Statistik och dataanalys I

F10: Stickprov och populationer, observationsstudier och experiment

Valentin Zulj

Statistiska institutionen

Stockholm University

Vad har vi gjort hittills?

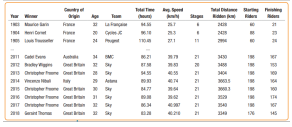
- Vi har gått igenom regressionsmodeller, och metoder för att utvärdera en modell på **testdata**
- Anledningen till att vi vill utvärdera modeller på testdata är att vi vill undersöka om den förklarar **världen omkring oss**
- Mer konkret vill vi veta om modellen förklarar ett mer generellt samband mellan variablerna, eller om den bara råkat hitta ett samband i träningsdata
- När vi utgår ifrån ett begränsat datamaterial, och extrapolerar till en mer generell population, utövar vi **statistisk inferens**
- Inferens blir ett viktigt inslag i del 2 av den här kursen, och vi ska börja lägga grunden för det redan idag

Vad ska vi göra nu?

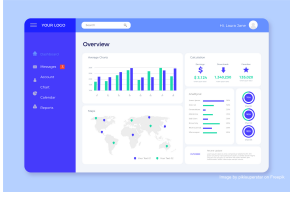
- Idag ska vi
  - Prata om vad som skiljer deskriptiv statistik (som vi har sysslat med hittils) från inferens
  - Introducera begrepp som populationer och populationsparametrar
  - Prata om slumpmässiga urval, bias, och statistikor
  - Diskutera skillnader mellan experimentella studier och observationsstudier

Deskriptiv statistik och inferens

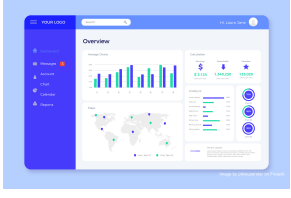
Deskriptiv statistik: Beskriva våra data på ett meningsfullt sätt

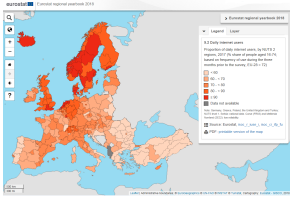


Year	Winner	Country of Origin	Age	Team	Total Time (hours)	Avg. Speed (km/h)	Stages	Total Distance Ridden (km)	Starting Riders	Finishing Riders
1903	Maurice Garin	France	32	Le Française	94.55	25.7	6	2428	60	21
1904	Henri Cornet	France	20	Cycles JC	96.16	25.3	6	2428	88	23
1905	Louis Trousselier	France	24	Peugeot	110.45	27.1	11	2994	60	24
...										
2011	Cadel Evans	Australia	34	BMC	86.21	39.79	21	3430	198	167
2012	Bradley Wiggins	Great Britain	32	Sky	87.58	39.83	20	3488	198	153
2013	Christopher Froome	Great Britain	28	Sky	84.55	40.55	21	3404	198	169
2014	Vincenzo Nibali	Italy	29	Astana	89.83	40.74	21	3663.5	198	164
2015	Christopher Froome	Great Britain	30	Sky	84.77	39.64	21	3660.3	198	160
2016	Christopher Froome	Great Britain	31	Sky	89.08	39.62	21	3519	198	174
2017	Christopher Froome	Great Britain	32	Sky	86.34	40.997	21	3540	198	167
2018	Geraint Thomas	Great Britain	32	Sky	83.28	40.218	21	3349	176	145



Inferens: Använda våra data för att dra slutsatser om världen utanför





- Vi går ut och frågar folk på stan om de tycker att regeringen gör ett bra jobb
- Med hjälp av svaren vi får kan vi lätt beräkna hur stor andel i **våra data** som tycker si eller så,

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

- Troligtvis är vi dock inte så intresserade av *enbart* personerna vi frågar, utan vi vill veta något om vad hela **befolkningen** tycker
- Åsikterna hos personerna vi frågar är *bara intressanta om vi tror att de är representativa för befolkningen*

- I statistik kallar vi den grupp som vi vill dra slutsatser om för en **population**

