

Interne notater

STATISTISK SENTRALBYRÅ

32/6

12. februar 1982

OM TREKKING AV UTVALG TIL ET SYSTEM AV LEVEKÅRSUNDERSØKELSER¹⁾.

av
Ib Thomsen^{*)}

INNHOLD

	Side
1. Innledning	1
2. En generell modell for et system av utvalgsundersøkelser	1
2.1. Innledning	1
2.2. Begreper og definisjoner	1
2.3. Optimal strategi	4
3. Momenter ved valg av strategi	5
3.1. Innledning	5
3.2. Strategier som tar sikte på å redusere innsamlingskostnadene ..	5
3.3. Reduksjon av usikkerheten til estimatorer for endringer over tid	5
3.4. Økte muligheter for analyse av data	6
3.5. Foreløpige konklusjoner	7
4. Praktiske problemer i forbindelse med panel og roterende utvalg	7
4.1. Innledning	7
4.2. Handtering av utvalget	8
4.3. Frafall	9
4.4. Målefeil	10
4.5. Spesielle problemer knyttet til husholdningsundersøkelser	13
Litteratur	14

1) Arbeidet er delvis finansiert av Forbruker- og administrasjonsdepartementet.

*) Tor Haldorsen har skrevet kapitlene 4.1-4.3

1. INNLEDNING

I de fleste tilfeller planlegges en undersøkelse som om den bare skal gjennomføres en gang, og som om det ikke foreligger klare planer for å gjenta den. Dessuten er det vanlig å se på utvalgsplanen for en undersøkelse uavhengig av planene fra andre undersøkelser. De to viktigste unntakene fra denne regelen i Byrået er arbeidskraftundersøkelsene og forbruksundersøkelsene. For begge disse undersøkelser har en laget en utvalgsplan som tar hensyn til at undersøkelsene gjentas med visse tidsmellomrom.

I forbindelse med levekårsundersøkelsene er det satt fram ønsker om å planlegge disse, ikke som en enkelt undersøkelse, men som en hel serie av undersøkelser, og i tillegg legge stor vekt på disse undersøkelsers sammenheng med annen statistikk og andre utvalgsundersøkelser. Spesielt er det framsatt ønske om å koordinere levekårsundersøkelsene med inntektsundersøkelsene. Se NOU 1980:20 : "Om arbeidet med levekårsspørsmål". Et av de viktigste redskaper i en slik koordinering er bruken av felles definisjoner og standarder i de ulike undersøkelsene. I dette notatet skal vi begrense oss til å se på muligheter og problemer som oppstår når en ønsker å koordinere trekningen av utvalgene til levekårsundersøkelsene med utvalgene til inntektsundersøkelsene. Dessuten skal vi se på hvordan utvalgene til framtidige levekårsundersøkelser bør trekkes.

I kap. 2 skal vi først formulere en generell modell for et system av utvalgsundersøkelser som en tenker seg strekker seg inn i framtiden. I kap. 3 skal det deretter presenteres noen argumenter for panel- og roterende utvalg. I kap. 4 skal behandles noen problemer knyttet til frafall, håndtering av utvalget, målefeil ved panel- og roterende utvalg.

2. EN GENERELL MODELL FOR ET SYSTEM AV UTVALGSUNDERSØKELSER

2.1. Innledning

For den videre diskusjon i notatet er det nyttig å ha en generell modell for et system av utvalgsundersøkelser. I et slikt system kunne en naturligvis tenke seg hele det statistiske systemet, men vi skal her konsentrere oss omkring en beskrivelse av et system av utvalgsundersøkelser.

Ettersom vi i dette notatet først og fremst skal behandle trekking av utvalg, skal de øvrige begreper og metoder bare nevnes for fullstendighetens skyld.

2.2. Begreper og definisjoner

I et system av utvalgsundersøkelser inngår tre hovedkomponenter: innsamlingsmetoder, metoder for trekking av utvalg og til slutt metoder for arkivering av data som gjør det mulig å foreta grundige analyser både innen Byrået og av personer utenfor Byrået.

Innsamlingsmetoder skal her oppfattes i bred betydning og dekke metoder for spørsmålsformulering, utdanning og drift av intervjuerkorpset, kodings- og revisjonsrutinger m.m. Vi definerer innsamlingsfasen som avsluttet når det foreligger en file som kan brukes som tabellgrunnlag, og som kan gå inn i et arkiv for utvalgsdata.

Viktige instrumenter for koordineringen av innsamlingsmetodene er spørsmålsarkiv, arkiv for kodings- og revisjonsinstruksjoner m.v.

Trekking av utvalg omfatter utvalgsplaner og registerrutiner for samtlige utvalgsundersøkelser som inngår i systemet.

Måter å arkivere data på. Det er viktig at data arkiveres på en måte slik at de er lett å få tak i for senere bruk. Arkivet bør derfor i tillegg til data inneholde en del informasjon om de innsamlings- og utvalgsmetoder som har vært nyttet.

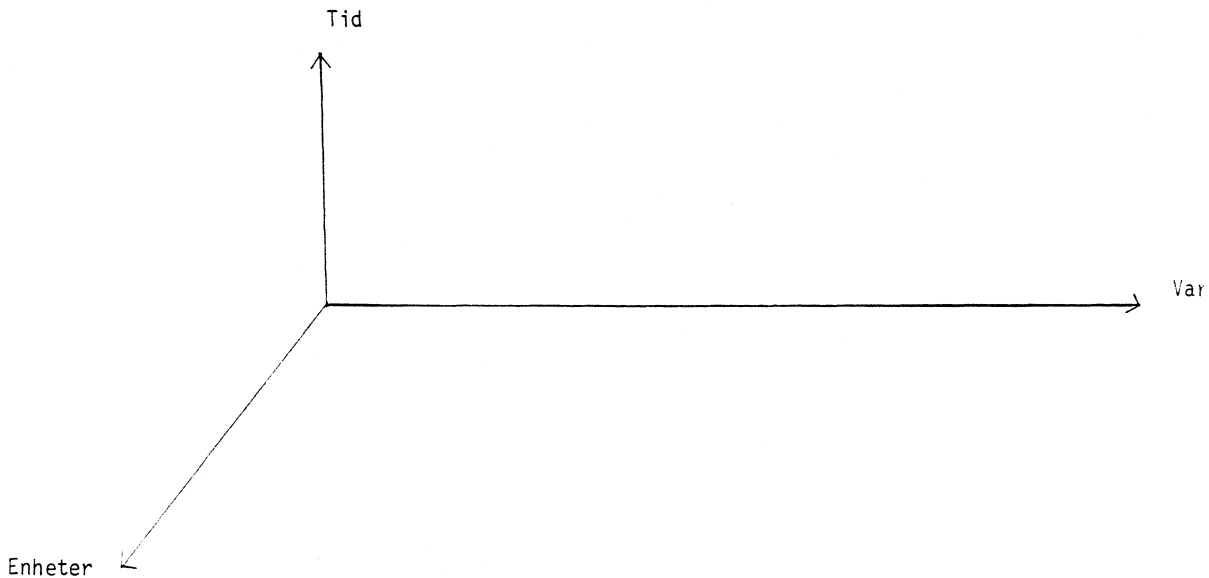
Universet som utvalgene skal trekkes fra kan beskrives som en tredimensjonal datamatrix som en tenker seg strekker seg inn i framtiden. Se fig. 1. Tallet på en bestemt plass i datamatriksen angir verdien til en bestemt variabel, på et bestemt tidspunkt samt for en bestemt enhet i universet. Hvert

