Data cleaning, checking, plausibility, missing values and weighting EUROSTUDENT V

Jakob Hartl (hartl@ihs.ac.at)
Martin Unger (unger@ihs.ac.at)
EUROSTUDENT / IHS-Institut für Höhere Studien

eurostudent.eu

Missing Values

- What is it about?
  - Should drop-outs be included?
  - Should a minimum of questions be answered?
- Why is it a problem ?
  - Large differences in valid cases per variable leads to differences in totals. E.g. Total is 5.000, but 1.000 miss in gender - total by gender is 4.000

Missing Values

“A case from the sample is only valid if there is logically consistent information on the variables **age, sex and qualification and at least two other variables for the remaining focus groups** (which would be study intensity, special groups, educational attainment of parents, migration, formal status or form of housing, see topic “Metadata”). If this “3 plus 2”-criterion is not met, exclude the case from the whole analysis.”

https://eurostudent.his.de/wiki/images/b/ba/Glossary_130810.pdf

Data cleaning, checking, plausibility

Not in the student target group (or at least suspicious)

Data cleaning, checking, plausibility

- What is it about?
  - Mistakes, implausible data?
  - Mistakes and fakes?
- Why is it a problem?
  - Data cleaning is crucial for the quality of data!
  - Exclusion of cases?

Data cleaning, checking, plausibility

- Mistakes
  - Time, life cycle
  - Zero, twisted numbers
- Fake answers
  - Open questions
  - Numbers
- Sensible questions (e.g. money)

Data cleaning, checking, plausibility

- Rules for data cleaning in E:IV
  - Define valid cases for analysis
  - Check distribution and cut off
  - Cross-check with other questions/answers
  - Replace values

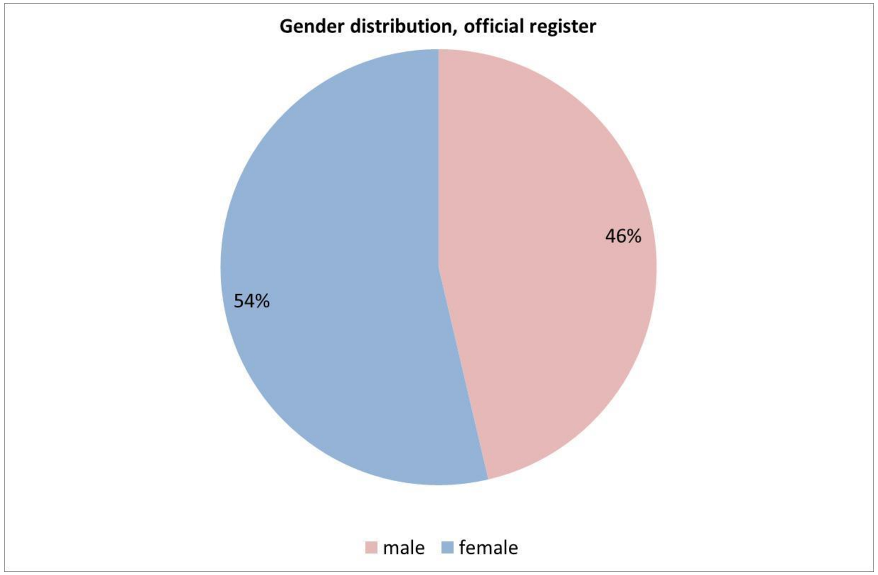
https://eurostudent.his.de/wiki/images/b/ba/Glossary_130810.pdf

