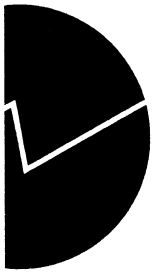
Statistics Norway



Statistisk sentralbyrå

Egil Heilund

Utvalgstrekking, usikkerhetsberegning og frafallsbehandling i inntekts- og kostnadsundersøkelsen for tannleger og fysioterapeuter

Avdeling for næringsstatistikk/Seksjon for Inntekts- og lønnsstatistikk

Innholdsfortegnelse

1. Innledning

Seksjon 420 har gjennomført inntekts- og kostnadsundersøkelser for privatpraktiserende tannleger og fysioterapeuter for henholdsvis inntektsåret 1993 og 1994. Tannlegeundersøkelsen ble gjennomført på oppdrag fra Den norske tannlegeforening og Administrasjonsdepartementet. Fysioterapeutundersøkelsen ble gjennomført på oppdrag fra Norske fysioterapeuters forbund, Kommunenes sentralforbund, Sosial- og helsedepartementet og Administrasjonsdepartementet. Undersøkelsene var forholdsvis like i innhold og form. Begge undersøkelsene tok utgangspunkt i de respektive foreningers medlemsregister, hvorfra det ble trukket et representativt utvalg med hensyn på alder, geografi og praksisform. Tannlege-undersøkelsen er dokumentert i sin helhet av Anne Sørbråten (95/43), og fysioterapeutundersøkelsen av Ann Cathrin Steen (publiseres i desember 1995).

Dette notatet tar spesielt for seg de metodene som er benyttet ved utvalgstrekking, frafallsbehandling og usikkerhetsberegning i de to undersøkelsene. Utvalgstrekkingen og usikkerhetsberegningene er gjort i SAS. Programmene er laget som makroer, og disse kan lett benyttes av andre i analoge situasjoner. Parametre som kreves for å kalle opp makroene beskrives i notatet, og eksempler på programkjøringer er tatt med.

Makroene er tilgjengelige på Unix, på området **$FELLES/sasprog**. Ved hjelp av makroene kan man på en enkel måte trekke et utvalg fra et trekkegrunnlag, eller beregne utvalgsvariansen for et realisert utvalg. Inputfilene (trekkegrunnlaget eller utvalgsfilen) må være på SAS-format. Det er ikke nødvendig med avanserte kunnskaper i SAS for å ta i bruk makroene, to programlinjer i SAS er alt som skal til for å trekke et utvalg eller gjennomføre usikkerhetsberegninger. Det er imidlertid viktig å være klar over hvilke situasjoner makroene kan håndtere, og hvilke parametre som kreves for å kalle opp de ulike makroene.

I **kapittel 2** er det beskrevet hvilke situasjoner makroene kan håndtere, og her forklares hvordan makroene kalles opp. I dette kapittelet er også gitt en kort oversikt over hvilke parametre som kreves for å kalle opp makroene. I **kapittel 3** beskrives makroene for utvalgstrekking, *Trek_sys* og *Trek_pps,* mer i detalj. Eksempler på kjøringer er tatt med. **Kapittel 4** tar for seg hvordan utvalgsvariansen i et enkelt tilfeldig utvalg kan estimeres ved hjelp av makroen *Des_var*, og hvilke forutsetninger som ligger til grunn for beregningene. I **kapittel 5** presenteres hvordan informasjon om ulike frafallstyper er benyttet til å korrigere populasjonen i de to undersøkelsene. De matematiske formlene som er benyttet ved usikkerhetsberegningene er gjengitt i **kapittel 6**. Dataflyt-diagram og SAS kildekode for makroene er tatt med som vedlegg.

2. Makroer for utvalgstrekking og usikkerhetsberegninger

2.1 Situasjoner som kan håndteres

Makroene har et begrenset bruksområde, og særlig gjelder dette makroen for usikkerhetsberegning. Utvalgsvariansen kan i utgangspunktet kun beregnes for et enkelt tilfeldig utvalg, men med noen forutsetninger kan variansen også anslås når man har foretatt et systematisk trekk. Dette blir nærmere omhandlet i kapittel 4 og i Appendiks A. Utvalg kan trekkes som et enkelt tilfeldig utvalg, som et systematisk trekk, eller med trekksannsynlighet proporsjonalt med størrelsen på en angitt variabel. For alle de nevnte situasjonene kan man ha ett eller flere strata. I tabellen er ulike typer utvalg listet opp sammen med hvilken makro som eventuelt kan brukes til utvalgstrekking og usikkerhetsberegning.

Utvalg	Makro ved utvalgstrekking	Makro ved usikkerhetsberegning
Enkelt tilfeldig utvalg, ett strata	Trek_sys	Des_var
Enkelt tilfeldig utvalg, flere strata	Trek_sys	Des_var
Systematisk utvalg, ett strata	Trek_sys	Des_var (med forutsetninger)
Systematisk utvalg, flere strata	Trek_sys	Des_var (med forutsetninger)
Proporsjonal trekksannsynlighet, ett strata	Trek_pps	Nei
Proporsjonal trekksannsynlighet, flere strata	Trek_pps	Nei
Andre typer utvalg (totrinnsutvalg, panel, etc.)	Nei	Nei

Selv i en situasjon der usikkerhetsberegninger i prinsippet kan utføres ved hjelp av *Des_var*, må makroen ikke brukes ukritisk. Det er kun mulig å måle den delen av usikkerheten som skyldes utvalgsstørrelse og stratifisering. Det blir ikke tatt hensyn til andre feilkilder som målefeil, registerfeil og skjevheter i frafallet. I tillegg vil dataene ofte bli revidert på en måte som vanskeliggjør usikkerhetsberegninger.

2.2 Kjøring av SAS-makroene

Filene *trek_sys.sas*, *trek_pps.sas* og *des_var.sas* inneholder kildekoden til de tre makroene. Filene ligger på området **$FELLES/sasprog**. Her ligger også dokumentasjonsfiler, henholdsvis *trek_sys.txt*, *trek_pps.txt* og *des_var.txt*. Før en makro kan kjøres, må man utføre en %INCLUDE-setning i SAS som 'henter' kildekoden. Dette gjøres slik:

```
%INCLUDE '$FELLES/sasprog/trek_sys.sas'
```

Når %INCLUDE-setningen er utført, kan makroen kjøres. Dette gjøres på denne måten: (en oversikt over parametrene i de tre makroene er gitt i neste avsnitt):

```
%Trek_sys (Fil_i, Fil_u, Ident, J, Gruppe, Fast, Z);
```

Hvilke parametre som kreves, vil avhenge av hvilken makro som kjøres. For utvalgsmakroene (*Trek_sys* og *Trek_pps*) vil programmet selv be om trekksannsynligheter (eller utvalgsstørrelse) i hvert stratum. Det er derfor ikke nødvendig å angi utvalgsstørrelser når selve makroen kalles opp.

2.3 Parametrene i *Trek_sys*, *Trek_pps* og *Des_var*

På neste side gis en kort oversikt over hvilke parametre som kreves for å kjøre de tre makroene. En mer detaljert beskrivelse er gitt i kapittel 3 og 4, sammen med eksempler på bruken av makroene.