dataelement kan skrives som

$$d = X(i, t_j, k)$$

hvor i er et tall som identifiserer enheten, t_j angir et tidspunkt og k er en variabelidentifikasjon. Universet vil forandre seg over tid, noen enheter vil "dø", mens andre vil bli "født".

Figur 1.



Med populasjonsparametre skal vi mene en funksjon av dataelementene. F.eks. er funksjonen

$$X(t_j, k) = \sum_{i \in U_{t_j}} x(i, j, k),$$

hvor $\sum_{i \in U_{t_j}}$ betyr summen over alle enheter til tiden t_j lik den totale verdi av en variabel k på tidspunkt t_j . Det tilsvarende gjennomsnittet fås ved å dividere med antall enheter til tiden t_j , N_{t_j} .

$$\bar{X}(t_j, k) = \frac{1}{N_{t_j}} \sum_{i \in U_{t_j}} x(i, j, k).$$

Populasjonsgjennomsnitt er en av de parametre som en er opptatt av å estimere ved hjelp av utvalgsøkkelser.

I tillegg til å estimere gjennomsnitt for hele populasjonen på ett bestemt tidspunkt, er det oftest også interessert i å estimere gjennomsnitt for deler av populasjonen, samt differanser mellom disse gjennomsnittene. Vi skal ikke se nærmere på dette problemet i denne sammenheng. De fleste traktninger i det følgende gjelder også når en skal estimere i delklasser. Et spesielt problem i praksis er estimering av geografiske delklasser. Dette skyldes at mange utvalg trekkes i flere men er uten spesiell interesse i denne sammenheng. Interesserte lesere henvises til Siring og The (1981).

Et annet eksempel på en parameter en ofte er opptatt av, er

$$\bar{X}(t_1, k) - \bar{X}(t_2, k),$$

som er forskjellen mellom gjennomsnittene til variabel k til tidspunktene t_1 og t_2 .

Modell

Det er ikke bestandig at de parametre en ønsker å estimere kan settes opp som funksjoner av elementene i datamatriksen. F.eks. ønsker en å estimere parametre i en modell som beskriver sammenhengene mellom to og flere variable på ett bestemt tidspunkt. Et annet eksempel er en modell for utviklingen over tid. En hyppig anvendt modell for slike formål er f.eks.

$$Y_t = a + bX_t + e_t, \quad t = 0, 1, 2, \dots$$

hvor Y_t er en endogen variabel til tidspunkt t , X_t er en eksogen variabel til tidspunkt t og e_t er et ikke observerbart restledd.

I denne modellen er a og b parametre, men de er ikke funksjoner av elementene i datamatriksen. De har ingen eksistens i noen endelig populasjon, men er hypotetiske størrelser som uttrykker sammenhengen mellom variablene Y_t og X_t .

I prinsippet er det visse forskjeller mellom populasjonsparametre og modellparametre, men i denne sammenhengen skal vi ikke komme nærmere inn på disse forskjeller, men bare gjøre oppmerksom på de to typene av parametre, som begge er aktuelle i forbindelse med levekårsundersøkelsene.

For å estimere parametrene trekkes det utvalg fra universet. Med et utvalg skal vi mene en delmengde av tallene i datamatriksen.

Med et statistisk utvalg skal vi mene utvalg fra datamatriksen som har den egenskap at enhver enhet har en kjent sannsynlighet for å være med til ethvert tidspunkt. Vi skal altså ikke snakke om statistiske utvalg av variable. I det følgende skal vi ofte omtale statistiske utvalg bare som utvalg når meningen er klar av sammenhengen.

Et utvalg bestående av verdiene til en variabel k fra et sannsynlighetsutvalg med størrelse n til tiden t_1 , kan nå skrives som

$$S_1 = \bigcup_{i=1}^n x(i, t_1, k).$$

På samme måten betyr

$$T_1 = \bigcup_{k=1}^L \bigcup_{i=1}^n x(i, t_1, k).$$

et utvalg bestående av verdiene til variablene 1 til L for et utvalg av individer til tiden t_1 . Et slikt sannsynlighetsutvalg kalles ofte et tverrsnitt.

Et panel er data om det samme utvalg av enheter til flere tidspunkter, og kan skrives som

$$P = \bigcup_{t=1}^T \bigcup_{k=1}^L \bigcup_{i=1}^n x(i, t, k).$$

Roterende utvalg er utvalg som delvis består av data om de samme personer på flere tidspunkter, og delvis data for personer på et enkelt tidspunkt. La utvalgene på to tidspunkter t_1 og t_2 bestå av n enheter på hvert tidspunkt, og la n_2 av enhetene være med på begge tidspunkter. Dette er et roterende utvalg, og kan skrives som

$$\left\{ \bigcup_{k=1}^L \bigcup_{i=1}^{n_1} x(i, t_1, k) \right\} \cup \left\{ \bigcup_{j=1}^2 \bigcup_{k=1}^L \bigcup_{i=1}^{n_1} x(i, t_j, k) \right\} \cup \left\{ \bigcup_{k=1}^L \bigcup_{i=n_2+1}^n x(i, t_2, k) \right\}.$$

Dette kan virke som en unødige komplisert måte å uttrykke et roterende utvalg på, men demonstrerer at det er mulig å bruke begrepene definert ovenfor. Vi skal ikke gjøre særlig bruk av denne formelle beskrivelse av et roterende utvalg.

