

Statistisk översiktscurs (SÖK)

VT 2025

Michael Carlson, Statistiska institutionen

Idag

1. Datakällor och datainsamlingsmetoder

- primär- och sekundärdata, registerdata

2. Fel i undersökningar

- feltyper, orsaker och effekter

3. Sveriges Officiella Statistik (SOS)

- statistikansvariga myndigheter (SAM)

4. Kvalitet i statistiska undersökningar

- lagstiftning och industristandard, kvalitetsredovisning

1. DATAKÄLLOR OCH DATAINSAMLING

Primärdata

- Nya data, primärt insamlat för undersökningens syften.
 - Kontroll över definitioner och avgränsningar, insamlingsprocesser.
 - Aktuella och relevanta data.
 - Dyrt!

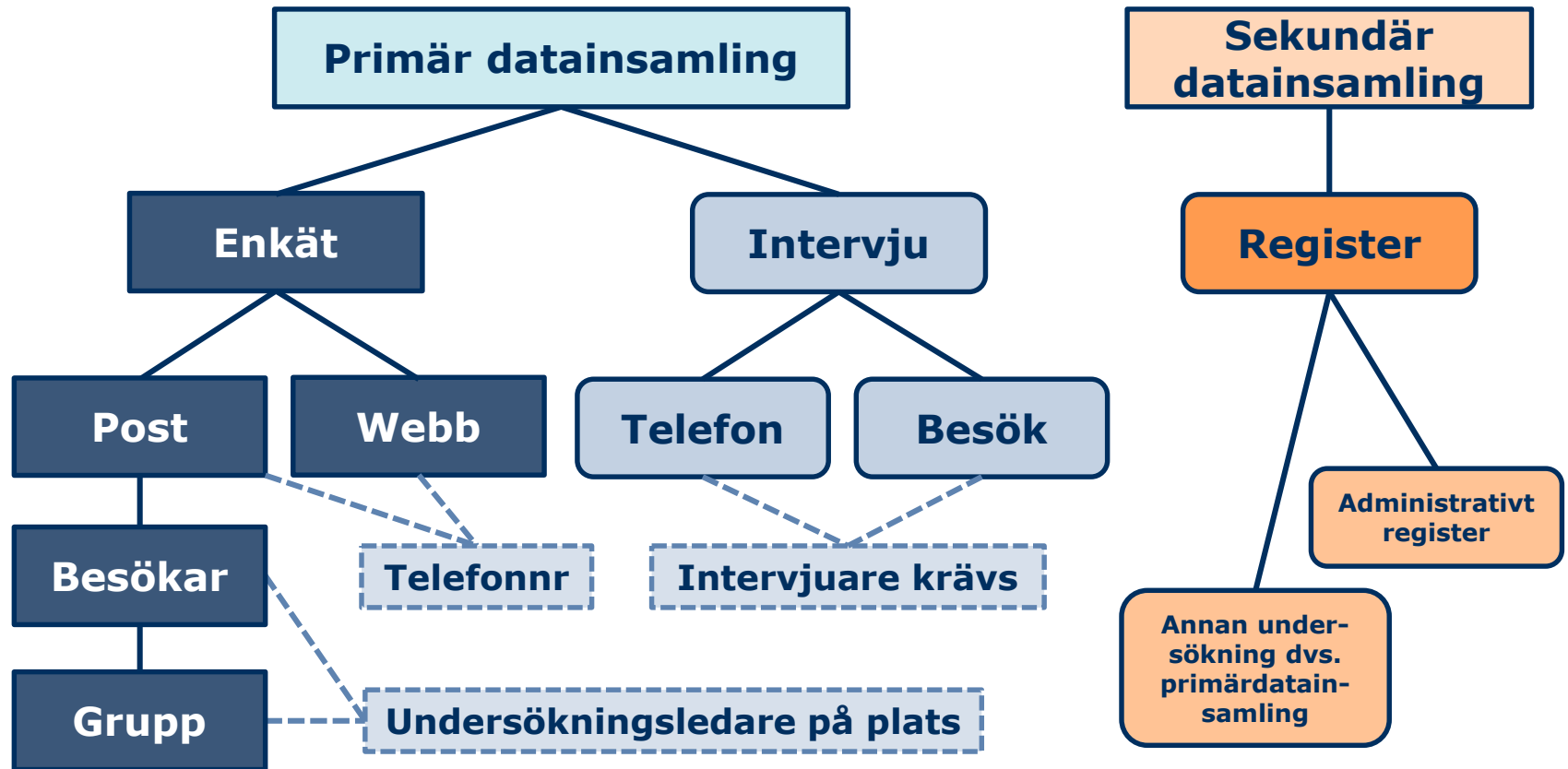
Sekundärdata

- "Andra" data, ofta insamlat för något annat syfte än din undersökning.
 - Ex. en annan undersökning, **registerdata**, administrativa räkningar.
 - Kan användas vid skrivbordsundersökningar, testa programvara osv.
 - Går inte att styra på samma sätt, oftast krävs bearbetning.
 - Billigt och kostnadseffektivt.

