Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 71/8 22. juni 1971

# NOEN FEILKILDER VED STATISTISKE UNDERSØKELSER MED SÆRLIG VEKT PÅ FRAFALL

Av Ib Thomsen

INNHOLD

## 1. Innledning

Hensikten med dette notatet er å gi en oversikt over en rekke viktige feilkilder ved statistiske undersøkelser. Det er lagt særlig vekt på å gruppere de ikke-tilfeldige feil. Bare for to kilders vedkommende skal jeg gå litt i detalj, nemlig utvalgsfeil og frafall.

I tiden framover vil vi identifisere og estimere betydningen av forskjellige feilkilder ved undersøkelser som utføres ved kontoret for intervjuundersøkelser. Hensikten med å studere disse feilene er både å kunne estimere den totale usikkerheten for de publiserte tall og kunne allokere de samlede ressurser fornuftig på de enkelte prosesser i innsamlingsrutinen.

## 2. Gruppering av feilkilder

Ennå finnes det ingen modell, som gir grunnlag for å optimalisere den totale utvalgsplan, med hvilket jeg mener en plan som tar hensyn til alle feilkildene ved en undersøkelse. Eksisterende modeller er enten for summariske [11] eller de behandler en eller to feilkilder isolert [7], [5].

En slik modell ville også gjøre det mulig å estimere den totale varians på de størrelser som publiseres.

I mangel av en slik modell skal jeg derfor gruppere feilkildene og deretter behandle dem hver for seg.

Følgende hovedinndeling er gjort i modellen utviklet ved Bureau of the Census [6], [7].

La U være en ukjent parameter, som vi ønsker å estimere ved vår telling eller utvalgsundersøkelse. La x være estimator for U. Da defineres det "totale middelverdiavvik" som

(1) $E(x-U)^2$ = Samplingvariansen + variansen som skyldes ikke tilfeldige feil + kovariansen mellom sampling og ikke-samplingfeil + $(\text{skjevhet})^2$.

Første ledd i summen ovenfor avhenger av utvalgsplanen og kan vanligvis inndeles i relevante komponenter. Estimeringen av disse utgjør sjelden et alvorlig problem sammenlignet med problemet å estimere de øvrige ledd av summen.

Ved en totaltelling er første ledd i summen naturligvis lik null.

Kovariansleddet i (1) skal jeg ikke komme nærmere inn på her, den er behandlet i bl.a. [5].

Nedenfor skal jeg konsentrere meg om de ikke tilfeldige feil. Disse kan med fordel inndeles i 3 hovedklasser:

1. Registerfeil og utvalgsfeil
2. Frafall
3. Målefeil

## 3. Registerfeil og utvalgsfeil

Når en trekker enheter til en statistisk undersøkelse er målet at alle enheter i en populasjon skal ha en kjent sannsynlighet for å komme med i utvalget. (Ved en totaltelling ønsker en utvalgssannsynligheten lik 1.) Men på grunn av manglende opplysninger om populasjonen, f.eks. dårlige registre, krever det både fantasi og mye arbeid å nå dette målet. Problemet har inntil nå vært lite behandlet i litteraturen, men en finner en god framstilling av problemet og mange løsningsforslag i [9]

Men det er ikke bare feil ved registeret som kan føre til skjevheter i utvalget. Mange feltrutiner er lagt opp slik at visse grupper av personer ikke har noen sjanse for å komme med i utvalget. Jeg skal nedenfor gi et eksempel på dette.

I Statistisk Sentralbyrå brukes ved personundersøkelser følgende utvalgsplan.

Fra registeret trekkes personer med navn og adresse. Men på grunn av flyttinger, som ikke er kommet med i registeret skjer det at når intervjueren når fram til adressen er det flyttet en ny familie inn. Tidligere ville en da trekke en ny person tilfeldig fra populasjonen som

erstatning for den flyttede. Denne regel er et eksempel på hvorledes en kan innføre skjevheter i utvalget under selve feltarbeidet, idet vi på denne måte utelater alle personer som nylig er flyttet.

