

K O M M E N T A R

af Hans Bay

Vejning/vægtning af stikprøver

I løbet af det 20. århundrede er brugen af stikprøver (ofte omtalt som repræsentative stikprøver) steget dramatisk, og store forretningsområder er baseret på brugen af stikprøver. Tilbage i 1936 fik G. Gallup¹ et stort gennembrud, da han korrekt forudsagde udfaldet af det amerikanske præsidentvalg med stor præcision på grundlag af en meget lille, men velvalgt stikprøve. Så stor var succesen, at politiske opinionsundersøgelser stadig oftest omtales som Gallup-undersøgelser. Et andet eksempel er "Nielsen", som er et meget stort internationalt analyseinstitut, der også producerer viden på basis af indsamlede stikprøver. I den brede offentlighed er "Nielsen" mest kendt for målinger af hvad befolkningen ser af TV-udsendelser. Men også inden for detailhandelen er Nielsen en stor aktør. Utallige rapporter har Nielsen produceret om markedsandele vedrørende Coca-Cola, Pampers, OTA solgryn, Merrild kaffe etc. Disse rapporter er sædvanligvis bygget på en lille stikprøve af butikker.

Også offentlige myndigheder og universiteter gør flittigt brug af stikprøver. Danmarks Statistik udarbejder løbende vigtige nøgletal, der er baseret på stikprøver. Mest kendt er tillidsindikatoren², der reflekterer befolkningens økonomiske forventninger til fremtiden. Et andet kendt eksempel er PISA-undersøgelserne. Undervisningsministeriet besluttede tilbage i 2000, at Danmark skulle deltage i de nu meget berømte PISA³ undersøgelser. PISA undersøgelser er en stikprøve på ca. 4.200 elever, der deltager i en to timers test. PISA undersøgelserne har givet stor omtale og dermed lagt en grobund for at foretage ændringer i folkeskolen.

Mange sociologer gør brug af stikprøver, og det er umuligt at liste alle disse stikprøver. Men det er vel i dag svært at komme uden om European Social Survey (ESS)⁴ og værdiundersøgelsen⁵. Disse undersøgelser beskæftiger

sig med befolkningens værdier og holdninger inden for områderne politik, religion, familier etc.

I tidens løb har der været mange eksempler på, at stikprøverne "har taget fejl". Særlig berømt er episoden ved præsidentvalget i 1948. Ved præsidentvalget benyttede Gallup "exit-poll" metoden. Dvs. at de spurgte vælgere, der havde afgivet deres stemme, hvad de havde stemt på. Resultatet blev, at det anerkendte Gallup Institut forudsagde, at republikaneren Dewey ville vinde det amerikanske præsidentvalg over demokraten Truman. Dette fejlagtige resultat bragte nogle aviser på deres forside den følgende dag. Især billedet af Truman, der viser en avis forside, der hævder at republikaneren har vundet præsidentvalget, er kendt⁶. Der har siden været en diskussion af, hvad Gallup gjorde galt. Dette er ikke helt afklaret, men for eksemplets skyld kunne man hævde, at Gallup havde foretaget telefoninterviews. Da telefoner i 1948 ikke var lige udbredt i befolkningen medførte dette uundgåeligt en skævhed i stikprøven.

I nyere tid kan nævnes det danske nej til Maastricht-traktaten i 1992. I samtlige aviser viste samtlige opinionsundersøgelser, at det ville blive et ja ved afstemningen. Regeringen og regeringsapparatet blev tydeligvis overrasket over, at det blev et nej, og efterfølgende fik SF en stor mulighed for at få indflydelse på Europapolitikken.

I 1998 var der tilsvarende stor tiltro til opinionsundersøgelser, der alle forudsagde, at det ville blive en borgerlig regering⁷. Men et meget tæt opgør blev afgjort til Socialdemokraternes side.

Det er derfor klart, at stikprøver kan være et endog meget effektivt redskab, men at det skal bruges korrekt. I det følgende vil der derfor blive fokuseret på begreberne repræsentativ stikprøve og vejning af en stikprøve.

