What are the rows and columns of a data table?

**rows:** observation, cases

**columns:** variables, features(?)

What are xsec, tseries and xt panel data? What’s an observation in each? Give an example for each.

* **xsec data:** Cross-Sectional data that have information on many units at the same time. The order does not matter so the rows can be changed, the information content would not change.
  + *example: information on employees of a firm*
* **tseries data:** Time series data that have information on a ***single*** unit observed many times. The order is important.
  + *example: Growth of net income in a company*
* **xt panel data:** Multi-dimensional (panel) data that have multiple dimensions.
  + - Many cross-sectional units observed many times
    - Units observed in different space
  + *example: countries observed repeatedly for several years*

What’s the validity and what’s the reliability of a variable? Give an example of variables with high validity and one with low validity.

* **Content** - what is the substance a variable captures. Always check details.
* **Validity** - is the content of variable close to intended content. "Durability" vs "Quality"
* **Reliability**. If we were to measure the same variable multiple times for the same observation it should give the same result.
* **Comparability** in measurement across observations.
* **Coverage**. Ideally complete coverage. In practice, they may not include all planned units (incomplete coverage).
* **Unbiased selection**. In incomplete coverage, observations that are included shouldbe similar to all observations that were intended to be covered.

**Validity:** the collected data is as close to the intended content as possible. for example: the time is relevant to our analysis or they are ’out-dated”. Do they measure what they supposed to do?

high validity: data about ages if the analysis want to measure it

low validity: data about a price of a product from 3 years ago

**Reliability:** the data should not change by each answ79er or data collection. It should be exactly the same.

What’s selection bias? Give an example of data with selection bias and are without.

Selection bias is the term used to describe the situation where an analysis has been conducted among a subset of the data (a sample) with the goal of drawing conclusions about the population, but the resulting conclusions will likely be wrong (biased), because the subgroup differs from the population in some important way. Selection bias is usually introduced as an error with the sampling and having a selection for analysis that is not properly randomized.

good example: random sampling

bad example: choosing the first 10 observation a table from up to down

List two common advantages of admin data and two disadvanatges.

* Many advantages

Often great coverage, few missing values, high quality content

Many well defined and documented variables

* Some disadvantages

Variables defined for business/government purposes. May not fit in analysis plans

Often not detailed/specific enough

Biggest problem is very limited access

How we can tell if a sample is representative of a population?

**Samples** have to represent the population. A sample is representative if the distribution of all variables in the sample are the same as, or very close to, their corresponding distribution in the population.

* The distribution of variables is the frequency of their values, e.g., fraction female, percent with income within a certain range. (More on this in Class 3. )
* Product prices in shops surveyed vs price in all shops.

01

List two sampling rules that likely lead to a representative sample and two sampling rules that don’t.

1. Benchmarking: comparing the distribution of those variables in the sample to the those in the population.
2. Asses the selection proccess (evaluating the sampling process)

The sample you collect is different to the population -> **Sample selecting bias**

Representative sample:

1. Random sampling: any random selection method that is unrelated to any varible in the dataset.
2. Benchmarking the few variables for which we know the distribution in the population.

* For instance, there may be some national statistics. Or very similar businesses collected data.
* Reality check always really useful

Non-representative:

1. picking the first 10% of the data set
2. picking consentrating part of the data set (for example age from all age)

List three common features of Big Data. Why does each feature make data analysis difficult?

1. Very large. Billions of observations. (Bigger than what fits into your computer.)

Warning: just because sample is large, it is not necessarily representative!!!!

1. Automatic collection. Not for your analytic purpose - unlike a survey. Data collected by apps, sensors.
2. Complex - text (video, music/noise), network, multidimensional, maps

An important principle for research is maintaining confidentality. How can we achieve that when we collect survey data?

Not giving information to third party.

Always communicate with the source owner(s) and or with legal professional if you are planning to use seemingly sensitive data!

You want to collect data on the learning habits of students in your data analysis class. List two survey methods that you may use and highlight their advantages and disadvantages.

Surveys collect data by asking people (respondents) and recording their answers.Answers to a questionnaire are short and easily transformed into variables.

* Major advantage: you can ask exactly what you want to know
* There are two major kinds of surveys: self-administered surveys and interviews.
* Choice of data collection approach matters a great deal.
* Self-administered survey
  + I cheap and efficient, can use visual aids.
  + I What could go wrong? -> not everyone responds, the wording of a question may result in different results from the respondents. May not be representative if only younger ones or internet users answer, etc
* Interview
  + more expensive and time consuming (training cdata collectors/interviewers)

You want to collect data on the friends network of students in a class. You consider two options:

1. collect their networks of Facebook usersr using data there (80% of them are on Facebook)
2. conduct an online survey where they are asked to mark their friends from a list of all students .

List arguments for each option, paying attention to representation, costs and ethical issues.

You consider surveying a sample of employees at a large firm. List of four selection methods and asses whether each would result in a a representative sample.