

Hvordan vælger jeg den rigtige størrelse på en stikprøve?

Er kvantitet altid kvalitet?

Peter Linde, konsulent, Peter@Brede.dk

Dagsorden

- Hvad tæller?
- Store stikprøver
- Tilfældige og systematiske fejl
- Optimering af stikprøver
- Forskerbeskyttelse – igen?
- Afslutning

Hvad tæller? (1)

Einstein:

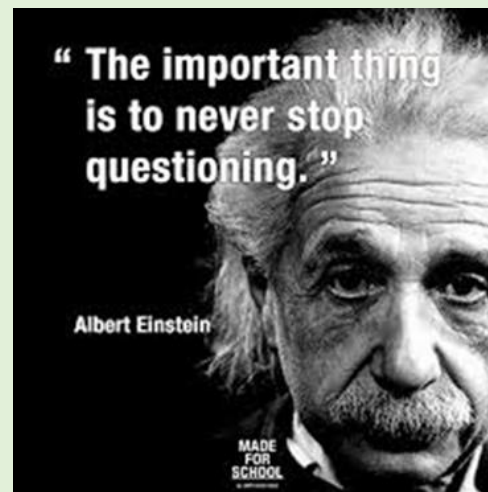
Ikke alt der kan tælles tæller

– og ikke alt der tæller kan tælles

Udfordringen ved surveys to hovedfejl:

- Det er svært at måle **bias**
- Den **tilfældige fejl** kan beregnes

Den tilfældige fejl falder med kvadratroden af antal svar, så det er noget der kan tælles



Hvad tæller? (2)

Tre punkter man altid skal have fokus på:

- **Udvalgsrammen, udvælgelsen og antal svar.** Sandsynlighedsudvælgelse fra hele populationen.
- **Hvordan man spørger.** Spørgsmålene og design af spørgeskemaet. Let og brugervenligt.
- **Bortfald og dataindsamlingen.** Mød de udvalgte på deres betingelser og gør det let. Brug tid og penge optimalt. Lille byrde i bred forstand er forudsætningen for mange deltager og gode svar. Det giver den mindste bias.

Antal svar er således ét af mange bidrag til kvaliteten



Store stikprøver (1)

- **Private firmaer** har adgang til digitalpost, hvis borgeren har tilmeldt sig firmaet aktivt.
- **Offentlige myndigheder** har adgang til digitalpost til borgerne til deres myndighedsopgaver.
- Universiteter, Danmarks Statistik og forskningsinstitutioner er juridisk myndigheder.
- Digitalpost med et link til spørgeskema koster op til 1 krone stykket.
- Et brev fra 12 til 24 kroner med et brev med et link eller papirskema.
- Normale offentlige myndigheder har pligt til at sende brev om deres myndighedsopgave, hvis borgeren er fritaget for digital post. Universiteter og forskningsinstitutioner har ikke oplevet at få krav herom. Kravet er knyttet til det er information til borgerne om myndighedsopgaver.
- Hvis man vælger at undlade opfølgning bliver stykprisen for opnåede svar endnu mindre.
- **Analysefirmaer har web-paneler**, hvor de sender til de, der står i basen. Basen kan være ret stor, med alle de, der gennem tiden har været i kontakt med firmaet eller selv henvendt sig.

Det er er blevet muligt at lave store stikprøver med en lille stykpris.

Store stikprøver (2)

Når man alle med digital post?

Omkring 6% er fritaget fra digitalpost fordi de har bedt herom. Typisk ældre, folk uden uddannelse eller indvandre. Hvis man ikke sender brev til dem giver det bias i udvælgelsen.

Læser alle deres digitale post?

Der er få data herom, selv om informationerne findes i e-Boks. Hvorfor?

Tænketanken JUSTITIA har en undersøgelse fra 2022:

6-7,5% er fritaget fra digital post.

10-15% har store problemer med at bruge digitale løsninger og digital post.

2,5% ved ikke hvordan man bliver fritaget for digital post.

Og så er der dem, der får hjælp fra kollegaer, ægtefælle eller familie. Blevet sværere med scanner genkendelse, hvis et andet medie bruges.