- **Exempel på populationer**
  - Om vi vill ta reda på hur stor andel av befolkningen som tycker att regeringen gör ett bra jobb, så är Sveriges befolkning vår population
  - Om vi vill veta hur många högskolepoäng en genomsnittlig student på SU har, är populationen alla studenter vid SU
  - Om vi vill veta genomsnittslönen för en nyutexaminerad statistiker, är populationen alla nyutexaminerade statistiker

Populationer

- När vi ska göra en studie eller analys är det viktigt att ställa frågan ***”vilken population är vi itnresserade av att dra slutsatser om?”***
- Svaret behöver vara så precist formulerat som möjligt, för att undvika fallgropar och snedvridande effekter

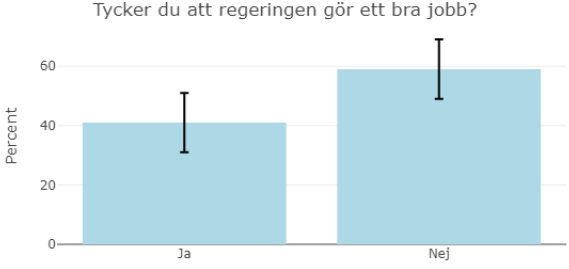
  - När vi pratar om Sveriges befolkning: menar vi bara medborgare, eller alla som bor här? Inkluderar vi bara myndiga, eller även personer under 18 år?
  - När vi pratar om alla studenter på SU: syftar vi enbart på heltidsstudenter, eller räknar vi också in någon som tar en enstaka distanskurs på deltid?

- Vår studie/analys kan behöva utföras på olika sätt, och ge olika svar/resultat, beroende på hur vi definierar populationen
- Det finns **inget “rätt svar”** på frågorna, men vi måste fatta ett beslut om vilka som ingår i populationen för att kunna genomföra studien på ett korrekt sätt

- Vi har på den här kursen pratat om dataset med observationer
- Vanligtvis är de data vi använder när vi gör inferens insamlade genom **stickprov** (*sample*), som är en delmängd av någon given population
- Med hjälp av ett stickprov vill vi dra slutsatser om hela den population som intresserar oss

- **Exempel på stickprov:**
  - Om vår population är hela Sveriges befolkning, kan vårt stickprov bestå av 1 000 personer som slumpvis valts ut ur befolkningsregistret
  - Om vi vill veta hur inkomster fördelar sig i Sverige kan vårt stickprov bestå av 1 000 inkomstuppgifter som vi begär ut från Skatteverket
  - Om vi vill veta hur bilars vikt påverkar deras bränsleförbrukning, kan vårt stickprov bestå av 32 bilar som någon har testat åt oss

- När vi talar om ett medelvärde *i våra data/vårt stickprov* kan vi ange ett exakt värde
- Om 410 av 1000 tillfrågade tycker att en regering gör ett bra jobb, så är andelen *i våra data* 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
- Efter Tobias delkurs kommer vi kunna säga att populationsandelen med viss säkerhet/**konfidens** ligger inom ett givet intervall (se pinnar i toppen av staplarna)



- Att mäta en egenskap hos en population är **svårt**
- Tabellen (från CNN) visar resultat av olika mätningar av förtoendet för Joe Biden i slutet av Juli 2023
- Mätningarna är gjorda i princip samtidigt, men ger ändå 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

Två felkällor

- Det finns två huvudsakliga felkällor när vi drar slutsatser om en population med hjälp av ett stickprov:
  - **Bias** (även känt som *systematiska fel*)
  - **Slumpmässiga variationer** (personerna i stockprovet väljs oftast slumpmässigt)
- Ett stickprov med bias är insamlat på ett sätt som **systematisk snedvrider resultatet** – någon som kan komma på ett exempel?
- De slumpmässiga variationerna **är oundvikliga** när vi drar ett stickprov
- När vi ser t.ex. en opinionsundersökning, kan vi utgå ifrån att den i någon mån påverkas av både bias och slumpmässiga variationer
- Detta gäller oavsett hur välgjord undersökningen är, och är troligtvis inte avsiktligt

