METHODS OF QUANTITATIVE DATA ANALYIS
MSR Course, 2011-2012

Harry B.G. Ganzeboom

Lecture 1.1: Introduction

April 2 2012

Course outline

- Quantitative versus qualitative
- Not about data collection, but only about data analysis
- Weekly routine: two lectures & assignment.
- Course schedule: see website.
- The book
- The computer program: SPSS or STATA.

Assignments

- This is not a writing course!
- However, it is a tabulation course: your tables have to be perfect, elaborate, 100% professional, and separate from a text.
- Text should preferably in bullet style.
- Assignments should be handed in at deadline (usually Thursday 21:59).

Treiman's book (1)

- Elementary, introductory..
- Despite its recent publication, it is surprisingly old-fashioned and traditional in parts.
- Treiman is my close colleague and does research that is close to mine: much about stratification and migration.
- Heavily oriented towards American examples. I will try to balance these a bit by using European (ISSP) data.
- The book assumes that you have mastered a good undergraduate inferential statistics course. I will recap these materials in my second lecture.

Cross-tabulations

- Treiman starts with three chapters on cross-tabulations (or: contingency tables). Important topics here are in particular:
  - How to present percentages: row, column.
  - Lay-out of a table: header, footer, body, panels, columns.
  - How to do controls (holding constant) in table analysis (elaboration).
  - Direct standardization (rather: adjustment).

Why not use cross-tabulations

- In general, I think that full cross-tabs are much less useful than Treiman thinks. They can be very confusing and are in fact hard to analyze.
- One point is really fundamental: always conceive of your analytical problem in terms of independent variable X (cause) and dependent variables Y (effect).
- Direction of taking percentages follows from the causal order: compare Y between values (categories) of X, so the X-categories should sum to 100%.
- In stead of cross-tabulation, I prefer conditional means tables: show a single value of Y within categories of X.

Why not use contingency tables

- Contingency tables invite simplifying the data by creating fewer categories; this is not harmless, you loose statistical power.
- Contingency table invite to represent problems as bivariate. Multivariate presentations are possible, but are complicated to read (unlike a linear model).

Conditional means of Y on X

- Y can take various forms:
  - Binomial: two possible outcomes (yes or no)
  - Multiple ordered (ranked) outcomes
  - Metric outcomes
  - Multinomial: Multiple nominal outcomes.
- Binomials can best be represented by a single percentage. Everything can be dichotomized into a binomial outcome.
- Ranked and metric outcomes can represented by a measure of central tendency (mean) or – if needed – a disperson measure (standard deviation).
- This is MUCH clearer than contingency tables; and it prepares for a regression model.

Elaboration

- Elaboration is the word that older sociologists use for causal analysis using cross-tabulations (or rather: means tables).
- Some of the terminology is arcane and is no longer in general use:
  - Explanation: confounding
  - Interpretation: mediation
  - Specification: moderation, interaction.

Spurious association

- Or spurious effect?
- Refers to the situation that an association between X and Y arises, because of a confounding variable Z, that influences both.
- An association can be shown to be spurious if Z can be assumed to be causally prior to X and Y and the association diminishes or disappears when Z is held constant (=‘controlled’).
- Note: a control variable should be causally prior to both X and Y and influence both. The is no need to control variables that only influence Y, but not X.

Suppressor effects

- Effects of X on Y can be suppressed if Z is causally prior to both X and Y and has reversed effects in X and Y.
- As a result, the effect of X on Y may not be (fully) visible in their association, this may even be zero.
- Logically, confounding and suppression are the same thing.

Intervening variables

- A more often used term is: mediation. I would prefer to name mediating variable M.
- If we control M, the association between X and Y may disappear (or increase), just like with confounders / suppressors.
- Whether a variable is a Z or an M, is not determined by any statistical analysis, it is in the causal order assumption – and in the research design.
- Mediation analysis show to what extend effect are direct or indirect (=‘explained’).

Interaction / moderation

- If the size of the effect of X on Y varies between categories of Z, we speak of interaction or moderation.
- The term ‘interaction’ is more generally used (although far from clear).
- Also: combination non-additive effects.