

Föreläsning 10

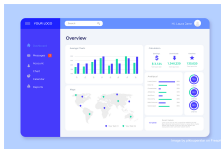
karl.sigfrid@stat.su.se

Vad har vi gjort hittills, och vad vi ska göra nu

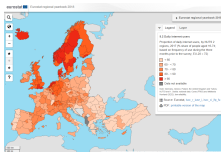
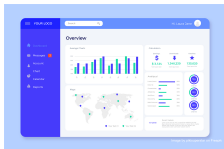
- ▶ Vi har gått igenom regressionsmodeller och metoder för att utvärdera en modell på **testdata**.
- ▶ Anledningen till att vi ofta utvärderar en modell på testdata är att vi vill veta om den förklarar **världen omkring oss**, dvs ett generellt samband, alltså inte bara förklarar observationerna i vårt dataset.
- ▶ Att dra mer generella slutsatser utifrån ett begränsat datamaterial kallas **inferens**.
- ▶ Inferens blir ett viktigt inslag i del 2 av den här kursen, men vi ska redan nu prata om vad som skiljer inferens från deskriptiv statistik.

Deskriptiv statistik och inferens

Deskriptiv statistik: Beskriv din data på ett meningsfullt sätt



Inferens: Dra slutsatser om världen utanför



Inferens

- ▶ Anta att vi fråga folk på stan om de tycker att regeringen gör ett bra jobb.
- ▶ Vi kan lätt räkna ut hur stor andel i **vårt dataset** som är av en viss uppfattning.

$$\text{Andel positiva} = \frac{\text{Antal positiva respondenter}}{\text{Totalt antal respondenter}}$$

- ▶ Men vi gör förmodligen inte undersökningen för att vi är intresserade åsikten hos just de personer som vi frågar.
- ▶ Vi är intresserade av **befolkningen** tycker.
- ▶ Åsikterna hos de personer som svara på enkäten är intresseranta för att vi tror att de **är representativa för befolkningen**.

- ▶ I statistik kallar vi den grupp vi är intresserade av att analysera för en **population**.

Exempel på populationer

- ▶ Om vi vill ta reda på hur stor andel av befolkningen som har en viss åsikt är vår population befolkningen.
- ▶ Om vi vill veta hur många poäng en genomsnittlig student på Stockholms Universitet har är populationen alla studenter vid Stockholms Universitet.

Det är viktigt att specificera sin population.

- ▶ När vi pratar om befolkningen i ett land, inkluderar vi enbart medborgare eller alla som är bostatta i landet? Inkluderar vi enbart myndiga eller även personer under 18 år?
- ▶ När vi pratar om alla studenter på Stockholms Universitet, syftar vi enbart på heltidsstudenter, eller räknar vi också in någon som tar en enstaka distanskurs på deltid?
- ▶ Det finns **inget rätt svar** på frågorna, men vi måste fatta ett beslut om vilka som ingår i populationen.

Inferens

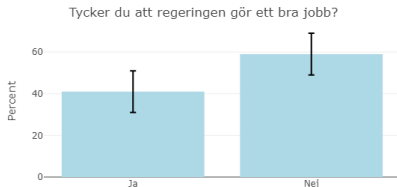
- ▶ Vi har på den här kursen pratat om dataset med observationer.
- ▶ Det dataset vi vanligen har tillgång till när vi gör inferens är ett **stickprov (sample)**, som är en delmängd av vår population.
- ▶ Med hjälp av vårt stickprov drar vi slutsatser om hela den population som intresserar oss.

Exempel på stickprov

- ▶ Om vår population är hela befolkningen i ett land kan vårt stickprov bestå av 1000 slumpvis utvalda invånare i landet.
- ▶ Om vi vill veta hur inkomsterna i Sverige fördelar sig kan stickprovet vara 1000 inkomstuppgifter som vi begär ut från Skatteverket.