## 4. Frafall

### 4.1. Årsaker til frafall

En feilkilde, som har mye til felles med den ovenstående er frafall. Et IO som tilhører den relevante populasjon og som er valgt ut for å delta i en undersøkelse regnes som frafall dersom samtlige kjennetegn hos vedkommende i undersøkelsen av en eller annen grunn ikke blir målt.

For å kunne si noe om størrelsen på frafallet og dets effekt på viktige resultater i undersøkelsen er det vanlig å inndele det etter årsaken til frafallet. En god klassifikasjon av frafallet avhenger av undersøkelsens art, men det nedenstående er nok blitt en temmelig standardisert inndeling i mange typer undersøkelser, kanskje først og fremst intervjuundersøkelser. [9].

1. Nektere
2. Ikke å treffe
3. Sykdom
4. Annen årsak

a. Nektere

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

Det er viktig å være oppmerksom på at disse på ingen måte er uforanderlige. Der finnes måter å motivere IO på f.eks. ved å sende et brev i forveien o.l. En annen viktig ting er at en person kan nekte fordi tidspunktet for besøket passer dårlig eller fordi tilliten til intervjueren ikke er stor nok. Ved valgundersøkelsen 1969 ble antall nektere redusert med ca. 20 % ved å bruke postal innsamling etter at intervjueren var blitt nektet intervju [13].

En må dog regne med at selv om en gjør store anstrengelser for å redusere gruppen av nektere, vil det alltid bli igjen en "hard kjerne", som det ikke er mulig å få tak i uten bruk av eventuelle lovbestemmelser,

b. Ikke å treffe

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

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

c. Sykdom

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

d. Annen årsak

Til den kategorien hører en rekke frafallsårsaker. En av de viktigste hos oss er sykdom eller ferie blant intervjuerne, slik at en ikke innen den fastsatte tiden eller ressursramme kan nå fram til alle uttrukne IO.

## 4.2. Effekter av frafall

For å studere effekten av frafall skal vi tenke oss populasjonen inndelt i to "strata".

Det ene stratum består av de personer som ville komme med i utvalget dersom de ble trukket og det andre stratum består av de personer som ikke ville komme med i utvalget selv om de ble trukket ut.

En slik inndeling er naturligvis litt for enkel, men er allikevel en måte å belyse problemet på.

Utvalget gir ingen informasjon om tilstanden i det annet stratum, men dersom en kan forutsette at tilstanden i stratum 2 er lik tilstanden i stratum 1, er naturligvis problemet løst. På den andre siden viser erfaringen oss at dette meget sjelden er tilfelle [13], [9].

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

$$(1) \quad E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} = \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2) = W_2(\bar{Y}_1 - \bar{Y}_2)$$

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

Bruttovariansen for $\bar{y}_1$ blir

$$E(\bar{y}_1 - \bar{Y})^2 = E(\bar{y}_1 - \bar{Y}_1 + \bar{Y}_1 - \bar{Y})^2$$

$$= E(\bar{y}_1 - \bar{Y}_1)^2 + E(\bar{Y}_1 - W_1\bar{Y}_1 - W_2\bar{Y}_2)^2$$

$$(2) \qquad + 2E(\bar{y}_1 - \bar{Y}_1)(\bar{Y}_1 - \bar{Y})$$

$$= \frac{S_1^2}{n_1} + W_2^2(\bar{Y}_1 - \bar{Y}_2)^2$$

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

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

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

Nedenfor skal vi se hvorledes (2) kan brukes til å bestemme en fornuftig allokering av de samlede ressurser på å redusere frafallet.

I tillegg til denne effekten kommer at frafallet ofte er ujevnt fordelt på samtlige strata. Dette medfører at hvert stratum ikke har det antall observasjoner som utvalgsplanen tilsier. Effekten av dette finner en beskrevet i [9].

