2.1 Hvad er registerdata?

I Danmark findes en række officielle databaser, hvor du kan hente kvantitative data. For eksempel Danmarks Statistik (Statistikbanken), Ny i Danmark og Undervisningsministeriet. I disse databaser er de indsamlede data såkaldte **registerdata**. Det vil sige, at data er resultatet af **totaltællinger** og er således ikke behæftet med den statistiske usikkerhed, som der er i forbindelse med stikprøver. Dog skal man ikke være blind for såkaldte mørketal. Sort arbejde indgår ikke i indkomststatistikken, illegale indvandrere indgår ikke i folketællinger osv.

Registerdata: Data indsamlet af myndighed eller offentligt kontor med administrative formål og vil derfor have en høj grad af troværdighed.

Totaltællinger: Statistik hvor alle data indgår. Fx er befolkningsstatistikken i Danmark baseret på en totaltælling.

Mange af disse databaser er interaktive. Det vil sige, at du som bruger selv kan udvælge, hvilke data du vil have fingre i. Endvidere er det som oftest muligt at hente data ned i et format – typisk regneark – således de kan viderebearbejdes og formidles.

Det er ikke muligt her at introducere eller nævne alle de relevante databaser, der findes. På bogens hjemmeside kan du finde en emneopdelt liste over de mest relevante databaser i forbindelse med de samfundsfaglige discipliner.

De vigtigste er listet op i skemaet nedenfor.

Database	Udgiver	Områder, der dækkes
Statistikbanken	Danmarks Statistik	Stort set alle områder: økonomi, sociale forhold, kriminalitet, indkomst, forbrug.
Eurostat	EU	Dækker statistik for de 27 medlemslande inden for stort set alle områder.
US Census Bureau	USA's regering	Indkomst, uddannelse, fattigdom, helbred for USA.
World Bank	Verdensbanken	Statistik for alle lande. Opdelt efter lande eller indikator.

Note: På Danmarks Statistiks hjemmeside *http://www.dst.dk/da/Statistik/international-statistik/statistikbureauerne* er der en omfattende liste over internationale statistikproducenter.

Hvis du søger kvantitative data i eksempelvis Google, vil det være en god ide at tilføje ordet tabel eller statistik (table eller statistics, hvis du søger på engelsk) i søgningen, således at du målrettet går efter tal. For eksempel vil søgningen:

„Illegal immigrants UK" give links til en række artikler om illegal indvandring, mens søgningen:

„Illegal immigrants UK statistics" vil give links til tabelmateriale.

2.2 Hvor sikre er stikprøvebaserede undersøgelser og meningsmålinger?

Ved folketingsvalget juni 2015, afstemningen om Brexit i UK i juni 2016 og præsidentvalget i USA november 2016 blev resultatet noget anderledes end forudsagt af meningsmålingerne. Det blev tydeligt, at disse

Kapitel 2

Databaser og talbehandling – *hvor finder jeg tal og hvad gør jeg ved dem?*

I forbindelse med valgkampene i 2016 (Brexit og præsidentvalget i USA) dukkede begrebet 'det postfaktuelle samfund' op i medierne. Adskillige påstande med enten et forkert eller fordrejet indhold indgik under og efter valgkampene. Det gjaldt også kvantitative data. Eksempelvis påstod Trump, at han vandt både flertallet i valgmandskollegiet (hvilket er sandt) og flertallet blandt samtlige vælgere (Hillary Clinton fik tre millioner flere stemmer end Trump). Netop problemet med falske nyheder ('fake news') gør, at man som bruger af data skal være ekstra opmærksom på de kilder, man anvender. Begrebet falske nyheder uddybes i kapitel 3. Dette kapitel beskæftiger sig med kvantitative datakilder og efterbehandlingen af disse.

Fokuspunkter i kapitel 2

1. **Hvad er registerdata?**
2. **Hvor sikre er stikprøvebaserede undersøgelser og meningsmålinger?**
3. **Hvilke beregninger skal man kunne til eksamen?**
4. **Hvad er lineær regression, og hvordan udføres den?**
5. **Hvordan måles økonomisk ulighed?**

åbenbart er behæftet med stor usikkerhed – en større usikkerhed end det fremgår af medierne, som kæmper om at være hurtigst til at bringe de nyeste målinger. Meningsmålinger er baseret på stikprøver. Det vil sige, at man kun spørger en brøkdel af den samlede population og dernæst forsøger at generalisere resultaterne til hele populationen på baggrund af stikprøven. Vi spørger få (stikprøven); men vil gerne udtale os om alle vælgerne!

Meningsmålinger

Meningsmåling: Stikprøvebaseret undersøgelse, hvor respondenterne kommer med en menings- eller holdningstilkendegivelse.