Inferens

- ▶ När vi talar om ett medelvärde i vår data kan vi ange **ett exakt värde**. Det kan vara 410 av 1000 deltagare i en undersökning som tycker att en regering gör ett bra jobb. Då är andelen i vår undersökning exakt 41 procent.
- ▶ När vi uttalar oss om hela befolkningen (dvs vår population) kan vi inte vara lika exakta, eftersom vi inte har mätt åsikten hos hela befolkningen.
- ▶ Det vi kan säga är att andelen av befolkningen med viss säkerhet, eller med en viss **konfidens**, ligger inom ett visst intervall.



Inferens

- ▶ Att mäta en egenskap hos en population är **svårt**.
- ▶ Den här tabellen visar förtoendet för Joe Biden i slutet av Juli 2023. Tabellen kommer från <https://edition.cnn.com/polling/approval-rating-poll-of-polls>.
- ▶ Mätningarna är gjorda i princip samtidigt. Ändå kommer de till olika resultat. Skillnaden mellan 44% och 38% är stor! Varför så olika resultat?

The 07/27/23 Poll of Polls is an average of the following polls:

Poll	Dates	Approve	Disapprove	Sample Size
Marquette Law School	July 7-12	42%	57%	1,005
CNBC	July 12-16	39%	55%	1,000
Quinnipiac University	July 13-17	38%	54%	2,056
Monmouth University	July 12-17	44%	52%	910
Reuters/Ipsos	July 7-9	40%	54%	1,028

Inferens - två felkällor

- ▶ Det finns två huvudsakliga felkällor när vi drar slutsatser om en population med hjälp av ett stickprov:
 - ▶ **Bias (Systematiska fel)**
 - ▶ **Slumpmässiga variationer**
- ▶ Ett stickprov med bias är insamlat på ett sätt som **snedvrider resultatet på ett systematisk sätt.**
- ▶ De slumpmässiga variationerna **är oundvikliga** när vi tar ett stickprov.
- ▶ När vi ser exempelvis en opinionsundersökning, oavsett hur välgjord den är, kan vi utgå från att den i någon mån är påverkad både av bias och av slumpmässiga variationer.

Inferens - bias

Ofta är den som samlar in data omedveten om att stickprovet har ett bias.

Exempel på bias

D Veaux et al (2021), sid 352, berättar om en tidning som försökte mäta sympatierna inför det amerikanska presidentvalet 1936. De valde slumpvis personer ur telefonkatalogen. Dock var det många med sämre ekonomi som saknade telefon denna tid, och den gruppen hamnade utanför stickprovet. Mätningens resultat blev därför helt fel.

Inferens - bias

Exempel på bias

Anta att en statistikstudent vill undersöka hur många timmar i veckan en heltidsstudent på Stockholms Universitet lägger på sina studier.

Studenten gör en enkät som 50 andra studenter på samma kurs besvarar.

Problemet med stickprovet är att statistikstudenter kanske pluggar mer, eller mindre, än studenter som läser andra ämnen. Undersökningen kan fortfarande vara värdefull om vi använder den som ett mått på hur mycket **statistikstudenter på SU** (inte studenter i allmänhet på SU) pluggar.

Inferens - bias

- ▶ Ett sätt att försöka undvika bias är att **göra ett slumpurval** inom populationen.
- ▶ Om vi gör ett slumpurval från populationen kan vi förvänta oss att stickprovet är någorlunda representativt.
- ▶ Men ett slumpurval är **ingen garanti** för att stickprovet är representativt. (De som gjorde den tidigare nämnda undersökningen inför presidentvalet 1936 **trodde** förmodligen att de hade ett representativt urval när de slumpade fram personer ur telefonkatalogen.)

