Q 2016
European Conference on Quality in Official Statistics

Inference for Statistics Based on Complete Enumerations?

25 – Methodology

3 June 2016

Johannes Klotz

Statistics Austria, Vienna

johannes.klotz@statistik.gv.at

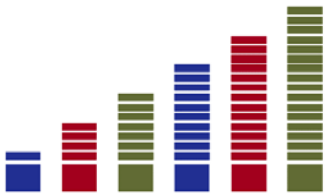
INE Instituto Nacional de Estadística eurostat

Madrid, May 31 - June 3

Q 2016
European Conference on Quality in Official Statistics

Background

- Shift in official statistics from sample surveys to administrative data = ‘complete enumerations’
- Does this mean statistics are ‘true values’ without sampling errors?
- ‘Complete enumerations’ are not necessarily complete for any application
- The issue has a long history in demography (census, vital events)



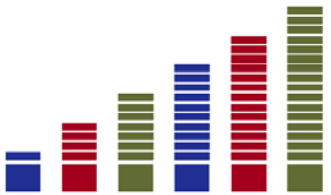
Q 2016
European Conference on Quality in Official Statistics

Quotations in Demography

"There can be stochastic variation in vital events even when the data do not come from a sample survey"
(Westergaard 1880)

"As a basis for scientific generalizations and decisions for action, a census is only a sample"
(Deming and Stephan 1941)

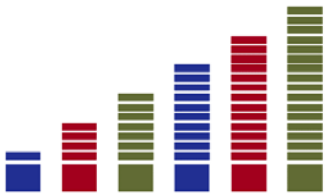
"Do two death rates differ by more than some level of natural fluctuations?"
(Brillinger 1986)



Q 2016
European Conference on Quality in Official Statistics

Randomness = ?

- Probability theory: no substantive definition of a ‘random experiment’
- Social science textbooks: Draw a random sample from a finite population (cf. measurement error, tolerance bands, dice rolls)
- More general perspective:
  Randomness = Lack of information
  (application-dependent)

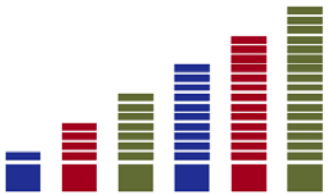


Applications

- Identification of (too) small populations
- Inference for superpopulations extending the enumerated population
- Efficient weighting in regression analyses
- Priority list in plausibility checking (operationalize implausbility by low probability)

Variance estimation

- Pseudo-replicates (areas, periods)
- Bootstrap resamples
- Parametric models



Example

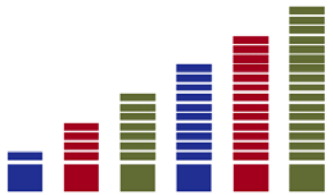
Total fertility rate

$$\mathrm{TFR} = \sum_{x=15}^{49} \frac{B_x}{F_x}$$

Small B_x values in small areas

Assume F_x nonstochastic, B_x Poisson distributed and B_a statistically independent of B_b ($a \neq b$)

$$\widehat{\mathrm{Var}}(\mathrm{TFR}) = \sum_{x=15}^{49} \frac{B_x}{F_x^2}$$



Summary

- ‘Complete enumeration’ does not necessarily mean ‘true values’ (descriptive vs. analytical approach)
- Understanding randomness in a general sense as a lack of information (not solely sampling from a finite frame)
- Many applications for which a stochastic perspective is beneficial
- Randomization out of the statistician’s control
- Variance estimation e.g. by parametric models