SAMFUNNSØKONOMISKE STUDIER

33

PRINSIPPER OG METODER FOR STATISTISK SENTRALBYRÅS UTVALGSUNDERSØKELSER

SAMPLING METHODS APPLIED BY THE CENTRAL BUREAU OF STATISTICS OF NORWAY

STATISTISK SENTRALBYRÅ
CENTRAL BUREAU OF STATISTICS OF NORWAY
OSLO 1977

PRINSIPPER OG METODER FOR STATISTISK SENTRALBYRÅS UTVALGSUNDERSØKELSER

SAMFUNNSØKONOMISKE STUDIER NR. 33

PRINSIPPER OG METODER FOR STATISTISK SENTRALBYRÅS UTVALGSUNDERSØKELSER

SAMPLING METHODS APPLIED BY THE CENTRAL BUREAU OF STATISTICS OF NORWAY

STATISTISK SENTRALBYRÅ
CENTRAL BUREAU OF STATISTICS OF NORWAY
OSLO 1977

FORORD

I 1967 ble det opprettet en egen intervjuorganisasjon i Statistisk Sentralbyrå. Denne har siden hatt ansvaret for planleggingen og gjennomføringen av de fleste utvalgsundersøkelser, kanskje særlig undersøkelser hvor trekkeenhetene er personer og husholdninger.

I denne publikasjonen beskrives de grunnleggende trekk ved utvalgsmetoden som er brukt siden 1975. I tillegg blir det gjort rede for de viktigste prinsipper som utvalgsplanen bygger på. Beskrivelsen er stort sett gjort uten bruk av matematiske symboler. Da innsamling og analyse av data fra utvalgsundersøkelser må bygge på metoder fra teoretisk statistikk, er noe av det statistiske grunnlaget gitt i vedlegget.

Publikasjonen er skrevet av forsker Ib Thomsen.

S t a t i s t i s k S e n t r a l b y r å, Oslo, 27. september 1977

Odd Aukrust

PREFACE

In 1967 the Central Bureau of Statistics of Norway established an interview survey division which has planned and carried out a number of surveys. In this publication the basic sample plan is described together with the main principles it is based upon. The description is given without extensive use of mathematical symbols. The collection and analysis of survey data must, however, be based on methods of mathematical statistics, and therefore the appendix provides some of the basic mathematical statistics used in such cases.

The publication has been written by Mr. Ib Thomsen.

Central Bureau of Statistics, Oslo, 27 September 1977

Odd Aukrust

INNHOLD

S t a n d a r d t e g n

0,0 Mindre enn en halv av den brukte enhet

CONTENTS

Explanation of Symbols

0.0 Less than half of unit employed

I. INNLEDNING

Helt siden daværende direktør i Statistisk Sentralbyrå, A.N. Kiær, i 1891 utførte en av de første utvalgsundersøkelser i verden, har Byrået mer eller mindre regelmessig gjennomført en rekke slike undersøkelser. I 1967 ble det opprettet en egen intervjuorganisasjon, som i løpet av de siste 10 årene har stått for planleggingen og gjennomføringen av de fleste utvalgsundersøkelser utført av Byrået. Utvalgsplanen som ble brukt fram til 1975 er beskrevet i Tamsfoss (1970).

På grunnlag av data fra Folke- og boligtellingen i 1970 ble det i 1975 laget en ny utvalgsplan som på flere punkter avviker vesentlig fra den tidligere, først og fremst på grunn av at det oppstod behov for et større antall intervjuere enn det en startet med. I denne publikasjonen er gitt en relativt detaljert beskrivelse av utvalgsplanen. I tillegg er det gitt en redegjørelse for de viktigste prinsipper som utvalgsplanen bygger på. Det har derfor vært nødvendig å gi en beskrivelse av begreper og metoder som bare indirekte er knyttet til utvalgsplanen.

Metodene er felles for alle typer statistiske utvalg, men i denne publikasjonen er det lagt særlig vekt på de metoder som brukes i forbindelse med utvalg av personer og husholdninger.

Kapittel II gir en enkel oversikt over de vanligste utvalgsmetoder, samt de viktigste begreper felles for alle sannsynlighetsutvalg, og tjener som innledning til kapitlene III og IV, hvor Byråets generelle utvalgsplan og de vanligste estimeringsmetoder er beskrevet. Utvalgsplanen er i mange henseende identisk med utvalgsplanene til andre liknende institusjoner rundt om i verden. Bare når det gjelder trekkingen fra registrene har Byrået måttet finne egne løsninger. Disse løsninger er beskrevet i avsnitt III.5.

I kapittel V er beskrevet noen metoder for hvordan en kan bestemme hvor stort utvalget må være for å tilfredsstille oppsatte krav til de resultater en ønsker.

Enhver statistisk undersøkelse er beheftet med feil som oppstår under innsamlingen og klargjøringen av data. I Byrået brukes en del metoder for å identifisere og kontrollere slike feilkilder. Disse metodene er beskrevet i kapittel VI.

Kapittel VII gir en oversikt over de forskjellige måter resultater fra utvalgsundersøkelsene publiseres på, samt hvilke regler som gjelder for å gjøre rede for de usikkerheter som er knyttet til resultatene.

I de siste årene har en i en viss utstrekning i Byrået gått over til å bruke mindre løpende undersøkelser, i stedet for større, mindre hyppige undersøkelser. Dette har reist en serie nye metodeproblemer, hvorav mange ennå er uløste. Måten noen av disse problemene er blitt behandlet

på er beskrevet i kapittel VIII.

Hensikten med denne publikasjonen er å gi en ikke altfor teknisk beskrivelse av Byråets utvalgsmetoder beregnet på alle de som kommer i kontakt med data samlet inn av Byråets intervjuorganisasjon. Framstillingen er derfor gjort uten bruk av matematiske formler. I vedlegg 1 er gitt en mer nøyaktig beskrivelse av metodene ved hjelp av begreper fra den matematiske statistikk. Deler av vedlegget er skrevet av konsulent Petter Laake.

I denne publikasjonen er det bare gitt få henvisninger til den enorme metodelitteraturen som finnes. Slike henvisninger kan en finne i boka til Moser og Kalton (1973). Derimot er det lagt vekt på å presentere det meste av det omfattende metodearbeid som de siste årene er utført i Byrået.

II. FORSKJELLIGE MÅTER Å TREKKE UTVALG PÅ

1. Innledning

En kan trekke utvalg fra en befolkning på mange måter. Noen av metodene er meget enkle, mens andre er karakterisert ved at det legges ned mye arbeid ved trekking av utvalget. Uansett hvordan utvalgene er trukket, vil det være et problem å avgjøre i hvilken grad resultatene i utvalget likner på de resultatene en ville ha fått dersom en hadde undersøkt hele befolkningen. Felles for alle utvalg som trekkes i Byrået er at de enheter som trekkes, enten det er personer, husholdninger eller bedrifter, alltid trekkes på en slik måte at enhver enhet har en kjent sannsynlighet for å komme med i utvalget. Slike utvalg kalles sannsynlighetsutvalg.

Den viktigste årsak til at en slik trekkemetode er valgt er at en på denne måten oppnår kontroll med kvaliteten av resultatene fra en undersøkelse. I slike tilfeller gir den matematiske statistikk en rekke metoder til vurdering av kvaliteten. Det har vist seg at disse teoretiske resultater stemmer godt overens med det en har konstatert i praksis.

Denne måten å trekke utvalg på medfører blant annet at hvis en person blir valgt ut for å delta i en undersøkelse, kan ikke denne personen byttes ut med en annen, f.eks. naboen. Denne utvalgsmetoden, som ikke er brukt av alle institusjoner som foretar utvalgsundersøkelser, fører naturligvis også til at det blir dyrere å samle inn data, enn hvis en lot intervjueren erstatte en uttrukken person med en annen som det er lettere å få tak i.

I dette kapitlet skal noen begreper som er sentrale i den videre framstilling defineres, og det skal gis en kort oversikt over de mest brukte måter å trekke utvalg på.

2. Viktige begreper felles for alle sannsynlighetsutvalg

Målepopulasjon og kjennemerkeverdi

I utvalgsteorien tenker en seg at det er gitt en bestemt endelig populasjon, f.eks. populasjonen av norske husholdninger. Til hvert element i populasjonen er det knyttet ett eller flere kjennemerker som en ønsker å studere. For eksempel ønsker en å estimere de samlede utgifter til mat i den norske befolkning. For å gjøre dette tas et utvalg fra populasjonen, og på grunnlag av resultatene i utvalget, estimeres det resultatet en ville ha fått dersom alle elementene i populasjonen hadde blitt undersøkt.

Utvalgsteorien er opptatt med å besvare følgende to spørsmål:

1. Hvordan skal en trekke utvalget?
2. Hvordan skal en på grunnlag av resultatet i utvalget kunne si noe om det en ville ha fått dersom alle elementer i populasjonen var blitt undersøkt?

Svaret er at hvis vi trekker ut de enheter som skal være med i utvalget på en slik måte at enhetene i populasjonen har kjente sannsynligheter for å komme med i utvalget, så kan vår utvalgsmetode beskrives ved en stokastisk modell. Numrene på de uttrukne enhetene oppfattes som stokastiske variable, og de størrelser i populasjonen vi vil vite noe om er parametre i sannsynlighetsfordelingen for disse kjennemerker. Ved å bruke metoder fra den teoretiske statistikken kan vi da estimere parametrene, og eventuelt teste hypoteser.

Utvalgsforventning og utvalgsvarians

For å gi en enkel beskrivelse av de sentrale begreper i teorien for statistiske utvalg, skal det gis et lite eksempel. La populasjonen bestå av fire husholdninger. Husholdningene tenkes nummerert slik at husholdning nr. 1 har en inntekt på kr 10 000, husholdning nr. 2 har kr 15 000, husholdning nr. 3 har kr 20 000 og husholdning nr. 4 har kr 30 000 i inntekt. Vi ønsker nå på grunnlag av et utvalg på to husholdninger å anslå den gjennomsnittlige inntekt i populasjonen av fire husholdninger. De to husholdninger som skal være med i utvalget, blir trukket tilfeldig blant de fire husholdninger.

Anta at vi trekker husholdningene med kr 10 000 og kr 20 000 i inntekt. Gjennomsnittet i dette utvalget er kr 15 000, og et naturlig anslag på gjennomsnittlig inntekt i hele populasjonen er derfor kr 15 000. Hvis vi derimot hadde trukket husholdningene med kr 20 000 og kr 30 000 i inntekt, ville den gjennomsnittlige inntekt i utvalget vært kr 25 000, og et naturlig anslag for samme tall i populasjonen ville også ha vært kr 25 000.

I dette tilfelle er den gjennomsnittlige inntekt i populasjonen kr 18 750, og ingen av de to trukne utvalg gir derfor et korrekt anslag på den gjennomsnittlige inntekt. Dette er typisk for de fleste utvalg. I tabell 1 nedenfor er listet alle seks mulige utvalg på to husholdninger av de fire husholdninger. En ser her at ingen av anslagene er nøyaktig lik den gjennomsnittlige inntekt i populasjonen. Hvis en derimot ser på gjennomsnittet av de seks mulige anslag, vil en se at dette gjennomsnitt er nøyaktig lik kr 18 750.

Tabell 1. Mulige utvalg ved enkel tilfeldig trekking av to husholdninger fra populasjonen på fire husholdninger

Mulige utvalg husholdningsnr.	Gjennomsnittlig inntekt i utvalget. Kr
1 og 2	12 500
1 og 3	15 000
1 og 4	20 000
2 og 3	17 500
2 og 4	22 500
3 og 4	25 000

Den regnemetode en anvender på kjennemerkeverdiene i utvalget for å anslå gjennomsnittet i hele populasjonen, kalles en estimator. I eksemplet ovenfor er estimatoren gjennomsnittlig inntekt i det trukne utvalg. I tabell 1 er listet de seks mulige verdier denne estimator kan ta avhengig av hvilket utvalg som blir trukket. En estimator som har den egenskap at gjennomsnittet av alle dens mulige verdier er lik gjennomsittet i populasjonen, sies å være utvalgsforventningsrett*), i det følgende kalt forventningsrett. I eksemplet ovenfor ses det at gjennomsnittet i et trukket utvalg er en forventningsrett estimator for gjennomsnittet i hele populasjonen.

I eksemplet ovenfor viser det seg at på tross av at estimatoren er forventningsrett, er ingen av dens mulige verdier nøyaktig lik den gjennomsnittlige inntekt i populasjonen. En kan derfor spørre hvorfor det er så viktig å trekke et tilfeldig utvalg, når en likevel ikke har noen garanti for at den estimatoren en bruker, gir anslag som ligger nær den verdi man

*) Dette utsagn er bare riktig dersom alle utvalg har samme sannsynlighet for å bli trukket. For en mer nøyaktig gjennomgåelse av begrepene i dette avsnittet henvises interesserte lesere til vedlegg 1.

ønsker å anslå. Det som mangler er et mål for hvor store avvik en i gjennomsnitt kan vente seg fra den riktige verdi. Det mest alminnelig brukte mål for spredningen i de mulige verdier, er utvalgsvariansen, i det følgende kalt variansen. Variansen i vårt eksempel regnes ut på følgende måte:

$$\{[(12\ 500-18\ 750)^2+(15\ 000-18\ 750)^2+(20\ 000-18\ 750)^2+ (17\ 500-18\ 750)^2+(22\ 500-18\ 750)^2+(25\ 000-18\ 750)^2]/6\}$$

$$= 18\ 229\ 166.$$

Et annet mål for spredningen er standardavviket, som er lik kvadratroten av variansen. I eksemplet ovenfor er standardavviket lik 4 269,6.

I praksis har en bare tilgjengelig resultatene fra et enkelt utvalg, og en kan derfor ikke på grunnlag av tilgjengelige data regne ut variansen slik som det er gjort ovenfor. I den teoretiske statistikk er det imidlertid vist at en kan lage en forventningsrett estimator for variansen på grunnlag av resultatene i utvalget. Se vedlegg 1. Intuitivt venter en at utvalgsvariansen går ned når en øker størrelsen på utvalget. Hvis vi i eksemplet ovenfor trekker tre husholdninger, er de mulige utvalg med tilhørende gjennomsnitt vist i tabell 2. Som en kunne vente seg er utvalgsvariansen mindre når en trekker tre husholdninger enn når en trekker to.

Tabell 2. Mulige utvalg ved enkel tilfeldig trekking av tre husholdninger

Mulige utvalg av husholdninger	Gjennomsnittlig inntekt i utvalget. Kr
1, 2 og 3	15 000
1, 2 og 4	18 333
1, 3 og 4	20 000
2, 3 og 4	21 667

Variansen viser seg å bli 6 077 443. Sammenhengen mellom utvalgsvarians og utvalgsstørrelsen er nærmere beskrevet i vedlegg 1.

I praksis ville en naturligvis ikke trekke utvalg av populasjoner som bare består av fire husholdninger slik som i det nevnte eksempel. Oftest er populasjonene meget store, f.eks. alle husholdninger i Norge, e.l. Utvalgene er vanligvis også mye større enn to. Likevel er utvalgsforventning og utvalgsvarians definert på akkurat samme måten som i eksemplet, uansett størrelsen på populasjonen og utvalget.

Det er i tilfeller med store populasjoner og utvalg en har virkelig gevinst ved å bruke sannsynlighetsutvalg. Det kan nemlig vises at

en på grunnlag av resultatene i utvalget kan konstruere et intervall, som med stor sannsynlighet dekker det sanne populasjonsgjennomsnitt. Hvis en f.eks. på grunnlag av et enkelt tilfeldig utvalg på 1 000 personer har estimert at 35 prosent vil stemme på et bestemt parti, er det naturligvis liten sjanse for at nøyaktig 35 prosent i populasjonen vil gjøre det. Men en kan si at intervallet mellom 32 og 38 prosent med stor sannsynlighet dekker den nøyaktige prosentandel i populasjonen som vil stemme på partiet. Ved å øke utvalgsstørrelsen kan en gjøre dette intervallet så smalt som en ønsker. Intervaller av denne typen kalles konfidensintervaller.

I det følgende skal det gis en beskrivelse av de mest anvendte metoder for trekking av utvalg fra en endelig populasjon.

3. Oversikt over de vanligste typer statistiske utvalg

Følgende tre krav stilles vanligvis til måten å trekke statistiske utvalg på:

1. Den bør inneholde en klar avgrensing av den populasjonen en ønsker å estimere parametre i.
2. Utvalgstrekkingen må være slik at en kan beregne estimatorer med kjente statistiske egenskaper.
3. Utvalgsplanen bør være økonomisk og praktisk.

For å unngå at utgiftene til en undersøkelse blir uforholdsmessig høye, kan det i praksis ofte bli nødvendig å fire på kravet om at estimatorene skal ha kjente statistiske egenskaper. Ved gjennomføringen av en undersøkelse kan en oppleve at det er så vanskelig å treffe en bestemt uttrukken person, at en velger å regne denne personen som frafall i stedet for å fortsette med å oppsøke ham. (Se nærmere om frafall i kapittel VI.) Det kan nevnes mange andre eksempler på hvorledes man i praksis må modifisere kravet om at en bør ha estimatorer med kjente statistiske egenskaper for å gjennomføre undersøkelsen. I dette og det etterfølgende kapitlet skal en imidlertid se bort fra slike vanskeligheter, og ta dem opp i kapittel VI. Først ser en på en meget enkel måte å trekke sannsynlighetsutvalg på, og deretter modifiseres denne, for å gjøre den mer økonomisk og praktisk anvendelig.

Enkel lotterisk trekking

La det være gitt en populasjon med N enheter, hvorfra en ønsker å gjøre et utvalg på n enheter. Dersom de n enheter blir trukket på en slik måte at alle $\binom{N}{n}$ mulige utvalg har samme sannsynlighet for å bli valgt ut, sier en at utvalget er enkelt tilfeldig trukket. Det er antakeligvis sjeldent at en i praksis trekker et enkelt tilfeldig utvalg, men det er likevel vanlig å behandle denne trekkemetoden inngående.

Det kan vises at utvalgsgjennomsnittet i slike utvalg er en forventningsrett estimator for gjennomsnittet i populasjonen. Under forutsetning av at populasjonsstørrelsen N er mye større enn utvalgsstørrelsen n er det gjennomsnittlige kvadratavvik i utvalget en forventningsrett estimator for variansen.

Årsakene til at en i praksis sjelden trekker et enkelt tilfeldig utvalg er mange, her skal vi nevne tre. Den første er at en med relativt enkle metoder kan trekke et utvalg som med samme utvalgsstørrelse vil gi mindre utvalgsvarians enn det en oppnår med et enkelt tilfeldig utvalg. Den andre er at en for å kunne trekke et enkelt tilfeldig utvalg må ha adgang til et register, som inneholder alle enhetene i den aktuelle populasjon. For det tredje vil enkel tilfeldig trekking føre til store reisekostnader når data samles inn av intervjuere, noe som ofte er tilfellet i Byrået. Nedenfor skal det gis eksempler på hvorledes en i praksis kan trekke utvalg, som forenkler innsamlingen av data.

Det er viktig å merke seg at det er trekkemetoden som bestemmer om et utvalg er enkelt tilfeldig eller ikke. Det er altså ikke mulig å avgjøre om et utvalg er trukket enkelt tilfeldig ved å studere ett bestemt trukket utvalg. Et enkelt tilfeldig utvalg kan vise seg å være lite representativt for populasjonen. For eksempel kan et utvalg av personer bosatt i Norge inneholde 70 prosent kvinner, og likevel være trukket enkelt tilfeldig. Sjansen for å trekke et slikt utvalg er imidlertid meget liten.

Stratifisering

Det er tidligere demonstrert hvordan en ved å øke utvalgets størrelse kan minske usikkerheten på estimatorene. Dette er imidlertid ikke den eneste måte, og i dette avsnittet skal det vises hvorledes en ved å inndele populasjonen i delpopulasjoner, såkalte strata, og deretter trekke enkelte tilfeldige utvalg innen hvert stratum, kan minske usikkerheten på estimatorene uten å øke utvalgsstørrelsen. La oss se på eksemplet med de fire husholdninger fra forrige avsnitt. En hadde her fire husholdninger med inntektene kr 10 000, kr 15 000, kr 20 000 og kr 30 000, henholdsvis. I stedet for å trekke et tilfeldig utvalg av to husholdninger blant de fire, deles nå populasjonen i to strata. Første stratum består av husholdningene med inntektene kr 10 000 og kr 15 000, mens det andre stratum består av husholdningene med kr 20 000 og kr 30 000 i inntekt. Innen hvert av de to strata trekkes én husholdning tilfeldig. Dette gir fire mulige utvalg. Som estimator for den gjennomsnittlige inntekt i hele populasjonen, brukes igjen utvalgsgjennomsnittet. De fire mulige verdier som estimatoren antar er kr 15 000, kr 20 000, kr 17 500 og kr 22 500.

Igjen ser en at gjennomsnittet av de fire mulige verdier er lik den gjennomsnittlige inntekt i populasjonen, nemlig kr 18 750. Dessuten er det lett å se at de fire mulige verdier i gjennomsnitt ligger nærmere denne riktige verdi enn tilfellet var da vi trakk et enkelt tilfeldig utvalg på to husholdninger. Denne trekkemetoden fører altså til mindre samplingsvarians enn estimatoren i avsnittet ovenfor. Det er heller ikke vanskelig å se hvorfor dette er tilfelle. Ved å stratifisere populasjonen før trekking, forhindres at utvalget skal bestå av to husholdninger med små inntekter, eller to husholdninger med store inntekter. Et stratifisert utvalg vil alltid bestå av én husholdning med liten inntekt og én husholdning med stor inntekt.

I vårt eksempel er det like mange enheter i begge strata, og dessuten hadde hver enhet samme trekkesannsynlighet i de to strata. I praksis er dette ikke nødvendig. Planleggeren av en undersøkelse er helt fri til å stratifisere sin populasjon nøyaktig som han ønsker det, og han kan velge forskjellige utvalgssannsynligheter i de forskjellige strata etter det formål han har med undersøkelsen. I praksis er det derfor viktig at en konstruerer strata hensiktsmessig og allokerer det samlede utvalg over strataene på en fornuftig måte.

En bør merke seg at formålet med stratifiseringen ikke alltid er å redusere variansen. Dersom formålet med en utvalgsundersøkelse er å gi estimatorer for små delgrupper i populasjonen, kan det være nødvendig å definere disse små delpopulasjoner som egne strata og deretter overrepresentere disse i utvalget. I neste kapittel skal en gi en beskrivelse av Byråets utvalgsplan, og en vil her komme inn på et tredje formål med stratifisering, som er spesielt viktig når utvalget trekkes i to trinn.

Når en skal trekke et stratifisert utvalg, er det vesentlig tre ting en må avgjøre: For det første hvilke kjennemerker en vil stratifisere etter, for det andre hvordan en skal konstruere strataene, og for det tredje hvordan det samlede utvalg skal allokeres på strataene.

Den variansreduksjon en kan oppnå ved en stratifisering, varierer mye fra den ene type undersøkelse til den andre. Dersom populasjonen en studerer er inhomogen, kan den stratifiseringsmetode en bruker være helt avgjørende for kvaliteten til utvalgsundersøkelsene. Hvis en f.eks. ønsker å estimere tall for populasjonen av alle norske bedrifter, er dette en så inhomogen masse at en kan oppnå betydelige reduksjoner i utvalgsvariansen ved hensiktsmessig stratifisering. Når det gjelder mange landsomfattende undersøkelser av personer, får stratifiseringen ofte en litt annen betydning enn tilfellet er ved de inhomogene populasjoner. Det kanskje viktigste formål med stratifisering i slike tilfelle er å sikre representativitet av utvalget ikke bare på hele populasjonen, men også på et mindre antall regioner. Se også diskusjonen om stratifisering i neste kapittel.

Valg av antall observasjoner innen strataene

I litteraturen om utvalgsundersøkelser finner en ofte beskrevet to måter en kan allokere det totale utvalg på strataene. Den ene måten er proporsjonal allokering, dvs. at antall observasjoner fra et stratum er proporsjonal med antall enheter i stratumet. Den andre metoden kalles optimal allokering, dvs. at en allokerer utvalget med sikte på å få minst mulig utvalgsvarians på de endelige resultater. En sammenlikning av variansgevinstene ved disse to allokeringsmetoder er gjort i Thomsen (1976).

Igjen bør en være oppmerksom på at andre allokeringsmetoder er mulig dersom formålet med stratifiseringen tilsier det. Dersom en f.eks. er opptatt av å gi tall for små deler av populasjonen i tillegg til tall for hele populasjonen, bør en overveie å oversample de deler av populasjonen en ønsker tall for i tillegg til tall for hele populasjonen. Hvis en f.eks. ved en landsomfattende undersøkelse ønsker å gi tall for Finnmark fylke, må en i de fleste praktiske tilfelle definere Finnmark som eget stratum, og trekke relativt mange enheter innen dette stratum.

Valg av kjennetegn en ønsker å stratifisere etter

Når en skal stratifisere en populasjon, bør en stratifisere etter kjennemerker som er høyt korrelert med den eller de variable en ønsker å studere i undersøkelsen. En rekke praktiske og teoretiske studier av effekten av stratifisering, antyder at det sjelden lønner seg å lage mer enn mellom 5 og 10 strata etter et bestemt kjennemerke. Dersom en på forhånd har opplysninger om flere kjennemerker som er høyt korrelert med de variable som studeres i undersøkelsen, lønner det seg å trekke flere kjennemerker inn ved stratifiseringen. I slike tilfelle ligger det fornuftige antall strata vesentlig høyere enn tilfellet er ved stratifisering etter ett kjennemerke. Thomsen (1977).

Klyngeutvalg og utvalg i flere trinn

De utvalgsmetoder som er nevnt til nå forutsetter at det finnes et register over alle enheter i populasjonen. Opprettelsen av slike registre vil i mange tilfelle være umulige eller meget kostbar, og det er derfor utviklet utvalgsmetoder som tar sikte på å redusere arbeidet med opprettelse av registre. Hvis en f.eks. skal trekke et landsomfattende utvalg av husholdninger i Norge, kan dette gjøres ved at man først trekker et utvalg av kommuner, og deretter trekker utvalg av husholdninger innen de uttrukne kommuner. På denne måten trenger en bare et register over de

husholdninger som finnes i de uttrukne kommuner. I forbindelse med intervjuundersøkelser har denne metoden også den fordel, at den reduserer innsamlingskostnadene vesentlig. Dette gjøres ved at en med sjeldne mellomrom, f.eks. hvert tiende år, trekker et utvalg av kommuner. De utvalg som skal brukes i forbindelse med intervjuundersøkelser, trekkes deretter innen de uttrukne kommuner. Ved å ansette en eller flere intervjuere innen de uttrukne kommuner reduserer en det geografiske området som en bestemt intervjuer må dekke, og reduserer derved reisekostnadene. En inndeling av populasjonen i innbyrdes ekskluderende grupper, kalles en klyngeinndeling. Selve gruppene kalles klynger (på engelsk "cluster"). Ovenfor er klyngene definert som kommunene i landet.

Som tilfellet var under stratifisering, står planleggeren fritt når det gjelder å velge hvorledes han ønsker å inndele populasjonen i klynger, og hvilke metoder han vil bruke under trekkingen av klynger og av enheter innen de uttrukne klynger. I noen tilfeller velger en å ta alle enheter i de uttrukne klynger.

Klyngeutvalg

Når en tar med i utvalget alle enhetene i en uttrukken klynge, kalles utvalget klyngeutvalg. Denne måten å trekke utvalg på er et viktig hjelpemiddel for planleggeren til å redusere kostnadene til en undersøkelse. I Byrået brukes klyngeutvalg i forbindelse med arbeidskraftundersøkelsene, hvor en trekker et utvalg av husholdninger, og deretter intervjuer samtlige medlemmer av husholdningen. På denne måten reduseres reisekostnadene vesentlig. En slik utvalgsmetode er det vanlig å bruke når undersøkelsen omfatter hovedsakelig fakta-spørsmål. I en holdningsundersøkelse er derimot denne utvalgsmetoden mindre heldig, da en ofte vil finne at holdningene innen en bestemt husholdning er korrelerte med hverandre. Dette fører ikke til noen typer skjevheter i utvalget, men det er lite økonomisk å foreta flere observasjoner innen en husholdning når en vet at observasjonene er høyt korrelerte. Også i forbindelse med fakta-spørsmål vil en ofte finne en viss korrelasjon innen husholdningene, men vanligvis er denne så liten at en kan se bort fra den.

For å se på konsekvensene på utvalgsvariansen av denne trekkemetoden, skal vi igjen ta for oss eksemplet med fire husholdninger.

La de fire husholdninger være inndelt i to klynger, der en klynge består av husholdningene med inntektene kr 10 000 og kr 15 000, og den andre klyngen av husholdningene med inntektene kr 20 000 og kr 30 000. For å trekke et utvalg på to husholdninger, trekkes nå en klynge tilfeldig blant

de to. Dette utvalg er et sannsynlighetsutvalg. Som estimator for gjennomsnittlig inntekt brukes fortsatt utvalgsgjennomsnittet. I dette tilfelle er det to mulige utvalg med de tilsvarende gjennomsnitt kr 12 500 og kr 25 000. Igjen ses det at gjennomsnittet av de to gjennomsnitt er lik gjennomsnittlig inntekt for alle fire husholdninger, men det ses også at den gjennomsnittlige avstand mellom utvalgsgjennomsnittene og det riktige gjennomsnitt er større nå enn tilfellet var ved stratifiserte utvalg og enkelt tilfeldig utvalg. Det er typisk at klyngeutvalg fører til større utvalgsvarians. Når en likevel i praktiske tilfelle bruker klyngeutvalg, skyldes dette at en ønsker å redusere kostnadene.

