Introduction to Survey Statistics – Day 1
Survey Methodology 101

Federico Vegetti
Central European University

University of Heidelberg

Goals of the course

By the end of this course you should have learned

- What are the main considerations behind the design of a survey
- Some basic concepts of sampling and weighting
- Some basic concepts of measurement and psychometrics
- How to implement these things with R

Organization

- Day 1: Theoretical Considerations + Introduction to R
- Day 2: Sampling and Weighting + Making survey weights
- Day 3: Measurement + Assess measurement quality

Reading material

This class draws mostly from the books:

- *Survey Methodology* (2nd edition, 2009) by Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau
- *Complex Surveys. A Guide to Analysis Using R* (1st edition, 2010) by Lumley

I will also cite other documents (journal articles, reports) that provide additional information, or put concepts in a nicer way

The course should be self-sufficient. Readings are meant just in case you want to study some of the things discussed here more in depth

On research

Why do we do research?

- To explain phenomena (academia)
- To inform decision-making (private sector)

In both cases we make **arguments**, theories about how the world works

To convince people that our aguments are valid, it helps to bring data in our support

On research (2)

Arguments can be:

- Descriptive
  - To answer **what** questions
  - Accounts, Indicators, Associations, Syntheses, Typologies (Gerring 2012)
- Causal
  - To answer **why** questions
  - Ideally addressed with experiments (but not only)

Here we discuss issues that are relevant both when the argument is causal and descriptive

However, making causal arguments requires dealing with a number of additional issues that are not covered here

Research in practice

- Usually our theories are about relationships between concepts
- Concepts are measured, so we test relationships between variables
- The validity of our conclusions depends in great extent on:

  1. Model specification & estimation
     - Can we find the hypothesized relationship in the data? Is it robust?
  2. Data quality
     - Can we trust the data at all?

     2.1 Measurement
     2.2 Representation

The model specification/estimation step

- This is what most statistics courses focus on
- Modeling implies
  1. Describing the process that generated the data
  2. Describing a relationship between indicators
- E.g. **linear regression**
  - Describes Y as a variable generated by a Gaussian process
  - Describes how a set of predictors X are associated with Y
  - Tells how well this description fits the data (R^2)
- It can be extended to include measurement as well (more on this later)

Working with surveys

- As social scientists, we are often interested in human **populations**
  - What is the difference in vote share for AfD between West and East Germany?
  - How many Italians believe that vaccines cause autism?
- A **survey** is a statistical tool designed to measure population characteristics
- Common tool for observational (descriptive) as well as experimental (causal) research
- Still the main data source in sociology and political science
  - (though “big data” are becoming more and more popular)

Complication

- When we work with survey data, odds are that we are working on a **sample**
- A sample is a subgroup of the population that we want to study
- We are rarely interested in the sample itself, but we use it to make a probabilistic inference about the population
- **Inference**: a guess that we make about a (general) state of the world based on the (particular) evidence that we have
- It is "probabilistic", because we make every guess with a certain (quantifiable) degree of confidence

Surveys and inference

- Every time we make an inference, we ask the reader to give us a little bit of *trust*
- When we do research using survey data, we do this twice:

1. We infer respondents' characteristics (often on abstract traits) from their answers to the survey's questions
2. We infer population characteristics from sample characteristics

- Many wars with reviewers are fought on these two fronts
- The higher the **quality** of our data, the easier it will be to buy the reader's (and the reviewer's) trust

Surveys and inference (2)