På grunnlag av data i utvalget eren nå interessert i å estimere parametrene i populasjonen og/eller i en eller flere modeller. For dette brukes en estimator, som er en funksjon av variabelverdiene i utvalget. F.eks. kan utvalgsgjennomsnittet til tidspunkt t_1 av variabel k skrives som

$$\sum_{i \in S} x(i, t_1, k),$$

hvor $\sum_{i \in S}$ betyr summen over alle enheter i utvalget. Også mer kompliserte estimatorene kan uttrykkes på lignende måte.

2.3. Optimal strategi

Dersom utvalget er et sannsynlighetsutvalg, vil det være mulig å bestemme forventning og varians til de fleste av de estimatorer som vi skal ta opp nedenfor.

Med uttrykket strategi skal vi mene valg av utvalgsplan samt valg av estimator for å estimere en eller flere parametre. Med optimal strategi skal vi mene den strategi som gir en forventningsrett estimator og minst varians når en ønsker å estimere et bestemt sett av parametre.

I praksis er det ofte vanskelig og uhensiktsmessig å søke etter en optimal strategi, men i noen tilfeller kan det være nyttig å bruke den optimale strategi som utgangspunkt når en lager utvalgsplaner og velger estimatorer.

I forbindelse med levekårsundersøkelsene skal det estimeres et stort antall parametre, og mange analyser skal ventelig utføres. Det er derfor en nokså håpløs oppgave å finne noen optimal strategi, selv for et enkelt tidspunkt. I det følgende skal vi derfor konsentrere diskusjonen rundt valg mellom bruk av roterende utvalg, panel og utvalg som blir trukket uavhengig av hverandre til hvert tidspunkt. Vi skal tenke oss noen typiske parametre, både populasjonsparametre og modellparametre, som det kan bli bruk for å estimere på grunnlag av levekårsundersøkelsene, og deretter se på hvordan utvalgene må trekkes for å estimere disse parametre. Dessuten skal vi til slutt se på de konsekvenser de forskjellige valg har for inntektsundersøkelsene.

3. MOMENTER VED VALG AV STRATEGI

3.1. Innledning

Som nevnt tidligere er det en uendelighet av måter å trekke utvalg på. Litteraturen om utvalgs-teori er omfattende når det gjelder å bestemme gode utvalgsplaner og estimeringsmetoder for trekking av tverrsnitt. I dette notatet er vi spesielt interessert i å finne måter å trekke en hel serie av utvalg på, og dessuten se på de problemer som oppstår når en ønsker å koordinere utvalgene til levekårsundersøkelsene med utvalgene til inntektsundersøkelsene. I dette kapittel skal vi spesielt ta opp spørsmål knyttet til de fordeler det er ved å trekke utvalgene til levekårsundersøkelsene som panel eller roterende utvalg. Konsekvensene for inntektsundersøkelsene vil bli tatt opp i avsnitt 4.5.

Når en skal diskutere bruk av panel eller roterende utvalg i forbindelse med levekårsundersøkelsene, lønner det seg å dele diskusjonen slik:

- (i) Reduksjon av kostnader
- (ii) Reduksjon av usikkerheten til estimatorer for endringer over tid
- (iii) Økte muligheter for analyse av data.

3.2. Reduksjon av kostnader

Langt de fleste person- og husholdningsutvalg som brukes i Statistisk Sentralbyrå trekkes i to trinn. I første trinn trekkes et utvalg av kommuner, og deretter trekkes et utvalg av personer eller husholdninger innen de uttrukne kommuner. Hensikten med å trekke utvalgene på denne måten er å konsentrere utvalgene innen et begrenset antall kommuner, hvor en har ansatt intervjuere for å samle inn data. De kommuner som blir trukket i første trinn forblir i utvalget for en lengre periode, slik at andre utvalg som trekkes i denne perioden trekkes innen de uttrukne kommuner. I dette tilfellet velger en altså i en bestemt periode å trekke samtlige husholdnings- og personutvalg innen et begrenset antall kommuner. Hensikten med denne koordineringen er å redusere innsamlingskostnadene.

Nå er det klart at en for å unngå å "slite ut" de mindre utvalgsområdene, er nødt til å skifte ut disse oftere enn tilfellet er med større utvalgsområder. Av hensyn til mulighetene for koordinering av utvalg i annet trinn, bør en vente så lenge som mulig med å skifte ut utvalgsområdene, dessuten bør utskiftingen skje løpende, slik at en ikke blir nødt til å skifte ut for mange kommuner på en gang.

I Byrået er det i dag ingen klare planer for om en bør lage en ny utvalgsplan etter folketellingen 1980, men det ser ut som om det ville være en fordel med mindre justeringer i løpet av første halvdel av 1980-årene, dersom intervjuerstaben fortsatt skal ha den størrelse den har i dag. Hvis det på den andre siden blir nødvendig å redusere antall intervjuere vesentlig, må utvalgsplanen legges vesentlig om, og det må trekkes nye utvalg også i første trinn. Dette ville kunne redusere mulighetene for å koordinere utvalg over tid vesentlig.

3.3. Reduksjon av usikkerheten til estimatorer for endringer over tid

En vanlig årsak til å gjenta en utvalgsundersøkelse er at en ønsker å estimere endringer over tid. I kapittel 2 definerte vi

$$\bar{X}(t_1, k) - \bar{X}(t_2, k), \quad (3.1)$$

som forskjellen mellom gjennomsnittene til variabel k på tidspunktene t_1 og t_2 . Når en skal estimere en eller flere parametre av typen (3.1) er det en stor fordel å bruke panel eller roterende utvalg. Når en estimerer endringer ved hjelp av panel eller roterende utvalg er utvalgsvariansen til endringen nesten alltid mindre enn om en estimerer ved hjelp av uavhengige utvalg. La $\bar{x}(t_1, k)$ og $\bar{x}(t_2, k)$ betegne utvalgsgjennomsnittene til variabel k på tidspunktene t_1 og t_2 respektive. Da er forskjellen mellom de to utvalgsgjennomsnittene ofte en forventingsrett estimator for (3.1), og det gjelder at

$$\text{var}(\bar{x}(t_1, k) - \bar{x}(t_2, k)) = \text{var}(\bar{x}(t_1, k)) + \text{var}(\bar{x}(t_2, k)) - 2 \text{cov}(\bar{x}(t_1, k), \bar{x}(t_2, k)) \quad (3.2)$$

Usikkerheten til forskjellen mellom de to gjennomsnitt, målt ved variansen, er altså ofte større enn usikkerheten til hvert av gjennomsnittene. Dersom en bruker helt uavhengige utvalg på de to tidspunktene, er

$$\text{cov}(\bar{x}(t_1, k), \bar{x}(t_2, k)) = 0,$$

og variansen til forskjellen blir altså summen av variansene til hvert av utvalgsgjennomsnittene. Dersom en bruker et panel, vil kovariansen mellom $\bar{x}(t_1, k)$ og $\bar{x}(t_2, k)$ være lik kovariansen mellom de individuelle svar på de to tidspunktene. Vanligvis vil svarene til de samme spørsmål på to følgende tidspunkter være positivt korrelerte, og derfor vil variansen være mindre når en bruker panelundersøkelse enn om en bruker uavhengig utvalg.

