



Användning av nya datakällor i produktion av officiell statistik

Thomas Laitila
Örebro universitet och SCB

Presentation, Surveyföreningens
kvalitetsseminarium 2019



facebook.com/statistiskacentralbyranscb



[@SCB_nyheter](https://twitter.com/SCB_nyheter)



[statistiska_centralbyran_scb](https://www.instagram.com/statistiska_centralbyran_scb)



www.linkedin.com/company/scb





Framtida efterfrågan, utmaningar och möjligheter

- Globalisering: Beslutsfattande och styrning mer komplext
- Efterfrågan av mer, snabbare och specialiserad statistik
- Traditionella urvalsundersökningar svårare att genomföra
 - Svarsbenägenheten minskar
- • Traditionella urvalsundersökningar tar tid
- • Alternativa datakällor
- • Ny teknik för datainsamling

Hur ta oss an framtiden?



Hur använda nya datakällor?

- Traditionella datakällor
 - Urvalsundersökningar
 - Administrativa data

- Nya datakällor
 - Big Data
 - Nya administrativa data
 - Observationsdata, sekundärdata, forskningsdata, etc.

- Kräver användning av modellbaserade ansatser

- Hur klara kvalitetskraven?





Kombinera datakällor

Traditionell idé om användning av datakällor

Urvalsundersökning





Kombinera datakällor

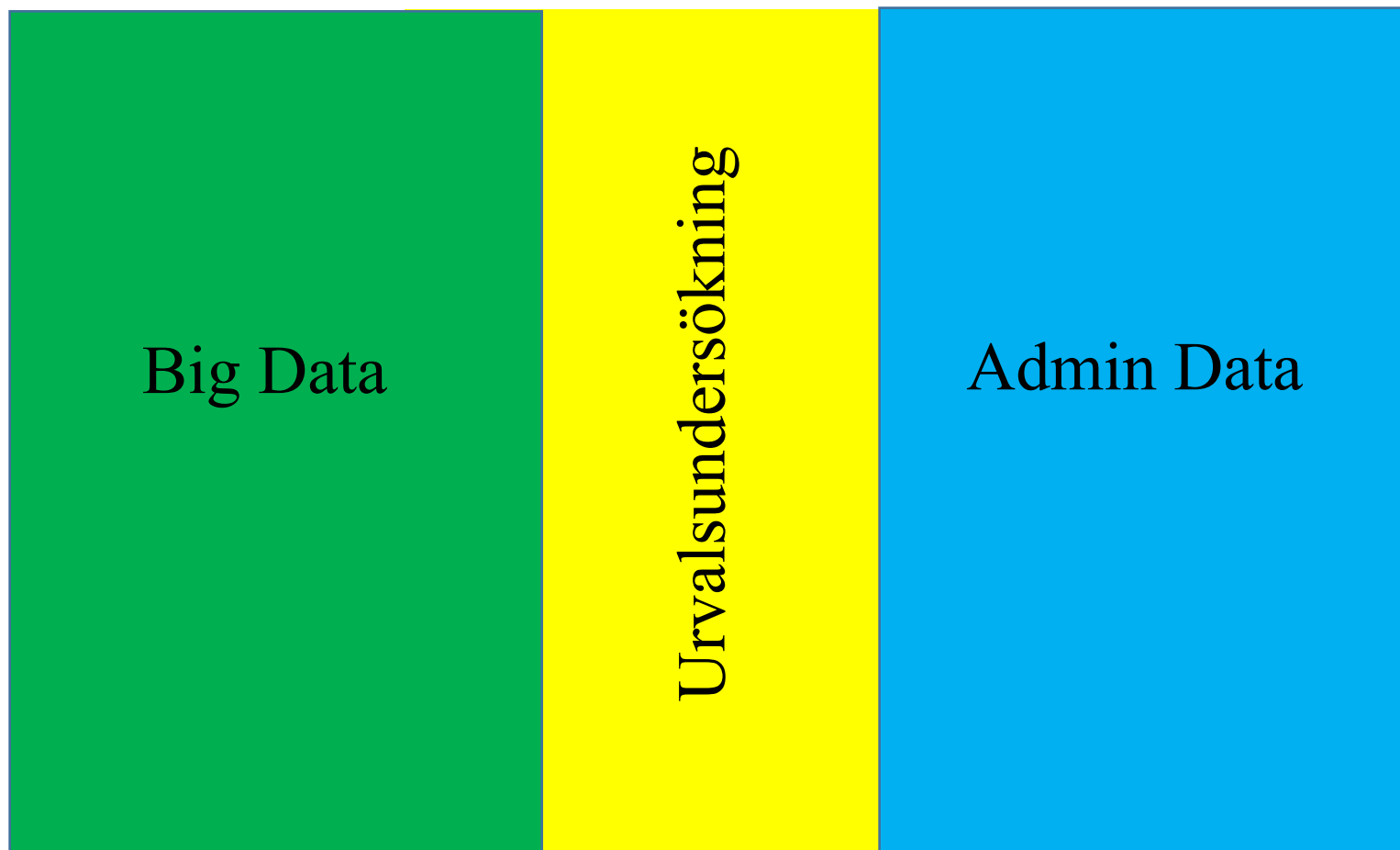
Traditionell idé om användning av datakällor





