Statistikproduktionssystem 4.0

Thomas Laitila
SCB och Örebro universitet

Presentation vid SCB/ÖU sommarskola 2021

Agenda

- Framtidens statistikproduktion
- Organisationen, framtidens krav
- Några reflektioner på ämnet för CES
- Selektionsproblemet
- SP 4.0
- Framtidens statistiker
- Till sist

Framtidens statistikproduktion

- Nya datakällor
  - minimera direktinsamling, dvs urvalsundersökningar
  - reducera kostnader
- Big Data
- Sensordata
- Scannerdata
- Paneldata
- …

Framtidens statistikproduktion

- Nya metoder för databearbetning
- Maskin-Inlärning (ML)
- Artificiell Intelligens
- Textanalys
- …

Obamas “Big Data Research and Development Initiative” 2012

(The IT Law wiki)

- ” To launch the initiative, six Federal departments and agencies announced more than $200 million in new commitments that,…”

- “Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and…”

(Nordstat, 2012)

15 år sedan Libanons premiärminister mördades

(Från SVT nyheter, 200821)

- **Gärningsmannen, som inte har identifierats, detonerade en bomb som motsvarade 2000- 3000 kilo TNT när Hariris bilkonvoj passerade. Hariri och 21 andra personer dödades och flera hundra människor sårades.**
- Domen slår fast att Hizbollah-medlemmen Salim Jamil Ayyash är skyldig till mord på den tidigare premiärministern Rafik al-Hariri.
- Domstolens viktigaste bevisning har varit telefontrafiken mellan Hizbollah-medlemmarna som kartlagts och som upphörde vid dådet.

Framtidens statistikproduktion

- Produktionen på 70-talet

	PRODUKTER					
M						
E						
T		Urvalsundersökningar				
O						
D						

Framtidens statistikprodukter

- Produktionen idag

PRODUKTER						
M						
E						
T	Register			Urvalsundersökningar		
O						
D						

Framtidens statistikproduktion

- Produktionen i framtiden

PRODUKTER						
M						
E						
T	Register			Nya Datakällor		Urvalsun.
O						
D						

Organisationen, framtidens krav

- Ny infrastruktur för hantering och bearbetning av data
  - Idag ”slimmad” för urvalsundersökningar och registerdata
- Nya kompetenser för analys och bearbetning av data
  - ML och AI
- UNECE, CES 2017, Geneve
  - Adressering av kompetensproblematiken
  - Session 2(2): ”The next generation of statisticians and data scientists”
  - Bidrag från Sverige: ”**Statistics production system 4.0**”

Några reflektioner på ämnet för CES

- Nu som då existerar inte ett enda exempel där en urvalsbaserad undersökning ersatts av nya datakällor.
- Trots ett drygt årtionde av forskning vet vi ännu inte hur nya datakällor ska hanteras.
- Vi vet inte vilka nya datakällor som ska användas.
- Hur kan vi då börja diskutera framtidens kompetenskrav på statistiker?

Några reflektioner på ämnet för CES

- Exemplet med turismstatistik baserad på mobilpositionsdata (mpd)
- NTTS 2011 – Exempel med mpd, fungerar inte
- NTTS 2017 – Exempel med mpd, fungerar inte nu heller!!!
- Julen 2020, statistik om stockholmarnas resor
  - mpd från telia
  - siffrorna avspeglar Telia - abonnemangens resor in och ut ur Stockholm

Några reflektioner på ämnet för CES

- Forskningen om mpd och statistik över resor, hmm…
- Vi vet på förhand att mpd inte fungerar för sig självt
- Ett tips är att titta på moderna metoder för Skogstaxering.

- Nya datakällor karaktäriseras av selektionsproblematik
  - Data täcker inte alla objekt i populationen
  - Data är inte ett randomiserat urval ur populationen
  - Se exempelvis Jäckle et al. (2017) A review of new technologies and data sources for measuring household finances: Implications for total survey error. University of Essex.

Selektionsproblemet

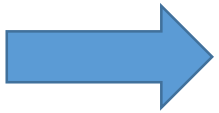
- Genererar bias i skattningar av populationsparametrar.
- Heckman (1979), modellbaserad ansats för korrigering av regressionsskattningar.

- Kan Heckman's ansats användas?
- Tveksamt
  - Bygger på helt annan inferensteori som tolkas annorlunda
  - Selektionsmekanismen anses känd, vilka variabler påverkar?
  - Modellval inte given och en modell är alltid fel.

SP 4.0

- Framtidens kompetensbehov ?

- Framtidens produktionssystem?



- SP 4.0

- Kompetensbehov

SP 4.0

- Grundidé

- Ny datakälla + Registerdata + Urvalsundersökning

- Registerdata definierar populationer, domäner, m.m. och bidrar med hjälpinformation
- Urvalsundersökning bidrar till att ge statistiken traditionella egenskaper

SP 4.0

- Exempel: Platsbanksdata (pbd) och vakansstatistik

- Använd pbd och registerdata för prediktion av vakanser över hela populationen av ftg.
- Använd pbd och registerdata för design av urvalsundersökning.
- Använd prediktioner som hjälpinformation vid skattning.