Inferens - Slumpmässiga variationer

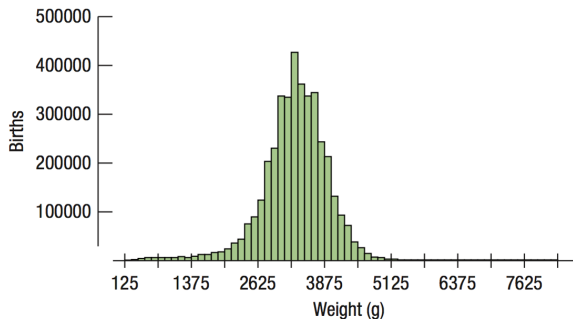
- ▶ Även stickprov som är helt utan bias kommer att påverkas av slumpen.
- ▶ Om vi tar flera stickprov kommer resultatet att variera mellan stickproven.

Exempel på slumpmässiga variationer

- ▶ Ett opinionsinstitut mäter andelen av befolkningen som sympatiserar med olika politiska partier. De gör två undersökningar med en månads mellanrum. Även om opinionen ser exakt likadan ut vi båda tidpunkterna så kommer resultatet förmodligen att skilja sig åt, åtminstone en aning. Det beror på att de slumpvisa personer som ingår i undersökningen har bytts ut.
- ▶ Två olika forskare försöker var för sig estimerar genomsnittsvikten för en viss fiskart. Resultaten kommer troligen att skilja sig åt eftersom urvalet av fiskar som de hittar är olika.

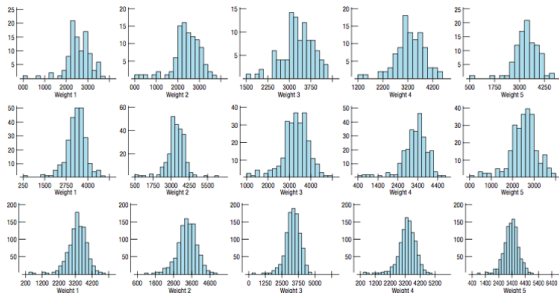
Inferens - Slumpvisa variationer

- ▶ Figur 10.1 i Deveaux et al (2021) visar viktfordelningen för nyfödda år 1998 i USA.
- ▶ $N=3,940,552$



Inferens - Slumpvisa variationer

- ▶ Här ser vi ett antal stickprov från fördelningen på förra bilden.
- ▶ Överst: $n = 100$, Mitten: $n = 250$, Nederst: $n = 1000$



Inferens - Stickprovets storlek

- ▶ En opinionsmätning har ofta ungefär 1000 respondenter, men det kan vara allt från några hundra till flera tusen.
- ▶ Hur stort ska stickprovet vara för att vara representativt? Det beror på hur stora **felmarginaler** vi accepterar.
- ▶ Ju större stickprov desto mindre felmarginaler.
- ▶ **Viktigt:** Populationens storlek avgör inte hur stort stickprovet måste vara! Att USA är ett mycket större land än Island betyder inte att stickproven måste vara större för att mäta den amerikanska väljaropinionen än för att mäta den isländska.

Inferens - Simple random sampling

- ▶ Ett vanlig metod för att välja ut vilka observationer som ska ingå i ett stickprov är **Simple random sampling**
- ▶ Simple random sampling betyder att **varje möjligt stickprov** har samma sannolikhet att bli valt.
- ▶ Att varje observation har samma sannolikhet att inkluderas i stickprovet är **inte tillräckligt** för att det ska vara ett Simple random sample.

Inferens - Simple random sampling

Exempel på Simple random sampling

- ▶ Vi har 5 observationer i en liten population: $\{x_1, x_2, x_3, x_4, x_5\}$. Vi ska ta ett stickprov som inkluderar 2 av dessa observationer.
- ▶ Följande stickprov är möjliga, och ska ha samma sannolikhet att bli valda:

$\{x_1, x_2\}, \{x_1, x_3\}, \{x_1, x_4\}, \{x_1, x_5\}, \{x_2, x_3\}, \{x_2, x_4\}, \{x_2, x_5\}, \{x_3, x_4\}, \{x_3, x_5\}, \{x_4, x_5\}$

Inferens - parametrar och statistika

- ▶ **Populationsparametrar** är nyckeltal som säger något om en *population*.
- ▶ Ofta kan vi inte **mäta eller observera** värdet på en parameter. Däremot kan vi estimerat värdet.

