

Statistisk översikt kurs - Föreläsning 3

Anders Fredriksson

Statistiska Institutionen
Stockholms Universitet

26 mars 2025



Stockholm
University

Föreläsningen ger en kort översikt och introduktion till statistiska undersökningar och datainsamling, med fokus på

- **Mål och syften** med statistiska undersökningar
- **Grundläggande begrepp** - Tvärsnitt, tidsserie, primärdata, sekundärdata, population, urval, etc.
- **Experimentella studier**
- **Observationsstudier**
- Mot slutet av kursen (F10) kommer Michael Carlson från Statistiska institutionen att gästföreläsa om officiell statistik, registerdata, etc.

Mål och syften med statistiska undersökningar

- **Beskrivande** - svarar på frågor som *"hur många"*, *"hur stor andel"*, *"hur mycket"*, *"hur länge"*?
 - Opinionsundersökningar - hur ser opinionen ut?
 - Marknadsundersökningar - utvärdering av olika reklamkampanjer
 - Kvalitetskontroll - andelen defekta varor vid produktion
- **Förklarande/utredande/analytiska** - hitta och förklara samband mellan variabler (orsak och verkan), testa hypoteser.
 - Hur ser sambandet mellan utbildning och lön ut?
 - Har andelen moderater ökat sedan senaste mätningen?
 - Är andelen defekta varor i produktionen signifikant högre än 5 procent?

- **Prognosticerande** - göra förutsägelser om framtiden
 - Valprognos
 - Olika ekonomiska prognoser, inkl. om räntan
 - Befolkningsprognos i en kommun
 - Väderprognos
 - Försäljningsprognos

- **Tvärsnittsdata** (cross-sectional) - de som ingår i studien observeras en gång.
 - Ett tvärsnitt av populationen, verkligheten vid en viss tidpunkt.
 - **Ex:** Alla Titanicfigurerna från föreläsning 2
- **Tidsseriedata, longitudinell data** (time series, longitudinal) - variabeln som ingår i studien observeras vid flera tidpunkter.
 - Förändringar över tid är i fokus.
 - **Ex:** Grafen över Riksbankens ränta från föreläsning 2
- **Paneldata** (panel) - flera olika variabler och objekt över tid.
 - **Ex:** Olika nyckeltal, för ett eller flera företag, för varje år.

- **Retrospektiv** - tillbakablickande i tid, sådant som redan har hänt.
 - **Ex:** En studie som görs idag, med tidigare insamlade data från sjukvården, om någon effekt av coronapandemin
- **Prospektiv** - framåtblickande i tid, innan data finns.
 - Kan ofta planeras och styras av forskaren.
 - Observations- eller experimentell studie, longitudinella studier.
 - **Ex:** Följ två grupper av individer framåt i tiden och samla in data om hälsovariabler.

■ Primärdata

- Nya data, primärt insamlat för ett visst syfte.
- Datainsamlingen anpassas till den aktuella undersökningsfrågan.
- **Ex:** Intervjuer med kaffebönder (föreläsning 1)
- **Ex:** Valu-undersökningen i sig (intervjuer utanför vallokal).

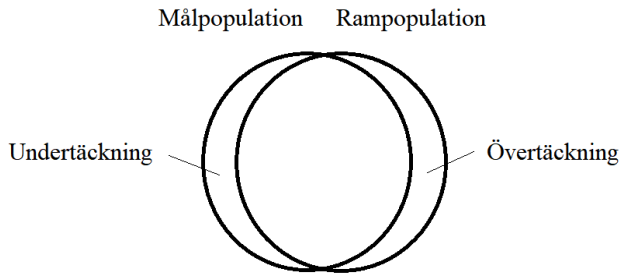
■ Sekundärdata

- Data som redan finns, insamlade för något annat syfte.
- Exv. en annan undersökning, registerdata, administrativa data.
- Billigare än att samla in egna data.
- **Ex:** Data från tidigare val som också används i valprognosen.