## 4.3. Publisering av frafall

Det er vanlig å publisere opplysninger om frafall ved å gi det totale frafall og fordele dette etter årsak og kanskje gi en aldersfordeling for frafallet. Etter min mening er dette ikke tilfredsstillende av to årsaker

1. Data om frafall står foran i publikasjonene og er derfor vanskelig å se i sammenheng med tabellverket forøvrig.
2. Det er sjelden å se frafallsdata, som kan hjelpe en under vurderingen av påliteligheten av publiserte størrelser. Med dette menger jeg at en sjelden får oppgitt frafallet spesifisert slik at (1) og (2) kan brukes av de som måtte ønske det.

Følgende eksempler vil belyse hva jeg mener.

Eksempel 1:

Tabell NN. Menn i alderen 20-25 år etter om de tilhører arbeidskraften eller ikke. Prosent

| | I arbeidskraften | Ikke i arbeidskraften | Total antall menn i utvalget | Frafall i prosent |
|---|---|---|---|---|
| Menn i alder 20-25 år | 84 | 16 | 2 087 | 25 |

Av tabellen ovenfor er en i stand til å finne ut frafallets innvirkning på tabellerte størrelser direkte.

Hvis en ikke ønsker frafallet med i hver tabell (det er nesten aldri mulig å finne frafallstall for samtlige tabeller i et tabellverk), må en analysere frafallet i forordet med sikte på at en skal kunne vurdere effekten av frafallet på de publiserte størrelser (hvor dette er mulig). Når jeg leser frafallsanalyser har jeg ofte inntrykk av at disse er tatt med for "å legge kortene på bordet".

## 4.4. Måter å redusere frafallet på

Når en skal behandle problemer med frafall i en undersøkelse, kan en angripe dem fra to sider, nemlig redusere størrelsen på frafallet

og/eller forsøke å redusere virkningen av frafallet. Vi skal ta for oss problemet med å redusere størrelsen på frafallet først og se på det andre problemet nedenfor. Følgende metoder kan brukes for å redusere størrelsen på frafallet:

4.4.1. En kan forbedre innsamlingsprosedyren generelt ved å

a) Garantere anonymitet overfor IO og kanskje gi dette anledning til å sende svaret direkte til Statistisk Sentralbyrå. En annen måte å sikre anonymiteten på, som vi ikke har forsøkt i Norge, er foreslått i [15]. Metoden består av at intervjueren gir et spørsmål, f.eks. "Snyter De i skatt?" I stedet for å forlange svar på spørsmålet, ber intervjueren IO om å trekke et kort fra en "velblandet" kortstokk hvor det står Ja eller Nei. IO blir da bedt om å opplyse om det som står på kortet er sant eller usant. Når en kjenner fordelingen på Ja- og Nei-kort i stokken, kan en deretter estimere hvor stor prosent som snyter i skatt.
Metoden er utviklet videre slik at det er mulig å bruke den ved flere svaralternativer enn to.

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

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

d) Sende bud i forveien til IO om at det vil bli oppsøkt av en intervjuer. (En bør her være forsiktig med opplysninger om undersøkelsens art, da en kan påvirke IO's adferd og dermed resultatet av undersøkelsen.) [14].

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

4.4.2. Gjenbesøk

Gjenbesøk eller purring er den mest alminnelige måten å redusere frafallet på, og da spesielt det frafall som skyldes "ingen å treffe". Problemet i denne forbindelsen er å avgjøre hvor mye ressurser en skal sette inn på gjenbesøk, skal en velge et stort utvalg og bare gå en gang, eller skal en velge et mindre utvalg og satse mere tid og penger på å redusere frafallet.

I [3] er gitt en modell, som skal hjelpe til ved løsningen av dette problemet. Jeg skal ikke komme inn på denne modellen her, men heller anvende resultatene fra avsnitt 4.2. Vi fant da

$$(3) \qquad E(\bar{y} - P)^2 = \frac{P_1(1 - P_1)}{M_1} + W_2^2(P_1 - P_2)^2$$

hvor $\bar{Y}_1$ og $\bar{Y}_2$ er erstattet med $P_1$ og $P_2$ henholdsvis, i det vi tenker oss at vi observerer en binær variabel.

