Hvor god er statistikken?

Alle tall har en usikkerhet. De fleste tallene fra Statistisk sentralbyrå er ikke feilfrie, men de er nyttige. Det kan faktisk være umulig å finne den absolutte sannheten. For å vurdere nytten av tallene, er det viktig å kjenne feilkildene og størrelsen på disse.

Aslaug Hurlen Foss

Registrene – grunnfjellet i statistikkproduksjonen

I Statistisk sentralbyrå er de offisielle registrene selve grunnfjellet i statistikkproduksjonen. De fleste tall har et utgangspunkt i ett eller flere registre. Statistisk sentralbyrå har tilgang til omtrent 140 offentlige registre. Registrene er ofte ikke laget for å produsere statistikk, men til offentlig administrasjon. Det kan derfor være forskjeller mellom registrene og den statistikken som blir produsert. Registrene blir kontrollert og bearbeidet før de blir brukt.

Riktige enheter

Et hovedproblem er at registrene skal bestå av de riktige enhetene. Mennesker fødes, dør, innvandrer og utvandrer, alt dette skal bli registrert så hurtig som mulig etter at det har skjedd. Det er omtrent 160 000 slike endringer i Folkeregisteret i løpet av et år. Andre registre har mange flere slike endringer i forhold til antall enheter i registeret. Det kan derfor være en stor og vanskelig oppgave å holde et register à jour til enhver tid.

Korrekte variabler

I tillegg er det viktig at hver enhet har den riktige informasjonen knyttet til seg. I Folkeregisteret skal for eksempel hver person være knyttet til riktig adresse og riktig familie. Hvis noen ikke melder flytting til Folkeregisteret, vil dette føre til at folketallet i kommunen de flyttet fra og til blir feil, selv om folketallet totalt i landet er korrekt. I Folkeregisteret blir det årlig registrert omtrent 200 000 flyttinger mellom kommuner og nesten 400 000 flyttinger innenfor kommunen.

Kontrollrutiner

Registrene blir kontrollert med forskjellige metoder. Mellom enkelte variabler i et register kan det ofte være en sammenheng. Det er derfor mulig å kontrollere disse variablene ved å kjøre dem opp mot hverandre. Dette kan for eksempel være variabler som moms og sysselsetting i Bedriftsregisteret. Sannsynlige feil blir funnet ved denne metoden. I andre tilfeller kan det bli funnet feil som vi vet med sikkerhet er feil. Hvis det for eksempel i Folkeregisteret blir registrert et giftermål for en 13 år gammel jente er dette med sikkerhet feil, siden dette ikke er mulig ifølge norsk lov. Antakelig har vigselsdatoen blitt tastet inn feil. En annen måte å kontrollere registre på, er å sammenligne med en annen datakilde med tilsvarende opplysninger. Ved avvik må det da avgjøres hvilken datakilde som antakeligvis er korrekt.

Utvalgsundersøkelser – en verden i miniatyr

Det er mye statistikk som er etterspurt som det ikke finnes noen informasjon om i registrene. For å få denne informasjonen finnes det to alternativer: Enten spørre alle eller bare spørre noen. Det kan være svært dyrt å spørre alle sammen, det er derfor sjelden dette blir gjort i Statistisk sentralbyrå. Det hyppigste er altså utvalgsundersøkelser som Statistisk sentralbyrå har svært mange av i løpet av et år. I utvalgsundersøkelser blir bare et fåtall av

Hovedregistrene i Statistisk sentralbyrå:
Det sentrale folkeregister

Bedrifts- og foretaksregisteret

Grunneiendoms-, adresse- og bygningsregisteret

Aslaug Hurlen Foss *er statistikkrådgiver i Statistisk sentralbyrå, Seksjon for statistiske metoder og standarder (aslaug.hurlen.foss@ssb.no).*

enhetene spurt. Hver enhet teller ikke bare for seg selv, men representerer også noen som ikke ble trukket. Vi lager et utvalg av enheter som skal representere alle i hele populasjonen – en verden i miniatyr. Grunnlaget for å trekke ut enheter til utvalgsundersøkelsene er i hovedsak et av basisregistrene, det vil oftest si Folkeregisteret eller Bedriftsregisteret. Feil og mangler i basisregistrene kan dermed føre til feil i utvalgsundersøkelsene.