Det er verd å merke seg at måten en konstruerer klyngene på har innflytelse på utvalgsvariansen. Hvis en f.eks. hadde valgt den ene klyngen til å bestå av husholdningene med inntektene kr 10 000 og kr 30 000 i en klynge og husholdningene med inntektene kr 15 000 og kr 20 000 i den andre klyngen, ville utvalgsvariansen reduseres vesentlig. Ved den siste inndeling i klynger satte en sammen husholdninger på en slik måte at de var mest mulig heterogene. Også dette er et typisk trekk ved klyngeutvalg. Når en konstruerer klynger, bør en derfor forsøke å få disse så heterogene som mulig.

Utvalg i to eller flere trinn

Dersom en trekker et utvalg av enhetene i en uttrukket klynge, kalles utvalget totrinns-utvalg. Slike utvalg er meget brukt når en skal trekke landsomfattende utvalg. I forhold til klyngeutvalg representerer denne utvalgstrekking bare ett nytt problem, nemlig valg av trekkesannsynlighet innen hver klynge. Den vanligste måte å velge trekkesannsynlighet på kan beskrives på følgende måte: La f_1 betegne sannsynligheten for at en klynge blir trukket i første trinn. La f_2 være trekkesannsynligheten en bruker innen den samme klyngen. En velger da f_1 og f_2 på en slik måte at produktet $f_1 \cdot f_2$ er konstant for alle klynger i populasjonen. Slike utvalg kalles selvveiende. Med denne utvalgsmetoden oppnår en at alle elementer i populasjonen vil få samme trekkesannsynlighet. Sannsynligheten for at en person blir trukket ut i et utvalg, er nemlig lik sannsynligheten for at den klyngen enheten tilhører blir trukket ut i første trinn, multiplisert med sannsynligheten for at personen blir valgt ut innen den klyngen vedkommende tilhører. De klynger som velges ut i første trinn kalles primære utvalgsområder.

Forskjellig måte å kombinere utvalgsmetoder på

I praksis består de fleste utvalgsplaner av at en kombinerer de fire metoder som er nevnt ovenfor. Den utvalgsmetode som brukes i Byrået

ved trekking av landsomfattende utvalg av husholdninger og personer, er en kombinasjon av alle fire utvalgsmetoder. Landet er først delt i klynger, som stort sett består av kommuner. Deretter er disse kommuner stratifisert etter størrelse og kommunetype. Innen hvert stratum trekkes en kommune som representant for alle de andre kommuner i samme stratum. Endelig trekkes personer og husholdninger enkelt tilfeldig innen den uttrukne kommunen. I neste kapittel skal vi gi en detaljert beskrivelse av den utvalgsmetode som brukes i Byrået til trekking av landsomfattende utvalg av personer og husholdninger.

III. BESKRIVELSE AV BYRÅETS GENERELLE UTVALGSPLAN

1. Innledning

Som nevnt i kapittel II, kan utvalg trekkes på mange måter. I Byråets generelle utvalgsplan har en kombinert de forskjellige metoder for å finne en løsning som er økonomisk og praktisk, samtidig som utvalget er et sannsynlighetsutvalg med de fordeler dette gir. Før en velger utvalgsplan er det nødvendig å sette opp en del grunnleggende forutsetninger, basert på kjennskap til hva en kan vente seg utvalgsplanen skal brukes til, samt de ressursrammer en må regne med å arbeide innenfor. De viktigste forutsetninger i forbindelse med Byråets utvalgsplan er følgende:

1) Utvalgsplanen skal gi grunnlag for undersøkelser av varierende art.
2) Antall intervjuere som brukes i forbindelse med en enkelt undersøkelse skal kunne variere mellom 100 og 350.
3) Primært tas det sikte på å kunne publisere tall for hele landet, men i forbindelse med større undersøkelser bør utvalget allokeres slik at en kan gi tall for visse geografiske områder i tillegg.
4) Utvalgene vil oftest være selvveiende.

2. Valg av primære utvalgsområder

Ut fra erfaringene gjort i Byrået, er det liten tvil om at langt de fleste utvalg med fordel kan trekkes i to trinn. Gitt at utvalget skal trekkes i to trinn, er størrelsen på de primære utvalgsområder, samt fordelingen av intervjuobjektene på intervjuerne i de primære utvalgsområdene, de viktigste valg ved konstruksjonen av utvalgsplanen. Ut fra ønsket om å få minst mulig varians, bør en velge mange små utvalgsområder. I praksis vil dette si at en med 350 intervjuere kan velge 350 primære utvalgsområder, og allokere en intervjuer pr. område. Ut fra et økonomisk administrativt

synspunkt er det derimot en fordel å konsentrere intervjuerne så mye som mulig. I prinsippet kan disse to krav avveies mot hverandre ved å se på hvordan kostnader og varianser varierer med forskjellige valg av utvalgsområder. Slike studier er gjort, men de resultater som kom fram om kostnader var av en slik art at en meget vanskelig kunne trekke sikre konklusjoner om kostnadene. En fant likevel en svak sammenheng mellom reisekostnader og gjennomsnittlig flateinnhold til utvalgsområdene.

Et tredje moment som spiller en rolle ved valg av utvalgsområder, er hvilke registre som er tilgjengelige, samt hvilke opplysninger disse inneholder. De geografiske områder som skal danne grunnlaget for et utvalgsområde, skal kunne identifiseres i et tilgjengelig register.

Ut fra en samlet vurdering kom en fram til å velge kommuner som utvalgsområde. Kommuner som hadde færre enn 3 000 innbyggere ved folketellingen i 1970 ble slått sammen med andre kommuner, slik at det er minst 3 000 personer innen hvert utvalgsområde. I gjennomsnitt regner vi med å trenge 3-4 intervjuere pr. utvalgsområde. Med et samlet antall intervjuere på ca. 350 vil det si at en måtte trekke 100 utvalgsområder i første trinn.

I forbindelse med mindre undersøkelser er det behov for å kunne utføre undersøkelsen uten å bruke samtlige intervjuere. Dette problem er løst ved å inndele hvert utvalgsområde i et passende antall mindre områder, som hver kan dekkes av en intervjuer, og som representerer hele utvalgsområdet. Ved mindre undersøkelser trekkes derfor utvalget i tre trinn. I første trinn trekkes et utvalgsområde, i andre trinnet trekkes en del av det utvalgte område, og til slutt trekkes enhetene til undersøkelsen. En beskrivelse av oppdelingen av utvalgsområdene er gitt i avsnitt III.6.

3. Stratifisering av de primære utvalgsområder

I stedet for å trekke utvalgsområdene tilfeldig fra hele landet, kan en med fordel stratifisere utvalgsområdene, og foreta uavhengig trekking innen hvert stratum. Formålene med stratifiseringen er to:

1) Å redusere utvalgsvariansen.
2) Å bli i stand til å lage regionale tall.

Dersom en ønsker å redusere variansen ved hjelp av stratifisering, bør strataene gjøres så homogene som mulig. Etter som denne utvalgsplanen skal brukes i forskjellige typer av undersøkelser, vil det være en fordel å finne stratifikasjonsvariable som har relevans for flest mulig av de undersøkelser som blir foretatt. De to mest brukte variable er geografisk beliggenhet og næringsstruktur.

Når det gjelder det andre formålet med stratifiseringen, nemlig ønsket om å kunne publisere regionale tall, må en ta andre hensyn ved

stratifiseringen. Et stratum er nemlig den minste geografiske enhet utvalget er representativt for, og derfor den minste enhet en kan gi tall for. Hvis vi ønsker å kunne gi tall for bestemte områder, bør stratumgrensene ikke krysse grensene for disse områder. Før stratifiseringen foretas, bør en derfor få klarhet over hvilke områder en ønsker tall for. Under planleggingen av utvalgsplanen kom det fram ønsker om å kunne spesifisere utvalget geografisk. Ønskene kan klassifiseres i fire typer:

Type A. En ønsker primært tall for hele landet, samt tall for et større antall geografiske områder (f.eks. fylke, grupper av fylker).

Type B. En ønsker primært tall for hele landet og i tillegg tall for et mindre antall geografiske områder (f.eks. handelsfelt, landsdel).

Type C. En ønsker tall både for hele landet og for kommunetyper.

Type D. Geografisk avgrensede undersøkelser.

Det viste seg vanskelig på forhånd å spesifisere de områder en venter å foreta lokale undersøkelser for. Dessuten vil en lokal undersøkelse ofte føre til at en må bruke flere intervjuere fra andre områder for å kunne foreta datainnsamlingen. En valgte derfor ikke å ta hensyn til ønsker om lokale undersøkelser ved konstruksjon av utvalgsplanen. Unntakelser er Oslo, Bergen og Trondheim, hvor det er mulig å foreta lokale undersøkelser uten altfor store omallokeringer av intervjuere. Det er spesielt i forbindelse med store undersøkelser at ønsker av type A er framkommet. Det er her vanligvis sterke ønsker om å gi tall for fylker eller grupper av fylker. For andre undersøkelsers vedkommende er det ikke kommet fram vesentlige krav om noen inngående regionale oppdelinger av utvalget. Ønsker av type B forekommer i forbindelse med de fleste utvalgsundersøkelser; ønsket er likevel sjelden så sterkt at en har valgt å overrepresentere visse deler av landet. Etter at den nye kommunegrupperingen er blitt ferdig, er det kommet fram ønsker om å kunne gruppere utvalget etter disse typene. Grupperingen av kommunene i kommunetyper er gitt i Rideng (1974).

Ønsket om homogenitet innen strata samt kravet om å kunne gi regionale tall viste seg å være bare delvis sammenfallende. Under stratifiseringen viste det seg f.eks. nødvendig helt å oppgi å inndele kommunene etter viktigste næring. Indirekte er det likevel tatt hensyn til næring da kommunene er stratifisert etter kommunetype.

Et litt spesielt problem i forbindelse med stratifiseringen gjelder valg av antall strata i forhold til antall primære utvalgsområder. Ut fra hensynet om å redusere utvalgsvariansen og å bli i stand til å lage regionale tall bør en lage så mange strata som mulig, og trekke ett primært utvalgsområde pr. stratum. En ulempe ved denne stratifisering er at en vanskelig kan estimere utvalgsvariansen, og da spesielt den del av utvalgsvariansen som skyldes variasjonen mellom primære utvalgsområder innen det samme

stratum. På tross av denne svakhet, valgte en å lage flest mulig strata, og trekke ett primært utvalgsområde innen hvert stratum.

Som konklusjon kom en fram til følgende stratifisering av de primære utvalgsområder:

Først ble landet inndelt i 5 landsdeler:	Innbyggere etter folketellingen i 1970
1. Oslo - Akershus	814 214
2. Resten av Østlandet	1 128 164
3. Sørlandet/Vestlandet (÷ Møre)	977 263
4. Møre/Trøndelag	589 972
5. Nord-Norge	463 381

Bortsett fra Oslo-Akershus er landsdelene inndelt i følgende regioner:

2.1.	Østfold - Vestfold	407 351
2.2.	Hedmark - Oppland	358 268
2.3.	Buskerud - Telemark	362 545
3.1.	Agder - Rogaland	492 346
3.2.	Hordaland - Sogn og Fjordane	484 917
4.1.	Møre og Romsdal	229 407
4.2.	Trøndelag	360 565
5.1.	Nordland	242 627
5.2.	Troms - Finnmark	220 754

De primære utvalgsområdene i regionene ble så stratifisert slik at enhver region kan dannes som en sammenslåing av strata. Det finnes altså ikke strata som har primære utvalgsområder for flere enn én region. Årsaken til dette er ønsker om oppdeling av utvalget til arbeidskraftundersøkelsene.

Byer med flere enn 30 000, samt noen få i tillegg, er stratifisert slik at de hver for seg utgjør det eneste primære utvalgsområdet i stratumet. De øvrige primære utvalgsområdene er stratifisert etter kommunetype og innbyggerantall. De to siste kriteriene er såkalte underordnede kriterier. Dvs. at en har vært nødt til å ha utvalgsområder av forskjellige typer innen samme stratum. Dette forhindrer likevel ikke at en kan gruppere utvalget etter kommunetype ved utkjøring av tabeller.

De kommuner som er trukket ut med sikkerhet, og derfor utgjør et eget stratum, er følgende:

		Innbyggere etter folketellingen i 1970
0219	Bærum	80 861
0220	Asker	33 774
0231	Skedsmo	33 456
0301	Oslo	408 301
0101	Halden	27 193
0103	Fredrikstad	29 396
0104	Moss	25 522
0805	Porsgrunn	31 584
0602	Drammen	50 565
0501	Lillehammer	21 046
0601	Ringerike	29 709
0706	Sandefjord	33 232
0806	Skien (+Siljan)	47 671
1102	Sandnes	32 891
1149	Karmøy	29 053
1101	Kristiansand	58 977
1103	Stavanger	84 334
1201	Bergen	214 470
1601	Trondheim	133 205
1501	Ålesund	40 683
1833	Rana	26 301
1804	Bodø	30 384
1902	Tromsø	42 246
	I alt	1 544 854

4. Trekking på første trinn

Innen hvert stratum som består av flere kommuner, er trukket et primært utvalgsområde proporsjonalt med innbyggertallet i 1970. En metode til å trekke utvalgsområder proporsjonalt med innbyggertallet er utførlig beskrevet i Raj (1968, side 47). Når kommunene er trukket på en slik måte på første trinn, får en ikke et representativt utvalg av kommuner, idet de store kommuner er overrepresentert. Denne trekkemetoden er anvendt fordi den reduserer innsamlingskostnadene og reduserer variansen i de fleste tilfelle. Den skjevhet en legger inn ved trekking i første trinn, oppveies deretter ved trekking i annet trinn. La oss se på et eksempel. Når en ønsker et 1-prosents utvalg av hele befolkningen, ønsker en å gi samtlige

personer en sannsynlighet på 1/100 for å komme med i utvalget. Innen de kommuner som er trukket med sikkerhet, dvs. alle kommuner listet opp på side 24, trekkes et 1-prosents utvalg tilfeldig. Innen de øvrige kommuner trekkes med en sannsynlighet som er slik at produktet av kommunens trekkesannsynlighet og personens trekkesannsynlighet er lik 1 prosent. Nes kommune er trukket med en sannsynlighet på litt mer enn 50 prosent, hvilket vil si at en innen Nes kommune må trekke ca. 2 prosent av befolkningen for at alle i Nes kommune skal få en trekkesannsynlighet på 1/100. På samme måten utregnes trekkesannsynligheten innen alle de øvrige uttrukne kommuner. I en tiårsperiode holdes resultatet av første trinns trekking konstant, og utvalget av personer og husholdninger trekkes innen disse kommuner i forbindelse med de forskjellige undersøkelser.

5. Trekking på annet trinn

a. Oppretting og ajourføring av utvalgsregisteret

Utvalgene trekkes fra et register som er laget spesielt med henblikk på denne utvalgsplanen. Grunnlaget for utvalgsregisteret er folkeregistrenes magnetbåndregister. Dette inneholder alle familier og personer bosatt i Norge, og det blir ajourført etter oppgaver fra de lokale folkeregistre. Sæbø (1976).

Som tidligere nevnt består hver primær utvalgsenhet av en kommune, eller i noen få tilfeller av to eller flere mindre kommuner. Ved mindre undersøkelser brukes som tidligere nevnt bare en tredjedel av utvalgsområdene, slik at en i de fleste uttrukne kommuner, unntatt de største byene, bruker en intervjuer. Oppdelingen i utvalgsområder innen kommunen er foretatt etter valgkretser.

Ajourføringen foregår ved at det opprettes et nytt register med ajourførte data fra folkeregistrene som grunnlag. De uttrukne primære utvalgsenheter holdes fast ved ajourføring av registeret. Det kan imidlertid være aktuelt å endre den videre oppdeling av disse. Formålet med dette er først og fremst å redusere feltkostnadene, men slike endringer kan også være nødvendige for å redusere utvalgsvariansen når en trekker utvalg fra deler av de primære utvalgsområder. En annen årsak er at inndelingen i valgkretser kan være endret siden siste ajourføring. Byrået får ajourførte bånd fra folkeregistrene i begynnelsen av hvert år, utvalgsregisteret ajourføres hvert år på grunnlag av disse.

b. Organisering av utvalgsregistrene

Utvalgsregistrene inneholder en record for hver person bosatt i utvalgsområdene. Opplysningene i registeret gir en del informasjon om personene

og familiene. Fødselsnummeret gir alder og kjønn. Familienummeret er fødselsnummeret til familiens hovedperson. Dessuten inneholder registeret opplysninger om personens stilling i husholdningen. En familie i folkeregisteret består aldri av mer enn 2 generasjoner. Barn som flytter hjemmefra blir tildelt eget familienummer. Dette beholdes selv om de seinere flytter tilbake til foreldrene. I praksis oppstår det derfor en del problemer med å trekke et utvalg av familier, disse er behandlet i avsnittet nedenfor. I tillegg til de nevnte opplysninger inneholder registeret opplysninger om ekteskapelig status. Denne informasjonen gir mulighet til å trekke utvalg av personer med bestemte kjennetegn. For eksempel går det an å trekke et utvalg av gifte kvinner.

c. Oppretting av delregistre

Utvalgsregisteret for landet utenom Oslo består av ca. 2 millioner records, mens Oslo har ca. 450 000 records. For å trekke et utvalg må en gjennomløpe hele registeret. Dette blir dyrt med så store registre, og en har derfor opprettet et mindre delregister for hvert register. Trekking av utvalg til undersøkelser foregår fra disse.

I Byråets utvalgsplan er de primære utvalgsområder og dermed utvalgsområdene som tidligere nevnt trukket med ulike sannsynligheter. For at det endelige utvalg skal bli selvveiende, må en derfor trekke familier eller personer med ulik sannsynlighet i de forskjellige utvalgsområdene. Det tas imidlertid hensyn til dette alt ved trekking av delregistrene, slik at disse består av selvveiende utvalg på 10 prosent av befolkningen. I områder som er trukket med sannsynlighet 1 i første trinn, f.eks. Oslo, trekkes 10 prosent, mens andelen fra andre områder er større. Delregisteret for landet utenom Oslo inneholder fra 10 prosent av befolkningen i de største byene, til nær 100 prosent i de mindre utvalgsområder. I alt består det av ca. 350 000 records, mens delregisteret for Oslo har med 45 000. Ved å trekke utvalgene fra dette delregisteret, regner en med å spare betydelige beløp ved trekking av utvalg.

d. Problemer i forbindelse med trekking av utvalg fra et register

I dette avsnitt tas opp noen av de vanskeligheter som oppstår når en trekker utvalg fra et personregister, og settes opp mulige regler for hvordan disse vanskeligheter skal takles. I praksis har det vist seg vanskelig å sette opp regler som skal følges for alle typer undersøkelser. I forbindelse med noen undersøkelser velger en én løsning, mens en i forbindelse med andre undersøkelser kan velge en helt annen løsning.

Problemene med trekking av utvalg fra et personregister skyldes i det vesentlige to ting:

1. Registeret er aldri helt å jour. Spesielt er flyttinger et problem.
2. De enheter som registeret inneholder, person og familieenhet, faller bare delvis sammen med de enheter vi trekker til utvalg.

Personene er gruppert i familieenheter, som er definert på følgende måte:

Personene i en kjernefamilie må være registrert bosatt på samme sted (bo i samme leilighet).

En familie vil alltid tilhøre en av følgende familietyper:

a. Mann og hustru uten hjemmeværende ugifte barn, som er registrert bosatt på samme sted.
b. Mann og hustru med hjemmeværende ugifte barn, som er registrert bosatt på samme sted.
c. Mor og hjemmeværende ugifte barn (moren kan være ugift, gift eller før gift).
d. Far og hjemmeværende ugifte barn (faren kan være gift, før gift eller ugift).
e. Person som er registrert bosatt uten hjemmeværende ugifte barn, ikke sammen med ektefelle, ikke sammen med foreldre.

En familie vil således aldri bestå av personer fra mer enn to generasjoner. Når en ugift mor bor hjemme hos foreldrene sammen med barnet, danner denne ugifte moren og barnet hennes en egen familieenhet. Ugifte barn som bor hos foreldrene utgjør, sammen med foreldrene, en familieenhet uansett barnas alder og uansett om barna har selvstendig inntekt eller ikke.

En skal i det følgende inndele utvalgsundersøkelsene i to typer:

Undersøkelser med person som trekkeenhet, personundersøkelse, og Undersøkelser med trekkeenheter som består av minst en familieenhet, husholdningsundersøkelser:

Undersøkelser med personer som trekkeenheter

Målet ved trekkingen er å gi alle personer bosatt innen et utvalgsområde samme sannsynlighet for å komme med i utvalget. For dette brukes registeret som ikke er helt å jour og derfor inneholder navn og adresser til personer som er flyttet ut av området. Dessuten må en vente at det er bygget nye hus hvis beboere ikke er med i registeret. Det finnes metoder en kunne bruke for å få med nye hus, men fordi problemet i praksis ikke er stort, har en valgt å se helt bort fra det i utvalgsundersøkelsene. Når det gjelder flyttinger ellers, har det vist seg at det har vært nødvendig å ha regler som tar sikte på å gi nyinnflyttede personer den samme trekkesannsynlighet

som de øvrige personer i utvalget. Dette gjøres ved at en tenker seg at personutvalget innen et utvalgsområde er trukket i to trinn; først trekkes adressen, og innen adressen trekkes en person. Dette gjøres på følgende måte:

1. Anta at hver person i personregisteret har sannsynlighet p for å bli trukket.
2. En adresse (leilighet) har da tilnærmet sannsynlighet $m_i p$ for å bli uttrukket, hvor m_i er antall ganger adressen forekom i registeret. m_i og p er kjente før trekking.
3. Hvis en person blir trukket og intervjueren finner ut at den leiligheten personen bor i, inneholder de samme personer som registeret, er den uttrukne personen med i utvalget.
4. Hvis den uttrukne personen er flytter til en ny adresse, skal intervjueren trekke en ny person, som blir med i utvalget i stedet for den personen som ble trukket sentralt. Under trekkingen av en ny person skal alle faktisk bosatte ha en sjanse på $1/m_i$ for å komme med i utvalget.

I praksis har det vist seg visse vanskeligheter med å trekke personen innen husholdningen, og en bruker derfor ofte å trekke en ny person ved å velge den personen i den nye husholdningen som likner mest mulig på den opprinnelig trukne personen. Det har også vist seg vanskelig å identifisere leiligheten på grunnlag av navn, spesielt gjelder dette i felleshusholdninger (pensjonat, internat, sykepleierbygg o.likn.). I slike tilfeller betraktes huset som trukket, og intervjueren trekker ut en person etter samme regler som gitt ovenfor.

En annen løsning som kan brukes i forbindelse med trekking av personutvalg består i at en finner ut den nye adressen til den uttrukne personen, og deretter intervjuer personen på den nye adressen. På denne måten får en et utvalg som er et sannsynlighetsutvalg fra populasjonen slik den så ut ved siste ajourføringsdato. Det er vanskelig å gi noen generelle regler for når man bør velge den ene eller den andre løsningen. Men det er helt klart at den siste løsningen fordyrer innsamlingen ganske vesentlig sammenliknet med den første metoden.

Undersøkelser hvor trekkeenheten består av minst en familieenhet

I husholdningsundersøkelser varierer trekkeenheten fra undersøkelse til undersøkelse. I forbruksundersøkelsene er f.eks. trekkeenheten kosthusholdningen, mens trekkeenheten i arbeidskraftundersøkelsene er bohusholdningen. I slike tilfeller kunne en gå igjennom hele registeret og samle familieenhetene i trekkeenheter, og deretter foreta trekkingen. En slik framgangsmåte er unødvendig kostbar, og er aldri blitt brukt i praksis. Vanligvis trekkes utvalget etter en av følgende regler: Først trekkes et utvalg av familieenheter fra registeret. Dersom en trekkeenhet består av

flere familieenheter, er hele trekkeenheten trukket dersom familieenheten med trekkeenhetens eldste medlem er trukket. Hvis den uttrukne familieheten ikke har trekkeenhetens eldste medlem, legges familieenheten tilbake i registeret. En oppnår dermed at alle trekkeenheter får samme trekkesannsynlighet. I praksis kan en først etter at intervjuingen er foretatt finne ut av om en trekkeenhet består av flere familieenheter eller ikke. I slike tilfeller kan en velge å kaste alle opplysningene dersom trekkeenheten inneholder en person som er eldre enn personene i den familieenhet som ble trukket ut, og førte til at trekkeenheten kom med i utvalget. En annen mulighet består i å ta med disse opplysninger, og gi personene en vekt som er omvendt proporsjonal med trekkesannsynligheten. En tredje mulighet består i å se bort fra at trekkesannsynligheten varierer, og ta med oppgavene uten å veie dem. Alle tre metoder er brukt i praksis, og det har vist seg at forskjellen mellom de to siste løsninger er uten særlig betydning, og derfor er det ofte at en velger å inkludere alle oppgavene i utvalget uten å veie.

Spesielle problemer når en bruker roterende utvalg

I forbindelse med roterende utvalg, dvs. utvalg hvor en del av de uttrukne personer har vært med i en tidligere liknende undersøkelse, har en det spesielle problem om en skal rotere på adresse eller på personer. Med dette menes om en skal la en del av et utvalg av adresser, eller en del av utvalget av personer fra en undersøkelse, være med i en seinere undersøkelse. Dette spørsmålet har egentlig ikke noe med registerets kvalitet å gjøre, men skyldes at utvalget trekkes i to trinn, og er altså en del av utvalgsplanen. I Byrået har en liten erfaring med slike problemer, men til nå har en brukt å rotere på personer, hvilket medfører at personer som flytter ut av utvalgsområdene mellom to undersøkelser følges opp. Det er klart at denne metoden gir bedre grunnlag for å kunne uttale seg om endringer fra den ene undersøkelse til den annen, men det er like klart at metoden er vesentlig dyrere enn den metode som består av å rotere på adresser.

Sluttmerknader til avsnitt

De fleste av de problemer som er tatt opp under avsnittet om spesielle problemer i forbindelse med trekking av utvalg fra registre har ikke noen klar entydig teoretisk, god løsning. Det viser seg da også at det i forbindelse med spesielle undersøkelser er behov for å justere de forslag til løsninger som er skissert ovenfor. Det er vanskelig å gi noen enkle mål for hvilken effekt de forskjellige løsningsmetoder har, men som regel er det lett å avgjøre hva som er enklest og mest praktisk.

6. Oppdeling av utvalgsområdene i mindre områder

Som tidligere nevnt har en til bruk ved mindre undersøkelser delt inn utvalgsområdene i mindre områder. I slike undersøkelser trekkes utvalget altså i tre trinn. Først trekkes utvalgsområdet, dernest trekkes en mindre del av det uttrukne område, og til slutt trekkes personen eller husholdningen som skal være med i undersøkelsen.

Hensikten med å innføre et tredje trinn i trekkingen, er å redusere reisekostnadene til intervjuerne, og redusere antall intervjuere som skal delta ved mindre undersøkelser. For å oppnå dette bør oppdelingen være slik, at hvert delområde kan dekkes av en intervjuer. Dessuten bør hvert område være representativ for hele kommunen.

Oppdelingen av utvalgsområdene er gjort på grunnlag av valgkretser. Hvert utvalgsområde er blitt delt i ca. 3 områder, som er så like som mulig i folketall, næringsstruktur og utstrekning.

Det er ikke mulig i denne sammenheng å gi en detaljert beskrivelse av oppdelingen av samtlige utvalgsområder, men i neste avsnitt skal gis en beskrivelse av oppdelingen av Karmøy kommune. De øvrige utvalgsområder er oppdelt på samme måten.

a. Oppdeling av Karmøy kommune

Karmøy kommune er den uttrukne kommune i stratum 62. Folketallet i 1970 var 27 635. Kommunen var inndelt i 28 valgkretser, nummerert fra 1 til 28.

Kommunen er inndelt på følgende måte:

Område I består av valgkretsene 01 og 02, 13-16 og 20. Antall personer i 1970 var 10 169, hvor 6 468 var 16 år og over.

Område II består av valgkretsene 03-12. Antall personer i 1970 var 10 059, hvor 6 393 var over 16 år.

Område III består av valgkretsene 17-19, 21-25 og 27-28. Antall personer i 1970 var 10 090, hvor 5 859 var over 16 år.

Alle personer 16 år og over fordelt etter viktigste kilde til livsopphold innen hvert område er gitt i tabell 3.

