ARTIKLER 37

OM BRUK AV STIKKPRØVER VED KONTORET FOR INTERVJUUNDERSØKELSER, STATISTISK SENTRALBYRÅ

Av Steinar Tamsfoss

ON THE USE OF SAMPLING SURVEYS BY THE CENTRAL BUREAU OF STATISTICS, NORWAY

OSLO 1970

STATISTISK SENTRALBYRÅ

ARTIKLER FRA STATISTISK SENTRALBYRÅ NR. 37

OM BRUK AV STIKKPRØVER VED KONTORET FOR INTERVJUUNDERSØKELSER, STATISTISK SENTRALBYRÅ

Av Steinar Tamsfoss

ON THE USE OF SAMPLING SURVEYS BY THE CENTRAL BUREAU OF STATISTICS, NORWAY

OSLO 1970

MARIENDALS BOKTRYKKERI A/S, GJØVIK

Forord

Statistisk Sentralbyrå har i de siste år gjennomført en rekke undersøkelser på grunnlag av data hentet inn ved hjelp av den intervjuorganisasjonen som ble bygd opp i 1967. Denne artikkelen gir en oversikt over utvalgsplanen som har vært nyttet.

Artikkelen tar også opp noen problemer som er av mer allmenn natur for intervjuundersøkelser. I framstillingen har en i størst mulig grad unngått å bruke matematiske formuleringer.

Statistisk Sentralbyrå, Oslo, 11. april 1970.

Petter Jakob Bjerve

Preface

The Central Bureau of Statistics of Norway established in 1967 a permanent organization for sampling surveys based on interviewing. This article describes the sample design which has been applied during the first years.

The article also deals with problems of a more general nature in interview surveys. Mathematical formulations have been avoided as far as possible.

Central Bureau of Statistics, Oslo, 11 April 1970.

Petter Jakob Bjerve

Innhold

Contents

I. Innledning

Et særpreg ved industriland i økonomisk vekst, er at dyptgripende endringer finner sted på nesten alle områder i samfunnslivet. En forutsetning for kontroll og ledning av denne utviklingen er kunnskap om samfunnet og de krefter som styrer det. Ervervelse av slik kunnskap kan foregå på utallige måter. En av de metodene som særlig i den elektroniske datamaskinens dager har fått økende betydning, er stikkprøvemetoden eller såkalte utvalgsundersøkelser.

Ved slike undersøkelser hentes det inn data fra bare en utvalgt del av den befolkningen en ønsker å vite noe om. Innhenting og bearbeiding av oppgavene kan derfor skje langt hurtigere og mye billigere enn tilfellet ville være dersom hele populasjonen skulle undersøkes.

Resultatene fra slike undersøkelser er alltid påvirket av en rekke feilkilder. Noen av disse feilkildene kan kontrolleres ut fra teorien om statistiske stikkprøver, men det finnes andre som ikke så lett lar seg kontrollere. I moderne samplingteori ser det ut til at de siste feilkildene blir ofret større oppmerksomhet enn hva som tidligere var tilfelle. De kontrollerbare feilkildene kalles gjerne for varians, og lar seg beregne når utvalgsmetodikken er kjent.

Utvalgsundersøkelser nyttes i en viss utstrekning for innhenting av oppgaver til offisiell statistikk. I Norge blir f.eks. deler av produksjons- og lønnsstatistikken og jordbruksundersøkelsene basert på innhenting av oppgaver fra et utvalg av telle-enheter. For en rekke problemstillinger er det påkrevd å hente inn oppgavene gjennom intervjuere. Dette har i en rekke land ført til opprettelsen av egne organisasjoner som driver med intervjuundersøkelser.

I januar 1967 opprettet Statistisk Sentralbyrå en egen intervjuorganisasjon. Før denne tid hadde imidlertid flere private institusjoner her i landet drevet med intervjuing basert på stikkprøver.

I denne artikkelen vil hovedvekten bli lagt på en beskrivelse av den utvalgsplanen som er utviklet i Byrået. Varianser og feilkilder av ikke-

statistisk natur vil bli behandlet, mens de kostnadsmessige aspektene bare blir berørt i liten grad.

I vedlegget vil noen av de statistiske problemene i artikkelen bli utdypet litt nærmere.

II. Grunnleggende forutsetninger

Et av hovedformålene ved en utvalgsplan for Statistisk Sentralbyrå er at undersøkelsene kan gi resultater som gjelder for hele landet. For at dette skal kunne skje, må utvalget være «representativt» for hele landet, dvs. at utvelgingen må skje fra hele populasjonen. Hvis det foretas en *enkel tilfeldig utvelging* fra totalpopulasjonen, vil en rekke problemer reise seg. Blant annet vil vi måtte ha et stort antall intervjuere, og disse ville få lange reiser for å oppsøke intervjuobjektene (IO). En slik arbeidsmetode vil derfor være både kostbar og tidkrevende.

Det finnes imidlertid andre metoder som er mer hensiktsmessige når det gjelder økonomi og arbeid. To-trinns-utvelging er en slik metode, og det er denne Statistisk Sentralbyrå har valgt. Metoden innebærer nemlig at intervjuarbeidet blir konsentrert til noen få geografisk avgrensede områder («utvalgsområder»).

Teoretiske, kostnads- og arbeidsmessige vurderinger har ført fram til en del grunnleggende forutsetninger:

i) Utvalgsplanen skal gi grunnlag for undersøkelser av varierende art (general purpose sample).
ii) «Utvalgsområdene» skal ha en samlet folkemengde som utgjør 5—6 prosent av landets totale folkemengde.
iii) Antall intervjuere skal være 100.
iv) Det tas sikte på å gi tall gjeldende for hele landet. Dessuten bør utvalgsplanen konstrueres slik at det kan gis spesifikasjoner for handelsfelt og de 3 største byene.
v) Dersom et «utvalgsområde» av en eller annen grunn skulle falle fra i en undersøkelse, skal det kunne erstattes av et annet («makkerområde»).
vi) Utvalget av telle-enheter skal være «selv-veiende», dvs. at alle enhetene skal bidra med samme vekt i estimatorene.

III. Primære utvalgsenheter (PUE)

I forbindelse med to-trinns-utvelging opererer en gjerne med to forskjellige utvalgsenheter — en enhetstype til hvert trinn. På første utvelgingstrinn kalles enhetene for *primære utvalgsenheter* som forkortet skrives pue. To-trinns-utvelgingen består da i å trekke intervjuenhetene innenfor et utvalg av primære utvalgsenheter. Intervjuenhetene som altså trekkes på andre trinn, kalles formelt for *sekundære utvalgsenheter* og forkortes sue. Disse vil bli nærmere behandlet i et senere avsnitt.

Hele landet er delt opp i primære utvalgsenheter, hver med et folketall på om lag 2 000 mennesker. Oslo, som på flere måter danner unntak fra de generelle reglene som er brukt, er delt opp i 40 slike enheter med et gjennomsnittlig folketall på ca. 12 500 (mellom 6 000 og 20 000 mennesker).

Grunnlaget for konstruksjonen av de primære utvalgsenhetene er tellingskretsene fra Folketellingen 1960. Tellingskretser er slått sammen eller delt slik at det ønskede folketall er oppnådd. Ved sammenslåing av kretsene er det sørget for at hver primær utvalgsenhet blir et geografisk sammenhengende område. Bare i noen få tilfelle er tellingskretser fra forskjellige kommuner satt sammen til primære utvalgsenheter.[1] Grensene til de 5 handelsfeltene[2] og de 3 største byene blir ikke i noen tilfelle krysset av grenser mellom primære utvalgsenheter.

Ved inndelingen av kommuner i tellingskretser har en alltid skilt ut tettbygde strøk som egne tellingskretser. Ved konstruksjonen av de primære utvalgsenhetene har en i størst mulig utstrekning prøvd å overføre egenskapene tett- og spredtbygde strøk til å bli karakteristika også for disse enhetene. Samtlige enheter kan derfor — om ønskelig — sorteres på disse kategoriene. Det totale antall primære utvalgsenheter er 1 541, hvorav 40 i Oslo.

Argumentene for å gjøre primærinndelingen på denne måten er flere. Teoretisk har en tatt hensyn til hvordan variansene på de mest brukte estimatorene avhenger av primær-enhetenes antall og størrelse. For flervariabel optimalisering av utvalgsplanen gir som regel en design med like store primær-enheter den beste optimalitet (dvs. minst varians). Dessuten er primær-enhetenes likhet i størrelse behagelig ut fra ønsket

---

Noter: [1] Om registeret og tellingskretsene henvises til kapittel VIII.
[2] Utvalgsplanen bygger på den gamle inndelingen av landet i 5 handelsfelt. Fra 1/1—68 er tallet på handelsfelt blitt 4 idet Søndre handelsfelt er delt mellom Østre og Vestre handelsfelt. Denne revisjonen har ingen praktiske konsekvenser for utvalgsplanen.

om oversiktlighet og enkle vurderinger av forskjellige samplingproblemer.

En mer kostnadsmessig begrunnelse er at dersom en intervjuer skal ha ansvar for og arbeide i større geografiske områder, vil relativt store utgifter gå med til ofte tidkrevende reiser. Størrelsen på våre primære utvalgsenheter er imidlertid slik at en intervjuer uten å spille for mye tid kan rekke over hele området til noenlunde rimelige kostnader pr. intervju. Like store områder bidrar også til at hver intervjuer får omtrent lik arbeidsbyrde.

Generelt er det én intervjuer pr. utvalgt primær-enhet, men i Oslo hvor enhetene er mye større enn ellers, har en to intervjuere pr. primærenhet.

Selv om stratifikasjonen ikke er omtalt ennå, kan det nevnes at i alt 93 primære utvalgsenheter er valgt ut. 7 av disse ligger i Oslo, så det totale antall intervjuere blir 100. Det samlede folketall i de utvalgte enhetene er ca. 255 000, eller mellom 6—7 prosent av totalbefolkningen (3,8 mill., 1968). De utvalgte primær-enhetene kalles gjerne for «utvalgsområder». Forutsetningene II ii) og iii) er dermed oppfylt.

IV. Representativitet

Dette avsnittet og neste er en presisering av punkt II i), iv), v) og vi). «Representativitet» er et meget brukt (og misbrukt) begrep. Det er derfor hensiktsmessig å definere hva som skal forstås med «representativt utvalg» i denne sammenheng.

I statistiske termer er et utvalg representativt hvis sannsynligheten for at en vilkårlig valgt (utvalgs-)enhet i populasjonen skal komme i utvalget, er *kjent*.

Forskjellige enheter kan gjerne ha ulik sannsynlighet for å bli trukket ut, hovedpoenget er at man vet hvilken sannsynlighet som er knyttet til den bestemte enhet. Hvis samtlige enheter har *like stor* sjanse til å velges ut, kaller vi utvalget for *selv-veiende*.

Et hyppig forekommende misbruk av begrepet representativitet er å karakterisere *store* utvalg som representative — selv om utvelgingssannsynligheten(e) ikke er kjent(e). Definisjonen ovenfor dekker imidlertid utvalg på bare én enhet av befolkningen. Representativiteten har derfor lite å gjøre med utvalgets størrelse. (Derimot er resultatenes *pålitelighet* i høy grad avhengig av størrelsen, men en objektiv estimering av den (påliteligheten) forutsetter et representativt utvalg.)