Store stikprøver (3)

- **Et web-panel** dannes i den bedste af alle verdener fra en repræsentativ CPR stikprøve.
- Danmarks Statistik indsamlede hver måned i godt to år deltagere til et web-panel. Godt 2.000 blev udvalgt hver måned repræsentativt fra CPR, 1.500 svarede og 600 heraf sagde ja til at indgå i et webpanel. I alt endte det med 20.000 i web-panelet:

Eksempler på bias i dette web-panel:

	Population	Web-panel
18-34 year	29,0%	23,1%
50-64 year	26,8%	30,6%
Elementary school	32,3%	22,2%
Long education	7,5%	10,4%
<100.000 DDK	18,1%	12,3%
>600.000 DDK	15,8%	20,9%

Store stikprøver (4)

Web-panelet var mere skævt end normalt bortfald. Kan man rette op på det i udvælgelsen? Danmarks Statistik trak 3.000 af de 20.000, så de var så tæt som muligt på fem faktorer:

	Population	Stikprøven	Opnåelsen
G: Men	50,1%	50,1%	53,7%
A: 18-34 year	29,0%	29,1%	22,7%
R: South DK	21,1%	21,0%	22,7%
E: Elementary	32,3%	31,6%	27,5%
F: Single parents	28,1%	28,1%	18,9%

Konklusion.

Web-panelet er skævt fra begyndelsen. Når der i udvælgelsen rettes op herpå, gentager skævheden sig. Selv om der udvælges blandt de, der tidligere har deltaget og sagt ja!

Tilfældige og systematiske fejl (1)

Ingen statistik er bedre end sit svageste punkt.

- Hvis udvælgelsen ikke er repræsentativ
- Hvis spørgsmålene ikke er til at svare på
- Hvis folk ikke vil svare

Så hjælper det ikke man skruer op for alt andet



Tilfældige og systematiske fejl (2)

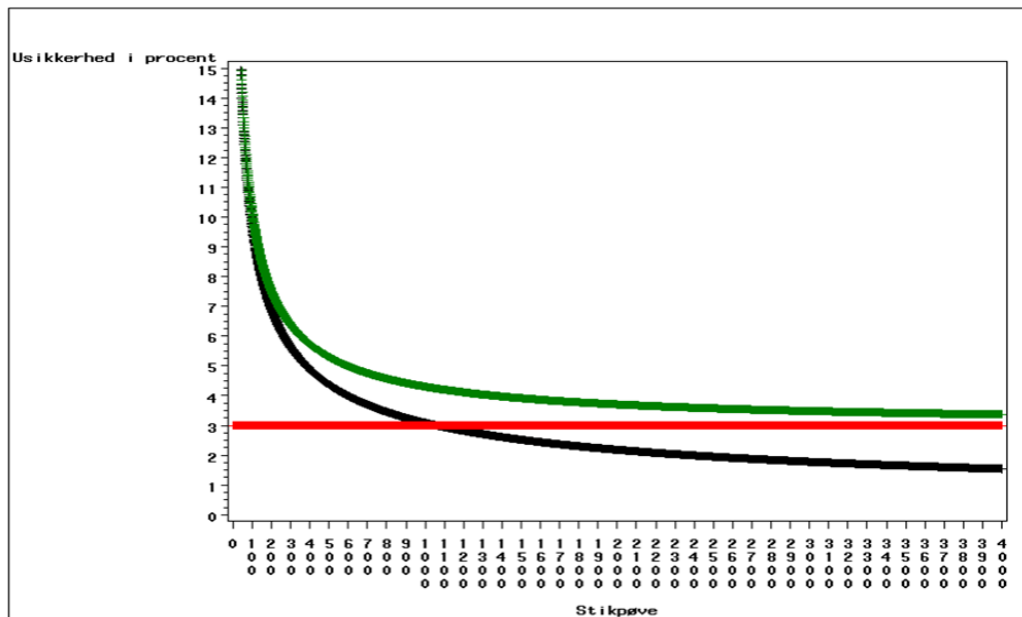
Bortfald vil altid betyde vi i data har flere integrerede, med lang uddannelse, ældre, i beskæftigelse, par og høj indkomst:

Indkomst i populationen:	210,235	
Indkomst i SRS stikprøve på 43,500:	210,735	(+0.2%)
Indkomst efter bortfald :	228,641	(+8.8%)

Bias og stort bortfald vil altid male billedet mere positivt: Bedre helbred, flere gør noget for miljøet og færre drikker eller ryger. Færre føler sig utrygge, flere ser positivt på fremtiden og økonomien er bedre.