Man kombinerar nästan alltid primär- och sekundärdata!



Datakällor och datainsamling, forts.



Papper vs Webb

Grov jämförelse – det finns alltid utrymme för debatt och diskussion.

Papper

- Dyrare - papper och porto
- Billig konstruktion(?)
- Formuläret är en för alla - störande hoppinstruktioner
- Bättre ramar – man skickar till adresser
- Hållbarhet? Papper

Webb

- Billigare distribution, snabbare
- Dyr konstruktion(?), förr kanske
- Formuläret kan anpassa sig automatiskt efter hur man svarar
- Ramar med e-postadresser svåra att hitta, första kontakt via post
- Hållbarhet? Elektronik och batterier

Enkäter vs Intervju

Grov jämförelse – det finns alltid utrymme för debatt och diskussion.

Enkäter - självadministrerade

- Respondenten känner sig inte speciell, det är opersonligt
- Respondentens frågor kan inte besvaras (nummer att ringa?)
- Billigare
- Man kan ta det i sin egen takt
- Ingen intervjuareffekt

Intervjuer – styrs av intervjuare

- Respondenten är speciell, personligt
- Intervjuare kan förklara otydligheter
- Dyrare, mycket dyrare!
- Måste göras "nu"
- Intervjuareffekter

Mixed mode

Vanligt numera att man gör en **mix av insamlingsmetoder** inom samma undersökning – anpassning till respondenternas villkor och beteenden.

- Kan ge ökad kvalitet (anpassat till respondenten)
- Fler som svarar – förhoppningsvis (anpassat till respondenten)
- Dyrare (lite?) – inte bara en insamling som ska designas
- Svårare skapa likvärdiga mätningar, ska helst ge samma svar oavsett mode:
 - Effekter av de olika sätten påverkar resultatet – s.k. **mode effect**

Datainsamling

Enligt min f.d. kollega P. Lundquist, SCB

Kriterium	Enkät	Webb	Telefon	Besök
Låga kostnader	++	++	-	--
Kort tid - snabbhet	-	+	+	-
Svarsandel (lågt bortfall)	-	-	+	++
Medverkan, kräver insats	-	-	+	++
Komplext ämne	-	-	+	++
Svaren finns i systemen	+	+	-	-
Känsliga frågor	+	+	-	-
Många svarsalternativ	+	+	-	+
Många öppna frågor	-	-	+	++
Många hopp (filterfrågor)	-	++	+	+
Visuella hjälpmedel	+	++	--	+



Vilken metod ska man välja?

- Beror på vad vi vill ha och vad vi kan få.
- Kompromisser alltid nödvändiga!



Registerdata

Definition:

Ett (data-)register med en (idealt) fullständig företeckning över samtliga objekt i en viss objektmängd (målpopulation). Det är önskvärt att flera bakgrundsvariabler finns tillgängliga i registret.

- Används som **urvalsramar** vid primär datainsamling.
- **Samkörning** av register och undersökningar:
 - primärdata kan kompletteras med sekundärdata.
 - billigare per undersökningsobjekt.
 - finns även möjliga nackdelar.
- I Sverige finns det gott om register!

Registerdata, forts.

Fördelar

- Billigare per undersökningsenhet än primärdatainsamling
- Möjlighet att följa specifika objekt över tiden
- Användning i kombination med primärdata
- Minskad uppgiftslämnarbörda (färre frågor krävs)

Nackdelar

- Registrets ursprungliga syfte ej detsamma som undersökningens
- Andra variabeldefinitioner
- Andra avgränsningar av populationen och element
- Krävs ofta bearbetning av sekundärdata
- Viss risk för inaktuella uppgifter, över- resp. undertäckning



Registerdata – exempel från SCB

- **Registret över totalbefolkningen (RTB)**
 - sedan 1968, förvaltas av SCB, utdrag ur folkbokföringsregistret som ansvaras av Skatteverket.
- **Företagsdatabasen (FDB)**
 - register över samtliga företag, myndigheter, organisationer och deras arbetsställen. Uppdateringar av databasen sker varje vecka.
- **Befolkningens utbildning (UREG)**
 - visar bl.a. individers utbildningsnivå och inriktning; samt kön, ålder och nationell bakgrund mm.
- **Yrkesregistret**
 - visar hur många som arbetar inom olika typer av yrken, yrkesutveckling inom olika branscher och samhällssektorer.