En annen fordel ved panelundersøkelsen er at en i tillegg til å kunne måle netto-endringer, også kan måle brutto-endringene. I arbeidskraftundersøkelsene, som er roterende utvalg, kan en i tillegg til å estimere endringer i sysselsettingsrater også estimere hvor mange som har skiftet verdi på variabelen sysselsatt/ikke sysselsatt. Slike bruttomål for endringene kan være av stor betydning i forbindelse med analyser, f.eks. prognoser for arbeidsstyrken.

3.4. Økte muligheter for analyse av data

Den største fordel ved å bruke panel ligger antakeligvis i de økte muligheter for analyse som slike utvalg gir. I det følgende skal vi kort se på noen av de typer av analyser, som forutsetter at en har observasjoner om de samme individer på flere tidspunkter, og referere noe av den omfattende litteratur som de siste årene er skrevet på dette feltet.

3.4.1. Markov-modeller

En type modeller som har vært mye brukt av sosiologer i forbindelse med analyse av panelundersøkelser er Markov-modeller. Markus (1979). I sin enkleste form uttrykker denne modellen at sannsynligheten for å ha en bestemt verdi på en variabel på tidspunkt $(t+1)$ bare avhenger av verdien til variabelen på tidspunkt t .

Hvis f.eks. en variabel bare antar verdiene 0 og 1, trenger en fire modell-parametre for å beskrive utviklingen fra t til $(t+1)$. Disse er

P_{11} : Sannsynligheten for å anta verdien 1 til tidspunkt $(t+1)$ gitt at verdien er 1 på tidspunkt t .

P_{10} : Sannsynligheten for å anta verdien 0 til tidspunkt $(t+1)$ gitt at verdien er 1 på tidspunkt t .

P_{01} : Sannsynligheten for å anta verdien 1 til tidspunkt $(t+1)$ gitt at verdien er 0 på tidspunkt t .

P_{00} : Sannsynligheten for å anta verdien 0 til tidspunkt $(t+1)$ gitt at verdien er 0 på tidspunkt t .

Nå gjelder det at

$$P_{11} + P_{10} = 1$$

og

$$P_{01} + P_{00} = 1,$$

slik at det er tilstrekkelig å kjenne to av parametrene.

Markov-modeller danner ofte utgangspunkt for analyse av panel og roterende utvalg. F.eks. er utviklingen av politiske holdninger i England studert ved slike modeller, Butler and Stokes (1974).

For å kunne estimere parametrene må en ha panel eller roterende utvalg. Dersom en i tillegg skal

teste forutsetningen om at parametrene er konstant over tid, må en dessuten ha data fra minst tre undersøkelser til sin rådighet.

I mange anvendelser viser det seg at Markov-modellen i sin enkleste form ikke viser seg å gi god tilpasning til data. Det viser seg da ofte nyttig å foreta modifikasjoner av modellen. De vanligste er

- (i) La parametrene variere mellom grupper i populasjonen. Coleman (1964)
- (ii) La parametrene variere over tid
- (iii) Definere et sett med latente variable (målefeil) og la modellen være oppfylt for disse. Wiggins (1973).

Uansett hvilken modifikasjon en bruker, må en ha data fra panel, eller roterende utvalg.

3.4.2. Kausal-analyse. (Regresjonsanalyse)

Dersom en ønsker å estimere utviklingen i fordelingene til et sett av levekårskomponenter, kan en delvis klare seg med utvalg som er trukket uavhengig av hverandre. De data en da vil ha etter noen gjentakelser vil da være svært lik de tidsrekke-data som økonometrikere har arbeidet med gjennom flere år. Se f.eks. Johnston, 1972, kap. 8. I de fleste tilfeller ønsker en å estimere den simultane utvikling i flere levekårskomponenter, ofte som funksjon av andre variable som kjønn, utdanning osv. Også for slike analyser kan en bruke data samlet inn ved uavhengige utvalg. Analysemetodene vil ofte bestå av å estimere i simultane ligningssystemer. Problemene i denne forbindelse er ofte vanskelige. En god introduksjon er gitt i Johnston, (1972), kap. 12 og 13 og Ostrom (1979) hvor det også finnes en omfattende referanseliste.

På tross av at det ofte ikke er strengt nødvendig med panel, eller roterende utvalg, kan det i noen modeller være nødvendig med panel eller roterende utvalg. Hvis en f.eks. ønsker å studere sammenhengen mellom levekårskomponentene vil uavhengige utvalg være utilstrekkelig i mange tilfeller. Hvis hensikten er å finne ut om arbeidsforhold har noen effekt på helse, kan en i et tverrsnitt dele utvalget inn i personer med dårlige arbeidsforhold og med gode arbeidsforhold, og sammenlikne helsetilstanden mellom de to gruppene.

En slik analyse er naturligvis lite tilfredsstillende fordi en ikke vet noe om helsetilstanden før arbeidsforholdet ble etablert. For å få slik informasjon er det nødvendig å kjenne helsetilstanden på et tidligere tidspunkt. Det ville naturligvis være en fordel hvis en hadde en fullstendig oversikt over arbeidsforholdene og helsen til ethvert tidspunkt, noe som kanskje kunne oppnås med retrospektive spørsmål, men en ville komme et godt stykke på vei med panel eller roterende utvalg. Selv med slike data er det knyttet betydelige problemer til å foreta en kausal analyse, men i denne sammenhengen er det tilstrekkelig å konstatere at kausal analyse er umulig uten panel eller roterende utvalg.

