Data analysis with R software

Data-analyysi R-ohjelmistolla

Tommi Härkänen

National Institute for Health and Welfare (THL), Helsinki
E-mail: tommi.harkanen@helsinki.fi

University of Helsinki, February 6, 2013

Contents

Comparison of groups

Observational vs. randomized studies

During past lectures groups (say A and B) were compared using e.g. t-test.

Are these tests adequate in all studies?

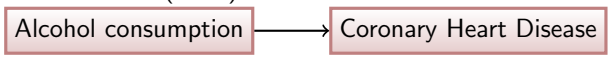
- The groups can be different in many respects.
  E.g. consider people with basic (group A) or high (university degree, B) education
  1. Subjects in group A are on average younger than in B
  2. Older subjects generally have more illnesses than young

  $\Rightarrow$ Subjects in group B have more illnesses, which may result from differences in age, not from education
- **Randomization** removes the differences of the distributions of all background factors between A and B,
  but education (and many other factors) cannot be randomized
- **Confounding effect** of age needs to be accounted for using e.g.
  - experimental design,
  - subset analyses or
  - adjustment using e.g. regression analyses

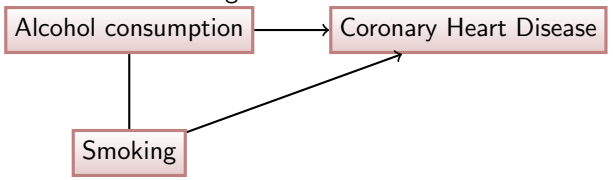
Causes

Causality relations are often depicted using graphs. **Nodes** are connected with **arrows**, which represent (possible) causality.

Example: What is the association of alcohol consumption and Coronary Heart Disease (CHD)?



Problem: People who consume larger quantities of alcohol tend to be smokers and smoking has direct effect on CHD.



(Another problem related to confounding: Maternal smoking, low birth weight and increased infant mortality Hernandez-Diaz *et al.* 2006, Wilcox 2006).

Confounders

Confounders – necessary conditions[1]. The factor must:

- C1 be a **cause of the disease, or a surrogate measure of a cause**, in unexposed people; factors satisfying this condition are **called risk factors** and
- C2 **not be an intermediate step** in the causal pathway between the exposure and the disease
- C3 **not be affected** by the exposure

Confounders usually need to be adjusted for in statistical analyses.

[1] http://oem.bmj.com/content/60/3/227.full

How to select potential confounders?

List **known risk factors** of the outcome.

- Usually based on earlier research (literature).
- Other (expert) information.

Omit the risk factors, whose values can change if the risk factor under study is modified.

- Adjusting for **intermediators** can produce biased results.

Test **the associations** of the remaining group risk factors and the risk factor under study. Omit the nonsignificant risk factors.

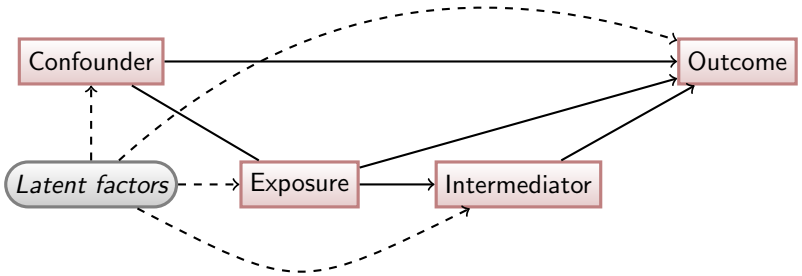
- Test associations of two variables using t-test, Mann-Whitney, χ^2-test, ...

Include the remaining risk factors into the **regression model** as **confounders** (covariates).

- The *adjusted* result is often considered more reliable than results which were not adjusted for confounsing.

Relationships of variables: Summary

Before building a (regression) model, the relations of different variables must be assessed with care.
Temporality can be of help: cause always precedes effect.



Effects of latent factors are difficult to assess. Randomization is often the only way to remove the confounding.