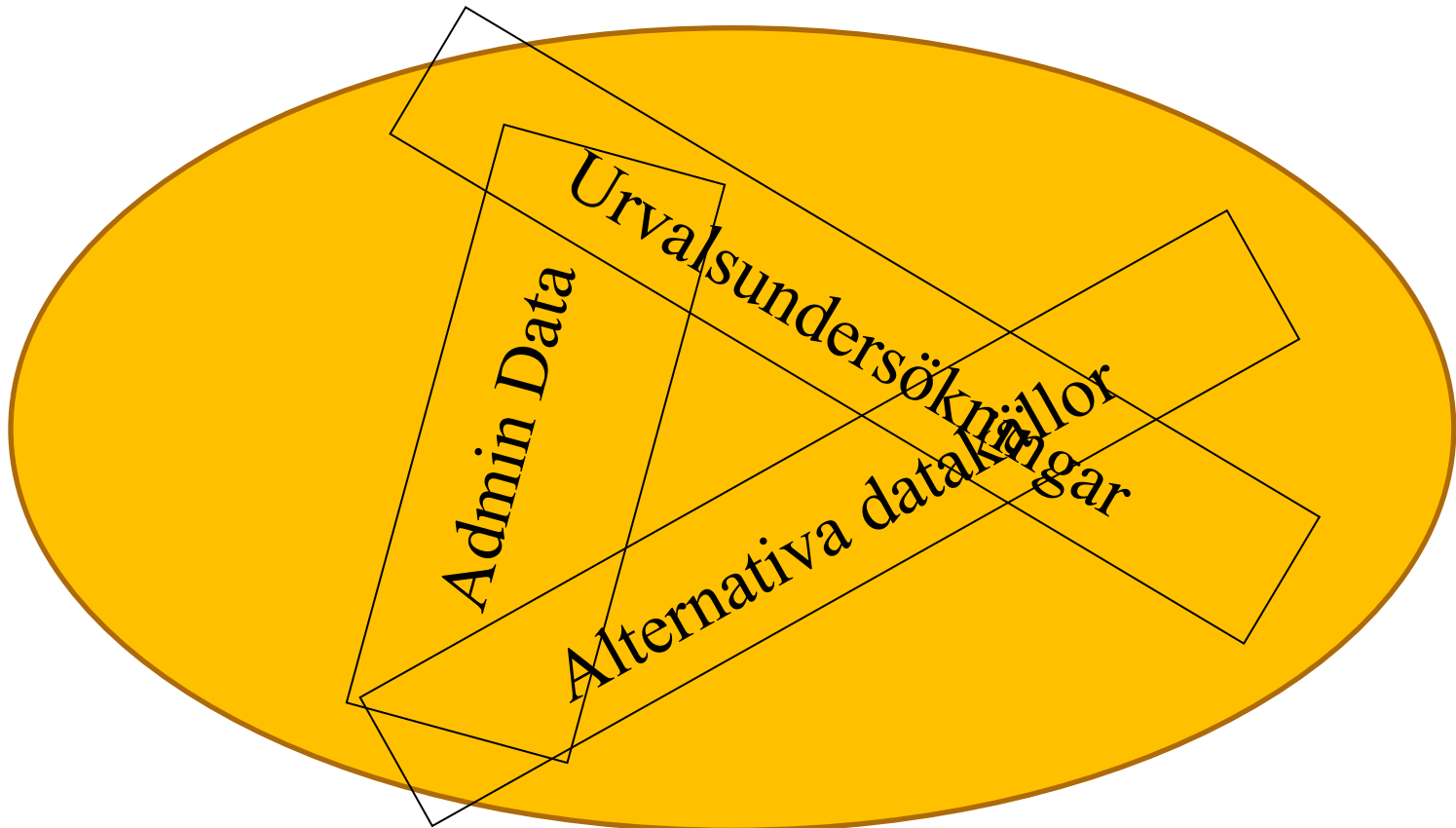
Kombinera datakällor

Traditionell idé om användning av datakällor

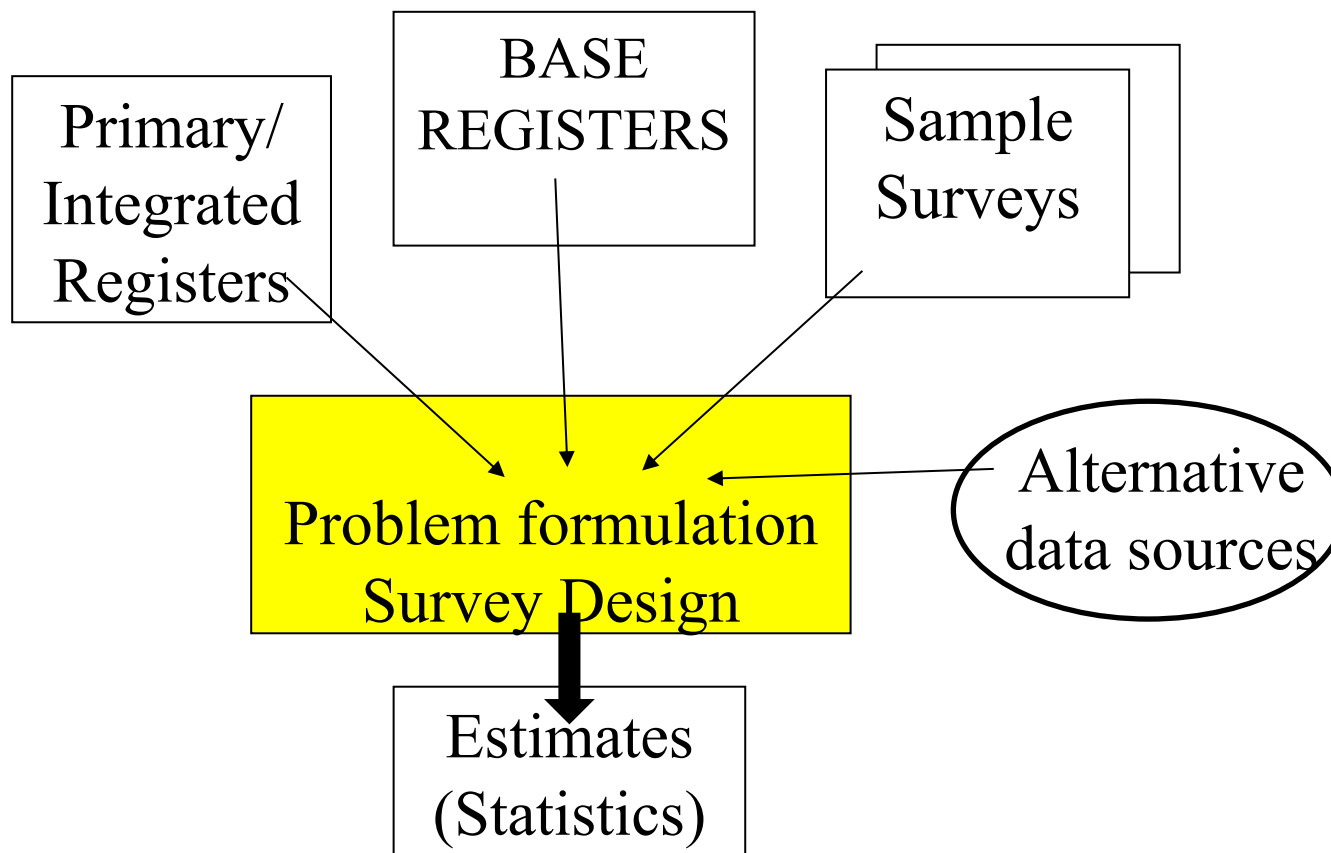


Kombinera datakällor

- Nyttja respektive datakällas styrka! Kräver ny design av undersökningar



Statistics Production System 4.0





Exempel: Turiststatistik

- Urval och intervjuer vid gränsen
- Design följer traditionell urvalsundersökning
 - Definition av population
 - Urval
 - Mätning av valda enheter



(Exempel: Turiststatistik, forts.)