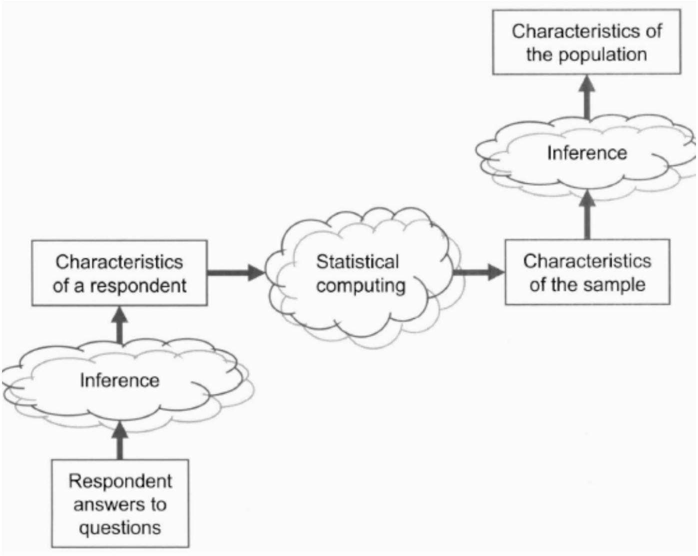


Figure 1: From Groves et al. (2009)

Data quality

- Definition: data has quality when they satisfy the requirements of their intended use
- Several dimensions (and some variation in the literature)
- OECD (2011) identifies 7 aspects:
  - *Accuracy, Relevance, Cost-efficiency, Timeliness, Accessibility, Interpretability, Credibility*
- Another dimension that is important with survey data is *Comparability*
- Maximizing some dimensions may imply minimizing others (given budget constraints)
- Some dimensions are more interesting for our purposes

Accuracy

- Definition: the extent to which the values that we observe for a concept deviate from the *true* values of the concept
- Higher deviation means higher **error**, hence lower accuracy
- When we make the two inferences that we saw above, we leverage on the accuracy of the data
    - The more accurate our data, the more credible our inference

Accuracy (2)

Because the concepts that we are interested in are population characteristics, there are two potential sources of error:

1. Measurement
   - The difference between the values that we observe for a given observation, and the true values for that observation
2. Representation
   - The difference between the values that we observe in the sample and the true values in the population

- The errors arise as we descend from **abstract** (concepts/populations) to **concrete** (responses/samples)

Sources of error

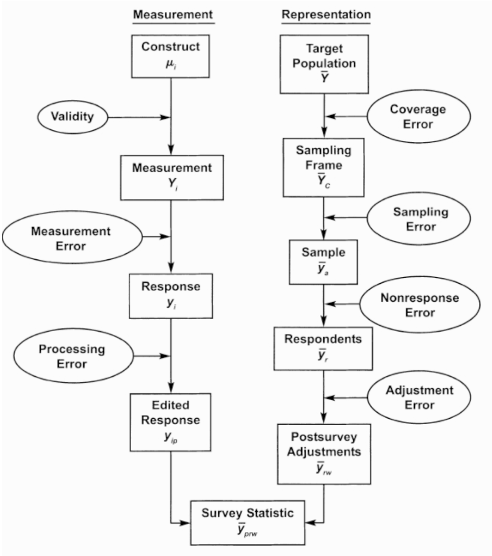


Figure 2: From Groves et al. (2009)

Measurement

- Measurement errors arise on the way from the concepts to the individual responses
- They are as many as the subjects in our study
- They depend to a certain extent on the clarity of the concepts in our head, and a lot on the **mode** of data collection
  - E.g. Telephone interviews are likely to produce different errors than face-to-face interviews

Construct validity

- Definition: the extent to which a measure is related to the underlying construct
  - In this case, *construct* = *concept*
- First of all, it is a theoretical matter
- Often times we end up using **proxies** for our concepts
  - E.g. voting for a right-wing party as a proxy for being ideologycally right-wing
- *Conceptual stretching* is what we do when we use a measure that is far from the concept
  - It may pose a validity problem
- It is our duty to convince the reader that our variable is a valid proxy for our concept

Construct validity (2)

- In statistical terms, the measurement Y is a function of the true value of the construct μ plus some error ϵ.

$$Y_i = \mu_i + \epsilon_i$$

- The validity of the measure is the **correlation** between Y and μ
- Note that validity is a property of the *covariation* between the construct and the measure, not of the congruence between the two
- When the measure draws a lot from other constructs that are unrelated to the one of our interest, ϵ overpowers μ, hence validity is poor

Measurement error

- Definition: the difference between the *true* value of the measurement as applied to a respondent, and the observed value for that respondent
  - For instance, we want to measure mathematical ability, so we give respondents 10 maths problems to solve
  - Jan is usually very good at maths, but that morning he has a terrible hangover, so he manages to solve only 2 problems
  - The value of mathematical ability that would be obtained by Jan on a different day would be much higher than the one we measured