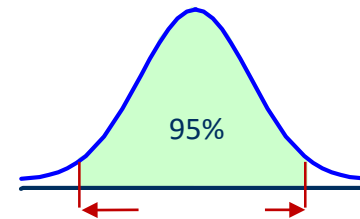
2. FEL I STATISTISKA UNDERSÖKNINGAR

- **Sampling error - Urvalsfelet**
 - man drar ett urval men man observerar inte alla
 - bra precision = liten varians = litet konfidensintervall
 - kan hanteras med statistisk teori
- **Non-sampling error – Icke-urvalsfel, systematiska fel**
 - andra fel än samplingfelet
 - ger typiskt bias, systematiska fel
 - kan hanteras med statistisk teori men modellberoende, antaganden

Sampling error – Urvalsfel

- Avvikelsen från det sanna och okända parametervärdet som beror på att vi inte har gjort en totalundersökning utan endast från ett **urval**.
 - vi har dragit **slumpmässigt** och sedan **skattat** parametern.
- Med **sannolikhetsurval** kan vi kvantifiera precisionen (**osäkerheten**)
 - samplingfel = variansskattning, standardfel, felmarginal

$$\underbrace{\bar{y}}_{\text{punktskattning}} \pm \underbrace{t_{(n-1)} \cdot \frac{s_y}{\sqrt{n}}}_{\text{felmarginal}}$$



När $n \rightarrow \infty$ så går
felmarginalen $\rightarrow 0$

Non-sampling errors – icke-urvalsfel

- **Fel som** inte beror på urvalsförfarandet utan **uppstår av andra skäl**.
- Kan medföra att **urvalsfelet över- eller underskattas**:
 - ex. de med extremvärden (stora och små) inte svarar,
 - ex. mätfel gör att observerade värden har för stor spridning.
- Kan framförallt medföra **systematiska fel = bias** i skattningen:
 - skattningar **är inte väntevärdesriktiga**,
 - i genomsnitt hamnar vi fel!
 - kan också medföra under- och överskattningar av urvalsfelet.

Fyra huvudtyper av icke-urvalsfel

1. Täckningsfel, ramfel

- ramen är inte överensstämmande med målpopulationen.

2. Mätfel

- mätningarna ger inte sanna värden.

3. Bearbetningsfel

- felregistrering, felkodning, felräkning, ...

4. Bortfallsfel

- alla svarar inte,
- objektsbortfall och partiellt bortfall.



1. Ram- och täckningsfel

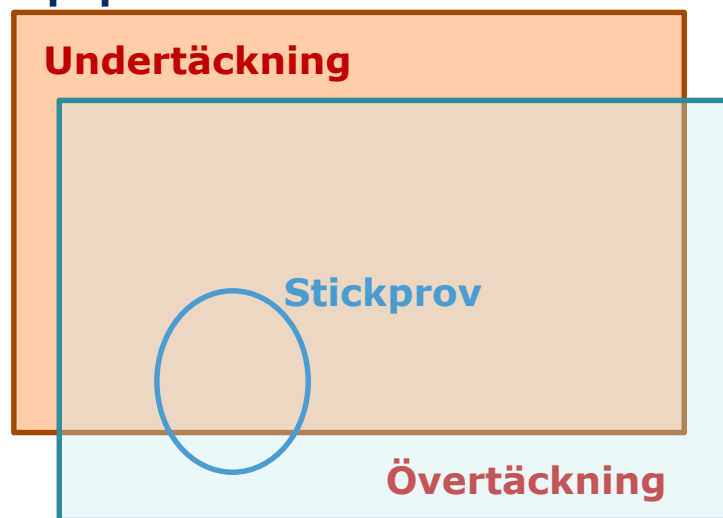
- **Urvalsram** eller bara **ram** (eng. *frame*)
 - Ram = en lista, datafil, register eller liknande som listar alla objekt i en ändlig population.
 - Man drar urvalet från ramen = **rampopulationen**.
 - Idealt ska det finnas en exakt matchning med de man vill undersöka = **målpopulationen**.
- Ibland är inte ramen uppdaterad; ibland är den bara dålig.
 - **Undertäckning** = alla de objekt som tillhör målpopulationen men som saknas i ramen.
 - **Övertäckning** = alla de objekt som finns i ramen men som inte tillhör målpopulationen.
 - De objekt i stickprovet som ingår i övertäckningen kan ofta identifieras genom kontrollfrågor och då läggas åt sidan.



Ram- och täckningsfel, forts.

- Om det finns strukturella skillnader mellan mål- och rampopulationerna och dessa skillnader samvarierar med det man vill undersöka så riskerar man systematiska fel.

Målpopulation = de vi vill undersöka



Rampopulation = de som man känner till och som kan dras till urvalet

Exempel:

- Nyfödda och nyligen immigrerade finns inte i registret ännu.
- Nyligen avlidna och emigrerade finns kvar i registret.
- Man undersöker behovet av förskoleplatser inom kommunen för de kommande 5 åren.



2. Mätfel

Mätfel = skillnaden mellan erhållet svar och sant värde.