Begrebet en "repræsentativ stikprøve" er hyppigt anvendt i litteraturen – ikke mindst i avisartikler. Det dækker over, at man har et mindre datasæt "repræsentativt" udtrukket fra et større datasæt, og at man derfor kan lave beregninger i det mindre datasæt, som uhindret kan overføres til det store datasæt.

En egentlig definition af en repræsentativ stikprøve findes ikke, men umiddelbart opfatter man med en "repræsentativ stikprøve", at man har sikret sig en geografisk, aldersmæssig, uddannelsesmæssig etc. variation, der afspejler den variation, der er i det store datasæt – universet.

Når man skal beskrive begrebet repræsentative stikprøver, så bruger man ofte begreberne:

Univers: Den målgruppe hvori stikprøven skal trækkes.

Det kan være alle statsborgere i Danmark, eller alle personer i Danmark, der har mulighed for at stemme til folketingsvalget, eller alle indskrevne studenter på Københavns Universitet etc.

Stikprøve (sample): Den del af universet, der skal bruges til at beskrive universet.

Respondenter: de personer i universet, der er blevet udvalgt til at være i stikprøven.

Udtrækningsandsynlighed: Sandsynligheden for at en given person i universet bliver udtrukket til stikprøven.

Den statistiske teori, der danner grundlaget for "repræsentative stikprøver", er stikprøveteorien^{8 9}. Her er udgangspunktet: *en simpel tilfældig udtrækning i universet*. Ved simpel tilfældig forstås, at udtrækningsandsynligheden er ens for alle personer i universet.¹⁰

Hvis en simpel tilfældig udvælgelse er foretaget, så gælder, at gennemsnittet i stikprøven er en middelret estimator¹¹ for gennemsnittet i universet. Med andre ord så kan man tillade sig at bruge gennemsnittet i stikprøven som udtryk for gennemsnittet i universet. Den klare betingelse for dette er, at stikprøven er trukket simpelt tilfældigt.

Det statistiske faktum, at gennemsnittet af en stikprøve kan bruges som gennemsnittet for universet, når stikprøven er udtrukket simpelt tilfældigt, leder derfor til følgende naturlige definition:

En repræsentativ stikprøve er en stikprøve, som er udtrukket simpelt tilfældigt i universet.

Her skal det bemærkes, at denne definition af "repræsentativ" ikke indeholder noget krav om størrelsen af stikprøven. Så der er ikke noget umiddelbart krav om, at repræsentative stikprøver skal have en vis størrelse. Men når man har en repræsentativ stikprøve, er det selvfølgelig naturligt at spørge om: "Hvor sikker er så denne stikprøve?" Svaret på dette spørgsmål kræver, at man kan sætte en ramme op, der gør, at man kan bruge en række statistiske argumenter.

Den usikkerhed¹², der knyttes til en simpel tilfældig stikprøve, er sammensat af tre komponenter:

1. Komponent: Stikprøvens størrelse i forhold til universets størrelse.
2. Komponent: Stikprøvens absolutte størrelse.
3. Komponent: Den variation der findes i universet.

Da det er bekosteligt at gennemføre surveys, er man interesseret i ikke at skulle spørge flere end nødvendigt. De tre ovennævnte komponenter giver mulighed for at beregne hvor stor en stikprøve man skal have, i forhold til hvor præcist man vil kunne bruge svarene. Så det spørgsmål, som man kan besvare, er: Hvor stor skal stikprøven være, hvis man ønsker, at usikkerheden på stikprøven skal være mindre end en given størrelse? Dette er muligt at beregne.

Eksempel 1 (tilslutning til et parti)

Hvis et politisk parti har 30 % tilslutning i befolkningen, vil en simpel tilfældig stikprøve på 1.000 respondenter have et 95 % konfidensinterval på $\pm 2,8$ %.

Det betyder, at målinger, der er i intervallet 27,2 % til 32,8 % (svarende til 30 \pm 2,8%), er udtryk for en *uændret* tilslutning til partiet. Så hvis partiet opnåede 30 % ved sidste folketingsvalg, så vil en opinionsundersøgelse, der viser en nuværende tilslutning på eksempelvis 27,5 %, være udtryk for en uændret tilslutning til partiet. Mange aviser vil sikkert falde for den fristelse at beskrive et "fald" fra 30 % ved valget til kun 27,5 % i meningsmålingen som udtryk for en "vælgerlusing". Men ud fra statistiske argumenter, er der ikke belæg for at konkludere, at tilslutningen er faldet.