%Trek_sys (Fil_i, Fil_u, Ident, J, Gruppe, Fast, Z);

Fil_i	Navn på innfilen (trekkegrunnlag)
Fil_u	Navn på utfilen (som vil bestå av de uttrukne observasjonene)
Ident	Navn på unik variabel. Numerisk eller alfanumerisk
j	Navn på stratumvariabel. Numerisk eller alfanumerisk. Settes til 0 dersom kun ett stratum
Gruppe	Sorteringsvariabel. Dersom dette ikke er aktuelt, oppgis 0. Når Gruppe settes til 0 og trekkegrunnlaget er tilfeldig sortert på forhånd, vil vi få et enkelt tilfeldig utvalg Sorteringsvariabelen kan være numerisk eller alfanumerisk
Fast	Settes til 0 (utvalgsstørrelsen skal angis i prosent) eller 1 (absolutte tall). Programmet vil selv be om utvalgsstørrelser eller trekksannsynligheter under kjøringen
Z	Startverdi

%Trek_pps (Fil_i, Fil_u, Ident, J, Prop, Fast, Z);

Fil_i	Navn på innfilen (trekkegrunnlaget)
Fil_u	Navn på utfilen (som vil bestå av de uttrukne observasjonene)
Ident	Navn på unik variabel. Numerisk eller alfanumerisk
j	Navn på stratumvariabel. Numerisk eller alfanumerisk. Settes til 0 dersom kun ett stratum
Prop	Variabelen som det skal trekkes proporsjonalt med hensyn på. Numerisk
Fast	Settes til 0 (utvalgsstørrelsen angis i prosent) eller 1 (absolutte tall). Programmet vil selv be om utvalgsstørrelser eller trekksannsynligheter under kjøringen
Z	Startverdi

%Des_var (Fil, J, V, X, Delp);

Fil	Utvalgsfil
J	Stratumvariabel. Settes lik 0 dersom kun ett stratum. Numerisk eller alfanumerisk
V	Navn på variabel som inneholder vektene. Numerisk
X	Navn på variabelen som skal analyseres. Numerisk
Delp	Navn på eventuell delpopulasjon. Numerisk eller alfanumerisk. Dersom dette ikke er aktuelt, oppgi 0

3. Utvalgstrekking

3.1 Enkelt tilfeldig utvalg og systematisk utvalg

Makroen *Trek_sys* utfører i utgangspunktet et systematisk utvalg, men kan også brukes til å trekke et enkelt tilfeldig utvalg. Forskjellen på tilfeldig og systematisk utvalg kan kort beskrives slik:

Tilfeldig utvalg: Hvis man lager en lapp for alle enhetene i populasjonen, legger lappene i en hatt og trekker 100 lapper, har vi et tilfeldig utvalg med utvalgsstørrelse 100. Skal man imidlertid trekke f.eks. fra folkeregisteret eller bedriftsregisteret, kan det være vanskelig å finne store nok hatter. En enkel måte å trekke et tilfeldig utvalg maskinelt er å tildele hver observasjon i populasjonen et tilfeldig tall r mellom 0 og 1, og så velge de 100 observasjonene som har lavest r. En slik rutine kan lett lages i SAS ved å bruke funksjonen `RANUNI()`.

Systematisk utvalg: Ordet *systematisk* kan virke noe misvisende, fordi et systmatisk utvalg også blir trukket tilfeldig. Selve prinsippet for trekkingen blir imidlertid noe annerledes. La oss si vi ønsker å ta et utvalg av størrelse 25 prosent av populasjonen. Når vi skal trekke systematisk, trekker vi hver fjerde. observasjon i populasjonen, etter først å ha trukket et tilfeldig startpunkt mellom 1 og 4. Trekker vi startpunktet 2, består utvalget vårt av observasjon nummer 2, 6, 10, 14, .. osv. En slik rutine er imidlertid noe vanskeligere å lage i SAS, i hvertfall på generell form.

Hensikten med å benytte systematiske utvalg er å sikre representativitet i større grad enn det en oppnår ved tilfeldige utvalg. Dersom en f.eks. sorterer observasjonene i hvert stratum etter fylke før en foretar et systematisk utvalg, er man sikret tilnærmet samme geografiske fordeling i utvalget som i populasjonen. Systematiske utvalg kan betraktes som en slags 'understratifisering', der enhetene innenfor ett stratum i de ulike 'understrataene' (fylkene) blir trukket med lik trekksannsynlighet. Denne måten å trekke på kan dermed begrense antall fysiske strata. En annen grunn til at systematiske utvalg blir brukt, er at den praktiske trekkingen er lett å utføre manuelt ved hjelp av et register. For eksempel blir meningsmålinger ofte gjennomført ved at man velger et tilfeldig startpunkt i telefonkatalogen, og så trekkes intervjuobjekter med et bestemt intervall.

Trek_sys kan også brukes til å trekke et enkelt tilfeldig utvalg. Da må trekkegrunnlaget på forhånd være sortert tilfeldig. Når makroen kalles opp, settes parameteren `Gruppe` til 0 (ingen sorteringsvariabel).

3.2 Kjøring av *Trek_sys*

Når *Trek_sys* kjøres, vil brukeren bli bedt om å oppgi utvalgsstørrelser i hvert stratum. Brukeren kan selv velge hvorvidt utvalgsstørrelsene skal oppgis i prosent eller antall. Når utvalget er ferdig trukket, vil en kjørerapport bli lagt sist i LOG-vinduet. Når makroen kalles opp, kreves følgende parametre:

```
%Trek_sys (Fil_i, Fil_u, Ident, J, Gruppe, Fast, Z);
```

`Fil_i`	Navn på innfilen (trekkegrunnlaget)
`Fil_u`	Navn på utfilen (som vil bestå av de uttrukne observasjonene)
`Ident`	Navn på unik variabel (f.eks. personnummer, bedriftsnummer). Dersom en slik variabel ikke eksisterer i datagrunnlaget, må denne opprettes før makroen kan kjøres. Variabelen kan være numerisk eller alfanumerisk
`J`	Navn på stratumvariabel. Dersom en kun har ett stratum og en slik variabel derfor ikke eksisterer, settes parameteren til 0. Variabelen som angir stratum kan være numerisk eller alfanumerisk
`Gruppe`	Variabel som populasjonen skal sorteres på før trekking. Dersom dette ikke er aktuelt, oppgis 0. Når `Gruppe` settes til 0 og trekkegrunnlaget er tilfeldig sortert på forhånd, vil vi få et enkelt tilfeldig utvalg. Variabelen kan være numerisk eller alfanumerisk
`Fast`	Indikerer om utvalgsstørrelsene skal oppgis i prosent eller antall 0 = prosent 1 = antall
`Z`	Startverdi for maskinens tilfeldige tallgenerator. Oppgi et positivt heltall mindre enn 2 147 483 647. Dersom makroen kjøres flere ganger med samme startverdi (og trekkgrunnlag), vil maskinen trekke det samme utvalget. Dersom `Z` settes lik 0, vil startverdien bestemmes av maskinens klokke. Da vil maskinen trekke forskjellig utvalg hver gang. Dersom man ønsker å kunne rekonstruere utvalgstrekkingen (og det kan ofte være nødvendig), må `Z` ikke settes lik 0

Eksempel 1:

Trekkgrunnlaget for fysioterapeutundersøkelsen 1994 var medlemsregisteret i Norske fysioterapeuters forbund pr. februar 1995. Registeret består bl.a. av følgende opplysninger:

`Medl_nr`	*Medlemsnummer*
`Navn`	*Navn*
`F_dato`	*Fødselsdato*
`Komm`	*Kommunenummer*
`Praksis`	*Praksistype (eier, leier, gruppepraksis eller andre)*

Medlemsnummeret er unikt, dvs. at det ikke eksisterer flere personer med samme medlemsnummer. I tillegg ble det konstruert en variabel `Stratum` *som en funksjon av praksisform og aldersgruppe:*

Stratum	Prasisform	Alder	Antall i medlemsregister
11	Eier	35 år eller yngre	51
12	Eier	36 - 50 år	484
13	Eier	51 år eller eldre	219
21	Leier	35 år eller yngre	262
22	Leier	36 - 50 år	472
23	Leier	51 år eller eldre	112
31	Gruppepraksis	35 år eller yngre	57
32	Gruppepraksis	36 - 50 år	203
33	Gruppepraksis	51 år eller eldre	34
41	Andre praksisformer	35 år eller yngre	64
42	Andre praksisformer	36 - 50 år	137
43	Andre praksisformer	51 år eller eldre	31
		Totalt	2 126