3.5. Foreløpige konklusjoner

Når en skal velge strategi for levekårsundersøkelsene, ville det være en fordel om en på forhånd hadde en fullstendig liste over alle anvendelser av undersøkelsene. En slik liste eksisterer ikke, og i dette kapittel har vi derfor vært nødt til å tenke oss til hvilke analyser som kan bli aktuelle, og vist behovet for panel eller roterende utvalg. Det har da vist seg, at mange av de analyser det er naturlig å vente seg vil bli aktuelle i forbindelse med levekårsundersøkelsene, forutsetter panel i roterende utvalg eller retrospektive spørsmål. Andre analyser kan utføres med uavhengige utvalg, men vil bli vesentlig bedre med andre strategier. I neste kapittel skal vi se på noen av de praktiske problemer som er knyttet til bruk av panel og roterende utvalg.

4. PRAKTISKE PROBLEMER I FORBINDELSE MED PANEL OG ROTERENDE UTVALG

4.1. Generelt

Momentene i kapittel 3 peker i retning av at roterende utvalg og panelutvalg er fordelaktig i en del sammenhenger. De gir muligheter for bedre estimering av endringer, og økte analysemuligheter, for noen problemstillinger er de strengt nødvendige. På den annen side medfører denne type utvalg en

del spesielle problemer. Disse må også taes i betraktning når en velger strategi. I dette kapitlet pekes på noen slike problemer og det knyttes kommentarer til noen av dem.

Når en undersøkelse gjentas flere ganger vil som regel ikke målpopulasjonen være identisk fra en gang til en annen. En vil ofte ha en viss avgang (død, flytting utenlands, passering av øvre aldersgrense, endret status o.l.) og en viss tilgang (fødsel, innflytting, passering av nedre aldersgrense, endret status o.l.). Det betyr at noen av enhetene i utvalget ikke blir med i undersøkelsen det antall ganger en på forhånd har bestemt. Problemene dette skaper er som regel lette å løse, unntak kan være undersøkelser med husholdning som enhet. Dette behandles i kapittel 4.5.

Administrasjonen av utvalget ved roterende utvalg og panelutvalg skaper en del ekstraarbeid. Før seinere undersøkelser må adressene til enhetene i utvalget ajourføres, og en bør helst ha mulighet til å sjekke om enheter har endret status så de har falt utenfor målpopulasjonen. Hvis en skal gjøre dette i felten, blir det økning i kostnadene.

Når utvalgsplanene bygger på faste intervjuere i bestemte områder, vil flytting i tiden mellom undersøkelsene skape problemer. En del personer/husholdninger vil ha flyttet til områder som ikke er dekket av noen intervjuere. Intervjuing av slike kan bli kostbar. Vi vil se nærmere på dette i kapittel 4.2.

Frafall må en regne med også i disse undersøkelsene. Det kan redusere den forventede gevinst ved roterende utvalg og panelutvalg. En må også ta i betraktning at frafallet kan bli ekstra stort ved slike undersøkelser da noen som svarer første gang (er) går lei og ikke vil være med lenger. Noen erfaringer på dette området drøftes i kapittel 4.3.

Når en har panelutvalg eller roterende utvalg, vil en som regel også analysere tverrsnittutvalgene. En bør derfor innrette seg slik at disse utvalgene får gode egenskaper. Det kan oppstå situasjoner der en må veie mellom hensynet til "panelegenskaper" og "tverrsnittsegenskaper". Et moment i denne forbindelse er at det kreves både kompetanse og ressurser for å dra full nytte av egenskapene til helt eller delvis overlappende utvalg.

Et praktisk moment i forbindelse med gjentatt intervjuing er at Dataloven stiller spesielle krav om informasjon til de som intervjues.

4.2. Handtering av utvalget

Det vesentlige problem i forbindelse med undersøkelser som går over lang tid er at noen flytter i tiden mellom to intervju. Det er derfor nødvendig å ajourføre adressene i utvalget før hver ny intervjurunde. Hvert år er det snaut 200 000 flyttinger over kommunegrensene og drøyt 200 000 flyttinger innen kommunene. Flyttetilbøyeligheten avhenger av alder og er på annen måte ulikt fordelt i befolkningen, så det er vanskelig ut fra disse tall å beregne hvor mange i et bestemt utvalg som har fått ny adresse i løpet av en bestemt periode. I Valgundersøkelsen 1981 ville en ha delvis overlapping med utvalget til tilsvarende undersøkelse i 1977. I alt 2 153 personer (alder 25-79 i 1981) fra 77-utvalget var aktuelle for 81-utvalget. Ved kontroll i 1981 viste det seg at av disse var 82 døde, 236 var bosatt i en annen kommune enn i 77. Minst 138 personer hadde en ny adresse innenfor 77-kommunen, (138 er et minimumstall fordi det viser antall med nytt postnummer innen samme kommune).

I utvalgsundersøkelser basert på personlig intervju vil flyttinger også skape et annet problem. Noen intervjuobjekter vil i tiden mellom to intervjuer ha flyttet til områder som ligger svært langt fra der de faste intervjuere er stasjonert. Det kan derfor bli svært dyrt å gjennomføre intervju med enkelte. Størrelsen på dette problemet avhenger bl.a. av hvor intervjuerne er plassert, flyttemønsteret og kommunikasjonsnettets. Utvalgsområdene til Byrådet dekker omlag 60 prosent av befolkningen. I disse er faste intervjuere som også kan oppsøke noen steder utenfor utvalgsområdene uten vesentlige mer-kostnader. Foran Valgundersøkelsen i 1981 fant en som nevnt at av 2 153 hadde 236 bosatt seg i annen kommune enn i 1977. Av de 236 bodde 100 utenfor Byråets utvalgsområder. En prøvde å vurdere ekstra kostnader for å nå disse 100 og fant at 30 bodde slik til at det ville bli minst 100 kr i ekstra-kostnader pr. stk. Flyttingene skapte altså ingen vesentlige vanskeligheter for denne undersøkelsen.

Hvis ekstrakostnadene ved å intervju alle som har flyttet blir så store at en må sløyfe noen av intervjuene, er det viktig å passe på at en ikke innfører skjevheter i utvalget. En strategi kan være å trekke et mindre tilfeldig utvalg blant flytterne og bruke pengene en har til å intervju alle som trekkes.