Bias – Exempel 1

- Detta exempel visar på att den som samlar in data omedvetet kan föra in bias i stickprovet
- På sid 352 i De Veaux et al (2021) ges följande exempel
  - En tidning ville mäta sympatier inför presidentvalet i USA 1936
  - Använde ett stickprov med personer som dragits slumpmässigt ur telefonkatalogen, i ett ärligt försök att ge hela populationen möjlighet att komma med i stickprovet
  - Vid denna tid saknade många ekonomiska förutsättningar att ha en telefon och fanns därför inte med i telefonkatalogen – de kom så inte med i stickprovet
  - Är det rimligt att anta att dessa personer till stor del sympatiserade med samma kandidat? I så fall blev mätningens resultat snedvridet

Bias – Exempel 2

- En statistikstudent vill undersöka hur många timmar i veckan en heltidsstudent på SU lägger på sina studier
- Som underlag samlar hen in data med hjälp av en enkät, som besvaras av 50 andra studenter på samma kurs
- Kan någon tänka sig en källa till bias i denna undersökning?
- Kan det vara så att statistikstudenter pluggar mer (eller mindre) än studenter som läser andra ämnen?
- I så fall är stickprovet inte representativt för alla studenter på SU
- Undersökningen kan dock fortfarande vara värdefull, om vi använder den som ett mått på hur mycket *statistikstudenter* på SU pluggar

Bias och slumpurval

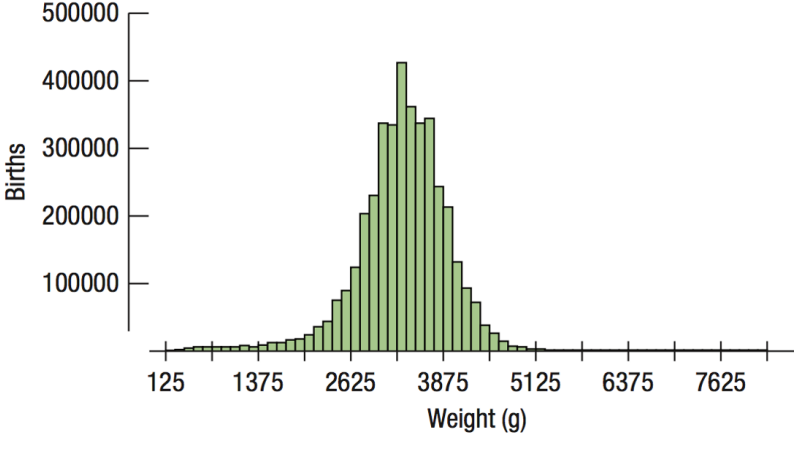
- Ett sätt att försöka undvika bias är att göra **ett slumpurval** från vår målpopulation
- Om vi gör ett slumpurval från populationen kan vi förvänta oss att stickprovet är någorlunda representativt
- Men ett slumpurval är **ingen garanti** för att stickprovet är representativt
- De som slumpmässigt valde personer ur telefonkatalogen trodde förmodligen att deras stickprov skulle bli representativt
- Anledningen till icke-representativitet var att vissa individer i målpopulationen hade 0% chans att komma med i stickprovet – och detta är en situation vi gärna vill undvika

Slumpmässiga variationer

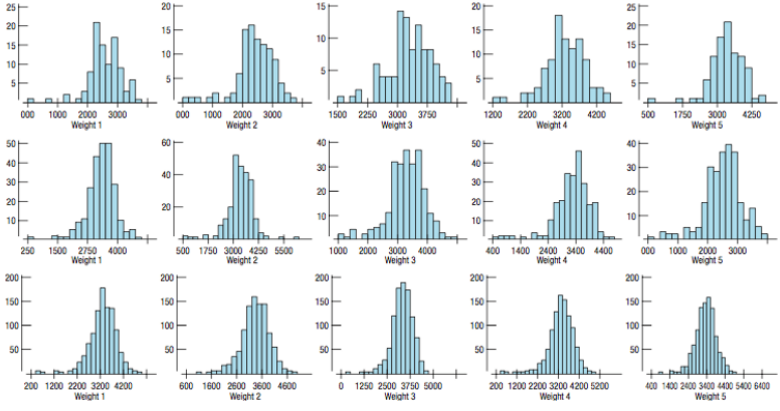
- Även stickprov som är helt utan bias kommer att påverkas av slumpen
- Om vi tar flera olika stickprov kommer resultaten variera mellan stickprov