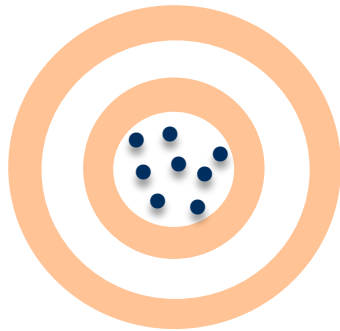
- Mätningen kan ge upphov till slumpmässiga fel (dåligt instrument):
 - variationen som uppstår om vi mäter samma personer flera gånger med samma metod förutsatt att det sanna värdet inte förändras
 - variationen är graden av tillförlitlighet = **reliabilitet**
- Mätningen kan även ge upphov till systematiska fel:
 - $Y = T + \varepsilon$ där Y = uppmätt värde, T = sant värde och ε = mätfel
 - om variansen för ε är liten \Rightarrow hög reliabilitet,
 - om variansen för ε är stor \Rightarrow låg reliabilitet
 - om $E(\varepsilon) \neq 0 \Rightarrow$ systematiskt fel, bias



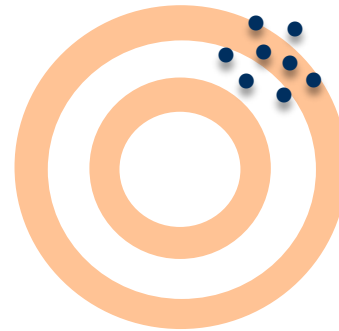
Validitet och reliabilitet

- **Validitet** betyder att mätningen mäter det man vill mäta, dvs. mätningens relevans.
 - Hög validitet kräver hög reliabilitet
 - **Hög reliabilitet är ingen garanti för validitet** (fel inställd våg),
 - Låg reliabilitet \Rightarrow låg validitet.
- Ett dåligt sätt att mäta på kan ge upphov till systematiska mätfel:
 - **systematiska mätfel som ger upphov till bias gör att validiteten är låg.**
- Läs gärna mer på <https://sv.wikipedia.org/wiki/Validitet>

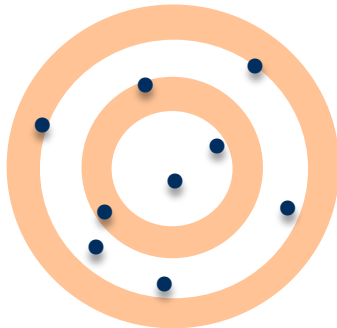
Reliabilitet och validitet



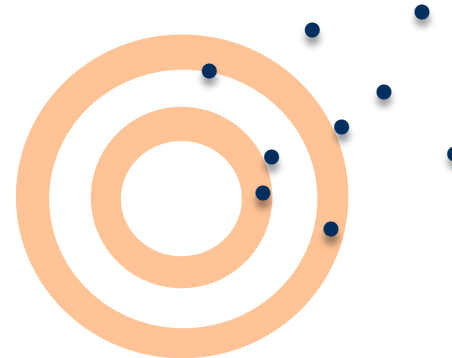
Hög reliabilitet, hög validitet



Hög reliabilitet, låg validitet



Lägre reliabilitet, lägre validitet

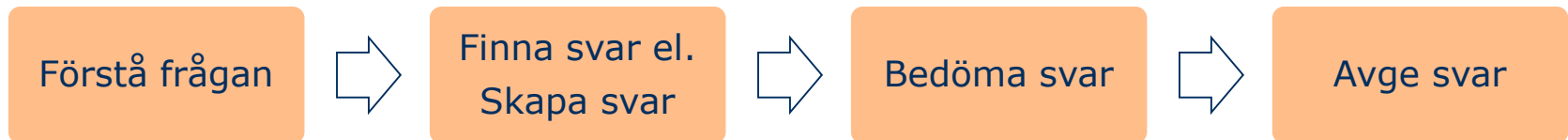


Låg reliabilitet, hög validitet

När uppstår mätfelet?

T.ex. vid individundersökningar:

- När som helst i den s.k. **CASM**-modellen*/4-steps modellen:



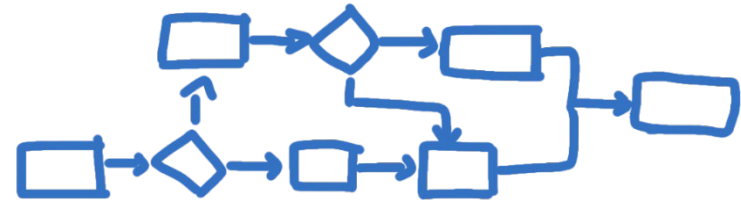
- Satisficing** – tillräckligt bra, bekväm väg genom formuläret
- Social önskvärdhet** – vill inte vara udda, prestigebias, integritet, känsliga frågor
- Aquiescence** – tendens att "hålla med", vara positiv

* CASM = Cognitive Aspects of Survey Methodology

3. Bearbetningsfel

Efter insamlingen ska rådata bearbetas och fel kan uppstå vid:

- **Kodning**
- **Registrering in till dator**
- **Datorbearbetning**



- Allmänt anses bearbetningsfelet som den **minst problematiska** och den feltyp som bidrar minst till det totala felet:
 - men det finns undantag ...