4.3. Frafall

Som ved vanlige tverrsnittundersøkelser skaper frafallet problemer ved panel og roterende utvalg. En får færre observasjoner enn ønskelig og taper informasjon. Frafall er en mulig kilde til skjevheter og gjør det vanskeligere å bruke de opplysninger en har. De komplikasjoner frafall medfører for bruken av kjente analysemetoder er kanskje større ved panel og roterende utvalg enn ved tverrsnittundersøkelser. Et problem av en litt annen karakter er om den gjentatte intervjuingen i seg selv skaper større frafall. Hvis det er slik at folk før eller siden går lei og blir frafall, kan frafallsproblemene bli svært store.

Det er rimelig at størrelsen av en slik effekt vil variere fra undersøkelse til undersøkelse. Den vil avhenge av undersøkelse(e)s art, antall intervjuer i alt, tidsavstand mellom intervjuene m.m. Vårt erfaringsmateriale er såpass lite at vi ikke kan beskrive denne variasjonen, men vi skal gjengi noen av resultatene fra Byråets undersøkelser.

Til Levekårsundersøkelsen 1973 ble utvalget på 3 874 personer over 16 år trukket tilfeldig fra utvalget til Forbruksundersøkelsen samme år. Den sistnevnte undersøkelsen ble gjennomført først. Tidsavstanden mellom intervjuene var høyst et år, som regel en god del mindre. Personene som var i utvalget for begge undersøkelser fordelte seg på følgende måte regnet i avrundede prosenter.

		Levekår 73		
		Intervjuet	Frafall	
Forbruk	Intervjuet	69	11	80
	Frafall	7	13	20
		77	23	100

For det første ser en av tallene at de som er frafall i én undersøkelse med stor sannsynlighet også er frafall i den andre. Det kommer av at noen nekter å delta i de fleste typer undersøkelser og at andre som regel er vanskelig å treffe hjemme. Frafallet er litt større i Levekårsundersøkelsen enn i Forbruksundersøkelsen. Av den erfaring en ellers har på 70-tallet om frafallet i ulike typer undersøkelser, ville en forvente det motsatte. Siden Levekårsundersøkelsen ble gjennomført sist, kan vi derfor ikke utelukke at en del av frafallet ved Levekårsundersøkelsen skyldes at utvalget nokså nylig hadde vært med i utvalget til Forbruksundersøkelsen.

I Arbeidskraftundersøkelsene (AKU) intervjues folk i alt fire ganger, først i to kvartal etter hverandre så i de samme to kvartal året etter. I hvert kvartal består tverrsnittsutvalget av 4 puljer som er med i undersøkelsen for henholdsvis første, annen, tredje eller fjerde gang. Personer i alderen 16 til 74 år deltar i undersøkelsen. En grovregning på puljene som startet i undersøkelsen i 1977 - 79 viser at frafallet var henholdsvis 12, 11, 12 og 12 prosent i de fire intervjurundene. I disse tallene er det ikke justert for en del mindre effekter som virker i forskjellige retninger i forhold til det en her skal belyse. Tallene viser ikke tegn til økt frafall i de senere intervjurunder. Men det bør her nevnes noen gunstige forhold ved AKU. Det er et kort intervju, praktisk talt likt fra gang til gang, så det er lett å gjennomføre indirekte intervju og/eller telefonintervju.

Til Valgundersøkelsen i 1981 besto om lag halvparten av utvalget av personer som hadde vært med i utvalget til tilsvarende undersøkelse i 1977. Siden en har en kontrollgruppe, den del av utvalget som er nye i 1981, kan data fra denne undersøkelsen belyse vårt problem. En begrenser seg til den del av utvalget som i 1981 var 24 - 79 år for å ha sammenliknbare tall. Da viser det seg at frafallet i 1981 var 31,2 prosent i paneldelen og 24,6 prosent i den andre delen av utvalget. Kun en liten del av forskjellen i frafall kan forklares ved hjelp av vanskeligheter med flyttere, så resultatene indikerer økt frafall ved gjentatt intervjuing. For paneldelen har en:

		1981		
		Intervjuet	Frafall	
1977	Intervjuet	594	178	772
	Frafall	78	127	205
		672	305	977

En har ikke nok resultater til å si noe bestemt om det blir økt frafall ved gjentatt intervjuing. Det er indikasjoner på at en slik tendens vil avhenge av bl.a. undersøkelsens art. Men resultatene gir all grunn til å regne med en slik mulighet ved planlegging og gjennomføring av undersøkelser.

4.4. Målefeil

Omfanget og betydningen av målefeil i forbindelse med longitudinelle studier varierer mye med innsamlingsmetode og de analyseteknikker som brukes. I motsetning til hva som er tilfellet med utvalgsvarianser er det meget vanskelig å si noe generelt om størrelsen på målefeilene av den enkle grunn at de varierer mye fra variabel til variabel, og fordi estimeringen av målefeil er meget komplisert og ressurskrevende sammenliknet med estimeringen av utvalgsvarianser.

Til nå har det vært utført noen få målefeilstudier i Byrået, og i framtiden regner vi med å legge fram et større arbeid med resultater fra disse studiene. I denne sammenheng skal vi innskrenke oss til å referere noen få eksempler på målefeil, for å se om de erfaringer vi har til nå bør ha noen konsekvenser for trekking av utvalg til levekårsundersøkelsene i framtiden.

Problemet har mange aspekter. For det første kan en spørre om påliteligheten av svarene til retrospektive spørsmål. Resultatene fra slike studier varierer mye fra forskningsrapport til forskningsrapport. En av de sammenhenger hvor dette spørsmålet er studert mest grundig, er studiet av skifte i sosioøkonomisk status mellom generasjoner. En har her konstruert modeller hvor målefeil inngår eksplisitt. Likevel varierer konklusjonene mye. Bielby (1977). I Byrået har vi ingen eller få opplysninger om kvaliteten til retrospektive spørsmål, og av den grunn og fordi problemet ikke er helt sentralt når det gjelder trekkingen av utvalget ved levekårsundersøkelsene, skal vi ikke komme nærmere inn på dette vanskelige spørsmålet her.

Et annet problem er kvaliteten til svarene på spørsmål som refererer seg til nåtid eller nærmeste fortid. I Byrået har det vært utført noen få forsøk på å vurdere omfanget av slike feil. Tre metoder har vært brukt:

- (i) Gjennom intervjuing er svarene på folke- og bolig tellingsskjemaene blitt kontrollert.
- (ii) Data samlet inn ved intervjuundersøkelser har blitt sammenliknet med data fra administrative registre.
- (iii) Det er foretatt forskjellige logiske kontroller på data.

