Statistical Methods, part 1
Module 3: Observational studies

Dan Hedlin

Stockholm University

December 2012

Stockholm University

The experimental method

- The experimental method is clearly hugely important to research and the growth of human knowledge.
- Randomised, double-blinded, well focused research questions, well established practices, sound statistical methods.
- In lieu of this? For example, the Swedish Prison and Probation Service conducts experiments on treatment methods.

Example

- Some prisoners get treatment A, some are controls.
- Outcome measure is recidivism, either binary (1=former prisoner is convicted to prison again) or time till they are convicted again (if they are)

ART

- ART (Aggression Replacement Training)
  1. Improve interpersonal skills
  2. Anger management
  3. Ethical issues
- Widely used, in the US, England, Wales, etc
- In one study in 2011 the Prison and Probation Service found a negative effect of ART compared to control. ART was discontinued.

Quasi-randomised

- Two issues:
  - Cannot randomise treatment, (e.g. ART)
  - Cannot be made double-blind
- If ART had come out as superior, the cause might have been high motivation among participants in ART group rather than ART per se.

Are experiments superior?

- Problem with non-randomised studies: risk for confounding. You never know for sure.
- One can view research at least in the dimensions

1. Accuracy, bias (*internal validity*)
2. if the results can be generalised (*external validity*)
3. If the results are broad or narrow

Threats to internal validity

Sources of invalidity constitute alternative explanations, rival hypotheses.

1. Extraneous variables, e.g. history (before experiment), seasonal effects, etc
2. Maturation (during experiment)
3. Instrumental decay/change of methods. Not uncommon, even among people who should know better

4. Reactivity. The experiments unduly changes characteristics of units, e.g. the Hawthorne effect and ’John Henry effect’, the latter concerns control group who discover that the treatment group got treatment and they did not
5. Bias in selection of units

6. Pre-tests (=baseline measure). E.g. learning effect, that make units perform ‘better’ when treated. IQ-test first for both treatment and control group, some ‘intelligence enhancing training’ as treatment, then an IQ-test again.

7. Regression towards the mean
   - Kahneman's example "punishment works". People learning to fly aircraft got better after having been told off for bad performance…
   - Affects self-selected units, or units selected because they are at the one end of the distribution, e.g. worse than average.
- Those were main effects. Also interaction effects

- ’Quasi-randomised’ experiments are not randomised. A treatment is administered to one group. E.g. ART.
- In observational studies, you study a ’treatment’ that happened anyway. E.g. smokers vs non-smokers.
- Analysis of survey data is rather like an observational study

Quasi-randomised vs randomised

- In terms of internal validity, differences are found mainly in interaction effects. For example, if ’change of methods’ is applied to both treatment and control group, there should be no difference in a quasi-randomised experiment
- … unless there is an interaction effect: Selection * other source of invalidity

External validity

- Are the results of the experiment possible to generalise?
- Short, brutal answer: no. Problem of induction.
- May be reasonable to generalise anyway.
- Few studies are of interest at all if results cannot be generalised.
- It is a judgement call. This is the main reason why you need several studies, not just one.

Threats to external validity

1. Reactivity (’I am a guinea pig’)
2. Things change. Extraneous variables. E.g. in SU’s course for teachers, one graph on effectiveness of lectures uses data from 1935.
3. ‘No-significance results’ discarded, the one significant result retained and reported
4. Two types of interaction effects with the treatment:

1. Interaction pre-test * treatment

- Interaction pre-test * treatment. E.g. if a pre-test affects emotionally, and interacts with treatment, e.g. sensitises treatment group
- Remedy: have a group with treatment-control subgroups but *without* pre-test, another group *with* pre-test and treatment-control subgroups

2. Interaction selection * treatment

- E.g. a researcher approaches nine schools to make them interested in his/her experiment on teaching methods. They decline. The tenth school is interested.
- This school may be special (More open to improvement? Less conservative?)
- Really, the researcher should write in her/his research paper that nine schools declined. Rarely done outside medical research.

Matching

- It is common in observational studies to match objects. Eg, research on effectiveness of CBT (cognitive behavioural treatment) on depression. 100 patients who have undergone CBT are selected. From a register 100 control patients are matched in pairs on gender, age, alcohol consumption etc to the treatment patients. Now, if the controls

are selected because they have scored low on the Montgomery Åsberg depression test, they may regress towards the mean when they are tested again after the experiment. The treatment group may also regress towards the mean (from above). Hence the treatment effectiveness is underestimated.

- Page 7-18 are to a large extent taken from Campbell and Stanley (1966)

Response schedule

- What further assumptions are needed to conclude a causal effect?
- Say that we estimate the regression

  $Y = a + bx + e$

assuming that the e's are iid random variables.

- Can we say that x causes y?
- Freedman (2009) uses the concept response schedule.

- Subtle but important difference. The regression model is a response schedule if
  - Parameters are constant across subjects (levels/ values of x) and
  - Parameters are constant whether you intervene with a new x or just take the x's that are there in nature. 'Stability under intervention'.
  - The model is true (ie agrees with real world)

- To deduce causal effect, need also assumption of exogeneity: x's chosen independently of e's
- Without exogeneity, it is possible that whenever e is large and positive, x will also be large. 'Selection bias', 'endogeneity bias'
- With these assumptions, 1. response schedule, 2. exogeneity, every time one particular value of x is chosen, a particular value of y will result (bar some independent random variation represented by e).
- In this sense, the x causes that y.