

Föreläsning 10: Population och stickprov. Experiment.

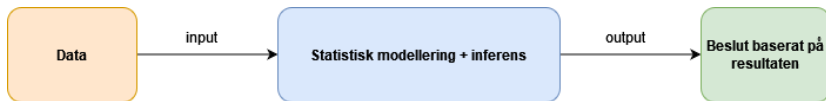
Matias Quiroz¹

¹Statistiska institutionen, Stockholms universitet

VT 2023

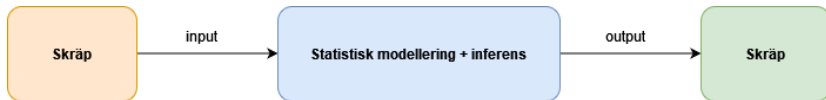
- ▶ Statistikerns roll.
- ▶ Från population till stickprov.
- ▶ Stickprov som inte generaliserar till populationen.
- ▶ Observationsstudier kontra experiment.
- ▶ Randomiserade experiment för att påvisa kausalitet.
- ▶ Randomiserade experiment i techindustrin.

Arbetsflödet i ett projekt



- ▶ Statistikern befinner sig ofta endast i Steg 2 av processen.
- ▶ Avancerad modellering av data (mha SDA I och förhoppningsvis fler av våra kurser 😊).
- ▶ Statistikern bör vara inblandad i Steg 1. Eller åtminstone förstå hur data har samlats in.
- ▶ Personer inblandade i Steg 3 förlitar sig att resultaten besluten baseras på är korrekta.

Arbetsflödet i ett projekt, forts.



- ▶ **Garbage in = Garbage out** oavsett hur bra modellen anpassar data.
- ▶ Teknologiutvecklingen möjliggör data av hög kvalitet. Sensorer, mobiler, banktransaktioner, aktivitet på webben osv.
- ▶ Ibland behöver man samla in data. Statistik hjälper oss att dra slutsatser på **populationsnivå utifrån ett urval av populationen**.
- ▶ Ett urval av populationen kallar vi för ett **stickprov**.
- ▶ För att generalisera från ett stickprov till populationen (som Villani kommer lära er allt om) måste **stickprovet vara representativt för populationen**.

Slumpmässiga urval representerar populationen

- ▶ Hur fås ett representativt urval? **Simple random sampling** (SRS)¹.
- ▶ SRS: Varje stickprov av storleken n , från en population med N element, dvs

$$(y_1, y_2, \dots, y_n)$$

har samma sannolikhet att uppkomma.

- ▶ Hur stort stickprov (dvs n) ska vi samla?
- ▶ Ett slående resultat: Det är antalet n som spelar roll, inte hur stor andel av N vi väljer n som.
- ▶ Oavsett storleken på N , så är $n = 1000$ ofta tillräckligt.

¹Andra urvalsmetoder presenteras i SDA II och SDA III.

Slumpmässiga urval representerar populationen, forts.

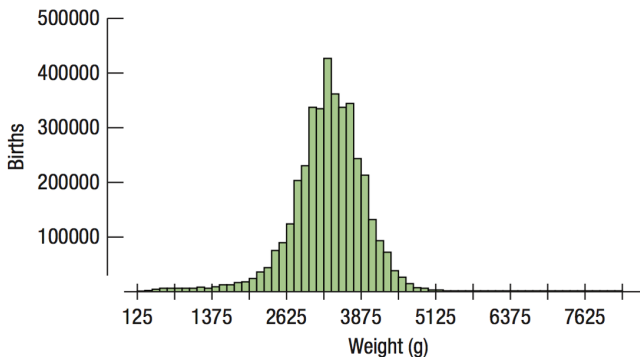


Figure 1: Figur 10.1 i De Veaux et al. (2021).

- Viktfördelning för populationen nyfödda år 1998 i USA. $N = 3940552$.

Slumpmässiga urval representerar populationen, forts.

