Lecture 2 Origins of Data

Tamer Çetin

Learning outcomes

- Understand the basic aspects of data
- Understand the most important data-collection methods
- Assess various aspects of data quality based on how the data was collected
- Understand some of the trade-offs in the design and implementation of data collection

What is data?

- Data is factual information used as a basis for reasoning, discussion, or calculation.
  - Measurements (factual) or statistics (information)
- Data is most straightforward to analyze if it forms a single data table (matrix format).
- A data table consists of *observations* and *variables*.
  - Observations are also known as cases, or rows
  - Variables are sometimes called features or covariates.
- In a data table the rows are the observations, columns are variables.
- A dataset is a collection of data tables, typically related / used in a project
  - 10 data tables, same topic for 10 different years

Data quality is key

- Data quality is key
- If our data is useless to answer our question the results of our analysis are bound to be useless...
- ... no matter how fancy method we apply to it.

Data quality and your question

Data quality is generally a subjective notion!

- First you have to specify what is your (business/research) question!
- What do you want to explore or understand?
- If you have a clear answer, then you can decide on your data quality!

However, there are some objective measures to decide if you have your question!

Data quality

1. Content - what is the substance a variable captures?

   Always check details

2. Validity - is the content of variable close to intended content?

   "Durability" vs "Quality"

1. Reliability - If we were to measure the same variable multiple times for the same observation it should give the same result.
2. Comparability in measurement across observations
3. Coverage - Ideally complete coverage. In practice, they may not include all planned units (incomplete coverage).
4. Unbiased selection - In incomplete coverage, observations included should be similar to all observations that were intended to be covered.

Sidenote

- This is not the type of class where you will have to memorize a list.
- But you should be able to judge the quality of variables in work.
- And you should always remember: **GIGO**: garbage in, garbage out.

Data analysts should know their data

- How data was born
- All details of measurement that may be relevant for their analysis
- To this end, consider having
  - README.txt that describes where dataset comes from
  - VARIABLES.xls that provides basic information on your variables

Data collection

- Automated data collection
- Survey
- Administrative / Census
- Big Data

Data collection: Digital

Automated data collection

- Application Programming Interface, or API – directly load data into a statistical software.
  - API is a software intermediary, or an interface,
  - It allows programs, or scripts, to talk to each other.
- API is widely used in many context.
  - Macro data: FRED - St Louis Fed at research.stlouisfed.org/docs/api/fred/, also World Bank, etc.
  - Micro data such as weather at: openweathermap.org/api
- Data collection limited to dataset.
- Typically additional info available.

Data collection: Digital

Automated data collection

- Web scraping - collecting data from online platform
- html code includes data, can be found, analyzed and collected
- Need extensive cleaning
- Once a procedure is ready (code, script), can be repeated
- Data collection limited to what is on a site

Data collection: Administrative

- Business transactions
- Government records, taxes, social security
- Often: census - records on the population
- 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

Finding a good deal among hotels: data collection

- The dataset on hotels in Vienna was collected from a price comparison website, by web scraping.
- On a specific date
- The purpose of the website is not facilitating data analysis...
- No other potential source
- Good quality, but noise, needed work to make it ready for analysis.
- Coverage is good but not full. Hotels advertising on these websites are not a random sub-sample. Which are the hotels that are left out?

Data collection: Survey

- 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.
- Web, telephone, in person, mix - computer aided interview.
- Choice of data collection approach matters a great deal.
- Self-administered survey
  - cheap and efficient, can use visual aids.
  - What could go wrong?

Sampling

- In many cases, we can collect data on all the people we care about (= the *population*). Often this is not possible...
- For cost/time reasons, we need to take a sample - this is the process of *sampling*.
- 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 Chapter 03.*)

Sample: Representativeness

- The difficulty is: whether a sample is representative is impossible to tell directly.
- There are two ways of assessing whether a sample is representative:
- Evaluating the data collection *process* - subjective with objective elements
- *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

Sampling: Random samples

- *Random sampling* is the process that most likely leads to representative samples.
- All observations in the population have the same chance of being selected into the sample.
- In practice: randomization rule (e.g. flip a fair coin)
- Any other methods that are not randomly picking observations may yield an unexpected bias thus preventing our sample from being representative.
- Practically just like random sampling include fixed rules that are unrelated to the distribution of variables in the data.
- Examples?

Sampling: Random samples

- In small samples (dozens-few hundred) anything is possible.
- Sample of a several thousand observations may equally well represent populations of fifty thousand or ten millions
- The required sample size depends on details of what you want to measure!
- More on this topic later

Sample selection bias

- The sample you collect is different to the population
- This difference is crucial in the story
- Example: Predicting presidential election
  - 1936: Literary Digest. FD Roosevelt vs Landon. 10m people asked. 2m replied. Biggest poll ever. Landon was predicted win 57%
  - What could have gone wrong?

What is different with Big Data?

- Big Data refers to: (i) massive (very large) datasets that are (ii) often automatically and continuously collected and stored, and (iii) may be of complex nature.