Tilfældige og systematiske fejl (3)

- Man kan ikke løse kvalitetsproblemer ved at spørge flere
- Kvantitet slår ikke over i ny kvalitet
- Hvis man med en metode med bias spørger flere er det eneste sikre, at survey-en mere sikkert er forkert. Med små stikprøver kan det rigtige godt ligge indenfor sikkerhedsintervallet.



Rød: Bias

Sort: Tilfældig fejl (SRS)

Grøn: Samlet fejl (MSE)

Optimering af stikprøver (1)

Et argument for en stor stikprøve er man vil kunne gå ned i undergrupper med pæn styrke.

Også små undergrupper. Fx små kommuner.

For at få nok svar i små undergrupper, fx 100, trækkes en meget stor stikprøve.

Man kan også stratificere efter disse undergrupper.

Optimering af stikprøver (2)

A: Hvis man vil optimere små grupper og sammenligning imellem grupper:
= træk lige mange i hvert gruppe (stratum)

B: Hvis man vil optimere alle grupperne under et, fx hele landet
= træk simpelt tilfældigt i hele populationen.

Hvis man optimerer efter A koster det for B pga. vægte. Tilsvarende er det sværere at analysere små grupper, hvis tilfældigt i hele populationen.

Optimering af stikprøver (3)

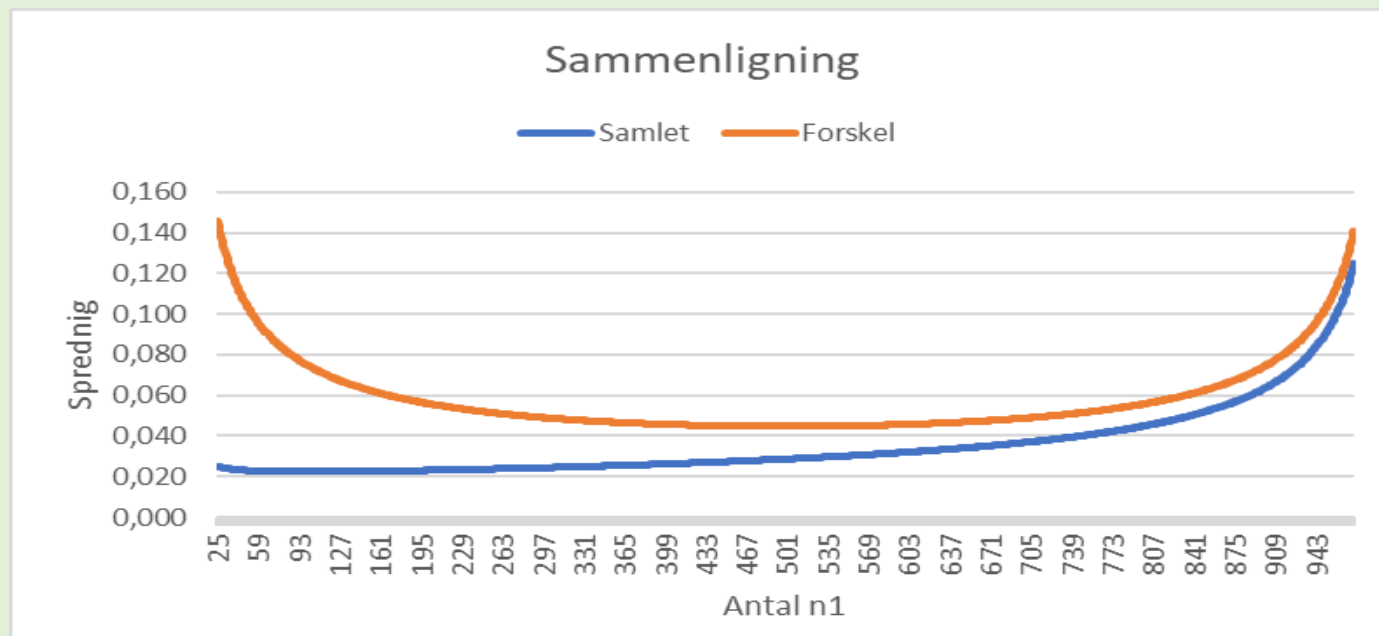
Løsningen på dilemmaet er ret simpelt.

