

Statistiska databaser och register

Masterutbildningen i Surveymetoder
och Officiell Statistik

och på

Avancerad nivå i statistik

7,5 ECTS-p, Vårterminen 2009

Kursansvarig: Daniel Thorburn

Kursdelar

- Inledning, 2 ggr, Daniel Thorburn
- Statistiska aspekter på registerstatistik och konstruktion
5 ggr, Anders och Britt Wallgren, (ca 2/5)
- Infologiska aspekter på register och databaser
6 ggr, Bo Sundgren och Alf Fyhrlund (?) (ca 2/5)
- Övrigt (ca 1/5)
bl a Datamining, Mats Lanner
Beslutsträdstekniker Mats Lanner
I privata ekonomiska sektorn, Per Weidenman,

Litteratur

- -Wallgren, A. och Wallgren, B. *Register-based Statistics – Administrative Data for Statistical Purposes*. Chichester, Wiley. (Finns att skaffa på Adlibris för 624 kr)
- Register-based statistics in the Nordic Countries – Review of best practices with focus on population and social statistics (2007), Geneva, United Nations (finns att ladda ner på sidan www.unece.org/stats/publications)
- Kompletterande material som utdelas i samband med undervisningen.
Obs detta är en viktig del av kurslitteraturen!
- Del 2 och 3 tenteras enbart med kompletterande material

Schema

- Inledning: Idag, 26/1 11-13 (gruppuppgifter)
- Del 1: 30/1 11-13, 3/2 13-15, 13/2 13-15, 17/2 10-12, 24/2 9-11. Muntlig tentamen 2/3 kl 10 -
- Del 2: preliminärt: 4/2 11-13, 9/2 13-15, 10/2 11-13, 20/2 9-11, 24/2 13-15, 3/3 13-15 (Eventuellt skriftlig hemtentamen)
- Del 3 3/3 9-11 (ML), 10/3 9-11 (ML), 17/3 13-15 (PW)
- Skrivning 20/3 kl 9-14

Examination

- Första delen muntlig tentamen på kursboken. Kursboken bör tas med på tentamen.
- Andra delen preliminärt
 - en/flera inlämningsuppgifter (arbete med verkliga databaser)
 - skriftlig tentamen med teorifrågor och en konstruktion av lämplig datamodell
- Sista delen, Tid för skriftlig tenta finns inbokad.

Examination

- Alla betygssättningen kommer att översättas till poäng (del1 max 40p, del 2 max 40p, del 3 max 20p).
- Poängen summeras och ger slutbetyget
 - A. Utmärkt – motsvarar 90 – 100 poäng
 - B Mycket bra – motsvarar 80 – 89 poäng
 - C Bra – motsvarar 70 – 79 poäng
 - D Tillfredsställande – motsvarar 60 – 69 poäng
 - E Tillräckligt – motsvarar 50 – 59 poäng
- Dessutom krävs godkänt på alla delar

Nästa gång

- Förberedelse inför Anders och Britts delar.
- Gruppuppgifter
- Diskussion och frågor till mig nästa gång -
 - Alla grupper redogör för hur långt de kommit

Statistiska Databaser

- Större delen av all svensk statistik tas fram baserad enbart på register
- Större delen av urvalsundersökningar utnyttjar på något sätt register t ex som ram, i skattningarna för precisionsförbättring (efterstratifiering, kalibrering, imputering) eller att uppgifter ur register förs på urvalet.

- Idag är sekundäranvändningen av statistisk grunduppgifter minst lika viktig som det ursprungliga målet
- Det är alltså viktigt att samla statistiken i databaser så att den är lätt att komma åt och använda.
- Detta är en minst lika viktig kunskap för en statistiker idag som andra kunskaper i beskrivande statistik