Utvalgsfeil – vi bommer på 1 av 20 undersøkelser

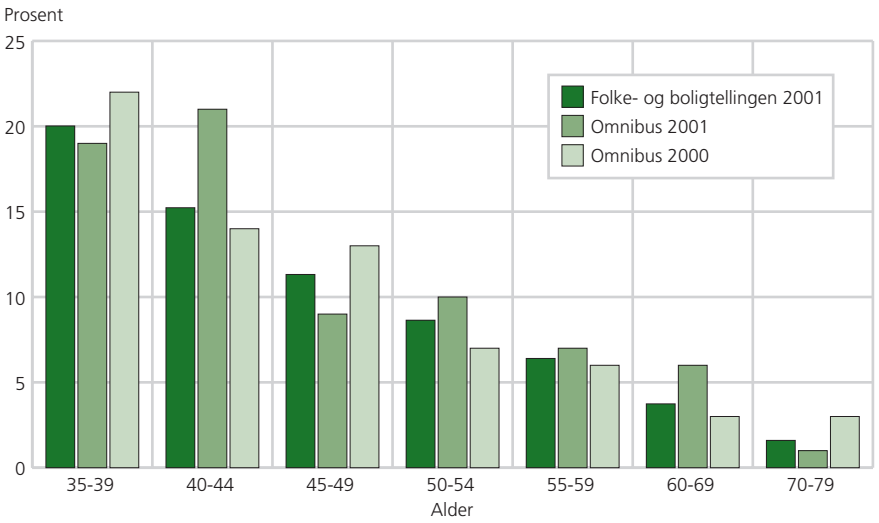
I mange utvalgsundersøkelser er det mulig å beregne konfidensintervall, det vil si det intervallet den sanne verdien vil ligge i med 95 prosent[1] sikkerhet. I praksis vil dette si at i 1 av 20 utvalgsundersøkelser ligger sannheten utenfor det konfidensintervallet vi har oppgitt. Utvalg vi har trukket tilfeldig er ikke representative for hele populasjonen i forhold til det vi skal måle. Det er to måter å oppdage at den sanne verdien sannsynligvis ligger utenfor konfidensintervallet. Den ene måten er hvis det finnes en tidsserie av dataene, da er det mulig å sjekke om tidsserien gjør noen merkelige hopp. En annen måte er å sjekke mot en annen kilde – hvis den finnes. Hvis vi ikke har noe å kontrollere mot, kan vi ikke vite hvilke undersøkelser vi bommer på.

I figur 1 er det et eksempel på utvalgsfeil i omnibusundersøkelsen i 2001. Andelen mannlige samboere mellom 40 og 44 år var i denne undersøkelsen mye høyere enn i den tilsvarende undersøkelsen året før og ifølge tallene fra folke- og boligtellingen samme år. Dette tyder på at det var en utvalgsfeil for denne gruppen i omnibusundersøkelsen i 2001. Den sanne verdien til andelen samboere for menn i alderen 40-44 år ligger utenfor konfidensintervallet.

Frafall – det er ikke alle som svarer på undersøkelsene

Det er alltid noen som ikke svarer på undersøkelser. Det kan være flere grunner til dette. En av dem kan være at de ikke er hjemme de gangene de får telefon fra Statistisk sentralbyrå. En annen kan være at de har problemer med å svare, eller ikke ønsker å svare. Frafallet er større i de frivillige undersøkelsene enn i de lovpålagte. For de lovpålagte undersøkelsene har Statistisk sentralbyrå lov til gi disse personene eller bedriftene en bot. Problemet med frafall blir løst på forskjellige måter. En måte å kompensere for frafall, er å la de personene (eller bedriftene) som er trukket ut telle for flere personer enn det som var planlagt. Det vil si at vi øker vektene på de som har svart for å kompensere for frafallet. En annen måte å kompensere for frafallet på, er å imputere verdier for de som ikke har svart. Å imputere er det motsatte av å amputere. Det vil si at personen eller bedriften som ikke svarte, blir tildelt verdier istedenfor de svarene de ikke ga. Det finnes mange måter å beregne hvilke verdier som skal bli gitt til en som ikke har svart.