Ønsker man at kunne afgøre, om forholdsvis små forskelle er udtryk for signifikante ændringer, er man nødt til at øge stikprøven. I nogle tilfælde vil den ønskede sikkerhed kræve relativt store stikprøver og dermed øge omkostningerne væsentligt¹³.

Der findes en række metoder til at reducere usikkerheden på, når man laver opinionsundersøgelser. Disse metoder vil ikke blive omtalt her.

Stikprøveudtrækninger i praksis

I Danmark har alle indbyggere et unikt CPR nummer, og præcis derfor er det muligt at udføre simple tilfældige stikprøveudtrækninger. Uden brug af CPR numre (eller tilsvarende) ville det være meget vanskeligt at kreere simple tilfældige stikprøver. Brug af telefonnumre vanskeliggøres af hemmelige telefonnumre, telefonnumre der er "forladt", manglende lister fra telefonselskaberne etc. Det må derfor konstateres, at det ofte er meget vanskeligt at trække simple tilfældige stikprøver i den danske (eller andre) befolkninger.

Ofte vil man uundgåeligt komme i den situation, at den indsamlede stikprøve er skæv, da der er nogle persontyper, der ikke har deltaget i undersøgelsen, eller udtrækket er skævt af anden grund.

Stikprøveindsamlinger er bekostelige og ofte krævende logistisk, og man er interesseret i at kunne bruge resultaterne til at sige noget generelt, og derfor vil en "skæv" stikprøve blive forsøgt "vejet/vægtet" på plads, således at den vejede/vægtede stikprøve kan betragtes som repræsentativ.

Vejning af stikprøver

Princippet i vejning af stikprøver er, at respondenter fra grupperinger, der er underrepræsenteret i stikprøven, tildeles en større vægt og vice versa.

Som illustration bruges den danske del af European Social Survey (ESS)¹⁴ indsamlet i år 2006.

ESS datamaterialet kan downloades fra nettet bl. a. fra Sociologisk Databank¹⁵. ESS er en omfattende survey (stikprøve) gennemført i en lang række europæiske lande. Formålet er at måle holdninger, værdier med mere blandt europæere. Et grundigt forarbejde blev foretaget inden landenes survey gik i

gang. ESS er blevet indsamlet hvert 2. år, første gang i år 2002.

Den danske del omfatter 1.505 realiserede interviews. Selve stikprøven er udtrukket simpelt tilfældigt baseret på CPR-numre og består af personer over 15 år. Der blev oprindeligt udtrukket 2.964 respondenter, og samlet indkom der besvarelser fra 1.505 respondenter. Der er dermed en opnåelse på 50,8 % og følgelig et bortfald på 49,2 %. En del af bortfaldet skyldes forskerbeskyttelsen, som vi i det følgende vil se bort fra.

For at gøre det kommende eksempel mere skarpt vil vi i det følgende indskrænke stikprøven til de respondenter, som har besvaret spørgsmålet: "Hvilket parti stemte De på ved folketingsvalget i år 2005?" Dermed har vi nu en realiseret stikprøve på 1.244.

Igen for eksemplets skyld, inddeler vi i 6 køns- & aldersgrupper/strata, og gør op, hvor mange personer vi kunne forvente svarede ifølge den store undersøgelse, i forhold til hvor mange der har svaret. Dette fremgår af tabel 1.

Tabel 1. Univers og stikprøve for den danske del af ESS år 2006.

		Univers	% Ford.	Stik	% Ford.	Vægte	DF stemmer i stikprøven uvægtet	DF stemmer i stikprøven vægtet
M	15-40	874.044	19,8	165	13,3	1,49	10	14,9
M	40- 69	1.057.619	24,0	368	29,6	0,81	30	24,3
M	70+	233.798	5,3	77	6,2	0,86	12	10,3
K	15-40	851.737	19,3	179	14,4	1,34	6	8,0
K	40-69	1.053.420	23,9	359	28,9	0,83	35	29,1
K	70+	340.962	7,7	96	7,7	1,00	4	4,0
	I alt	4.411.580	100	1244	100		97	90,6

Det fremgår af ovenstående tabel 1, at stikprøven er skæv. Skævheden er især opstået i de unge aldersgrupper, hvor der er få realiserede interviews. De unge mænd (alder 15 til 40) udgør 19,8 % af universet, men i stikprøven udgør de kun 13,3 %. Der er altså tale om, at de unge mænd er underrepræsenterede.

