** between a **sample** and the **population**.
- **Explain** and **recognise** examples of **sampling bias** and the notion of a **representative sample**.
- **Outline the stages involved in the process of sample design**.
- **Distinguish** between **non-probabilistic** and **probabilistic sampling methods**.
- **Summarise** the nature of the following non-probabilistic sampling methods – judgemental, quota, snowball and convenience sampling.
- **Explain** the **concept** of a sampling frame and be able to apply this to a study of your design.

Sampling

- It is often simply **impractical to collect all data relating to a task for logistical reasons including budgets, time and access**.

EXAMPLE: We cannot

- interview every adult within an electoral district,

- measure the girth of every tree within a forest or

- count every grain of sand on a beach.

- This would not be a very **efficient** way of going about your task even if a full survey (or census) was feasible.
- Largely, we do not need to gather a complete and full set of data relating to our task in order to draw **general conclusions** about the processes or phenomena we are investigating.
- There are always **exceptions to any generalisation**, and there are certain circumstances where it is possible and feasible to deal with a complete population, although these situations are rare.
- You could for example interview every individual on a planning committee regarding a particular decision. More normally in research we may **gather a sample** (or **subset**) **of data** ***from our target population, from which we may draw robust conclusions or develop scientific theories***.

Target population

- The target population can be defined as **the complete set of measurements that might hypothetically be recorded in a particular study context that are relevant to the study**.

EXAMPLE: In a research context this could include

- all people or vegetation within a defined area at a certain time, or

- the distributed activities of a single business unit operating from one location.

- Often, the notion of population is more theoretical than quantifiable.

- A population falling outside either the geographical area or context of a study is sometimes referred to as an **out-of-scope population**.

Representative sample

This is a sample that matches (or represents) the statistical characteristics of the overall target population.

The **characteristics of a sample** are known as its **statistics**.

Sampling

- The characteristics of a sample of a population are known as **statistics**.

- A sample that ***matches*** (or ***represents***) the population is known as a **representative sample**.

- Achieving a representative sample is a **key issue when adopting probabilistic statistical methods in your research**, but also applies when adopting other research methodologies whether qualitative or quantitative in approach.

- However, not all research traditions require that the outcomes are representative of the whole, as some work is **illustrative** in nature. It is important to respect these differences across the discipline; there is a place for the in-depth case study, particularly in regard to gaining more detailed knowledge about how a process or phenomenon emerges.

The process of sampling

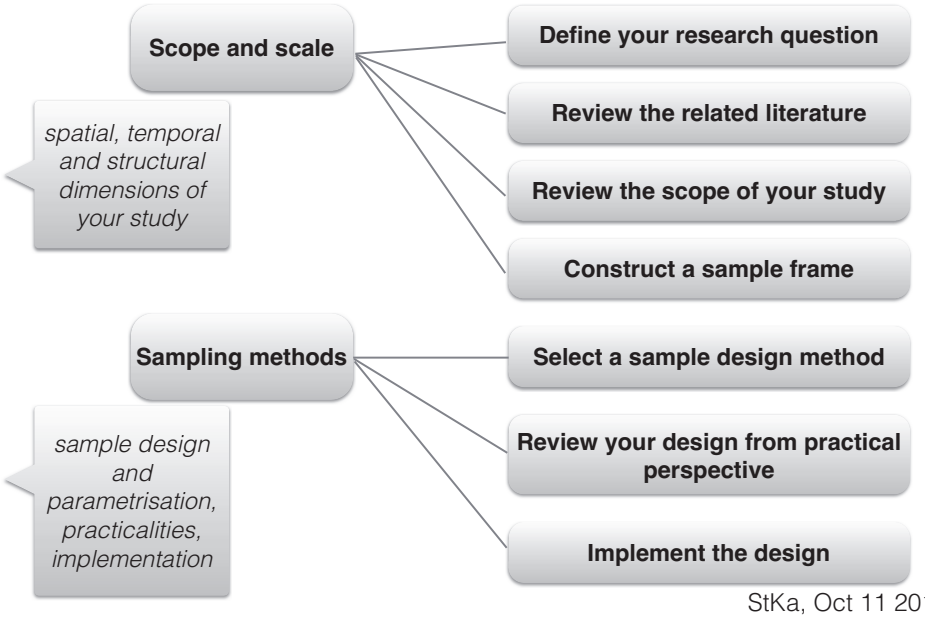
- Considering ***which design best meets the goals of your study*** forms an important next step in the process, with the underlying **aim** behind the process of sample design being to ensure that the ***subset of data collected reflects the characteristics of the overall target population***.

- Sub-questions here include ***how the design should be parameterised***, and ***whether you should first conduct a pilot study***.

- ***Theoretical ideals and practical realities conflict*** – practical pointers that may suggest that your sampling strategy needs modification before you go out in the field.

- And there is something more to consider – **any analysis is dependent on data and whether they are fit for purpose**.

The process of sampling



Formulating the research question

- What is your research question?
- What processes or phenomena do you expect to influence that question?
- What data should you be collecting?
- Have you reflected on whether the question is tractable given the time and resources at your disposal?
- How much do you already know about the data and underlying geographical process? Rather than making unreasonable and ungrounded assumptions, it would be appropriate to undertake a **pilot study** before continuing with your main sampling campaign if necessary.

EXAMPLE: Formulate your own research question, and answer other questions above. Discuss the answers in the class.

