Determining what to observe

Overview

- Why is case selection important?
- Different types of inference
- Sample surveys
- Some examples of sampling plans
- How many cases?
- Great sampling disasters
- Sampling and non-sampling errors

Why is case selection important?

- Variability of human populations
  - Oak leaves
  - Frenchmen
- If variability is small, case selection is easy
  - “Anyone will do.”
- In human populations variability is great
  - typically we assume that short term variability is small
    - It doesn’t matter if value of X on A is recorded today and value of X on B is recorded next week

Inference

- Case selection or sampling is a vital link in the chain of inductive logic
  - Sampling method determines whether valid inferences from sample to population can be made.
    - Describe sample in order to infer something about the population from which it has been drawn.

- NB Inference problem does not always arise in this form.
  - Sometimes you are only interested in (or the method of case selection only permits) description of the case(s) selected – ie a convenience sample, case study
    - No formal apparatus of inference is needed
    - You describe what you see
    - **BUT**: Evidential basis for description may be imperfect (requiring inference in a different sense)

Inference continued

- What can be inferred from the surviving historical “relics” about Henry VIII’s motives for splitting from the Church of Rome?
  - Primary evidence
    - documents; eye-witness accounts; letters; nearly contemporaneous interpretations
  - What remains will be a sample of what existed
    - Random selection - fire, flood and rodents
    - Non-random selection - suppression and misrepresentation

Inference continued

- Historical craftsmanship
  - Source criticism
    - How close was the writer to the events?
    - What were their motivations in writing?
    - Do both sides of the correspondence survive?
  - Inference aided by high level generalisations about:
    - Human behaviour; what makes sense etc.
- NB inferences **in this sense** are about singular events

Inference continued

- **Sample survey**
  - inference from sample to finite population
  - estimation of % Jedi Knights in UK
    - point estimate;
    - interval estimate
  - sampling process should be constructed to ensure **long run** “representativeness” ie reflect the variability of the population with regard to relevant characteristics
- Usually not possible to know **with certainty** whether any **particular** sample is “representative”

Inference continued

- **Experiment/clinical trials**
  - Statistical inference is from a particular randomisation to the population of all possible randomisations **with those (or equivalent) subjects**
  - External Validity
    - What is the population from which the subjects have been drawn? To which population can we generalize the results?
  - Random sampling from population + randomisation to experimental conditions = external+internal validity

Inference continued

- Inference when apparently the whole population is observed
  - ie annual time series on welfare state spending and economic openness in all OECD states 1950-2001
    - can regard observed outcomes as “sample” from a hypothetical population of all the outcomes that might have been observed - **conditional on the truth of a particular model of the data generation process**
- Sample selection issues about which states and which time periods are included

Summary

- **Historian**
  - tries to describe what has happened in a particular case
  - inferences about the case from partial and incomplete sources
  - disadvantaged compared to contemporary case study/ethnographic observation in that historian's data are finite
- **Sample Survey**
  - random sampling of cases to facilitate inferences from sample values to finite population values
- **Experimentation**
  - randomisation ensures valid inference, but not external validity
- **Model based inference**
  - depends on formulation of an explicit stochastic model, observed outcomes are one "sample" from it

Sample Surveys

- Probabilistic
  - Every element in target population has a non-zero probability of selection
    - **Simple Random Sampling SRS**
      - selection probabilities are equal
      - uses no auxiliary information about population structure
    - **Stratified Random Sample**
      - incorporates auxiliary information about structure of the population into the sampling process
      - selection probabilities can be equal or unequal
    - **Clustered Random Sample**
      - Data form naturally occurring or artificially constructed groups
        - classes in schools; postcode areas

Sample Surveys continued

- Common to all probabilistic methods
  - **Human judgement plays no part in the selection of cases/elements drawn into the sample**
    - If this condition is not satisfied then all attempts to apply standard statistical inference models are **invalid**
    - “What we can’t say we can’t say, and we can’t whistle it either.” ! Frank Ramsey
  - Human judgement may enter into the choice of auxiliary information to incorporate