V. Stratifikasjon

Statistiske metoder blir som regel bedre hvis en gjør bruk av a priori kunnskap om den populasjon som skal utforskes. Slik kunnskap ligger f.eks. til grunn når populasjonen blir stratifisert, dvs. at de utvalgsenhetene som er mest mulig like med hensyn på et eller annet kjennetegn, blir samlet i grupper eller strata. («Kjennetegn» = verdi på en variabel.) Gevinsten av stratifikasjonen blir som regel større jo sterkere sammenheng det er mellom stratifikasjonsvariabelen og den variabel som utforskes.

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. Videre er det klart at det ville være nokså meningsløst å gi seg inn på en stratifikasjon av sekundærenhetene så lenge utvalgsplanen skal være generell. Sekundærenhetene (egentlig analyseenhetene) skal kunne variere uten at opplegget endres. I denne situasjonen er det derfor bare mulig å stratifisere utvalgsenhetene på 1. trinn.

De primære utvalgsenhetene er gruppert etter to variable: *geografisk beliggenhet* og *nærings-struktur.*

Skalaen for geografisk beliggenhet er definert som handelsfelt og de 3 største byene. Vi har altså *8 geografiske strata* (handelsfelt + 3 største byene). De 3 byene er selvsagt «tatt ut» av sine respektive handelsfelt og blir således ikke medregnet når handelsfeltene er Østre, Vestre eller Midtre.

Innen hvert geografisk stratum (unntatt Oslo) er det igjen foretatt en gruppering av primær-enhetene etter hvilken næringsstruktur de har. På denne måten blir enhetene samlet i homogene grupper som er våre endelige strata.

Næringsstruktur er en sammensatt variabel med 4 dimensjoner (grupper av næringer), nemlig

1) Jord- og skogbruk,
2) Industri m.v.,
3) Fiske og fangst og
4) Tjenesteyting.

For hver primær-enhet er frekvensfordelingen av yrkesbefolkningen over disse gruppene beregnet. De enhetene som har mest mulig lik næringsstruktur er så slått sammen innenfor hvert geografisk stratum. I praksis har det vært tilstrekkelig å bare ta hensyn til noen av de 4

dimensjonene i næringsstrukturen ved stratifikasjonen. Næringsskalaen er forskjellig for hvert geografisk stratum.

Stratumgrensene er satt slik at hvert stratum i «handelsfeltene» inneholder omtrent like mange primære utvalgsenheter, mellom 30—38. For Trondheim og Bergen er tallene mellom 15—18.

Oslo er som nevnt ikke stratifisert etter næringsstruktur. Her er det foretatt en geografisk gruppering av primær-enhetene til i alt 7 strata eller grupper. Hvert stratum i Oslo er således sammensatt av 5—7 enheter.

Skjematisk illustrasjon til stratifikasjonen

I de nedenstående figurer blir det gitt noen mer konkrete oppgaver over stratifikasjonen. Hver hovedblokk representerer et geografisk stratum, og er delt opp i mindre enheter — nærings-strataene. Næringsstrataene er (for oversiktens skyld) tegnet som sammenhengende blokker av primære utvalgsenheter. Dette må imidlertid ikke oppfattes som noe geografisk grensefellesskap mellom de enhetene som tilhører samme stratum. For hvert stratum er næringsstrukturen angitt for de variable som har vært betydningsfulle. Dessuten er det angitt hvor mange primær-enheter vedkommende stratum inneholder. For Oslos vedkommende er ikke næringsstrukturen beregnet, så her er bare angitt antall enheter i hvert stratum.

Fra hvert stratum er det trukket to utvalgsområder — unntatt i Trondheim, Bergen og Oslo hvor bare ett utvalgsområde er trukket pr. stratum. De to utvalgsområdene fra hvert stratum — utenom de tre største byene — kalles *makkerområder*. En må regne med at et utvalgsområde kan falle fra i en undersøkelse fordi f.eks. intervjueren er syk. For å sikre at vedkommende stratum blir *representert* i utvalget, er det hensiktsmessig å ha to utvalgsområder i hvert stratum. I Oslo, Bergen og Trondheim vil det være lettere å få erstatning for intervjueren, og det er derfor trukket bare ett område fra hvert stratum. For å estimere varianser er det også nødvendig at minst to primær-enheter er valgt ut fra hvert stratum.

Hvis et utvalgsområde skulle falle fra, blir oppgavene fra makkerområdet fordoblet slik at «ingen» skjevheter oppstår av den grunn.

Illustrasjon til stratifikasjonen

Stratum nr.: 1 Antall pue: 37 Industri: 0—18 %	Stratum nr.: 6 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: 20—45 % Tjenesteyting: over 15 %	Stratum nr.: 12 Antall pue: 38 Industri: 40—60 % Jord-skogbruk: over 1 % Tjenesteyting: 0— 9 %
Stratum nr.: 2 Antall pue: 37 Industri: 19—39 % Jord-skogbruk: 10—19 % Tjenesteyting: 0—27 %	Stratum nr.: 7 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: over 45 % Tjenesteyting: 5—10 %	Stratum nr.: 13 Antall pue: 38 Industri: 40—60 % Jord-skogbruk: over 1 % Tjenesteyting: 10—13 %
Stratum nr.: 3 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: 10—19 % Tjenesteyting: over 27 %	Stratum nr.: 8 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: over 45 % Tjenesteyting: over 10 %	Stratum nr.: 14 Antall pue: 37 Industri: 40—60 % Jord-skogbruk: over 1 % Tjenesteyting: over 13 %
Stratum nr.: 4 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: 20—45 % Tjenesteyting: 0—11 %	Stratum nr.: 9 Antall pue: 37 Industri: 40—60 % Jord-skogbruk: 0— 1 % Tjenesteyting: 0—14 %	Stratum nr.: 15 Antall pue: 35 Industri: 61—65 %
Stratum nr.: 5 Antall pue: 38 Industri: 19—39 % Jord-skogbruk: 20—45 % Tjenesteyting: 12—15 %	Stratum nr.: 10 Antall pue: 37 Industri: 40—60 % Jord-skogbruk: 0— 1 % Tjenesteyting: 15—19 %	Stratum nr.: 16 Antall pue: 36 Industri: over 65 %
	Stratum nr.: 11 Antall pue: 37 Industri: 40—60 % Jord-skogbruk: 0— 1 % Tjenesteyting: over 19 %	

Østre handelsfelt
ekskl. Oslo

Stratum nr.: 17 Antall pue: 35 Industri: 0—30 %
Stratum nr.: 18 Antall pue: 36 Industri: 31—40 %
Stratum nr.: 19 Antall pue: 35 Industri: over 40 %

Søndre handelsfelt

Stratum nr.:	20
Antall pue:	35
Industri:	0—23 %
Fiske-fangst:	0— 7 %

Stratum nr.:	21
Antall pue:	35
Industri:	0—23 %
Fiske-fangst:	over 7 %

Stratum nr.:	22
Antall pue:	35
Industri:	24—36 %
Jord-skogbruk:	over 30 %
Fiske-fangst:	0— 3 %

Stratum nr.:	23
Antall pue:	35
Industri:	24—36 %
Fiske-fangst:	over 3 %

Stratum nr.:	24
Antall pue:	35
Industri:	24—36 %
Jord-skogbruk:	0—30 %
Fiske-fangst:	0— 3 %

Stratum nr.:	25
Antall pue:	35
Industri:	37—57 %
Jord-skogbruk:	over 12 %

Stratum nr.:	26
Antall pue:	35
Industri:	37—44 %
Jord-skogbruk:	0—12 %

Stratum nr.:	27
Antall pue:	36
Industri:	45—57 %
Jord-skogbruk:	0—12 %

Stratum nr.:	28
Antall pue:	35
Industri:	58—81 %

Vestre handelsfelt
ekskl. Bergen

Stratum nr.:	29
Antall pue:	33
Industri:	0—27 %
Jord-skogbruk:	over 51 %

Stratum nr.:	30
Antall pue:	33
Industri:	0—27 %
Fiske-fangst:	over 16 %

Stratum nr.:	31
Antall pue:	34
Industri:	0—27 %
Jord-skogbruk:	0—16 %
Fiske-fangst:	0—16 %

Stratum nr.:	32
Antall pue:	34
Industri:	28—44 %
Jord-skogbruk:	over 31 %

Stratum nr.:	33
Antall pue:	33
Industri:	28—44 %
Jord-skogbruk:	0—31 %

Stratum nr.:	34
Antall pue:	33
Industri:	45—70 %

Midtre handelsfelt
ekskl. Trondheim

Stratum nr.:	35
Antall pue:	31
Industri:	0—19 %
Jord-skogbruk } Fiske-fangst }	over 40 %

Stratum nr.:	36
Antall pue:	30
Industri:	0—19 %
Jord-skogbruk } Fiske-fangst }	0—40 %

Stratum nr.:	37
Antall pue:	30
Industri:	20—34 %
Jord-skogbruk:	over 22 %

Stratum nr.:	38
Antall pue:	30
Industri:	20—34 %
Fiske-fangst:	over 18 %

Stratum nr.:	39
Antall pue:	30
Industri:	20—34 %
Jord-skogbruk:	0—22 %
Fiske-fangst:	0—18 %

Stratum nr.:	40
Antall pue:	30
Industri:	35—74 %

Nordre handelsfelt

Stratum nr.:	41
Antall pue:	18
Industri:	0—34 %

Stratum nr.:	42
Antall pue:	18
Industri:	35—39 %

Stratum nr.:	43
Antall pue:	18
Industri:	over 39 %

Bergen

Stratum nr.:	44
Antall pue:	16
Industri:	0—32 %

Stratum nr.:	45
Antall pue:	16
Industri:	33—37 %

Stratum nr.:	46
Antall pue:	15
Industri:	over 37 %

Trondheim

Stratum nr.:	37
Antall pue:	7

Stratum nr.:	48
Antall pue:	6

Stratum nr.:	49
Antall pue:	5

Stratum nr.:	50
Antall pue:	5

Stratum nr.:	51
Antall pue:	5

Stratum nr.:	52
Antall pue:	6

Stratum nr.:	53
Antall pue:	6

Oslo

VI. Trekking på 1. trinn. Utvalgsområder

På 1. trinn trekkes *utvalgsområdene* stratumvis. Sannsynligheten for at en vilkårlig valgt primær utvalgsenhet i stratum nr. i skal komme i utvalget (dvs. bli et utvalgsområde), kalles for *utvalgsbrøk 1. trinn, stratum i,* og skrives

f_{1i}; i = 1, 2, . . . , s. hvor s er tallet på strata.
La videre
M_i = totalt antall pue i stratum nr.i.
m_i = antall utvalgte pue i stratum nr.i.
Vi har da at: $f_{1i} = \frac{m_i}{M_i}$