I tabell 1 er gjengitt noen resultater fra kvalitetskontrollundersøkelsen 1970. En ser at sysselsettingen stort sett blir underestimert i Folke- og bolig tellingen 1970, og at underestimeringen er størst blant kvinner.

Tabell 1. Personer 16 år og over i folketellingen og kvalitetskontrollundersøkelsen 1970 i grupper for kjønn, etter samlet arbeidstid i inntektsgivende arbeid. Prosent

	Kvalitetskontroll-undersøkelsen	Folke- og bolig-tellingen 1970
I alt	100,0	100,0
Uten inntektsgivende arbeid	31,6	38,9
Med inntektsgivende arbeid	68,4	61,1
Samlet arbeidstid mindre enn 500 timer	8,9	4,9
Samlet arbeidstid 500 timer og over	59,5	56,2
MENN		
I alt	100,0	100,0
Uten inntektsgivende arbeid	14,1	18,0
Med inntektsgivende arbeid	85,9	82,0
Samlet arbeidstid mindre enn 500 timer	5,9	3,7
Samlet arbeidstid 500 timer og over	80,0	78,3
KVINNER		
I alt	100,0	100,0
Uten inntektsgivende arbeid	48,6	59,2
Med inntektsgivende arbeid	51,4	40,8
Samlet arbeidstid mindre enn 500 timer	11,8	6,1
Samlet arbeidstid 500 timer og over	39,5	34,7

Nå er årsakene til at målefeilene er så store i dette tilfellet naturligvis at folketellings-skjemaet er enkelt i forhold til det det er mulig å gjøre ved hjelp av et spørreskjema og intervjuere. Resultatene i tabell 1 er derfor av begrenset betydning når det gjelder å få innsikt i de målefeil vi kan vente oss i forbindelse med levekårsundersøkelsene.

I forbindelse med fruktbarhetsundersøkelsen i det følgende kalt FU77, sammenlignet en nettoinntekt oppgitt i intervjuundersøkelsen i tabell 2 kalt IFU, med nettoinntekt i skatteregisteret i tabell 2 kalt ISR, Siring (1980). I tabell 2 er det gjengitt noen av resultatene fra denne undersøkelsen.

Tabell 2. Inntektstall fra skatteregisteret, og differansen mellom inntektstall fra fruktbarhetsundersøkelsen 1977 og inntektstall fra skatteregisteret i grupper for alder. Gjennomsnittstall i 1 000 kr

Kvinnens alder	Gifte kvinners egeninntekter			Ektemenns inntekter		
	ISR	IFU-ISR	Antall personer	ISR	IFU-ISR	Antall personer
Alle	15,2	2,5	2 727	62,1	6,0	2 274
18-19 år	10,4	1,6	25	45,0	4,6	6
20-24 "	15,4	2,1	359	46,5	6,8	235
25-29 "	14,8	2,1	712	55,4	6,7	619
30-34 "	14,4	2,6	695	65,5	6,6	617
35-39 "	15,8	3,3	511	68,4	6,1	443
40-44 "	16,6	2,4	425	71,0	2,9	354

De gifte kvinnene oppga i gjennomsnitt 6 000 kroner mer i inntekt for ektemennene sine i FU77 enn det vi har fra skatteregisteret, mens de i gjennomsnitt oppga 2 500 kroner mer i egeninntekt i FU77 enn det de hadde oppgitt til likningsmyndighetene. Det siste tallet er omtrent av samme størrelse som det tilsvarende tallet for de ugifte.

Også andre lignende studier foreligger fra andre undersøkelser, men i likhet med resultatene ovenfor gir de liten innsikt i hvilke effekter målefeil har på studiet av utviklingen over tid, samt hvilke konsekvenser dette bør få for hvordan utvalgene bør koordineres i framtiden.

Når det gjelder å illustrere målefeilsproblemer knyttet til analyse av gjentatte undersøkelser, er resultatene fra folketellingen 1970 samt to "folketellingslignende undersøkelser" i 70-åra nyttig. I 1975 ble det sendt et folketellingskjema til et utvalg av personer, og i 1978 ble det utført en prøveundersøkelse til Folke- og bolig tellingen 1980. I det følgende vil Folketellingen 1970 og de to øvrige undersøkelsene i 70-åra bli kalt tellinger. Definisjonene i 1970, 1975 og 1980 var stort sett de samme av hensyn til sammenlignbarheten. I forbindelse med alle tre tellinger ble det utført en kvalitetskontrollundersøkelse ved hjelp av intervjuere, som gjennom et omfattende spørreskjema kartla sysselsettingsforholdene mer nøyaktig enn tilfellet var i tellingene hvor skjemaene nødvendigvis måtte være enkle. Resultatene fra de tre tellingene med tilhørende kvalitetskontrollundersøkelser, kalt KU i det følgende, er gitt i tabell 3 Haldal (1980). En nærmere analyse av tabell 3 er gitt i Haldal (1980), men her er det tilstrekkelig å konstatere at endringer mellom 1970 og 1978 målt ved hjelp av tellingene er helt forskjellig fra de samme endringer målt ved hjelp av kvalitetskontrollundersøkelsene.

Tabell 3. Personer 16-73 år i grupper for kjønn etter samlet årlig arbeidstid i TELLINGEN-70, KU-70, TELLINGEN-75, KU-75, TELLINGEN-78 og KU-78.

	Alle				Menn				Kvinner									
	TEL- LIN- GEN 70	KU 70	MINI FoB 75	TEL- KU 75	TEL- LIN- GEN 78	KU 78	TEL- LIN- GEN 70	KU 70	TEL- LIN- GEN 75	KU 75	TEL- LIN- GEN 78	KU 78						
Med inntektsgivende arbeid i alt	64,9	72,5	70,1	72,2	72,9	74,7	86,5	90,2	85,5	86,3	88,9	88,8	43,7	55,0	55,7	58,8	58,7	62,0
Samlet arbeidstid																		
1- 99 timer ...	0,6	1,3	1,7	1,6	2,2	.	0,4	0,5	0,7	0,7	1,0	.	0,8	2,1	2,6	2,5	3,3	.
100-499 " ...	4,6	8,2	6,4	8,3	6,0	.	3,4	5,6	4,2	4,9	3,8	.	5,8	10,7	8,5	11,5	8,0	.
500-999 " ..	6,5	7,7	7,0	8,8	7,9	.	3,8	4,8	3,9	4,5	2,2	.	9,2	10,6	10,0	12,9	12,0	.
1 000 timer eller mer	53,2	55,3	54,1	53,4	54,2	.	78,9	79,4	75,7	76,3	78,4	.	27,8	31,5	33,7	31,9	32,5	.
Uoppgitt	-	-	1,0	0,0	3,1	.	-	-	1,1	0,0	3,4	.	-	-	0,8	0,0	2,9	.
Antall svar	5 106		8 293		1 846		2 533		4 021		873		2 573		4 272		973	

Resultatene i tabell 3 gir et godt bilde av de begrensinger som målefeil legger på mulighetene for å studere utviklingen over tid ved hjelp av gjentatte undersøkelser. En bør imidlertid være klar over at skjemaene brukt i de tre tellingene er meget grove sammenlignet med de målemetoder som brukes i forbindelse med levekårsundersøkelsene.