Trekkegrunnlaget kalles for `GRUNNLAG` *og legges under katalogen* `'/SSB/KS3/UTVALG/'`*. For eksempelets skyld ønsker vi å trekke et utvalg med 30 observasjoner fra hvert stratum, og utvalgsfilen skal legges på den temporære filen* `UTFIL`*. I tillegg til variablene det allerede er stratifisert på, ønsker vi å sikre geografisk representativitet i utvalget. Dette utføres med følgende programlinjer (%INCLUDE-setningen er overflødig dersom den er utført tidligere):*

```
LIBNAME PERM '/SSB/KS3/UTVALG';

%INCLUDE '$FELLES/sasprog/trek_sys.sas';
%TREK_SYS (PERM.Grunnlag, Utfil, Medl_nr, Stratum, Komm, 1, 33343);
```

Når makroen kjøres, vil brukeren bli bedt om å oppgi utvalgsstørrelser i hvert stratum. Utvalget vil så bli trukket automatisk og lagt på filen `UTFIL`*. En del nyttige opplysninger blir lagt sist i LOG-vinduet, som vil se slik ut:*

```
************************************************************
Dato og klokkeslett                    28JUL95 13:04
Innfil (trekkgrunnlag)                 perm.grunnlag
Antall observasjoner i populasjonen          2126
Utfil                                  utfil
Antall observasjoner i utvalget               360
Startverdi                             33343
************************************************************
```

Eksempel 2:

Utvalg kan også trekkes med trekksannsynligheter i hvert stratum i stedet for absolutte utvalgsstørrelser. Nå ønskes et utvalg av denne typen:

Stratum	Utvalgsstørrelse
11	15 %
12	15 %
13	10 %
21	50 %
22	40 %
Øvrige	10 %

Makroen kalles da opp med parameteren `Fast` *lik 0. Her velges dessuten en ny startverdi, mens de øvrige parametrene forblir uendret:*

```
%TREK_SYS (PERM.Grunnlag, Utfil, Medl_nr, Stratum, Komm, 0,544437);
```

Trekksannsynlighetene i hvert stratum legges inn på forespørsel fra programmet, og nedenfor er kjørerapporten i LOG-vinduet gjengitt:

```
************************************************************
Dato og klokkeslett                  28JUL95 13:04
Innfil (trekkgrunnlag)               perm.grunnlag
Antall observasjoner i populasjonen          2126
Utfil                                utfil
Antall observasjoner i utvalget               480
Startverdi                           544437
************************************************************
```

3.3 Proporsjonale trekksannsynligheter

Det er ofte ønskelig å la trekksannsynligheten variere med et størrelsesmål, spesielt er dette vanlig i bedriftsutvalg. Et slikt størrelsesmål kan for eksempel være opplysning om sysselsetting eller omsetning. Bakgrunnen for dette er at man generelt ønsker å minimere antall bedrifter i utvalget samtidig som man ønsker å inkludere så mange ansatte (eller så stor del av omsetningen) som mulig. Dette oppnås ved at man overrepresenterer store bedrifter. Se f.eks. *Cochran (1977)* for mer teori omkring pps-trekking (Probability Proportional to Size).

En metode for dette er å stratifisere etter størrelse, for eksempel 'store', 'mellomstore' og 'små' bedrifter. Trekksannsynligheter kan da bestemmes direkte for hvert stratum. Man har da fremdeles kun én trekksannsynlighet i hvert stratum, og makroen *Trek_sys* kan brukes på vanlig måte.

En annen måte er å la enheter i samme stratum få ulik trekksannsynlighet, avhengig av størrelsen på en gitt variabel. Prinsippet er at dersom bedrift A er dobbelt så stor som bedrift B, skal denne ha dobbelt så stor sannsynlighet for å bli trukket, osv. Det kan imidlertid ofte bli vanskelig å holde på dette prinsippet. Ta for eksempel et stratum med tre bedrifter, med størrelsesmål (sysselsetting) som vist i tabellen. Tabellen viser sannsynligheten for at hver av de tre bedriftene skal inkluderes i utvalget, når utvalgsstørrelsen er henholdsvis en, to og tre bedrifter.

Bedrift	Sysselsetting	$Pr_{n=1}$	$Pr_{n=2}$	$Pr_{n=3}$
A	2	0,033	0,173	1,000
B	10	0,161	0,839	1,000
C	50	0,806	0,988	1,000
Sum trekksannsynligheter		1,000	2,000	3,000

Med n=1 ser vi at bedrift C har fem ganger så stor trekksannsynlighet som bedrift B, som igjen har fem ganger så stor trekksannsynlighet som bedrift A. Når n=2 har bedrift C 1,2 ganger så stor trekksannsynlighet som bedrift B, mens forholdet mellom bedrift B og A er 4,8. Når n=3 har selvsagt alle bedriftene trekksannsynlighet 1.

I praksis vil det være umulig å beregne eksakte trekksannsynligheter når n er stor. Dette vil igjen lett føre til en skjevhet i estimatene, fordi trekksannsynlighetene ofte benyttes til å blåse opp observasjonene i utvalget. Følgende kan tjene som tommelfingerregler når man trekker proporsjonalt med størrelse:

- Metoden er lite egnet i situasjoner med høy trekksannsynlighet (ref. eksempelet over).
- Observasjonene bør ikke ha for stor spredning i størrelsesmålet. For eksempel vil det være lite formålstjenlig å trekke 100 bedrifter proporsjonalt med antall sysselsatte blant alle bedrifter i Norge. Man vil da med stor sikkerhet ikke ha småbedrifter representert i utvalget. En slik situasjon løses ofte ved at de største bedriftene behandles for seg. Med n=2 i eksempelet over kunne trekksannsynligheten settes lik 1 for bedrift C, mens den siste observasjonen til utvalget ble trukket pps blant A og B.

3.4 Kjøring av *Trek_pps*

Makroen *Trek_pps* trekker et utvalg der trekksannsynlighetene er tilnærmet proporsjonale med størrelsen på en angitt variabel. Parametrene som kreves er nesten identiske med parametrene for *Trek_sys*. Eneste

forskjell er at i stedet for å angi eventuell sorteringsvariabel, angis en numerisk variabel som trekksannsynlighetene skal bestemmes av.

```
%Trek_pps (Fil_i, Fil_u, Ident, J, Prop, Fast, Z);
```

`Fil_i`	Navn på innfilen (trekkegrunnlaget)
`Fil_u`	Navn på utfilen (som vil bestå av de uttrukne observasjonene)
`Ident`	Navn på unik variabel (f.eks. personnummer, bedriftsnummer). Dersom en slik variabel ikke eksisterer i datagrunnlaget, må denne opprettes før makroen kan kjøres. Variabelen kan være numerisk eller alfanumerisk
`J`	Navn på stratumvariabel. Dersom en kun har ett stratum og en slik variabel derfor ikke eksisterer, settes parameteren til 0. Variabelen som angir stratum kan være numerisk eller alfanumerisk
`Prop`	Variabelen som det skal trekkes proporsjonalt med hensyn på. Variabelen må være numerisk
`Fast`	Indikerer om utvalgsstørrelsene skal oppgis i prosent eller antall. 0 = prosent 1 = antall
`Z`	Startverdi for maskinens tilfeldige tallgenerator

Som i *Trek_sys* vil en også her få spørsmål om utvalgsstørrelse (eller trekksannsynlighet) i hvert stratum.