Exempel på parametrar

Vi har redan talat om populationsparametrar, men utan att kalla dem så:

- ▶ Om vi betraktar Sveriges befolkning som en population är **andelen som tycker att regeringen gör ett bra jobb** en populationsparameter.
- ▶ Om vi betraktar alla älgar som en population är **den genomsnittliga vikten hos en älg** en populationsparameter, liksom **standardavvikelsen** för älgarnas viktfordelning.

Inferens - parametrar och statistika

- ▶ När vi estimerar en populationsparameter gör vi det med hjälp av en **statistika (statistic)**. En statistika kan vi observera i vår data.

Exempel på statistika

- ▶ **Proportionen av respondenterna** till en enkät som tycker att regeringen gör ett bra jobb är en statistika. Denna statistika kan användas som ett estimat för hur stor del av *befolkningen* som tycker att regeringen gör ett bra jobb.
- ▶ **Medelvikten** hos ett antal älgar som vägs kan användas som ett estimat för medelvikten för alla älgar.

Inferens - parametrar och statistika

- ▶ Ofta använder vi **grekiska bokstäver** för att symbolisera populationsparametrar.
- ▶ Tabellen nedan, ur De Veaux et al (2021), listar uttrycken för ett antal vanliga statistika och deras korresponderande populationsparametrar.

Name	Statistic	Parameter
Mean	\bar{y}	μ (mu, pronounced “meeoo,” not “moo”)
Standard Deviation	s	σ (sigma)
Correlation	r	ρ (rho, pronounced like “row”)
Regression Coefficient	b	β (beta, pronounced “baytah” ⁵)
Proportion	\hat{p}	p (pronounced “pee” ⁶)

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.
- ▶ En konsekvens av studien blev krav på att fler elever borde få musiklektioner.

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.
- ▶ En konsekvens av studien blev krav på att fler elever borde få musiklektioner.
- ▶ Utifrån det vi har lärt oss hittills på den här kursen, kan vi utgå från att fler elever kommer att lyckas bättre om fler får lära sig att spela ett instrument?

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.
- ▶ En konsekvens av studien blev krav på att fler elever borde få musiklektioner.
- ▶ Utifrån det vi har lärt oss hittills på den här kursen, kan vi utgå från att fler elever kommer att lyckas bättre om fler får lära sig att spela ett instrument?
- ▶ **Nej**, ett observerat samband betyder inte att vi har hittat kausalitet.

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.
- ▶ En konsekvens av studien blev krav på att fler elever borde få musiklktioner.
- ▶ Utifrån det vi har lärt oss hittills på den här kursen, kan vi utgå från att fler elever kommer att lyckas bättre om fler får lära sig att spela ett instrument?
- ▶ **Nej**, ett observerat samband betyder inte att vi har hittat kausalitet.
- ▶ Det kan finnas andra faktorer som gör att vissa elever **både** är mer benägna att spela ett instrument och mer benägna att få höga betyg.

Observationsstudier och experiment

- ▶ I en amerikansk studie fann forskarna att studenter som spelade musikinstrument hade bättre betyg än andra studenter.
- ▶ En konsekvens av studien blev krav på att fler elever borde få musiklektioner.
- ▶ Utifrån det vi har lärt oss hittills på den här kursen, kan vi utgå från att fler elever kommer att lyckas bättre om fler får lära sig att spela ett instrument?
- ▶ **Nej**, ett observerat samband betyder inte att vi har hittat kausalitet.
- ▶ Det kan finnas andra faktorer som gör att vissa elever **både** är mer benägna att spela ett instrument och mer benägna att få höga betyg.
- ▶ Kanske är elever som har det lättare i skolan mer benägna att vilja lära sig ett instrument. Kanske har elever som spelar instrument mer engagerade föräldrar.