I tabel 2.1 er vist en **meningsmåling** fra dagbladet Børsen gennemført december 2016. I tabellen kan resultatet fra folketingsvalget året forinden aflæses i procent og i mandater. I kolonne 4 og 5 kan resultatet af meningsmålingen aflæses – også i procent og antal mandater. Endelig er der i kolonne 6 angivet usikkerheden på det pågældende partis procentvise andel. For eksempel skal Socialdemokraternes andel læses som 26,7 % ± 2,5 %, det vil sige, at partiets andel med stor sandsynlighed vil ligge mellem 24,2 % og 29,2 % af stemmerne.

Ifølge meningsmålingen vil blå blok få 45,4 % af stemmerne, og rød blok 53,4 % af stemmerne. Umiddelbart et klart flertal til rød blok, som dog er behæftet med stor usikkerhed.

Af noten til tabellen fremgår det, at 70 % af svarene er indsamlet telefonisk og 30 % via nettet. Hvorfor denne fordeling er valgt af Børsen,

Tabel 2.1. Meningsmåling. Børsen, den 30. december 2016

Parti	Valget 18.6.2015		Meningsmåling 16.-21.12. 2016		
	Procent	Mandater	Procent	Mandater	Usikkerhed
A - Socialdemokratiet	26,3	47	26,7	48	2,5
B - Det Radikale Venstre	4,6	8	7,0	12	1,4
C - Det Konservative Folkeparti	3,4	6	4,1	7	1,1
F - Socialistisk Folkeparti	4,2	7	4,6	8	1,2
I – Liberal alliance	7,5	13	6,1	11	1,3
K - Kristendemokraterne	0,8	-	0,5	-	0,4
O - Dansk Folkeparti	21,1	37	16,1	29	2,1
V - Venstre	19,5	34	15,9	28	2,0
Ø - Enhedslisten	7,8	14	8,7	16	1,6
Å - Alternativet	4,8	9	6,4	11	1,4
Nye Borgerlige	-	-	2,7	5	0,9
Andre partier	-	-	1,2	-	0,6

Note: Undersøgelsen er gennemført i perioden 16.-21.12 2016 og bygger på telefoniske (70 %) og webbaserede (30 %) svar fra 1230 personer udvalgt tilfældigt blandt personer på 18 år og derover. 14,6 % af vælgerne ved ikke, hvilket parti de vil stemme på.

sikkerhed end
gst til at bringe
prøver. Det vil
ulation og der-
ationen på bag-
l gerne udtale

en gennemført
valget året for-
kan resultatet af
ndater. Endelig
ende partis pro-
s andel læses
tor sandsynlig-

mmerne, og rød
til rød blok, som

er indsamlet tele-
valgt af Børsen,

6.-21.12. 2016	
ater	Usikkerhed
	2,5
	1,4
	1,1
	1,2
	1,3
	0,4
	2,1
	2,0
	1,6
	1,4
	0,9
	0,6

og webbaserede (30 %)
e ved ikke, hvilket parti

fremgår ikke, men det kan være for at sikre repræsentativiteten – at de cirka 1.230 personer er en afspejling af befolkningen (se side 93).

Som tidligere nævnt har meningsmålingerne ikke været særlig dygtige til at forudsige resultaterne af valg i en række lande, hvilket giver anledning til at stille spørgsmålet: Hvilke usikkerheder er der forbundet med meningsmålinger?

Fejlkilder: Forhold som kan give anledning til at resultaterne i en stikprøvebaseret undersøgelse bliver mindre pålidelige.

Helt grundlæggende er der tre **fejlkilder** i forbindelse med meningsmålinger:

- Den første drejer sig om matematikken
- Den anden er fejl i metoden og
- Den tredje er menneskets natur

Fejlkilder ved meningsmålinger

Den første fejl skyldes matematikken. En meningsmåling er behæftet med usikkerhed. Der gælder nogle tommelfingerregler. For det første skal udvælgelsen af respondenter være tilfældig. For det andet gælder, at jo større stikprøven er, jo mindre bliver usikkerheden. I Danmark gælder en tommelfingerregel om, at det normalt er nok at spørge ca. 1000 respondenter. For det tredje: Desto større sikkerhed man vil udtale sig med, jo større bliver usikkerheden. Det sidste lyder umiddelbart selvmodsigende, men vil blive forklaret nedenfor.