Et av de målefeilsproblemer som er viktige i forbindelse med valg av rotasjonsplan for levekårsundersøkelsene, er om svarene et individ gir på et spørsmål avhenger av om det samme spørsmål ble stilt i en tidligere undersøkelse. I Byrådet har vi ikke studert slike målefeil, men resultater fra U.S.A. tyder på at det finnes slike typer målefeil, f.eks. er det en tendens til å svare arbeidsløs feilaktig på en arbeidskraftundersøkelse, når en tidligere har vært arbeidsløs. Dersom slike typer feil er omfattende også i forbindelse med levekårsundersøkelsene, bør en naturligvis være forsiktig med å beholde individer i utvalget fra en undersøkelse til den neste. Foreløpig vet vi ingenting om slike målefeil, og den eneste måten å eventuelt identifisere og estimere dem på, er å lage roterende utvalg.

En annen fordel med roterende utvalg når det gjelder estimering av målefeil, er at det er konstruert en serie markov-modeller med målefeil, hvor målefeilene er gjort til en eksplisitt del av modellen. For å kunne estimere i slike modeller under hensyntaken til målefeil, er det nødvendig med roterende utvalg. Det er verd å merke seg at for å kunne estimere parametrene i de fleste av disse modeller, er det nødvendig med minst tre gjentakelser av undersøkelsene.

Som konklusjon kan en altså si at målefeil i likhet med utvalgsvariansen legger vesentlige begrensninger på muligheten for å studere utviklingen over tid ved hjelp av gjentatte undersøkelser. Mens vi er i stand til å kontrollere størrelsen på utvalgsvariansen ved hjelp av størrelsen på utvalget, er problemene knyttet til å kontrollere målefeil mye større, og det vil først i framtiden bli mulig å kvantifisere betydningen av målefeil.

Når det gjelder hvilken effekt målefeil bør ha på rotasjonsplanen for levekårsundersøkelsene, har vi til nå ikke greid å identifisere målefeil som på noen måte diskvalifiserer roterende utvalg, snarere tvertimot. Skal vi øke våre kunnskaper om målefeil i forbindelse med levekårsundersøkelsene, og skal levekårsundersøkelsene studeres ved hjelp av modeller hvor målefeilene tas inn som en del av modellen, er det helt nødvendig at en bruker panel eller roterende utvalg.

4.5. Spesielle problemer knyttet til rotasjon av husholdningsundersøkelser

Til nå har levekårsundersøkelsene vært basert på personundersøkelser. Likevel kan det vise seg nødvendig å ta opp noen problemer knyttet til rotasjon av husholdningsundersøkelser på grunn av koordineringen med inntektsundersøkelsene. Vi skal ikke i dette notat ta opp alle spørsmål i forbindelse med rotasjon av husholdningsutvalg, men bare peke på problemene og antyde hvordan en eventuelt kan løse dem. Dersom det blir nødvendig å rotere inntektsundersøkelsene må spørsmålet drøftes nærmere, og en del detaljer i framgangsmåten må diskuteres særskilt.

Problemet med rotasjon av husholdningsundersøkelser skyldes at populasjonen av husholdninger endrer seg over tid. Noen husholdninger går i oppløsning, og andre etableres slik at det er vanskelig til en hver tid å bestemme seg for hvordan en husholdning skal følges opp. På den andre siden er det opplagt at det er mange fordeler knyttet til å beholde noen husholdninger gjennom to eller flere tidspunkter både for analyseformål og for å redusere utvalgsvariansen til endringstallene.

Det er neppe tvil om at rotasjon av husholdningsutvalget til inntektsundersøkelsen vil bety at både utvalgstrekkning og estimering blir forholdsvis komplisert. En annen mulighet er å rotere bare en del av utvalget til levekårsundersøkelsen og samordne den andre delen med inntektsundersøkelsen.

Arbeidet med den konkrete utvalgsplan til levekårsundersøkelsen vil imidlertid bli tatt opp i forbindelse med levekårsundersøkelsen i 1983. Denne utvalgsplanen vil bli presentert i en bearbeidet versjon av det foreliggende notat.

LITTERATUR

- Bielby W., Hauser R. and Foatherman D. (1977): *Reponse errors of black and nonblack males in models of the intergenerational transmission of socioeconomic status*. American Journal of Sociology 82, nr. 6.
- Butler, D.E. and D. Stokes (1974): *Political Change in Britain*, 2nd. ed. London: Macmillan.
- Coleman, J.S. (1964): *Introduction to Mathematical Sociology*. New York: Free Press.
- Heldal, J. (1980): *Proveundersøkelse i Folke- og Boligtellingen 1980*. Rapport fra kvalitetskontrollundersøkelsen. IN 80/3 Statistisk Sentralbyrå.
- Johnston, J. (1972): *Econometric Methods* 2nd. ed. McGraw-Hill.
- Markus, Gregory (1979): *Analyzing panel data*. Sage University Paper Series on Quantitative Applications of the Social Sciences, 07-001, Beverly Hills and London: Sage Publications.
- Ostrom, Charles W. Jr. (1978): *Time Series Analysis: Regression Techniques*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-009. Beverly Hills and London: Sage Publications.
- Siring, Erling og Ib Thomsen (1981): *Metoder for estimering av tall for fylker ved hjelp av utvalgsundersøkelser*. Statistisk Sentralbyrå. Rapporter 81/6.
- Siring, E. (1980): *Sammenlikning av inntektstall oppgitt i fruktbarhetsundersøkelsen 1977 med inntektstall fra skatteregisteret*. IN 80/24 Statistisk Sentralbyrå.
- Wiggins, L.M. (1973): *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. San Francisco: Josey-Bass.