Eksempel 3:

En populasjon på 50 000 bedrifter er opprettet med filnavn `Pop_fil`*, med 10 000 bedrifter på henholdsvis 10, 20 , 30, 40 og 50 ansatte. For enkelhets skyld er alle bedriftene plassert i samme stratum. Populasjonen inneholder dermed kun variablene:*

`Ident`	*Løpende, unikt identifikasjonsnummer*
`Ansatte`	*Antall ansatte, 10, 20, 30, 40 eller 50.*

Dersom man nå ønsker å trekke et utvalg med utvalgsstørrelse 5 000 og kalle utvalgsfilen for `UTV_FIL`, kaller man makroen opp med følgende programlinjer i SAS (%INCLUDE-setningen er overflødig dersom den er utført tidligere):

```
%INCLUDE '$FELLES/sasprog/trek_pps.sas';
%Trek_pps (Pop_fil, Utv_fil, Unik, 0, Ansatte, 0, 12345);
```

og den ønskede utvalgsstørrelsen oppgis når programmet spør om det. Antall bedrifter i utvalget fordeler seg slik med hensyn på størrelse (PROC FREQ i SAS er brukt):

```
+OUTPUT-----------------------------------------------------------------+
|Command ===>                                                           |
|                                                                       |
|The SAS System                                                      1  |
|                                                                       |
|                                 Cumulative   Cumulative               |
|ANSATTE   Frequency   Percent    Frequency    Percent                  |
|-----------------------------------------------------                  |
|     10         332       6.6          332        6.6                  |
|     20         637      12.7          969       19.4                  |
|     30         954      19.1         1923       38.5                  |
|     40        1383      27.7         3306       66.1                  |
|     50        1694      33.9         5000      100.0                  |
|                                                                       |
|                                                                       |
|                                                                       |
|                                                                       |
+-----------------------------------------------ZOOM----I--------------+
                                              KS4  14:39:15
```

Utvalget gir et riktig forhold mellom store og små bedrifter, for eksempel er det rundt fem ganger så mange bedrifter med 50 ansatte som med 10 ansatte, osv.

4. Usikkerhetsberegninger

4.1 Innledning

Resultatene av en undersøkelse vil alltid være påvirket av usikkerhet. Selv om vi gjennomfører en totaltelling uten frafall, kan vi ikke gardere oss mot registerfeil, målefeil, feilpunching, etc. For enkelhets skyld betegnes denne type usikkerhet her som *målefeil*. Denne størrelsen er som regel svært vanskelig å anslå.

Dersom vi trekker et utvalg, vil usikkerheten i tallene øke. Denne ekstra usikkerheten kalles *utvalgsvarians* eller *designvarians* og er mulig å beregne under gitte forutsetninger. Det er imidlertid

viktig å være klar over at vi med utvalgsvarians ikke mener den totale usikkerheten, men kun usikkerheten som følge av at vi har gjennomført et utvalg fremfor totaltelling.

Makroen *Des_var* beregner utvalgsvariansen i et enkelt tilfeldig utvalg. Med noen forutsetninger kan variansen også anslås for systematiske utvalg. Makroen gir estimater (gjennomsnitt og oppblåst sum), standardfeil og relativt standardavvik til en gitt variabel. Usikkerheten kan også beregnes for delpopulasjoner som kjønn eller aldersgrupper. Selve variansformlene som er benyttet i makroen, er lagt til Appendiks A.

4.2 Utvalgsvarians i systematiske utvalg

Det er i prinsippet ikke mulig å beregne den eksakte utvalgsvariansen til et systematisk utvalg når man kun har kjennskap til de observerte verdiene i det realiserte utvalget. Dette fordi variansen mellom de mulige utvalgene inngår i variansuttrykket. (Dersom man trekker hver femte observasjon, har man fem mulige utvalg avhengig av startpunktet). Utvalgsvariansen kan likevel tilnærmes i noen situasjoner ved å bruke variansuttrykkene for enkelt tilfeldig utvalg. Et viktig resultat *(Cochran s. 208)* sier at et systematisk utvalg er mer presist enn et enkelt tilfeldig utvalg dersom den innbyrdes variansen i de k mulige utvalgene er større enn populasjonsvariansen. Dette vil normalt være tilfelle når populasjonen før trekkingen blir sortert på en slik måte at en i alle de k mulige utvalgene er sikret en god spredning i analysevariabelen. Følgende tommelfingerregler kan være nyttige:

- Dersom sorteringsvariabelen er uavhengig av målevariabelen, vil variansen være tilnærmet lik ved systematisk og enkelt tilfeldig utvalg. Da kan variansen beregnes tilnærmingsvis ved hjelp av *Des_var*.

- Dersom sorteringsvariabelen er svakt positivt eller negativt korrelert med målevariabelen, vil variansen være mindre ved et systematisk utvalg enn ved et enkelt tilfeldig utvalg. Da kan usikkerheten anslås ved hjelp av *Des_var*, og en vil være sikker på at den virkelige variansen 'i hvert fall ikke er høyere' enn vårt estimat. Dette vil imidlertid være en dårlig løsning dersom sorteringsvariabelen og målevariabelen er sterkt korrelerte, fordi den virkelige variansen da vil være langt lavere enn våre anslag.

- Dersom man har periodisitet i dataene som faller sammen med intervallet i utvalget, vil variansen i et systematisk utvalg være større enn variansen i et enkelt tilfeldig utvalg. I slike tilfeller bør en ikke foreta systematiske utvalg, og variansen kan ikke estimeres.

Av det siste punktet følger at man bør være forsiktig med å bruke systematisk utvalg når populasjonen er sortert etter tiden. For eksempel vil det være lite hensiktsmessig å samle inn omsetningstall fra en bedrift hver sjuende dag dersom man ønsker å estimere den gjennomsnittlige daglige omsetningen. Man vil da sitte igjen med et sett observasjoner som er fra samme ukedag, og som ikke er representativt for hele uken. Derimot kan man for eksempel observere hver tredje dag eller hver femte dag, og utvalget vil da bli representativt dersom målingene går over mange nok uker. I **kapittel 6** er noen matematiske sider ved utvalgsvariansen i et systematisk utvalg diskutert mer i detalj.

4.3 Kjøring av *Des_var*

For å kunne regne ut utvalgsvariansen ved hjelp av *Des_var*, må følgende forutsetninger være oppfylt:

- Innenfor ett stratum har alle enhetene i populasjonen like stor sannsynlighet for å bli trukket til utvalget. Følgelig kan *Des_var* ikke brukes til å beregne utvalgsvariansen ved pps-sampling.
- Gjennomsnittet av analysevariabelen er tilnærmet normalfordelt. Vanligvis er tilnærmelsen brukbar hvis en på forhånd vet at analysevariabelen tar verdier som alle opptrer relativt hyppig. Store utvalg gir en bedre tilnærming enn små utvalg.
- Utvalget må ha minst to observasjoner i alle strata. Dersom vi ønsker usikkerhetstall for en delpopulasjon, må utvalget ha minst to observasjoner i alle kombinasjoner av (delpopulasjon * stratum) som forekommer i populasjonen. Denne forutsetningen vil makroen selv sjekke.
- Samtlige observasjoner må være tildelt en vekt (numerisk), og observasjoner i samme stratum må ha samme vekt.