Det antages generelt om repræsentative stikprøver, at disse er normalfordelt (klokkeformede) omkring gennemsnittet. Det betyder, at hvis man gennemførte en stribe meningsmålinger, ville mange af disse vise en tilslutning til Venstre omkring de 15,9 % (se tabel 2.1). Nogle ville være længere fra de 15,9 %, og nogle ville være tættere på. For en normalfordeling kan det påvises, at i 95 % af tilfældene falder resultatet inden for gennemsnittet af tællingerne ± 1,96 gange standardafvigelsen. Intervallet benævnes 95 %-**konfidensintervallet**. Endvidere gælder det, at for 99 % af tilfældene falder resultatet indenfor ±2,57 gange med standardafvigelsen. Intervallet benævnes 99 %-konfidensintervallet.

Konfidensinterval: Ved et 95-procent konfidensinterval er der 95 % sandsynlighed for at resultatet vil ligge indenfor intervallet, som udregnes efter formlen til højre herfor.

Statistisk usikkerhed: Den usikkerhed som skyldes stikprøvens størrelse.

Formlen for den **statistiske usikkerhed** er:

$$\text{Usikkerhed} = \pm\ 1{,}96 * \sqrt{\frac{p * (100-p)}{n}}$$

, hvor n er stikprøvens størrelse, p er andelen i procent.

Et eksempel fra tabel 2.1.: For Socialdemokraterne er usikkerheden:

$$\text{Usikkerhed}_{\text{Socialdemokraterne}} = \pm\ 1{,}96 * \sqrt{\frac{26{,}7*(100-26{,}7)}{1230}} = \pm\ 2{,}5.$$

Socialdemokratiets andel vil altså i 95 % af tilfældene ligge mellem 24,2 % og 29,2 % af stemmerne.

Der er således tre faktorer, der afgør stikprøveusikkerheden:

a) Stikprøvens størrelse (n). Jo større stikprøve, jo mindre usikkerhed
b) Hvor stor er andelen? Bemærk, at usikkerheden vokser med andelen
c) Hvor sikre vil vi være? Et 99 %-konfidensinterval giver større sikkerhed end 95 %-konfidensinterval.

Opgave: *Beregn usikkerheden*

a) Prøv at ændre stikprøvens størrelse til 2000. Dvs. de 1230 erstattes med 2000. Hvordan bliver usikkerheden så for Socialdemokraterne i forhold til en stikprøve på 1230?
b) Prøv at udregne usikkerheden for Liberal Alliance.
c) Prøv at udregne usikkerheden for Socialdemokraterne med et 99 % konfidensinterval.

Et sidste spørgsmål: Kan du være sikker på at Socialdemokraterne på baggrund af meningsmålingen vil gå frem ved et kommende valg? Nej er svaret. Socialdemokraterne står til at få mellem 24,2 % og 29,2 % af stemmerne. De fik 26,6 % ved valget i 2015. Så det er bestemt ikke sikkert at partiet vil få fremgang, da der er overlap mellem det faktiske valgresultat og konfidensintervallet.

Den anden fejl skyldes stikprøven. Det klassiske eksempel er magasinet The Literary Digest, der i 1936 forudsagde republikansk (Alfred Landon) jordskredssejr over demokraten Franklin D. Roosevelt. Stikprøvestørrelsen var med 2,4 millioner svar astronomisk. Forudsigelsen holdt

Sådan gør Greens Analyseinstitut

Univers: Undersøgelsens univers er de godt 4,1 millioner valgberettigede personer i alderen 18 år og opefter – med bopæl i 2,59 millioner private husstande.

Stratificering: Der er stratificeret efter geografi. Det vil sige, at stikprøven afspejler vælgernes geografiske sammensætning. Bor 40 % af vælgerne i Jylland, skal der være 40 % jyder i stikprøven.

Vejning: Materialet er dynamisk vejet efter geografi samt vejet for skævheden i stikprøven i forhold til universet. I vægtningen indgår afgivet stemme ved sidste folketingsvalg.

Metode: Undersøgelsen er gennemført som telefoninterviews (70 %) og web-interviews (30 %) med et tilfældigt udsnit af vælgerbefolkningen i perioden: Der er foretaget op til tre genopkald, dersom svarpersonen ikke var at træffe. Med et gennemsnitligt antal respondenter på 1200 kan man med 95 % sikkerhed kalkulere med, at forskelle på 1 %-point eller derover repræsenterer reelle forskydninger for de mindre partier. For de større partier skal forskellene i procent være højere, før de er signifikante.

