Statistical methods in public health
Confounding and standardization

Tommi Härkänen

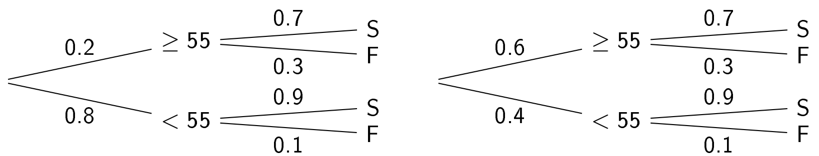
National Institute for Health and Welfare (THL)
Department of Health (TERO)

September 22, 2015

Contents

Confounding by age

Unexposed subjects | Exposed subjects



Unequal age distribution

$\mathbb{P}\{\text{Age} < 55\} = 0.8$ | $\mathbb{P}\{\text{Age} < 55\} = 0.4$

Probability of failure

$0.8 \times 0.1 + 0.2 \times 0.3 = 0.14$ | $0.4 \times 0.1 + 0.6 \times 0.3 = 0.22$

Equal age distribution

$\mathbb{P}\{\text{Age} < 55\} = 0.6$ | $\mathbb{P}\{\text{Age} < 55\} = 0.6$

Probability of failure

$0.6 \times 0.1 + 0.4 \times 0.3 = 0.18$ | $0.6 \times 0.1 + 0.4 \times 0.3 = 0.18$

Comparison of groups

Observational vs. randomized studies

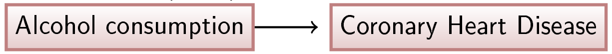
- The groups can be different in many respects.
  E.g. consider people with basic (group A) or high (university degree, B) education
  1. Subjects in group A are on average younger than in B
  2. Older subjects generally have more illnesses than young

  $\Rightarrow$ 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

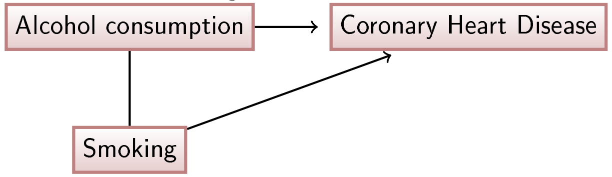
Causes

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 large quantities of alcohol tend to be smokers and smoking has direct effect on CHD.



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.

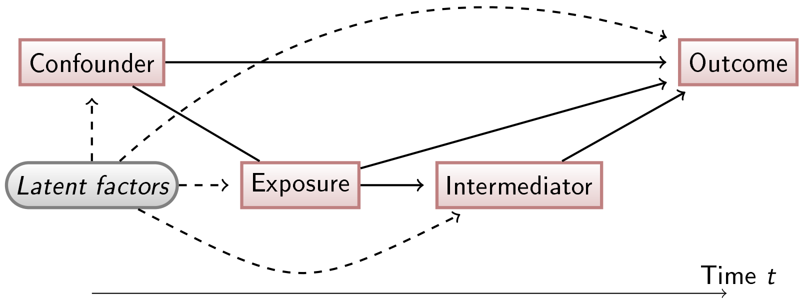
Omit the risk factors, whose values can change if the risk factor under study is modified.
- Adjusting for **intermediators** can produce biased results.

Test **the associations** of the remaining group of risk factors and the risk factor of interest. 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 can be the only efficient way to remove the confounding.

Direct standardization

Study cohort with 337 men exposed (energy intake < 2,750 kcal/d) or unexposed (≥ 2,750 kcal/d), outcome is ischaemic heart disease (IHD).

	Exposed				.... Unexposed			
age	cases	p.years	prop.p.y	rate	cases	p.years	prop.p.y	rate
40-49	2	311.9	0.17	6.41	4	607.9	0.22	6.58
50-59	12	878.1	0.47	13.67	5	1272.1	0.46	3.93
60-69	14	667.5	0.36	20.97	8	888.9	0.32	9.00
Total	28	1857.5		15.07	17	2768.9		6.14

Crude rate estimate for exposed is
$(0.17\times6.41)+(0.47\times13.67)+(0.36\times20.97)=15.06$.

directly adjusted (DA) estimate (assuming equal distribution of person years in each age group)
$(0.33\times6.41)+(0.33\times13.67)+(0.33\times20.97)=13.67$.

DA estimate for unexposed is 6.5.

Cox's proportional hazards model

Poisson likelihood and baseline hazard

Recall lecture 2: Time-dependent hazard rate λ_t and short time bands $h > 0$.
In lecture 3 $\lambda.$ was reparameterized as $\lambda_i = \exp\{X_i\beta\}$.

Now we reparameterize $\lambda.$ so that it depends on both individual factors X_i and time t:

$$\lambda_{it} = \lambda_{0t}\exp\{X_i\beta\}. \qquad (1)$$

$\lambda_{0.}$ is called the *baseline hazard*.

Note that also here $\exp\{\beta\}$ has the interpretation of risk ratio!

The Poisson loglikelihood terms corresponding to (1) are $\delta_{it}\log(\lambda_{it}) - t_{it}\lambda_{it}$ and the Poisson loglikelihood becomes

$$\ell(\lambda_{o\cdot}, \beta; \Delta, T, X) = \sum_{i,t}\delta_{it}\log(\lambda_{it}) - t_{it}\lambda_{it} \qquad (2)$$

Cox's proportional hazards model

Profile loglikelihood

Now recall lecture 1: profile likelihood.
Given β (and the data), the likelihood is maximized by

$$\hat\lambda_{0t} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n t_i\exp\{X_i\beta\}},$$

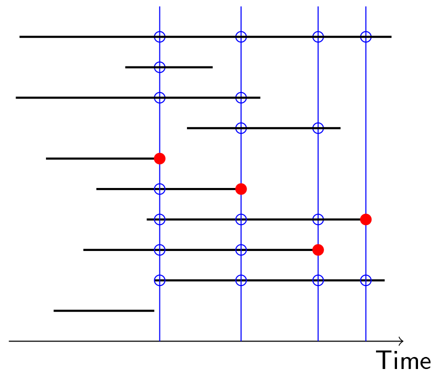
which can be substituted into (2)

$$\ell(\beta; \hat\lambda_{o\cdot}, \Delta, T, X) = \sum_{j,t}\delta_{jt}\log\left(\frac{\exp\{X_j\beta\}}{\sum_i t_{it}\exp\{X_i\beta\}}\right). \qquad (3)$$

If we let $h\downarrow 0$, follow-up times t_{it} become either 0 (not at risk) or h (at risk), and (3) becomes

$$\sum_{j,t}\delta_{jt}\log\left(\frac{\exp\{X_j\beta\}}{\sum_i ht_{it}\exp\{X_i\beta\}}\right) = \sum_{j,t}\delta_{jt}\log\left(\frac{\exp\{X_j\beta\}}{\sum_i t_{it}\exp\{X_i\beta\}}\right) - \sum_{i,t}\delta_{it}\log(h).$$

Risk sets for the observed failures in Cox model



Follow-up data with 10 observations and 4 observed failure times.
At each observed failure time •
denominator of the log likelihood consists of the individuals at risk ○.