- Vad är problemet?
- Behov av statistik till satellitkonton i NR

- Skatta $Y_j = \sum_U y_{kj}$

med $\hat{Y}_j = \sum_s d_k y_{kj}$



(Exempel: Turiststatistik, forts.)

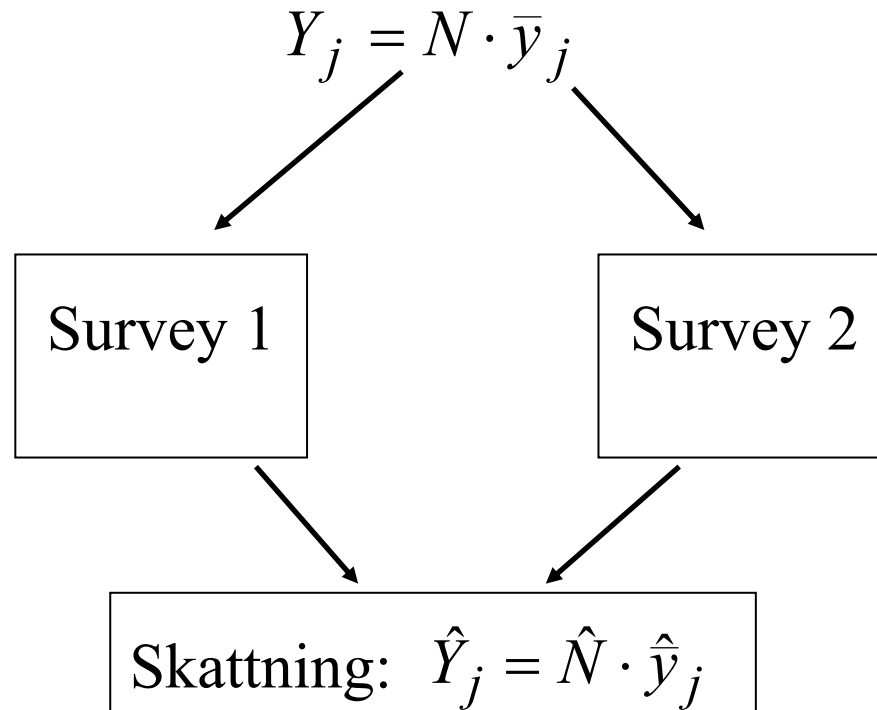