Hvis man tester om stikprøvens fordeling svarer til universets fordeling, så må det konkluderes, at skævheden er meget signifikant¹⁶.

Det vil så ofte føre til, at man konstruerer nogle vægte, der kan *veje* stikprøven på plads. Konstruktionen af vægte bygger på, at man inddeler universet i en række strata, i ovenstående eksempel er der tale om 6 strata (som altså er baseret på køn og alder). Man vægter så de strata *op*, som har for få respondenter, og vægter de strata *ned*, som har for mange.

Den egentlige konstruktion af vægtene svarer til, at man tager forholdet mellem universets andel og den tilsvarende andel for stikprøven. I det første

stratum er vægten udregnet til 1,49, som er forholdet mellem 19,8 og 13,3. At vægten er større end 1 betyder, at respondenter i dette stratum tæller for mere end en person, helt præcist for 1,49 personer. Så de unge mænds svar i dette stratum tillægges en større værdi for at kompensere for, at de er underrepræsenterede.

I den statistiske litteratur vil vægtning ofte blive omtalt som en form for (efter)-stratifikation¹⁷. I anden sammenhæng vil denne vejning blive omtalt som "demografisk vejning"¹⁸, da stratifikationen er baseret på demografiske kriterier.

Ovenstående teknik er meget operationel, og det ligger derfor lige for at indføre denne vægtning som en form for standard ved indsamling af stikprøver. Man kan efterfølgende meget naturligt hævde, at man har "vægtet" stikprøven på plads, men man kan stille spørgsmålstegn ved nytten af denne vægtning.

Tester man hvordan en vægtning virker vedrørende stemmeafgivningen ved sidste folketingsvalg, var der i stikprøven 97 personer af 1.244 der sagde, at de havde stemt på Dansk Folkeparti (DF) ved sidste valg, svarende til 7,8%. Ved Folketingsvalget i 2005 fik DF i realiteten 13,3% af de afgivne stemmer, så der er ingen tvivl om, at DF er underrepræsenteret i den aktuelle stikprøve. Formålet med en vægtning skulle være at afhjælpe denne skævhed. Det viser sig dog, at hvis man anvender vægtene på de afgivne stemmer på DF, vil man få, at det samlede antal DF stemmer i den vægtede stikprøve bliver 90,6 stemmer svarende til 7,3%'s andel. På trods af vægte, fik man altså ikke vejret DF's andel på plads. Dette skyldes, at skævheden i antallet af personer, der angiver at have stemt på/ henholdsvis ikke at have stemt på DF, ikke kan forklares med demografiske faktorer. Eksemplet viser, hvorledes andelen af stemmer på DF trods vejningen forbliver skæv.

At vejning ikke altid afhjælper skævheder i stikprøven underbygges også i litteraturen vedrørende stikprøveteorien:

...I mange praktiske eksempler viser det sig tilsvarende, at det kun er til besvær at betragte inddelinger i strata efter irrelevante kriterier. Så en bevidstløs inddeling efter køn, urbaniseringsgrad etc. kan ikke anbefales som en generel fremgangsmåde.... (Milhøj).¹⁹

Ved at inddrage stikprøveteorien kan man vise, at variansen på de to estimatorer (7,8% og 7,3%) stort set er identiske, hvilket er en konsekvens af, at den aktuelle stratifikation er irrelevant.

Vægtning er dog ikke altid problematisk, hvilket et andet eksempel fra Milhøjs²⁰ bog, der handler om personer, der læser ugebladet Anders And, viser. Her er inddelt i to strata, som er baseret på alder.

Tabel 2.