- **Exempel på slumpmässiga variationer:**
  - Två opinionsinstitut mäter andelen av befolkningen som sympatiserar med olika partier, två separata undersökningar under samma tidsperiod
  - Resultaten kommer skilja sig mellan undersökningarna, ibland mer ibland mindre
  - Detta beror på att personerna som ingår i undersökningen är slumpmässigt utvalda, och byts ut mellan undersökningarna

- Figur 10.1 i Deveaux et al (2021) visar viktfördelningen för **alla** nyfödda år 1998 i USA, där vi har $N = 3\ 940\ 552$



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



Stickprovets storlek

- Opinionsmätningar har ofta ungefär 1000 deltagare, men det kan variera från några hundra till flera tusen
- **Men:** hur stort måste stickprovet vara för att vara representativt?
- Det beror helt på vad det är vi vill undersöka, och hur vi drar stickprovet
- Drar vi stickprovet helt slumpmässigt, eller ser vi till att dela upp vår dragning så att vi är säkra på att få med individer från relevanta grupper, etc?
  - Ett jättestort stickprov kan i princip vara värdelöst på grund av systematiska fel i insamlingen
  - Ett jämförelsevis pyttelitet stickprov kan dock vara extremt informativt, om det är väl genomtänkt och insamlat
- Detta är en hel vetenskap i sig, men lite överkurs just nu
- **Viktigt:** Populationens storlek avgör inte hur stort stickprovet måste vara!
  - USA:s befolkning är mycket större än Islands, men vi behöver inte nödvändigtvis dra större stickprov när vi mäter opinion i USA
  - Den **absoluta stickprovsstorleken** spelar roll, men inte den **relativa stickprovsstorleken**

Simple random sampling (obundet slumpmässig urval)

Stockholm University

- Ett vanlig metod för att välja ut vilka observationer som ska ingå i ett stickprov är **simple random sampling (SRS)**
- SRS betyder att **varje möjligt stickprov** har samma sannolikhet att bli valt
- Det är **inte tillräckligt** att alla individer har samma sannolikhet för inklusion, utan det måste gälla **alla möjliga stickprov**

Simple random sampling – Exempel

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

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

- Under kriteriet att alla individer ska ha samma sannlokhet skulle det räcka att bara välja bland dessa stickprov

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

- Men här finns förhöjd risk för bias, eftersom vissa kombinationer inte går att dra – jämför med telefonkatalogen

- **Populationsparametrar** är nyckeltal som beskriver någon aspekt av en population
- Vikan sällan **mäta eller observera** värdet på en parameter, utan vår uppgift som är vanligtvis att skatta värdet på den

- **Exempel på parametrar**
  - Om vi betraktar Sveriges befolkning som en population, så är *andelen som tycker att regeringen gör ett bra jobb* en populationsparameter
  - Om vi betraktar alla älgar i Sverige som en population, så är *den genomsnittliga vikten hos en älg* en populationsparameter
  - På samma sätt är *standardavvikelsen* för alla älgarnas viktfördelning en populationsparameter

- När vi skattar en populationsparameter gör vi det med hjälp av en **statistika** (*statistic*)
- En statistika är något som vi kan beräkna med hjälp av våra data

- **Exempel på statistikor:**
  - *Andelen i våra data* som tycker att regeringen gör ett bra jobb är en statistika
  - Denna statistika kan användas som en skattning av hur stor del av *befolkningen* som tycker att regeringen gör ett bra jobb
  - *Medelvikten i ett ett stickprov av älgar* är en annan statistika
  - Denna statistika kan användas för att uppskatta medelvikten för alla älgar