Färre människor inblandade ⇒ färre *human errors*!

Fler automatiserade steg ⇒ svårare att upptäcka fel?

Bearbetningsfel – exempel

(DN 24 mars 2010-03-24)

”Under 2008 gjorde SCB det så kallade skofelet i konsumentprisindex (KPI), som gav upphov till en felaktigt hög inflationssiffra inför Riksbankens sista räntehöjningar precis före Lehman Brothers-kraschen.

Skofelet, där felaktiga prisangivelser på skor rörde till det, fick även effekter på beräkningen av ersättningsnivåer från försäkringskassan, vilket i sin tur hade en negativ effekt på statskassan.”

Allt detta pga. en nolla för mycket i en Excel-cell ...

4. Bortfallsfel

Definition:

- De objekt i ramen som tillhör målpopulationen och som ingick i urvalet men som man inte fick något svar ifrån.

Man skiljer på:

- **Individbortfall**, objektsbortfall – svarar inte alls, ingen kontakt
- **Partiellt bortfall**, variabelbortfall – svarar inte på vissa frågor

Bortfallskategorier

- **Ej anträffbara**

- bytt namn, flyttat, bortresta, screenare, kontaktuppgifter saknas.

- **Vägrare**

- av princip, rädsla för intrång i privatlivet, orkar inte, ointresse för själva undersökningen.

- **Övrigt**

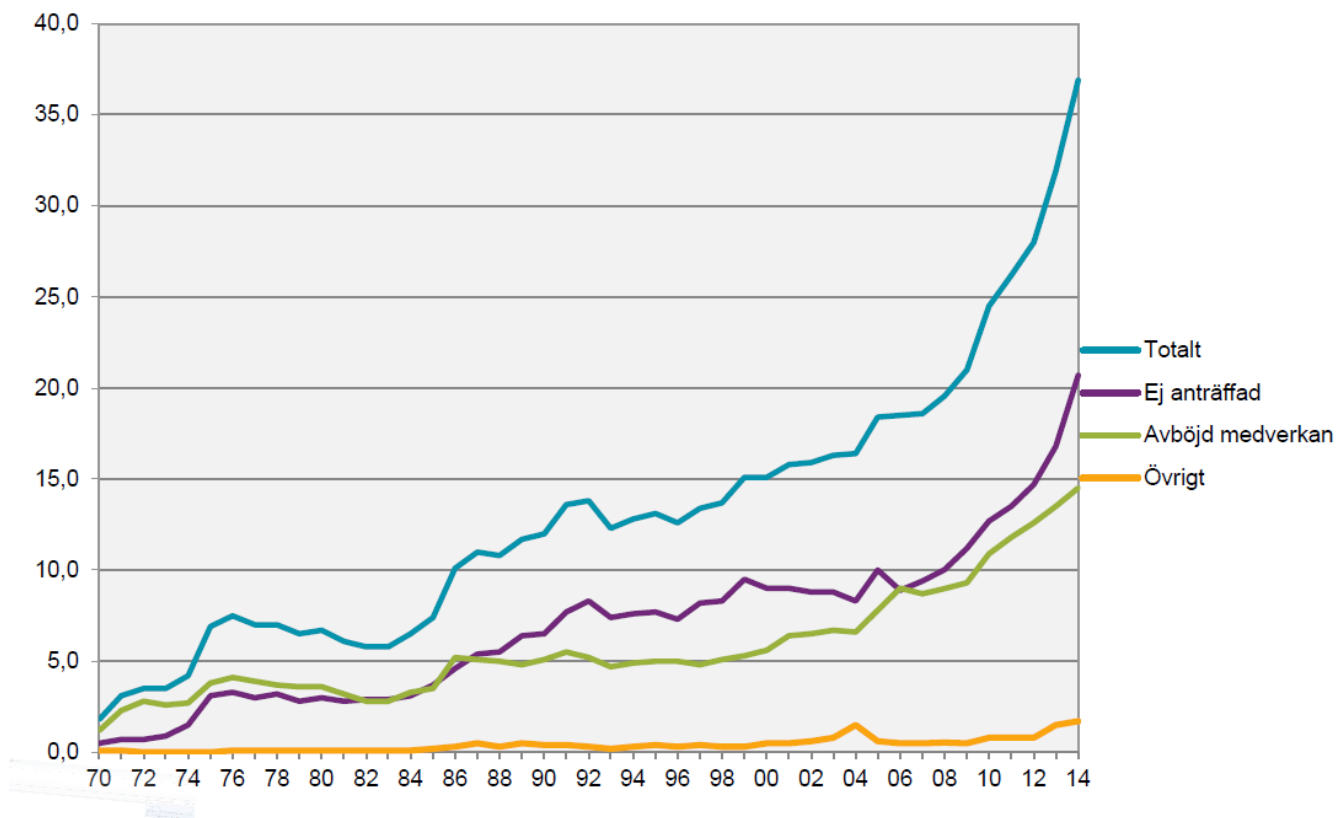
- språk, förstår inte, gammal, sjuk, intagen,
- webbformuläret hängde sig, tekniska problem,
- missar i hanteringen, borttappade brev och enkäter.