VII. Trekking på 2. trinn

1. Definisjon av enheter

Som regel går en intervjuundersøkelse ut på å foreta et intervju med én person som gjerne kan representere f.eks. en hel husholdning eller familie. Intervjuet blir da foretatt på vegne av en veldefinert enhet som selvsagt kan variere i ulike undersøkelser. Denne enheten, som oppgavene intervjuet innbringer, skal gjelde for, kalles for *analyse-enhet.* Ofte er det imidlertid slik at en del vansker er forbundet med å velge ut analyse-enhetene direkte, kanskje særlig når analyse-enheten er et individ. Befolkningen er stadig i «bevegelse», det forekommer fødsler, dødsfall, skilsmisser, ekteskap, flyttinger osv. En spesifisert analyseenhet kan derfor være vanskelig å finne når utvelgingen skjer fra et register som ikke i tilfredsstillende grad er ajourført med hensyn til navngitte personer. Registeret er derimot bedre når det gjelder adressefortegnelse. Adressene er stabile unntatt tilfelle av brann, sanering eller nybygging (nye adresser). Hvis adressen defineres ved hjelp av sted, vei/gate, oppgang og *leilighet,* vil en nesten alltid finne at adressen rommer en analyse-enhet (ved de fleste av våre undersøkelser én og bare en analyse-enhet). Vi velger derfor å se bort fra *hvem* som bor i leiligheten, og velger ut selve leiligheten. Intervjuobjektet eller analyse-enheten blir da det IO som på intervjutidspunktet bor i leiligheten. Dette er altså en form for indirekte utvelging av IO. De enhetene som direkte velges. ut, leilighetene, kaller vi for *sekundære utvalgsenheter (sue).*

Ved våre undersøkelser blir det alltid sørget for at overensstemmelsen mellom analyse-enheter og sekundær-enheter er en-entydig, dvs. at det i en sekundær utvalgsenhet er bare en analyse-enhet, og at alle analyseenheter tilhører en eller annen sekundær-enhet.

2. Utvalgs-sannsynligheter

Vi definerer følgende størrelser:

N totalt antall sue i populasjonen
n = totalt antall sue i utvalget
N_i = totalt antall sue i stratum i
n_i = antall sue i utvalget fra stratum i
N_{ij} = totalt antall sue i pue (i, j)
n_{ij} = antall utvalgte sue fra pue (i, j)

Her er innført skrivemåten «pue (i, j)» som betyr «pue nr. j i stratum nr. i». Skrivemåten generaliseres også til å gjelde sue, og da betyr «sue (i, j, k)»: «sue nr. k i pue nr. j i stratum nr. i».

La videre:

f = sannsynligheten for en vilkårlig valgt sue i populasjonen å komme i utvalget (= «total utvalgsbrøk»)
f_i = sannsynligheten for en vilkårlig valgt sue i stratum nr. i å komme i utvalget (= «utvalgsbrøk stratum i»)
f_{2i} = betinget sannsynlighet for en vilkårlig valgt sue i stratum nr. i å komme i utvalget, gitt at vedkommende sue er i et vilkårlig valgt utvalgsområde i samme stratum
f_{2ij} = betinget sannsynlighet for en vilkårlig valgt sue i stratum i å komme i utvalget, gitt at vedkommende sue er i pue (i, j)

Vi har her at:

$i = 1, 2, \ldots, s$ og $j = 1, 2, \ldots, M_i$.

Ut fra den generelle teori for multivariabel, stratifisert sampling vil stikkprøven gjerne ha en optimal fordeling på strata og innen strata når utvelgingen foretas proporsjonalt fra strata og fra utvalgsområder innen strata.

Anta at de m_i utvalgte primær-enhetene fra stratum i er pue'ne $(i, j_1), \ldots, (i, j_{m_i})$. Setter vi $f_{2ij} = f_{2i}$ for $j = j_1, j_2, \ldots, j_{m_i}$, så vil utvelgingen foregå proporsjonalt fra utvalgsområdene innen samme stratum. f_{2i} kalles da for *utvalgsbrøk, 2. trinn, stratum i.*

Hvis vi også lar $f_i = f$ for $i = 1, 2, \ldots, s$, vil utvelgingen fra hvert stratum foretas proporsjonalt.

Ifølge definisjon av betinget sannsynlighet har vi nå følgende sammenheng mellom f, f_i, f_{1i} og f_{2i}:

$$f = f_i = f_{1i} \cdot f_{2i}$$

Den følgende framstillingen kan nærmest kalles for en «bokføringsmessig» redegjørelse for hvordan stikkprøven fordeles på strata og utvalgsområder. I vedlegget blir det gitt en mer statistisk beskrivelse av dette problemet.

Av de foran gitte definisjoner finnes nokså enkelt at de forskjellige utvelgings-sannsynlighetene må være:

$$f = \frac{n}{N},\ f_i = \frac{n_i}{N_i},\ f_{2i} = \frac{n_{ij}}{N_{ij}};\ j = j_1, \ldots, j_{m_i}.$$

Fra før har vi at $f_{1i} = \frac{m_i}{M_i}$, og idet $f = f_{1i}\, f_{2i}$, følger at n_{ij} bestemmes ut fra ligningen:

$$n_{ij} = \frac{f}{f_{1i}} N_{ij}$$

Dermed blir: $n_i = \sum_{\nu=1}^{m_i} n_{ij_\nu}$ og $n = \sum_{i=1}^{s} n_i$

Når f og f_{1i} er fastsatt, er utvalgets størrelse tilnærmet bestemt. (Dette gjelder ikke generelt, men slik vår utvalgsplan er konstruert, må en regne med at påstanden holder. Se ellers vedlegget.) Siden vi spesielt har valgt å sette alle $f_i = f$ ($i = 1, 2, \ldots, s$), vil utvalget være *selv-veiende*. Dette gjør blant annet estimeringsproblemene enklere enn hva som ellers ville ha vært tilfelle.

VIII. Registeret

Utvalgene trekkes fra et register som er laget spesielt med henblikk på denne utvalgsplanen. Utvalgsområdene er trukket «en gang for alle». En har derfor kunnet innskrenke registeret til bare å omfatte utvalgsområdene.

For å få kartlagt utvalgsområdene så nøyaktig som mulig, har oppbyggingen av registeret vært foretatt i nært samarbeid med de respektive folkeregistre. I denne sammenheng gjør visse problemer seg gjeldende. Som nevnt tidligere, er de primære utvalgsenhetene konstruert på grunnlag av tellingskretsene ved Folketellingen 1960, men da tellingskretsene ikke har noen aktualitet for folkeregistrene, blir ikke tellingskretsenes numre punchet inn på hullkortene.

Høsten 1966, før feltarbeidet ble satt igang, ble et fullstendig register over alle bosatte i utvalgsområdene etablert. I listene for hver tellingskrets fra Folketellingen 1960, var adressene nøyaktig oppført i form av gårds- og bruksnr. eller gate- og husnr. Byrået sendte de ulike folkeregistre en nøyaktig fortegnelse over hvilke gårds- og bruksnr./gate- og husnr. som ifølge disse oppgaver hørte med til utvalgsområdet. Det ble anmodet om å få utlånt folkeregisterets hullkort for personer som pr. d.d. bodde i de nevnte hus. Samtidig bad en om at mulige nye bosteder innenfor *de geografiske grenser* av nevnte tellingskretser skulle tas med.

De utlånte kort ble reprodusert i Byrået, folketallet ble sammenliknet med folketallet i 1960, og større avvik ble nærmere gransket.

Etter etableringen av registeret har ajourholdet skapt et kontinuerlig problem. Byrået får oppgaver over flyttinger, dødsfall, fødsler o.l. for en kommune som helhet. I disse meldinger finnes ingen opplysning om tellingskretsens nr. Etter 1960 har en også hatt en rekke sammenslåinger av kommuner, hvilket har skapt ytterligere komplikasjoner.

Et visst ajourhold av navngitte personer på oppgitte adresser kan gjennomføres ved at listen over uttrukne IO sendes folkeregistrene foran hver ny undersøkelse. (Se avsn. XI 2.)

For å tilfredsstille behovet for mer fullstendige rettinger, blir registeret med jamne mellomrom (ca. 2 år) fornyet i sin helhet ved at folkeregistrene sender den tidligere omtalte kortmasse til Byrået. Ved denne prosedyre får en også tilført nybygg i utvalgsområdene. Likevel hender det ikke så sjelden at når intervjueren kommer ut i marken, så har det skjedd endringer i form av flyttinger, dødsfall m.v., og det blir da intervjuerens oppgave å gi beskjed om dette.

Registeret inneholder opplysninger om følgende kjennetegn for personer:

— navn, adresse
— personnr. (inklusive fødselsdato og kjønn)
— ekteskapelig status
— statsborgerskap og
— stilling i husholdningen

Som regel er den sekundære utvalgsenheten i våre undersøkelser en *boligenhet* (leilighet, adresse). Personene som hører til de ulike boligenhetene, kan av og til være vanskelig å markere på listene. Grupperingen av personer som hører til samme boligenhet, må derfor til en viss grad foretas skjønnsmessig. Registerets opplysninger om navn, adresse, ekteskapelig status, fødselsår og stilling i husholdningen er selvsagt til god hjelp i dette arbeidet. For enkelte utvalgsområders vedkommende (Oslo) finnes det også opplysninger om oppgang og etasje i adressene.

IX. Feilkilder

En utvalgsundersøkelse kan inndeles i 3 generelle faser:

— Planlegging og trekking av utvalg
— Målinger
— Databehandling

I hver av disse fasene må en regne med at det foreligger kilder til feil.

I det følgende vil bare de feilkilder som knytter seg til vår spesielle utvalgstype bli behandlet.

Også feilene kan vi på tilsvarende måte inndele i 3 uavhengige kategorier:

1. Utvalgsfeil
2. Målingsfeil
3. Databehandlingsfeil

Alle disse feiltypene bidrar til totalfeilen. Det kan selvsagt inntreffe at denne er null, men dette må i de aller fleste tilfelle tolkes slik at de forskjellige feilene har opphevet hverandre.

1. Utvalgsfeil

a) Varians

Variansen er et mål på den spredningen av enkeltobservasjonene som gir seg til kjenne i et materiale. Når vi kjenner variansen til en størrelse (stokastisk variabel), så vet vi hvilke grenser vi kan vente at en enkeltmåling av vedkommende størrelse ligger innenfor.

I sammenheng med stikkprøver, er det spesielt variansen på de estimatorene vi bruker, som er av interesse. Anta at gjennomsnittet av observasjonene *i utvalget*, $\bar{x}$, er estimator for gjennomsnittet av den samme størrelse *i populasjonen*, $\overline{X}$. Variansen til $\bar{x}$, var $\bar{x}$, er da et uttrykk for de variasjoner i $\bar{x}$ vi vil finne ved å gjenta målingen av $\bar{x}$ ved flere stikkprøver.

Varians (eller standard-avvik) er altså ikke noen «feil» i egentlig forstand. Mer adekvate begreper er usikkerhet, slumpvariasjon, tilfeldige variasjoner eller spredning. Varians blir behandlet nærmere i et senere avsnitt.

b) Utvalgs-skjevhet (Utvalgsbias)

Denne feiltypen er resultatet av et systematisk avvik fra den fastlagte samplingsprosedyre. Slike avvik kan framkomme på forskjellige måter:

— Trekkingsmekanismen virker ikke som foreskrevet

— Andre avvik fra den ønskede plan

Utvalgene i våre undersøkelser blir trukket ved hjelp av «tilfeldige tall» (random digits) som er fullt ut brukbare til våre formål.