Weighting

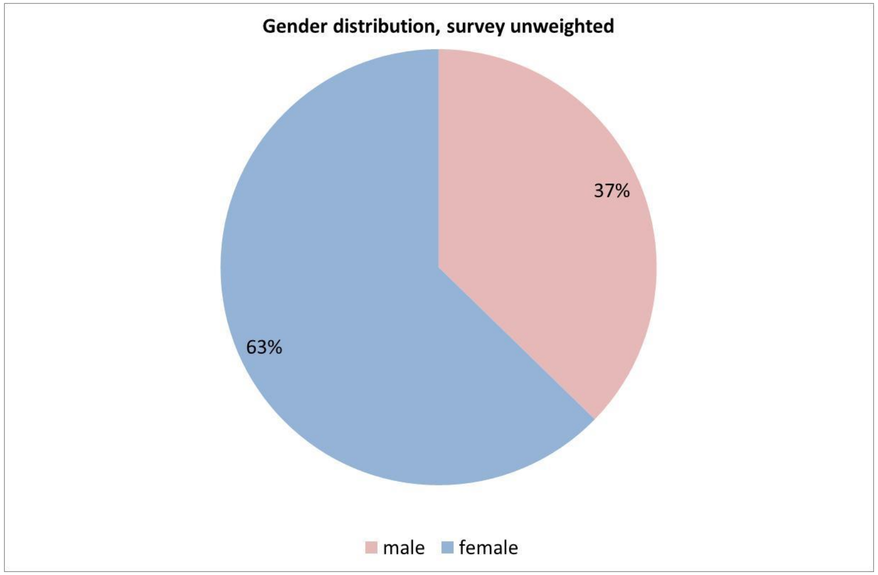
- What is it about?
  - Mapping the total sample in your random sample
  - “correcting” your sample
- Why is it a problem?
  - Different participation patterns/ answer behavior
  - Representation of the students who are hard to approach

Weighting

- Example: Gender distribution of university students in Austria



Gender distribution, official register



Gender distribution, survey unweighted

Need for weighting!

Weighting

“There are no central conventions defined for the weighting of data. We ask you to deal with weighting in the most appropriate way relating to your sample and the demands from Eurostudent. Weighting of data should ensure that the sample is representative for the standard target group to be covered by Eurostudent. If you weighted your data, please enter the frequencies after weighting into the Data Delivery Module. We will ask you to comment briefly on your weighting scheme for our manager’s report at the end of the project.”

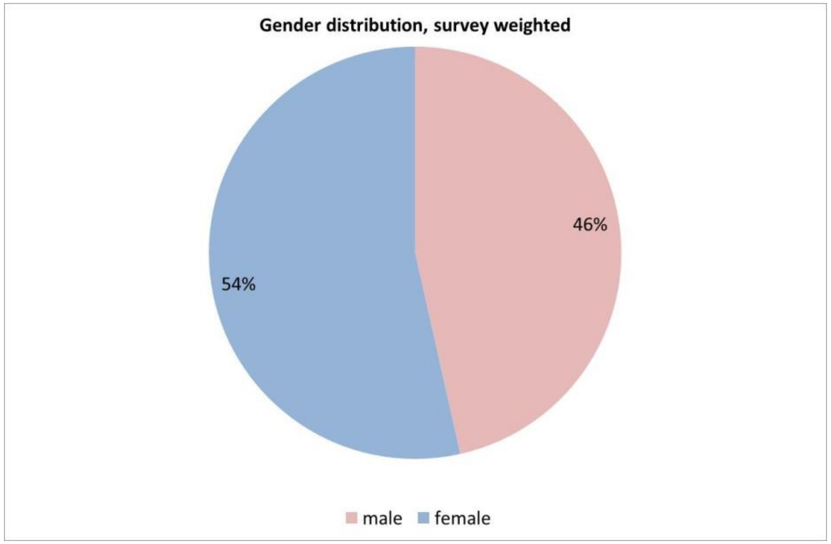
https://eurostudent.his.de/wiki/images/b/ba/Glossary_130810.pdf

Weighting

- Best: official register, you can work with
- Second Best: official register, you can order tables from
- Also possible: marginal distributions
- No Sampling is perfect – weighting **is** crucial

Weighting with cell frequencies

- Weighting scheme: gender, age, study programme, HE institution
- Variables are interdependent, combine characteristics (e.g. gender x age)
- Trim weights (e.g. >5) to control for outliers



Weighting with marginal distributions

- Problem: Variables are interdependent
- Solutions:
  - Iteration manual or with raking
- Recommended introduction (ppt-slides)
  - David R. Johnson, Using weights in the Analysis of Survey Data

help.pop.psu.edu/help-by-statistical-method/weighting/Introduction%20to%20survey%20weights%20pri%20version.ppt/at_download/file