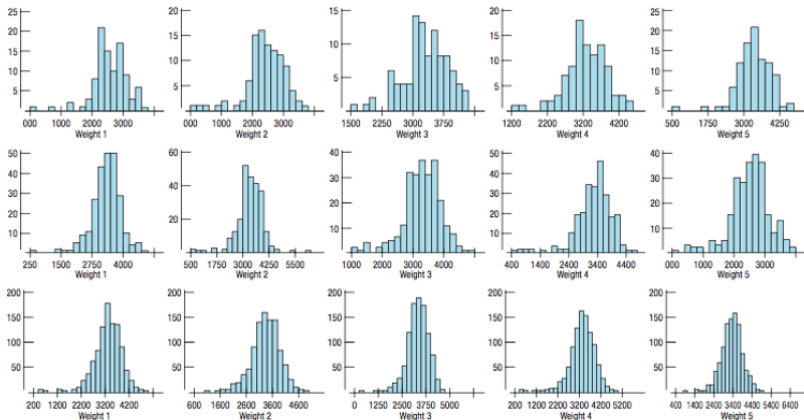


Figure 2: Figur 10.2 i De Veaux et al. (2021).

- SRS: Översta $n = 100$, mittersta $n = 250$, understa $n = 1000$.

Urval som inte representerar populationen

- ▶ **Voluntary response sample:** Individer bestämmer själva om de vill ingå.
- ▶ Exempel: Webbundersökningar. Självrekruterande paneler (YouGov).

YouGov

**Dela din åsikt.
Tjäna pengar.
Forma nyheterna.**

Gör som miljontals andra och dela din åsikt om politik,
sport, underhållning och mer

Start

Figure 3: Från YouGows hemsida.

- ▶ Deltagandet är inte slumpmässig, individer som känner starkare för en viss fråga mer benägna att delta. Stickprovet representerar inte populationen.

Bortfall räknas!

- ▶ SR Ekot rapporterade 2010 att varannan kvinnlig skådespelare har drabbats av sexuella trakasserier på arbetsplatsen baserat på ett urval med $n = 1672$.
- ▶ Det framkom senare att 4477 enkäter hade skickats ut.
- ▶ Om de 2805 **bortfallna svaren var slumpmässiga** hade resultaten kunnat generalisera till populationen.
- ▶ Det är rimligt att anta att:
 - ▶ De som har blivit utsatta är mer benägna att svara på enkäten.
 - ▶ De som inte har blivit utsatta är mindre benägna att svara på enkäten.
- ▶ **Stickprovet får en överrepresentation** av kvinnliga skådespelare som har drabbats av sexuella trakasserier på arbetsplatsen.
- ▶ Tyvärr handlade efterföljande diskussioner om SR Ekots inkompetens istället för ett verkligt problem...
- ▶ ... man hade kunnat uppmärksamma att sexuella trakasserier verkar vara ett större problem i teaterbranschen än andra branscher (se nästa slide).

Omstridd Ekot-undersökning – Sextrakasserier trots allt vanliga

PUBLICERAD 27 APRIL 2010

Trots vetenskapliga brister i Ekots undersökning kring sextrakasserier inom teatern är det ingen tvekan om att problemet är vanligare här än inom övriga branscher.

“Varannan kvinnlig skådespelare säger att hon blivit sextrakasserad på jobbet visar en ny undersökning. Det är Sveriges Radios ekoredaktion som har frågat medlemmar i Teaterförbundet”. Så inleddes Kulturnyheterna den 15 februari i år.

Det borde ha formulerats så här: “Varannan kvinnlig skådespelare som svarat på Ekots undersökning säger att hon blivit sextrakasserad på jobbet.”

Problemet med Ekots undersökning var att för få svarade. Av de 4477 scenkonstare, alla medlemmar i Teaterförbundet, som fick enkäten svarade bara 1672.

Därmed saknas statistisk representativitet för att kunna säga att varannan kvinnlig skådespelare i hela kåren blivit utsatt för trakasserier.

Men, faktum kvarstår, av de tillfrågade kvinnorna i Ekots undersökning svarade 320 personer att de utsatts för sexuella trakasserier. Det är 12,6 % av de tillfrågade.