Hvis (3) skal reduseres ved gjenbesøk er det ikke først og fremst økningen i n, som betyr noe, men en eventuell redusering av $W_2^2(P_1 - P_2)^2$. Dette kunne en tenke seg ved å redusere $W_2$ alene og forutsette at $(P_1 - P_2)^2$ ikke vokser, eller en kunne, hvis en hadde et godt kjennskap til populasjonen, oppsøke frafallsgrupper med en meget atyprisk adferd og derved redusere både $W_2$ og $(P_1 - P_2)$. Det typiske er dog at en arbeider på å redusere $W_2$. Dvs. at en ut fra erfaringer om svarprosenter og omkostninger ved gjenbesøk kan vurdere kostnader ved gjenbesøk og sette dette opp mot gevinsten estimert ved (3).

Eksempel:

Ved å velge 2 gjenbesøk istedet for 1 kan vi tenke oss at $W_2$ erfaringsmessig endrer seg fra 19 % til 16 %. En estimator for gevinsten er da $0{,}19^2(P_1 - P_2)^2 - 0.16^2(P_1 - P_2)^2 = 0{,}01(P_1 - P_2)^2$, hvor $(P_1 - P_2)$ forutsettes uendret. Reduksjonen i bruttovarians er altså maximalt lik 1 %.

Eksempel:

I valgundersøkelsen 1969 fant vi $P_1 - P_2 = 0{,}05$, altså liten forskjell. $P_1$ som er stemmefrekvensen i utvalget var 0,90. Frafallet var 9,9 %.

$$E(\bar{y} - \bar{Y})^2 = \frac{0{,}90 \cdot 0{,}10}{2702} + (0{,}099)^2 \cdot (0{,}05)^2$$

$$= 0{,}000033 + 0{,}000025$$

$$= 0{,}000058$$

Den delen av bruttovariansen som skyldes frafall utgjør altså ca. 43 %.

Hvis en her øker antall IO til 4 000 eller $n_1$ til $4000 \cdot 0{,}099$ vil andre leddet i bruttovariansen ikke endre seg, mens det første leddet

blir 0,000025 og bruttovariansen dermed 0,000050. Altså en meget liten reduksjon. Hvis en derimot brukte tid og penger på å redusere $W_2$ f.eks. til 0,08 ville bruttovariansen bli redusert til 0,000049.

#### 4.4.3. Utvalgsundersøkelse på frafallet

Dersom det er uøkonomisk å foreta gjenbesøk fordi frafallet er meget stort og/eller meget spredt, kan det komme på tale å utføre utvalgsundersøkelse på frafallet. Denne metoden var først brukt i forbindelse med en postundersøkelse fulgt opp av intervjuere på et utvalg av frafallet. Problemet å bestemme optimal utvalgsstørrelse og hvor stor del av frafallet som skal følges opp er behandlet i [7].

#### 4.4.4. Erstatninger

En ser ofte at frafallet reduseres ved å erstatte dette med andre IO. Det ekstreme eksempel er kvotasampling, hvor en holder på til en finner noen hjemme. (Det er vel et spørsmål om en kan tale om frafall ved slike utvalgsplaner.) Ved å erstatte er det klart at utvalgsstørrelsen øker, og en derved reduserer første ledd av bruttovariansen. Reduksjonen er dog minimal (se eksempel under pkt. 4.4.2.). Skal en redusere andre leddet av bruttovariansen er erstatninger verdiløse.

### 4.5. Måter å redusere effektene av frafallet på

Selv etter store anstrengelser for å redusere størrelsen på frafallet, vil dette alltid ha en viss størrelse etter at innsamlingen er avsluttet. I visse undersøkelser er frafallet så stort at en må gjøre noe for å redusere effekten av det.

#### 4.5.1. Veiing av observasjonene

Veiing av delutvalg omvendt proposjonalt med sjansen for å være med i utvalget er aktuelt når frafallet er ulikt fordelt på grupper av befolkningen og disse grupper er meget forskjellige. For å belyse effekten av veiing skal vi se på en populasjon inndelt i to delpopulasjoner, som for enkelhets skyld er like store. Vi tenker oss at vi utfører en utvalgsundersøkelse for å finne hyppigheten av et bestemt kjennetegn P. La $P_1$ og $P_2$ være hyppigheten i delgruppe 1 og 2 henholdsvis. Vi skal

dessuten anta at vi ikke har frafall i delgruppe 2, men at vi i delgruppe 1 har frafall. For delgruppe 1 har vi altså en lignende modell som ovenfor.