Man trækker en del af stikprøven efter princip A og den resterende stikprøve efter princip B. Man har så ikke helt små undergrupper og mindre vægte for de generelle estimater for hele populationen.

Og regner så på de forskellige alternativer.

Optimering af stikprøver (4)

I eksemplet nedenfor er der to grupper, en på 10% og en på 90% af populationen. Varianser antages ens. Den samlede stikprøve er 1.000. Stikprøven n_1 til stratummet på 10% varierer fra 25 til næsten 950.

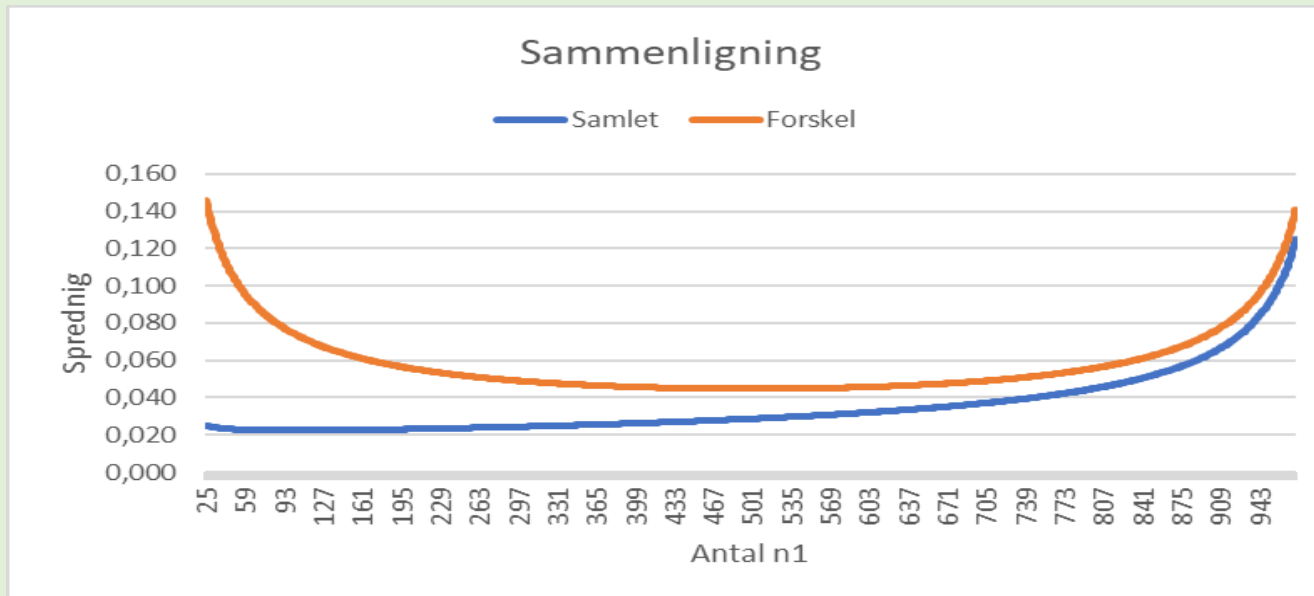


Optimering af stikprøver (5)

Spredningen for **forskellen** er ret flad på en pænt stykke omkring $n_1=500$.

Hvis der stratificeret efter undergrupper er spredning stejl i begge ender.

Spredningen for det **generelle estimat** kan ikke lide at komme langt væk fra $n_1=100$. Falder (lidt) ned til 100, stiger derefter fladt og derefter mere stejlt.



Optimering af stikprøver (6)

Man kunne vælge at fordele halvdelen (500) af stikprøven efter det samlede estimat
Og den anden halvdel (500) efter det, der var optimalt for sammenligningen.

Det ville give $n_1=100/2+500/2=300$. Og dermed $n_2=700$.