Den siste forutsetningen fører til at usikkerheten ikke kan beregnes ved hjelp av *Des_var* dersom en f.eks. har redusert vektene til ekstremobservasjoner. Selve makroen utføres med følgende programsetning:

```
%Des_var (Fil, J, V, X, Delp);
```

`Fil`	Navn på utvalgsfil
`J`	Navn på stratumvariabel
`V`	Navn på variabel som angir vekt
`X`	Navn på variabelen som skal analyseres
`Delp`	Navn på eventuell delpopulasjon. Dersom dette ikke er aktuelt, oppgi 0

Eksempel 4:

Utvalgsfilen for fysioterapeutundersøkelsen lyder navnet `Fysio`. *Filen inneholder stratumvariabelen* `Stratum`, *vektvariabelen* `Vekt`, *analysevariabelen* `Inntekt` *og en variabel* `Agru` *som angir en av tre aldersgrupper. For å analysere gjennomsnittlig og total inntekt i populasjonen med tilhørende standardfeil fordelt på aldersgruppe, er det nå tilstrekkelig med følgende programlinjer:*

```
%INCLUDE '$FELLES/sasprog/des_var.sas';
%DES_VAR (Fysio, Stratum, Vekt, Inntekt, Agru);
```

og resultatet fremkommer i OUTPUT-vinduet.

```
+OUTPUT-------------------------------------------------------------------------+
|Command ===>                                                                   |
|                                                                               |
|Estimat og standardfeil for snitt og sum, variabel Inntekt                  16 |
|                                                                               |
|        Antall  Oppblåst Estimert Estimert Relativ   Estimert    Estimert      |
|AGRU   i utvalg  antall  gj.snitt std.feil std.feil     sum      std.feil      |
|                                                                               |
|  -35      72       338   374 831   17 944    0.048  126 689 289  6 065 041   |
|36-50     181     1 160   369 632    9 145    0.025  428 858 204 10 610 053   |
|51-        91       372   312 649   10 984    0.035  116 261 608  4 084 695   |
|Totalt    344     1 870   359 241    6 890    0.019  671 809 102 12 885 755   |
|                                                                               |
|                                                                               |
|                                                                               |
|                                                                               |
|                                                                               |
|                                                                               |
+-------------------------------------------------------------ZOOM----I--------+
                                                              KS3  10:36:17
```

Under vektberegningen er det tatt hensyn til ulike frafallsgrunner, se kapittel 5. Det oppblåste populasjonsantallet stemmer derfor ikke overens med antall i medlemsregisteret i avsnitt 3.2.

Standardfeilen kan så benyttes til å angi konfidensintervall for parameteren. Et 95 prosent konfidensintervall for gjennomsnittlig inntekt for hele massen er [345 701 kr, 372 697 kr]. Teori omkring standardfeil og konfidensintervall kan f.eks. leses i Cochran (1977).

Eksempel 5

For å illustrere bruken av makroen ytterligere, gis et nytt eksempel. Denne gangen ønskes et estimat og standardfeil på årsresultat (variabel `Ov_skudd`*) fordelt på landsdel (variabel* `Region`*). For noen landsdeler er det ikke mulig å beregne usikkerheten. Dette fordi noen kombinasjoner av* (`Stratum * Region`) *kun har én observasjon. Beregningen utføres ved hjelp av programlinjen*

```
%Des_var (PERM.Fysio, Stratum, Vekt, Ov_skudd, Region);
```

(eventuelt en %INCLUDE-setning i forkant dersom denne ikke er utført).

```
+OUTPUT-------------------------------------------------------------------------+
|Command ===>                                                                   |
|                                                                               |
|Estimat og standardfeil for snitt og sum, variabel p935                      1 |
|                                                                               |
|                    Antall      Oppblåst     Estimert     Estimert             |
|REGION             i utvalg      antall      gj.snitt     std.feil             |
|                                                                               |
|Agder/Rogaland          48          247       292 442            .            |
|Nord-Norge              20          119       218 657            .            |
|Oslo og Akershus        44          251       207 016       13 308            |
|Trøndelag               24          116       204 001       15 229            |
|Vestlandet              55          301       249 821            .            |
|Østlandet ellers       153          837       242 494        7 933            |
|Totalt                 344        1 870       241 591        5 168            |
|                                                                               |
|                                                                               |
|                   Relativ       Estimert        Estimert                      |
|REGION             std.feil         sum          std.feil                      |
|                                                                               |
|Agder/Rogaland          .       72 151 181              .                      |
|Nord-Norge              .       26 035 448              .                      |
|Oslo og Akershus    0.064       51 936 211    3 338 730                        |
|Trøndelag           0.075       23 715 079    1 770 429                        |
|Vestlandet              .       75 108 815              .                      |
|Østlandet ellers    0.033      202 848 241    6 635 929                        |
|Totalt              0.021      451 794 975    9 665 150                        |
|                                                                               |
|                                                                               |
|                                                                               |
|                                                                               |
+-------------------------------------------------------------ZOOM----I--------+
                                                              KS4  14:21:20
```

I neste tabell er estimater på gjennomsnittlige driftsinntekter, driftskostnader og årsoverskudd vist sammen med estimert standardavvik og relativt standardavvik i de to undersøkelsene[1].

	Tannleger			Fysioterapeuter		
Variabel	Gj.snitt	Std.feil	Rel std.feil	Gj.snitt	Std.feil	Rel. std.feil
Driftsinntekter	1 050 144	28 148	0,027	359 241	6 890	0,019
Driftskostnader	605 244	18 171	0,030	115 099	4 457	0,039
Årsresultat	434 512	13 424	0,031	241 591	5 168	0,021

[1] For fysioterapeuter avviker datagrunnlaget noe fra de publiserte tallene.

5. Frafall og populasjonskorrigering

I dette kapittelet forklares hvordan informasjon om frafallsgrunn er benyttet til å korrigere populasjonen i de to undersøkelsene. For å holde notasjonen på et oversiktlig nivå, behandles her kun en situasjon med ett stratum. Prinsippene kan imidlertid overføres direkte til et utvalg med flere strata.

5.1 Ulike frafallsgrunner

La N være antall enheter i populasjonen (i ett gitt stratum), og n_r nettoutvalget i det samme stratumet. Ved etterstratifisering er det vanlig å gi observasjonene en vekt v, som er lik forholdet mellom antall i populasjonen og antall observasjoner i utvalget:

$$v = \frac{N}{n_r}$$

Dersom vi har kunnskap om frafallsgrunn, kan vi benytte denne kunnskapen til å konstruere bedre vekter. Vet vi f.eks. at halvparten av frafallet skyldes at enhetene faller utenfor populasjonen (gått over til ny selskapsform, nedlagt praksis etc.), virker det fornuftig å nedjustere N. Omfanget av slikt frafall vil rimeligvis avhenge av kvaliteten på registeret vi trekker fra. Vi skiller for anledningen mellom tre typer frafall:

Type 1: Frafall der intervjuobjektet (IO) faller innenfor populasjonen, men ikke svarer av andre årsaker (bortreist, sykdom, vil ikke svare, etc)

Type 2: Frafall der IO faller utenfor populasjonen som følge av dårlig registerkvalitet (ny selskapsform, ny næring, død, nedlagt virksomhet etc)

Type 3: Frafall med ukjent frafallsgrunn

Undersøkelsene for tannleger og fysioterapeuter var begge basert på frivillighet. I tilfeller der IO ikke svarte, ble vedkommende bedt om å returnere et frafallsskjema med påført frafallsgrunn. Personer som leverte frafallsskjema, ble gruppert i type 1 eller type 2. Personer som ikke returnerte frafallsskjema, ble satt til type 3 (ukjent frafall). Nå innføres:

N	Antall i populasjonen, ifølge register/trekkgrunnlag
n_s	Bruttoutvalg
n_{s-r}	Totalt frafall
f_j	Frafall av type j, j=1, 2 eller 3. $\sum_{j=1}^{3} f_j = n_{s-r}$

Vi forutsetter nå følgende:

1. Sannsynligheten for at en ikke-respondent faller utenfor populasjonen er uavhengig av om vedkommende sendte inn frafallsskjema eller ikke.
2. Dersom vi hadde foretatt en totaltelling, ville vi fått samme frafallsprosent som vi har fått i utvalget.
3. Blant de enhetene som ikke hadde respondert i en (eventuell) totaltelling, kan vi anta at andelen som faller utenfor populasjonen er den samme som blant de som sendte inn frafallsskjema til undersøkelsen.