Measurement error (2)

Two types of measurement error

1. Systematic
   - When the distortion in the measurement is directional
   - E.g. our maths problems are too easy to solve, so everyone gets the highest score
   - When this is the case, the measurement is said to be **biased**
2. Random
   - The measured quantity may be instable, so the same person would provide different answers in different times
   - E.g. *How much do you generally agree with your partner about political matters?*
   - The episodes that you recall when you think of an answer are likely to vary over time
   - This type of error inflates the **variability** of the measure

Processing error

- Definition: all the error arising from the way the values have been coded or recoded
- Not such a big problem when using standardized questionnaires
- However, some values may be regarded as implausible when cleaning the data, and erroneously coded as missing

Sources of error (reprise)

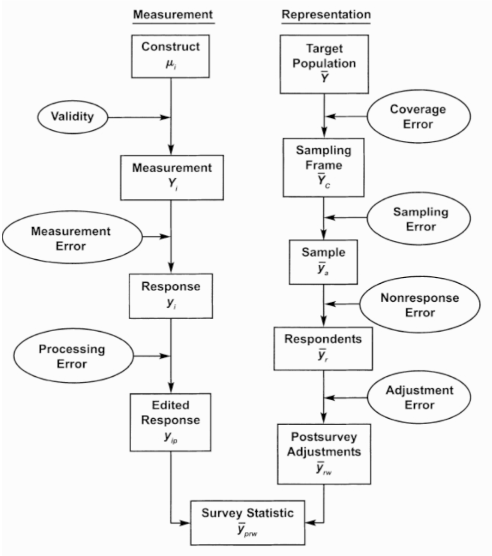


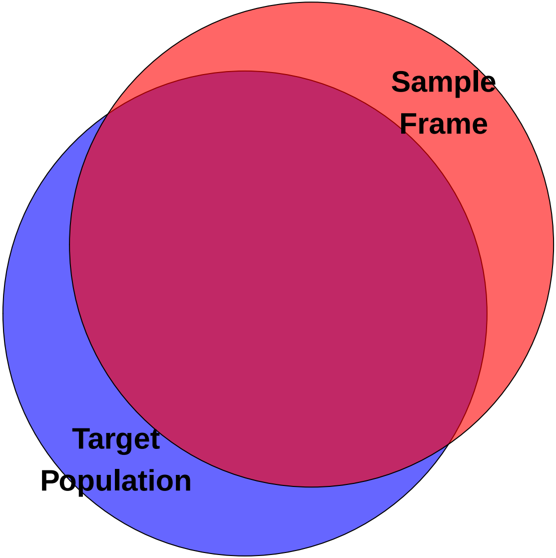
Figure 3: From Groves et al. (2009)

Representation

- Representation errors emerge when we move from an abstract concept of *population* (the Italians) to a concrete pool of data
- They are as many as the statistics that we extract from the data
  - E.g. The mean income in our data will have a different error than the variance of left-right self placement
- They depend on the adherence of our data to the target population, which in turn depends a lot on **survey mode**
  - E.g. If we do an online survey we will be able to reach only the internet users

Coverage error

- Definition: the deviation between the target population and the sample frame
- *Target population*: the entire set of individuals for which we make an inference
- *Sample frame*: the actual list of individuals that we use to draw our sample
- Example:
  - Target population: all German citizens
  - Sample frame: registered telephone users in Germany


Sample
Frame
Target
Population

Coverage error (3)

- Coverage error is likely to produce a **bias** (i.e. directional error)
- It is quantifiable (theoretically) and it depends on what statistic we are interested in
- Example: mean age in an online survey
  - Among internet users: 41
  - Among internet non-users: 48
  - Share of internet non-users: 10%

```
0.1 * (41 - 48)
```

```
## [1] -0.7
```