Sextrakasserier är därmed vanligare inom teatern än i andra branscher. Enligt Arbetsmiljöverket har 2 % av hela arbetsmarknadens kvinnor utsatts för sextrakasserier.

Figure 4: Från SVT kulturs hemsida.

Vad används det slumpmässiga urvalet till?

- ▶ När vi har ett stickprov som är representativt för populationen använder vi det för att uppskatta populationsparametrar. **Statistisk inferens**.
- ▶ Populationen representeras av en teoretisk fördelning som har parametrar. Vi **skattar/estimerar** dessa utifrån stickprovet.
- ▶ Vi har sett åtskilliga exempel på detta (utan att fundera djupare på det):
 - ▶ \bar{y} och s är skattningar av μ och σ i populationen $N(\mu, \sigma)$.
 - ▶ När vi räknar en andel i en tabell, säg \hat{p} , är det en skattning av sannolikheten p i en såkallad binomial population $\text{Bin}(n, p)$ (Villani berättar mer).
 - ▶ b_0 och b_1 är skattningar av β_0 och β_1 i populationen $N(\beta_0 + \beta_1 x, \sigma)$.
- ▶ Nu kommer vi inte längre, vi behöver **sannolikhetsteori** för att kunna förstå det vi praktiskt har jobbat med under Del 1 av kursen.
- ▶ Innan jag lämnar över ska vi diskutera två olika typer av data:
 - ▶ Observationsstudier (**observational data**).
 - ▶ Experiment (**experimental data**).

Observationsstudier kontra experiment

- ▶ Statistiker är ofta intresserade av att mäta den **kausala effekten** av en behandling (**treatment** på engelska).
- ▶ Det klassiska (och tråkiga!) exemplet är att mäta effekten av ett läkemedel.
- ▶ En gammal föreställning är att statistiker spenderar mycket tid i laboratorier och utför experiment.
- ▶ Teknologiutvecklingen har gjort all interaktion mellan människa och maskin till ett laboratorium!
- ▶ Exempel: Vart ska man placera en knapp i en mobilapp för att maximera antalet tryckningar?
- ▶ Ett annat exempel: Vill kunder ha en ny tjänst? Experiment med mindre antal kunder innan man skalar upp tjänsten (om den är efterfrågad).

Observationsstudier kontra experiment, forts.

- ▶ Observationsstudier är data där statistikern inte har någon kontroll över vilka det är som får behandlingen.
- ▶ Exempel: En observationsstudie gjordes i USA där man jämförde studieresultaten för studenter i en musikklass mot studenter i en vanlig klass.
- ▶ Behandlingen (**treatment** på engelska) är “att spela ett instrument”.
- ▶ Behandlingsgruppen (**treatment group** på engelska) består av studenter i musikklassen.
- ▶ Kontrollgruppen (**control group** på engelska) består av studenter i den vanliga klassen.
- ▶ Data:
 - ▶ Kontrollgruppen: GPA 2.91. 5% fick betyget A.
 - ▶ Behandlingsgruppen: GPA 3.59. 16% fick betyget A.
- ▶ Skillnaderna är **statistiskt signifikanta**. Slutsats: Lär studenterna att spela ett instrument så får de högre betyg!

Observationsstudier kontra experiment, forts.

- ▶ Slutsatsen stämmer om det inte finns andra skillnader mellan grupperna som skulle kunna förklara skillnaden i betyg.
- ▶ Alternativa förklaringar till högre betyg:
 1. Musikstudenter kanske har bättre disciplin.
 2. Mer stöd från föräldrar.
 3. Musiktalang kan korrelera med annan typ av intelligens.
- ▶ Vi kan inte dra slutsatsen att det är behandlingen som orsakar skillnaden.
- ▶ Att fastställa kausalitet utifrån observationsstudier är svårt. Viktigt fält inom Ekonometri där observationsstudier ofta är det enda alternativet.
- ▶ I ett experiment så bestämmer statistikern vilka som tillhör kontrollgrupp samt behandlingsgrupp.
- ▶ Hur skulle experimentversionen av observationsstudien ovan se ut?

Observationsstudier kontra experiment, forts.