Det svake punkt ved utvelgelsen av IO er imidlertid at adresse- eller personregisteret som brukes, ikke er helt tilfredsstillende ajourført. Dette kan ha en viss innflytelse på størrelsen og sammensetningen av frafallet som vil bli behandlet mer inngående i et senere avsnitt.

2. Målingsfeil

a) Fastlegging av hva som skal måles

Grunnlaget for fastlegging av hva som skal måles er oppdragsgiverens målsetting for undersøkelsen. For å få opplegget i samsvar med oppdragsgiverens intensjoner er det viktig at målsettingen er tilstrekkelig presisert f.eks. i form av konkrete problemstillinger eller utkast til tabeller. Målsettingen får avgjørende betydning når det gjelder å avgrense problemfeltet for undersøkelsen, hvilke begreper som skal nyttes og hvordan disse skal defineres. I mange tilfelle er det imidlertid ikke mulig for oppdragsgiveren på forhånd å fastlegge en tilstrekkelig klar målsetting. Han må derfor vurdere om den begrepsmodell en kommer fram til, tilfredsstiller hans intensjoner for undersøkelsen.

Av og til, særlig når begrepene som skal brukes, refererer til uklare eller abstrakte fenomener, kan det være vanskelig å operasjonalisere begrepene. Dette kan selvsagt føre til at en undersøker problemer som ikke helt er de samme som intensjonene tilsier.

b) Konstruksjon av måleinstrument

Måleinstrumentet i våre undersøkelser er *spørsmål.* Spørsmålene er en operasjonalisering av begrepsmodellen og avhenger derfor av hvor «god» denne er. Det er dessverre et faktum at språket ikke alltid er så entydig og presist. Det kan meget fort oppstå kjedelige feil hvis spørsmålene inneholder ord og formuleringer som ikke er omhyggelig vurdert. Som regel blir spørsmålene etterfulgt av fast oppsatte svaralternativer. Det hender ikke så rent sjelden — særlig når det gjelder holdningsspørsmål — at IO finner å måtte svare «blankt», fordi ingen av de oppsatte alternativer passer for dets mening. Likeså kan det forekomme at IO i et spørsmål får seg forelagt en problemstilling som det aldri har tenkt over. At svarene i slike tilfelle blir nokså vilkårlige — hvis en i det hele tatt får noe svar — er et faktum som kan telle sterkt med. Spørsmål som påvirker eller provoserer IO, er det helt forkastelig å ha med i en god undersøkelse. Det kan settes opp noen generelle krav til et godt spørsmål:

— Enkel formulering som alle forstår
— Entydig formulering som ikke kan misforstås
— Fremmedord eller andre vanskelige ord bør ikke forekomme
— Menings-ledende formuleringer må utelates
— Suggestive ord/formuleringer må unngås

— Spørsmålet bør ikke være av en slik art (eller formulert slik) at det representerer en problemstilling som er helt ny for IO.

Av svaralternativene må det kreves at de gir rom for så mange nyanser som mulig eller rimelig. En annen mulighet som imidlertid byr på flere praktiske problemer, er å la spørsmålene være åpne, dvs. at IO's svar fylles inn ordrett på intervjuskjemaet.

c) Målingen

Uansett hvor mange ganger en måling foretas, vil omstendighetene i målingsøyeblikket aldri være de samme. Dette reiser derfor spørsmålet om variasjoner i omstendighetene er store nok til å påvirke måleresultatet i en eller annen retning. Feil av denne typen er som så mange andre feilkilder, nesten umulig å rette opp etterpå. Det må derfor legges en viss vekt på disse problemene når resultatet av en undersøkelse skal vurderes.

De faktorer som bidrar til slike feil, kan være en eller flere av følgende:

i) Målingen selv
ii) Måleren
iii) Hva som måles
iv) Den som måles

i) Målingen selv

Denne feilkilden kan best illustreres ved et eksempel. En opinionsundersøkelse skal finne ut hvilke meninger folk har om et spørsmål. Hvis IO får seg forelagt et spørsmål som innebærer en problemstilling han aldri har tenkt over, kan IO tvinges til å ta et standpunkt han kanskje ikke hadde på forhånd.

Et annet og kanskje mer illustrerende eksempel for denne feiltypen, finner vi i undersøkelser hvor det kreves en viss innsats av IO ved f.eks. å føre husholdningsregnskap. Slik aktivitet er vel for de fleste IO et avvik fra deres vanlige rutine, noe som åpenbart kan medføre at rutinen, og dermed måleresultatet, blir noe annet enn det tilsiktede.

ii) Måleren

Intervjueren kan være årsak til mange feil, ikke bare «fusk», men også en rekke andre feiltyper som han bevisst eller ubevisst kan gjøre seg skyldig i. Noen viktige punkter i denne sammenheng angis for oversiktens skyld:

— Fusk og motivene for dette
— Måten spørsmålet uttales på
— Kontroll-effekt (Intervjueren er f.eks. til stede, fremmed osv.)
— «Biased viewpoint-effect»[1]
— Inklusjon av uvedkommende ting i spørsmålet (eksplisitt eller implisitt)
— Eksklusjon av relevante ting i spørsmålet («tretthets-slurv»)
— Konversasjon (kan påvirke IO)
— Intervjuerens utseende, kjønn, alder osv. i forhold til IO

iii) Hva som måles

Feil knyttet til hva som måles kan bero på blant annet følgende faktorer som her bare angis stikkordmessig:

— Spørsmål som krever gjenkallelse fra hukommelsen
— Prestisje-faktorer som årsak til bevisst forvrengning av svarene
— Tabu-faktorer som er knyttet til en rekke felter i det sosiale liv (sex, «privatlivets indre anliggender», kriminalitet, osv.)

iv) Den som måles

Den som intervjues kan i enda større grad enn *hva* som måles forårsake usikre resultater. «Svars-variabilitet» er knyttet til omtrent alle størrelser. Selv såkalte «objektive data» som f.eks. alder, sivilstand, yrke, utdannelse osv. kan i enkelte tilfelle «variere» i løpet av utrolig kort tid. Svarsvariabilitet trenger ikke å bero på bevisst løgn, som regel blir de faktiske forhold forvrengt ubevisst, avhengig av en rekke omstendigheter både ved IO selv og ved omgivelsene og situasjonen. Presisjonsgraden i spørsmålet vil også ha betydning i denne sammenhengen.

3. Databehandlingsfeil

a) Akseptering og forkasting av individuelle data

Feilaktigheter eller slurv-feil forekommer av og til i råmaterialet fra en intervju-undersøkelse. Som regel avgjøres det skjønnsmessig hvorvidt en feil er av en slik art at den kan rettes, eller om opplysningen må forkastes. Svakheten ved dette er åpenbar. Det menneskelige skjønn er

---

Note: [1] «Biased viewpoint-effect» er den effekt som skyldes at intervjueren innehar en rolle i samhandlingen med IO. Intervjueren og IO ser begge på situasjonen ut fra *sine* respektive roller.

variabelt, og det kan gi systematiske utslag i en eller annen retning når det brukes som sorteringsmekanisme. Relativt sett er feil av denne typen bare av små dimensjoner.

b) Korrigering av individuelle data

De opplysninger (feil) soom ikke forkastes, blir korrigert. Som regel dreier det seg her om «åpenbare» feil som stort sett blir riktige etter korrigeringen. Imidlertid kan det enkelte ganger reises tvil når det er foretatt korrigeringer av såkalte «inkonsekvente» eller motstridende opplysninger. Det kan nemlig være vanskelig å avgjøre hvilken av to motstridende opplysninger som er den riktige, og forekomsten av uriktige «korrigeringer» kan derfor ikke utelukkes helt.

Hvis det er stor tvil om hva som er de riktige svar, blir som regel IO konsultert på nytt — dette gjelder også for punkt a) ovenfor.

X. Estimatorer og deres varians

Sampling-modellen som er beskrevet foran, kan i korte trekk sies å være et selv-veiende, stratifisert to-trinns-utvalg. Hvordan estimering foretas under denne modellen, vil her bare bli skissert i hovedtrekkene. De som har interesse av en dypere teoretisk innføring, henvises til standardverker om sampling-teori.[1]

1. Definisjoner

Noen av de størrelsene som forekommer i det følgende, er definert i avsnitt VII. De vil likevel bli repetert og sammenholdt med andre størrelser i en oversiktstabell. Det gjøres oppmerksom på skillet mellom størrelser som gjelder for (total-)populasjonen (store bokstaver), og dem som gjelder for utvalget (små bokstaver).

La generelt den variabel som observeres betegnes med X.

Note: [1] Spesielt kan nevnes:

(1) Hansen, Hurwitz and Madow: «Sample Survey Methods and Theory», Vol. I—II. John Wiley & Sons, Inc., New York 1953.

(2) Ernst Lykke Jensen: «Repræsentative undersøgelsers teori og metode», bind II, København 1960 (stensil).

(3) William G. Cochran: «Sampling Techniques». 2nd edition. John Wiley & Sons, Inc., New York, London.

	Populasjonen	Utvalget
X-verdi for sue (i, j, k)	X_{ijk}	x_{ijk}
antall sue i alt	N	n
antall sue i det i-te stratum	N_i	n_i
antall sue i pue (i, j)	N_{ij}	n_{ij}
antall pue i det i-te stratum	M_i	m_i
stratum-indeks	$i = 1, 2, \ldots, s$	$i = 1, 2, \ldots, s$
pue-indeks	$j = 1, 2, \ldots, M_i$	$j = 1, 2, \ldots, m_i$
sue-indeks	$k = 1, 2, \ldots, N_{ij}$	$k = 1, 2, \ldots, n_{ij}$
X-sum for pue (i, j)	$X_{ij\cdot}$	$x_{ij\cdot}$
X-sum for i-te stratum	$X_{i\cdot\cdot}$	$x_{i\cdot\cdot}$
X-sum totalt	$X\ldots$	$x\ldots$
X-gjennomsn. pr. sue i pue (i, j)	$\overline{X}_{ij}$	$\bar{x}_{ij}$
X-gjennomsn. pr. pue, stratum i	$\overline{X}_i$	$\bar{x}_i$
X-gjennomsn. pr. sue i alt	$\overline{X}$	$\bar{x}$
gjennomsnittlig antall sue pr. pue i det i-te stratum	$\overline{N}_i$	$\overline{n}_i$

De forskjellige X-summene, X-gjennomsnittene, $\overline{N}_i$ og $\overline{n}_i$ er definert slik:

Populasjonen	Utvalget
$X_{ij\cdot} = \sum_{k=1}^{N_{ij}} X_{ijk}$	$x_{ij\cdot} = \sum_{k=1}^{n_{ij}} x_{ijk}$
$X_{i\cdot\cdot} = \sum_{j=1}^{M_i} X_{ij\cdot}$	$x_{i\cdot\cdot} = \sum_{j=1}^{m_i} x_{ij\cdot}$
$X\ldots = \sum_{i=1}^{s} X_{i\cdot\cdot}$	$x\ldots = \sum_{i=1}^{s} x_{i\cdot\cdot}$
$\overline{X}_{ij} = \frac{X_{ij\cdot}}{N_{ij}}$	$\bar{x}_{ij} = \frac{x_{ij\cdot}}{n_{ij}}$
$\overline{X}_i = \frac{X_{i\cdot\cdot}}{M_i}$	$\bar{x}_i = \frac{x_{i\cdot\cdot}}{m_i}$
$\overline{X} = \frac{X\ldots}{N}$	$\bar{x} = \frac{x\ldots}{n}$
$\overline{N}_i = \frac{N_i}{M_i}$	$\overline{n}_i = \frac{n_i}{m_i}$

2. Estimering

Siktepunktet i våre undersøkelser er som regel å estimere visse kjennetegns fordeling i populasjonen, gjennomsnittlige størrelse pr. sekundære utvalgsenhet og totaler for populasjonen som helhet eller deler av den. For ikke å skape uklarhet i begrepene, presiseres det at «kjennetegn» betyr:

i) *diskrete* variable: Et kjennetegn er her en av de kategoriene variabelen kan klassifiseres i.

ii) *kontinuerlige* variable: Slike variable kan oppdeles i intervaller. Med et kjennetegn menes da et slikt intervall.

iii) Et kjennetegn kan også være en kombinasjon av to eller flere kjennetegn definert ved i) eller ii).

a) Estimering av frekvenser og gjennomsnitt

Når oppgaven er å estimere med hvilken hyppighet et kjennetegn — f.eks. A — opptrer, er det naturlig å definere X_{ijk} som en binær variabel («dummy»-variabel). Da settes X_{ijk} lik:

$$X_{ijk} = \begin{cases} 1 \text{ hvis sue (i, j, k) har A} \\ 0 \text{ hvis sue (i, j, k) ikke har A} \end{cases}$$

La generelt s være antall strata populasjonen er inndelt i. «Populasjonen» kan gjerne være en *del* av total-populasjonen, f.eks. ett av våre geografiske strata («handelsfelt»), eller det kan være total-populasjonen selv. Betraktningene blir de samme uansett hvilken populasjon det gjelder. Det er bare s som influeres av dette.

Antall sue som har kjennetegnet A, er åpenbart $X\ldots$, og siden populasjonens størrelse er N, må frekvensen for A være:

$$\overline{X} = \frac{X\ldots}{N}$$

Verbalt uttrykt kan en si at $100 \cdot \overline{X}$ prosent av populasjonen har kjennetegnet A.

Hvis X er en kontinuerlig variabel, og vi er interessert i gjennomsnittet pr. sekundær-enhet av X (f.eks. gjennomsnittlig årsinntekt pr. sue), så er også dette gjennomsnittet lik $\overline{X}$.

Likegyldig hvilken type variabel $\overline{X}$ er, så er altså frekvensen eller gjennomsnittet lik $\overline{X}$, og det er denne størrelsen vi skal estimere.

Siden vårt utvalg er selvveiende (og «populasjonen» alltid må omfatte bare *hele* strata), vil følgende uttrykk være en konsistent (asymptotisk forventningsrett) estimator for $\overline{X}$:

$$\hat{g} = \bar{x}$$

b) Estimering av populasjons-totaler

Med «populasjons-total» menes summen av alle X-verdier over hele populasjonen (slik denne er definert i punkt 1 foran). Denne summen er åpenbart lik $X\ldots$.

En forventningsrett estimator for $X\ldots$ er:

$$\hat{t} = \frac{1}{f} x\ldots$$

c) Varians og estimering av varians

Variansene til $\hat{g}$ og $\hat{t}$ gis her uten bevis:

$$\operatorname{var} \hat{g} = \frac{1}{N^2} \sum_{i=1}^{s} \frac{(1 - f_{1i})}{f_{1i}} M_i \, S_{1i}^2 + \frac{1}{N^2} \sum_{i=1}^{s} \frac{1 - f_{2i}}{j} \sum_{j=1}^{M_i} N_{ij} \, S_{2ij}^2$$

og $\operatorname{var} \hat{t} = N^2 \operatorname{var} \hat{g}$

hvor:

$$S_{1i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (X_{ij.} - \overline{X}_i)^2 \text{ og}$$

$$S_{2ij}^2 = \frac{1}{N_{ij} - 1} \sum_{k=1}^{N_{ij}} (X_{ijk} - \overline{X}_{ij})^2$$

Som estimator for var $\hat{g}$ kan vi bruke:

$$s^2(\hat{g}) = \frac{1}{n^2} \sum_{i=1}^{s} (1 - f_{1i}) \, m_i \, s_{1i}^2 + \frac{f}{n^2} \sum_{i=1}^{s} (1 - f_{2i}) \, m_i \sum_{j=1}^{m_i} N_{ij} \, s_{2ij}^2$$

hvor:

$$s_{1i}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (x_{ij.} - \bar{x}_i)^2 \text{ og}$$

$$s_{2ij}^2 = \frac{1}{n_{ij} - 1} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij})^2$$

var $\hat{t}$ estimeres ved:

$$s^2(\hat{t}) = N^2 \, s^2(\hat{g})$$

Vanligvis er f meget liten (mindre enn 1 %). Dette medfører at også $\frac{f}{n^2}$ blir svært liten. Ved en vanlig undersøkelse hvor husholdning er sue, vil $f \approx \frac{1}{400}$ og $n \approx 3000$. Dette innebærer at $\frac{f}{n^2} = \frac{1}{3{,}6}\ 10^{-9} \approx 0$. Altså kan vi uten å endre så mye på s^2 ($\hat{g}$) sløyfe det siste leddet i summen, leddet hvor $\frac{f}{n^2}$ inngår som faktor. Sløyfing av dette leddet

$$-\frac{f}{n^2} \sum_{i=1}^{s} (1 - f_{2i})\, m_i \sum_{j=1}^{m_i} N_{ij}\, s^2_{2ij} -$$

forutsetter at summen som $\frac{f}{n^2}$ skal multipliseres med, er begrenset — hvilket den i praksis vil være.

Som en god tilnærmelse kan vi derfor estimere var $\hat{g}$ med:

$$s^2 = \frac{1}{n^2} \sum_{i=1}^{s} (1 - f_{1i})\, m_i\, s^2_{1i}$$

og tilsvarende estimeres var $\hat{t}$ med:

$$s_t^2 = N^2\, s_g^2$$

For at s_g^2 og s_t^2 skal kunne brukes, må $m_i \geqq 2$ for alle i. Vår utvalgsplan er slik at $m_i = 2$ for alle strata utenfor Oslo, Bergen og Trondheim, hvor $m_i = 1$. Dette problemet er løst på følgende måte:

Oslo:

Vi har i alt 7 strata. To og to strata slås sammen til såkalte «collapsed» strata. Derved fås 3 «strata» à 2 utvalgsområder. Ett av de opprinnelige strataene blir imidlertid stående igjen. Men idet Oslo har to intervjuere pr. utvalgsområde, kan materialet som hver av de to intervjuerne har samlet inn, betraktes som stammende fra hvert sitt utvalgsområde. Dermed har vi i alt 4 «strata» à 2 utvalgsområder.

Bergen og Trondheim:

Hver av disse byene har 3 strata med ett utvalgsområde i hver. Hvis et stratum fra hver by slås sammen, får vi 3 nye «collapsed» strata à 2 utvalgsområder.

Dermed er $m_i = 2$ for alle i. Varians-estimatoren blir da noe enklere:

$$s_g^2 = \frac{2}{n^2} \sum_{i=1}^{s} \left(1 - \frac{2}{M_i}\right) s_{1i}^2$$

hvor s^2_{1i} nå blir:

$$s^2_{i1} = (x_{i1.} - \bar{x}_i)^2 + (x_{i2.} - \bar{x}_i)^2$$

XI. Frafall

Ved statistiske undersøkelser som baseres på intervju av (helt eller delvis) spesifiserte personer, oppstår det alltid situasjoner hvor intervju ikke kommer i stand. Det er dette vi kaller *frafall* eller mer presist *ikke-respons fra på forhånd utvalgte IO.*

Årsakene til frafall er mange, men vi kan skille mellom to ulike typer av frafall:

— IO ikke til stede.

— Intervju avslås eller kommer ikke i stand av andre grunner.

Når IO ikke er til stede, kan dette skyldes tre ting: IO er *flyttet,* IO er *midlertidig fraværende* eller IO er (nylig) *avgått ved døden.*

Avslag på intervju kan ha varierende årsaker. Enkelte IO kan være konsekvente nektere som aldri lar seg intervjue, mens andre nekter i visse intervjutyper. Det kan hende at IO blir oppsøkt på et tidspunkt hvor det passer svært dårlig å bli intervjuet, mens noen IO kan være i en slik psykisk og/eller somatisk forfatning at intervju ikke lar seg gjennomføre.

Det er alltid mulig å minske frafallet, men dette kan kreve både mye tid og store omkostninger. Et midlertidig fraværende IO kan f.eks. oppsøkes på dets nåværende oppholdssted. *Flere* besøk hjelper som regel.

Med nektere er det verre, som regel er problemene her av psykologisk art.

En del spesielle undersøkelser som har vært foretatt i andre land omkring frafallsproblemet, har f.eks. vist at intervjuere med lang øvelse har et frafall som bare er på ca. 10 prosent av hva intervjuere med mindre øvelse har.

De tiltak som kan settes i verk for å redusere frafallet, vil bli behandlet i et senere avsnitt. I det følgende avsnitt vil en drøfte hvilken betydning frafallet kan ha for resultatene. Framstillingen bygger delvis på eksempler hentet fra undersøkelser som Statistisk Sentralbyrå har foretatt.

1. Teoretiske synspunkter

Anta at vi har en populasjon på N enheter (symbolene i dette avsnittet avviker en del fra dem som er brukt foran). Vi er interessert i å bestemme hvor stor prosent P av denne populasjonen som har et eller annet kjennetegn. Da er det altså (100 — P) prosent som ikke har vedkommende kjennetegn. Vi vil anslå P med en nøyaktighet som ikke avviker mer enn Q_0 prosent fra den «sanne» verdi. Anta videre at vi vil undersøke samtlige N enheter, men at vi etter å ha foretatt undersøkelsen har et *frafall på Q_f prosent av populasjonen.* For enkelhets skyld kan vi også anta at den eneste feilkilden som har virket inn på bestemmelsen av P, er frafallet. La:

N(f) = antall enheter som faller fra
N(m) = N-N(f) = antall enheter som medvirker
P(m) = den prosentvise andel av N(m) som har kjennetegnet vi er interessert i
P(f) = den prosentvise andel av N(f) som har kjennetegnet.

Resultatet av vår undersøkelse er at vi får bestemt N(f), N(m) og P(m), men ikke P(f). Spørsmålet er nå om P(m) er et godt nok anslag for P. Dette avhenger selvsagt av hvor stor P(f) er, og det vet vi ikke. Hva en generelt ikke bør gjøre, er å anta at P(f) = P(m), dvs. å hevde at frafallet har vært «tilfeldig» eller at «frafallet ikke har virket på undersøkelsesmaterialets representativitet». Det mest forsvarlige i denne situasjonen er å anslå ekstremalverdiene for P. Dette gjør vi ved å anta henholdsvis P(f) = 0 prosent (nedre grense for P; P(min)) og P(f) = 100 prosent (øvre grense for P; P(max)).