Bortfall

- Orsakerna kan vara många.
 - Svår att nå (mobiler), trött på okända tel.nr., "survey fatigue", ...
- Om benägenheten att svara eller inte svara, direkt eller indirekt beror på undersökningsvariabeln så införs **bias**:
 - Ex. storstad svarar mindre än glesbygd, unga mindre än äldre osv.
- Om bortfallet är litet ($< 10\%$) kan det kanske accepteras.
 - Många undersökningar idag har betydligt större bortfall ($> 50\%$)

Ex. Bortfall

Figur 4. Bortfallet i AKU 1970–2014, åldersgruppen 16–64 år, ovägt i procent på årsbasis



Historiskt exempel

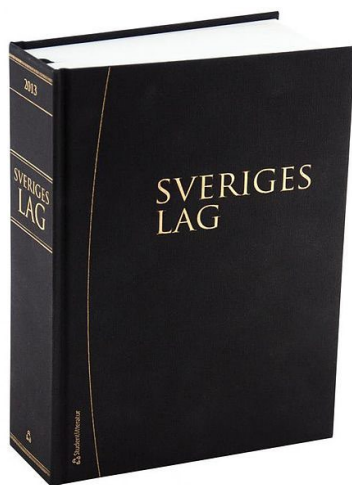
En klassiker:

- USA:s presidentval 1936,
Landon (R) mot Roosevelt (D)
- **Literary Digest** hade förutsett vinnaren i de fem senaste valen.
 - 10M enkäter utskickade, 2.3M kom in (**67 %** bortfall).
 - **Urvalsramar**: lista över Literary Digests prenumeranter, bilägarregister och telefonkataloger.
 - Uppenbara **bortfalls-** och **täckningsproblem!**



3. KVALITET I STATISTISKA UNDERSÖKNINGAR

- Kvalitetskriterierna för Sveriges Officiella Statistik är reglerade i **Lag (2001:99) om den officiella statistiken:**



© Studentlitteratur AB

- Relevans
- Noggrannhet, tillförlitlighet
- Aktualitet
- Punktlighet
- Tillgänglighet, tydlighet
- Jämförbarhet
- Samstämmighet

Kvalitetskriterier

- **Relevans**

- Ändamål och informationsbehov och användares informationsbehov
- Statistikens innehåll
 - statistiska mått
 - redovisningsgrupper
 - referenstider

- **Tillförlitlighet**

- Osäkerhetskällor

- urvalsfel
- ramtäckning
- mätning
- bortfall
- bearbetning



Feltyperna som vi precis har pratat om

- modellantaganden **Feltyp som vi inte har pratat om**



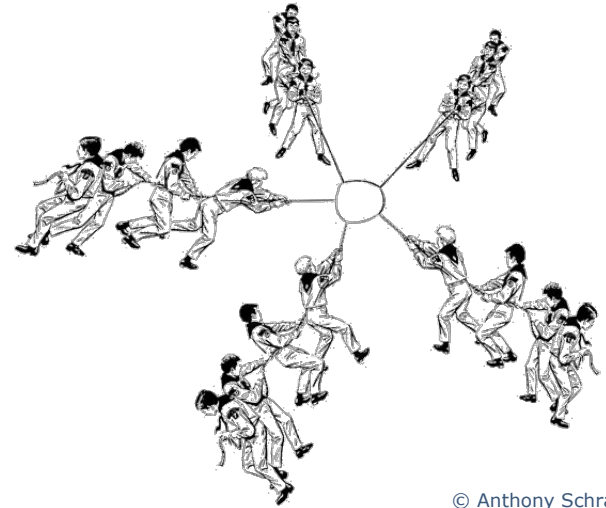
Kvalitetskriterier, forts.