- ▶ **Randomisera vilka som ingår i respektive grupp** (behandlingsgrupp och kontrollgrupp)!
- ▶ Då skulle variablerna som vi inte mäter (och som kan förklara bra betyg) balanseras jämnt mellan grupperna.
- ▶ Den enda skillnaden mellan grupperna är då behandlingen.
- ▶ När vi mäter skillnaden mellan grupperna så är allt (förutom behandlingen) balanserat mellan grupperna.
- ▶ Får vi en signifikant skattning (skillnaden mellan grupperna beror inte på slumpen) så har vi en kausal effekt av behandlingen.
- ▶ Denna ansats kallas för **randomiserade experiment**.

Randomiserade experiment

- ▶ I många experiment finns det inte möjlighet att dra ett slumpmässigt urval från populationen.
- ▶ Ofta efterfrågar man försökspersoner genom till exempel annonser.
- ▶ Eller så är försökspersonerna givna utifrån förutsättningarna, till exempel de har en viss sjukdom man vill testa ett läkemedel för.
- ▶ Ett **slumpmässigt urval från populationen är nödvändigt för att generalisera till populationen**.
- ▶ **Randomiserade experiment används för att påvisa kausalitet**, dvs att behandlingen har en inverkan.
- ▶ Randomiserade experiment behöver replikeras flera gånger (genom olika experiment) innan slutsatser på populationsnivå kan dras.

Randomiserade experiment i techindustrin

- ▶ Exempel på statistiska experiment kommer (nästan) alltid från medicin. Kliniska studier (**clinical trials** på engelska).
- ▶ Försöksplanering (**experimental design** på engelska) kallas fältet i statistik som sysslar med experiment².
- ▶ Techindustrin har skapat nya laboratorium! Platser där användare interagerar med produkter är ett labb där experiment kan utföras.
- ▶ Exempel: Hemsidor, app-interface, en robots navigeringssystem.
- ▶ Det bästa av två världar:
 - ▶ Slumpmässiga urval.
 - ▶ Estimera kausala effekter mha randomiserade experiment.
- ▶ Inom machine learning kallas försöksplanering för **A/B testing**.

²Institutionens ger kursen Försöksplanering på masternivå.

Randomiserade experiment i techindustrin, forts.

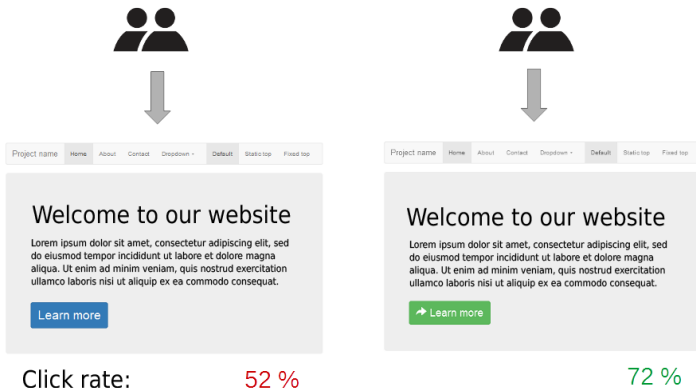


Figure 5: Källa: Wikipedia, figur av Maxime Lorant.

Randomiserade experiment i techindustrin, forts.

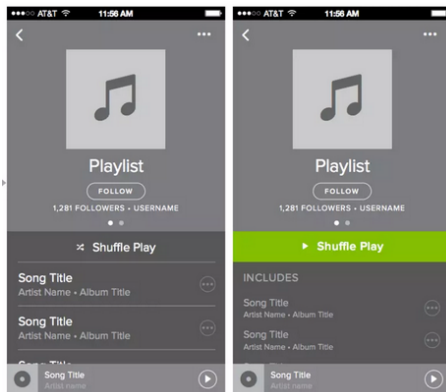


Figure 6: Resultat av ett randomiserat experiment på Spotify.

References I

De Veaux, R. D., Velleman, P., and Bock, D. (2021). *Stats: Data and Models*. Pearson, Harlow, United Kingdom, fifth edition.