Vi har da

$$P = \tfrac{1}{2}(P_1 + P_2)$$

$P_1 = W_1P_1' + W_2P_1''$, hvor $P_1'$ er hyppigheten i den del av delgruppe 1, $W_1$, vi får tak i og $P_1''$ er hyppigheten i den del av delgruppe 1, $W_2$, vi ikke får tak i ved den valgte innsamlingsteknikk.

Det følger nå

$$P = \tfrac{1}{2}\left[(W_1P_1' + W_2P_1'') + P_2\right],$$

som er den ukjente estimand vi ønsker å estimere.

En uveiet estimator for P er

$$\hat{P} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n} y_i}{n_1 + n}$$

hvor n er antall observasjoner fra hver delgruppe. (Om n tenkes fremkommet ved først å stratifisere populasjonen og deretter trekke n fra hver delgruppe, eller om n er resultat av poststratifering er underordnet her.)

$n_1 = W_1 \cdot n$, altså størrelsen på utvalget fra delgruppe 1.

$x_i = 1$ hvis $i^{te}$ utvalgsmedlem fra delgruppe 1 har det bestemte kjennetegn

$x_i = 0$ ellers

og $y_i$ definert som $x_i$ for delgruppe 2.

Ved å bruke samme metode som under punkt 4.2 finner vi nå

$$E(\hat{p} - P) = P_1'\left(\frac{W_1}{W_1+1} - \frac{W_1}{2}\right) - \frac{W_2}{2}P_1'' + P_2\left(\frac{1}{W_1+1} - \tfrac{1}{2}\right)$$

og

$$E(\hat{p} - P)^2 = \frac{W_1^2}{(W_1+1)^2}\left[\frac{P_1'(1-P_1')}{W_1 n} + W_2^2(P_1' - P_1'')^2\right] + \left(\frac{1}{W_1+1}\right)^2 \frac{P_2(1-P_2)}{n}$$

En ser at selv når $P_1' = P_1''$, altså når frafallet i delgruppe 1 ikke adskiller seg fra resten av delgruppe 1, er ikke $\hat{p}$ forventningsrett og at skjevheten øker når forskjellen mellom $P_1$ og $P_2$ øker.

La nå $$\hat{u} = \frac{\frac{1}{W_1}\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n} y_i}{\frac{1}{W_1} n_1 + n} .$$

Da er $\hat{u}$ framkommet ved å veie observasjonene fra delgruppe 1 og en får

$$E(\hat{u} - P) = \tfrac{1}{2} W_2 (P_1' - P_1'')$$

og $$E(\hat{u} - P)^2 = 1/4 \left[ \frac{P_1'(1-P_1')}{n_1} + W_2^2 (P_1' - P_1'')^2 + \frac{P_2(1-P_2)}{n} \right]$$

En ser at skjevheten nå bare avhenger av $(P_1' - P_1'')$ og $W_2$. Denne estimator kan altså med fordel brukes når $P_1' - P_1''$ er liten og $P_1 - P_2$ er stor.

Sammenligning mellom bruttovariansene er noe mer komplisert. I [9] vil en finne at variansen øker ved en slik veiing, men av det ovenstående følger at bruttovariansen ikke nødvendigvis øker ved veiing.

Eksempel:

La $P_1' = 0{,}50$

$P_2'' = 0{,}54$

$W_2 = 0{,}50$

Vi får da:

| | Skjevhet | Brutto-varians |
|---|---|---|
| $\hat{u}$ | 0,01 | 0,00019 |
| $\hat{p}$ | 0,023 | 0,00012 |

### 4.5.2. Politz's metode

Hensikten med denne metoden er å redusere antall gjenbesøk og i stedet korrigere observasjonene med sannsynligheten for å få svar.