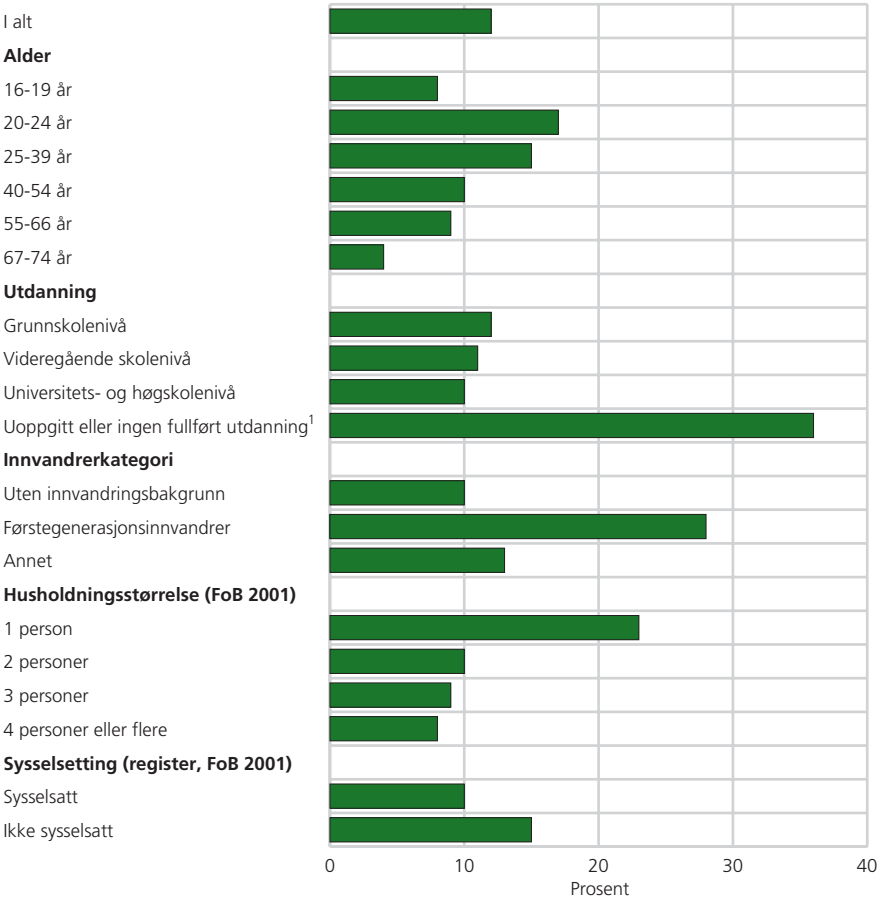
Figur 1. Andelen samboere. Menn. Prosent



Kilde: Folke- og boligtellingen 2001, Omnibusundersøkelsene 2000 og 2001.

I figur 2 er det vist hvordan frafallet i arbeidskraftundersøkelsen var i 4. kvartal 2001 (Foss 2004). Denne undersøkelsen er obligatorisk, men

Figur 2. Frafallet i arbeidskraftundersøkelsen. 4. kvartal 2001. Prosent



[1] Denne gruppen er liten og består hovedsakelig av førstegenerasjonsinnvandrere med kort botid.