Vi får derved anslått et intervall, [P(min), P(max)], hvor P må ligge. Den eksakte verdi av P er gitt ved følgende uttrykk:

$$P = P(m) \cdot \left(1 - \frac{Q_f}{100}\right) + P(f) \cdot \frac{Q_f}{100}$$

P blir av denne formelen angitt som et prosenttall (P = P %). Settes P(f) = 0 % fås:

$$P(min) = P(m) \cdot \left(1 - \frac{Q_f}{100}\right)$$

og med P(f) = 100 % fås:

$$P(max) = P(m)\left(1 - \frac{Q_f}{100}\right) + Q_f$$

Herav ser vi at

$$P(max) - P(min) = Q_f$$

Det vil si at den maksimale prosentvise feil som skyldes frafallet, ikke overstiger frafallsprosenten.

Et tall-eksempel vil illustrere disse forholdene: Vi tabulerer P(max), P(min) og P(max) — P(min) = Q_f for ulike verdier av P(m) og forutsetter Q_f = 10 %.

Tabell 1. Ekstreme virkninger av frafall på estimatet (P(m)).

P(m)	P(max)	P(min)	P(max)—P(min)
0	10	0	10
20	28	18	10
50	55	45	10
80	82	72	10
100	100	90	10

Tallene er prosenter.

Et studium av ovenstående tabell forteller umiddelbart at frafallet kan *forskyve* resultatet til dels betydelig. Hvis vi fastsetter at Q_0 (tillatt feilmargin) skal være f.eks. 5 prosent i hver retning omkring den «sanne» verdi, P (som vi har estimert med P(m)), så ser vi at dette kravet bare er oppfylt når P(m) = 50 prosent.

Når Q_f vokser, dvs. frafallet blir større, så vokser også intervallet [P(min), P(max)], og anslaget P(m) for P blir derfor mer usikkert.

Det som er nevnt ovenfor, er de *mest ekstreme* konsekvensene av frafall. I praksis vil en til en viss grad kunne *vurdere* hvor stor P(f) kan være. Generelt kan en si at jo større kunnskap vi har om frafallet, jo bedre grunnlag har vi for å vurdere hvor god den estimerte verdi (P(m)) av P er.

Det eksisterer flere metoder for å analysere frafallet. En metode er å sammenlikne sammensetningen av den gruppen som har medvirket i undersøkelsen, med sammensetningen av den gruppen som ikke har medvirket — frafallsgruppen. En slik sammenlikning er likevel bare av begrenset verdi idet en ikke kan trekke noen statistisk holdbare konklusjoner av den. Hvis det f.eks. skulle vise seg at de to gruppene er omtrent likt sammensatt («representativt frafall»), så kan en *ikke konkludere* med at P(m) $\approx$ P(f). En bør også være varsom med å si at det er *større sannsynlighet* for at en slik konklusjon er riktig.

Hvis det derimot skulle vise seg at det er *betydelige ulikheter* i sammensetningen av de to gruppene, så kan vi med stor rimelighet ta dette som et indisium på at frafallet har virket forstyrrende på resultatet. Av det som her er sagt om sammenlikning av de to gruppene, kan vi altså trekke den slutning at *sammenlikninger av denne art kan «betvile» en hypotese (P(m) = P), men ikke bekrefte den.* Som eksempel på en slik «struktursammenlikning», kan vi ta «Boligundersøkelsen, oktober 1967».[1] Nedenstående tabell viser den prosentvise fordeling av forskjellige variable i gruppen som medvirket, og i gruppen som ikke medvirket i undersøkelsen (frafallet).[2] Enhet er husholdning.

Tabell 2. Struktur-sammenlikning av medvirker- og frafallsgruppen

Variabel	Medv. prosent	Frafall prosent	Variabel	Medv. prosent	Frafall prosent
Type strøk:			*Geografiske distrikter:*		
Tettbygd	52	74	Østre handelsfelt (Ekskl. Oslo)	34	31
Spredtbygd	48	26	Oslo	14	19
			Søndre og Vestre h.felt	28	28
Husholdningstype:			Midtre h.felt	15	10
En-fam.hush.			Nordre handelsfelt ..	9	12
u/ugifte barn ..	22	19			
» m/1 » » ..	21	16	*Hovedpersonens alder:*		
» m/2 » » ..	19	10	Under 25 år	3	5
» m/3 » » ..	11	4	25—34 år	14	10
Flerfam. husholdning	2	0	35—44 år	18	15
» m/4 » » ..	6	3	45—59 år	32	30
Enslige	16	43	60—69 år	19	23
Øvrige husholdninger	3	5	70—79 år	11	12
			80 år og over	3	5

Som vi ser av tabellen, kan avvikene være betydelige for enkelte variable. Når det gjelder «type strøk», så er prosentdifferansen hele 22 prosent. Hvor stor virkning dette avviket kan ha for de tallene som skal publiseres, vil avhenge av hvor sterk sammenhengen er mellom den størrelsen som det publiseres tall for, og «type strøk». Jo sterkere sammenheng, desto mer usikre blir estimatene. Det samme gjelder selvsagt for andre variable.

Noter: [1] Publisert som rapport fra Kontoret for intervjuundersøkelser, nr. 3.

[2] I denne undersøkelsen er det til en viss grad foretatt *erstatning* for frafalte. Tallene som presenteres her er de endelige, altså *etter* at erstatning er foretatt. (Se ellers avsnitt om erstatninger (XI 3)).

I Boligundersøkelsen er det i bestemte tilfelle foretatt *erstatning* for å minske frafallet. Konsekvensene av en slik framgangsmåte kan variere, men i prinsippet er det følgende som skjer: Når en enhet faller fra, erstattes denne med en ny (reserve). Hvis vi foretar erstatning N(e) ganger, så øker *totalutvalget* fra N til N+N(e) enheter. Det en *ikke* oppnår, er å få i stand intervju med dem som faller fra. Hvis vi oppnår intervju med samtlige N(e) enheter, så har vi altså i alt undersøkt N(m) + N(e) enheter. Men N(f) er like stor som før. Konklusjonen på dette blir derfor at vi ved å foreta erstatning *ikke oppnår noen kunnskap om frafallet.* Derimot øker vi antall enheter som undersøkes, hvilket betyr at variansen på estimatene blir mindre. Det kan inntreffe at en ved å foreta erstatning *øker* usikkerheten som skyldes frafallet. Anta f.eks. på grunnlag av ovenstående tabell at det store frafallet av *enslige* er blitt kompensert ved erstatning av flerpersonhusholdninger. Det er videre nokså rimelig å anta at «enslige husholdninger» i mange henseender avviker en del fra flerperson-husholdninger. Dette kan medføre — idet forholdet mellom enslige og andre er blitt *dobbelt* forskjøvet (ved frafall og erstatning) i forhold til det opprinnelige utvalg — at de estimatorene som vi bruker («selv-veiende») blir dårlige. Paradoksalt nok kan det faktisk inntreffe at *estimatene blir «dårligere»* samtidig som *variansen blir mindre.* Forklaringen er selvsagt at feilen på grunn av frafallet er blitt forstørret; det utvalg som undersøkes, er fremdeles et sannsynlighetsutvalg hvor sannsynligheten for at en vilkårlig valgt enhet skal komme i utvalget er kjent, men sannsynligheten er *ikke lenger lik* for alle enheter i populasjonen. En slik forandring av de opprinnelige forutsetninger kan få konsekvenser for de estimatorene som brukes. Dersom frafallet er spesielt stort innen en eller flere grupper, bør det overveies å erstatte de «selv-veiende» estimatorene med andre som er laget på grunnlag av *faktorer med ulike vekter.* En kan da konstruere konsistente eller forventningsrette estimatorer, men det kan hende at variansen på disse blir større enn på de opprinnelige.

Hvis en foretar en *veiing* av estimatorene, så betyr dette at det er innført en ny stratifikasjon. Om vi f.eks. skiller mellom *enslige* og *ikke-enslige* husholdninger i estimatorene, blir stratifikasjonen å regne etter disse to kjennetegn. Det bør understrekes at en slik «ekstra-stratifikasjon» som regel fører til et stort merarbeid; en må nemlig estimere (ved fullstendig eller delvis *telling)* antall enslige og ikke-enslige husholdninger i hvert av de opprinnelige strata.

2. Reduksjon av frafallet

Det har tidligere vært nevnt at frafallet kan reduseres ved blant annet bedre ajourføring av registeret som brukes ved trekkingen. Ideelt sett bør utvalget trekkes samme dag som intervjuingen finner sted, og trekkingen må foretas på grunnlag av lister som er korrekte for denne dagen. I praksis foretas trekkingen på et tidspunkt som ligger noen måneder forut for intervjuingen, og de listene (registeret) som brukes, er ikke ajourførte. Dette medfører at de forandringer ved befolkningen (flytting, dødsfall, ekteskap, fødsler osv.) som har funnet sted etter det tidspunkt da registeret kunne regnes som korrekt, ikke blir tatt til følge uten spesielle anstrengelser. Og det er disse tiltakene som ved korreksjon av utvalget tar sikte på å minske frafallet.

Med utgangspunkt i den ideelle situasjonen som er beskrevet ovenfor, er det logisk hensiktsmessig å dele behandlingen av frafallsreduksjon til to tidsintervaller: T_1 = fra det tidspunkt da registeret var korrekt til tidspunktet for trekking

T_2 = tiden fra trekking til intervju.

Når det gjelder T_1, kan problemet som et ufullstendig register representerer, løses ved at det kontinuerlig innløper oppgaver fra folkeregistrene om endringer i befolkningen. Skal en slik ordning bli effektiv, må for det første folk straks melde fra om eventuelle forandringer, og for det andre må folkeregistrene umiddelbart sende disse opplysningene til registeret i Byrået.

Den vesentligste hindringen for fortgang i denne prosessen er nok det faktum at folk unnlater eller venter lenge med å gi folkeregisteret beskjed om forandringer som har funnet sted. I hovedsaken skyldes dette at mange simpelthen ikke er oppmerksomme på at de plikter å melde fra om enkelte forandringer.

Registeret ville muligens bli noe mer korrekt om folkeregistrene *straks* meldte fra til Byrået når en forandring hadde funnet sted i kommunen. Men i relasjon til det førstnevnte problemet er dette tidstapet som regel av mindre betydning.

Når det gjelder T_2, så er feilmengden i utvalget direkte avhengig av lengden på T_2. Jo tidligere vi trekker et utvalg før feltarbeidet, desto flere feil vil oppstå. Utvalget må imidlertid trekkes en tid forut for det tidspunkt da intervjuingen finner sted, og de tiltakene som settes i verk for å redusere feilmengden, kan oppsummeres i fire punkter:

i) Når utvalget til en undersøkelse er trukket, lages det spesielle IO-lister for hvert utvalgsområde.

ii) Listene sendes til de respektive folkeregistre til kontroll og eventuell retting. Familier som er flyttet, strykes av listen, og familier som eventuelt er flyttet inn i deres sted, føres opp.
iii) Når listene kommer i retur, føres rettelsene inn i vårt register, og nye lister lages for intervjuerne.
iv) De feil som på denne måten *ikke* blir rettet, registreres og rettes av intervjuerne etter regler som er fastlagt av Byrået.