- The sampling frame is 0.7 years younger than the target population

Sampling error

- Same logic as with coverage error, just in this case our sample is but one of many possible realizations
- A given statistic in our sample will most likely deviate from the same statistic in the sampling frame
- However, we can exert some control
- Two sources of error: **sampling bias** and **sampling variance**
- The first is systematic, the second is random

Sampling bias

- Sampling bias arises when all possible samples we could draw consistently fail to select some members of the sampling frame
  - E.g. People in working age who have a phone but are never at home
- It is a function of how the probability to be selected is distributed among frame members
- It can be removed by giving all members an equal chance of selection

Sampling variance

- Sampling variance is the variability of a given statistic across all possible sample realizations
- E.g. the mean age in our sample will be different from the mean age in the sampling frame
- However, if we could draw *many* samples, the mean of the means of the samples will approximate the mean in the sampling frame
- This is due to the **central limit theorem**
- `Here` and `here` are two good visual demonstrations

Sampling variance (2)

- Remember, in most cases we only have one sample, so we are going all-in for it!
- Sampling variance can be reduced in three ways:

1. Drawing a larger sample
2. Using stratification
3. Avoiding cluster sampling

Stratified sampling

- We divide the population into internally-homogeneous, mutually-exclusive and collectively-exhaustive groups
- We sample randomly within the groups
- The **weighted mean** of this sample is then closer to the mean of the sample frame than the mean of a random sample
- Different from "*quota sampling*", where the number of observations in each stratum is based on specific proportions

Cluster sampling

- We divide the population into groups that are as similar as possible to one another
- We sample groups, and we can:
  - Observe all individuals within the groups (single-stage)
  - Sample again within groups (multistage)
- It allows to save costs of data collection, especially in case of surveys conducted face-to-face
- However, since observations within the same cluster tend to be correlated to one another, cluster samples produce less precise estimates

Nonresponse error

- Nonresponse error arises when we do not collect data for some sample elements, because we fail to reach them or because they refuse to take the survey
- **Nonresponse bias** arises when the group of respondents is systematically different from the group of nonrespondents
  - Example: personal income question, where richer people are less likely to respond than others
- High nonresponse rate is not a problem in itself (although it reduces our sample size) as long as it does not come with bias

Other quality criteria: Relevance

- Definition: the extent to which a given data source is useful for our purposes
- It depends on our research question
- Often we end up doing conceptual stretches because the variables that we use do not measure the exact concept that we are studying
- This may posit a validity problem

Comparability

- Definition: the extent to which observed differences among different countries, cultures, etc., can be attributable to differences in population true values and not to different functioning of the measurement
- This is a particularly relevant problem with cross-country survey data
  - ESS, WVS, EES, CSES
- There are methods in psychometrics to estimate measurement equivalence

Relevance vs. Accuracy

- Relevant data contain all the variables that we need
- Some times we need *a lot* of variables
  - E.g. very long multi-item indexes, very complex explanations
- Survey respondents are willing to spend a limited amount of time before they give up
- Very long surveys have larger drop out rates

Relevance vs. Accuracy (2)

- We may provide incentives for respondents to stay until the end
  - E.g. we pay only when the questionnaire is complete
- However, after a certain amount of time, respondents may lose concentration
- The longer a survey, the larger drop of accuracy in variables collected later

Comparability vs. Accuracy

- Example: We have a survey that is held every year in Germany since 1960
- At a certain point, somebody comes out with a question that captures welfare state attitudes much better than the one used in previous waves of the survey
- Should we change the question wording in the next wave of the survey?

Final remarks

- Survey design is a struggle to reduce the error in two domains:
    1. Measurement
    2. Representation
- As data users, how is this useful for us?
    - Surveys usually come with **weights**: it helps to know what is their purpose, and how they work
    - There are many diagnostics to assess the **quality of measurement** in survey data: it is useful to master some of them
- In the next two days we will focus on these two aspects

References

Gerring, John. 2012. "Mere Description." *British Journal of Political Science* 42 (4): 721–46.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2 edition. Hoboken, N.J: Wiley.

OECD. 2011. "Quality Dimensions, Core Values for OECD Statistics and Procedures for Planning and Evaluating Statistical Activities." http://www.oecd.org/std/21687665.pdf.