Kilde: Frafallet i Arbeidskraftundersøkelsen 4. kvartal 2001 påkoblet variabler fra Folke- og boligtellingen 2001 (Foss 2004).

likevel var det totalt 12 prosent som ikke svarte. Disse ble ikke ilagt bot, selv om Statistisk sentralbyrå har lov til å gi dem det. Frafallet var størst blant dem mellom 20 og 25 år, her var det hele 17 prosent som ikke svarte på undersøkelsen. Frafallet minket ettersom alderen økte, for de mellom 67 og 74 år var frafallet bare på 4 prosent. I tillegg er det en tendens til at flere personer med høyere utdanning svarer på undersøkelser i forhold til personer med lav utdanning. Frafallet var svært stort for enpersonhusholdninger, der var det en fjerdedel som ikke svarte på undersøkelsen. For husholdninger med fire personer eller mer var frafallet bare 8 prosent. Det er svært mange førstegenerasjonsinnvandrere som ikke svarer på arbeidskraftundersøkelsen, andelen er hele 28 prosent. En grunn til dette kan være problemer med språket. I arbeidskraftundersøkelsen er ett av hovedformålene å anslå antallet arbeidsledige. Da er det ganske uheldig at frafallet i denne undersøkelsen er større for de som ikke er sysselsatt i forhold til de som er sysselsatt. (Informasjonen om sysselsetting er hentet fra register.) Hvordan frafallet er fordelt i denne undersøkelsen er trolig nokså typisk for de fleste personundersøkelser i Statistisk sentralbyrå.

Seksjon for metode og standarder

Statistisk sentralbyrå har en egen seksjon som skal hjelpe de som produserer statistikk til å få en best mulig kvalitet på sine produkter. Seksjonen har 13 personer med kompetanse innenfor matematisk statistikk. De jobber på fulltid med å hjelpe statistikkprodusentene med utvalgsplaner, frafall, vekting, revisjonsmetoder, analysemetoder og beregning av usikkerhet. De jobber i hovedsak ut fra forespørsler fra de som trenger hjelp, men de gir også kurs i statistiske metoder.

Målefeil – det er ikke alltid enkelt å svare korrekt

Det kan være vanskelig å svare på spørsmål, spesielt å svare riktig på spørsmålene. Det kan være mange grunner til dette. Ord og uttrykk kan ha en annen betydning for de som svarer enn for de som stiller spørsmålet. En måte å avdekke dette er ved kognitiv kartlegging eller ved fokusgrupper. Disse metodene prøver å avdekke hvordan spørsmålene blir oppfattet av en gruppe som ligner på de som skal svare på undersøkelsen.

I noen tilfeller er spørsmålene upresist formulert. Dette vil som regel føre til at svaret også blir upresist. I noen undersøkelser er det lange forklaringer til et upresist spørsmål. Det blir ofte ikke bedre, siden mange har problemer med å huske lange setninger.

Det er ikke alle detaljer som er vanlig å huske til enhver tid. Det kan derfor være at vi svarer ganske omtrentlig på enkelte spørsmål. Dette gjelder spesielt hendelser som ligger langt tilbake i tid eller ting som er kompliserte.

Tabell 1. Antall rom i folke- og boligtellingen og boforholdsundersøkelsen[1]. 2001. Prosent

Antall rom i folke- og boligtellingen	Antall	Antall rom i boforholdsundersøkelsen						
		I alt	1	2	3	4	5	6 rom eller flere
1	83	100	36	28	16	10	7	4
2	216	100	2	69	15	8	3	2
3	460	100	0	13	50	19	10	7
4	617	100	0	2	20	46	16	16
5	621	100	1	2	6	34	33	24
6 rom eller flere	714	100	0	1	2	11	24	61

[1] Det er bare tatt med de som sannsynligvis har svart for samme bolig i folke- og boligtellingen og boforholdsundersøkelsen. Det vil si at studenter og personer som har eller planlegger å flytte, er tatt ut av datasettet.