3. Erstatninger

I enkelte tilfelle kan det være tvil om hva som skal regnes som frafall. I egentlig forstand er frafall å regne som «ikke oppnådd intervju» med IO som er trukket ut. Etter det som er nevnt i det foregående, er det klart at de lister som IO trekkes fra, ikke stemmer overens med virkeligheten på alle punkter. Ideelt skulle jo utvalget trekkes fra den *virkelige* populasjonen, men dette er altså ikke mulig. Hvis vi derfor skal regne frafall etter den egentlige definisjon, betyr dette — logisk sett — at vi definerer populasjonen som vårt register og ikke som virkeligheten. Vi sier altså at det er feil ved de virkelige forholdene og ikke ved vårt register — hvilket åpenbart ikke er i overensstemmelse med sunn fornuft.

For at ikke denne logikken skal få råde, er det tatt i bruk erstatninger for visse typer «frafall». I enkelte tilfelle kan intervjueren få oppgitt et nytt IO til erstatning for et IO som ikke kan intervjues.

Et *uvillig* IO erstattes ikke. Det er klart *frafall.*

Erstatning for et tildelt IO må alltid gis av hovedkontoret, intervjueren kan aldri velge seg en reserve selv.

Erstatningsgrunnlag vil vanligvis være:

— Tomme leiligheter — leiligheten er *ubebodd* på det tidspunkt undersøkelsen foregår
— Nedbrente — nedrevne hus
— For husholdningsundersøkelser: medlem av en felleshusholdning (anstalt)
— For personundersøkelser: psykisk syke personer (åndssvake)
— Langvarige sykehusopphold (pleiepasienter) hvor vi har grunn til å tro at vedkommende ikke kommer tilbake til sin bolig (regnes som tomme leiligheter)
— Sjøfolk i langfart (for husholdningsundersøkelser, med tom leilighet hjemme)
— Langvarige utenlandsopphold

Graden av opplysninger og lengden av antatt fravær kan være avgjørende.

Forhold som *ikke* gir grunnlag for erstatninger:

— Sykdom i hjemmet
— Vanlige sykehusopphold
— Kortere feriefravær
— Ikke truffet hjemme
— Liten tid

Etter det som ble nevnt om erstatninger i avsnitt XI 1, kan det være verdt å undersøke erstatningene i noen av de undersøkelsene som er foretatt.

Det mest interessante er å sammenlikne *erstatningsgruppen* (dem som medvirker) med *erstattet-gruppen* (dem som er gått ut av utvalget). Dermed kan eventuelle strukturforandringer fra det opprinnelige utvalg vurderes.

Vi vil her ta for oss to undersøkelser: Boligundersøkelsen 1967 og Ferieundersøkelsen 1967/68.

I de nedenstående tabellene er frekvensfordelingen for hovedpersonens alder og husholdningsstørrelsen (antall medlemmer i husholdningen) angitt for de to gruppene.

Boligundersøkelsen 1967: Totalt antall enheter som er erstattet, er 120 husholdninger.

Tabell 3. Aldersfordeling for erstattet- og erstatningsgruppene

Hovedpersonens alder	Erstattet-gruppe	Erstatningsgruppe
15—29 år	15	10
30—39 »	13	20
40—49 »	8	19
50—59 »	15	19
60—69 »	20	19
70 år og over	29	13
I alt	100	100

Tabell 4. Husholdningsstørrelse for erstattet- og erstatningsgruppene

Husholdningsstørrelse	Erstattet-gruppe	Erstatningsgruppe
1	60	20
2	20	24
3	9	19
4	8	19
5 og over	3	18
I alt	100	100

I tabell 3 ser vi at nokså mange over 70 år er blitt erstattet med husholdninger som er noe «yngre». Dette er også hva vi kan vente ved å bruke en erstatningsmetode av vår type. Frafallet (eller avgangen) av eldre er av forståelige grunner nokså stort, og siden hovedmengden av *tilfeldig* utvalgte erstatninger vil ligge mellom 30—69 år, må vi få en skjev representasjon. I de øvrige aldersgruppene er forskjellene ubetydelige.

I Boligundersøkelsen ble det undersøkt i alt 2 815 husholdninger. Erstatningsgruppen utgjør 120 enheter, dvs. ca. 4 prosent av totalmassen. Hvis erstatninger ikke hadde vært foretatt, måtte vi regne nesten alle de 120 erstattede husholdninger som rent frafall, og det er som tidligere nevnt, ikke helt realistisk. Erstatning kan også oppfattes som en korreksjon av utvalget på grunn av uoverensstemmelsen mellom register og virkelighet. Men som korrigeringsmetode er erstatningsreglene neppe de beste. Dette framgår tydelig av tabell 4 hvor vi ser at erstatningene går markant i favør av flerpersonhusholdninger.

Ferieundersøkelsen 1967/68: I alt 89 erstatninger. Enhet: Husholdning.

Tabell 5. Aldersfordeling for erstattet- og erstatningsgruppene

Hovedpersonens alder	Erstattet-gruppe	Erstatningsgruppe
15—29 år	15	6
30—39 »	14	15
40—49 »	14	25
50—59 »	17	19
60—69 »	14	21
70 år og over	26	14
I alt	100	100

Tabell 6. Husholdningsstørrelse for erstattet- og erstatningsgruppene

Husholdningsstørrelse	Erstattet-gruppe	Erstatningsgruppe
1	61	15
2	18	24
3	8	15
4	11	22
5 og flere	2	24
I alt	100	100

Hvis vi sammenlikner tabell 5—6 med tabell 3—4, oppdager vi en slående likhet. De kommentarene som ble gitt til Boligundersøkelsen, kan derfor også gis til Ferieundersøkelsen.

Det ser ut til at den gruppen som må erstattes ved husholdningsundersøkelser, har omtrent samme struktur ved hver undersøkelse. Dette går nokså tydelig fram av tabell 7 og 8 som viser simultan-fordelingen for hovedpersonens alder og husholdningens størrelse for erstattet-gruppene.

Tabell 7. Erstattet-gruppens fordeling med hensyn på husholdningsstørrelse og hovedpersonens alder. Boligundersøkelsen

Husholdnings-størrelse	Hovedpersonens alder						
	15—29 år	30—39 år	40—49 år	50—59 år	60—69 år	70 år og over	I alt
1	10	7	2	7	14	20	60
2	2	1	4	4	1	8	20
3	2	1	0	2	3	1	9
4	1	3	1	2	1	0	8
5 og over	0	1	1	0	1	0	3
I alt	15	13	8	15	20	29	100

Tabell 8. Erstattet-gruppens fordeling med hensyn på husholdningsstørrelse og hovedpersonens alder. Ferieundersøkelsen

Husholdnings-størrelse	Hovedpersonens alder						I alt
	15—29 år	30—39 år	40—49 år	50—59 år	60—69 år	70 år og over	
1	12	4	9	10	5	21	61
2	1	3	1	2	7	4	18
3	1	1	3	2	1	0	8
4	1	5	1	2	1	1	11
5 og flere	0	1	0	1	0	0	2
I alt	15	14	14	17	14	26	100

To undersøkelser er selvsagt i minste laget til å si noe allment om erstattet-gruppens struktur. Men siden overensstemmelsen er så god i de to tabellene, kan vi med støtte i dem og det vi generelt kan forvente om frafall av denne art, si at det frafallet som erstattes, virker sterkest reduserende på gruppen av eldre enslige. I noen mindre grad reduseres også gruppen av unge (15—29 år) enslige. Størsteparten av reservene (erstatningsgruppen) som settes inn i utvalget, tilhører ikke disse to gruppene, og skjevheten blir derfor forstørret ved erstatning. Hvor store konsekvenser dette har for utvalgets selv-veiings-egenskap, er avhengig av antall erstatninger. For bolig- og ferieundersøkelsen utgjør erstatningene 3—4 prosent av de medvirkende husholdningene. Dette kan resultere i strukturforskyvninger på mellom 0—4 prosent for en eller flere grupper sammenliknet med de opprinnelige utvalg. I praksis kan vi imidlertid regne med at forskyvningene ikke er større enn ca. 2 prosent for noen grupper, og dette er såpass lite at estimerte størrelser bare i liten grad blir påvirket av forskyvningene. Vi kan likevel ikke overse dette feilmomentet, særlig på bakgrunn av at erstatningene reduserer variansen og i enda større grad frafallsfeilen.

XII. Sluttord

Utvalgsplanen som her er beskrevet, har vært brukt til gjennomføring av undersøkelser med varierende sekundære utvalgsenheter. En har oppnådd en viss erfaring med å løse de praktiske problemene som alltid oppstår, særlig i forbindelse med populasjonsregisteret. Imidlertid kan mange problemer løses på en mer tilfredsstillende måte hvis det kan

bygges opp et register (f.eks. på grunnlag av Folketellingen 1970) hvor en tar hensyn til de spesielle faktorer som kan skape problemer ved stikkprøver. Et slikt register burde også gjøre det mulig å ta i bruk maskinell utvelging istedenfor som nå, manuell.

Videre foreligger det grundige oppgaver over de kostnader som er forbundet med utvalgsplanen. Av særlig betydning er oppgavene for utgiftene ved feltarbeidet. Slike erfaringer bør utnyttes til å gjennomføre såkalte kostnadsvarians-analyser med henblikk på å finne et optimalt forhold mellom stikkprøvens fordeling og kostnadene. Alternative utvalgsplaner bør da også vurderes.

Vedlegg

Appendix

La f, f_i, f_{1i} og f_{2i} (f_{2ij}) ha samme betydning som definert i kapittel VII. For øvrig gjør vi nå en liten forandring i terminologien idet vi innfører en stokastisk variabel, Z, som defineres ved:

Z_{ij} = totalt antall sue i pue (i, j) *i utvalget.*

Z_{ij} er da en stokastisk variabel fordi pue'ne velges *tilfeldig*. Størrelsen på Z_{ij} vil derfor avhenge av hvilken pue (dvs. hvilken (i, j) eller bare j) som blir valgt ut.

Hvis vi på Z_{ij} lar indeks j markere hvilken trekning i stratum i Z_{ij} observeres, så vil med symbolene fra avsnitt VII observasjonsresultatet være:

For Z_{i1} observeres N_{ij_1},

for Z_{i2} observeres N_{ij_2},

. . . . . .

og for Z_{im_i} observeres $N_{ij_{m_i}}$.

(Rekkefølgen er underordnet, — her fastsettes bare en konvensjon.)

Enhver realisasjon, N_{ij_v} ($v = 1, 2, \ldots . M_i$), av Z_{ij} har like stor sannsynlighet, nemlig $\frac{1}{M_i}$, ved en enkel trekning.