Antall i populasjonen, N, justeres nå ved at

$N^* = N$ – (Frafallstype 2 i utvalget)
– (Antatt frafallstype 2 av dem som har ukjent frafallsgrunn)
– (Antatt frafallstype 2 blant enhetene utenfor utvalget).

$$N^* = N - f_2 - f_3 \frac{f_2}{(f_1 + f_2)} - (N - n_s)\frac{n_{s-r}}{n_{bru}}\frac{f_2}{(f_1 + f_2)}$$

$$= N - f_2 - \frac{f_2}{(f_1 + f_2)}\left[f_3 + \frac{Nn_{s-r}}{n_s} - n_{s-r}\right]$$

$$= N - f_2 - \frac{f_2}{(f_1 + f_2)}\left[\frac{Nn_{s-r}}{n_s} - (f_1 + f_2)\right]$$

$$= N - f_2 + f_2 - \frac{f_2}{(f_1 + f_2)}\frac{Nn_{s-r}}{n_s}$$

$$= N\left[1 - \frac{f_2 n_{s-r}}{(f_1 + f_2)n_s}\right]$$

og de justerte vektene v^* er gitt ved

$$v^* = \frac{N^*}{n_r}$$

5.2 Populasjonskorrigeringer i tannlege- og fysioterapeutundersøkelsen

I de følgende tabellene er vekter beregnet uten og med populasjonskorrigering i tannlege- og fysioterapeutundersøkelsen. Vektene gjelder for resultatregnskapet i de to undersøkelsene.

Tannleger

Stratum	Popul. N	Bruttoutv. n_s	Nettoutv. n_r	Svar-prosent	Frafall n_{s-r}	Type 1 f_1	Type 2 f_2	Type 3 f_3	Vekt uten korr.	Vekt med korr.
1	47	47	13	28	34	2	7	25	3,62	1,58
2	517	155	46	30	109	12	2	95	11,24	10,11
3	445	133	55	41	78	5	2	71	8,09	6,74
4	442	132	59	45	73	10	5	58	7,49	6,11
5	256	102	49	48	53	12	7	34	5,22	4,22
91	108	54	8	15	46	8	5	33	13,50	9,08
92	44	44	15	34	29	1	5	23	2,93	1,32
	1859	667	245	37	422	50	33	339		

Fysioterapeuter

Stratum	Popul. N	Bruttoutv. n_s	Nettoutv. n_r	Svar-prosent	Frafall n_{s-r}	Type 1 f_1	Type 2 f_2	Type 3 f_3	Vekt uten korr.	Vekt med korr.
11	51	38	17	45	21	7	2	12	3,00	2,63
12	484	121	56	46	65	31	3	31	8,64	8,23
13	219	109	46	42	63	43	1	19	4,76	4,70
21	262	131	36	27	95	35	12	48	7,28	5,93
22	472	118	53	45	65	28	8	29	8,91	7,82
23	112	56	18	32	38	29	5	4	6,22	5,60
31	57	42	11	26	31	11	4	16	5,18	4,16
32	203	101	48	48	53	25	3	25	4,23	3,99
33	34	34	15	44	19	15	2	2	2,27	2,12
41	64	48	8	17	40	12	15	13	8,00	4,30
42	137	68	24	35	44	17	14	13	5,71	4,04
43	31	31	12	39	19	8	5	6	2,58	1,97
	2126	897	344	38	553	261	74	218		

Hvordan korrigeringene slår ut, vil selvsagt variere fra undersøkelse til undersøkelse. Totalsummer vil generelt bli estimert lavere, fordi observasjonene får en lavere vekt. Estimerte gjennomsnittsverdier vil

kun påvirkes dersom andelen av type 2 frafall fordeler seg ulikt på strataene. I så fall kan de estimerte verdiene bli påvirket i både positiv og negativ retning.

I neste tabell er gitt noen hovedstørrelser fra resultatregnskapet i de to undersøkelsene. Tabellen viser estimerte gjennomstnittstall uten og med frafallskorrigering, og prosentvis avvik mellom estimatene[2].

	Tannleger (1993)			Fysioterapeuter (1994)		
	Uten korr.	Med korr.	Prosent avvik	Uten korr.	Med korr.	Prosent avvik
Driftsinntekter	1 050 144	1 051 527	0,1	355 928	359 241	0,9
Driftsresultat	444 900	446 215	0,3	242 246	244 099	0,8
Årsresultat	434 512	436 126	0,4	239 797	241 591	0,7

Som det fremgår av tabellen, har populasjonskorrigeringen størst relativ effekt i fysioterapeutundersøkelsen. I de offisielle tallene fra undersøkelsen er også slik korrigering gjort. Neste tabell viser hvor mye populasjonsantallet er redusert i hvert stratum. I tannlegeundersøkelsen er populasjonskorrigering ikke gjort i de offisielle tallene.

Fysioterapeutundersøkelsen

Stratum	N	N*
11	51	44,7
12	484	460,4
13	219	215,9
21	262	213,0
22	472	412,7
23	112	100,8
31	57	45,8
32	203	190,7
33	34	31,8
41	64	34,4
42	137	96,6
43	31	23,5
Totalt	2 126	1 870,1

Modellen for populasjonskorrigering bygger på ulike forutsetninger. Realismen i forutsetningene kan selvsagt diskuteres, og vil avhenge av undersøkelsens art. Spesielt den første forutsetningen,

[2] For fysioterapeuter avviker datagrunnlaget noe fra de publiserte tallene.

sannsynligheten for at en ikke-respondent faller utenfor populasjonen er uavhengig av om vedkommende sendte inn frafallsskjema eller ikke, vil ikke holde i alle tilfeller. Dersom en alternativ frafallsgrunn er 'død', vil dette i de fleste tilfellene bli regnet som et type 2 frafall. Av de som er døde, vil imidlertid de færreste returnere frafallsskjema. Forutsetningene i modellen vil da ikke være til stede (dersom vi da ikke kan identifisere døde personer ved hjelp av andre kilder).

6. Formler for utvalgsvarians

Her blir formlene som er benyttet i makroen *Des_var* gjengitt. Resultatene er hentet fra *Cochran (1977)* og *Thomsen (SØS 33).*

6.1 Enkelt tilfeldig utvalg

Med et enkelt, tilfeldig utvalg menes at alle enhetene i populasjonen har like stor sannsynlighet til å bli trukket. La N være antall enheter i populasjonen, mens n er antall enheter i utvalget. Vi ønsker å estimere $\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j$ eller $Y = \sum_{j=1}^{N} Y_j$. En forventningsrett estimator for $\overline{Y}$ er

$$\overline{y} = \frac{1}{n}\sum_{i \in s} y_i$$

der det med notasjonen $i \in s$ menes alle y_i som er inkludert i utvalget. Variansen til $\overline{y}$ er gitt ved

$$V(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \quad \text{, der}$$

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(y_j - \overline{Y}\right)^2$$