Tabell 3. Personer 16 år og over etter viktigste kilde til livsopphold og næring i forskjellige områder i Karmøy kommune. Prosent

		Viktigste kilde til livsopphold										
			Inntekt eget arbeid									
				Næring								
	I alt	I alt	Jordbruk, skogbruk, fiske m.v.	Industri m.v., bygge- og anl. virksomhet	Varehandel	Samferdsel	Tjenesteytende næringer	Uoppgitt	Pensjon, trygd	Formue, lån, stipend m.v.	Husarbeid i hjemmet, forsørget
Område I	100	48,02	6,23	20,01	5,57	8,67	7,54	-	16,72	1,39	33,87
Område II	100	47,47	8,34	19,33	4,22	8,17	7,29	0,12	18,69	1,27	32,57
Område III	100	48,73	5,22	21,86	5,79	7,97	7,89	-	14,25	1,56	35,46

En valgkrets, nemlig nr. 26 med 75 innbyggere i 1970, er sløyfet helt. Dvs. at personer bosatt her ikke har noen mulighet for å komme med i noe utvalg. Dette fører til en meget liten skjevhet, mens det reduserer kostnadene vesentlig, da kretsen består av en liten øygruppe et stykke fra fastlandet.

IV. OVERSIKT OVER DE VANLIGSTE ESTIMERINGSMETODENE

1. Innledning

Formålet med utvalgsundersøkelser er å estimere visse størrelser i populasjonen på grunnlag av resultatene i utvalget. De krav en setter til en god estimator er at den skal være forventningsrett, eller tilnærmet forventningsrett, i den forstand som er beskrevet i kapittel II. Dessuten bør den ha minst mulig varians for gitt utvalgsstørrelse. Estimatoren blir derfor avhengig av den utvalgsplan som brukes, men da det i Byrået oftest brukes selvveiende utvalg, skal en her innskrenke seg til å beskrive estimatorer som er forventningsrette i slike tilfeller. Dessuten skal det forutsettes at en ønsker å estimere gjennomsnitt eller totaler for hele populasjonen. I praksis ønsker en ofte å estimere gjennomsnitt og totaler for delpopulasjoner, f.eks. personer i en viss aldersklasse, i tillegg til gjennomsnitt for hele populasjonen. I noen tilfeller kan dette reise spesielle estimeringsproblemer, men disse skal ikke tas opp her. Interesserte lesere henvises til Kish (1965).

2. Utvalgsgjennomsnitt

Den mest anvendte estimeringsmetode er å bruke utvalgsgjennomsnitt som estimator for gjennomsnittet i populasjonen. Denne estimator er tilnærmet forventningsrett for alle selvveiende utvalgsplaner. Estimatoren er lett å beregne, og har visse andre fordeler som skal beskrives nærmere nedenfor. Endelig er det forholdsvis enkelt å estimere variansen til denne estimator for de fleste utvalgsplaner som er i bruk. I vedlegg 1 er gitt en nærmere beskrivelse av egenskapene til utvalgsgjennomsnittet.

3. Rateestimatoren

Ofte er det slik at statistikeren, i tillegg til data fra utvalget, har en del relevant informasjon om den populasjonen han arbeider med. Dette kan f.eks. være kjønnsfordelingen, aldersfordelingen, eller det kan være tidligere gjorte observasjoner av det kjennemerket en studerer i en bestemt undersøkelse. Slik tilleggsinformasjon kan brukes på mange måter for å redusere usikkerheten til resultatene fra en undersøkelse. En måte å gjøre det på er beskrevet i kapittel II, og består av å stratifisere populasjonen før en foretar trekking av utvalget. Som tidligere nevnt vil dette ofte føre til mindre usikkerhet enn et rent tilfeldig utvalg. En annen måte å utnytte slik tilleggsinformasjon på er å bruke en rateestimator for å estimere gjennomsnittet i populasjonen. For å kunne bruke rateestimatoren må følgende betingelser være oppfylt: For det første må det til hver enhet være knyttet to kjennemerkeverdier, som er høyt korrelerte. Dessuten må verdiene til det ene kjennemerket være kjent for samtlige enheter i populasjonen. La $\bar{x}$ betegne utvalgsgjennomsnittet for den variabel vi er interessert i å estimere gjennomsnittstallet for, la $\bar{y}$ betegne utvalgsgjennomsnittet for den andre kjennemerkeverdien, og la $\bar{Y}$ betegne populasjonsgjennomsnittet til hjelpevariabelen. Rateestimatoren ser da ut som følger:

$$R = \frac{\bar{x}}{\bar{y}} \bar{Y}.$$

Den antakeligvis mest kjente bruk av rateestimatoren er i forbindelse med de politiske valgbarometre. $\bar{x}$ er da andelen av velgere som sier at de vil stemme på et bestemt parti ved neste valg, mens $\bar{y}$ betegner andelen av velgere som sa at de stemte på samme parti ved siste valg. $\bar{Y}$ blir da andelen av velgere som stemte på partiet ved siste valg. $\bar{Y}$ er altså kjent. En kan si at rateestimatoren på en måte korrigerer det vanlige utvalgsgjennomsnitt. I mange tilfeller viser det seg at variansen til rateestimatoren kan være betydelig mindre enn variansen til det vanlige gjennomsnitt. På den andre siden er det viktig å være oppmerksom på at en i rateestimatoren kan

få betydelige skjevheter, dersom den forventede verdi til $\bar{y}$ er forskjellig fra $\bar{Y}$. I avsnittet ovenfor ble det påstått at utvalgsgjennomsnittet alltid vil være tilnærmet forventningsrett for alle selvveiende utvalgsplaner. For slike utvalgsplaner kan en derfor vente seg at den forventede verdi til $\bar{y}$ ofte vil være lik med $\bar{Y}$. På grunn av målefeil kan det imidlertid inntreffe at dette ikke er tilfelle, og en bør da vise stor forsiktighet ved bruk av rateestimatoren. I forbindelse med de politiske valgbarometre har det f.eks. vist seg, at personer glemmer hva de stemte på ved siste valg, hvilket naturligvis medfører at den forventede verdi til $\bar{y}$ ikke nødvendigvis blir lik $\bar{Y}$. Det kreves derfor betydelig innsikt i en bestemt type undersøkelser for å kunne foreta et fornuftig valg mellom utvalgsgjennomsnittet og rateestimatoren. Thomsen (1977). I praksis har det ofte vist seg at stratifisering gir samme reduksjon i variansen som rateestimatoren. Ved stratifisering risikerer en ikke å innføre skjevheter på samme måte som tilfellet er med rateestimatoren, og derfor bør en som hovedregel heller stratifisere populasjonen snarere enn å bruke rateestimatoren. I praksis er det dessverre ofte umulig å stratifisere i stedet for å bruke rateestimatoren, for eksempel i forbindelse med de ovennevnte politiske valgbarometre. I Byrået brukes rateestimatoren bl.a. til å estimere omsetningsindeksen.

4. Etterstratifisering

Tidligere er det vist hvordan en ved hjelp av stratifisering kan redusere usikkerheten på en estimator. I mange tilfeller i praksis kan det være vanskelig eller umulig på forhånd å avgjøre hvilket stratum et bestemt element tilhører. Hvis en f.eks. ønsker å stratifisere personene i en by etter alder, er det lett å få tall for aldersfordelingen i byen, men uten adgang til et godt register som gir alderen til hver person vil stratifisering ikke være mulig. I slike tilfeller kan en med fordel stratifisere etter at data er samlet inn, også kalt etterstratifisering. Anta at en i en utvalgsundersøkelse ønsker å estimere gjennomsnittlig inntekt for en populasjon, og at en ønsker å stratifisere etter alder, men ikke kan gjøre det på grunn av manglende register. I et slikt tilfelle kan en trekke et tilfeldig utvalg innen byen. De innsamlede data kan deretter grupperes etter alder, og gjennomsnittlig inntekt utregnes for hver aldersgruppe. Antallet N_i i aldersgruppene i populasjonen er kjent, og gjennomsnittlig inntekt for hele populasjonen kan nå estimeres ved

$$\sum_{i=1}^{h} \frac{N_i}{N} \bar{X}_i,$$

hvor h er antall aldersgrupper, $N = \Sigma N_i$ og $\bar{X}_i$ er gjennomsnittlig inntekt i aldersgruppe i. I et selvveiende utvalg vil de forskjellige grupper ikke

bli representert nøyaktig på samme måten som i den endelige populasjonen på grunn av tilfeldige variasjoner. Hensikten med etterstratifiseringen er å gi hvert gruppegjennomsnitt nøyaktig samme vekt i utvalget som det har i populasjonen. Forutsatt at utvalget er så stort at en kan vente å få et visst antall observasjoner i alle etterstrata, gir etterstratifiseringen nesten samme gevinst som stratifisering og proporsjonal allokering. Etterstratifisering fører imidlertid til en litt mer komplisert regnerutine, slik at en normalt bør bruke vanlig stratifisering hvis det er mulig.

En viktig egenskap ved etterstratifisering er at en reduserer effekten av frafall, spesielt når frafallsprosentene varierer mye fra etterstratum til etterstratum. Dette problem er behandlet litt videre i kapittel VI.

I Byrået brukes etterstratifisering i forbindelse med de kvartalsvise arbeidskraftundersøkelser.

5. Andre estimeringsmetoder

I løpet av de siste årene har det skjedd en veldig utvikling innen teorien for statistiske utvalg. Denne nye teorien atskiller seg fra den klassiske gjennom at en forutsetter at den endelige populasjonen er tenkt generert av en stokastisk mekanisme. Idéen er ikke ny, men interessen har tatt seg opp de siste årene. Et utvalg tenkes her framkommet i to trinn: først er den endelige populasjonen trukket fra et stort utvalg fra en uendelig populasjon, og deretter er det trukket et mindre utvalg fra den endelige populasjonen. På grunnlag av observasjonene i utvalget kan en gjøre generaliseringer til den endelige populasjonen, eller til superpopulasjonen. Fordelen ved en slik betraktningsmåte er blant annet at en kan formulere modeller for den endelige populasjonen, og bruke denne informasjonen i forbindelse med estimering. Metodene er ennå lite brukt i Byrået, men i forbindelse med arbeidskraftundersøkelsene er de blitt brukt. Dagsvik (1975, 76). En må regne med at metodene vil være i utstrakt bruk om noen år.

6. Estimering av utvalgsvarianser

Enhver statistisk undersøkelse er forbundet med en rekke usikkerheter. Noen av disse kan kontrolleres i meget høy grad, mens andre kun i mindre grad kan kontrolleres av statistikeren. En av de usikkerheter som relativt enkelt kan kvantifiseres og kontrolleres, er den usikkerhet som skyldes at estimatorene er utregnet på grunnlag av et utvalg fra hele populasjonen. Eksempler på feilkilder som er mer vanskelig å kontrollere er

det faktum at folk glemmer, svarer feil bevisst, eller helt nekter å svare på et spørsmål. Disse siste typer feil er behandlet i kapittel VI. I løpet av de siste årene er det i Byrået nedlagt mye arbeid for å få et innblikk i de utvalgsvarianser som resultatet fra den generelle utvalgsplanen er forbundet med. I dette avsnittet skal resultatene fra dette arbeid beskrives, men først skal det gis en kort innføring i problemstillingen.

Dersom formålet er å estimere andelen av en populasjon som har bil, her kalt p, og en har et enkelt tilfeldig utvalg fra populasjonen, er utvalgsvariansen tilnærmet gitt ved

$$p(1-p)/n,$$

hvor n er størrelsen på utvalget. Utvalgsvariansen estimeres enkelt ved å innsette en estimator for p. En person som leser resultatene fra en undersøkelse, kan altså lett regne ut en estimator for utvalgsvariansen på grunnlag av de publiserte relative hyppigheter. Problemene oppstår fordi en vanligvis ikke har enkle tilfeldige utvalg. Ved to- og flertrinnsutvalg er utvalgsvariansen vesentlig mer komplisert enn tilfelle er ved enkel tilfeldig trekking. Dessuten er det en del problemer knyttet til estimering av variansen.

Når en har adgang til data for hele populasjonen, er det mulig å regne ut eksakte uttrykk for utvalgsvariansen. Dette følger av utledningene gitt i vedlegg 1. Data fra Folke- og boligtellingen 1970 gir data for hele den norske befolkning et bestemt år, og gir derfor grunnlag for å foreta nøyaktige beregninger av utvalgsvariansen for de kjennemerker som er med i folketellingen. Ved studiet av utvalgsvarianser er en særlig opptatt av forholdet mellom variansen når utvalg trekkes etter Byråets generelle utvalgsplan, og variansen når utvalg trekkes som et enkelt tilfeldig utvalg. Dette forhold kalles design-effekten. I tabell 4 er gitt en rekke design-effekter for forskjellige kjennemerker, og for forskjellig valg av utvalgsstørrelser. Disse beregninger er gjort på grunnlag av data fra 1970. Sæbø (1976).

En bør merke seg følgende av tabell 4: Forskjellige kjennemerker har forskjellige design-effekter. Kjennskap til design-effekten for ett kjennemerke gir i prinsippet ikke noen informasjon om design-effektene til andre kjennemerker. Design-effekten øker med utvalgsstørrelsen. Sist men ikke minst ses det at design-effekten er stort sett større enn 1. Ved å bruke den generelle utvalgsplanen i stedet for et enkelt tilfeldig utvalg, øker en altså usikkerheten på resultatene noe. Hensikten med en generell utvalgsplan er å redusere kostnadene. Tallene i tabell 4 forteller da om hvor mye en taper i sikkerhet gjennom å bruke en slik utvalgsplan. At

Måten å publisere utvalgsvarianser på varierer mye fra institusjon til institusjon. Byråets praksis på dette området er beskrevet i kapittel VII. Utledningen av utvalgsvariansen er gitt i vedlegg 1, hvor det også er foreslått en estimator for den.

V. HVOR STORT BØR UTVALGET VÆRE?

1. Innledning

Da utvalgets størrelse har avgjørende innflytelse på kostnadene til en utvalgsundersøkelse, er spørsmålet om hvor stort utvalget bør være blant de første som blir stilt i forbindelse med planlegging av undersøkelsen. De fleste statistikere har opplevd at det virker overraskende på de fleste brukere når de ikke får svar på dette enkle spørsmål allerede tidlig under planleggingen, og at mange blir irritert når de gjøres oppmerksom på at størrelsen på utvalget helt avhenger av hva en ønsker å få ut av undersøkelsen. I dette avsnittet skal først gis en forklaring på hvorfor utvalgsstørrelsen en nær knyttet sammen med formålet med undersøkelsen, deretter skal en beskrive noen enkle framgangsmåter som kan brukes for å komme fram til en fornuftig utvalgsstørrelse i visse tilfelle. Til slutt i avsnittet er beskrevet mer formelle metoder ved hjelp av hvilke man kan bestemme optimale utvalgsstørrelser.

2. Hvorfor varierer utvalgsstørrelsen så mye fra én undersøkelse til en annen?

Når en person med mindre erfaring fra utvalgsundersøkelser plutselig blir nødt til å ta stilling til hvor stort utvalg som trengs til en bestemt undersøkelse, er det naturlig å forsøke å finne ut av hvor store utvalg en har brukt ved tidligere liknende utvalgsundersøkelser. Vedkommende vil da ofte finne ut at størrelsen på utvalgene varierer betydelig fra én undersøkelse til en annen. I Byrået utføres det i dag utvalgsundersøkelser på grunnlag av færre enn 1 000 intervjuobjekter, mens andre utvalgsundersøkelser krever så mye som 12 000 intervjuobjekter. For å se nærmere på årsakene til denne variasjonen i utvalgsstørrelse, skal det settes opp tre forhold som spiller en avgjørende rolle for hvor stort utvalg en trenger:

a. Den nøyaktighet en ønsker på resultatene.
b. Homogeniteten i populasjonen.
c. Antall oppsplittinger av utvalget en ønsker å gjøre.

Ad. pkt. a. I praksis er det som regel meget vanskelig på forhånd å si mye om hvor nøyaktige resultater en ønsker. Dette henger sammen

Tabell 4. Design-effekt for estimert sysselsetting etter næring ved forskjellige utvalgsstørrelser. Tall for svensk utvalgsplan i parentes

Næring	Utvalgsstørrelse								
	1 000		2 000		5 000		10 000		12 000
Jordbruk, skogbruk	1,01	(0,97)	1,07	(1,04)	1,25	(1,24)	1,54	(1,58)	1,64
Fiske, fangst ..	1,08	-	1,19	-	1,50	-	2,03	-	2,24
Industri m.v. ..	1,01	(1,00)	1,04	(1,04)	1,13	(1,14)	1,29	(1,32)	1,35
Bygg, anlegg ...	1,00	(1,02)	1,01	(1,05)	1,05	(1,12)	1,11	(1,25)	1,13
Varehandel	0,99	(0,99)	0,99	(1,00)	1,01	(1,03)	1,04	(1,07)	1,05
Samferdsel	1,00	(1,01)	1,00	(1,02)	1,03	(1,06)	1,06	(1,13)	1,08
Tjenester m.v. .	0,99	(1,00)	1,01	(1,01)	1,06	(1,07)	1,14	(1,17)	1,18
Sysselsetting ..	1,00	(1,01)	1,01	(1,02)	1,04	(1,07)	1,08	(1,14)	1,10

design-effekten ligger mellom 1,0 og 1,5 er noe som er observert i forbindelse med andre liknende utvalgsplaner. Kemsley (1966) kommer fram til liknende konklusjoner for engelske forbruksundersøkelser. Kish (1965) gir også liknende resultater fra en serie undersøkelser i USA. Dessuten ser det ut som om de svenske resultater som er gitt i parantes i tabell 4 stemmer godt med Byråets resultater.

Selv om design-effekten til et kjennemerke i prinsippet ikke sier noe om design-effekten til et annet kjennemerke, viser empiriske studier at en øker variansen med mellom 0 og 50 prosent ved å bruke generelle utvalgsplaner av den samme type som Byråets i stedet for tilfeldige utvalg.

Det er ikke mulig å beregne eksakte varianser for kjennemerker som ikke er med i en eller annen totaltelling. I slike tilfeller er det behov for på grunnlag av resultatene i utvalget å kunne estimere utvalgsvariansen. De forsøk som er gjort på dette i Byrået, har imidlertid vist at det er vanskelig å finne gode estimatorer for utvalgsvariansen i mange tilfeller. De metoder en til nå har forsøkt, har vist seg enten å overestimere eller å underestimere utvalgsvariansen. Når det er behov for å angi utvalgsvariansen i forbindelse med publiserte tall, brukes derfor en av følgende framgangsmåter:

i) Når det gjelder variansen til relative hyppigheter, anbefales leseren å bruke uttrykket $1,5 \cdot p \cdot (1-p)/n$. Med de vanligste utvalgsstørrelser vil denne antakeligvis i de fleste tilfelle overestimere den virkelige utvalgsvarians.

ii) En kan estimere utvalgsvariansen ved hjelp av et av de tilgjengelige programmer. Også i slike tilfelle kan en regne med å få resultater som overestimerer den riktige utvalgsvarians.

med at en meget sjelden kan kvantifisere konsekvensene av å basere beslutninger på grunnlag av unøyaktige estimatorer. Det er f.eks. vanskelig å si noe om kostnadene ved at sysselsettingen i en bestemt næring blir anslått unøyaktig. Det er likevel viktig at en forsøker å komme fram til visse uttrykk for hvor gode resultater en ønsker, spesielt i situasjoner hvor en på forhånd har en del informasjon. Hvis en i slike tilfelle velger for liten utvalgsstørrelse, kan en risikere at resultatene fra utvalgsundersøkelsene blir helt verdiløse. Ved valg av utvalgsstørrelse i slike situasjoner bør en derfor bruke en av de mer formelle metoder som er beskrevet i avsnitt V.3. nedenfor.

Ad. pkt. b. Det sier seg selv at en trenger et relativt lite utvalg i tilfeller hvor det er liten variasjon i kjennemerkeverdiene i populasjonen. I det ekstreme tilfellet hvor alle elementene har samme verdi på det kjennemerke en ønsker å studere, trenger en bare en observasjon for å anslå gjennomsnittlig verdi i populasjonen. Hvis en derimot ønsker å anslå gjennomsnittlig inntekt i Norge, trenger en et større utvalg for å få brukbare tall. I avsnitt V.3. nedenfor er det foreslått metoder en kan bruke for å finne nødvendige utvalgsstørrelser.

Ad. pkt. c. Hvis en i en landsomfattende undersøkelse i tillegg til tall for hele landet også ønsker tall for deler av populasjonen, må en ta hensyn til dette ved valg av utvalgsstørrelse. Hvis en for eksempel er kommet fram til at 100 observasjoner er tilstrekkelig for å gi gode tall for hele landet, vil en i de fleste tilfelle måtte trekke 200 personer dersom en ønsker gode tall for menn og kvinner i tillegg til gode tall for hele landet. Dersom en ønsker å oppsplitte materialet ytterligere, bør en sikre seg at det er ca. 100 observasjoner innen hver oppsplitting. I praksis vil en ofte finne at det er ønsket om oppsplitting av materialet som setter de største krav til utvalgsstørrelsen. Når det i Byrået skal gis et raskt, foreløpig anslag på utvalgsstørrelse, finner en ofte ut hvilke oppsplittinger av materialet en ønsker seg, deretter regnes utvalget ut på en slik måte, at vi er sikre på å få minst 40 intervjuobjekter innen hver oppsplitting. Etterhånden som formålene med undersøkelsene blir mer klargjorte, justeres dette anslaget ved hjelp av mer formelle metoder beskrevet i neste avsnitt.

3. Formelle metoder en kan bruke for å finne fram til fornuftige utvalgsstørrelser

a. Innledning

For i det hele tatt å kunne foreta et fornuftig valg av utvalgsstørrelser, må en på en eller annen måte komme fram til utsagn om den nøyaktighet en ønsker resultatene skal ha. Ofte er det vanskelig å avgjøre hvor mye usikkerhet som kan aksepteres, men etter en del diskusjon kan en komme fram til at en ønsker et standardavvik som gir et resultat som er "korrekt pluss/minus 5 prosent". Det vil si at dersom utvalget viser en prosentandel på 50, er en relativt sikker på at det riktige tall ligger mellom 45 og 55. En kan naturligvis aldri være helt sikker på at dette utsagnet er riktig, men en kan garantere for at sjansen for at utsagnet er riktig er meget høy. Etter å være kommet fram til et slikt utsagn om nøyaktighet, er det mulig å si noe om hvor stort utvalg en må ha. For enkelhets skyld skal vi anta at både populasjonen og utvalget er stort og at anslaget i utvalget er normalfordelt. La p betegne anslaget i utvalget, og P andelen i populasjonen. Siden p er antatt normalfordelt rundt P, vil intervallet $(p \pm 2 \cdot \sigma)$ inneholde P med ca. 95 prosent sjanse. Dessuten er

$$\sigma = \sqrt{\frac{P(1-P)}{n}}$$

Det følger altså at

$$2\sqrt{\frac{P(1-P)}{n}} = 5 \text{ eller } n = \frac{4P(1-P)}{25} .$$

I formelen for n inngår den ukjente størrelsen P, som vi ønsker å estimere. Dette vil alltid være tilfelle når en skal anslå utvalgsstørrelsen i en undersøkelse. En enkel måte å løse dette problemet på består i å anta at P er 50 prosent. I dette tilfelle antar telleren i formelen for n sin maksimale verdi, og en vil derfor sikre seg at n blir tilstrekkelig stor. Hvis en på den andre siden vet at P ligger rundt 20 prosent, bør en bruke dette anslaget i formelen for n, for derved å spare en del observasjoner.

I tilfeller hvor P ligger nær null, kan formelen for n virke forvirrende. En vil nemlig finne at en bare trenger et veldig lite antall observasjoner for å oppfylle det kvalitetskrav som er satt ovenfor. Dersom en på forhånd ved at P ligger nær null, vil en naturligvis ikke akseptere at svaret er riktig innenfor 5 prosent, men forlange at svaret skal

være riktig innenfor f.eks. ½ prosent. Ved å innsette dette nye kvalitetskrav i formelen for n ovenfor, vil man komme fram til utvalgsstørrelser som stemmer med det en kunne vente seg.

b. Formel for bestemmelse av n

Tankegangen i avsnittet ovenfor kan generaliseres på følgende måte: En ønsker å anslå hyppigheten av et bestemt kjennemerke i en populasjon, P. En er villig til å akseptere at avviket mellom P og anslaget kan bli større enn d med en liten sannsynlighet, α. En har altså at

$$P(|p-P| \geq d) = \alpha.$$

Dersom det trekkes et enkelt tilfeldig utvalg, antas p å være normalfordelt med forventning P og standardavvik gitt ved

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{P(1-P)}{n}} .$$

En finner da at den verdi n minst må ha er gitt ved

$$n = \frac{t^2 P\ (1-P)/d^2}{1+\frac{1}{N}\left(\frac{t^2 P(1-P)}{d^2} - 1\right)} ,$$

hvor t er (1-d) - fraktilen i normalfordelingen.

I praksis må en sette inn et anslag for P for å bestemme nødvendig utvalgsstørrelse.

I forbindelse med landsomfattende undersøkelser er N så stor at en kan bruke følgende tilnærmelse for n

$$n \approx \frac{t^2 P(1-P)}{d^2} .$$

c. Valg av utvalgsstørrelse når en ønsker å finne ut om det er skjedd endringer i forhold til resultatene fra en tidligere foretatt undersøkelse

Etter hvert som flere og flere områder er blitt dekket av undersøkelser, blir det mer vanlig med gjentakelser, eller oppfølging av gamle undersøkelser. Ett av de viktigste formålene med slike gjentatte undersøkelser er å få anslaget for de endringer som har skjedd over tid. I slike tilfeller bør en legge ned spesielt mye arbeid i valg av utvalgsstørrelse. Problemene med å estimere endringer henger sammen med to forhold. For det første vil estimatoren for en endring i de fleste tilfelle være differansen

mellom to estimatorer, og variansen vil derfor bli summen av variansen til hver av de to estimatorer. (En ser her bort fra panelundersøkelser.) Grovt regnet kan en si at variansen til en estimator for endringen er dobbelt så stor som variansen til de tilsvarende nivåtallene. For det andre gjelder det at endringene ofte er små, hvilket gjør det nødvendig med stor nøyaktighet for å få brukbare anslag på endringene.

For å komme fram til en fornuftig utvalgsstørrelse kan en gå fram som i avsnittet ovenfor. En må nå bestemme seg for en ønsket nøyaktighet på endringstallet, samt gi et anslag for den endring en venter seg. Disse størrelser settes deretter inn i en formel tilsvarende til formlene ovenfor.

Nedenfor skal en se på en annen metode som en kan bestemme en fornuftig utvalgsstørrelse med. Andelen av personer som tilhører en bestemt gruppe er anslått i en tidligere utvalgsundersøkelse. La P_0 betegne denne andelen i populasjonen, og la $\hat{P}_0$ betegne det tidligere anslag. Tilnærmet kan variansen skrives opp som

$$\text{var}\,(\hat{P}_0) \approx \frac{P_0(1-P_0)}{n_o},$$

hvor n_o er antallet elementer i utvalget. En har mistanke om at P_0 har endret seg, og etter en tid ønsker en å utføre en undersøkelse for å teste om dette er tilfelle. Spørsmålet er da hvor stort utvalg en trenger for å være relativt sikker på å oppdage eventuelle endringer. Det er rimelig å regne med at dersom det har vært store endringer, trenger en ikke et så stort utvalg som i tilfeller hvor en regner med at endringene er små. I det følgende skal vi angi en metode som en kan fastslå nødvendig utvalgsstørrelse med, for å være relativt sikker på å oppdage endringer av en bestemt størrelse.

Når en på grunnlag av de nye resultater skal avgjøre om det har foregått endringer eller ikke, er det naturlig å basere denne konklusjonen på en hypotesetesting, hvor null-hypotesen er $P_1 = P_0$, altså ingen endring. Det nye utvalg bør være så stort at sjansen for å forkaste hypotesen at $P_1 = P_0$ når denne er riktig skal være liten, f.eks. lik ε, og samtidig ønsker vi stor sannsynlighet, s, for å forkaste den samme hypotesen når $P_1 - P_0$ er stor, f.eks. lik c. En kan nå sette opp følgende formel til bestemmelse av utvalgsstørrelsen:

$$n_1 \approx P_1(1-P_1)\Big/\left\{\frac{c^2}{(f\varepsilon - fs)^2} - \frac{P_0(1-P_0)}{n_o}\right\},$$

hvor f_ε og f_s betegner ε-fraktilen og s-fraktilen i normalfordelingen.

Når en skal bestemme utvalgsstørrelsen som er nødvendig, må en sette inn verdier for ε, s, P_0 og c i formelen ovenfor.

En bør merke seg at høyre side i formelen for n_1 kan være negativ. Dette henger sammen med at det er mulig å sette opp verdier for parametrene i formelen, slik at ingen utvalgsstørrelse er stor nok til å oppfylle kravene. Selv om den nye undersøkelsen består av en totaltelling, er det grense for hvor fine endringer en kan måle, når den første undersøkelse er en utvalgsundersøkelse. Når en i praksis kommer opp i en slik situasjon, bør en overveie å bruke panelundersøkelser, det vil si at en helt eller delvis bruker det samme utvalg to ganger. En skal ikke her komme nærmere inn på bruken av panelundersøkelser, men henviser til kapittel VIII.