- **Population, målpopulation** (target population, population of interest)
 - Vid en statistisk undersökning vill man studera en (mål-)population; en mängd individer eller objekt med minst en gemensam egenskap - det man ska mäta tex längd, vikt, partisympati.
 - **Ex 1:** Alla röstberättigade
 - **Ex 2:** Mönstringsskyldiga
- **Urval** ("Sample")
 - En del av populationen (oftast en liten del). Individerna/objekten kallas urval. Man efterstavar **slumpmässiga urval** (mer om slumpen följer).
 - Kan också kallas **stickprov**
 - **Ex 1:** 1054 personer som deltagit i en opinionsundersökning ("Om det vore val idag, vilket parti")
 - **Ex 2:** Gymnasieelever (år 3) på en informationsdag om försvaret

- **Totalundersökning** (census, total survey)
 - Alla individer eller objekt i en population observeras.
 - **Ex:** En kommun vill veta vad rektorerna i skolor tycker i en viss fråga (typiskt en liten grupp)
 - **Ex:** Mönstring kan ses som en totalundersökning
- **Urvalsundersökning** (survey, sample survey)
 - Ett urval individer eller objekt i en population observeras.
 - **Ex:** Opinionsundersökningar
 - **Ex:** Marknadsundersökningar
 - **Ex:** TV-tittande ("Hur många såg melodifestivalen?")

Begrepp - Målpopulation, rampopulation, täckning



■ Ram, rampopulation ("frame")

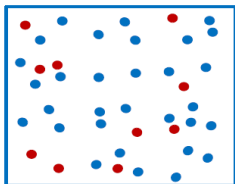
- Ett register eller liknande, för att dra ett urval
- Ram- och målpopulation (de vi vill undersöka) kanske inte matchar
- **Ex:** Skyddad identitet, i individundersökning baserad på befolkningsregister (**undertäckning**)
- **Ex:** Oregistrerade brott, i brottsregisterundersökning (undertäckning)
- **Ex:** Individer över 18 år i bilförarundersökning, baserad på befolkningsregister (**övertäckning**, alla har inte körkort)

Sannolikhetsurval (probability sampling)

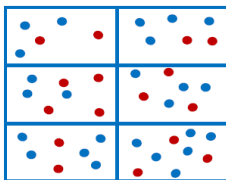
- Alla individer eller objekt ska ha en känd sannolikhet (större än noll) att komma med i urvalet.
- Urvalet kan väljas på olika sätt i en urvalsundersökning.

Urvalsdesigner från en ändlig population

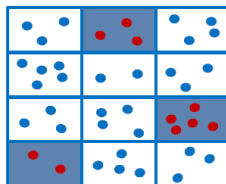
Exempel:



Obundet slumpmässigt urval – dra några från populationen



Stratifierat urval – dra några från varje grupp (delpopulation)



Klusterurval – dra hela grupper och undersök samtliga objekt i dessa

● = valda och undersökta objekt

- **Obundet slumpmässigt urval (OSU)** (Simple random sampling)
 - Alla individer eller objekt i populationen har samma sannolikhet att bli dragna.
 - Urvalet blir en mindre "kopia" av populationen.
 - Enheterna dras med t.ex. en slumpvalsgenerator (mer om slumpen kommer i föreläsning 4).

- **Systematiskt urval** (Systematic sampling)

- Exempel: Var 10:e objekt i listan från urvalsramen dras.
- Startobjektet slumpas fram.
- Kan ge upphov till systematiska fel om listan inte är slumpmässigt ordnad.
- Om listan är slumpmässigt ordnad är OSU och systematiskt urval likställda.

- **Stratifierat urval** (Stratified sampling)
 - Populationen delas in i grupper (**strata**).
 - Objekt inom varje **stratum** är lika varandra med avseende på någon relevant egenskap.
 - Ett OSU eller systematiskt urval dras sedan från alla grupper.
 - Säkerställer att alla strata blir representerade i urvalet.

- **Klusterurval** (Cluster sampling)
 - Urvalsramen består av grupper av objekt.
 - Slumpmässigt urval av grupper, sedan undersöks samtliga objekt i utvalda grupper.
- Skillnad mot stratifierat urval:
 - Stratifierat: Strata ska vara homogena inom sig och heterogena mellan sig.
 - Kluster: Grupperna ska vara heterogena inom sig eftersom endast ett urval av grupper kommer med.