Anta nå at utvelgings-sannsynlighetene er fastsatte. La z_{ij} være antall sue i utvalget fra pue (i, j) i utvalget. z_{ij} er da lik:

$z_{ij} = f_{2i} Z_{ij}$, — altså en stokastisk variabel siden Z_{ij} er det. Stikkprøvens størrelse blir derfor også en stokastisk variabel, dvs. en sum av stokastiske variable:

$$z = \sum_{i=1}^{s} z_i = \sum_{i=1}^{s} \sum_{j=1}^{m_i} z_{ij} = \sum_{i=1}^{s} \sum_{j=1}^{m_i} f_{2i} Z_{ij}$$

Variasjonen i z avhenger av variasjonene i f_{2i} og Z_{ij} (dvs. N_{ij_v}).

Betrakt nå den forventede utvalgs-størrelse, Ez:

$$Ez = \sum_{i=1}^{s} \sum_{j=1}^{m_i} f_{2i} EZ_{ij}.$$

$$\text{Nå er } EZ_{ij} = \sum_{v=1}^{M_i} N_{ij_v} \cdot P(Z_{ij} = N_{ij_v}) = \frac{1}{M_i} \sum_{v=1}^{M_i} N_{ij_v}, \text{ men siden}$$

$$\sum_{v=1}^{M_i} N_{ij_v} = N_i, \text{ må } EZ_{ij} = \frac{N_i}{M_i}, \text{ og dermed fås:}$$

$$Ez = \sum_{i=1}^{s} \sum_{j=1}^{m_i} f_{2i} \frac{N_i}{M_i} = \sum_{i=1}^{s} f_{2i} \cdot f_{1i} \cdot N_i = \sum_{i=1}^{s} f\, N_i = f\, N.$$

Altså har vi at $f = \frac{Ez}{N}$.

Med de symbolene som ble brukt i kapittel VII, er altså n *forventet* utvalgsstørrelse, — dvs. n = Ez:

En helt konsistent språkbruk vil derfor kreve at vi kaller f for *forventet total utvalgsbrøk*, og

$f_i = \frac{Ez_i}{N_i}$ for *forventet utvalgsbrøk, stratum i.*

En bør merke seg at brøkene ikke fastlegger størrelsen på den konkrete stikkprøve og hvordan denne fordeles på strata. Utvalgsbrøkene angir bare hva vi kan vente å få i det lange løp ved gjentatte utvelginger, — da forutsatt at utvalgsbrøkene hele tiden holdes konstante.

English summary

The Central Bureau of Statistics has since 1967 had a permanent organization for sampling surveys based on intervie-wing. The article contains a description of the sample design which has been applied during the first years.

In order to keep the cost of country-wide sample surveys within reasonable limits, the C.B.S. decided to use a two-stage sampling system.

The primary sampling units (psu's) are of approximately equal size, 2000 inhabitants in average. The occupational structure of the population has been calculated for each psu, and the psu's have been stratified according to occupational structure and geographical location. Each stratum contains approximately the same number of psu's (30—38).

The theory of multivariate sampling gives some support to the choice of equally sized psu's and strata. From this theory, it might be argued that such a design carries some general optimum properties when the sample is to be a general purpose one. It should be added that to approach an optimum design in the multivariate case the over-all sampling ought to be taken proportionately.

At the first stage two psu's have generally been selected within each stratum. One reason why two psu's have been selected—and not only one—is to maintain the possibility of estimating variances within strata.

The first stage sample is permanent, and one interviewer has been employed in each of the sample-areas (i.e. selected psu's).

The over-all sampling fraction is usually kept constant. By doing this, we obtain a proportionate, selfweighting sample which is convenient for estimation and data-processing purposes.

One paramount problem in sampling is connected with the register from which the sample is to be selected. The article deals with some of the main difficulties of keeping the register up to date. To a certain extent, errors due to faults in the register can be avoided by careful definition of the second stage sampling units. For instance, in household surveys the sampling unit may be defined as the residence (flat) of the household. The persons to be interviewed are those living in the selected flat at the moment of interviewing. Thus, the problem of migration is avoided.

The article also maintains some points of view on more general problems in sampling. The effects of non-response and other non-sampling errors are discussed.

Utkommet i serien ART

Issued in the series Artikler fra Statistisk Sentralbyrå (ART)

Nr. 1 Odd Aukrust: Investeringenes effekt på nasjonalproduktet *The Effects of Capital Formation on the National Product* 1957 28 s. Utsolgt

» 2 Arne Amundsen: Vekst og sammenhenger i den norske økonomi 1920—1955 *Growth and Interdependence in Norwegian Economy* 1957 40 s. Utsolgt

» 3 Statistisk Sentralbyrås forskningsavdeling: Skattlegging av personlige skattytere i årene 1947—1956 *Taxation of Personal Tax Payers* 1957 8 s. Utsolgt

» 4 Odd Aukrust og Juul Bjerke: Realkapital og økonomisk vekst 1900—1956 *Real Capital and Economic Growth* 1958 32 s. Utsolgt

» 5 Paul Barca: Utviklingen av den norske jordbruksstatistikk *Development of the Norwegian Agricultural Statistics* 1958 23 s. kr. 2,00

» 6 Arne Amundsen: Metoder i analysen av forbruksdata *Methods in Family Budget Analyses* 1960 24 s. kr. 5,00

» 7 Arne Amundsen: Konsumelastisiteter og konsumprognoser bygd på nasjonalregnskapet *Consumer Demand Elasticities and Consumer Expenditure Projections Based on National Accounts Data* 1963 44 s. kr. 5,00.

» 8 Arne Øien og Hallvard Borgenvik: Utviklingen i personlige inntektsskatter 1952—1964 *The Development of Personal Income Taxes* 1964 30 s. kr. 5,00

» 9 Hallvard Borgenvik: Personlige inntektsskatter i sju vesteuropeiske land *Personal Income Taxes in Seven Countries in Western Europe* 1964 16 s. kr. 5,00

» 10 Gerd Skoe Lettenstrøm og Gisle Skancke: De yrkesaktive i Norge 1875—1960 og prognoser for utviklingen fram til 1970 *The Economically Active Population in Norway and Forecasts up to 1970* 1964 56 s. Kr. 6,00

» 11 Hallvard Borgenvik: Aktuelle skattetall 1965 *Current Tax Data* 1965 38 s. kr. 6,00

» 12 Idar Møglestue: Kriminalitet, årskull og økonomisk vekst *Crimes, Generations and Economic Growth* 1956 63 s. kr. 7,00

» 13 Svein Nordbotten: Desisjonstabeller og generering av maskinprogrammer for granskning av statistisk primærmateriale *Decision Tables and Generation of Computer Programs for Editing of Statistical Data* 1965 11 s. kr. 4,00

» 14 Gerd Skoe Lettenstrøm: Ekteskap og barnetall — En analyse av fruktbarhetsutviklingen i Norge *Marriages and Number of Children — An Analysis of Fertility Trend in Norway* 1965 29 s. kr. 6,00

» 15 Odd Aukrust: Tjue års økonomisk politikk i Norge: Suksesser og mistak *Twenty Years of Norwegian Economic Policy: An Appraisal* 1965 38 s. kr. 6,00

» 16 Svein Nordbotten: Long-Range Planning, Progress- and Cost-Reporting in the Central Bureau of Statistics of Norway *Lang-*

tidsprogrammering, framdrifts- og kostnadsrapportering i Statistisk Sentralbyrå 1966 17 s. kr. 4,00

Nr. 17 Olav Bjerkholt: Økonomiske konsekvenser av nedrustning i Norge *Economic Consequences of Disarmament in Norway* 1966 25 s. kr. 4,00

» 18 Petter Jakob Bjerve: Teknisk revolusjon i økonomisk analyse og politikk? *Technical Revolution in Economic Analysis and Policy?* 1966 23 s. kr. 4,00

» 19 Harold W. Watts: An Analysis of the Effects of Transitory Income on Expenditure of Norwegian Households 1968 28 s. kr. 5,00

» 20 Thomas Schiøtz: The Use of Computers in the National Accounts of Norway *Bruk av elektronregnemaskiner i nasjonalregnskapsarbeidet i Norge* 1968 28 s. kr. 5,00

» 21 Petter Jakob Bjerve: Trends in Quantitative Economic Planning in Norway *Utviklingstendensar i den kvantitative økonomiske planlegginga i Norge* 1968 29 s. kr. 5,00

» 22 Kari Karlsen og Helge Skaug: Statistisk Sentralbyrås sentrale registre *Registers in the Central Bureau of Statistics* 1968 24 s. kr. 3,50

» 23 Per Sevaldson: MODIS II A Macro-Economic Model for Short-Term Analysis and Planning *MODIS II En makroøkonomisk modell for korttidsanalyse og planlegging* 1968 40 s. kr. 4,50

» 24 Olav Bjerkholt: A Precise Description of the System of Equations of the Economic Model MODIS III *Likningssystemet i den økonomiske modell MODIS III* 1968 30 s. kr. 4,50

» 25 Eivind Hoffmann: Prinsipielt om måling av samfunnets utdanningskapital og et forsøk på å måle utdanningskapitalen i Norge i 1960 *On the Measurement of the Stock of Educational Capital and an Attempt to Measure Norway's Stock of Educational Capital in 1960* 1968 60 s. kr. 5,00

» 26 Hallvard Borgenvik: Aktuelle skattetall 1968 *Current Tax Data* 1969 40 s. kr. 7,00

» 27 Hallvard Borgenvik: Inntekts- og formuesskattlegging av norske kapitalplasseringer i utlandet *Income and Net Wealth Taxes of Norwegian Investment in Foreign Countries* 1969 40 s. kr. 7,00

» 28 Petter Jakob Bjerve og Svein Nordbotten: Automasjon i statistikkproduksjonen *Automation of the Production of Statictics* 1969 30 s. kr. 7,00

» 29 Tormod Andreassen: En analyse av industriens investeringsplaner *An Analysis of the Industries Investment Plans* 1969 26 s. kr. 5,00

» 30 Bela Balassa og Odd Aukrust: To artikler om norsk industri *Two Articles on Norwegian Manufacturing Industries* 1969 40 s. kr. 5,00

» 31 Hallvard Borgenvik og Hallvard Flø: Virkninger av skattereformen av 1969 *Effects of the Taxation Reform of 1969* 1969 35 s. kr. 7,00

» 32 Per Sevaldson: The Stability of Input-Output Coefficients *Stabilitet i kryssløpskoeffisienter* 1969 40 s. kr. 7,00

» 33 Odd Aukrust og Hallvard Borgenvik: Inntektsfordelingsvirkninger av skattereformen av 1969 *Income Distribution Effects of the Taxation Reform of 1969* 1969 29 s. kr. 7,00

» 34 Odd Aukrust og Svein Nordbotten: Dataregistrering, dataarkiver og samfunnsforskning *Data Registration, Data Banks and Social Research* 1969 43 s. kr. 7,00

» 35 Odd Aukrust: PRIM I A Model of the Price and Income Distri-

bution Mechanism of an Open Economy *PRIM I En modell av pris- og inntektsfordelingsmekanismen i en åpen økonomi* 1970 59 s. kr. 7,00

Nr. 36 Arne Amundsen: Konsumets og sparingens langsiktige utvikling *Consumption and Saving in the Process of Long-Term Growth* 1970 18 s. kr. 5,00

Publikasjonen utgis i kommisjon hos
H. Aschehoug & Co., Oslo, og er til salgs hos alle bokhandlere
Pris kr. 7,00