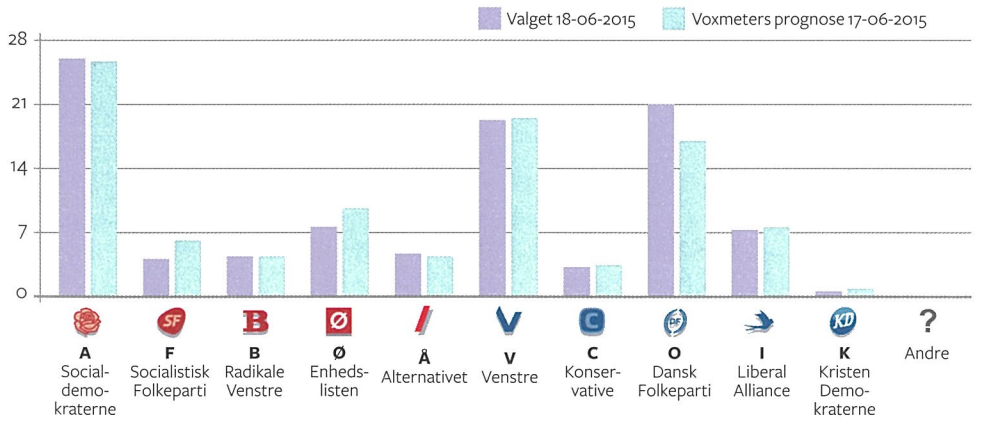
ikke stik. Roosevelt vandt med i alt 60,8 % af stemmerne. Problemet var, at man blandt andet kun havde spurgt egne læsere, bilejere og telefonabonnenter, hvilket udelukkede borgere med de laveste indkomster. Der var altså tale om en skæv stikprøve. Dette skrækeksempel har man lært af. Man gør alt for at minimere skævheden. Eksempelvis hos Greens Analyseinstitut, der gennemfører politiske meningsmålinger for blandt andet dagbladet Børsen (se boks).

Endelig er der for det tredje den menneskelige natur som fejlkilde. I politiske undersøgelser taler man om et såkaldt 'mørketal'. Det er betegnelsen for den andel af respondenterne, der enten bevidst eller ubevidst siger, at de vil stemme på et andet parti, end de rent faktisk har tænkt sig eller ender med at gøre. Dette metodiske problem kan opstå,

Voxmeter forudsagde 7 ud af 10 partier korrekt

Ved folketingsvalget den 18. juni 2015 var der ingen analyseinstitutter, som forudsagde Dansk Folkepartis eksplosive fremgang. Det var der mange årsager til, men den væsentligste var, at en massiv andel af „tvivlerne", da de stod i stemmeboksen, valgte at sætte kryds ved Dansk Folkeparti. Voxmeter brugte ugerne efter folketingsvalget på at analysere årsagerne og ringede op til samtlige „tvivlere", herunder til de vælgere, der havde sagt, at de stemte på SF og Enhedslisten. Studiet har medført, at Voxmeter fremover stiller en række kontrolspørgsmål, som øger træfsikkerheden. Årsagen til, at Voxmeter ikke ramte plet på SF og Enhedslisten, var primært, at en stor del af deres vælgere valgte at blive hjemme. Hyppigt med en begrundelse i kategorien „uanset hvad vil der blive ført blå politik".

Figur 2.1. Voxmeters prognose sammenholdt med valgresultatet juni 2015. Procent.



Kilde: http://voxmeter.dk/index.php/meningsmalinger/

hvis konkrete holdninger eller partier bliver genstand for massiv og konsekvent kritik eller opfattes som politisk ukorrekte i store dele af medierne og samfundet. Jo mere en person har oplevet situationer, der har været ubehagelige som følge af omverdenens fordømmelse i denne sammenhæng, desto mere tilbøjelig vil han/hun være til at prøve at undgå at svare. Denne problematik uddybes i boksen omhandlende Voxmeter.

Den usikkerhed, der er ved meningsmålinger, kan naturligvis overføres på alle typer af undersøgelser – også dine egne eller klassens undersøgelser. Forudsætningen er dog, at stikprøven er tilfældigt udvalgt.

Også hvis du bruger data fra Surveybank, skal du være opmærksom på usikkerheden i forbindelse med tolkning af resultaterne. Til den skriftlige prøve i samfundsfag skal du kunne beregne og tolke den statistiske usikkerhed.

Surveybank – her kan du lege forsker

En række institutioner stiller resultaterne af deres undersøgelser til brug for offentligheden. Det gælder for eksempel Surveybanken på Aalborg Universitet, hvor blandt andet resultaterne af flere års valgundersøgelser er frit tilgængelige. I det følgende vil der blive givet en kort introduktion til brugen af Surveybank. Hvorfor nu det? I forbindelse med vælgeradfærd er Surveybank simpelthen en guldgrube af informationer, hvor der er mulighed for at teste diverse hypoteser om vælgeradfærd. Samtidig er det muligt at kigge forskerne over skulderen og se de spørgsmål, der danner baggrund for resultaterne. Også resultaterne i Surveybank er baseret på stikprøver, det vil sige, at de samme forbehold som ved meningsmålinger vedrørende usikkerhed er gældende.

For at komme i gang klikker du på **START SURVEYBANKEN**.