	Univers	Univers- fordeling	Stikprøve	Antal, der læser Anders And	Stik- prøve- fordeling	Vægte	Vægtet stikprøve
Unge 13-19	528.000	0,12	2.000	800	0,33	0,36	288
Ældre 20+	3.872.000	0,88	4.000	400	0,67	1,32	528
	4.400.000	1,00	6.000	1200	1,00		816

I tabel 2 fremgår universet, som samlet er på 4,4 mio. personer, svarende til antallet af personer der er ældre end 13 år. En stikprøve er udtrukket på samlet 6.000 personer. Tallene viser, at stikprøven er skæv, da der er en overrepræsentation af unge personer. De unge udgør 33% af stikprøven, mens de i universet kun udgør 12%.

Der er derfor konstrueret vægte, og vægten for de unge er på 0,36 og altså klart mindre end 1, hvilket skyldes at man ønsker at mindske de unges svar. Antallet af personer i den samlede (uvægtede) stikprøve, der læser Anders And, er på 1200, mens antallet i den vægtede stikprøve er på 816. Vægten betyder, at den samlede andel i populationen, der læser Anders And, er på 13,6 %, hvilket var et godt estimat i 1990'erne. Såfremt tallene ikke vejes, er resultatet, at ca. 20% af befolkningen over 13 år i 1990'erne læste Anders And, og dette tal er for højt.

Konklusionen er, at vægtning afhænger af, hvilken problemstilling man betragter. Kvalificeret vægtning beskrives som noget af en kunst, der kræver stor indsigt i emnerne og det undersøgte univers samt det brugte stikprøvedesign.

At en stikprøve er vægtet gør den ikke "repræsentativ" forstået på den måde, at den vægtede stikprøve er et billede af universet. Hvis det er nødvendigt at vægte en stikprøve, skal der argumenteres for, at den udførte vægtning er relevant.

Noter

1. Se www.Wikipedia.org samt <http://www.denstoredanske.dk/>
2. Se www.dst.dk
3. Se www.pisa.oecd.org og www.dpu.dk
4. Om ESS se <http://www.europeansocialsurvey.org/>
5. <http://www.sociology.ku.dk/vaerdi/ddvhome.html>
6. Se www.wikipedia.org
7. Det skal noteres, at exit polls (der blev offentliggjort efter at valget var afsluttet) klart forudsagde, at den socialdemokratiske blok (ekskl. Grønland og Færøerne) ville vinde valget.

8. Milhøj, Anders (1994) Stikprøvetæori for økonomi og erhverv. Bogen er udsolgt fra forlaget, men kan hentes på nettet <http://www.econ.ku.dk/milhoj/stik/stik.htm>
9. Cochran, W. G. 1977: *Sampling techniques*. Wiley.
10. Mere præcist, N er antallet af respondenter i universet og n er stikprøvestørrelsen. Så skal alle kombinationer af n udtaget blandt N være lige sandsynlige.
11. Andersen, Jensen og Kousgaard 1987: *Teoretisk statistik for økonomer*. Akademisk forlag.
12. Formel for usikkerhed på en simpel tilfældig stikprøve er: $\text{var}(\bar{y}_{st}) = \frac{N-n}{nN} S^2$, den første komponent er $\frac{N-n}{N}$, som også kaldes endelighedskorrek-
 tionen, det er denne kom-
 N
 ponent, der forsvinder når der tages stikprøver i store universer. Variationen S^2 kan i det alternative tilfælde skrives som $p^*(1-p)$, hvor p angiver andelen i universet.
13. En stikprøvestørrelse på 8.000 giver et 95% konfidensinterval, der er $\pm 1\%$.
14. Om ESS se <http://www.europeansocialsurvey.org/>
15. Sociologisk Databank se <http://databank.soc.ku.dk/>
16. χ^2 teststørrelsen (5 frihedsgrader) bliver 73, hvilket er klart signifikant.
17. Se kapitel 3 i Milhøj.
18. Søren Risbjerg Thomsen 1995: *Politica*, bind 27, side 71.
19. Milhøj side 175.
20. Milhøj side 177, eksemplet er oprindeligt fra Dansk Medieindeks i starten af 1990'erne. Eksemplet er modificeret for at tydeliggøre pointen.