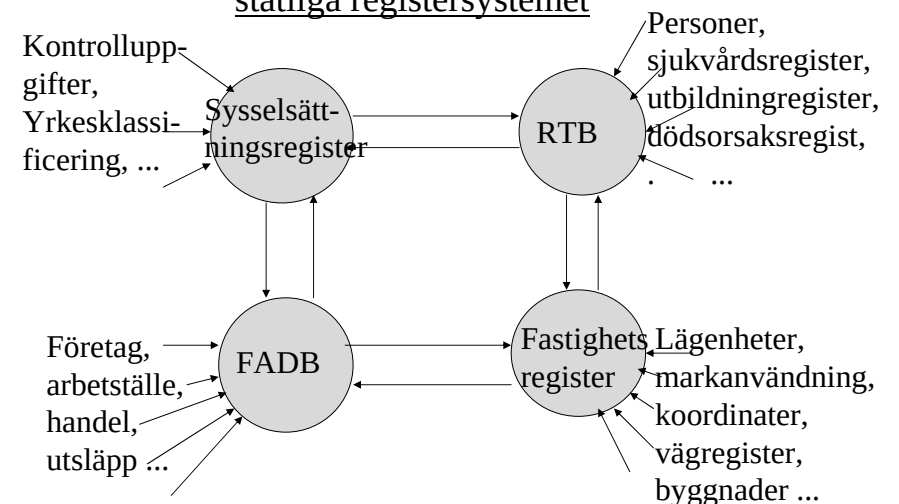
Exempel: Arbetskraftsundersökningen

- November år 0: dras det nya stickprovet för nästa år (12 paneler) ur registret för totalbefolkningen (RTB), Hjälpinformation: LAPS (arbetsgivaruppgifter, år -1) Stratifierat efter län, ålder, kön, födelseland och arbetskraftsstatus.
- Samkörs med diverse register för att hitta bl a arbetsgivaretelefonnummer till hushåll (föräldrar, sambo, ...) och annat som kan underlätta kontakten
- År 1-3: Intervjuer genomförs 8 gånger med 3 månaders mellanrum. En panel börjar varje månad.
- Skattning varje månad med aktuell hjälpinformation: senaste LAPS, näringsgren från företagsregistret, arbetskraftsstatus i AMS, senaste RTB... . Kalibrering med fyra hjälpvektorer med ca 40-80 grupper i varje.
- År 3: Sista intervjun frågas om hushållets sammansättning. (Uppgifter om hushållsmedlemmarnas arbetskraftsstatus hämtas ur LAPS, AMS, Skatterregistret mm. Data sammanställs för EU som kräver AKU på hushållsnivå.
- År 2 och framåt. Forskare får tillgång till data (bl a i Mona) och kopplar uppgifterna med andra register, t ex företagsregistret eller slutenvårdsregistret för arbetsmiljöstatistik

Exempel: FoB

- Beslut fattades för länge sedan att FoB 90 skulle vara den sista enkätbaserade räkningen
- FoB 2000 genomfördes aldrig. Trots åtagande till både FN och EU
- Nu skall den genomföras 2012 (senaste året enligt FN och EU för 2010 års FoB) baserad enbart på register, främst: RTB, Utbildningsregister, Fastighetsregister, Företagsregister, Kontrolluppgifter och Lägenhetsregister. Det sista skapas för närvarande. Varje lägenhet skall få en unik adress och så småningom skall all folkbokföring ske på lägenhet,
- Hur få detta att hänga ihop? Skapa hushåll, studera pendlingsmönster, trångboddhet, fattigdom etc.
- Danmark är föregångsland men även Finland och även Norge ligger före oss.

Schematisk bild över det planerade statliga registersystemet



Identitet

- Ofta försöker man skapa s.k. flata filer för varje register.
 - Dvs data ges i form av en matris.
 - Individ är en rad och variabel är en kolumn.
 - Längden på raderna är konstant

- En rekommendation är att identitetsnumret skall vara informationslöst.
- Det svenska personnumret skapades för länge sedan och är informationsbärande
- 130416-4893 t ex ger information om födelsedag, kön, födelseort/folkbokföringsort när numret infördes 1949 (I detta fall Uddevalla).
- Varje förändring leder till problem eftersom numret måste ändras, t ex vid könsbyte, invandrare som får bättre info om sin födelsedag eller vid skyddad identitet om födelsedag följer med.
- Datum som inte finns används för företagsnummer.

Identitet

- Det finns så få svenskar att antalet personnummer borde räcka utan problem. Nu har det dock tillkommit några problem.
- Varje person som kommer in i landet och arbetar under en begränsad tid får ett nummer och sedan är det upptaget även om han/hon lämnar Sverige efter ett år.
- Många invandrare vet inte sin födelsedag. De väljer då vissa dagar t ex 1 januari. Den dagen räcker inte de 500 numren till.
- Även skyddad identitet leder till att nummer förbrukas, liksom att många invandrare fått byta bort 9000-nummer.
- Därför har man fattat beslut om en ändring i det svenska personnummersystemet.

Statistiska problem

- Samordna begrepp och definitioner
- Samordna tidpunkter. Med levande register kommer olika undersökningar att redovisa olika värden.
- Över- och undertäckning, multipla förekomster
- Nycklar för koppling mellan register.
(Personnummer fungerar för personer men hur koppla till arbetsplats, bostad etc)
- Registerkvalitet blir viktigt

Infologiska problem

- Hur lägga upp systemet så att det fungerar smidigt och utan orimliga svarstider
- Sekretessaspekter.
- Relationsdatabaser?
- SQL
- Ex: MONA (forskardatalager)
- Ex: Mikrosimulering (undersök effekter av politiska beslut på en verklig population)

Data-mining

- Det finns stora databaser i t ex företag. T ex kunddatabaser (t ex Med Mera kort)
- Hur utnyttja detta?
- Här är inte den vanliga precisionen sällan problemet – men väl att man letar och letar och letar tills man hittar något. Hur värdera detta.
- Hur leta effektivt
- Går inte att säga att data får bara användas en gång. Paradigmskifte.
- Syftet kanske inte heller är att bevisa något vetenskapligt utan bara bestämma en lista för reklamutskick som man hoppas ger valuta för pengarna

Hälso-området

- Hälso-området
- Samköra register från olika tider för att hitta riskfaktorer eller farliga miljöer.
- Specialregister: Dödsorsaksregistret, Cancerregistret, Missbildningsregistret, Slutenvårdsregistret
- Allmänna. Mönstringsdata, RTB, FoB, etc
- Kombinera med olika data
 - följa upp behandlingar
 - Urvalsundersökningar, t ex Case control