- Definiera om problemet

$$Y_j = N \cdot \bar{y}_j$$



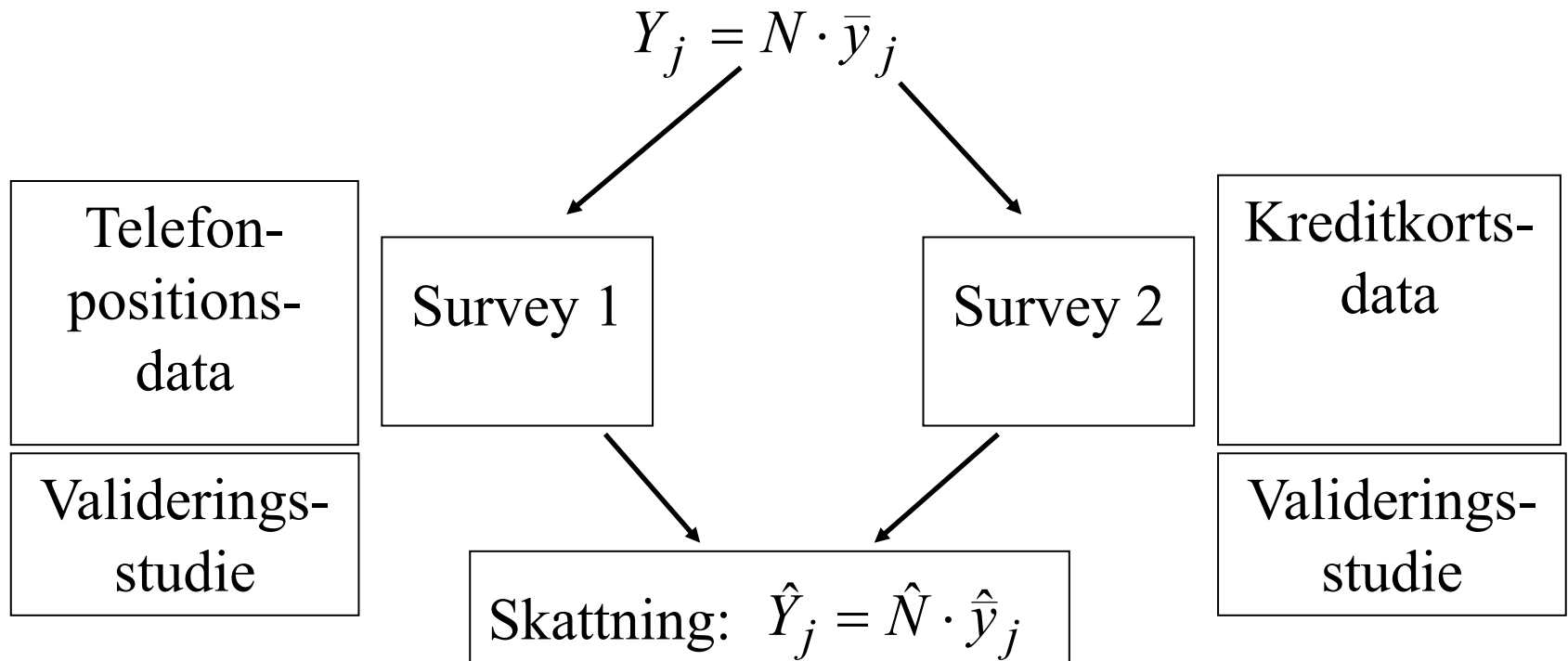
(Exempel: Turiststatistik, forts.)

- Definiera om problemet



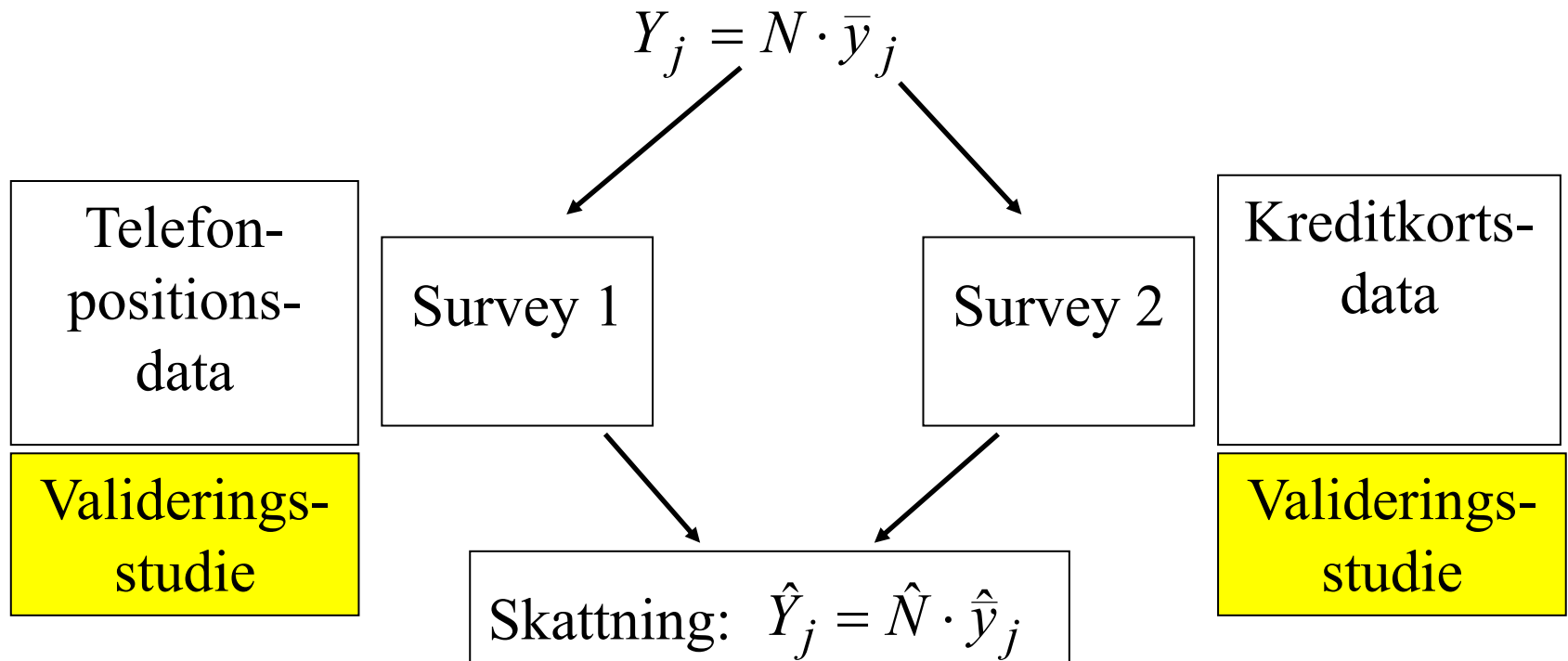
(Exempel: Turiststatistik, forts.)