Et spørsmål som har vist seg å være vanskelig å svare på er: Hvor mange rom har du i boligen? Dette skulle en tro var ganske enkelt å svare på, men det er det ikke. Det er en svært spesiell definisjon av rom som det er ønske å få svar på. Nemlig hvor mange lovlige soverom og oppholdsrom (stuer) som finnes i boligen. Det vil si at ganger, bad, wc, vaskerom og kjøkken ikke skal telle med. For mange kan det være vanskelig å finne ut hvilke rom som skal telles med i denne definisjonen.
Høsten 2001 ble det spurt om antall rom både i boforholdsundersøkelsen og i folke- og boligtellingen. Det er til dels svært stort sprik mellom hva de samme personene[2] svarte på disse to undersøkelsene (Foss 2003b). Det var omtrent 40 prosent som svarte likt i disse to undersøkelsene. 37 prosent svarte ett rom mer eller mindre i disse to undersøkelsene. De resterende, altså nesten en fjerdedel, hadde større avvik enn ett rom.

Noen ønsker kanskje ikke å fortelle sannheten?

I noen tilfeller kan det være slik at enkelte ikke ønsker å gi de korrekte opplysningene i et intervju. I en testing av nye husholdningsspørsmål ble det gjennomført to intervjuer (Fosen et al 2001). Hver av partnerne fikk spørsmålet om hvem de bodde sammen med og hvilket forhold de hadde til disse personene. Det viste seg at partneren til 13 av 53, som hadde svart at de var samboere, svarte at de ikke var det. Dette kan tyde på at noen synes det er sensitivt å oppgi at de er samboer over telefon, eller at de har en annen forståelse av begrepet samboer enn partneren.

Systematisk målefeil – skjevhet

Det er ikke farlig hvis det i en undersøkelse er like mange som svarer for mye som for lite. Da vil gjennomsnittet bli det samme. Det farlige er hvis det er enten flere som svarer for mye enn for lite – eller omvendt. Da vil gjennomsnittet enten bli trukket opp eller ned. Det finnes da en systematisk målefeil i

Figur 3. Størrelsen på feilene når vi går fra utvalgsundersøkelse til fulltelling

Tabell 2. Antall husholdninger, etter størrelse. 2001

Husholdnings-størrelse	Folke- og bolig-tellingen 2001	Levekårsunder-søkelsen 2001	Inntekts- og formues-undersøkelsen 2001
I alt	1 962 000	2 116 000	2 080 000
1 person	740 000	845 000	867 000
2 personer	535 000	634 000	582 000
3 personer	269 000	243 000	243 000
4 personer	266 000	250 000	248 000
5 personer eller flere	151 000	144 000	139 000

undersøkelsen – en skjevhet. I en utvalgsundersøkelse vil denne skjevheten kunne ligge innenfor konfidensintervallet for estimatet. I en fulltelling eksisterer det ikke utvalgsusikkerhet, fordi det ikke er et utvalg. Derimot eksisterer den tilfeldige støyen og skjevheten. Den tilfeldige støyen er et uttrykk for den tilfeldige variasjonen i dataene og vil være svært liten fordi antallet enheter er svært stort. Skjevheten der-imot vil ikke endre seg når vi går fra utvalgsundersøkelser til fulltelling, fordi den ikke er avhengig av antall enheter. Den vil få en mye større betydning i fulltelling fordi de andre feiltypene er blitt svært små.

Seksjon for datafangstmetoder

I Statistisk sentralbyrå finnes det en egen seksjon som i hovedsak jobber med utforming av spørsmål og spørreskjema. Dette gjelder alle typer undersøkelser, det vil si papirskjema, telefonintervju og elektronisk rapportering. Gruppen består av ni personer. De tester ut spørsmålene og spørreskjemaene ved intervju av brukere. Intervjuteknikken som de bruker kalles kognitiv kartlegging. Ved slik testing kan feil og mangler i skjema bli oppdaget, og det er da mulig å prøve å finne en bedre løsning på spørsmål og spørreskjema.