VI. FEILKILDER

1. Innledning

Arbeidet med en statistisk undersøkelse, enten det er en utvalgsundersøkelse eller en totaltelling, kan med fordel inndeles i tre faser: Planlegging og trekking av utvalg, innsamling av data og databehandling. I hver av disse fasene må en regne med at det foreligger kilder til feil. Feilene skal inndeles i fire typer:

1. Register- og utvalgsfeil.
2. Målefeil.
3. Databehandlingsfeil.
4. Frafall.

2. Register- og utvalgsfeil

Når en trekker enheter til en statistisk undersøkelse, er målet at alle enheter i en populasjon skal ha en kjent sannsynlighet for å komme med i utvalget. Ved en totaltelling ønsker en utvalgssannsynligheten lik 1. På grunn av manglende opplysninger om populasjonen, f.eks. dårlige register, krever det både fantasi og mye arbeid å nå dette målet. En kan si at et register er perfekt dersom alle analyseenheter foreligger i registeret bare en gang og med gode identifikasjoner. Slike register er sjeldne, og det typiske er at en må utbedre alvorlige svakheter før en trekker utvalg fra et register. Det foreligger et virvar av eksempler på problemer som oppstår ved trekking av utvalg fra register. Nedenfor skal gis en systematisk beskrivelse som dekker de fleste registerproblemer som kan oppstå, og en skal si litt om hva som kan gjøres for å løse disse problemer. Feilene inndeles

i fire typer:

Type 1. Registeret mangler elementer. Med dette menes at populasjonen inneholder elementer som ikke forekommer i registeret.

Type 2. Elementene forekommer i klynger i registeret. Dette er tilfellet når en ønsker å trekke personer, men bare har adgang til et register over adresser.

Type 3. Fremmedelementer i registeret. Det vil si at registeret inneholder elementer som ikke tilhører den populasjonen en ønsker å lage en undersøkelse i.

Type 4. Dubletter. Med dette menes at det samme element forekommer flere steder i registeret.

Før en tar et register i bruk til trekking av utvalg, bør en sikre seg å få oversikt over hvilke av de fire feiltyper som forekommer i det aktuelle tilfelle. Etter at en har kartlagt de feiltyper som foreligger, kan en velge å se bort fra dem. Dette kan en f.eks. gjøre i tilfeller hvor hyppigheten av feil er liten. Hvis en f. eks. vet at det mangler elementer i registeret, men at det bare gjelder noen ganske få elementer, kan en velge å se bort fra denne feilen. Tilsvarende for de øvrige feiltyper.

Nedenfor skal vi gi noen enkle metoder for helt eller delvis å løse de problemer som skyldes feil ved registeret.

Manglende elementer

Dersom det mangler elementer i registeret, kan en forsøke å løse dette problemet på to måter. Den ene måten består i at man skaffer seg andre registre som inneholder de elementer som ikke er med i det første registeret, og lager et eget stratum for disse elementer. Den andre metoden består av at man kobler manglende elementer til elementer i registeret. Hvis en f.eks. har et godt register over voksne i en bestand, kan en trekke et utvalg av barn ved å trekke et utvalg av voksne, og betrakte barna til de uttrukne voksne som det trukne utvalget av barn. Ethvert barn vil på den måten få samme trekkesannsynlighet som sine foreldre, og dermed har vi sikret oss at alle barna har kjent sannsynlighet for å bli trukket.

Klynger av elementer

Denne typen problemer oppstår dersom en f.eks. har et register over adresser og ønsker et utvalg av personer. Vi skal nevne tre metoder en kan bruke for å løse slike problemer.

Den første metoden består av at en inkluderer hele klyngen av uttrukne elementer i utvalget. På denne måten får hvert element samme sannsynlighet

for å bli trukket som den klyngen elementet tilhører, og på denne måten kan en sikre seg et sannsynlighetsutvalg av elementer. Ved Byråets arbeidskraftundersøkelser brukes en slik teknikk. En trekker her et selvveiende utvalg av adresser og intervjuer samtlige personer over 15 år innen de uttrukne adresser.

Den annen metode består av at en tilfeldig velger ett element innen hver av de uttrukne klynger tilfeldig. Sannsynligheten for at et element blir med i utvalget vil ikke være den samme for alle elementer. Det vanlige er i slike tilfeller at en veier observasjonene med en faktor som er omvendt proporsjonal med størrelsen på klyngen.

Den siste metoden består av at en trekker et stort utvalg av klynger, og oppretter et register for disse. Fra denne listen trekker en deretter et sannsynlighetsutvalg av de elementer en ønsker å ha med i utvalget.

Fremmedelementer i registeret

Dette problem, som er hyppig forekommende i praksis, løser en oftest ved rett og slett å se bort fra de fremmede elementer. Hvis en f.eks. ønsker å trekke et enkelt tilfeldig utvalg av elementer, kan en trekke et enkelt tilfeldig utvalg fra registeret og deretter kaste de fremmede elementene en har trukket. Det utvalg en da har igjen vil være et enkelt tilfeldig utvalg av elementer fra målepopulasjonen. En mister imidlertid kontrollen med utvalgsstørrelsen, spesielt hvis en på forhånd ikke vet noe om hyppigheten av forekomsten av fremmedelementer i registeret. Dersom det skulle vise seg at det er flere fremmedelementer i registeret enn ventet, og en derfor vil få for få elementer fra målepopulasjonen ved første trekking, kan en legge tilbake hele utvalget til registeret og trekke et nytt utvalg.

I praksis trekker en ofte tilfeldig utvalg ved systematisk trekking fra et register. Det er ofte en ser anbefalt en trekkemetode som går ut på å ta naboelementet dersom det uttrukne element ikke tilhører målepopulasjonen. En slik prosedyre bør brukes med forsiktighet, idet den gir hvert element en trekkesannsynlighet som er proporsjonal med antall blanke som ligger foran elementet på registeret. Dette kan en unngå ved å bruke metoden nevnt ovenfor.

Dubletter

Dette problem oppstår når en f.eks. skal trekke et utvalg av husholdninger på grunnlag av et personregister. Avhengig av den informasjon en har om registeret, finnes det to måter å løse dette problemet på. Den første metode lar seg lettest beskrive ved et eksempel. Hvis en ønsker å trekke et utvalg av husholdninger med samme sannsynlighet fra et personregister, kan en

på forhånd lage en regel om at husholdningen er trukket bare i de tilfellene hvor eldste person innen husholdningen blir trukket. En trekker deretter et likt sannsynlighetsutvalg av personer. For hver av de uttrukne personene avgjøres det om denne person er eldst innen den husholdning personen tilhører, og i slike tilfeller betraktes husholdningen for trukket. Dersom den uttrukne personen ikke er eldst innen den husholdning personen tilhører, betraktes ikke husholdningen som trukket. Det på denne måten framkomme utvalg av husholdninger, vil være utvalg av husholdninger med hver husholdning trukket med lik sannsynlighet.

Den andre metoden forutsetter at en på forhånd vet hvor mange dubletter det finnes for hver person i registeret. Dersom en trekker et likt sannsynlighetsutvalg fra registeret, vil en ha trukket et ikke likt sannsynlighetsutvalg av personer fra målepopulasjonen. Det vanlige er da at en veier hver person med en faktor som er omvendt proporsjonal med antall ganger personen forekommer i registeret.

En må huske på at det ikke bare er feil ved registeret som kan føre til skjevheter i utvalget. Mange feltrutiner er lagt opp slik at visse grupper av personer ikke har noen sjanse for å komme med i utvalget. La oss se på et eksempel. I Statistisk Sentralbyrå brukte en tidligere følgende utvalgsplan:

Fra personregisteret trekkes et utvalg av personer med navn og adresse. På grunn av flyttinger som ikke er kommet med i registeret, forekommer det at en ny familie er flyttet inn på adressen når intervjueren kommer fram. I slike tilfeller var regelen tidligere at en trakk en ny person tilfeldig fra populasjonen som erstatning for den flyttede. Denne regel er et eksempel på hvorledes en kan innføre skjevheter i utvalget under selve feltarbeidet, idet vi på denne måten utelater alle personer som nylig er flyttet.

3. Målefeil

a. Innledning

I dette avsnittet skal gis en systematisk oversikt over de viktigste målefeil. En skal særlig ha i tankene målefeil som forekommer i forbindelse med intervjuundersøkelser.

For å definere hva som menes med målefeil, skal det antas at det for enhver enhet finnes en sann verdi for det observerte kjennemerke. Det kan i noen tilfeller by på store problemer å definere hva som menes med sann verdi, spesielt når det er tale om holdninger, ønsker og liknende, men dette skal ikke tas opp her. Denne verdi tenker en seg er helt uavhengig av måten undersøkelsen utføres på, f.eks. av måten en spør på og hvem som spør. I en statistisk undersøkelse er det denne sanne verdi en ønsker å måle. Av forskjellige

årsaker vil det observerte svar avvike fra den sanne verdi i mange tilfeller. Dette avvik skal kalles det individuelle svarsavvik. Det skal gis noen eksempler på hvordan slike avvik oppstår i praksis:

Dersom en intervjuer av skjødesløshet avmerker feil svar på spørreskjemaet, oppstår det et svarsavvik. Det er lite sannsynlig at feil som skyldes skjødesløshet er systematisk. I ett tilfelle vil intervjueren registrere en alder som ligger over den riktige, og i andre tilfeller en alder som ligger under. Når en tar gjennomsnittet over flere observasjoner, vil slike feil ofte oppveie hverandre, slik at effekten på den gjennomsnittlige alder er liten. I motsetning til hva som var tilfelle i eksemplet ovenfor, er det situasjoner hvor intervjuerens personlighet fører til at den intervjuede svarer på en bestemt måte. Et eksempel er tendensen til å svare sosialt ønskelig ut fra de normer respondenten selv identifiserer seg med, og kanskje i høy grad ut fra de forventninger som tillegges intervjueren. For en bestemt intervjuer vil slike feil ikke oppveie hverandre, men fører til en systematisk skjevhet i svarene. Til slutt skal nevnes et eksempel tatt fra forbruksundersøkelsene. En har her observert at forbruket av tobakk og alkohol blir sterkt underrapportert, uansett intervjuer. Slike avvik, som skyldes at det ikke er sosialt akseptert å nyte for mye tobakk og alkohol, vil heller ikke oppveie hinannen innen utvalget.

For å komme videre i studiet av målefeil, kan en bruke en modell som er svært lik målemodeller en bruker innen f.eks. fysikken. En tenker seg at svaret på et spørsmål er en tilfeldig variabel, som ved repeterte målinger kan gi forskjellige resultater. Målingene tenker en seg utført under de samme generelle forhold, det vil si samme spørsmål, samme tidspunkt og samme betingelser ellers. Hvert svar betraktes da som en realisasjon av den stokastiske variable. Gjennomsnittet av slike gjentatte målinger, i den teoretiske statistikk kalt forventningen til den stokastiske variable, kan være lik den sanne verdi, eller kan være forskjellig fra den. Avviket mellom dette gjennomsnitt og den sanne verdi kalles svaravviket. Variasjonene i målingene, i den teoretiske statistikk kalt variansen til den stokastiske variable, kalles svarsvariansen. Da noen individuelle svarsavvik vil være positive og andre negative, er det mulig at de opphever hverandre over hele utvalget. Dette kan imidlertid ikke skje med svarsvariansen, da variansen alltid er positiv.

b. Kilder til svarsvarians og skjevheter

Et intervju er et resultat av et samspill mellom intervjueren, spørreskjema og den intervjuede, og en kan derfor på en måte si at det er urealistisk å diskutere de forskjellige kilder til feil. For å gjøre

diskusjonen mer konkret, skal vi likevel liste opp en del kilder til svarsfeil og svarsvarians. I Sagberg (1976) er det gjort et studie om feilkilder knyttet til interaksjon mellom intervjuer og respondent. Dette arbeid ligger også til grunn for deler av dette avsnittet.

Egenskaper ved intervjueren

Det har lenge vært kjent at personlige karakteristika ved intervjueren, kjønn, alder, utdannelse og sosiale status, kan influere på de svar vedkommende får. En rekke studier har vist forskjellige resultater avhengig av intervjuerens egenskaper. Wilkens (1949) har vist at veteraner fra den annen verdenskrig rapporterte flere medaljer til eldre kvinnelige intervjuere enn til andre. I Sagberg (1976) er det vist at folk som intervjues om psykisk helse rapporterer flere problemer overfor kvinner enn overfor menn.

Egenskaper ved intervjueren påvirker altså respondentens svar på flere måter. Mer og mindre åpenbare egenskaper som kjønn, alder, sosial status m.v. kan influere på respondenten ved at han tillegger intervjueren bestemte forventninger, avhengig av de nevnte karakteristika, og at hans svar tenderer i retning av konformitet med disse forventninger. Hva som respondenten opplever som sosialt ønskelig, kan også variere avhengig av om intervjueren er mann eller kvinne. Også intervjuerens holdninger kan via de antipatier eller sympatier som vekkes hos respondenten virke inn på resultatet. Muligheten for at slike forhold kan påvirke resultatene, avhenger selvsagt av hvor strukturert spørreskjemaet er.

I Hyman (1954) fant en at svarene avhenger mye av de forventninger som intervjueren hadde til respondenten. En fant at svarene til spørsmålene tidlig i intervjuet gav intervjueren et inntrykk av respondentens holdning. Dette igjen hadde innflytelse på intervjuerens tolkning av svar på spørsmål seinere i intervjuet.

Tendens til å svare sosialt ønskelig

For mange spørsmål er svaralternativene ulike med hensyn til hvor akseptable eller ønskelige de er ut fra sosiale normer. På spørsmålet: "Snyter De på skatten?" er åpenbart det sosialt ønskelige svaret: "Nei", og en må derfor anta at det velges oftere enn de faktiske forhold skulle tilsi. En forventer at respondenten snarere svarer det som betraktes som mest ønskelig i stedet for det som best karakteriserer ham, dersom det er forskjell mellom disse to alternativene. Denne svartendensen har vært undersøkt i forbindelse med holdningsmålinger og personlighetstester, Edwards (1957), og det ser ut til å være en betydelig variasjon i denne effekten. Betydningen av slike effekter er naturligvis avhengig av hvor klare normene er for hva som er ønskelig.

Respondentens oppfattelse av spørsmålet

Forekomsten av svarsavvik avhenger også av skjemaets form og innhold. Av erfaring vet en at kvaliteten på svarene avhenger av hvor klar, vesentlig eller relevant problemstillingen er for respondenten. Stort sett regner en med at spørsmål om konkrete atferdsaspekter vil være mindre utsatt for svarsavvik enn spørsmål om meninger, holdninger, ønsker osv.

Andre feilkilder

Respondenten kan gi feil svar fordi han mangler nødvendige kunnskaper, eller fordi han rett og slett ikke husker det riktige svaret.

c. Identifisering og estimering av målefeil

På bakgrunn av det store antall kilder til målefeil som er nevnt ovenfor, er det naturlig at en må legge ned mye arbeid for å finne ut hyppigheten av målefeil i de enkelte undersøkelser. Hvilken effekt har disse målefeil på en vanlig undersøkelse? Er det slik at typisk forekommende feil er av en slik art at de opphever hverandre over hele samplet? Kan forekomsten av målefeil fullstendig ødelegge kvaliteten til en undersøkelse?

På tross av at det har vært nedlagt en del arbeid på å besvare slike spørsmål, er svarene fortsatt foreløpige, og karakterisert ved at de bare gjelder en viss type undersøkelser, foretatt under visse betingelser. De observasjoner en har gjort om målefeil, lar seg vanskelig generalisere til andre områder og andre undersøkelser, enn der hvor observasjonene er gjort.

Nedenfor skal beskrives noen metoder som brukes i Byrået for å oppdage målefeil i datamaterialene. Metodene skal inndeles i to typer, kontroll på individnivå og kontroll på aggregerte data. Ved å kontrollere individ for individ får en et uttrykk for totalt antall målefeil, ofte kalt bruttoavvik, mens en ved å kontrollere aggregerte data får et uttrykk for nettovirkningen av feilene.

Kontroll på individnivå

Som ved enhver annen liknende institusjon, er det vanlig i Byrået at skjemaene blir revidert når de kommer inn. Hensikten med denne revisjonen er først og fremst å oppdage om det er gjort formelle feil ved utfyllingen av skjemaene, om det mangler svar på noen av spørsmålene, og å oppdage klare inkonsistenser i skjemaet. På den ene side er det klart at denne revisjonen øker kvaliteten på resultatene vesentlig, men på den annen side er det like klart at kun ganske få typer feil kan oppdages under en slik revisjon. I Byrået er det til nå ikke gjort noen forsøk på å finne ut

hvor effektiv denne revisjonen er, men et visst forsøk er gjort i Edvards (1956).

Kontroll mot annen statistikk

I prinsippet bør en alltid, når det er mulig, kontrollere opplysninger gitt på et skjema mot annen informasjon en har fra andre kilder. I Byrået er dette blant annet blitt gjort i forbindelse med valgundersøkelsene. Thomsen (1971).

I tillegg til at det sjelden er mulig å kontrollere opplysningene med annen statistikk, er det spesielt to vanskeligheter ved å gjennomføre slik kontroll i praksis. For det første er kontrollen uhyre ressurskrevende, og for det andre er det sjelden at den statistikk en kontrollerer mot er helt nøyaktig. Tolkningen av avvik blir derfor vanskelig, og resultatene er av begrenset verdi.

Gjentatt intervjuing

I de seinere år er det blitt mer alminnelig å foreta gjenintervjuing i forbindelse med store undersøkelser. I USA har en foretatt kvalitetsundersøkelser i forbindelse med folketellingene siden 1940.

I forbindelse med Folke- og boligtellingen 1970 ble det i Byrået utført en kvalitetsundersøkelse. For dette formål ble det trukket et utvalg av personer som alle innen en måned etter Folketellingen ble intervjuet om de samme forhold som var kartlagt i tellingen. Resultatene fra denne kvalitetsundersøkelse tyder på at en underestimerer antall personer med inntektsgivende arbeid i folketellingen. Denne underestimering er kraftigst for kvinnene, og spesielt kraftig for kvinner med deltidsarbeid. Statistisk Sentralbyrå (1976).

Ved tolkning av resultatene fra en kvalitetsundersøkelse står en overfor liknende problemer som de som ble nevnt i avsnittet ovenfor. For å sikre seg at resultatene fra kvalitetsundersøkelsen gir uttrykk for en svarsskjevhet, må en legge ned mye arbeid for at svarene på kvalitetsundersøkelsen skal bli sanne. I praksis vil det aldri være mulig å sikre seg slik at en får helt korrekte svar i kvalitetsundersøkelsen, men en kan gjøre mye for at svarene blir så korrekte som praktisk mulig. I forbindelse med Byråets kvalitetsundersøkelse av folketellingen, satte en i verk følgende tiltak for å sikre seg så korrekte svar som mulig:

1. Intervjuerne ble spesielt utdannet for å foreta kvalitetskontrollundersøkelsen.

2. Intervjuerne ble instruert om bare å hente informasjon fra den person som var valgt ut til undersøkelsen. I motsetning til hva som er tilfelle i en folketelling, ble ingen av skjemaene fylt ut på grunnlag av indirekte intervju.
3. I kvalitetsundersøkelsen hadde en et stort antall spørsmål som hadde til hensikt å kartlegge respondentens aktiviteter gjennom hele året. Dessuten inneholdt skjemaet en del kontrollspørsmål.

I tillegg til dette ble det lagt ned mye arbeid i koding og revisjon av data, og i tvilstilfelle ble respondenten kontaktet på nytt. En kan derfor regne med at svarene i kvalitetsundersøkelsen er så korrekte som det er praktisk mulig å få dem.

Intervjuerne var ikke utrustet med de svar som respondenten hadde gitt til folketellingen. Det er stor uenighet blant statistikerne om det er fornuftig å opplyse intervjuerne om det opprinnelige svar. Marks, Mauldin og Nisselson (1953). I de seinere år er det imidlertid blitt mer alminnelig å opplyse intervjueren om det opprinnelige svar før kvalitetsundersøkelsen foretas.

Kontroll på aggregerte data

Mange av de vanskeligheter som ble nevnt i forbindelse med kontroll på individnivå, er vesentlig mindre når det gjelder kontroll av gjennomsnitt mot annen statistikk. Det er da også vanlig praksis i Byrået å sjekke alders- og kjønnsfordeling i utvalgene mot alders- og kjønnsfordelingen tatt fra personregisteret. Bortsett fra kontroll på enkle demografiske kjennemerker, har det likevel vist seg en rekke vanskeligheter når en skal sammenlikne resultater fra én undersøkelse med tilsvarende resultater fra en annen. Ofte er definisjoner og klassifikasjoner ikke identiske, slik at sammenlikningene styrker eller svekker tilliten til nøyaktigheten, men kan sjelden brukes til en effektiv kontroll. Høst (1975).

En annen begrensning ved kontroll på aggregerte data er at de ikke dannen grunnlag for å vurdere kvaliteten til tall på lavere aggregeringsnivå. Dette problemet blir enda alvorligere dersom en har målefeil på de kjennemerker man bruker under disaggregeringen. Selv om to kjennemerker hver for seg ikke viser noe alvorlig avvik på aggregert nivå, kan resultatene som kommer fram ved en krysstabulering av de to kjennemerker være beheftet med meget store feil. Slike feil vil ikke bli oppdaget ved å teste de to kjennemerkene hver for seg på aggregert nivå.

Intervjuervarians

Det har lenge vært kjent at forskjellige intervjuere kan få forskjellige svar på de samme spørsmål. De første arbeider som tok sikte på å estimere effekten av intervjuervarians er beskrevet i Mahalanobis (1946). Siden

er det gjort en hel serie av undersøkelser om intervjuervariansen. Interesserte lesere vil finne en fyldig liste med referanser i Moser og Kalton (1975). Også i Byrået er det utført en studie av intervjuervariansen, Sagberg (1976).

En av vanskelighetene ved estimering av intervjuervarians er at en bestemt intervjuers arbeid stort sett er konsentrert innen et bestemt geografisk område. Den observerte variasjonen mellom svarene fra intervjuerne skyldes derfor delvis en intervjueffekt og delvis en geografisk effekt. Dersom en ønsker å måle den del av variasjonen som skyldes intervjuerne, må en allokere de uttrukne respondenter tilfeldig på intervjuerne. Dette kan føre til at intervjueren må reise langt, hvilket øker utgiftene til undersøkelsen vesentlig.

d. Kontroll av målefeil

I motsetning til hva som er tilfelle med utvalgsvariansen, har det i praksis vist seg meget vanskelig å kontrollere svarsvariansen. De anbefalinger en finner i litteraturen er delvis uklar, og delvis motstridende. I Hansen (1953) finner en følgende utsagn:

> "The Paucity of dependable data on response errors is unquestionably the greatest present obstacle to sound survey design."

Skjønt det er foretatt en serie studier av målefeil siden den gang, er hovedinnholdet i sitatet fortsatt riktig.

I Byrået har en stort sett brukt tre metoder for å kontrollere målefeil.

Arbeid med spørreskjema

Det sier seg selv at en av de viktigste kilder til svarfeil er spørreskjemaet. Det legges derfor ned mye arbeid på å vinne erfaring med spørsmål som forekommer hyppig. For dette formål er det i Byrået opprettet et arkiv av spørsmål, hvor en samler de erfaringer man har gjort med spørsmål.

Ansettelse, utdanning og kontroll av intervjuerne

Spørsmål om hvilke personer som egner seg særlig godt til å bli intervjuere, har opptatt mange institusjoner, Sudman (1967). De resultater en er kommet fram til, varierer meget, men det ser ut som om voksne kvinner har visse fortrinn. I Byrået ansettes intervjuerne ved at en først velger ut en del egnede kandidater. Kriterier for denne utvelging er bosted, eventuelt yrke, alder, om de disponerer bil og telefon m.v. Kandidatene får

deretter tilsendt de to første brevene av et brevkurs på 4 brev. Etter besvarelsene på skriftlige oppgaver i brevene velges et mindre antall ut for brev 3, og etter en ny utsiling utvelges den eller de intervjuere som får tilbud om ansettelse. Nærmere om opplæringen av intervjuerne. Brevkurset som alle Byråets intervjuere skal igjennom, består av følgende 4 brev.

Brev 1

Etter et innledende kapittel om Statistisk Sentralbyrå som institusjon og generelle betraktninger om nytten av statistikk, redegjøres det inngående om skillet mellom fullstendige tellinger og utvalgsundersøkelser. De vanligste statistiske begreper som telleenhet, statistisk masse, tilfeldig utvalg, frafall m.v. beskrives og forklares gjennom eksempler. Generelle metoder for å redusere frafall omtales.

Brev 2 viser gangen i en utvalgsundersøkelse, med beskrivelse av de forberedende drøftinger, planleggingsfasen, oppbyggingen av spørreskjema, bearbeiding og resultater. Intervjuerens viktige rolle, betydningen av nøyaktighet og en nøytral holdning i forholdet til IO inngår som et ledd i dette.

Brev 3 er viet det praktiske arbeid, med omtale og delvis repetisjon av sentrale begreper. I tillegg behandles rene administrative spørsmål, herunder arbeidsavtalen (bilag 2), kontakten med Byrået, intervjuernes opptreden og ansvar og Byråets kontroll av intervjuernes arbeid.

Brev 4 er for en stor del viet arbeidskraftundersøkelsene, med en detaljert gjennomgåing av definisjoner i tilknytning til spørsmålene og forklaring av den tekniske framgangsmåte ved utfyllingen av skjema. Selv om brevet i det vesentlige tar sikte på utfylling av det spesielle skjema for Arbeidskraftundersøkelsene, er de vanlige regler for skjemautfylling, med henvisningsmønster for ulike grupper og individer og betydningen av nøyaktighet med dette, like relevant for andre skjematyper.

I brevet inngår ellers opplæring i yrkeskoding i henhold til standard, med eks. og øvelser.

Hvert enkelt av de fire undervisningsbrevene avsluttes med prøveoppgaver, dels med spørsmål som har tilknytning til undervisningsstoffet, dels praktiske øvelser med utfylling av ulike skjematyper. Oppgavene rettes

og bedømmes og resultatene regnes som avgjørende for det endelige tilbud om engasjement.

For Oslo og nærmeste omegn innkalles kandidatene til personlig konferanse før de får tilbud om engasjement.

Samlingskurs - internatkurs

Et 3-dagers internatkurs er obligatorisk for alle intervjuere, uansett utdanning og/eller praksis.

Samlingskurs for nye intervjuere holdes 3-4 ganger i året, tilpasset avgang og tilgang, hvert kurs med 20-25 kursdeltakere. En stor del av undervisningen foregår i grupper på 6-8 deltakere. Faste gruppeledere er funksjonærer ved Intervjukontoret med lang erfaring i rettleiing av intervjuere. Tre av gruppelederne har hatt studietur til Sverige hvor de bl.a. har deltatt i tilsvarende kurs. Nye gruppeledere deltar ellers som observatører ved ett eller to samlingskurs før de tildeles egne grupper. Gruppelederne deltar også fra tid til annen i datainnsamling (intervjuing). Det er for tiden (1977) 5 kvalifiserte gruppeledere ved kontoret.

I gruppene gjennomføres praktiske øvelser, med skjemautfylling og intervju. I tillegg repeteres og utdypes de viktigste avsnitt i brevkurset. Det kjøres lydbånd, med eksempler fra aktuelle intervjusituasjoner, delvis innspilt av aktive intervjuere, kandidatene intervjuer også hverandre og gruppelederne. Det legges særlig vekt på å trekke fram vanlige motforestillinger hos IO og å finne argumenter for å bryte disse ned. Intervjuernes taushetsplikt, opptreden og medansvar når det gjelder materialets kvalitet er et gjennomgangstema.

Det er, både i brevkurset og ved samlingskursene, lagt stor vekt på å motivere kandidatene for intervjuarbeidet ved å søke å klarlegge for dem nødvendigheten av statistikk, bl.a. for planleggingsformål. Den personlige kontakt en oppnår i gruppene på samlingskurset er av vesentlig betydning for det videre samarbeid mellom kontoret og intervjuerne.

Selv om gruppeundervisningen foregår mer som "samtale over bordet" enn som forelesning, foreligger det manus for alle gruppetimer.

I tillegg til gruppeundervisningen holdes det 1-2 foredrag pr. dag. Det nyttes planleggere og andre forelesere til gjennomgåing av utvalgsplan, intervjuteknikk, gangen i en statistisk undersøkelse, vurdering av utvalgenes størrelse, frafallsårsaker og konsekvenser av frafall.

Gruppetimer og forelesninger foregår i tiden kl. 0830-1230 og 1400-1730. Kveldene nyttes dels til studier, dels til samtaler om kontorets virksomhet og til spørsmål av mer lokal art.

Spesialopplæring

Ved større og mer kompliserte undersøkelser gjennomføres endagskurs a ca. 8-9 undervisningstimer. Dagskurs ble blant annet gjennomført i forbindelse med Helseundersøkelsen 1975.

Løpende informasjon og kontroll

Intervjukontoret har løpende kontakt med de 365 intervjuerne, gjennom korrespondanse og over telefon. Det arrangeres også fra tid til annen kontaktmøter med mindre grupper av intervjuere, hvor formålet bl.a. er å utveksle erfaringer og å stimulere innsatsen. (Unngå frafall.) Det finnes videre et meldingsblad, "Intervjuer'n" som utkommer med 3-4 nr. i året.