Formulating the research question

- Before jumping into specific methods for sampling, we need to be sure that ***the sample we are collecting is representative*** **for our purpose**.
- At this point, we need to stand back from the overall problem.
- First, ***we need to be absolutely clear about the research question we are trying to answer***.
- Having established this fundamental step in the overall, and much larger, research design process, we need to ask ***which environmental or human factors influence the process or phenomenon we wish to investigate***.
- We also need to consider ***the scope of the specific*** *study.*
- Is it **feasible**, in terms of both the **complexity of the question** and **geographical coverage** intended that you could collect **sufficient data** to answer the question **in the time you have available**?
- It is also important to consider ***at what scale*** you might expect the process to occur or pattern to materialise at this early stage in the study and ***what associated assumptions*** it might be reasonable to make regarding the data.

Review the relevant literature

There is no substitute for detailed background literature work prior to developing a research methodology and statistical sampling design.

EXAMPLE: Commonly encountered factors, subject to the study in question, might include:

gender, sexual preference, family background, cultural background, economic background, educational background, computer literacy and access, age, niche personal interests, political preferences, subsurface or deep geology, longer-term environmental history of an area (for example, flooding, fire, uplift), meteorology and/or climate, depth, elevation, soil type, land cover, land use, basin size, distance from the sea, aspect, slope.

This long, but hardly comprehensive, list serves to illustrate ***the need to constrain the scope of your study***.

The tighter your reference question, the easier it is to avoid a biased sample.

The scope of the study

***Formulating precisely a research question that is as specific as possible is the easiest way of managing* the scope of your study.**

The scope of your study can be further tightened in two main ways:

- Firstly by **reducing the research extent** of your study (narrowing your focus).
- Secondly by **choosing sample sites** such that a number of potentially important factors affecting the process you are observing are held steady. This second practice is known as **controlling your variables**.

Controlling the variables

EXAMPLE: A solid piece of research looking at the effect of geology on river incision is better managed by **comparing two basins of similar size and rainfall patterns** than trying to undertake a complex model that accounts for varying basin size, rainfall volume and rainfall intensity in one go. This is because achieving a representative sample in the latter case would be a sizeable task before you could achieve any statistical certainty in your results, owing to the multiple potential interactions between variables.

EXAMPLE: You might choose to investigate **the impact of an environmental variable** (for example, distance from a power station or mobile phone mast) **on health** by comparing groups of similar age, economic and genetic background who have lived in the area for similar lengths of time.

Research extent

A look at the titles or abstracts of journal papers in areas of the discipline that particularly interest you can be instructive here.

EXAMPLE: Taking one example of how you might construct a focused dissertation title, consider the following title: '**The geography of shoplifting in a British city: evidence from Cardiff**' (Bromley and Thomas 1999).

The title shows that the authors are interested in the phenomenon of British shoplifting as a whole, but indicates with clarity that their evidence in regard to the complete country is partial and geographically specific.

Sampling frame

The **sampling frame** contains all possible data (or **sampling units**) to be selected (the population); *it frames or outlines the data set*.

Only data within the sampling frame may subsequently be selected for analysis.

Sampling frame is the actual list from which your sample items are drawn.

EXAMPLE: It might be the **GIS data grid** subdividing your study area in the context of a *physical geography* example, or the rather more tangible **electoral roll** (register of voters) in the case of a *human geography* task.

Hierarchy of terms in probability sampling

We have outlined the population, target population, sample frame and sample. These relate as follows:

Scenarios and questions – scope, scale and extent

EXAMPLE: If you are looking at **biodiversity indicators**, do you need to record the individual proportions of named species in your sample notebook? Might it be better to cover a wider geographical range, noting number of different species, instead?

EXAMPLE: You are seeking data about **shoplifting** in the UK. Should you rely on secondary data concerning recorded shoplifting incidents, given that these are suggested to be but a small sample of the whole (Bromley and Thomas 1999), or should you weigh up other strategies to gather your data?

Sampling frame

Subject related sampling frame – electoral roll for an area, the list of students matriculated at a university or school, or the complete police record of all shoplifting crime reported in a particular place or timespan.

Geography related sampling frames

– the electoral ward or enumeration district, a list of postcodes in an area or another form of regional boundary

– the list of operating meteorological stations or pollution control instruments in an area, a watershed relating to a river feature under investigation or a digital map showing geological or elevation classifications for the study area under review

Sample units should be defined by **size** (for example, quadrat extent in a bio-geographical study), the **location** and **time** of records.

Scenarios and questions – scope, scale and extent

EXAMPLE: You are interested in the **impact of green lanes on butterfly distributions**. Butterflies are in general most active at high levels of sunshine. Is it really worth undertaking a sampling campaign at both 9am and 12pm, or would one set of observations at 12pm be just as effective?

EXAMPLE: You have been investigating **the effects of heavy industry on the quality of river water**. While you are there, you decide it might be interesting to look at lev“levels of phosphorus and nitrogen also. After you have analysed the different chemicals, at considerable time and expense, you realise that you have no data concerning the history of agricultural land use in the area and cannot make good use of your nitrogen and phosphorus data. Further, because you have undertaken more analysis than planned, your project may run late and runs the risk of not being completed. Think back, what might you have done differently?

Scenarios and questions – scope, scale and extent

In general, these scenarios arise for a variety of reasons.

1. First, there is doubt regarding the **research question** itself or a lack of consideration concerning the processes and variables influencing your research question.

2. Secondly, **focus** and **discipline** are required as part of an effective sampling campaign in the fieldwork. This second issue is more likely to trap the enthusiastic and diligent researcher, determined to succeed. There is also a danger in collecting data because they are what has been collected in the past, affordable, familiar or easy to collect.