Selecting the sample

- Probabilistic sampling requires a frame/list/ or some other mechanism to generate a sample from
    - Desirable properties
        - Covers all the target population
        - Each case identifier appears once and only once
        - No relevant sequencing in the list
        - Contains contact information
- No sampling frame/mechanism is perfect

UK frames of choice

- For sampling households or individuals
  - Small users Postal Address File (PAF)
    - List of dwellings receiving less than 30 items of mail per day
    - Supplemented by “point of interview” randomisation tool (Kish Grid) to select
      - households in dwelling
        - individuals in households
- For sampling establishments or workplaces
  - Interdepartmental Business Register (IDBR)
    - In 1992 - 340375 separately identifiable workplaces
  - BT’s Business Database
    - 1.7m locations with a business telephone line

Random Digit Dialling

- Widely used in USA for drawing probability sample for telephone interviewing
  - Typically several stages
    - Numbers are blocked into groups by their first 7 digits
    - Block is chosen at random
    - Number from the block is dialled at random
    - If it is a domestic (business) number, the block is retained
    - Numbers in block are called until a fixed number of interviews are achieved

1998 Workplace Employment Relations Survey

- Target population
  - All British workplaces with 10+ employees
- Sampling frame
  - IDBR
- Stratification (disproportionate) by
  - Number of employees; Industry
    - sampling fractions
      - Workplaces with less than 25 employees 1/545
      - Workplaces with 500+ employees 1/21

Working in Britain 2000

- Target population
  - Employed and Self-employed population of Britain aged 20-60.
- Sampling frame
  - Small user PAF
- Primary sampling units
  - Postcode sectors (South of Caledonian Canal)
  - Stratified by population density and % in SEG 1 or 2
  - Sectors selected with probability proportional to size
- 40 (47 in London) addresses selected in each sector
- Households at address and individual in household selected with Kish grid

How many cases?

- Outside of context of probability sampling to make population inferences the question has no well defined meaning
  - But...holding quality constant, more information is usually better than less
- In the context of studies using probability sampling sensible answers can usually be given in relatively straightforward circumstances (though the details may be a little involved)
- Answer depends on:
  - Magnitude of effect you want to detect ie the size of a difference between two proportions
  - The probability with which you want to detect the effect (if it is real)
  - Amount of variability of the attribute you are interested in
  - The reliability with which you can measure the attribute
  - The size of the subgroups you want to measure the effect in
  - The extent to which you can control for exogenous influences

Great Sampling Disasters I

- Literary Digest Presidential Election Polls
  - Correctly predicted result in 1920, 1924, 1928 and 1932
- 1936 F. D. Roosevelt (incumbent) versus Alf Landon

National Archives

ALF M. LANDON
Republican Nominee for President
"That leadership along the trail
Which we have loved long since,
And lost awhile,
Has come to us again"

- LD poll had Landon winning by 57% to 43%.
- Actual result- Roosevelt 61%
- Sample size 2 million (out of 10 million ballots sent out by LD)
- Sampling frame
  - Lists of telephone subscribers and car owners

Great Sampling Disasters II

- **George Gallup** got the 1936 election result correct
- Also got 1940 and 1944 right
- Pioneered the use of **quota sampling**

- 1948 Presidential race was between Truman and Dewey

- Gallup predicted Dewey would win
- In fact it was too close to call
- Truman retained the Presidency

Great Sampling Disasters III

- April 1992 UK General Election
- Eve of Election polls by the big 5: (Lab%:Con%)
  - Harris (40:38)
  - Mori (39:38)
  - NOP (42:39)
  - Gallup (38:38.5)
  - ICM (38:38)
- Actual Result Lab 35.2: Con 42.8
- Causes
  - Non-response handling; late swing (?); poor quota controls
- See JRSS Series A, 159, 1996

Sampling and Non-sampling errors

- Probability sampling and only probability sampling permits the valid estimation of sampling variability
    - No probability sampling=no standard errors
- Other types of error are also important
    - Bias from unit non-response
    - Bias from item non-response
    - Measurement and classification errors
- In the UK unit non-response is reaching worryingly high levels