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

Observationsstuder och experiment

Motiverande exempel

- I en amerikansk studie fann forskare att skolelever som spelade musikinstrument hade bättre betyg än andra elever
- Efter studien kom krav på att fler elever borde få musiklektioner
- Utifrån det vi har lärt oss hittills på den här kursen, kan vi utgå ifrån att elever kommer att lyckas bättre i skolan om de får lära sig att spela ett instrument?
- **Nej!** Ett observerat samband är inte nödvädigtvis kausalt!
- Det kan finnas andra faktorer som gör att vissa elever är **både** mer benägna att spela ett instrument, **och** mer benägna att få höga betyg
- Kan elever som har det lättare i skolan ha mer energi över till att lära sig spela musik?
- Kan elever som spelar musik ha mer engagerade föräldrar, som också stöttar inlärning?

Observationsstudier

- Studien som beskrevs i föregående slide är ett exempel på en **observationsstudie**
- I en observationsstudie jämför forskare olika grupper, utan att själva kunna styra vilka som ska ingå i de grupper som jämförs
- Vi kan tänka oss att barn som väljer att lära sig spela instrument kan ha annorlunda förutsättningar än barn som inte vill spela musik, och att detta skiner igenom i resultaten
- Det kan alltså finnas en lång rad olika faktorer som skiljer grupperna åt, och vi **måste** ta hänsyn till dessa om vi ska få en rättvisande jämförelse
- Annars kan effekten av att spela musik delvis blandas ihop med t.ex. effekten av att ha föräldrar som kan engagera sig i sina barns fritid

Experiment

- Antag nu att vi verkligen vill veta om musiklektioner bidrar till bättre betyg
- Det kan vara svårt att skatta en sådan effekt i en observationsstudie, men det kan vara något enklare att anordna ett **experiment**
- Ett **experiment** går ut på att mäta hur en viss **faktor** påverkar en **responsvariabel**
- För att isolera just den faktor vi är intressrade av försöker vi skapa jämförelsegrupper som är **så lika som möjligt i övrigt**
- Om grupperna verkligen är lika i övrigt kan vi dra slutsatsen att skillnader mellan grupperna faktiskt beror på den faktor vi studerar
- Det finns i så fall inga bakomliggande faktorer som förvränger våra resultat

Skillnad mellan observationsstuder och experiment

- Notera skillnaden mellan de två exempel vi tagit upp
- I det första exemplet använde forskarna jämförelsegrupper som de inte kunde kontrollera, utan barnen (ev. deras föräldrar) valde själva om de ville spela instrument
- Forskarna kunde inte veta om det fanns bakomliggande faktorer, som påverkar både barnens betyg och deas benägenhet att spela instrument
- Detta är typiskt för en observationsstudie
- I ett experiment har forskarna själva kontroll över hur grupperna skapas
- De kan då se till att den enda skillnaden (i genomsnitt) mellan grupperna är att det spelas instrument i den ena, men inte i den andra
- Det blir så mycket lättare att isolera effekten av att spela musik

- Anta att vi vill göra ett experiment för att se om sömnbrist hos skolelever påverkar deras läsförståelse
- Vi vill testa om elever som sovit fyra timmar det senaste dygnet presterar sämre på ett läsförstående-test än vad elever 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
- Vad skulle kunna hända om vi lät deltagarna välja "sömngrupp" själva?
- Deltagare med vissa egenskaper/personlighetsdrag skulle kunna välja att sova mindre, och deltagare med andra egenskaper välja att sova mer
- Grupperna skulle inte bli likdana i alla andra aspekter utöver just sömn
- I verkligheten går det inte alltid att genomföra experiment
- Varken elever eller föräldrar skulle nog acceptera att vissa elever tvingas spela instrument, samtidigt som andra totalförbjuds från det
- Därför är observationsstudier i vissa sammanhang det enda alternativet
- I sådana fall går det att använda mer avancerade statistiska metoder som är överkurs här (den intresserade kan läsa mer om *kausal inferens*)
- I allmänhet **måste** vi vara försiktiga med vilka slutsatser som faktiskt går att dra när vi arbetar med observationsstudier

Tack för idag!!