Ellers følger det med en detaljert instruks ved alle undersøkelser. Instruksen gir beskrivelse av formålet og nødvendig forklaring til de enkelte spørsmål. Så langt tiden tillater det, gjennomgås innkomne skjemaer for å avdekke mulige misforståelser i skjemautfyllingen. Intervjuerne får umiddelbart beskjed hvis slike feil oppdages.

Det foretas regelmessig sammenlikning av frafall, for landet, for ulike landsdeler og for det enkelte utvalgsområde. Resultatet meddeles blant annet intervjuere med høyt frafall.

4. Frafall

Et av de største problemer i forbindelse med gjennomføringen av en utvalgsundersøkelse er frafallet, dvs. de elementer en ikke får kontakt med, men som er uttrukket til undersøkelsen. I dette avsnittet skal vi se litt på utviklingen av størrelsen på frafallet i Byrået, samt nevne noen av de tiltak som er gjennomført for å redusere størrelsen på frafallet og dens innvirkning på resultatene.

Et av de mest karakteristiske trekk ved frafall er at det har økt kraftig i løpet av de siste 10 år. Økningen varierer litt fra en type undersøkelser til en annen, men stort sett ser det ut som om det for de fleste undersøkelsers vedkommende er skjedd en fordobling av frafallet i løpet av de siste 10 år. I Byråets Stortingsvalgundersøkelser var frafallsprosentene i 1969 og 1973 henholdsvis 9,9 og 19,4 prosent. I Boforholdsundersøkelsene steg frafallsprosenten fra 9,9 prosent i 1967 til 22,9 prosent i 1973. Liknende økninger er observert ved statistiske sentralbyråer i andre land. På tross av at det er lagt ned mye arbeid for å finne årsakene til denne økning, er det stor uenighet blant statistikerne om hvilke forhold som er de viktigste. To årsaker som ser ut til å spille en vesentlig rolle, er en stadig større vegring blant publikum mot å gi svar i utvalgsundersøkelser, samt at store deler av befolkningen tilbringer stadig mindre tid

hjemme, noe som gjør det vanskelig å treffe folk. En må derfor regne med at en for å holde frafallet på et rimelig nivå, må øke kostnadene til datainnsamling.

Årsaker til frafall

For å kunne vurdere frafallets effekt på viktige resultater i en undersøkelse, er det vanlig å inndele dette etter årsaken til frafallet. For dette formål er det utarbeidd et prekodet frafallsskjema til bruk under intervjuernes feltarbeid. Skjemaet inneholder 10 frafallsårsaker, men det er vanlig at en før publisering slår sammen til noen få årsaker. Den klassifikasjon som publiseres varierer fra undersøkelse til undersøkelse, men det nedenstående er blitt en temmelig standardisert inndeling i mange typer undersøkelser, kanskje først og fremst intervjuundersøkelser.

1) Nekter
2) Ikke å treffe
3) Sykdom
4) Annen årsak

N e k t e r

Ved nesten alle statistiske undersøkelser vil en ha en gruppe av personer som nekter å gi de ønskede opplysninger. Dette skyldes i det vesentlige to faktorer, egenskaper ved IO og den intervjuteknikk som brukes.

Det er viktig å være oppmerksom på at disse på ingen måte er umulig å påvirke. En må likevel regne med at selv om en gjør store anstrengelser for å redusere gruppen av nektere, vil det alltid bli igjen en hard kjerne, som det ikke er mulig å få tak i uten bruk av eventuelle lovbestemmelser.

I k k e å t r e f f e

Mens nektere er spredt utover hele utvalget, viser det seg at folk i byene er vanskeligere å treffe enn folk i mindre tettbygde strøk. Dessuten er det lettere å treffe en eller annen i en husholdning enn et bestemt medlem av en husholdning.

Intervjutidspunktet har også mye å si. Vanligvis er kvelden den beste tiden når en ønsker å få tak i andre enn husmødre og eldre. Forarbeid som f.eks. forhåndsavtale eller brev sendt IO på forhånd kan redusere frafallet. Sist, men ikke minst, vil gjenbesøk ofte øke svarprosenten betraktelig.

S y k d o m

Sykdom hos IO selv eller i nærmeste familie er ofte årsak til frafall. Årsaken til å skille ut denne kategorien er at den har en helt spesiell virkning på resultatene i mange undersøkelser, fordi sykdom har en viktig innflytelse på mange andre kjennetegn for personen.

Effekter av frafall

For å studere effekten av frafall skal en tenke seg populasjonen inndelt i to "strata". Det ene stratum består av de personer som ville komme med i utvalget dersom de ble trukket, og det andre stratum består av de personer som ikke ville komme med i utvalget selv om de ble trukket ut. En slik inndeling er naturligvis litt for enkel, men er allikevel en måte å belyse problemet på.

Utvalget gir ingen informasjon om tilstanden i det andre stratum, men dersom en kan forutsette at tilstanden i stratum 2 er lik tilstanden i stratum 1, er naturlig vis problemet løst. På den andre siden viser erfaringene at dette meget sjelden er tilfelle, Thomsen (1971), Kish (1965).

La N_1 og N_2 være antall enheter i de to strata og la $W_1 = \frac{N_1}{N}$, $W_2 = \frac{N_2}{N}$, hvor $N = N_1 + N_2$. Anta at et tilfeldig utvalg er trukket fra hele populasjonen. Etter intervjuingen har en data for stratum 1, men ingen fra stratum 2. La $\bar{Y}$ være gjennomsnittet i hele populasjonen. Da er

$$E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2) = W_2(\bar{Y}_1 - \bar{Y}_2),$$

hvor $\bar{y}_1$ er gjennomsnittet i utvalget og $\bar{Y}_1$ og $\bar{Y}_2$ er populasjonsgjennomsnittene i strataene 1 og 2, dessuten er $n_1 = W_1 n$, altså antall observasjoner en får tak i.

Bruttovariansen for $\bar{y}_1$ blir

$$\begin{aligned} E(\bar{y}_1 - \bar{Y})^2 &= E(\bar{y}_1 - \bar{Y}_1 + \bar{Y}_1 - \bar{Y})^2 \\ &= E(\bar{y}_1 - \bar{Y}_1)^2 + E(\bar{Y}_1 - W_1\bar{Y}_1 - W_2\bar{Y}_2)^2 \\ &\quad + 2E(\bar{y}_1 - \bar{Y}_1)(\bar{Y}_1 - \bar{Y}) \\ &= \frac{S_1^2}{n_1} + W_2^2(\bar{Y}_1 - \bar{Y}_2)^2 \end{aligned}$$

hvor $$S_1^2 = \frac{\sum_{i=1}^{n_1} (y_i - \bar{Y}_1)^2}{n_1 - 1}$$

Bruttovariansen er altså sammensatt av et variansledd og et ledd som skyldes forventningsskjevhet. En ser at skjevheten er uavhengig av n, dvs. en kan ikke redusere skjevheten ved å øke n.

Hvis det kjennetegn en undersøker er en binær variabel og $\bar{Y}_1$ og $\bar{Y}_2$ derfor er hyppigheter, er $(\bar{Y}_1 - \bar{Y}_2)^2 \leq 1$, dvs. bruttovariansen er mindre enn eller lik $\frac{S_1 2}{n_1} + W_1^2$. Et frafall på 15 prosent vil altså maksimalt gi et tillegg på bruttovariansen på 2,3 prosent.

Måter å regulere frafallet på

Når en skal behandle problemer med frafall, kan en angripe dem fra to sider, nemlig redusere størrelsen på frafallet og/eller forsøke å redusere virkningen av frafallet. Vi skal ta for oss problemet med å redusere størrelsen på frafallet først og se på det andre problemet nedenfor. Følgende metoder kan brukes for å redusere størrelsen på frafallet:

En kan forbedre innsamlingsprosedyren generelt ved å

a) Garantere anonymitet overfor respondenten og kanskje gi denne anledning til å sende svaret direkte til Statistisk Sentralbyrå. En annen måte å sikre anonymiteten på, som vi ikke har forsøkt i Byrået, er foreslått i Warner (1965). Metoden består av at intervjueren gir et spørsmål, f.eks. "Snyter De i skatt?" I stedet for å forlange svar på spørsmålet, ber intervjueren respondenten om å trekke et kort fra en "velblandet" kortstokk hvor det står Ja eller Nei. Respondenten blir da bedt om å opplyse om det som står på kortet er sant eller usant. Når en kjenner fordelingen på Ja- og Nei-kort i stokken, kan en deretter estimere hvor stor prosent som snyter i skatt.
   Metoden er utviklet videre slik at det er mulig å bruke den ved flere svaralternativer enn to.

b) Forsøke å motivere respondenten til å svare gjennom annonser o.l.

c) Forsøke å engasjere respondenten ved f.eks. å åpne med fornuftige spørsmål.

d) Sende bud i forveien til respondenten om at han/hun vil bli oppsøkt av en intervjuer. (En bør her være forsiktig med opplysninger om undersøkelsens art, da en kan påvirke respondentens atferd og dermed resultatet av undersøkelsen.) Thomsen (1971).

e) Velge "fornuftige" intervjutidspunkter og kanskje lage forhåndsavtaler over telefon.

Gjenbesøk

Gjenbesøk eller purring er den mest alminnelige måten å redusere frafallet på, og da spesielt det frafall som skyldes "ingen å treffe". I arbeidskraftundersøkelsene har det eksempelvis vist seg at svarprosenten øker fra 74 til 92 prosent når en foretar 4 gjenbesøk.

I Boforholdsundersøkelsen 1973 er det foretatt en vurdering av effekten av gjenbesøk. Laake (1975). En har undersøkt hvor stor del av utvalget som bor i de ulike hustypene, og tabell 5 viser fordelingen av husholdningenes boligtype fordelt etter hvilket besøk intervjueren oppnådde intervju i.

Tabell 5. Husholdningenes boligtype etter nummer på besøket. Prosent

Besøk nr.		Antall IO som blir besøkt	Våningshus	Enebolig	Rekkehus	2-4-mannsbolig	Blokk	Annet
I alt	100,0		17,8	38,2	7,8	14,9	16,6	4,6
1	100,0	1 535	24,3	39,3	6,6	12,4	13,2	4,2
2	100,0	910	12,5	37,3	9,8	16,5	18,5	5,5
3	100,0	314	7,0	37,6	8,3	19,4	23,6	4,1
4	100,0	82	1,2	30,5	6,1	26,8	30,5	4,9
5	100,0	25	4,0	28,0	12,0	16,0	32,0	8,0
6	100,0	7	0,0	42,9	14,3	14,3	28,6	0,0
7	100,0	1	0,0	0,0	100,0	0,0	0,0	0,0
8	100,0	3	0,0	33,3	0,0	0,0	33,3	33,3
Uoppgitt ...	100,0	29	20,7	51,7	3,4	17,2	6,9	0,0

Tabellen viser at i første besøk svarer 24,3 prosent at de bor i våningshus. Blant de personer som intervjues i annet besøk synker denne andelen til 17,8 prosent. En ser tilsvarende at husholdninger som bor i blokker og 2-4-mannsboliger har en tendens til å falle fra i første besøk. Dersom man ikke korrigerer for at spesielle husholdningstyper har stor sannsynlighet for å falle fra i første besøk, kan estimatorene bli meget skjeve. I Boforholdsundersøkelsen har man altså korrigert skjevheten ved å gjenbesøke husholdninger som falt fra.

Utvalgsundersøkelse på frafallet

Dersom det er uøkonomisk å foreta gjenbesøk fordi frafallet er meget stort og/eller meget spredt, kan det komme på tale å utføre utvalgsundersøkelse

på frafallet. Denne metoden var først brukt i forbindelse med en post-undersøkelse fulgt opp av intervjuere på et utvalg av frafallet. Metoden er lite brukt i Byrået.

Erstatninger

En ser ofte at frafallet reduseres ved å erstatte dette med andre respondenter. Det ekstreme eksempel er kvotasampling, hvor en holder på til en finner noen hjemme. (Det er vel et spørsmål om en kan tale om frafall ved slike utvalgsplaner.) Ved å erstatte er det klart at utvalgsstørrelsen øker, men en må regne med at erstatningene likner mer på dem som en allerede har funnet hjemme, enn dem de er tenkt å erstatte. Erstatninger brukes derfor bare i liten utstrekning i Byrået.

Måte å redusere effekten av frafall på

Selv med store anstrengelser for å redusere størrelsen på frafall, vil det alltid ha en viss størrelse etter at innsamlingen er avsluttet. I visse undersøkelser er frafallet så stort at en må gjøre noe for å redusere effekten av det. Det er derfor ikke uvanlig at en i mange lærebøker finner anbefalinger om å veie de innsamlede data. Den mest vanlige metoden består i å inndele utvalget i grupper etter noen få kjennemerker, og deretter veie de forskjellige gruppegjennomsnitt med en faktor som er omvendt proporsjonal med frafallsprosenten innen gruppen. I Byrået brukes en slik metode ved forbruksundersøkelsene, hvor en har observert et større frafall blant små husholdninger enn blant store. For å korrigere for denne skjevhet i frafall, veier en observasjonene i de små husholdningene med en faktor som er omvendt proporsjonal med frafallsprosenten. En analyse av slike metoder synes å tyde på at en generelt ikke kan vente seg vesentlig reduksjon av virkningene av frafallet. Thomsen (1973). En annen veiemetode er foreslått i Bartholomew (1961). Denne metoden går ut på at observasjonene veies avhengig av hvilket besøk intervjueren oppnådde intervju ved. Metoden er forsøkt brukt i Byrået, men tallene er ikke blitt publisert. Årsaken er at en også her har funnet så små avvik mellom veide og uveide tall, at en har valgt ikke å publisere de veide tall. En tredje veiemetode er foreslått i Polits and Simons (1949). Metoden er ikke blitt forsøkt brukt i Byrået, men resultatene fra andre liknende institusjoner tyder på at også denne veiing har liten effekt på resultatene.

På grunnlag av de analyser som er foretatt, synes det mer fruktbart å arbeide for å redusere størrelsen på frafall enn å utvikle estimeringsmetoder for å motvirke virkningene av frafallet.

VII. PRESENTASJON AV RESULTATENE FRA EN UTVALGSUNDERSØKELSE

1. Innledning

Resultatene fra en intervjuundersøkelse offentliggjøres vanligvis i en eller flere av følgende av Byråets publiseringsserier:

N o r g e s o f f i s i e l l e s t a t i s t i k k (NOS). Her publiseres resultatene fra en undersøkelse vesentlig i form av tabeller. I tillegg gis en del generell informasjon av materialet.

S t a t i s t i s k e a n a l y s e r (SA). Denne serien omfatter publikasjoner der tabeller og figurer blir fulgt av utfyllende tekst, slik at publikasjonene får et mer analytisk preg.

S a m f u n n s ø k o n o m i s k e s t u d i e r (SØS). I denne serien publiseres undersøkelser som ikke er av rent statistisk karakter, blant annet historiske og analytiske studier om økonomiske og sosiale forhold.

A r t i k l e r f r a S t a t i s t i s k S e n t r a l b y r å (ART). Her publiseres kortere arbeider; i første rekke analytiske og historiske studier av mindre omfang enn de undersøkelser som blir offentliggjort i serien SØS.

R a p p o r t e r f r a U n d e r a v d e l i n g e n f o r i n t e r v j u u n d e r s ø k e l s e r (RAPP). Denne serien omfatter primærstatistikk og resultater fra spesielle undersøkelser ved Underavdelingen for intervjuundersøkelser utført på oppdrag fra andre institusjoner.

A r b e i d s n o t a t e r f r a S t a t i s t i s k S e n t r a l b y r å (ANO). I denne serien utgis dokumenter m.v. som er av en slik art at de bør være tilgjengelige i mangfoldiggjort form og som ikke innen rimelig tid kan innpasses i noen av Byråets øvrige publikasjonsserier eller ikke høver for disse. Serien omfatter to rekker som hver betegnes med to store bokstaver:

IB Interne dokumenter som ikke distribueres utenfor Byrået.

IO Notater m.v. som ikke inngår i rekke IB.

Andre publiseringsmåter kan komme på tale, men vanligvis brukes en eller flere av de nevnte serier. Det foreligger retningslinjer for hva som skal inngå i en publikasjon fra Byrået.

Vanligvis presenteres resultatene først i en beskrivende form enten i serien NOS eller serien RAPP, med noen få kommentarer til hovedresultatene fra undersøkelsen. I innledningen til disse publikasjonene gis en kort beskrivelse av formålet, de viktigste ledd i gjennomføringen av undersøkelsen, samt mål for kvaliteten til de publiserte tall i den utstrekning det er mulig. Årsaken til dette er at det primære formål med mange av Byråets undersøkelser er å gi en enkel statistisk beskrivelse av en bestand.

I tillegg til å publisere rene beskrivelser, er det blitt mer alminnelig at det publiseres statistiske analyser, som har som formål å gå i dybden innen mer avgrensede problemstillinger. Det er ingen skarp grense mellom de beskrivende og de analytiske publikasjoner. Gjennom tabellering forsøker en å kaste lys over ulike størrelser eller over utviklingen i tid. Slike resultater publiseres i NOS eller RAPP. Når en derimot publiserer resultater på en måte som gir utfyllende tekst, eller når det blir nyttet matematisk-statistiske metoder, publiseres resultatene ofte i serien SA eller ART. Byrået har i de siste år satset mye på å bruke forskjellige analyseteknikker for å utnytte data best mulig. Dette arbeidet er ennå i sin begynnelse, men en regner med at det i løpet av noen få år vil være i bruk en hel serie matematisk-statistiske metoder for å analysere multivariable sammenhenger i datamaterialet.

2. Beskrivelse av bestand og utvalgsmetode

Uansett i hvilken serie resultatene publiseres, må det gis en nøyaktig beskrivelse av den populasjonen undersøkelsen er tenkt representativ for, samt hovedtrekkene ved den utvalgsmetode og estimeringsmetode som er brukt. I noen tilfeller kan dette avsnittet være så omfattende at en må skrive en egen teknisk rapport og henvise til denne i publikasjonen. De fleste undersøkelser som Byrået utfører, er av en slik art at opplegget av undersøkelsen kan beskrives i samme publikasjon som inneholder data. Lesere med særlig teknisk interesse kan da henvises til relevante tekniske beskrivelser utgitt i seriene ANO eller Artikler. En viktig hensikt med slike beskrivelser er at leseren skal bli gjort oppmerksom på forskjellige begrensninger i materialet. En opplyser derfor om hvilke registre en har trukket utvalgsenhetene fra, samt gir de detaljer ved registeret som er relevante for å forstå trekkemetoden. Dessuten opplyses om tidspunktet for intervjuingen og den tidsperiode opplysningene gjelder for. Detaljer om hvem som kan svare på spørsmålene innen en husholdning, og antall besøk intervjueren skal foreta før en person regnes for frafall, skal også nevnes. I visse tilfeller kan det være nødvendig å referere større deler av instruksen til intervjuerne.

3. Diverse feilkilder og mål for usikkerhet

Alle resultater fra en statistisk undersøkelse er beheftet med usikkerheter eller feil. For å unngå misbruk av statistikk, er det viktig å gi leserne et inntrykk av påliteligheten av de tall en publiserer. Både usikkerheten som skyldes at vi har et utvalg, kalt utvalgsvariansen, og de usikkerheter som skyldes målefeil, frafall, dataoverføringer o.l. burde

ideelt beskrives i alle detaljer. I praksis er dette umulig. Det finnes i dag få veletablerte metoder til å publisere den samlede usikkerheten på publiserte tall, og publiseringsmetodene varierer derfor mye fra land til land. I Byrået konsentrerer en seg vanligvis om å gi summariske beskrivelser av feil som oppstår ved at undersøkelsesbestanden ikke er identisk med målepopulasjonen, av de feil som er oppstått under datainnsamlingen og bearbeidingen. I tillegg gis en fyldig redegjørelse for utvalgsvarians og frafall.

a. Utvalgsvarians

I forbindelse med publiseringen av tall fra en utvalgsundersøkelse er det alminnelig å forklare hva som menes med utvalgsvarians. Det er også vanlig at det beregnes utvalgsvarianser for et passende utvalg av de viktigste tall i publikasjonen. Spesielt gjelder dette når variansene lett lar seg beregne ved hjelp av eksisterende program. Usikkerheten i resultatene kan publiseres på flere måter:

1) I mange tilfeller synes det å være tilstrekkelig å gi leseren et omtrentlig anslag på usikkerheten. Vi kan da (ved to-trinns person/husholdningsundersøkelse) gi en tabell over størrelsen $\sqrt{\frac{1{,}5 \cdot p \cdot (1-p)}{n}}$, der p angir prosenttallet i tabellen og n angir antall observasjoner som prosentfordelingen baseres på, og forklarer at dette er et grovt anslag for usikkerheten på prosenttallet.

Ideelt sett burde en kanskje estimere usikkerheten på samtlige tall i en tabell. Imidlertid er arbeidet med å produsere slike tall, samt arbeidet med å lese dem så stort, at en forenkling som den ovenstående synes fornuftig. Det har dessuten vist seg vanskelig å finne gode estimatorer for utvalgsvariansen.

2) Ved noen undersøkelser er det ikke mulig å bruke den enkle formen under pkt. 1). En kan da velge å publisere variansene for visse hovedtall og publisere disse i tabellene, gjerne med parentes omkring. Unntaksvis kan det lages en egen publikasjon med utvalgsvarianser. Denne siste løsning er ennå ikke blitt valgt i Byrået.

3) I tillegg til å si noe om utvalgsvariansen kan det gis metoder for å anslå variansene til differansen mellom to tall i publikasjonen. Slike metoder beskrives vanligvis i et vedlegg.

I tillegg til slike beregninger, er det ofte ønskelig at en i forbindelse med tabellene markerer tall som har spesielt stor utvalgsvarians. Dette gjøres ofte ved at tall markeres med parentes. Som kvalitetsmål nyttes

den relative utvalgsvarians, dvs. utvalgsvariansen dividert med kvadratet av gjennomsnittet av de observerte størrelsene. Dette kvalitetsmål inneholder ikke noe om systematiske feil.

Følgende regler brukes i så fall:

1) Gjennomsnitt, oppblåste tall og fordelinger som baserer seg på færre enn 25 observasjoner, publiseres ikke. Symbolet (:) nyttes i slike tilfeller.

2) Gjennomsnitt og oppblåste tall som baserer seg på mer enn 25 observasjoner, men som har en relativ utvalgsvarians på mer enn 40-45 prosent, settes i parentes. Hvis det i en fordeling i tabellen bare er ett tall som bør ha parentes etter regelen nevnt foran, fjernes parentesen.

Det redegjøres i hvert enkelt tilfelle for hvilken regel som nyttes.

Bakgrunnen for disse regler er å hindre alvorlige mistolkninger av data uten at dette fører til at mange tall ikke blir publisert bare fordi de har en stor relativ usikkerhet. I visse tabellverk ønsker en naturligvis å spalte opp utvalget etter visse kjennemerker, og gi gjennomsnitt for så små grupper som mulig. På den annen side ønsker en ikke å spalte opp materialet i en slik grad at de publiserte tallene er beheftet med for stor usikkerhet. Da kravet til nøyaktighet kan variere fra leser til leser, har Byrået valgt å offentliggjøre fordelinger som baserer seg på så lite som 25 observasjoner. Ofte vil slike tall være av en meget dårlig kvalitet, men de kan danne grunnlag for hypoteser som kan testes gjennom framtidige undersøkelser. Dessuten kan brukeren sitte inne med kunnskaper som kan være av stor verdi når de kombineres med resultatene fra en utvalgsundersøkelse.

b. Frafall

Uansett hvor mye arbeid som legges ned under innsamling, vil det alltid forekomme frafall i forbindelse med statistiske undersøkelser. En bør derfor presentere frafallet på en måte som gjør det mulig for leseren å sjekke frafallets innvirkning på så mange kjennemerker som mulig. At frafallet ikke fører til skjevheter for bestemte kjennemerker, utelukker naturligvis ikke at det kan ha ført til skjevheter på andre kjennemerker. I praksis er det aldri mulig å sjekke frafallet på alle kjennemerker, men ved å kontrollere frafallet på så mange kjennemerker som mulig, gir en likevel leseren et inntrykk av om frafallet kan føre til vesentlige skjevheter.

I tillegg til å fordele frafallet etter visse kjennemerker, fordeles det som oftest etter årsak. I noen tilfeller hvor frafallet er særlig stort, eller hvor frafallet har ført til vesentlige skjevheter i aldersfordeling eller kjønnsfordeling, blir resultatene veid på en måte som tar sikte

på å rette opp den innførte skjevheten. I slike tilfeller beskrives og begrunnes den brukte veiemetoden. Som tidligere nevnt er effekten av slik veiing meget liten.

c. Innsamlings- og bearbeidingsfeil

I langt de fleste tilfeller vil en ha liten oversikt over de feil som oppstår under innsamlingen og under bearbeidingen av data. I spesielle tilfeller kan en likevel, fra tidligere undersøkelser, eller fra spesielle metodestudier, ha kjennskap til viktige feilkilder. I slike tilfeller bør leseren opplyses om dette. Videre kan det være grunn til å nevne faren for misvisende resultater som følge av den måten oppgavene registreres eller grupperes på.

VIII. LØPENDE UNDERSØKELSER

1. Innledning

I Byrået utføres en del undersøkelser som har til formål å beskrive utviklingen i forskjellige kjennemerker over tiden. Slike undersøkelser skal vi kalle løpende undersøkelser. Eksempler på slike er de kvartalsvise arbeidskraftundersøkelser, løpende forbruksundersøkelser fra 1974, og undersøkelsene om lytter- og seervaner. Planleggingen av løpende undersøkelser skiller seg i prinsippet ikke vesentlig fra planleggingen av en enkeltstående undersøkelse, men det er likevel en del spesielle forhold en bør være oppmerksom på. Noen av disse problemer er behandlet i Høst (1975), mens vi i dette avsnitt særlig skal behandle de spesielle utvalgsproblemer som oppstår, samt si litt om de spesielle tolknings- og presentasjonsproblemer en har ved repeterte undersøkelser.

2. Valg av utvalgsmetode

En kan med fordel inndele mulige utvalgsmetoder i forbindelse med løpende undersøkelser i 3 typer:

1. Uavhengig utvalg. Med dette menes at en trekker nye utvalg for hver undersøkelse.
2. Panelundersøkelser. Dvs. at en beholder de samme intervjuobjekter i alle undersøkelser.
3. Roterende utvalg. Dvs. at en lar noen av intervjuobjektene fra en undersøkelse være med i flere etterfølgende undersøkelser.

I Byrået med lett adgang til registeret er metoden som består av uavhengige utvalg den enkleste å gjennomføre, idet en ikke trenger å dra med seg intervjuobjekter fra tidligere undersøkelser. Ulempene ved denne metoden er for det første at den ikke gir mulighet for å følge et individ over tid. En annen ulempe er at variansen til endringstallene normalt blir vesentlig større når en bruker uavhengige utvalg enn når en bruker en av de to andre nevnte metoder. Etter den tilnærmelsesformelen som brukes i Byrået er standardavviket på differansen mellom to prosenttall, p' og p", gitt ved formelen

$$S_{(\hat{p}'-\hat{p}'')} = \sqrt{\left(\frac{p'(1-p')}{n'} + \frac{p''(1-p'')}{n''}\right) \quad 1.5},$$

hvor n' og n" er utvalgsstørrelsen ved henholdsvis første og andre undersøkelsen. Dersom forskjellen mellom p' og p" er liten og utvalgsstørrelsene er like i de to undersøkelser kan dette noe tilnærmet skrives som

$$S_{(\hat{p}'-\hat{p}'')} = \sqrt{\frac{p(1-p)}{n} \quad 3}$$

I tabellen nedenfor er det beregnet standardavvik etter denne tilnærmingsformelen for en del verdier av p og utvalgsstørrelse n.

Tabell 6. Standardavvik for endringer

Tallet på spurte, N	Prosent som danner utgangspunkt for beregningen av endringer				
	5	10	20	30	50
100	3,8	5,2	6,9	7,9	8,7
500	1,7	2,3	3,1	3,6	3,9
1 000	1,2	1,6	2,2	2,5	2,7
2 000	0,8	1,2	1,5	1,8	1,9
3 000	0,7	0,9	1,3	1,4	1,6
5 000	0,5	0,7	1,0	1,1	1,2

Tatt i betraktning at de endringer en ønsker å måle ofte er små, ses det at den relative usikkerhet på endringstallene kan bli meget stor. Erfaringer i Byrået tyder på at det er meget vanskelig å måle mindre endringer ved hjelp av slike uavhengige utvalg.

For å rette på dette bruker en i noen tilfeller panelundersøkelser. I tillegg til at en på denne måten ofte kan redusere usikkerheten på endringstallene vesentlig, gir en slik utvalgsmetode muligheter for å følge de enkelte individer over tiden.