Observationsstudier och experiment

- ▶ Studien som fann att elever som spelade instrument hade bättre betyg var en **observationsstudie**.
- ▶ I en observationsstudie jämför forskare olika grupper, men utan att själva styra vem som ska ingå i de grupper som jämförs.
- ▶ Det kan alltså finnas en lång rad olika faktorer som skiljer grupperna åt.

Men om vi nu verkligen vill veta om musiklektioner bidrar till bättre betyg, då kan vi inte förlita oss på en observationsstudie, utan vi måste anordna ett **experiment**.

Observationsstudier och experiment

- ▶ I ett **experiment** vill vi se hur en viss **faktor** påverkar en **responsvariabel**.
- ▶ Eftersom vi vill veta om just denna faktor är avgörande försöker vi skapa olika grupper **som är lika i övrigt**.
- ▶ Om grupperna verkligen är lika i övrigt kan vi dra slutsatsen att skillnader mellan grupperna faktiskt beror på den faktor vi studerar.

Observationsstudier och experiment

- ▶ Anta att vi vill göra ett experiment för att se om sömnbrist påverkar läsförståelsen.

Observationsstudier och experiment

- ▶ Anta att vi vill göra ett experiment för att se om sömnbrist påverkar läsförståelsen.
- ▶ Vi vill testa om individer som sovit fyra timmar senaste dygnet presterar sämre på ett prov som testar läsförståelse än vad individer som sovit åtta timmar gör.

Observationsstudier och experiment

- ▶ Anta att vi vill göra ett experiment för att se om sömnbrist påverkar läsförståelsen.
- ▶ Vi vill testa om individer som sovit fyra timmar senaste dygnet presterar sämre på ett prov som testar läsförståelse än vad individer som sovit åtta timmar gör.
- ▶ **Kontrollfråga:** Vilken är vår faktor och vilken är vår responsvariabel?

Observationsstudier och experiment

- ▶ Anta att vi vill göra ett experiment för att se om sömnbrist påverkar läsförståelsen.
- ▶ Vi vill testa om individer som sovit fyra timmar senaste dygnet presterar sämre på ett prov som testar läsförståelse än vad individer som sovit åtta timmar gör.
- ▶ **Kontrollfråga:** Vilken är vår faktor och vilken är vår responsvariabel?
- ▶ **Faktor:** antal timmars sömn senaste dygnet.

Observationsstudier och experiment

- ▶ Anta att vi vill göra ett experiment för att se om sömnbrist påverkar läsförståelsen.
- ▶ Vi vill testa om individer som sovit fyra timmar senaste dygnet presterar sämre på ett prov som testar läsförståelse än vad individer som sovit åtta timmar gör.
- ▶ **Kontrollfråga:** Vilken är vår faktor och vilken är vår responsvariabel?
- ▶ **Faktor:** antal timmars sömn senaste dygnet.
- ▶ **Responsvariabel:** Testresultat på test i läsförståelse.

Observationsstudier och experiment

- ▶ I ett experiment bör deltagarna fördelas **slumpvis** mellan grupperna.
- ▶ Om vi låter deltagarna i sömnstudien själva välja grupp, då skulle det kunna bli så att deltagare med vissa egenskaper väljer gruppen som sover fyra timmar och deltagare med andra egenskaper som väljer att sova åtta timmar. Då har vi inte längre två likadana grupper.
- ▶ I verkligheten är det inte alltid möjligt att genomföra experiment. Elever och föräldrar skulle knappast acceptera att vissa elever tvingas spela ett instrument och att andra elever förbjuds att spela ett instrument.
- ▶ Därför är observationsstudier i vissa sammanhang det enda alternativet. Då är det viktigt att vara försiktig med vilka slutsatser vi kan dra.