Andra urvalsmetoder än sannolikhetsurval

- När vi i kursen arbetar med exv. inferens förutsätter vi att vi har ett sannolikhetsurval
- Metoderna nedan behandlas inte vidare men kan vara bra att känna till
 - **Kvoturval** - en intervjuare får i uppdrag att exv. intervjua fem kvinnor och fem män - stor risk att det inte blir ett sannolikhetsurval
 - **Självrekrytering**, exv. webbpaneler (vem väljer att gå med i panelen?)
 - **Bekvämlighetsurval**, exv. intervjua utanför universitetets tunnelbana om vi vill nå gruppen "unga"
 - **Snöbollsurval** de som redan är med i en studie rekryterar andra



- **Experimentella studier** - Studera eventuella skillnader vid olika aktivt valda förhållanden och behandlingar.
- Försöket kan upprepas önskat antal gånger.
- Försöksbetingelserna kan kontrolleras.
 - **Ex:** effekter av kosttillskott
 - Slumpa in individer i två grupper, **experiment** (treatment) och **kontroll** (control), ge en grupp kosttillskott och en grupp placebo
 - Bakgrundsvariabler i medel lika mellan de två grupperna (men andra problem kan uppstå)
- Ger möjlighet att studera **kausala effekter**.
- Samhällsvetenskaplig forskning, exempelvis utvecklingsekonomi och arbetsmarknadsekonomi, använder också experiment.

■ Randomisering

- Indelning av experimentobjekt slumpmässigt i en kontrollgrupp och en experimentgrupp.
- Eventuella skillnader mellan grupperna beror på behandlingarna, inte på bakgrundsvariabler.
- Skapar två jämförbara grupper baserade på bakgrundsvariabler.

■ Förväxlingsvariabler (Confounding variables)

- Okända eller okontrollerade bakgrundsvariabler som kan påverka resultatet.
- Slumpmässig indelning minskar sannolikheten för påverkan från förväxlingsvariabler.

■ **Kvasiexperiment** (Quasi-experiment)

- Ibland kan inte randomisering utföras, ofta av etiska skäl (Ex: mediciner, droger).
- För att utvärdera effekten av drogen bildar man en experimentgrupp med personer som redan tagit drogen (frivilligt).
- Jämförelsegruppen bildas av individer som är så lika de i experimentgruppen som möjligt med avseende på bakgrundsvariablerna.

Exempel - experimentell studie

- Ge arbetssökande fler möten med arbetsförmedlare för att studera om chansen att få ett jobb påverkas
- Målgrupp - inskrivna på arbetsförmedlingskontor
- Randomisering - vilka kontor som skulle vara med i studien och vilka av de individer som var inskrivna på de aktuella kontoren som fick ytterligare möten med arbetsförmedlare (något förenklat)
- **Kontrollgrupp(er)**: individer som fick fortsätta med möten som tidigare.
- **Behandlingsgrupp/Experimentgrupp**: individer som fick fler möten.
- I aktuell studie*, som visade att fler möten förbättrade möjligheten att få ett jobb, studerades också undanträngningseffekter.

* Cheung, M., et al, 2024. Effekter av förstärkta förmedlingsinsatser – lärdomar från ett randomiserat experiment. Ekonomisk Debatt - ingår INTE i kursen



- **Observationsstudier** / Icke-experimentella studier - Passivt observera utan att påverka.
- Av olika anledningar, exv. etiska, ekonomiska, praktiska, kan vi inte utföra ett experiment
- Observationsstudier ofta enklare att genomföra än ett experiment
- **Ex:** Enkätundersökning om kosttillskott, om individers konsumtion och (upplevda) effekter.
- Bakgrundsvariabler kan påverka både vem som svarar på enkäten, huruvida personen tar kosttillskott, samt effekten av kosttillskotten.
- Vi kan eventuellt samla in data på vissa bakgrundsvariabler och "kontrollera för dessa".

- Två grupper följs över tid, exv. rökare/icke-rökare, de som tar/inte tar en medicin
- Vi är intresserade av att studera (exv.) hälsoutfall, men kan inte välja vem som ingår i respektive grupp
- Ett sätt att göra grupperna likartade är att jämföra ("matcha") individer med samma karakteristika (exv: liknande ålder, kön, inkomst)
- En utmaning är att vi inte känner till alla faktorer som påverkar och kanske inte heller har eller ens kan få data

Denna version av dokumentet: 2025-03-26

Materialet i Statistisk översikt kurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, med flera.