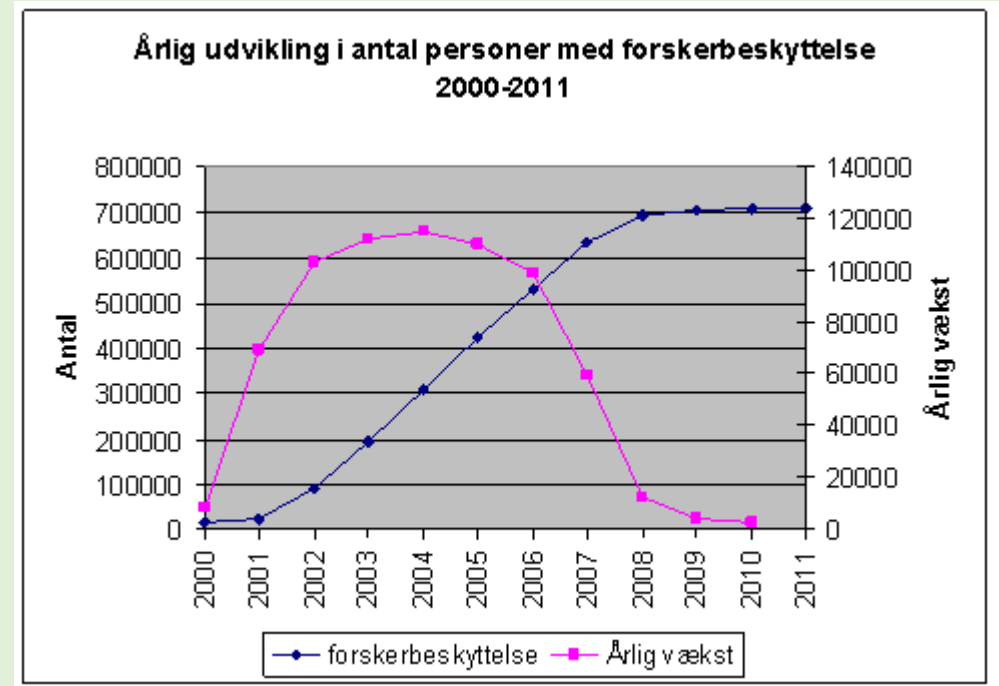
$n_1=300$ koster for begge mål kun lidt (9%) i forhold til deres optimale!

n1	Samlet		Forskel	
100	0,022361	100	0,074536	167
300	0,024398	109	0,048795	109
500	0,028636	128	0,044721	100

Grunden til 500/500 går godt er, at spredninger kun falder med kvadratroden af stikprøven. Og er ret flad omkring det optimale. Så det koster lidt at tage noget fra de store grupper og vinder meget at give det til de mindre grupper.

Forskerbeskyttelse (1)

- Fra 1995 var det muligt at blive fritaget for udvælgelse til statistiske undersøgelser. 24 havde i 1999 brugt denne mulighed.
- I 2000 overgik registreringen til kommunerne og de satte det på flytteblanketten.



Forskerbeskyttelse (2)

- Knap 800.000 have tilmeldt sig i 2014. 16% af den voksne befolkning.
- Det blev ophævet med lov 1. april 2014 efter nogen debat:
- *To medlemmer af Etisk Råd kan gå ind for den foreslåede ophævelse i CPR om forskerbeskyttelse. Andre (14) af Rådets medlemmer er af den opfattelse, at muligheden for markering i CPR af forskerbeskyttelse skal bibeholdes.*



Forskerbeskyttelse (3)

Tal om forskerbeskyttelsen 1. januar 2011 (14%)

Alder		Socioøkonomisk	
0-9 år	8,5%	Selvstændig	12,1%
10-19	12,5%	Lønmodtager	15,3%
20-29	21,2%	Arbejdsløs	21,4%
30-39	23,7%	Uddannelse	13,1%
40-49	14,4%	Pensioneret	7,5%
50-59	9,0%	Uden arbejds.	12,2%
60-69	6,5%		
70+	4,9%		

Forskerbeskyttelsen gav ekstra skævhed.
Og internationalt meget lave responsrater.

Opsamling

- Bias er i praksis større end den tilfældige fejl. Specielt i store stikprøver.
- Høj opnåelse, rykkere, flere indberetningsmuligheder, gennemtestede skemaer med lav byrde betyder meget.
- Man kan stratificere og få en mere effektiv stikprøve.
- De mange store stikprøver koster generelt på opnåelsen i Danmark.
- Hvis vi ikke passer på er vi på vej mod en ny forskerbeskyttelse.

Det er ikke nok, at én pil rammer i centrum.

Det er dem længst fra centrum, der sætter kvaliteten i en survey.

