Module 1. Methods of statistical inference: design-based, model-based and algorithm-based
Introduction

The context of Official Statistics

• Budget restrictions to carry out traditional surveys
• Increasing concern for response burden
• Increasing non-response
• New sources of available data:
  ➢ Administrative data
  ➢ Big Data sources: traffic sensors, M2M transactions, social media, satellite images...
• Development of mathematical-statistical methods and IT tools that allow for other forms of data treatment
Introduction
The objectives of statistical inference

- The purpose of statistical inference is to obtain information about a population (finite or infinite) from a sample from this population.

- Stochastic assumptions about the individual observations and/or the population are made.

- Statistical information of interest includes totals, means, proportions, ratios, quantiles, etc. or the probability distribution of a random variable.
Introduction
Overview of different modes of inference (paradigms)

- Design-based
- Model-assisted
- Model-based
- Algorithm-based

predictive
Traditionally used by National Statistical Institutes

- Use of surveys to collect data
- NSIs prefer not to rely on model assumptions, particularly if they are not verifiable
- Statistical (mathematical) models may be difficult to understand, communicate or even calculate in a production environment
- The concepts of random sample, sampling error, weighting observations, etc. are familiar to (educated) users of Official Statistics
Design-based inference

Design-based estimation

- Estimators (of a mean, a total, a proportion) are obtained by expanding or weighting the observations in the sample with survey weights
  - Survey weights are derived from the sample design and available auxiliary information

- The statistical properties of estimators are based on the probability distributions from the sampling design
  - Design-based estimators have «good» statistical properties such as asymptotic unbiasedness
Horvitz-Thompson estimator of a total

\[ \hat{Y}_{HT} = \sum_{i \in S} \frac{1}{\pi_i} y_i \]

where \( \pi_i \) is the probability of selection of unit \( i \), and \( 1/\pi_i \) is the weight of unit \( i \) calculated on the basis of the design:

- Stratification (auxiliary variables that define the strata)
- Sample size
- Corrections for non-response, calibration, etc.
Design-based inference

Limitations

- Design-based inference may not be suitable when
  - samples are small
  - in presence of non-sampling errors
  - discontinuities in survey design (e.g. change in data collection mode, new classifications, methodological change of concepts)
    - Design-based estimators do not take into account the changes and cannot separate the «real» change from the methodological change
Model-assisted inference

Introduction

- Design-based estimators of the parameter of a variable can be improved by using auxiliary information and modelling the relationship between the variable and the auxiliary information (=model-assisted)
Model-assisted inference

Model-assisted estimation – theoretical example

- HT estimator obtained from a linear regression model that relates the parameter to auxiliary information
  - Observed \((x_k; y_k)\) for a sample \(S\) (e.g. administrative and survey data), \(x\) are observed for the whole \(U\) universe
  - \(\hat{X}_{HT} = \sum_{i \in S} \frac{1}{\pi_i} x_i\) is the grossed-up total of observed auxiliary \(x\) values
  - \(X = \sum_{i \in U} x_i\) is the known total of auxiliary \(x\) values
  - \(\hat{Y}_{HT} = \sum_{i \in S} \frac{1}{\pi_i} y_i\) is the Horvitz-Thompson estimate
  - \(\hat{Y}_R = \hat{Y}_{HT} + b \cdot (X - \hat{X}_{HT})\) is the regression (=model-based) estimate based on the regression model \(y = a + b \cdot x\) estimated from the sample of observed \((x_k; y_k)\)
Model-based and model-assisted inference

Official statistics: examples of application

- Generalised regression estimator (GREG) widely used by NSIs for calibration
  - Adjusts totals for sub-populations (consistency across tables)
  - Adjusts to known totals
- Small Area Estimation (estimation borrowing strength over space)
- Surveys based on panels (estimation borrowing strength from the past)
- Modelling survey discontinuities
- Integration of sources in National Accounts
- Hedonic Price Indices
- Seasonal adjustment of statistical series
Algorithm-based inference

Introduction

• In the algorithmic approach, the equivalent of fitting a model is \textbf{tuning an algorithm, so that it predicts well}

• It is generally impossible to express algorithmic methods analytically in terms of a mathematical expression

• In the algorithmic approach, the data for which both $x$ and $y$ are known is split into two parts
  • \textbf{TRAINING SET:} part is used to tune the algorithm
  • \textbf{TEST SET:} part used to evaluate – or test – the predictive capabilities of the trained algorithm
Algorithm-based inference

Types of data

- collected from *units* through a **targeted survey** (e.g. Structural Business Survey, Labour Force survey)

- collected from *units* in support of some **administrative process** (e.g. tax records, unemployment benefits)

- other types, registering *events* (e.g. a transaction, an e-mail, a Tweet) generated as by-products of **processes unrelated to statistics or administration**
**Algorithm-based inference**

**Types of data - characteristics**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Survey data</th>
<th>Admin data</th>
<th>Other data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records are units of a target population</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Target variables are directly available</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Auxiliary variables are directly available</td>
<td>Yes</td>
<td>Often</td>
<td>No</td>
</tr>
<tr>
<td>Data preparation/ conversion is needed</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Data covers the complete target population</td>
<td>No</td>
<td>Often</td>
<td>Rarely</td>
</tr>
<tr>
<td>Data are (almost) representative</td>
<td>Usually</td>
<td>Usually</td>
<td>No</td>
</tr>
<tr>
<td>Susceptibility to measurement error</td>
<td>High</td>
<td>Medium</td>
<td>low</td>
</tr>
</tbody>
</table>

**Source:** Buelens *et al.* (2012)
Algorithm-based inference

Theoretical examples

• Similar to the model-based estimator, the algorithmic estimator is

\[ \hat{Y}_{Alg} = \sum_{k \in S} y_k + \sum_{k \in R} F(x_k) \]

• For some function \( F() \) which maps the observed \( x \) to the corresponding \( y \) within \( S \) (training set of units for which \( y \) is known), the set \( R \) contains the population units with unknown \( y \).

• Uncertainty of this estimator arises from the imperfect predictive power of the algorithm, and is assessed on the test set using some cost function.
Algorithm-based inference

Examples in official statistics

- **Central Statistics Office of Ireland**: automatic coding system for Classification of Individual Consumption by Purpose (COICOP) assignment for their Household Budget Survey, using previously coded records as training data.

- **Statistics New Zealand**: Support Vector Machines (SVM) to improve coding of variables Occupation and Post-school Qualification, using two disjoint sets of observations, each of size 10,000, from Census 2013 data for training and testing (50% correctness).

- **Statistics Portugal**: classification trees (a type of decision trees whose response variables are categorical) for error detection in foreign trade transaction data.

- **US Department of Agriculture**: hierarchical clustering to reduce the number of Quarterly Agriculture Survey (QAS) questionnaire versions (states x crops).
Algorithm-based inference

Examples in official statistics (2)

- **Italian National Institute of Statistics**: substituting (fully or partially) ICT in Enterprises surveys by collecting data via website scraping and extracting information using machine learning methods.

- **Statistics Canada**: use of satellite imaging data to assist with estimation of crop yields. Field surveyors were sent to corresponding actual locations to ascertain crop types and yields; these were used as response variables. Probabilistic image processing algorithms were used to learn and predict the field observations based on the satellite data.
References

- CROS Portal on MEMOBUST:
  - Generalised Regression Estimator (Method)
  - Calibration (Method)