La oss igjen se på standardavviket og differansen mellom to tall. Det kan vises at standardavviket til forskjellen mellom p' og p" kan skrives som

$$S_{(\hat{p}'-\hat{p}'')} \approx \sqrt{1,5\ (\frac{p'(1-p')}{n} + \frac{p''(1-p'')}{n} - \frac{2n_{12}}{n}\ (p_{12}-p'p'')}$$

I formelen betyr n_{12} tallet på personer som er med i begge utvalgene, mens p_{12} er andelen av personene som ikke skifter kategori mellom de to undersøkelsestidspunktene. Dersom vi som tidligere kan anta at p_1 og p_2 begge har samme størrelsesorden, p, blir uttrykket redusert til

$$S_{(\hat{p}'-\hat{p}'')} \approx \sqrt{3\frac{p(1-p)}{n} - \frac{2n_{12}}{n}\ (p_{12}-p^2)}\ .$$

Når en bruker panelundersøkelser er en n_{12} lik n, og en får følgende uttrykk for variansen til endringstall

$$S_{(\hat{p}'-\hat{p}'')} \approx \sqrt{3\frac{p(1-p)}{n} - \frac{2}{n}\ (p_{12}-p^2)}$$

En har altså redusert usikkerheten på endringstallene vesentlig ved å bruke panelutvalg.

De tre største problemer knyttet til bruken av panelutvalg er frafall, panelrepresentativitet, og det faktum at enhetene i panel kan endre sin atferd fordi de er medlem i panelet. Frafallsproblemet kommer naturligvis i forbindelse med de fleste utvalgsundersøkelser, men på grunn av større belastning på intervjuelementet i en panelundersøkelse, blir frafallet etter noen tid gjerne vesentlig større. I Byrået brukes panelutvalg blant annet i forbindelse med innsamling av prisdata. Utvalget av de butikker som rapporterer priser til Byrået hver måned er konstant fra måned til måned. Frafallet i et slikt panel er ca. 10 prosent om året.

Når en bruker det samme panel over lengre tid, vil den populasjon panelet er tenkt å være representativt for endre seg. Dermed ødelegges panelets representativitet. Ved Stortingsvalgundersøkelsen 1973 brukte en stort sett samme utvalg som ved EF-undersøkelsen i 1972. For å gjøre utvalget representativt måtte en derfor trekke et tilleggsutvalg av personer som hadde fått stemmerett i tidsrommet mellom de to undersøkelsene. Utvalgets størrelse ble redusert på grunn av at personer som var med i undersøkelsen i 1972 døde, eller emigrerte. Slik avgang avspeiler den naturlige avgang i den opprinnelige populasjonen, og det ble ikke foretatt erstatninger for disse personer.

I Byrået er det ikke gjort noen studie for å finne ut om personer som er med i panelet endrer sin atferd på grunn av at de er med i panelet. Frykten for at en slik effekt er til stede, er en av grunnene til at undersøkelsene om

lytter- og seeratferd utføres ved hjelp av uavhengige utvalg.

For å bøte på noen av svakhetene ved panelundersøkelsesmetoden, bruker en ofte å skifte ut en del av panelet fra en undersøkelse til en annen. På denne måten reduserer en frafallet, sikrer representativiteten, og gjør det mulig å teste om de personer som er med i flere undersøkelser endrer sin atferd fordi de er med flere ganger. Et slikt opplegg er brukt i flere undersøkelser i Byrået. Av spesiell interesse er de kvartalsvise arbeidskraftundersøkelser, som er lagt opp slik at halvdelen av utvalget er felles for to etterfølgende undersøkelser, og halvdelen av utvalget er felles for undersøkelsene i samme kvartal i to på hinannen følgende år. Med denne rotasjonsplan tar en sikte på å få gode endringstall på kvartaler som følger etter hinannen og for endringer fra et kvartal til samme kvartal året etter.

Av formlene ovenfor ses det at så vel økende overlapping mellom utvalgene som økt sammenheng mellom målene på de to tidspunktene vil føre til mindre usikkerhet på endringstallene.

3. Bruk av spesielle estimeringsmetoder

Uansett hvilken utvelgelsesmetode en velger, vil usikkerheten på endringstallene alltid være store i forhold til de endringer en ønsker å måle. Reduksjonen av den relative usikkerheten ved hjelp av økte utvalgsstørrelser vil gjerne falle dyrt, fordi utvalget skal økes ved samtlige undersøkelser. Det er ikke tilstrekkelig å øke nøyaktigheten for en av undersøkelsene. Det er derfor rimelig å spørre seg om det ikke er mulig å utvikle estimeringsmetoder som for gitt utvalgsstørrelse kan redusere usikkerheten i forhold til vanlig gjennomsnitt. Slike metoder finnes og kalles vanligvis sammensatt estimering. En sammensatt estimator er en veiet sum av estimatorer ved flere tidspunkter som dels er basert på hele utvalget ved det aktuelle tidspunkt, dels på den delen av utvalget som er felles ved flere tidspunkter. Vekten er valgt slik at en sammensatt estimator blir forventningsrett. Studier som er blitt gjort i Byrået tyder på at presisjonsgevinsten ved å bruke sammensatte estimatorer i arbeidskraftundersøkelsene, er den samme som en ville ha oppnådd ved å øke utvalgsstørrelsen fra 12 000 personer til ca. 16 000, Dagsvik (1975).

En annen estimeringsmetode som er brukt i Byrået, gjør bruk av at vi vet at der er sesongvariasjoner i sysselsetting. Ved å trekke inn slik a priori kunnskap kan en finne estimatorer, som ofte er bedre enn sammensatt estimering. Dagsvik (1976), Thomsen (1976).

4. Noen erfaringer om bruken av løpende undersøkelser

Problemet som oppstår når en vil estimere endringer ved hjelp av gjentatte undersøkelser, er drøftet utførlig som ledd i planleggingen av de løpende forbruksundersøkelser, Skarstad (1974). Stort sett kan en si at erfaringene i Byrået tyder på at det er uhyre vanskelig å måle mindre endringer ved hjelp av

utvalgsundersøkelser. I opplegget av arbeidskraftundersøkelsene er det lagt stor vekt på mulighetene for å estimere endringene. Utvalget til undersøkelsene er ca. 12 000 personer, altså langt mer enn det er vanlig for Byråets intervjuundersøkelser. Dessuten er rotasjonsplan og estimeringsmetode skreddersydd for disse undersøkelser. Til tross for dette har det likevel vist seg at de endringstallene som blir beregnet ut fra arbeidskraftundersøkelsene ikke er nøyaktige nok til å tilfredsstille behovet for opplysninger. I andre land hvor en legger større vekt på å kunne måle endringer i sysselsettingen ved hjelp av arbeidskraftundersøkelsene, er utvalgene opptil 60 000 personer i hver undersøkelse.

Når en tenker på å repetere en undersøkelse for å anslå endringene fra forrige undersøkelse, er det derfor veldig viktig at en nøye tenker gjennom hvor store endringer en ønsker å estimere. En slik overveielse kan føre til at en gir opp planene om å forsøke å estimere endringen ved hjelp av en ny undersøkelse. Hvis en derimot utfører undersøkelser periodisk over lengre tid, vil en kunne få et bilde av utviklingen over tid. Ved å se på resultatene fra alle undersøkelsene samtidig, kan en få et klart bilde av utviklingstendensene i flere kjennemerker. Endringene fra en undersøkelse til en annen er derimot normalt av veldig liten betydning, da de er beheftet med altfor store feil. Dette har ført til at en alltid publiserer resultatene fra en undersøkelse sammen med resultatene fra alle de tidligere undersøkelsene, gjerne i et diagram, som gir leserne en mulighet til å oppdage eventuelle tendenser i utviklingen. For de kjennemerker hvor hyppighetene går opp og ned, uten klar tendens, kan leseren få et visuelt inntrykk av den tilfeldige variasjon i resultatene.

Rent administrativt er det store fordeler ved å utføre mindre hyppige undersøkelser, sammenliknet med større og sjeldnere undersøkelser, idet en på denne måten får mer kontinuitet i arbeidet. Ved å bruke mindre løpende undersøkelser slipper en å bygge opp en stab for planlegging og gjennomføring av en større undersøkelse, for å bygge denne ned etter at undersøkelsen er gjennomført. Ved at de samme personer over lengre tid arbeider med identiske eller nesten identiske undersøkelser, vil en få mye bedre datakvalitet og tolkningen av data vil bli bedre. Dette er en av grunnene til at forbruksundersøkelsene i de siste årene er utført som mindre løpende undersøkelser istedenfor sjeldnere og store undersøkelser. Ennå er det nok vanskelig å gi klart svar på hva som er best, men det er visse tegn på at en i forskjellige statistiske sentralbyråer går mer over til mindre løpende undersøkelser. Det er f. eks. foreslått å nedlegge de ti-årlige folketellinger for å erstatte disse med større årlige utvalgsundersøkelser.

LITTERATUR

[1] Bartholomew, D. J. : A method of allowing for "not-at-homes" bias in sample surveys. Applied statistics, 1 (1961).

[2] Cochran, W. : Sampling techniques. J. Wiley (1963).

[3] Dagsvik, J. : Prosjektskisse til prosjektet "estimering i AKU ved hjelp av metoder fra tidsserieanalysen". Stensil JDa/MFo, 9/7-75 Statistisk Sentralbyrå.

[4] Dagsvik, J. : Estimering og prediksjon ved bruk av metoder fra tidsserieanalysen i Byråets arbeidskraftundersøkelser. Arbeidsnotat IO 76/4. (1976) Statistisk Sentralbyrå, Oslo.

[5] Des RAJ. : Sampling Theory. Mc Graw-Hill Book Company (1968).

[6] Edwards, F. : Readings in market research: A selection of papers by british authors. The British Market Research Bureau. (1956) London.

[7] Edwards, A.L. : Techniques of attitude scale constructions. Appleton-century-crofts (1957) New York.

[8] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. Sample survey methods and theory. Vol. I and vol. II. (1953) Wiley, New York.

[9] Hoem, J.M. : Statistisk Sentralbyrås utvalgsundersøkelser. Elementer av det matematiske grunnlaget. Artikler fra Statistisk Sentralbyrå nr. 58. (Oslo: Statistisk Sentralbyrå 1973).

[10] Hyman, H.H. and others: Interviewing in social research. University of Chicago Press. (1954) Chicago.

[11] Høst, S. : Gjentakelse av intervjuundersøkelser: Noen kommentarer basert på erfaringer fra planleggingen av Ferie- og friluftslivundersøkelsen 1974. Arbeidsnotater fra Statistisk Sentralbyrå, IO 75/8. (Statistisk Sentralbyrå 1975).

[12] Høst, S. : Ja-effekten (Acqiescence response set) og bruken av påstander i intervjuundersøkelser. Arbeidsnotater fra Statistisk Sentralbyrå, IO 75/34. (Statistisk Sentralbyrå 1975).

[13] Høst, S. : Omfanget av massemediabruk og teaterbesøk: En vurdering av noen opplysninger fra Tidsnyttingsundersøkelsen 1971/72. Arbeidsnotater fra Statistisk Sentralbyrå. IO 75/39. (Statistisk Sentralbyrå 1975).

[14] Kemsley, W.F.F.: Sampling Errors in the Family Expenditure Survey. Applied Statistics, 15, 1-14, (1966).

[15] Kish, L. : Survey Sampling (Wiley, New York, (1965)).

[16] Laake, Petter (1974): "Estimering av variansen til estimatoren for populasjonsverdien a_0 for Oslo i Byråets intervjuundersøkelser." Arbeidsnotat IO 74/7.

[17] Laake, P.: Frafallsproblemet i Byråets intervjuundersøkelser. En oversikt over utviklingen og noen forslag til metoder som kan redusere frafallet eller virkningen av frafallet. Arbeidsnotater fra Statistisk Sentralbyrå, IO 15/20. (Statistisk Sentralbyrå 1975).

[18] Laake, P. : Estimering av totaler med en to-trinns utvalgsplan der de primære utvalgsområdene trekkes med ulik sannsynlighet i første trinn. Arbeidsnotater fra Statistisk Sentralbyrå, IO 74/49. (Statistisk Sentralbyrå 1974).

[19] Laake, P. : Estimering av sysselsetting i geografiske regioner: Om estimatorenes skjevhet, varians og bruttovarians. Artikler fra Statistisk Sentralbyrå nr. 88. (Statistisk Sentralbyrå 1976).

[20] Laake, P. : An evaluation of synthetic estimates of employment. Stensil PLå/GHu, 28/3-77, Statistisk Sentralbyrå.

[21] Laake, P. : A note on a prediction approach to synthetic estimators. Stensil PLå/GHu, 29/3-77. Statistisk Sentralbyrå.

[22] Mahalanobis, P.C. : Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 109, (1953).

[23] Moser, C.A. and Kalton, G. : Survey Methods in social investigations. Heinemann Educational Book Ltd. (1973). London.

[24] Marks, E.S., Mauldin, W.P. and Nisselson, H.: The Post Enumeration Survey of the 1950 census: A case history in survey design. Journal of the American Statistical Association, 48, (1953).

[25] Politz, A. and Simmons, W. : I. An attempt to get the "not-at-homes" into the sample without call-backs. II. Further theoretical considerations regarding the plan for eliminating call-backs. Journal of the American Statistical Association, 44, (1949).

[26] Raj, D. : Sampling theory. Mc Graw-Hill Book Company (1968). New York.

[27] Rideng, A.: Klassifisering av kommunene i Norge 1974. Artikler fra Statistisk Sentralbyrå nr. 67. (Statistisk Sentralbyrå 1974).

[28] Sagberg, F. : Om validiteten av intervjuundersøkelser: Feilkilder knyttet til den sosiale interaksjon mellom intervjuer og respondent. Arbeidsnotater fra Statistisk Sentralbyrå, IO 76/21. (Statistisk Sentralbyrå 1976).

[29] Skarstad, O. : Systematiske målefeil ved registrering av forbruksutgifter ved regnskapsføring. Arbeidsnotater fra Statistisk Sentralbyrå, IO 74/39. 1974.

[30] Statistisk Sentralbyrå : Folke- og boligtelling 1970. Hefte VI. Kontrollundersøkelse. Norges offisielle statistikk A 823. (Oslo: Statistisk Sentralbyrå, 1976).

[31] Sudman, S. : Reducing the cost of surveys. National Opinion Research Center. Monographs in Social Research, No. 10. (1967) Aldine Chicago.

[32] Sverdrup, E. : Lov og tilfeldighet. Bind I Universitetsforlaget, Oslo etc.

[33] Sæbø, H.V. : Varianser og designeffekter for sysselsettingstall estimert ved bruk av Byråets nye utvalgsplan. Arbeidsnotater fra Statistisk Sentralbyrå, IO 76/1. 1976.

[34] Sæbø, H.V. : Utvalgsregistre og rutiner for trekking og klargjøring av utvalg ved Underavdelingen for intervjuundersøkelser. Arbeidsnotater fra Statistisk Sentralbyrå. IO 76/22. 1976.

[35] Tamsfoss, S. : Om bruk av stikkprøver ved kontoret for intervjuundersøkelser. Artikler fra Statistisk Sentralbyrå nr. 37. (Oslo Statistisk Sentralbyrå, 1970).

[36] Thomsen, I. : On the effect of non-response in the Norwegian election survey 1969. Statistisk Tidsskrift 1971: 3. (Statistiska Centralbyrån, Stockholm. 1971).

[37] Thomsen, I. : A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. Statistisk Tidsskrift. 1973: 4. (Statistiska Centralbyrån, Stockholm, 1973).

[38] Thomsen, I. : A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation, Metrika 23, 1976, pp 15-25.

[39] Thomsen, I. : Bruk av superpopulasjonsmodeller ved innsamling og analyse av data fra utvalgsundersøkelser. Arbeidsnotat IO 76/26. Statistisk Sentralbyrå, 1976.

[40] Thomsen, I. : On the effect of stratification when two stratifying variables are used. Journal of the American Statistical Society, 1977, Vol. 72, nr. 357, pp 149-153.

[41] Thomsen, I. : Et forsøk på en enkel, teoretisk vurdering av de estimeringsmetoder som brukes i forbindelse med de politiske meningsmålinger. Sosialøkonomen, september 1977.

[42] Wilkins, L.T. : Prediction of the demand for campaign stars and medals. (Government social survey, No. 109) Central Office of Information. London (1949).

ENGLISH SUMMARY

In this publication the new design of the basic sampling plan used by the Central Bureau of Statistics of Norway is presented. In addition to giving a presentation of the design, a number of concepts and basic methods are given in Chapter 2 as an introduction to the description of the sample design given in Chapter 3.

The basic sample design is a two-stage sample with stratification in the first stage and one primary sampling unit (psu) selected from each stratum.

The construction of psu is based on the municipalities of Norway. Those with less than 3 000 inhabitants according to the Census in 1970 are collapsed such that every psu has at least 3 000 inhabitants. Before selection the primary sampling units are stratified according to size, geographical region, and type of economic activity. The largest cities form their own strata and are so-called selfrepresenting. In each of the other strata one psu is selected with a probability proportional to the size according to the Population Census 1970. The selected primary sampling units become the primary areas and are used repeatedly year after year, preserving the continuity and investment of a core of trained interviewers and of established frames.

In the second stage a sample of persons or households is selected from the Cental Register of Persons.

In many aspects the sample is designed in the same way as similar surveys are designed in other statistical agencies. Only the selection in the second stage differs from what is done elsewhere, because of the existence of registers of persons in Norway. Problems connected with selection of persons and households from a register are mainly caused by the following two deficiencies:

1) A register is never completely up-to-date. In particular, instructions for handling people who has moved from one municipality to another are needed.

2) Units that can be identified in the register, namely persons and families, only partly coincide with the units needed in surveys.

When a sample of persons is selected, the aim is to give every person a selection probability which is known in advance, and usually it

is constant for all persons. When the interviewer finds out that a selected person has moved to another dwelling, a new person is usually selected within the selected dwelling. If the number of persons moving into the dwelling differs from the number of persons that have moved out, this may change the selection probabilities. In many surveys this is neglected, while in other surveys one chooses to weight the observations. In some surveys it is chosen to follow up persons in their new dwellings. This solution is, of course, much more costly than the first two.

In household surveys the selection unit varies from one survey to another. In some surveys the family is selection unit, while in other surveys the selection unit may consist of more than one family. The unit that can be identified in the register is family, and if several families may consist of one selection unit, one could go through the whole register and define the selection units centrally. This process is costly and has never been used. Usually, one of the following methods is used:

First a sample of families is selected. If a selection unit consists of several families, the unit is selected if the selected family has the oldest person. If the selected family does not have the oldest person, the selection unit is considered not selected. In this way all selection units get a correct selection probability. Another method consists of disregarding that a selection unit may consist of several families, and allow the selection units to be selected with different probabilities. In these cases, data are weighted during the tabulations.

In Chapter 4 is given a discription of the most commonly used estimators, which are the usual mean, the ratio estimator, post-stratification, and synthetic estimators. A short discussion of the robustness of estimators using supplementary information is given. Special attention is given to the estimation of sampling errors. The meaning of design-effect is defined, and the table of calculated design-effects is given. These are calculated on the basis of data from a census and therefore they are exact and not estimates. Some of the experiences concerning estimation of sampling errors are given. The fact that only one psu is selected from each stratum has given some difficulties concerning estimation of variances. Several estimation methods have been tried, and the results compared with the exact numbers mentioned above. None of the methods tried out so far have been very useful because they seem to be seriously over- or underestimating the true variance.

In Chapter 5 are given the methods actually used to estimate the sample size necessary to give results with specified accuracy. Rough estimates as well as more formal methods are given.

It is a well known fact that estimates based on a survey are disturbed by other errors than sampling errors. In Chapter 6 is given a description of methods used at the Central Bureau of Statistics to identify and control these errors. The chapter is divided into four sections. After an introduction follows a section about selection biases, imperfect frames, and methods to avoid them. In the third section is given a definition of response variance, and examples of such errors in actual surveys are presented. Methods to identify and estimate response errors are given with examples. Finally, is given a description of what is done to control these errors with special emphasis on recruitment and training of interviewers. In the fourth and last section of this chapter non-response is treated. The response rates have shown a declining trend in Norway as in many other countries. In surveys where a satisfactory completion rate is achieved this has been done at a much higher cost. At the Central Bureau of Statistics the rate of completion has declined from around 90 to 95 per cent in 1969 to around 80 to 85 percent to-day. Little is known about the reasons for this decline, but two reasons seem important. Firstly, people seem to be less willing to respond to surveys now than previously, and secondly it seems more difficult to reach people at home to-day than previously. It seems as if the non-response divides about equally between "not-at-homes" and refusals, independent of the completion rate. To reduce the effects of non-response after the collection of data is completed, weighting of the observations has been used in connection with some surveys. The efficiency of such weighting seems in many cases to be very small.

Chapter 7 presents different ways in which the results from a sample survey are publised from the Central Bureau of Statistics. In particular, rules are given for how different errors should be published to avoid misuse of the results.

In all publications the actual population should be defined in all details, and an outline of the sample design must be given, eventually with reference to a more technical description if necessary. Also the frame should be described together with the main problems involved in selection of the sample.

We have not yet developed methods to present the overall quality of the results, but a careful description of non-response is given. Conserning

publication of sampling errors, two presentation methods are used:

1. In many cases it is sufficient to give the reader a rough estimate of the sampling errors. In such cases the sample variance is calculated as if the sample was a simple random sample, and the result multiplied with 1.5. In most surveys and for a majority of variables this estimate will slightly overestimate the true sampling error. In surveys where the majority of variables are quantitative, a table is given of the values of $1.5 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$, which leads to a very rough but simple presentation of sampling errors.

2. In cases where the methods described above are not sufficient, one can choose to estimate sampling errors by means of a general computer programme available. Usually, this is only done for a limited number of main results.

In addition to calculation of the sampling errors, the author can choose to mark numbers with large sampling errors in the tables, to make the reader aware of the low quality of some of the numbers. Relative frequences and means based on less than 25 observations are not published.

In Chapter 8 some special problems in connection with repeated surveys are discussed. Two surveys are of particular interest in Norway, namely the Labour Force Surveys and the Family Expenditure Surveys. In both surveys a rotation design is applied to improve the accuracy of estimates of changes over time. Even with such tailor-made designs it has proven very difficult to measure changes over time because the relative variances are very large. Some methods used to increase the accuracy is discussed, namely the use of composite estimates, and the use of time series models. The latter is attempted used in Norway, but the results are not yet published.

The main part of this publication is meant to be readable to all persons who get in contact with the results from surveys done by the Central Bureau of Statistics. Therefore, the description is given without use of mathematical symbols. In appendix 1 basic concepts and methods from the theory of sampling are given. This appendix is meant to serve as a link between the verbal description given in this publication, and the large number of text-books treating the theory of sampling.

V e d l e g g 1

ELEMENTER AV TEORIEN FOR SANNSYNLIGHETSUTVALG

1. INNLEDNING

I kapitlene II, III og IV foran er det gitt en verbal oversikt over de mest brukte metoder for trekking av utvalg og estimering av tall for hele populasjonen på grunnlag av resultatene i utvalget. I dette vedlegget skal det gis en mer nøyaktig beskrivelse ved hjelp av metoder og begreper fra den matematiske statistikken. Dette vedlegg inneholder ingen nye resultater, og er tatt med for fullstendighetens skyld. For den som er interessert i en videre studie av teorien for utvalgsundersøkelser kan dette vedlegget tjene som et mellomledd mellom den rene verbale beskrivelse gitt ovenfor, og de lærebøker som gir en omfattende behandling av teorien for statistiske utvalg. Eksempler på slike bøker er Cochran (1963), Kish (1965), Des Raj (1968). Framstillingen nedenfor er basert på Sverdrup (1964, kapittel XI), Hoem (1973) og Laake (1974).

2. ENKELT TILFELDIG UTVALG

En antar at det foreligger et register hvor det totale antall enheter er N. Til hver enhet i populasjonen er knyttet et kjennemerke a. Verdiene for de N enhetene betegnes med

$$a_1, a_2, \ldots, a_N.$$

For hver enhet kan verdien til kjennemerket bestemmes uten nevneverdig måle-

feil, og oppgaven er å bestemme $a = \sum_{j=1}^{N} a_j$ på grunnlag av resultatene i et utvalg.

Definisjon

Et utvalg av n enheter er enkel tilfeldig trukket fra populasjonen bestående av N enheter dersom alle $\binom{N}{n}$ mulige utvalg har samme sannsynlighet for å bli valgt ut.

Nå innføres N binomisk fordelte variable $\delta_1, \ldots, \delta_N$, definert som

$$\delta_i = \begin{cases} 1 & \text{dersom enhet nr. i er med i utvalget.} \\ 0 & \text{ellers.} \end{cases}$$

Setning 2.1

Når utvalget er trukket som et enkelt tilfeldig utvalg, er

$$E(\delta_i) = \frac{n}{N}, \quad \text{var}(\delta_i) = \frac{n}{N}\left(1-\frac{n}{N}\right) \text{ og } \text{cov}(\delta_i\delta_j) = -\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right).$$

Bevis:

Sannsynligheten for å utvelge et vilkårlig sett av enheter er ifølge definisjonen på et enkelt tilfeldig utvalg $1/\binom{N}{n}$. Antall mulige utvalg som inneholder element i er $\binom{N-1}{n-1}$. En har altså at

$$E(\delta_i) = P(\delta_i=1) = \binom{N-1}{n-1}/\binom{N}{n} = \frac{n}{N}$$

$$\text{var}(\delta_i) = E(\delta_i^2) - (E(\delta_i))^2$$

$$= \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1-\frac{n}{N}\right).$$

$$\text{cov}(\delta_i\delta_j) = E(\delta_i\delta_j) - E(\delta_i)\,E(\delta_j)$$

$$= P((\delta_i = 1) \cap (\delta_j = 1)) - \left(\frac{n}{N}\right)^2.$$

Da antall utvalg som inneholder både element i og j er $\binom{N-2}{n-2}$, har en at

$$P((\delta_i = 1) \cap (\delta_j = 1)) = \binom{N-2}{n-2}/\binom{N}{n} = \frac{n(n-1)}{N(N-1)},$$

hvorav følger at

$$\text{cov}(\delta_i\delta_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) \square$$

Det er to grunner for å innføre δ_i som hjelpevariabel. Den ene er at den forenkler bevisene for den følgende setning. Den andre er at den gjør det klart hva som er stokastisk, og hva som ikke er stokastisk. Slik problemet er formulert her er det stokastiske knyttet til δ_i, altså til om en enhet er med i utvalget eller ikke, mens det ikke er knyttet usikkerhet til kjennemerket a. Ved å ha kontroll med fordelingen til δ_i har en full kontroll over modellen, noe som ikke er tilfelle når modellen er knyttet til tilstanden i "naturen". Slik δ_i er definert kan utvalgsgjennomsnittet $\bar{X}$ nå skrives som

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} \delta_i a_i = \frac{1}{n}\sum_{i\in s} a_i,$$

hvor $\sum_{i\in s}$ står for summen over de enheter som er med i utvalget s.

Setning 2.2

Ved et enkelt tilfeldig utvalg er

$$E(N\bar{X}) = a, \quad \text{var}(N\bar{X}) = N(N-n)\frac{S^2}{n},$$

hvor $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a_i-\bar{a})^2$

Bevis:

$$E(N\bar{X}) = N\frac{1}{n}\sum_{i=1}^{N} a_i E(\delta_i) = \sum_{i=1}^{N} a_i .$$

$$\text{var}(N\bar{X}) = N^2\frac{1}{n^2}\text{var}\left(\sum_{i=1}^{N} a_i \delta_i\right)$$

$$= N^2\frac{1}{n^2}\left\{\sum_{i=1}^{N} a_i^2 \text{var}(\delta_i) + \sum_{i\neq j} a_i a_j \text{cov}(\delta_i\delta_j)\right\}$$

$$= \frac{N^2}{n^2}\left\{\sum_{i=1}^{N} a_i^2 \frac{n}{N}(1-\frac{n}{N}) - \sum_{i\neq j} a_i a_j \frac{n}{N(N-1)}(1-\frac{n}{N})\right\}$$

$$= N(N-n)\frac{S^2}{n}. \quad \square$$

Dersom utvalget trekkes med tilbakelegging fås et litt annet resultat. En definerer nå som ovenfor en hjelpevariabel

t_i = antall ganger enhet nr. i blir trukket.

Da sannsynligheten for at i-te element trekkes i hver trekking er $\frac{1}{N}$, følger at

$$E(t_i) = \frac{n}{N}, \quad \text{var}(t_i) = n(\tfrac{1}{N})(1-\tfrac{1}{N}),$$

og $\text{cov}(t_i\, t_j) = -\frac{n}{N^2}$

Ved samme framgangsmåte som i setning 2.2 kan det nå vises at

$$E(N\bar{X}) = a \text{ og } \text{var}(N\bar{X}) = N(N-1)\frac{S^2}{n}.$$

Variansen når en trekker med tilbakelegging er altså litt større en når en trekker et enkelt tilfeldig utvalg. Forskjellen er imidlertid forsvinnende når N er stor i forhold til n. Noe som alltid er tilfelle ved landsomfattende undersøkelser.