Forskjellig definisjon – forskjellig resultat

Det er ofte det finnes forskjellige definisjoner på det samme begrepet. Dette vil føre til at enkelte statistikker ikke blir sammenlignbare. Dette gjelder for eksempel husholdningsdefinisjonen i utvalgsundersøkelser og i folke- og boligtellingen (Foss 2003a). I folke- og boligtellingen er det satt krav om at det bare er personer som er registrert bosatt i samme bolig (adresse) som kan inngå i en husholdning. Dette er i motsetning til utvalgsundersøkelser der det er opp til intervjuobjektet å definere hvem som er bosatt i boligen. Det er svært mange unge mennesker, spesielt studenter, som ikke bor der de er registrert. Dette vil føre til en forskjell i husholdningsstatistikken mellom utvalgundersøkelser og folke- og boligtellingen. Ingen av disse statistikkene er perfekte. I utvalgsundersøkelser kan man trekke et utvalg som ikke er rep-

resentativt, og det er ikke sikkert at intervjuobjektene svarer riktig på hvem som bor i boligen.

Det finnes ifølge levekårsundersøkelsen og inntekts- og formuesundersøkelsen omtrent 100 000 flere husholdninger enn i folke- og boligtellingen. I disse utvalgsundersøkelsene er det i hovedsak flere små husholdninger og færre store husholdninger. Store husholdninger i folke- og boligtellingen har blitt til to mindre husholdninger i utvalgsundersøkelsene. Som tidligere nevnt er grunnen til dette at unge mennesker, spesielt studenter, ikke har meldt flytting til Folkeregisteret.

Det ideelle

Det ideelle hadde vært om alle som skulle svare på undersøkelser fra Statistisk sentralbyrå gjorde det. Slik er ikke virkeligheten. Det vil alltid være noen som ikke kan svare på grunn av sykdom eller noe annet. I tillegg vil det være noen som ikke *vil* svare. Det kan også være feil i trekkegrunnlaget, det kan være personer som er trukket ut til å delta i undersøkelsen som har utvandret uten at dette er registrert. I tillegg ville det ideelle være at alle svarte korrekt på alle spørsmål. Det vil si at alle hadde den samme forståelsen av spørsmålet som Statistisk sentralbyrå og at de hadde en perfekt hukommelse, slik at de svarte riktig på alle spørsmål. Slik er ikke virkeligheten. Ingen statistikk er perfekt.

På Statistisk sentralbyrås internettside har alle statistikkfelt en egen side som heter "Om statistikken". På denne siden blir datakilder, feilkilder, kvalitet og mye annet som angår statistikkfeltet beskrevet.
De som lurer på kvaliteten på de forskjellige statistikkfeltene bør kikke på disse sidene. Statistisk sentralbyrå har i de siste årene satt inn store ressurser for å forbedre og effektivisere statistikkproduksjonen. Det har blitt utdannet 35 kvalitetsloser som er spesialtrent til å kvalitetssikre og effektivisere statistikkproduksjonen. Arbeidet med å heve kvaliteten og beskrive den vil fortsette i årene som kommer.

[1] Det kan også bli valgt andre grenser, for eksempel 67 eller 99 prosent konfidensintervall.

[2] Eller en i samme familie svarte.

Referanser

Fosen, Johan, Anne Gro Hustoft og Bengt Oscar Lagerstrøm (2001): Ny spørresekvens for å identifisere husholdninger i utvalgsundersøkelser. Notater 2001/25. Statistisk sentralbyrå.

Foss, Aslaug Hurlen (2003a): Kvaliteten i husholdningsdelen i Folke- og boligtellingen 2001. Notater 2003/78. Statistisk sentralbyrå.

Foss, Aslaug Hurlen (2003b): Kvaliteten i boligdelen i Folke- og boligtellingen 2001. Notater 2003/47. Statistisk sentralbyrå.

Foss, Aslaug Hurlen (2004): Kvaliteten i arbeidsmarkedsdelen i Folke- og boligtellingen 2001. Notater 2004/4. Statistisk sentralbyrå.