(i) Very large. Billions of observations. (Bigger than what fits into your computer.)
  - Warning: just because sample is large, it is not necessarily representative!!!!

(ii) Automatic collection. Not for your analytic purpose - unlike a survey. Data collected by apps, sensors.

(iii) Complex - text (video, music/noise), network, multidimensional, maps

Data collection: hard, time-consuming, costly.

- Collecting data is tedious task, costly as well.
- Usually it is not as simple as you think...
- Collect your experience with the data collecting assignment!

Variable Types: Qualitative vs Quantitative

- Data can be born (collected, generated) in different form, and our variables may capture the quality or the quantity of a phenomenon
- **Quantitative** variables are born as numbers. Typically take many values.
  - also called numeric variables
  - special case is time (date)
- **Qualitative** variables, also called categorical variables, take on a few values, with each value having a specific interpretation (belonging a category)
  - Another name used is categorical or factor variable
  - binary variable (YES/NO) is special case

Data wrangling (data munging)

Data wrangling is the process of transforming raw data to a set of data tables that can be used for a variety of downstream purposes such as analytics.

1 Understanding and storing

- start from raw data
- understand the structure and content
- create tidy data tables
- understand links between tables

[2] Data cleaning

- understand features, variable types
- filter duplicates
- look for and manage missing observations
- understand limitations

Data wrangling: common steps

1. Write a code - it can be repeated and improved later
2. Understand the structure of the dataset, create data tables, recognize links. Draw a schema.
3. Start by looking into the data table(s) to spot issues
4. Store data in tidy data tables. Make sure one row in the data is one observation and manage duplicates
5. Get each variable in an appropriate format
6. Have a description of variables
7. Make sure values are in meaningful ranges; correct non-admissible values or set them as missing
8. Identify missing values and store them in an appropriate format. Make edits if needed.
9. Document every step of data cleaning

The tidy data approach

A useful concept of organizing and cleaning data is called the *tidy data* approach:

1. Each observation forms a row.
2. Each variable forms a column.
3. Each type of observational unit forms a table.
4. Each observation has a unique identifier (ID)

Advantages:

- standard data tables that turn out to be easy to work with.
- finding errors and issues with data are usually easier with tidy data tables
- transparent, which helps other users to understand
- easy to extend. New observations added as new rows; new variables as new columns.

Data Structures

- Cross-sectional (xsec) data have information on many units observed at the same time.
- Time series (tseries) data have information on a single unit observed many times.
- Multi-dimensional (panel) data have multiple dimensions.
  - Many cross-sectional units observed many times
  - Units observed in different space

Simple tidy data table

Table: A simple tidy table

	Variables/columns		
	hotel_id	price	distance
Observations/rows	21897	81	1.7
	21901	85	1.4
	21902	83	1.7

Source: hotels-vienna data. Vienna, 2017 November week- end.

Data structures

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...
1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

Data structures

Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Data structures

A bit more on multi-dimensional - panel (xt) data

- A common type of panel data has many units, each observed multiple times.
- Such data is sometimes called *longitudinal data*, or cross-section-time-series data, sometimes abbreviated as *xt data*.
- Example: countries observed repeatedly for several years
- In xt data tables observations are identified by two ID variables: one for the cross-sectional units, one for time.
- xt data is *balanced* if all cross-sectional units are observed at the very same time periods. It is called unbalanced if some cross-sectional units are observed more times than others.

Displaying immunization rates across countries

Country	Year	imm	gdppc
India	2015	87	5743
India	2016	88	6145
India	2017	88	6516
Pakistan	2015	75	4459
Pakistan	2016	75	4608
Pakistan	2017	76	4771

Note: Tidy format of country-year panel data, each row is one country in one year. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: world-bank-vaccination data.

Data structures

Balanced Panel Data						Unbalanced Panel Data					
Eco.	Year	ITU	WB	WEF	IMD	Eco.	Year	ITU	WB	WEF	IMD
1	2007	4.3	4.0	1.2	1.8	1	2007	4.3	4.0	1.2	1.8
1	2008	4.4	3.7	1.4	1.3	1	2008	4.4	3.7	1.4	1.3
1	2009	4.3	3.9	2.0	0.7	2	2007	8.4	9.4	7.5	7.4
2	2007	8.4	9.4	7.5	7.4	2	2008	8.7	9.2	N/A	7.6
2	2008	8.7	9.2	7.5	7.6	2	2009	8.4	9.1	7.9	8.2
2	2009	8.4	9.1	7.9	8.2	3	2007	8.0	N/A	7.8	7.7

Summary

How is your data?

- Data quality, such as poor coverage (large share of missing observations), will determine what you can do with the data.
- Data may come from existing sources (such as tax authority, World Bank) or you may need to carry out a survey. Surveys may be more befitting but expensive and time consuming.
- Representative sample is essential for any any analysis. Even with big data.
- To respect data confidentiality is a key ethical rule to follow.