Anta nå at en er interessert i å bestemme et konfidensintervall for a, på grunnlag av observasjonene i utvalget. For å finne dette trenger en sannsynlighetsfordelingen til $N\bar{X}$ det vanlige er å anta at $\bar{X}$ er normalfordelt. Generelle regler for når denne tilnærmelsen er god er det vanskelig å sette opp. Vanligvis er tilnærmelsen brukbar hvis en på forhånd vet at a-verdiene i populasjonen antar verdier som alle opptrer relativt hyppig. Hvis en antar at $\bar{X}$ er normalfordelt, vil

$$\frac{N\bar{X} - a}{\sqrt{\text{var } (N\bar{X})}}$$

være tilnærmet normal med forventning 0, og varians 1. Hvis c er $(1-\frac{\varepsilon}{2})$ - fraktilen for den normerte normalfordelingen, finner en følgende konfidensintervall for a:

$$N\bar{X} - c\sqrt{\frac{N(N-n)}{n}}\ S < a < N\bar{X} + c\sqrt{\frac{N(N-n)}{n}}\ S.$$

Som regel vil S^2 være ukjent. Nedenfor er gitt en forventningsrett estimator for S^2, som kan settes inn i utrykket for konfidensintervallet.

Setning 2.3

Ved enkle tilfeldige utvalg gjelder at

$$E\ (\frac{1}{n-1} \sum_{ies} (a_i - \bar{X})^2) = S^2.$$

Bevis:

$$E\ (\frac{1}{n-1} \sum_{ies} (a_i - \bar{X})^2) = E\ (\frac{1}{n-1} \sum_{i=1}^{N} \delta_i (a_i - \bar{X})^2)$$

$$= E\ (\frac{1}{n-1} \sum_{i=1}^{N} \delta_i\ (a_i - \frac{1}{n} \sum_{j=1}^{N} \delta_j\ a_j)^2)$$

$$= E\left\{\frac{1}{n-1} \left(\sum_{i=1}^{N} \delta_i\ a_i^2 + \frac{1}{n} \left(\sum_{j=1}^{N} \delta_j\ a_j \right)^2 - 2\left(\sum_{i=1}^{N} \delta_i\ a_i \right) \left(\frac{1}{n} \sum_{j=1}^{N} \delta_j\ a_j \right) \right) \right\}$$

$$= E\left\{\frac{1}{n-1} \left\{ \sum_{i=1}^{N} \delta_i\ a_i^2 - n\ (\frac{1}{n} \sum_{j=1}^{N} \delta_j\ a_j)^2 \right\}\right.$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{N} \frac{n}{N}\ a_i^2 - n\ (\frac{1}{N} \sum_{i=1}^{N} a_i)^2 - n\ \text{var } (\bar{X}) \right\}$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{N} \frac{n}{N}\ a_i^2 - n\ (\frac{1}{N} \sum_{i=1}^{N} a_i)^2 - \frac{n(N-n)}{N}\ \frac{S^2}{n} \right\}$$

$$= \frac{1}{n-1} \left\{ \frac{n(N-1)}{N}\ S^2 - \frac{N-n}{N}\ S^2 \right\} = S^2. \quad \square$$

Anta spesielt at hver enhet er karakterisert ved at et kjennemerke A enten forekommer eller ikke forekommer. En kan da sette $a_i = 1$ hvis enhet nr. i har kjennemerket A, og $a_i = 0$ ellers. I slike tilfeller kan formlene for forventning og varians skrives som

$$E(\bar{X}) = P \text{ og var } (\bar{X}) = \frac{N-n}{N-1} \frac{P(1-P)}{n} .$$

I forbindelse med landsomfattende undersøkelser brukes ofte følgende tilnærmelse for variansen:

$$\text{var } (\bar{X}) = \frac{P(1-P)}{n} ,$$

hvor P er andelen av elementene i populasjonen med kjennemerket A. Estimatoren til variansen fås ved å innsette $\bar{X}$ som estimator for P i formlene ovenfor. I forbindelse med enkle tilfeldige utvalg er det altså enkelt å estimere variansen til en estimator i slike tilfeller.

2.1 Estimatorer for gjennomsnitt i delpopulasjoner

I praksis er en ofte interessert i å estimere totalen og gjennomsnitt for deler av populasjonen i tillegg til tall for hele populasjonen. I forbindelse med de fleste landsomfattende utvalg publiseres tall for deler av landet, for bestemte grupper av personer, osv. Spørsmålet er nå om en trenger en helt ny teori for å kunne takle dette problemet. Svaret er at stort sett er det slik at de resultater som gjelder for hele populasjonen også gjelder for deler av populasjonen. Dette kan vises på flere måter, her skal gis en framstilling som i det vesentlige faller sammen med avsnittene 6.1 og 6.2 i Hoem (1973).

La populasjonen bestå av N trekkeenheter, som det trekkes et enkelt tilfeldig utvalg på n enheter fra. La antall trekkeenheter som tilhører den aktuelle delpopulasjon være M, og la X være antall trekkeenheter fra delpopulasjonen i utvalget. Av de M enheter har M_i verdien i på et gitt, annet kjennemerke, og X_i av disse er med i utvalget. En har altså at $M = \Sigma M_i$ og $X = \Sigma X_i$. La

$$P = M/N \quad \text{og} \quad q_i = M_i/M,$$

og en ønsker å estimere $q_1, q_2, \ldots$.

Det er nå klart at

$$P\left\{ X = x \right\} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

og

$$P\left\{ X_i = x_i \mid X = x \right\} = \frac{\binom{M_i}{x_i}\binom{M - M_i}{x - x_i}}{\binom{M}{X}} .$$

I begge tilfeller antas det at de hypergeometriske uttrykkene ovenfor kan erstattes med de tilsvarende binomiske, slik at en med god tilnærmelse kan regne som om

$$P\left\{ X = x \right\} = \binom{n}{x} p^x (1-p)^{n-x} \quad x=0, 1, \ldots, n,$$

$$P\left\{ X_i = x_i \mid X = x \right\} = \binom{x}{x_i} q_i^{x_i} (1-q_i)^{x-x_i}; \; x_i = 0, 1, \ldots, x.$$

Den naturlige estimator for q_i er

$$q_i^* = X_i/X,$$

med mindre X er "for liten". For små X vil en antakelig avstå fra å estimere q_i. Når X er stor er den numeriske forskjellen mellom q_i^* og

$$\hat{q}_i = X_i/(X + 1)$$

neglisjerbar. I det følgende vil alle resonnementer bli basert på $\hat{q}_i$. Det ses lett at

$$E(\hat{q}_i \mid X = x) = \frac{x}{x + 1} q_i,$$

så når $X = x$ og $x/(x+1) \approx 1$, er $\hat{q}_i$ tilnærmet forventningsrett. Dessuten er

$$\text{var}(\hat{q}_i \mid X = x) = \frac{x}{(x + 1)^2} q_i (1-q_i),$$

som er tilnærmet lik $q_i(1-q_i)/(x + 1)$.

Hvis n og p er så pass stor at

$$P\left\{ \frac{X}{X + 1} \approx 1 \right\} \approx 1,$$

blir

$$E\,\hat{q}_i \approx q_i \,,$$

og

$$\operatorname{var}\,\hat{q}_i \approx q_i\,(1-q_i)\,E\left\{\frac{1}{X+1}\right\} \,.$$

Nå er

$$E\left\{\frac{1}{X+1}\right\} = \sum_{x=0}^{n} \frac{1}{x+1}\binom{n}{x} p^x\,(1-p)^{n-x}$$

$$= \frac{1}{n+1}\sum_{y=1}^{n+1}\binom{n+1}{y} p^y\,(1-p)^{n+1-y}$$

$$= \frac{1}{n+1}\left\{\sum_{y=0}^{n+1}\binom{n+1}{y} p^y\,(1-p)^{n+1-y} - (1-p)^{n+1}\right\}$$

$$= \frac{1-(1-p)^{n+1}}{p(n+1)} \approx \frac{1}{p(n+1)} \approx \frac{1}{np} \,.$$

Det følger da at

$$\operatorname{var}\,(\hat{q}_i) \approx \frac{q_i(1-q_i)}{np} \,.$$

En opplagt estimator for var $(\hat{q}_i)$ er

$$q_i^{*}\;(1-q^{*})/X.$$

Dette er den samme estimatoren som den en ville bruke dersom X var ikke-stokastisk.

Det er dermed vist at resultatene som gjelder når en estimerer i hele populasjonen, stort sett gjelder for delpopulasjoner. En må forutsette at delpopulasjonene har en viss størrelse for at tilnærmelsene skal være gode.

3. SYSTEMATISK UTVALG

En enkel måte å trekke utvalg på som er mye brukt i Byrået er systematisk utvelging. En forutsetter her at enhetene er nummerert fra 1 til N. La dessuten N = nk, hvor n er størrelsen på det utvalg en vil trekke.

(For enkelhets skyld forutsettes det at k er et helt tall.) Et tall velges nå tilfeldig blant tallene 1 til k. La det valgte tall være i, da er følgende enheter utvalgt: i, i+k, i+2k, ... i+(n-1)k. Utvalget er altså trukket ved å velge den første enhet tilfeldig og deretter hver k^{te} enhet.

Formelt kan en si at denne trekkemetoden svarer til å inndele populasjonen i k klynger, hver med n enheter, og deretter trekke en av disse klynger tilfeldig.

Sannsynligheten for at et element er med i utvalget er 1/k, hvorav følger at utvalget er et sannsynlighetsutvalg. Derimot er utvalget ikke et rent tilfeldig utvalg.

Setning 3.1

La N, n og k være som definert ovenfor. En forventningsrett estimator for $a = \sum_{i=1}^{N} a_i$ er da gitt ved

$$\bar{X} = k\, T_i ,$$

hvor T_i er summen av a-verdiene i den uttrukne klynge. Variansen til $\bar{X}$ er gitt ved

$$\operatorname{var}(\bar{X}) = k \sum_{i=1}^{k} \left(T_i - \frac{\sum_{i=1}^{N} a_i}{k}\right)^2$$

Bevis:

Igjen innføres en indikatorvariabel

$$\delta_i = \begin{cases} 1 & \text{hvis i-te klynge velges ut.} \\ 0 & \text{ellers.} \end{cases}$$

Ettersom bare en klynge velges ut, er

$$E(\delta_i) = \frac{1}{k} \qquad i = 1, 2, \ldots, k,$$

$$\operatorname{var}(\delta_i) = \frac{1}{k}\left(1 - \frac{1}{k}\right),$$

og $$\operatorname{cov}(\delta_i \delta_g) = -\frac{1}{k^2} .$$

Estimatoren $\bar{X}$ kan skrives som

$$\bar{X} = k \sum_{i=1}^{k} \delta_i T_i ,$$

hvorav følger at

$$E(\bar{X}) = \sum_{i=1}^{k} k \frac{1}{k} T_i = \sum_{i=1}^{k} T_i = \sum_{i=1}^{N} a_i .$$

Dessuten er

$$\text{var}(\bar{X}) = E(\bar{X} - \sum_{i=1}^{N} a_i)^2 = \sum_{i=1}^{k} \frac{1}{k} (k T_i - \sum_{i=1}^{N} a_i)^2$$

$$= k \sum_{i=1}^{k} (T_i - \frac{\sum_{i=1}^{N} a_i}{k})^2 \square .$$

Det er verdt å merke seg at variansen avhenger av hvordan populasjonen av enheter er ordnet før trekking. Dersom en ordner populasjonen på en annen måte, vil en forandre alle T_i, som igjen vil føre til en annen varians.

Dessuten bør en merke seg at variansen ikke nødvendigvis vil øke når en øker utvalgets størrelse.

Det største problemet knyttet til bruken av systematiske utvalg skyldes likevel at en ikke uten videre er i stand til å estimere variansen til estimatoren for $\sum_{i=1}^{N} a_i$. Det skyldes at en bare har en observasjon av T_i. Den letteste måten å løse dette problemet på, består av å ta flere uavhengige systematiske utvalg.

4. TREKKING MED FORSKJELLIGE SANNSYNLIGHETER

I praksis kan en ofte med fordel bruke utvalgsmetoder, som ikke nødvendigvis gir alle elementer i populasjonen samme sannsynlighet for å bli trukket ut. Ofte ønsker en f. eks. å gi store bedrifter en større sannsynlighet for å bli trukket ut enn en mindre bedrift. Det kan vises at dette ofte fører til mindre varians enn et enkelt tilfeldig utvalg. Dette kan gjøres ved å stratifisere, men også andre metoder kan brukes. Her skal behandles en enkelt av disse, ofte kalt utvelging med sannsynlighet proporsjonal med størrelsen.

En tenker seg her at det til hver enhet i populasjonen er knyttet et kjent positivt tall i tillegg til den ukjente a-verdien. Disse tall er normerte slik at summen av dem er 1. Tallet knyttet til i-te enhet er p_i. Det trekkes nå et utvalg på n enheter ved i hver trekking å gi enhet i sannsynligheten p_i for å bli trukket. Utvalget trekkes med tilbakelegging.

Setning 4.1

La utvalget trekkes på ovennevnte måte. Da er

$$X = \frac{1}{n} \sum_{i \in s} a_i/p_i$$

en forventningsrett estimator for $\sum_{i=1}^{N} a_i$, med varians

$$\text{var } (X) = \frac{1}{n} \left(\sum_{i \neq j} p_i p_j \left(\frac{a_i}{p_i} - \frac{a_j}{p_j} \right)^2 \right)$$

Bevis:

I stedet for som tidligere å innføre en indikator variabel, innføres nå en variabel t_i definert som antall ganger enhet nr. i er med i utvalget. Dette er nødvendig fordi trekkingen foretas med tilbakelegging, og hver enhet kan da trekkes flere ganger. Da hver trekking kan betraktes som et binomisk forsøk der i-te enhet blir trukket eller ikke, er

$$E(t_i) = np_i$$

$$\text{var } (t_i) = np_i (1-p_i).$$

Dessuten er

$$\text{cov } (t_i t_j) = - np_i p_j.$$

Estimatoren kan nå skrives som

$$X = \frac{1}{n} \sum_{i=1}^{N} t_i a_i/p_i.$$

Innsettes uttrykket for $E(t_i)$, fås

$$E(X) = \frac{1}{n} \sum_{i=1}^{N} E(t_i)\, a_i/p_i = \sum_{i=1}^{N} a_i .$$

Tilsvarende fås for variansen

$$\text{var } (X) = \frac{1}{n^2} \text{ var } \left(\sum_{i=1}^{N} t_i a_i/p_i \right)$$

$$= \frac{1}{n^2} \left(\sum_{i=1}^{N} \frac{a_i^2}{p_i^2} \text{ var } (t_i) + \sum_{i \neq j} \frac{a_i a_j}{p_i p_j} \text{ cov } (t_i t_j) \right),$$

$$= \frac{1}{n^2} \left(\sum_{i=1}^{N} \frac{a_i^2}{p_i^2} np_i (1-p_i) - \sum_{i \neq j} \frac{a_i a_j}{p_i p_j} np_i p_j \right)$$

$$= \frac{1}{n} \sum_{i \neq j}^{N} p_i p_j \left(\frac{a_i}{p_i} - \frac{a_j}{p_j} \right)^2 \square$$

En bør merke seg at var (X) = 0 dersom a_i/p_i er konstant for samtlige enheter i populasjonen. En bør altså velge p_i på en måte som gjør at forholdet a_i/p_i er tilnermet konstant. Sammenlikningen mellom variansene ved enkelt tilfeldig utvalg og utvalg proporsjonal med størrelsen er vanskelig. Seinere skal det vises at denne trekkemetoden bør kombineres med bruken av en annen type estimator enn den som er foreslått her, nemlig rateestimatoren, som skal behandles i neste avsnitt.

Det er et stort antall andre måter å trekke utvalg på som gir enhetene forskjellige sannsynligheter for å komme med. Disse skal ikke tas opp her. Eventuelt interesserte henvises til Des Raj (1968).

5. RATEESTIMATOREN

Forutsetningene er stort sett de samme som i avsnittet foran. Til hvert element i populasjonen er det i tillegg til kjennemerket a_i, knyttet et kjent tall b_i. Dette kan f.eks. være en tidligere gjort observasjon av kjennemerket a i en totaltelling. På samme måten som ovenfor ønsker en å bruke denne tilleggsinformasjonen for å lage et anslag for $\sum_{i=1}^{N} a_i$.

Det trekkes et enkelt tilfeldig utvalg. For de uttrukne enhetene observeres kjennemerkeverdiene til de to kjennemerker a og b. Rateestimatoren til $\sum_{i=1}^{N} a_i$ er da definert som

$$\hat{a} = \frac{\bar{X}}{\bar{Y}} b,$$

hvor $\bar{X} = \frac{1}{n} \sum_{ies} a_i$,

$$\bar{Y} = \frac{1}{n} \sum_{ies} b_i$$

og

$$b = \sum_{i=1}^{N} b_i.$$

Denne estimatoren kan skrives som $\hat{a} = \hat{R}\, b$,

hvor $\hat{R} = \bar{X}/\bar{Y}$.

Setning 5.1

Dersom utvalget trekkes som et enkelt tilfeldig utvalg, gjelder at

$$E(\hat{R}) \approx a/b = R ,$$

og

$$\text{var}(\hat{R}) \approx \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{(S_a^2 + R^2 S_b^2 - 2R\rho S_a S_b)}{\bar{b}^2} ,$$

hvor $$\rho = \text{cov}(\bar{X}\,\bar{Y}) = \frac{N-n}{n(N-1)N} \sum_{i=1}^{N} (a_i - \bar{a})(b_i - \bar{b}).$$

Bevis:

For store n kan en sette

$$\hat{R} = \frac{\bar{X}}{\bar{Y}} = \frac{\bar{a}}{\bar{b}} \frac{1 + \frac{\bar{X} - \bar{a}}{\bar{a}}}{1 + \frac{\bar{Y} - \bar{b}}{\bar{b}}} \approx \frac{\bar{a}}{\bar{b}} \left(1 + \frac{\bar{X} - \bar{a}}{\bar{a}} - \frac{\bar{Y} - \bar{b}}{\bar{b}}\right) .$$

Setningen bevises da enkelt ved å ta forventning og varians til tilnærmingen til R. □

Det følger nå at

$$\text{var}(\hat{a}) = b^2 \text{var}(R) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)(S_a^2 + R^2 S_b^2 - 2R\rho S_a S_b).$$

Da estimatoren $\hat{R}$ ikke er forventningsrett, kan det være en fordel å vite litt om skjevheten til $\hat{R}$.

Setning 5.2

Når utvalget er enkelt tilfeldig, gjelder at

$$E(\hat{R} - R) = - \left[E(\bar{X})\right]^{-1} \text{cov}(\hat{R}, \bar{X}).$$

Bevis:

Da

$$\text{cov}(\hat{R}, \bar{X}) = E(\bar{Y}) - E(\bar{Y}/\bar{X})\, E(\bar{X}),$$

fås

$$\bar{a}\, E\left(\frac{\bar{Y}}{\bar{X}}\right) = \bar{b} - \text{cov}(\hat{R}\,\bar{X}),$$

eller

$$E(\hat{R}) - R = - \frac{\text{cov}(\hat{R}_1 \bar{X})}{\bar{X}} \quad \square$$

Resultatet i setning 5.2 er av liten praktisk verdi da det ikke er mulig å estimere cov $(\hat{R}, \bar{X})$ ut fra resultatene fra et utvalg. I følgende setning skal derimot gis et resultat av større verdi i praksis.

Setning 5.3

Under samme forutsetninger som i setningen ovenfor gjelder at

$$E\ (\hat{R}-R) \leq \sqrt{\text{var}\ (\hat{R})}\ \sqrt{\frac{\text{var}\ (\bar{X})}{\bar{a}}}\ .$$

Bevis:

Av setning 5.2 følger at

$$E\ (\hat{R}-R) = -\ \sqrt{\text{var}\ (\hat{R})}\ \sqrt{\text{var}\ (\bar{X})}\ \ \frac{\rho\ (\hat{R}, \bar{X})}{\bar{X}}\ ,$$

hvor $\rho\ (\hat{R}, \bar{X})$ er korrelasjonskoeffisienten mellom $\hat{R}$ og $\bar{X}$. Videre følger at

$$\frac{E\ (\hat{R}-R)}{\sqrt{\text{var}\ (\hat{R})}} = -\ \rho\ (\hat{R}, \bar{X})\ \frac{\sqrt{\text{var}\ (\bar{X})}}{\bar{X}} \leq \frac{\sqrt{\text{var}\ (\bar{X})}}{\bar{X}}\ ,$$

hvorav setningen følger. □

Når en i praksis ønsker å sikre seg at skjevheten er liten, må en velge n så stor at $\sqrt{\text{var}\ (\bar{X})}\ /\ \bar{X}$ er liten.

Hvis en kan se bort fra skjevheten til rateestimatoren, får en følgende sammenlikning mellom det vanlige gjennomsnittet og rateestimatoren.

$$\text{var}\ (N\bar{X}) = \frac{N(N-n)}{n(N-1)}\ \sum_{i=1}^{N}\ (a_i - \bar{a})^2\ ,$$

og med litt omskriving fås at

$$\text{var}\ (\hat{a}) = \frac{N(N-n)}{n(N-1)}\ \sum_{i=1}^{N}\ (a_i - \frac{\bar{a}}{\bar{b}}\ b_i)^2\ .$$

Variansen til rateestimatoren er altså mindre enn variansen til gjennomsnittet, dersom

$$\sum_{i=1}^{N}\ (a_i - \frac{\bar{a}}{\bar{b}}\ b_i)^2 < \sum_{i=1}^{N}\ (a_i - \bar{a})^2\ .$$

Venstre siden måler hvor meget a-verdiene avviker fra å være proporsjonale med b-verdiene. Stort sett kan en altså påstå at rateestimatoren bør brukes

når de to kjennemerker er proporsjonale.

Det finnes flere måter en kan gjøre rateestimatoren forventningsrett. Det kan gjøres ved å trekke utvalget på en bestemt måte, og endelig finnes det en forventningsrett estimator for a/b ved enkel tilfeldig trekking. En skal ikke komme nærmere inn på dette her, men henviser til Des Raj (1968). Før en velger å bruke rateestimatoren bør en sikre seg at det ikke forekommer nevneverdig målefeil på b. Rateestimatoren kan nemlig i slike tilfeller være meget dårlig.

6. STRATIFISERING

Som nevnt i kapittel II kan en redusere variansen for gitt utvalgsstørrelse ved å inndele populasjonen i homogene delpopulasjoner, kalt strata, og deretter trekke uavhengige utvalg innen hvert stratum. Måten en trekker utvalget på kan variere fra stratum til stratum, men her skal en tenke seg at det trekkes et enkelt tilfeldig utvalg innen hvert stratum.

En tenker seg at populasjonen bestående av N enheter blir delt opp i L strata, med N_h enheter i h-te stratum, hvor summen av a-verdiene er a_h. Innen stratum h trekkes et enkelt tilfeldig utvalg på n_h enheter. Når en skal anslå a brukes følgende estimator:

$$\hat{a} = \sum_{h=1}^{L} N_h \bar{X}_h ,$$

hvor $\bar{X}_h$ er gjennomsnittet innen utvalget trukket fra stratum h.

Av setning 2.2 følger at

$$E(\hat{a}) = \sum_{h=1}^{L} E(N_h \bar{X}_h) = \sum_{h=1}^{L} a_h = a .$$

Dessuten er

$$\text{var}(\hat{a}) = \sum_{h=1}^{L} \text{var}(N_h \bar{X}_h) = \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} ,$$

hvor

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (a_{ih} - \bar{a}_h)^2 , \text{ og } a_{ih} \text{ er } i^{\text{te}}$$

element i h-te stratum.

En estimator for gjennomsnittet av a-verdiene er

$$\hat{\bar{a}} = \hat{a}/N = \sum_{h=1}^{L} (N_h/N) \bar{X}_h = \sum_{h=1}^{L} W_h \bar{X}_h .$$

Hvert stratum gjennomsnitt $\bar{X}_h$, får vekt etter hvor mange enheter det er i det tilhørende stratum og ikke etter hvor mange observasjoner det bygger på.

Hvis betingelsene er til stede for at alle $\bar{X}_s$ er tilnærmet normale, vil også $\hat{a}$ være tilnærmet normal og et konfidensintervall kan konstrueres som ovenfor.

6.1 Allokering av utvalget på strataene

Den reduksjon av varians en oppnår avhenger av hvordan en velger n_h. Vanligvis velger en enten antall observasjoner proporsjonal med antall enheter i populasjonen, eller en velger n_h med sikte på å gjøre variansen minst mulig.

En har at

$$\text{var}(\hat{a}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} - K,$$

hvor K er en størrelse uavhengig av n_h. En skal nå bestemme det sett, av $n_1, n_2, \ldots, n_L$, som minimerer

$$Q = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}$$

under bibetingelsen at $\sum_{h=1}^{L} n_h = n$. Innsettes $n_L = n - n_1 - n_2 - \ldots - n_{L-1}$ i uttrykket for Q, fås

$$Q = \frac{1}{N^2} \sum_{h=1}^{L-1} N_s^2 \frac{S_h^2}{n_s} + \frac{1}{N^2} \frac{N_L^2 S_L^2}{n - n_1 - n_2 - \ldots - n_{L-1}}.$$

En ønsker å finne de verdier av $n_1, n_2, \ldots, n_L$, som oppfyller likningene

$$\frac{\partial Q}{\partial n_h} = 0; \quad h = 1, 2, \ldots, L-1.$$

En får da

$$-\frac{N_h^2}{n_h^2} S_h^2 + \frac{N_L^2 S_L^2}{(n - n_1 - n_2 - \ldots - n_{L-1})^2} = 0.$$

Dette kan skrives som

$$n_h = \frac{N_h S_h}{N_L S_L} n_L, \quad h = 1, 2, \ldots, r.$$

Summeres alle likninger fås

$$n = \frac{\Sigma\, N_h\, S_h}{N_L\, S_L}\, n_L ,$$

eller

$$n_L = \frac{N_L\, S_L}{\Sigma\, N_h\, S_h}\, n .$$

Innsettes dette i uttrykket for n_h ovenfor, fås

$$n_h = \frac{N_h\, S_h}{\Sigma\, N_j\, S_j}\, n .$$

I praksis er S_h ukjent. Likevel kan ovennevnte regel for valg av n_h hjelpe til ved valg av n_h når en vet noe om de relative variasjonene av a-verdiene i de forskjellige strata. Hvis en f.eks. vet at mindre bedrifter har mindre variasjon i produksjonen enn større bedrifter, kan en velge å overrepresentere de store og underrepresentere de små bedrifter tilsvarende.

I noen tilfeller vet en lite om de relative størrelser av S_h. En velger da ofte følgende allokering:

$$n_h = \frac{N_h}{N}\, n ,$$

som er optimal dersom S_h er identisk for alle strata. Denne allokering kalles proporsjonal allokering. Det kan vises at variansen ved proporsjonal allokering aldri er større enn ved enkel tilfeldig utvelging.

6.2 Etter- stratifisering

Ofte har en ikke mulighet for å stratifisere hele populasjonen, men en har tabeller over størrelsene, N_h, på samtlige strata. I slike situasjoner kan en velge å sortere enheter i utvalget etter hvilket stratum de tilhører, og deretter i utvalget beregne gjennomsnittene innen hvert stratum, $\bar{X}_1, \ldots, \bar{X}_L$. En kan deretter bruke estimatoren

$$\hat{\bar{a}}^{*} = \sum_{h=1}^{L} \frac{N_h}{N}\, \bar{X}_h .$$

Dersom en har valgt strataene så store at sannsynligheten for å få observasjoner innen hvert stratum er stor, kan det vises at $\hat{\bar{a}}^{*}$ er en forventningsrett, estimator for $\bar{a}$, og at variansen til $\hat{\bar{a}}$ er tilnærmet den samme som den en ville ha fått ved vanlig stratifisering og proporsjonal allokering. En annen viktig egenskap ved $\hat{\bar{a}}^{*}$ er at den reduserer effekten av frafall dersom frafallsprosenten varierer mye fra stratum til stratum.

7. KLYNGEUTVALG

Til nå er det forutsatt at det foreligger et register over de enheter en ønsker å studere. Ofte er dette ikke tilfelle, og kostnadene ved å opprette et register kan ofte være så store at det ikke lønner seg å trekke enhetene direkte. Hvis en f. eks. ønsker å trekke et utvalg av personer som er ansatt i en bestemt type bedrift, har en kanskje et register for bedriftene, men ikke for de ansatte. En mulig måte å trekke et utvalg av personer på er da å trekke et utvalg av bedrifter, og deretter undersøke alle ansatte i de uttrukne bedrifter. Et slikt utvalg kalles et klyngeutvalg. En kan naturligvis velge å trekke et utvalg av ansatte i de uttrukne bedrifter, et slikt utvalg kalles et to-trinnsutvalg. Slike utvalg vil bli behandlet i neste avsnitt.

En antar at det er M klynger. Den j-te klynge har N(j) enheter, og det k-te element har a-verdien a (j, k).

La $$a(j) = \sum_{k=1}^{N(j)} a(j, k) .$$

Setning 7.1

Hvis en trekker et enkelt tilfeldig utvalg på m klynger, gjelder at

$$\hat{a} = \frac{M}{m} \sum_{j \in s} a(j)$$

er en forventningsrett estimator for a.