- **Aktualitet** och **Punktlighet**
 - Framställningstid (dröjer det länge?)
 - Frekvens (hur ofta?)
 - Punktlighet (publiceras vid utlovat klockslag)
- **Tillgänglighet, tydlighet**
 - Tillgång till statistiken (publicering, webb, mm.)
 - Möjlighet till ytterligare statistik
 - Presentation och dokumentation
- **Jämförbarhet** och **samanvändbarhet**
 - Jämförbarhet över tid
 - Jämförbarhet mellan grupper
 - Samanvändbarhet i övrigt, med andra undersökningar
 - Numerisk överensstämmelse

Potentiella målkonflikter

7 kriterier ger 21 olika par att jämföra

- Snabb och slarvig eller pedantisk och långsam?
 - **punktlighet** kontra **noggrannhet**
- Nytt och dynamiskt eller konstant och stabilt?
 - **relevans och samstämmighet** kontra **jämförbarhet över tid**
- Man måste bestämma vad som är viktigast i en given kontext
 - Ibland måste man kompromissa.
 - Hur mycket kvalitet får man för en given summa pengar.



© Anthony Schrag

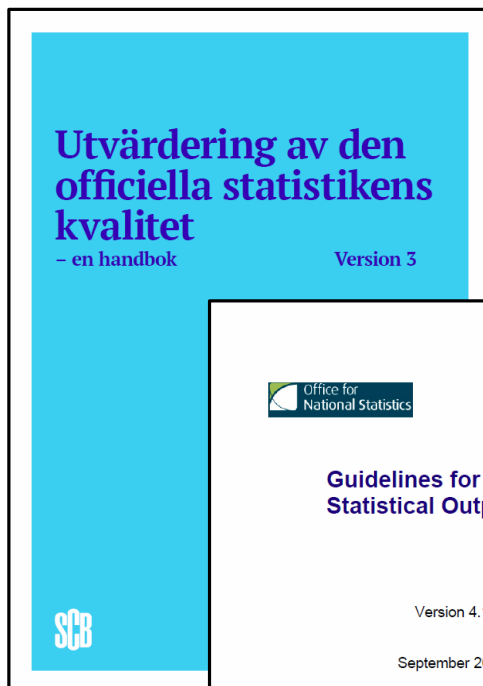
Kvalitetsredovisning

- Bör/ska alltid finnas med när statistik tillgängliggörs.
 - jämför med när du ska köpa en ny diskmaskin, du vill ju veta vilka egenskaper den har, inte bara förlita dig på färgen och en logga.
- En statistikprodukts **värde** = dess **användbarhet**.
- En kvalitetsredovisning ska
 - uppmärksamma användare på eventuella fel som begränsar värdet,
 - ge bra beslutsunderlag, minska risken för fel beslut,
 - göra det möjligt för användare att precisera sina krav på kvalitetsförbättringar.
- **Mål:** objektivt sammanställd information om ett samhälles tillstånd.

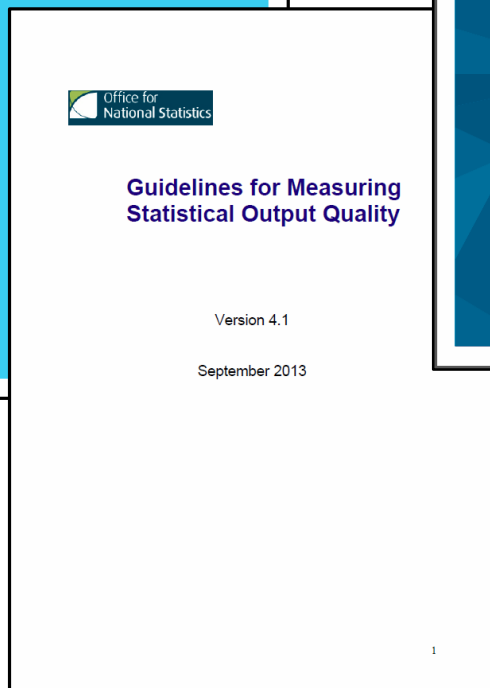
Exempel

[Länk](#)

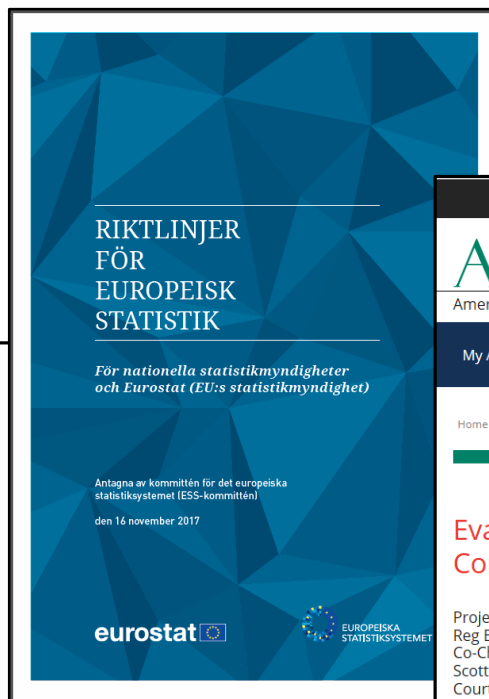
[Länk](#)



[Länk](#)



[Länk](#)



Stockholms
universitet

4. SVERIGES OFFICIELLA STATISTIK (SOS)


Officiell statistik – produceras av offentliga myndigheter.