IO blir spurt om i hvor mange av k liknende tidsperioder han/hun ville ha vært hjemme. Hvis svaret er r blir observasjonen veid med $(k + 1)/(r + 1)$. Ofte velges $k = 5$. Hvis IO svarer at han var til stede i $r = 1$ tilfelle får han vekten $6/2$. Metoden er beskrevet i [12].

Merk to viktige forutsetninger: For det første må vi forutsette at opplysninger om $r$ er korrekte, for det andre er vektene avhengig av $k$ og den tidsperioden på dagen en velger.

I [4] har en gjort en sammenligning mellom veide resultater og resultater etter gjenbesøk. En fant her at de veide resultater lignet mere på resultatene etter første besøk enn resultatene etter gjenbesøk.

### 4.5.3. Bartholomew's metode

Hensikten med denne metoden er å redusere antall besøk til to og veie resultatene fra de to besøk sammen.

Anta at vi ønsker å estimere hyppigheten av forekomsten av et kjennetegn i det utvalg som opprinnelig er trukket ut. La hyppigheten være P og den opprinnelige utvalgsstørrelse være N. Vi ønsker en estimator for P basert på de to første besøk. La antall hjemme ved første besøk være $N_1$, la $n_1$ være antallet av disse som har det bestemte kjennetegn og la $N_2$ og $n_2$ være de tilsvarende tall ved annet besøk.

Da er
$$P = \frac{N_1}{N} P_1 + (1 - \frac{N_1}{N}) \cdot P_r$$

hvor
$$P_r = \frac{(NP - n_1)}{N - N_1} \quad \text{og} \quad P_1 = \frac{n_1}{N_1}$$

$N_1$, $N$ og $P_1$ er kjent etter første besøk og vi mangler kun en estimator for $P_r$, som er hyppigheten av forekomsten av kjennetegnet i den del av utvalget som ikke ble kontaktet ved første besøk.

Hvis vi forutsetter at $n_2$ er et tilfeldig utvalg av $N_2$ er en naturlig estimator for $P_r$, $\hat{P}_r = n_2/N_2$, hvorav en dermed har en estimator.

$$\hat{P} = \frac{N_1}{N} P_1 + (1 - \frac{N_1}{N}) \hat{P}_r.$$

Denne estimator er foreslått av Bartholomew [1], som også har gjort visse empiriske studier, som ser svært lovende ut.

Den viktigste forutsetning er at $n_2$ er et tilfeldig utvalg av $N_2$. Denne forutsetning virker sikkert rar på mange. I [15] er estimatoren vist å være mindre skjev enn den vanlige under mere realistiske forutsetninger.

## 5. Målefeil

Med målefeil menes her forskjellen mellom den korrekte verdi $\zeta_i$ for intervjuobjekt i og den verdi $y_i$ som blir brukt ved beregning av de publiserte størrelser. Det er natulig å dele opp denne størrelsen i feil som er gjort under selve registreringen (med eller uten intervjuer), og feil som er gjort under koding og revisjon. En kunne ta med databehandlingsfeil, men jeg har valgt å utelate denne komponenten isolert.

I fysikk, kjemi og biologi har en lenge arbeidet med målefeil. En mengde informasjon er samlet inn om både måleinstrumenter, de personer som måler og sammenhengen mellom dem. Adskillig mindre er kjent når det gjelder å registrere svar, som er gitt verbalt fra et intervjuobjekt til en intervjuer. Da samspillet her spiller en stor rolle, kan feilanalysene blir meget kompliserte. I Norge har vi kun ganske få kunnskaper om disse feiltypene [14], men vi håper at kontrollundersøkelsen av Folketellingen 1970 skal gi oss en del opplysninger på dette feltet.

I de senere år er det utviklet en rekke modeller for studiet av målefeil [5], [6], [10].

### 5.1. Registreringsfeil

Med registreringsfeil menes de avvik fra den "sanne" verdi som innføres under selve registreringen (intervjuingen).

Følgende matematiske modeller brukes ofte ved studier av målefeil. Vi antar at vi utfører et stort antall repetisjoner av målingen av $i^{te}$ individ. La $y_i$ være måleresultatet ved måling nummer $\alpha$.