Dessuten er

$$\text{var}(\hat{a}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{\sum_{j=1}^{M} (a(j) - \frac{a}{M})^2}{M - 1}$$

Bevis:

Setningen følger av setning 2.2, hvor a_i erstattes med a (j). Det ses at variansen til $\hat{a}$ avhenger av variasjonen i a (j) - verdiene. Når klyngene varierer mye i størrelse, kan a (j) variere mye, hvilket fører til at $\hat{a}$ får stor varians sammenliknet med det vanlige gjennomsnittet. Det er derfor av stor viktighet å trekke klyngene på en måte som reduserer variansen til a (j). To måter er vanligvis brukt, nemlig stratifisering og trekking av klynger proporsjonal med størrelsen. Se Des Raj (1968) side 112-113.

8. ESTIMERING AV TOTALER NÅR BYRÅETS GENERELLE UTVALGSPLAN BRUKES *

Som nevnt i kapittel III er Byråets utvalgsplan trukket i to trinn. Før trekking av de primære utvalgsområder, er disse stratifisert, og innen hvert stratum er det trukket et utvalgsområde med en sannsynlighet proporsjonal med innbyggerantallet i 1970. Trekkingen i første trinn er altså gjort ved å kombinere flere av de utvalgsmetoder som er tatt opp tidligere i dette vedlegget. I andre trinnet er det trukket et enkelt tilfeldig utvalg innen de uttrukne utvalgsområdene. (I de siste årene foregår trekkingen i andre trinnet ved systematisk utvelging, men her skal variansen utledes som om utvalget fortsatt er enkelt tilfeldig.)

Først skal en betrakte en utvalgsplan, hvor de primære utvalgsområder er stratifisert, og det trekkes minst to utvalgsområder fra hvert stratum. Grunnen til dette er at en bare i slike tilfeller kan finne forventningsrette estimatorer for variansen.

Det antas at det er M_i utvalgsområder i i-te stratum. Det j-te av disse har $N_i(j)$ trekkeenheter, og den k-te trekkeenheten har verdien $a_i(j, k)$ på det vi måler. La

$$N_i = \sum_j N_i(j),$$

$$N = \sum_i N_i,$$

$$a_i(j) = \sum_k a_i(j, k),$$

$$\bar{a}_i(j) = a_i(j)/N_i(j),$$

$$a_i = \sum_j a_i(j),$$

$$\bar{a}_i = a_i/M_i,$$

og

$$a = \sum_i a_i = \sum_i \sum_j \sum_k a_i(j, k).$$

En ønsker å estimere totalverdien a.

I stratum i trekkes ut m_i av de M_i utvalgsområdene. La $\pi_i(j)$ være sannsynligheten for at utvalgsområde j i stratum i blir trukket ut, og la

* Dette avsnittet er skrevet av konsulent Petter Laake.

$\pi_i(j, k)$ være sannsynligheten for at både utvalgsområdene j og k i stratum i skal bli trukket ut.

La $J_{i1}, J_{i2}, \ldots, J_{im_i}$ være numrene på de utvalgsområdene som blir trukket ut i stratum i og la $\underset{\sim}{J}=(J_{i1},..,J_{im_i}, J_{21},\ldots)$ være numrene på alle de uttrukne utvalgsområdene. Fra hvert uttrukne utvalgsområde trekkes et gitt antall trekkeenheter rent lotterisk. La $n(\underset{\sim}{J})$ og $n_{ij}(\underset{\sim}{J})$ være totalt antall trekkeenheter i henholdsvis hele utvalget og i utvalget i j-te utvalgsområde i stratum i.

Numrene på de enhetene som blir trukket ut fra utvalgsområde j i stratum i, betegnes med $K_{ij1}, K_{ij2}, \ldots$, og la

$$X_{ijs} = a_i(j, K_{ijs}),$$

og

$$\overline{X}_{ij} = \sum_s X_{ijs}/n_{ij}(\underset{\sim}{J}).$$

Definer

$$I_{ij} = \begin{cases} 1 \text{ dersom utvalgsområde j i stratum i er i utvalget,} \\ 0 \text{ ellers,} \end{cases}$$

og lar $\overline{X}_{ij}$ være gjennomsnittet for de $n_{ij}(\underset{\sim}{J})$ uttrukne trekkeenhetene i utvalgsområde j i stratum i. Dersom $I_{ij} = 0$, er $n_{ij}(\underset{\sim}{J}) = 0$, og $\overline{X}_{ij}$ defineres lik null.

En estimator for totalen er

$$\hat{a} = \sum_i \sum_j \{I_{ij} N_i(j) \overline{X}_{ij}/\pi_i(j)\}$$

Setning 8.1: $\hat{a}$ er forventningsrett for a.

Bevis: La

$$V_{ij} = N_i(j) \overline{X}_{ij}.$$

Siden $E\{V_{ij}|I_{ij} = 1\} = a_i(j)$,

og

$$E\{I_{ij} V_{ij}|I_{ij}\} = I_{ij} a_i(j),$$

er

$$E\{I_{ij}\, V_{ij}\} = \pi_i(j)\, a_i(j).$$

Herav følger at

$$E\{I_{ij}\, V_{ij}/\pi_i(j)\} = a_i(j),$$

og satsen er dermed bevist. □

Sett nå

$$\hat{a}_i = \sum_j \{I_{ij}\, V_{ij}/\pi_i(j)\}.$$

Da er

$$\hat{a} = \sum_i \hat{a}_i.$$

La videre

$$\sigma_i^2(j) = \frac{1}{N_i(j)-1} \sum_k \{a_i(j,k) - \bar{a}_i(j)\}^2,$$

og

$$\tau_{ij}^2(\underset{\sim}{J}) = \frac{\sigma_i^2(j)}{n_{ij}(\underset{\sim}{J})} \frac{N_i(j)-n_{ij}(\underset{\sim}{J})}{N_i(j)}.$$

Dersom $I_{ij} = 0$, er $n_{ij}(\underset{\sim}{J}) = 0$, og $\tau_{ij}^2(\underset{\sim}{J})$ er da udefinert.

Setning 8.2: La $\hat{a}$ være definert som ovenfor. Da er

$$\operatorname{var} \hat{a} = \sum_i \left[\sum_j \sum_k \frac{\pi_i(j,k)-\pi_i(j)\,\pi_i(k)}{\pi_i(j)\,\pi_i(k)}\, a_i(j)\, a_i(k) + \sum_j \{\eta_i(j)/\pi_i(j)\} \right],$$

hvor $\eta_i(j) = E\{N_i^2(j)\, \tau_{ij}^2(\underset{\sim}{J}) \mid I_{ij} = 1\}.$

Bevis: Av definisjonen av $\hat{a}_i$ følger umiddelbart at

$$\operatorname{var} \hat{a}_i = \sum_j \operatorname{var}\{I_{ij}\, V_{ij}\}/\pi_i^2(j) + \sum_{j \neq k} \frac{1}{\pi_i(j)} \frac{1}{\pi_i(k)} \operatorname{cov}\{I_{ij}\, V_{ij},\, I_{ik}\, V_{ik}\}.$$

Videre er

$$\text{var}\{I_{ij} V_{ij} \mid I_{ij} = 1, \underset{\sim}{J} = j\} = N_i^2(j)\, \tau_{ij}^2(\underset{\sim}{j}),$$

og

$$\text{var}\{I_{ij} V_{ij} \mid I_{ij} = 1\} = E\{N_i^2(j)\, \tau_{ij}^2(\underset{\sim}{J}) \mid I_{ij} = 1\}$$

$$= \eta_i(j).$$

Herav følger at

$$\text{var}\{I_{ij} V_{ij} \mid I_{ij}\} = I_{ij}\, \eta_i(j).$$

Da $E\{I_{ij} V_{ij} \mid I_{ij}\} = I_{ij}\, a_i(j)$, er

$$\text{var}\{I_{ij} V_{ij}\} = \pi_i(j)\, \eta_i(j) + a_i^2(j)\, \pi_i(j)\{1 - \pi_i(j)\}.$$

Det gjenstår å finne

$$\text{cov}\{I_{ij} V_{ij}, I_{ik} V_{ik}\} = E\{I_{ij} V_{ij} I_{ik} V_{ik}\} - E\{I_{ij} V_{ij}\} E\{I_{ik} V_{ik}\}.$$

En har umiddelbart at

$$E\{I_{ij} V_{ij} I_{ik} V_{ik} \mid I_{ij}, I_{ik}\} = I_{ij} I_{ik}\, a_i(j)\, a_i(k),$$

slik at

$$E\{I_{ij} V_{ij} I_{ik} V_{ik}\} = \pi_i(j,k)\, a_i(j)\, a_i(k).$$

Videre er

$$E\{I_{ij} V_{ij}\} E\{I_{ik} V_{ik}\} = \pi_i(j)\, a_i(j)\, \pi_i(k)\, a_i(k),$$

som gir

$$\text{cov}\{I_{ij} V_{ij}, I_{ik} V_{ik}\} = a_i(j)\, a_i(k)\{\pi_i(j,k) - \pi_i(j)\, \pi_i(k)\}.$$

Ved et resonnement tilsvarende det i Hoem (1973, side 12) finner vi at

$$\text{cov}(\hat{a}_i, \hat{a}_j) = 0 \quad \text{for } i \neq j.$$

Ved innsettelse blir da

$$\text{var } \hat{a} = \sum_i \{\sum_j [\pi_i(j)\{1-\pi_i(j)\} a_i(j) + \pi_i(j)\,\eta_i(j)]/\pi_i^2(j)$$

$$+ \sum_{j\neq k} \frac{1}{\pi_i(j)} \frac{1}{\pi_i(k)} a_i(j)\, a_i(k) [\pi_i(j,k) - \pi_i(j)\,\pi_i(k)]\}$$

$$= \sum_i [\sum_j \sum_k \frac{\pi_i(j,k)-\pi_i(j)\pi_i(k)}{\pi_i(j)\,\pi_i(k)} a_i(j)\, a_i(k)$$

$$+ \sum_j \{\eta_i(j) / \pi_i(j)\}],$$

og dermed er satsen bevist. □

8.1 Estimering av variansen

La

$$S_{ij}^2(\underset{\sim}{J}) = \frac{I_i(j)}{n_{ij}(\underset{\sim}{J})-1} \sum_s \{X_{ijs} - \bar{X}_{ij}\}^2,$$

og

$$T_{ij}^2(\underset{\sim}{J}) = \frac{S_{ij}^2(\underset{\sim}{J})}{n_{ij}(\underset{\sim}{J})} \frac{N_i(j)-n_{ij}(\underset{\sim}{J})}{N_i(j)}.$$

Da er

$$E\{S_{ij}^2(\underset{\sim}{J}) \mid \underset{\sim}{J} = j,\ I_i(j) = 1\} = \sigma_i^2(j),$$

og

$$E\{T_{ij}^2(\underset{\sim}{J}) \mid \underset{\sim}{J} = j,\ I_i(j) = 1\} = \tau_{ij}^2(j).$$

Dermed er

$$E\{N_i^2(j)\, T_{ij}^2(\underset{\sim}{J}) \mid I_{ij}\} = I_{ij}\,\eta_i(j),$$

og

$$E\{N_i^2(j)\, T_{ij}^2(\underset{\sim}{J})\} = \pi_i(j)\,\eta_i(j).$$

Setning 8.3: En forventningsrett for var $\hat{a}$ er gitt ved

$$\text{est}_1 \text{ var } \hat{a} = \sum_i \{\sum_j \sum_k \frac{\pi_i(j,k)-\pi_i(j)\pi_i(k)}{\pi_i(j,k)} \cdot \frac{I_{ij}\, V_{ij}}{\pi_i(j)} \frac{I_{ik}\, V_{ik}}{\pi_i(k)}\}$$

$$+ \sum_i \sum_j \{N_i^2(j)\, T_{ij}^2(\underset{\sim}{J}) / \pi_i(j)\}.$$

Bevis: Da $E\{N_i^2(j)\, T_{ij}^2(\underset{\sim}{J})\} = \pi_i(j)\,\eta_i(j)$, er

$$E\left[\sum_i \sum_j \{N_i^2(j)\, T_{ij}^2(\underset{\sim}{J}) / \pi_i(j)\}\right] = \sum_i \sum_j \eta_i(j).$$

Ved sammen med uttrykkene for var $(I_{ij}\, V_{ij})$ og cov $(I_{ij}\, V_{ij}, I_{ik}\, V_{ik})$ å bruke at

$$E\{I_{ij}\, V_{ij}^2\} = \text{var}\{I_{ij}\, V_{ij}\} - \{E[I_{ij}\, V_{ij}]\}^2,$$

finner en

$$\sum_j \sum_k \frac{\pi_i(j,k)-\pi_i(j)\,\pi_i(k)}{\pi_i(j,k)} \frac{1}{\pi_i(j)} \frac{1}{\pi_i(k)} E\{I_{ij}\, V_{ij}\, I_{ik}\, V_{ik}\}$$

$$= \sum_j \sum_k \frac{\pi_i(j,k)-\pi_i(j)\,\pi_i(k)}{\pi_i(j)\,\pi_i(k)} a_i(j)\, a_i(k)$$

$$+ \sum_j \{\eta_i(j) / \pi_i(j)\} - \sum_j \eta_i(j).$$

Av uttrykket for var $\hat{a}$ følger da at

$$E\ \text{est}_1\ \text{var}\ \hat{a} = \text{var}\ \hat{a},$$

og satsen er bevist. □

La

$$I_{ijk} = \begin{cases} 1 \text{ dersom både utvalgsområdene j og k i stratum i} \\ \quad \text{er i utvalget,} \\ 0 \text{ ellers.} \end{cases}$$

Ved en enkel omforming kan $\text{est}_1\ \text{var}\ \hat{a}$ dermed skrives

$$\text{est}_1\ \text{var}\ \hat{a} = \sum_i \{ \sum_{j<k} \frac{\pi_i(j,k)-\pi_i(j)\,\pi_i(k)}{\pi_i(j,k)} I_{ijk} \left[\frac{V_{ij}}{\pi_i(j)} - \frac{V_{ik}}{\pi_i(k)}\right]^2 \}.$$

8.2. Selvveiende utvalg

En sier at utvalget er selvveiende dersom estimatoren $\hat{a}$ kan skrives på formen

$$\hat{a} = \frac{1}{b(\underset{\sim}{J})} \sum_i \sum_j I_{ij} \sum_s X_{ijs}.$$

La

$$n = \sum_i \sum_j n_{ij}(\underset{\sim}{J})$$

være gitt. Dersom $I_{ij}=0$, er $n_{ij}(\underset{\sim}{J})=0$. En oppnår at utvalget blir selvveiende dersom

$$b(\underset{\sim}{J}) = \pi_i(j)\, n_{ij}(\underset{\sim}{J}) \,/\, N_i(j),$$

$$b_i(\underset{\sim}{J}) = b(\underset{\sim}{J}) \,/\, \pi_i(j),$$

og

$$n_{ij}(\underset{\sim}{J}) = b_i(\underset{\sim}{J})\, N_i(j) \text{ for alle } j \text{ slik at } I_{ij}=1.$$

Da $n = \sum_i \sum_j n_{ij}(\underset{\sim}{J})$, er

$$b(\underset{\sim}{J}) = n \,/\, \sum_i \sum_j \{I_{ij}\, N_i(j) \,/\, \pi_i(j)\}.$$

8.3 Særtilfellet med kommuner som utgjør egne strata

Som nevnt i kapittel III er alle byer med flere enn 30 000 innbyggere tatt ut som egne strata. I disse strataene trekkes enhetene ut rent lotterisk. Anta at det er N_{0i} trekkeenheter i stratum i. Den k-te enheten i stratum i har verdien $a_{0i}(k)$ på det som måles. Tilsvarende til avsnittet foran innføres

$$a_{0i} = \sum_k a_{0i}(k),$$

og

$$\bar{a}_{0i} = a_{0i} \,/\, N_{0i}\ .$$

Av de N_{0i} enhetene trekkes et utvalg på

$$n_{0i}(\underset{\sim}{J}) = b(\underset{\sim}{J})\, N_{0i}$$

trekkeenheter. De $n_{0i}(\underset{\sim}{J})$ trekkeenhetene har numrene K_{0i1}, K_{0i2}, ...

La da

$$X_{0is} = a_{0i}(K_{0is}),$$

og

$$\bar{X}_{0i} = \sum_s X_{0is} \,/\, n_{0i}(\underset{\sim}{J}).$$

$$\hat{a}_{0i} = N_{0i}\, \bar{X}_{0i}$$

blir altså en estimator for totalen.

Tilsvarende til avsnittet foran defineres

$$\sigma_{0i}^2 = \frac{1}{N_{0i}-1} \sum_k \{a_{0i}(k) - \bar{a}_{0i}\}^2,$$

og

$$S_{0i}^2 = \frac{1}{n_{0i}-1} \sum_s \{X_{0is} - \bar{X}_{0i}\}^2.$$

Ifølge Hoem (1973, side 17) er

$$\text{var } \hat{a}_{0i} = \frac{\sigma_{0i}^2}{n} N_{0i} (N-n).$$

Laake (1974, side 3) har vist at

$$\text{est var } \hat{a}_{0i} = S_{0i}^2 N_0 (\frac{\hat{N}}{n} - 1),$$

der

$$\hat{N} = \sum_i \sum_j \{I_i(j) N_i(j) / \pi_i(j)\}$$

er en forventningsrett estimator for var $\hat{a}_{0i}$.
En estimator for totalen i alle kommunene som utgjør egne strata er

$$\hat{a}_0 = \sum_i \hat{a}_{0i}.$$

En forventningsrett estimator for totalen i hele landet blir dermed

$$\hat{\tilde{a}} = \hat{a} + \hat{a}_0.$$

Siden cov $(\hat{a}, \hat{a}_0) = 0$, er en forventningsrett estimator for var $\hat{\tilde{a}}$ gitt ved

$$\text{est var } \hat{\tilde{a}} = \text{est}_1 \text{ var } \hat{a} + \sum_i \text{est var } \hat{a}_{0i}.$$

8.4 Estimering av variansen når $m_i=1$ for alle i

I utvalgsplanen er antall strata så stort at en ikke kan trekke mer enn ett utvalgsområde innenfor hvert stratum. I dette tilfellet kan en ikke bruke estimatoren gitt ovenfor for variansen. Derfor slås strata sammen slik at hvert av de nye strataene inneholder minst to uttrukne utvalgsområder. Denne sammenslåingen må foretas før utvalgsområdene trekkes. En slik samling av sammenslåtte strata kalles en gruppe. Anta at det er H grupper og at der er L_h strata i gruppe h.
Da er

$$\hat{a}_i = \sum_j \{I_{ij} V_{ij} / \pi_i(j)\}$$

en forventningsrett estimator for totalen i stratum i. Som estimator for

variansen foreslås nå

$$\text{est}_2 \text{ var } \hat{a} = \sum_{h=1}^{H} \frac{L_h}{L_h - 1} \sum_{i=1}^{L_h} (\hat{a}_i - \frac{1}{L_h} \sum_{g=1}^{L_h} \hat{a}_g)^2.$$

Estimatoren er basert på at trekkingen på første trinn i hver gruppe foregår med tilbakelegging. Estimatoren er derfor ikke forventningsrett for variansen.

Setning 8.4: Skjevheten til estimatoren est_2 var $\hat{a}$ er gitt ved

$$E \text{ est}_2 \text{ var } \hat{a} - \text{var } \hat{a} = \sum_{h=1}^{H} \frac{L_h}{L_h - 1} \cdot \sum_{i=1}^{L_h} (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2.$$

Bevis: La

$$W_i^2 = (\hat{a}_i - \frac{1}{L_h} \sum_{g=1}^{L_h} \hat{a}_g)^2.$$

Denne observatoren har forventning

$$EW_i^2 = E\{(\hat{a}_i - E\hat{a}_i) - \frac{1}{L_h} (\sum_{g=1}^{L_h} \hat{a}_g - \sum_{g=1}^{L_h} a_g)$$

$$+ (E\hat{a}_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)\}^2 = E(\hat{a}_i - E\hat{a}_i)^2$$

$$+ E(\frac{1}{L_h} \sum_{g=1}^{L_h} \hat{a}_g - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2 + (E\hat{a}_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2$$

$$- 2E(\hat{a}_i - E\hat{a}_i) \frac{1}{L_h} \sum_{g=1}^{L_h} (\hat{a}_g - \sum_{g=1}^{L_h} a_g)$$

$$= \text{var } \hat{a}_i + \text{var } (\frac{1}{L_h} \sum_{g=1}^{L_h} \hat{a}_g) - 2 \text{ cov } (\hat{a}_i, \frac{1}{L_h} \sum_{g=1}^{L_h} \hat{a}_g)$$

$$+ (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2.$$

Da cov $(\hat{a}_i, \hat{a}_j)=0$ for $i \neq j$, er

$$EW_i^2 = \text{var } \hat{a}_i + \frac{1}{L_h^2} \sum_{g=1}^{L_h} \text{var } \hat{a}_g - \frac{2}{L_h} \text{ var } \hat{a}_i$$

$$+ (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2.$$

Det følger da at

$$\text{E est}_2 \text{ var } \hat{a} = \sum_{h=1}^{H} \frac{L_h}{L_h - 1} \sum_{i=1}^{L_h} (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2 + \sum_{h=1}^{H} \sum_{i=1}^{L_h} \text{var } \hat{a}_i ,$$

og en har derfor

$$\text{E est}_2 \text{ var } \hat{a} - \text{var } \hat{a} = \sum_{h=1}^{H} \frac{L_h}{L_h - 1} \sum_{i=1}^{L_h} (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g)^2 . \square$$

Vi ser altså at skjevheten til estimatoren for variansen er avhengig av differansen mellom populasjonstotalen i de strataene vi slår sammen. Dersom altså

$$\sum_{i=1}^{L_h} (a_i - \frac{1}{L_h} \sum_{g=1}^{L_h} a_g) = 0 \text{ for alle h,}$$

vil $\text{est}_2 \text{ var } \hat{a}$ være forventningsrett for $\text{var } \hat{a}$.

8.5 Forandringer i formlene når utvalget er selvveiende

Dersom utvalget er selvveiende, er

$$\hat{a}_i = \frac{1}{b(\underset{\sim}{J})} \sum_j I_{ij} \sum_s X_{ijs},$$

slik at $\text{est}_2 \text{ var } \hat{a}$ reduseres til

$$\text{est}_2 \text{ var } \hat{a} = \frac{1}{b^2(\underset{\sim}{J})} \sum_{h=1}^{H} \frac{L_h}{L_h - 1} \sum_{i=1}^{L_h} \{\sum_j I_{ij} \sum_s X_{ijs} - \frac{1}{L_h} \sum_{g=1}^{L_h} \sum_k I_{ik} \sum_r X_{ikr}\}^2 .$$

En estimator for variansen til estimatoren for totalen i hele landet er gitt ved

$$\text{est var } \hat{\underset{\sim}{a}} = \text{est var } \hat{a}_0 + \text{est}_2 \text{ var } \hat{a}.$$

Utkommet i serien SØS

Issued in the series Samfunnsøkonomiske studier (SØS)

Nr. 1 Det norske skattesystems virkninger på den personlige inntektsfordeling *The Effects of the Norwegian Tax System on the Personal Income Distribution* 1954 Sidetall 103 Pris kr 3,00

\- 2 Skatt på personleg inntekt og midel *Tax on Personal Income and Capital* 1954 Sidetall 120 Pris kr 3,00

\- 3 Økonomisk utsyn 1900-1950 *Economic Survey* 1955 Sidetall 217 Pris kr 4,00

\- 4 Nasjonalregnskap. Teoretiske prinsipper *National Accounts. Theoretical Principles* 1955 Sidetall 123 Pris kr 3,00

\- 5 Avskrivning og skattlegging *Depreciation and Taxation* 1956 Sidetall 85 Pris kr 3,00

\- 6 Bedriftsskatter i Danmark, Norge og Sverige *Corporate Taxes in Denmark, Norway and Sweden* 1958 Sidetall 101 Pris kr 4,00

\- 7 Det norske skattesystemet 1958 *The Norwegian System of Taxation* 1958 Sidetall 159 Pris kr 6,50

\- 8 Produksjonsstruktur, import og sysselsetting *Structure of Production, Imports and Employment* 1959 Sidetall 129 Pris kr 5,50

\- 9 Kryssløpsanalyse av produksjon og innsats i norske næringer 1954 *Input-Output Analysis of Norwegian Industries* 1960 Sidetall 614 Pris kr 10,00

\- 10 Dødeligheten og dens årsaker i Norge 1856-1955 *Trend of Mortality and Causes of Death in Norway* 1962 Sidetall 246 Pris kr 8,50

\- 11 Kriminalitet og sosial bakgrunn *Crimes and Social Background* 1962 Sidetall 194 Pris kr 7,00

\- 12 Norges økonomi etter krigen *The Norwegian Post-War Economy* 1965 Sidetall 437 Pris kr 15,00

\- 13 Ekteskap, fødsler og vandringer i Norge 1856-1960 *Marriages, Births and Migrations in Norway* 1965 Sidetall 221 Pris kr 9,00

\- 14 Foreign Ownership in Norwegian Enterprises *Utenlandske eierinteresser i norske bedrifter* 1965 Sidetall 213 Pris kr 12,00

\- 15 Progressiviteten i skattesystemet 1960 *Statistical Tax Incidence Investigation* 1966 Sidetall 95 Pris kr 7,00

\- 16 Langtidslinjer i norsk økonomi 1865-1960 *Trends in Norwegian Economy* 1966 Sidetall 150 Pris kr 8,00

\- 17 Dødelighet blant spedbarn i Norge 1901-1963 *Infant Mortality in Norway* 1966 Sidetall 74 Pris kr 7,00

\- 18 Storbyutvikling og arbeidsreiser En undersøkelse av pendling, befolkningsutvikling, næringsliv og urbanisering i Oslo-området *Metropolitan Growth, Commuting and Urbanization in the Oslo Area* 1966 Sidetall 298 Pris kr 12,00

\- 19 Det norske kredittmarked siden 1900 *The Norwegian Credit Market since 1900* Sidetall 395 Pris kr 11,00

\- 20 Det norske skattesystemet 1967 *The Norwegian System of Taxation* 1968 Sidetall 146 Pris kr 9,00

Nr. 21 Estimating Production Functions and Technical Change from Micro Data. An Exploratory Study of Individual Establishment Time-Series from Norwegian Mining and Manufacturing 1959-1967 *Estimering av produktfunksjoner og tekniske endringer fra mikro data. Analyser på grunnlag av tidsrekker for individuelle bedrifter fra norsk bergverk og industri* 1971 Sidetall 226 Pris kr 9,00

- 22 Forsvarets virkninger på norsk økonomi *The Impact of the Defence on the Norwegian Economy* 1972 Sidetall 141 Pris kr 9,00

- 23 Prisutvikling og prisatferd i 1960-årene En presentasjon og analyse av nasjonalregnskapets prisdata 1961-1969 *The Development and Behaviour of Prices in the 1960's Presentation and Analysis of the Price-Data of the Norwegian National Accounts* 1974 Sidetall 478 Pris kr 15,00

- 24 Det norske skattesystemet I Direkte skatter 1974 *The Norwegian System of Taxation I Direct Taxes* 1974 Sidetall 139 Pris kr 9,00

- 25 Friluftsliv, idrett og mosjon *Outdoor Recreation, Sport and Exercise* 1975 Sidetall 114 Pris kr 8,00

- 26 Nasjonalregnskap, modeller og analyse En artikkelsamling til Odd Aukrusts 60-årsdag *National Accounts, Models and Analysis To Odd Aukrust in Honour of his sixtieth Birthday* 1975 Sidetall 320 Pris kr 13,00

- 27 Den representative undersøgelsesmethode *The Representative Method of Statistical Surveys* 1976 Sidetall 64 Pris kr 8,00

- 28 Statistisk Sentralbyrå 100 år 1876-1976 *Central Bureau of Statistics 100 Years* 1976 Sidetall 128 Pris kr 9,00

- 29 Statistisk Sentralbyrås 100-årsjubileum Prolog og taler ved festmøtet i Universitetes aula 11. juni 1976 *Central Bureau of Statistics Prologue and Addresses at the Centenary Celebration, University Hall* 1976 Sidetall 32 Pris kr 7,00 ISBN 82-537-0637-5

- 30 Inntekts- og forbruksbeskatning fra et fordelingssynspunkt - En modell for empirisk analyse *Taxation of Income and Consumption from a Distributional Point of View - A model for Empirical Analysis* 1976 Sidetall 148 Pris kr 9,00 ISBN 82-537-0647-2

- 32 Inntekt og forbruk for funksjonshemma *Income and Consumer Expenditure of Disabled Persons* 1977 Sidetall 166 Pris kr 13,00 ISBN 82-537-0732-0

- 33 Prinsipper og metoder for Statistisk Sentralbyrås utvalgsundersøkelser *Sampling Methods Applied by the Central Bureau of Statistics of Norway* 1977 Sidetall 105 Pris kr 11,00 ISBN 82-537-0771-1

Publikasjonen utgis i kommisjon hos
H. Aschehoug & Co. og Universitetsforlaget, Oslo, og er til salgs
hos alle bokhandlere.
Pris kr 11,00.

ISBN 82-537-0771-1
Engers Boktrykkeri A/S - Otta