- Potential till reducering av urvalsundersökning och snabbare statistik

SP 4.0

- Exempel: Utländska turisters konsumtion i Sverige.

- Konsumtion totalt = Antal turistdagar X Genomsnittlig konsumtion/dag

- Antal turistdagar: Kombinera mpd med på-plats mätningar
- Genomsnittlig konsumtion: Kombinera kortdata med på-plats mätningar

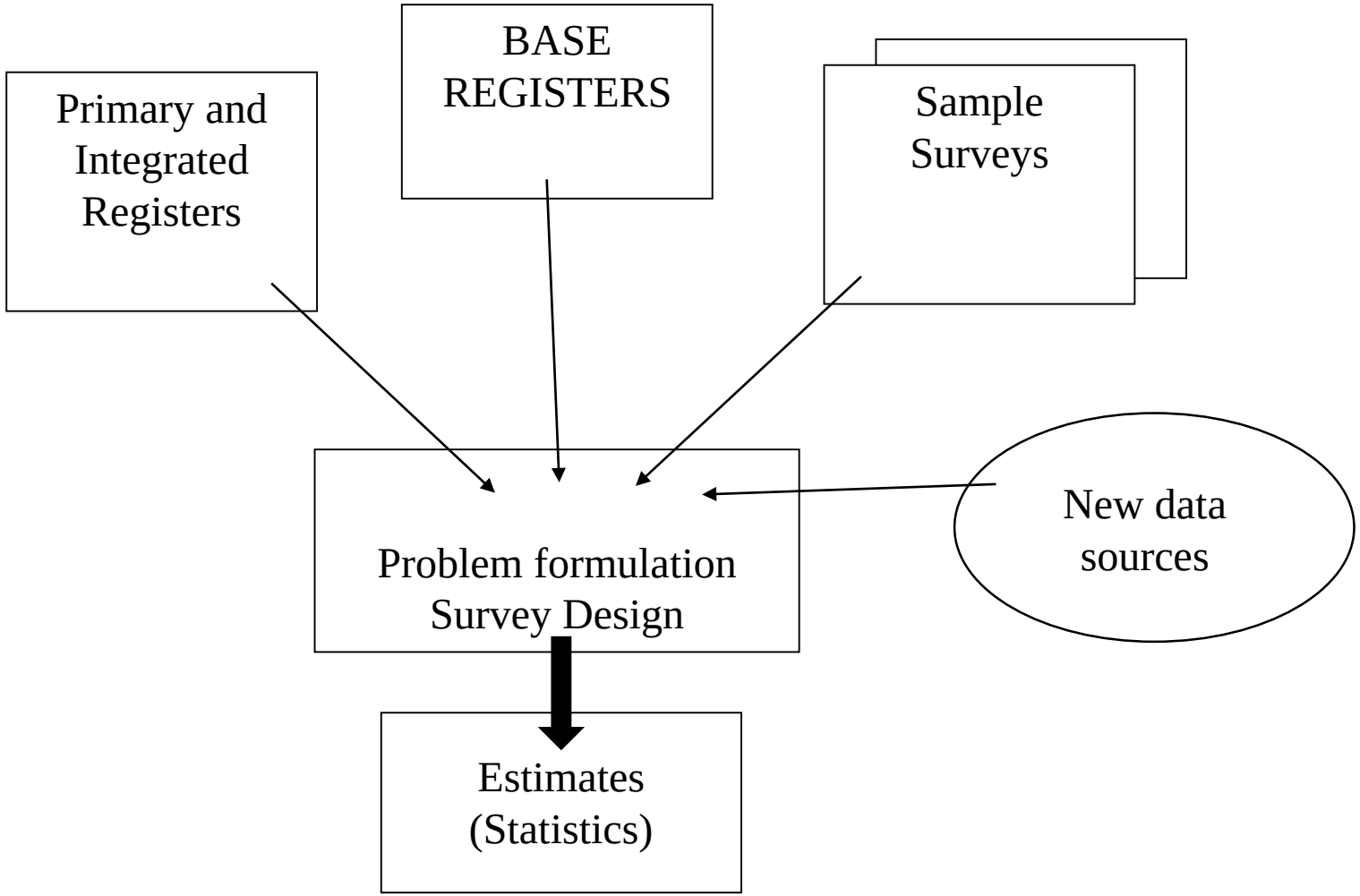


Figure 1: Statistics production system 4.0 in action.

SP 4.0

- Flytta fokus
- från ”Nya datakällor”
- Till ”Design av undersökningar”

Framtidens statistiker

- Behovet är inte i första hand kompetens att analysera nya datakällor, med exempelvis ML, eller tillämpning av AI.

- Viktigare är kompetensen att se möjligheter till kombinationer av nya datakällor med traditionella urvals- och registerbaserade undersökningar. Dvs hitta en design där nya datakällor används och resulterande statistik uppfyller traditionella kvalitetskrav.

Till sist, citat från Neyman (1934), s. 563

- “… The solution of the problem which I described as the problem of confidence intervals has been sought by the greatest minds since the work of Bayes I50 years ago. Any recent book on the theory of probability includes large sections concerning this problem. These sections are crowded with all sorts of " paradoxes," etc. The present solution means, I think, not less than a revolution in the theory of statistics. …”

Tack för att ni lyssnade!