- **Officiell statistik ska ge en bild av samhället:**
 - den ska berätta något viktigt om vårt samhälle, något som är av allmänintresse (relevans och aktualitet).
- **Ska ha hög kvalitet, den får inte vara opålitlig:**
 - kräver hög kompetens och bra metoder, god statistisk infrastruktur (noggrannhet och tillförlitlighet).
- **Ska vara objektiv, transparent och tillgänglig för alla samtidigt:**
 - fri från politisk påverkan, kvalitetsdeklarerad och öppen (tillgänglighet och tydlighet),
 - ska framställas regelbundet och med en långsiktig plan (aktualitet och punktlighet).



Lag (2001:99) om den officiella statistiken

Officiell statistik skall:

- finnas för allmän information, utredningsverksamhet och forskning. Den skall vara objektiv och allmänt tillgänglig;
- framställas och offentliggöras med beaktande av behovet av skydd för fysiska och juridiska personers intressen;
- dokumenteras, kvalitetsdeklareras, samt utan avgift offentliggöras och hållas allmänt tillgänglig i elektronisk form genom ett allmänt nätverk.
- När officiell statistik görs tillgänglig skall den vara försedd med beteckningen Sveriges officiella statistik:  Sveriges officiella statistik

Länk: [Sveriges officiella statistik \(scb.se\)](https://scb.se)

29 statistikansvariga myndigheter

Arbetsmiljöverket	Medlingsinstitutet	
Brottsförebyggande rådet	Myndigheten för familjerätt och föräldraskapsstöd	
Centrala studiestödsnämnden	Myndigheten för kulturanalys	
Domstolsverket	Myndigheten för tillväxtpolitiska utvärderingar och analyser	
Ekonomistyrningsverket	Naturvårdsverket	
Finansinspektionen	Pensionsmyndigheten	
Folkhälsomyndigheten	Riksgäldskontoret	Sveriges lantbruksuniversitet
Försäkringskassan	Skogsstyrelsen	Tillväxtverket
Havs- och vattenmyndigheten	Socialstyrelsen	Trafikanalys
Kemikalieinspektionen	Statens energimyndighet	Universitetskanslersämbetet
Konjunkturinstitutet	Statens jordbruksverk	Statistiska centralbyrån (SCB)
Kungliga biblioteket	Statens skolverk	

24 statistikområden

Arbetsmarknad

Befolkning

Bostäder och byggande

Demokrati

Energi

Finansmarknad

Folkhälsa

Handel med varor och tjänster

Hushållens ekonomi

Hälsa- och sjukvård

Jordbruk- och skogsbruk, fiske

Kultur och fritid

Levnadsförhållanden

Miljö

Miljövård och naturresurshushållning

Nationalräkenskaper

Näringsverksamhet

Offentlig ekonomi

Priser och konsumtion

Rättsväsende

Socialförsäkring

Socialtjänst

Transporter och kommunikationer

Utbildning och forskning

Officiell statistik, forts.

- **Vad ska den officiella statistiken publicera och i vilken form?**

- en avvägning mellan att berätta för användaren vad statistiken säger och låta användaren läsa själv.

- **Planerings- och beslutsunderlag** för myndigheter och företag

- ex. riksdagens utredningstjänst, rättsväsendet, kommuner och landsting, banker och företag.

- **Samhällsdebatten**

- journalister använder officiell statistik frekvent
- "väckarklockor"; ex. USA-valet, PISA, arbetslösheten, ...

PISA = Program for International Student Assessment

- **Forskning:**

- universiteten är stora användare av offentlig/officiell statistik.



Tre viktiga officiella statistikprodukter

- **Konsumentprisindex (KPI)**

- mäter prisutvecklingen på varor och tjänster i landet, dvs. inflationstakten
- styr bl.a. storleken på sociala förmåner, används som underlag i avtal (prisjusteringar, löneavtal mm.)
- indelad i olika del-KPI för olika grupper av varor och tjänster och för olika marknader

- **Arbetskraftsundersökningen (AKU)**

- underlag för arbetsmarknadspolitiska beslut och bedömning av konjunkturutvecklingen i ekonomin.

- **Nationalräkenskaperna (NR)**

- sammanfattar och beskriver landets ekonomiska aktiviteter; värdet av produktionen av varor och tjänster, inkomstbildning, omfördelning, transaktioner med utlandet mm.



Lästips

- **K Dahmström: Från datainsamling till rapport**
 - framförallt Kap 5, 6 och 12 har tagits upp idag
- **J Bethlehem: Applied Survey Methods**
 - finns att ladda ner gratis via SU Biblioteket
- Finns en hel del intressant att ladda ner och läsa på www.scb.se