Da antar vi

$$Y_{i\alpha} = \zeta_i + g_i' + c_{i\alpha}$$

hvor

$\zeta_i$ er den korrekte verdi (forventningen tatt over alle målinger).

$g_i'$ er en skjevhetskomponent

$c_{i\alpha}$ stokastisk variabel med forventning 0.

Vi kan gjøre forskjellige forutsetninger om fordelingen for $c_{i\alpha}$. $g_i'$ kan være konstant og vi kan forutsette uavhengighet mellom målingene. Men det mest alminnelige er en mer kompleks sammenheng mellom målefeielene som f.eks. intervjuereffekt. For videre studier av dette henvises til [5], [10].

### 5.2. Kodings- og revisjonsfeil

Med kodings- og revisjonsfeil menes de feil som innføres under koding og revisjon. Den matematiske modell ovenfor kan også brukes her og i stedet for en intervjuereffekt får en her en kode- og revisjonsleder-effekt. Estimeringen av denne kan gjøres på samme måte som en estimerer intervjuereffekt. Arbeidet med randomisering er dog vesentlig lettere ved estimering av kodings- og revisjonseffekten enn ved estimering av intervjuereffekten.

## 6. Sluttord

En rekke problemer innen estimering av størrelsen på de ikke tilfeldige feil er fortsatt uløste. Særlig gjelder dette problemer omkring målefeil. Det finnes dog flere metoder for å estimere en del av målefeilene. I tiden framover må vi regne med at flere holdningsspørsmål vil inngå i undersøkelsene våre, og da vil målefeilene utgjøre en større del av bruttovariansen enn de har gjort i de undersøkelser vi har utført til nå. Mens feilkilder som utvalgsfeil og frafall har vært underkastet analyse og behandling til nå ved kontoret, er det rimelig at analyse av målefeil vil bli behandlet i tiden som kommer.

## 7. Litteratur

[1] Bartholomew, D.J. (1961). A method of allowing for "not-at-home" bias in sample surveys. Applied statistics, 52-59.

[2] Dalenius, Tore (1957). Sample in Sweden. Almquist & Wiksell, Stockholm.

[3] Deming, W.e. (1953). On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponsen. J.A.S.A. 48 743-772.

[4] Durbin, J. and Stuart, A. (1954). Call-backs and clustering in sample surveys. JRSS (A) 387-428.

[5] Fellegi, J.p. (1964). Response variance and its estimation. Journal of the American Statistical Association.

[6] Hansen, M.H. and Tepping, B.j. (1969). Progress and problems in surveys methods and theory illustrated by the work of The United States Bureau of the Census. New developments in survey sampling. Wiley-Interscience, New York.

[7] Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. J.A.S.A. 41 517-529.

[8] Hansen, M.H., Hurwitz, W.N. and Pitzker, L. (1967). Measurement errors and statistical standards in the bureau of the census.

[9] Kish, Leslie (1963). Survey Sampling. John Wiley & Sons, Inc.

[10] Koch, G.G. and Horuitz, D.G. (1969). The effect of response errors on measures of association. New Developments in Survey Sapling. Wiley-interscience. New York.

[11] Nordbotten, Svein (1957). On errors and optimal allocation in a census. Skandinavisk aktuarietidsskrift. 1-10.

[12] Politz, A. and Simmons, W.R. (1949). An attempt to get the "not-at-homes" into sample without call-backs. J.A.S.A.

[13] Thomsen, Ib (1971). On the effect of non-response in the Norwegian election survey 1969. Statistisk Tidsskrift.

[14] Thomsen, Ib (1971). Ikke-tilfeldige feil i valgundersøkelsen 1969. Arbeidsnotat. Statistisk Sentralbyrå, Oslo.

[15] Thomsen, Ib (1971). Some comments to Bartholomew's method of allowing for "not-at-home" bias in sample surveys. Under publisering.

[16] Warner, S.l. (1965). Randomized response: A Survey Technique for Eliminating Evasive Answer Bias. JASA.