Når S^2 er ukjent, erstattes denne med

$$s^2 = \frac{1}{n-1}\sum_{i \in s}\left(y_i - \overline{y}\right)^2 \text{ , slik at}$$

$$\hat{V}(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

En estimator for summen Y samt tilhørende varians er gitt ved

$$\hat{Y} = N\overline{y}$$

$$\hat{V}(\hat{Y}) = N^2\,\hat{V}(\overline{y})$$

6.2 Stratifiserte utvalg

Populasjonen deles inn i L strata. Fra hvert stratum trekkes et enkelt, tilfeldig utvalg. Vi ønsker å estimere den gjennomsnittsverdien eller summen for hele massen,

$$\overline{Y} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{Y}_h$$

$$Y = N\overline{Y} = \sum_{h=1}^{L} N_h \overline{Y}_h$$

der N_h = antall i populasjonen i stratum h, $\sum_{h=1}^{L} N_h = N$

$\overline{Y}_h$ = gjennomsnittsverdi for Y i stratum h

En forventningsrett estimator for $\overline{Y}$ med tilhørende variansuttrykk er gitt ved

$$\overline{y}_{ST} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_h$$

$$V(\overline{y}_{ST}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

der n_h = antall i utvalget i stratum h, $\sum_{h=1}^{L} n_h = n$

$\overline{y}_h$ = gjennomsnittsverdi for Y i utvalget, stratum h

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)^2$$

Når S_h^2 er ukjent, kan variansen til $\overline{y}_{ST}$ estimeres ved

$$\hat{V}(\overline{y}_{ST}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} \quad \text{, der}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2$$

En forventningsrett estimator for den totale summen Y med tilhørende uttrykk for den estimerte variansen er gitt ved

$$\hat{Y}_{ST} = N\bar{y}_{ST}$$

$$\hat{V}(\hat{Y}_{ST}) = N^2 \hat{V}(\bar{y}_{ST})$$

6.3 Delpopulasjoner

Formlene i avsnitt A.2 kan brukes direkte til å beregne varians for delpopulasjoner. Hver delpopulasjon betraktes da som en selvstendig populasjon, med tilhørende utvalgsstørrelser i hvert stratum, antall i populasjonen, osv.

6.4 Systematiske utvalg

I et systematisk utvalg vil det være *k* mulige utvalg, alle med utvalgsstørrelse n. Variansen i et systematisk utvalg er da gitt ved

$$V(\bar{y}_{SY}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{WSY} \qquad \text{, der}$$

$$S^2_{WSY} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\bullet}\right)^2$$

For enkelhets skyld behandles her et utvalg med ett stratum. Uttrykkene kan overføres til et utvalg med flere strata, etter samme prinsipp som i avsnitt A.2.

Med notasjonen y_{ij} menes enhet nummer *j* i utvalg nummer *i* (i = 1, .. , k). For å beregne $V(\bar{y}_{SY})$ må alle y-verdier i populasjonen være kjent, noe som svært sjelden er tilfelle. Ved å sammenligne uttrykkene for $V(\bar{y})$ og $V(\bar{y}_{SY})$, kommer en imidlertid frem til et nyttig resultat:

> *Variansen i et systematisk utvalg er mindre enn variansen i et enkelt tilfeldig utvalg hvis og bare hvis* $S^2_{WSY} > S^2$.

Resultatet sier at et systematisk utvalg er mer presist enn et enkelt tilfeldig utvalg dersom den innbyrdes variansen i de *k* mulige utvalgene er større enn populasjonsvariansen. Dette vil normalt være tilfelle når populasjonen før trekkingen blir sortert på en slik måte at en i alle de *k* mulige utvalgene er sikret en god spredning i analysevariabelen. Følgende tommelfingerregler kan være nyttige:

- Dersom sorteringsvariabelen er uavhengig av målevariabelen, dvs. sorteringen kan sies å være tilfeldig, vil $V(\bar{y}_{SY}) \approx V(\bar{y})$. Da kan variansen beregnes som om utvalget er et enkelt tilfeldig utvalg.

- Dersom sorteringsvariabelen er positivt eller negativt korrelert med målevariabelen, vil $V(\bar{y}_{SY}) < V(\bar{y})$. Da kan usikkerheten beregnes som om utvalget er et enkelt tilfeldig utvalg, og en vil være sikker på at den virkelige variansen ikke er høyere enn vårt estimat. En sterk korrelasjon mellom sorteringsvariabel og målevariabel vil gi store avvik mellom $V(\bar{y}_{SY})$ og $V(\bar{y})$.

- Dersom man har periodisitet i dataene, vil $V(\bar{y}_{SY}) > V(\bar{y})$. I slike tilfeller bør en ikke foreta systematiske utvalg, og variansen kan ikke estimeres ved hjelp av det realiserte utvalget.

Av det siste punktet følger at man ikke bør bruke systematisk utvalg når populasjonen er sortert etter tiden. For eksempel vil det være lite hensiktsmessig å samle inn omsetningstall fra en bedrift hver sjuende dag dersom man ønsker å estimere den gjennomsnittlige daglige omsetningen.

Referanser

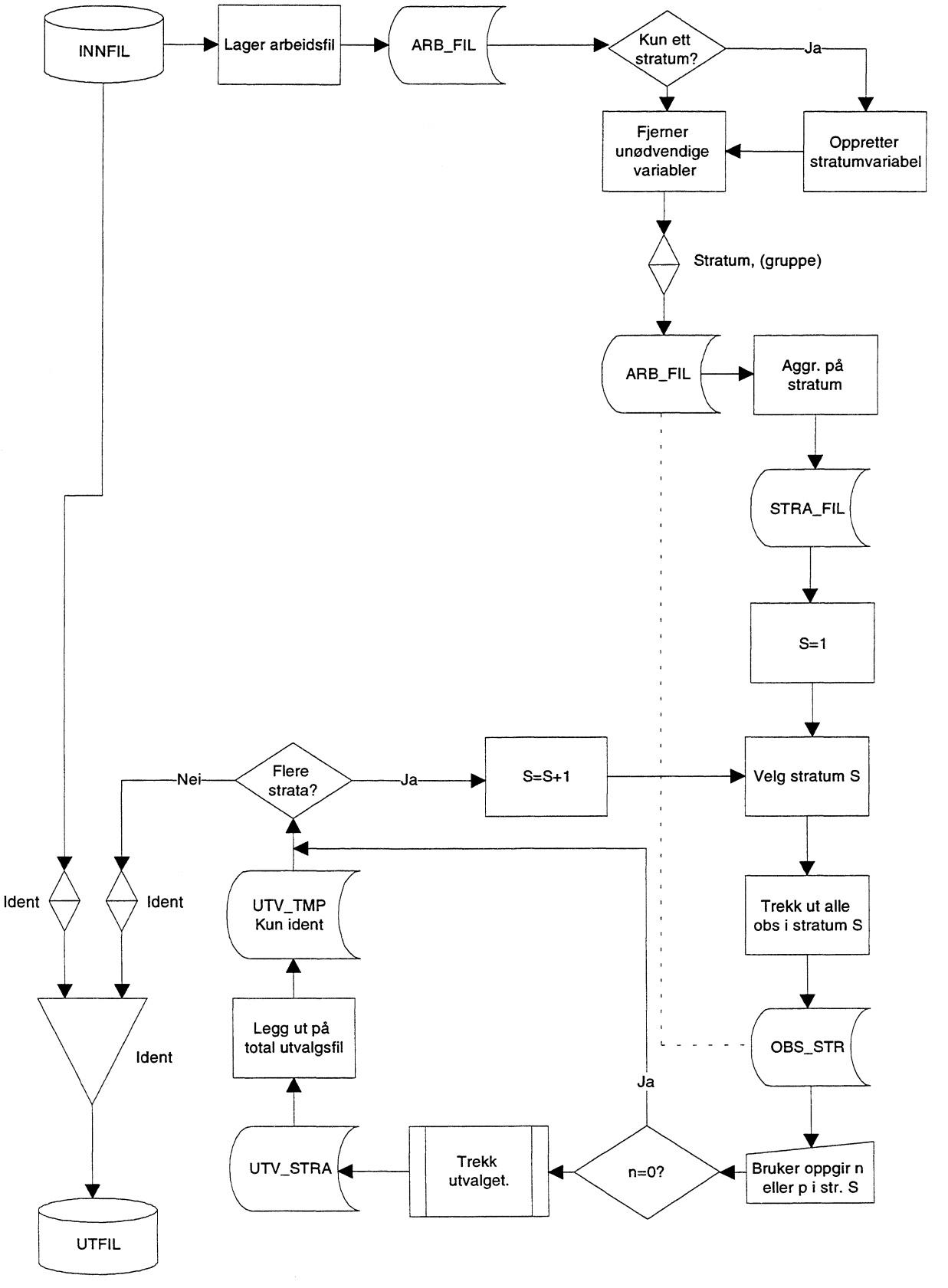
Cochran, W. G. (1977): *Sampling Technicues*, Wiley & Sons

Steen, A. C. Dokumentasjon av fysioterapeutundersøkelsen. Publiseres desember 1995.