- Definiera om problemet



(Exempel: Turiststatistik, forts.)

- Definiera om problemet





Vi måste tänka på statistikens kvalitet

- Teoretiska utgångspunkter en bortglömd kvalitetsaspekt
- Olika skolor för statistisk inferens
 - Randomiseringsteori (urvalsundersökningar)
 - Frekventistisk (modellbaserad)
 - Baeysiansk (modellbaserad)
- Skillnad i tolkning av sannolikhetsbegreppet
- Skillnad i tolkning av statistik
- Även om siffrorna är desamma säger de olika saker!



Randomiseringsteorin

- J Neyman (1930 talet) (A Kiær, 1895)
 - Sannolikhetsurval och konfidensintervall
- Fundament för officiell statistikproduktion
- Objektiv inferens
 - Ingen betingning på antaganden



Table: Change (%) in number of employees by industry.

| Industry | 95% CI | | Industry | 95% CI |
|---------------------|------------------|--|------------------|------------------|
| <i>Frmng/Hntng</i> | -2,5--14,5 | | Transport. | -3,5--2,5 |
| <i>Electricity</i> | -0,6--11,8 | | Banking | -2,5--4,5 |
| <i>Construction</i> | 4,5--12,3 | | Real Estate | 1,2--8,2 |
| Retail | 2,2--7,2 | | <i>Education</i> | 0,3--12,5 |
| <i>Hotel</i> | 8,1--22,7 | | Health Care | -1--10,4 |
| | | | | |

Source: Statistics Sweden, AM 63 SM 0703

Table: Change (%) in number of employees by industry.

| Industry | 95% CI | | Industry | 95% CI |
|---------------------|------------------|--|------------------|------------------|
| <i>Frmng/Hntng</i> | -2,5--14,5 | | Transport. | -3,5--2,5 |
| <i>Electricity</i> | -0,6--11,8 | | Banking | -2,5--4,5 |
| <i>Construction</i> | 4,5--12,3 | | Real Estate | 1,2--8,2 |
| Retail | 2,2--7,2 | | <i>Education</i> | 0,3--12,5 |
| <i>Hotel</i> | 8,1--22,7 | | Health Care | -1--10,4 |
| | | | | |

Source: Statistics Sweden, AM 63 SM 0703

Hur tolka resultaten om statistiken är modellbaserad?
(frekventistisk, Baeysiansk)

Sammanfattning

- Nya datakällor o nya tekniker för datainsamling
 - Möjlighet till ny statistik och
 - ny design av gamla undersökningar
- Kvalitetskrav på officiell statistik måste beaktas
 - Metodens "objektivitet"
 - Direkta och indirekta konsekvenser av modellbaserad statistik?
- Kombinera datakällor i design av undersökningar
- Ny roll för traditionella urvalsundersökningar



Nya datakällor som används av SCB?

- Försök till ny design av HUT, design med kassaregisterdata i kombination med demografiska data och surveydata
- webscraping, prisstatistik
- AIS, positioneringsdata (register)
- Miljöstatistiken (olika datakällor)
- m.m.





Tack!

(thomas.laitila@scb.se)

