Some current challenges for statistical methodology

Tamas Rudas

Hungarian Academy of Sciences

Centre for Social Sciences

Eötvös Loránd University

Faculty of Social Sciences

The current situation

Often seen as big data era / revolution

A relative concept

The presence of big data is the consequence of a new mode of operation of many societies

Different societal mechanisms

Algorithmic decisions

Machine readable platforms

Process produced / organic data?

Changing role of data collection

Fewer data sets collected with the purpose of analysis

More date sets picked up from the owner

Straightforward example: register based census

The fundamental challenge

Because of the changing societal mechanisms, the questions statistics has to answer change

The determination of the new questions is a political process

Which requires expert input

Some obvious examples with rapidly changing meanings:

Working time -- free time

Personal networks

Consumption of news

Political participation

Subcultures

But the societal changes are much deeper

Does the European Union have the infrastructure to quickly find out the opinions of 300 million people about a current political issue?

It does not

Does facebook have the infrastructure to quickly find out the opinions of 300 million people about a current political issue?

It does

Is the EU becoming a facebook group?

There are dangers, too

Does the European Union have an infrastructure to signal epidemies?
It does

Does Google have the capability to signal epidemies?
It does

Which one is faster?
Google

In one year it works, it does not in the other

Changing concept of quality

Data collection:

no sampling error

uncontrolled non-sampling error

Statistical inference:

lack of theory of inference based on big data

Inference from big data

The data do not speak for themselves

The findings do depend on choices of the analysts

Constructed truth

Deeper inferential problems do not disappear when one has complete data

What is true in every month, may not be true for the whole year

Official statistics has to adjust, too

Eurostat principles to implement the European statistics code of practice

How do they work for big data?

What do they mean, if one analyses twitter messages in a particular topic?

Institutional environment

1. Professional independence -- anybody can do it
2. Mandate for data collection -- the user agreed long time ago
3. Adequacy of resources -- 2 hours for a student
4. Commitment to quality -- simple software does it, error free
5. Statistical confidentiality -- people want publicity
6. Impartiality and objectivity -- remains relevant in choosing the questions and the results

Statistical processes

7. Sound methodology -- no protection against arbitrariness
8. Appropriate statistical procedures -- a matter of judgment
9. Non-excessive burden on respondents -- 'respondents' are eager to twit
10. Cost effectiveness -- costs almost nothing

Statistical output

11. Relevance -- of the data or of the method or of the results?
12. Accuracy and reliability -- is what it is, and cannot be repeated
13. Timeliness and punctuality -- a few hours for a good student
14. Coherence and comparability -- these are hard to interpret
15. Accessibility and clarity -- the results could be made accessible on the same platform

Challenges

Modified role

New questions

Modified institutional setting

New theories

New methodologies

Partially happening, but more active management is needed