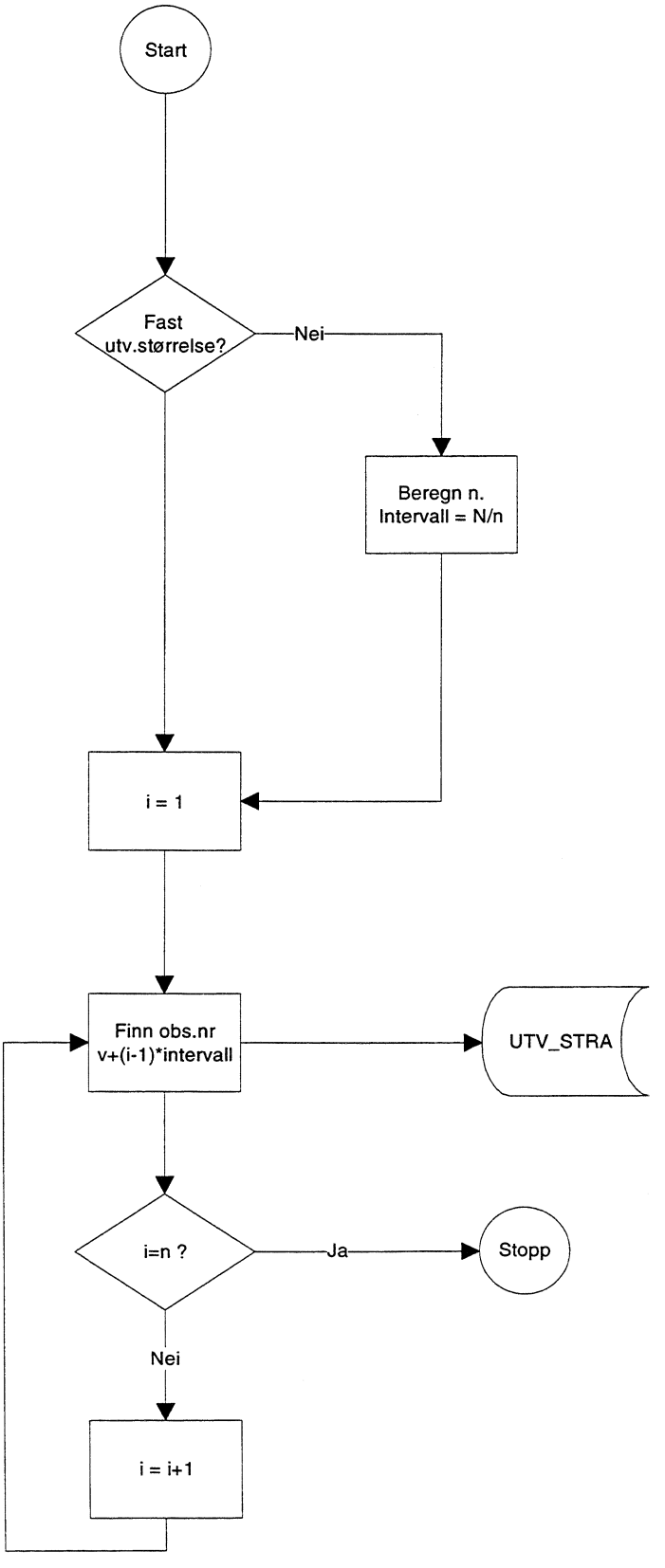
Sørbråten, A., *Inntekts- og kostnadsundersøkelsen for privatpraktiserende tannleger 1995,* Notater 95/43, Statistisk sentralbyrå.

Thomsen I. (1977): *Prinsipper og metoder for statistisk sentralbyrås utvalgsundersøkelser*, Sosiale og økonomiske studier 33.

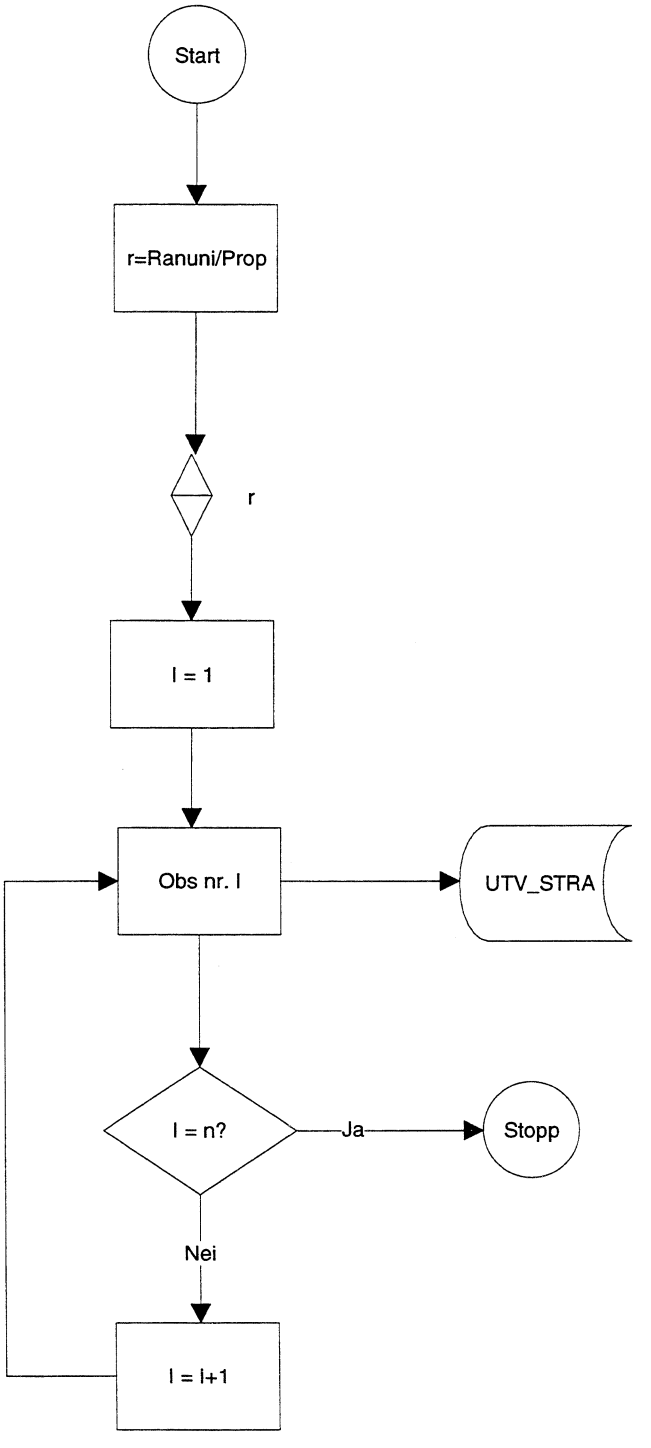
Dataflytdiagram - Trek_sys og Trek_pps



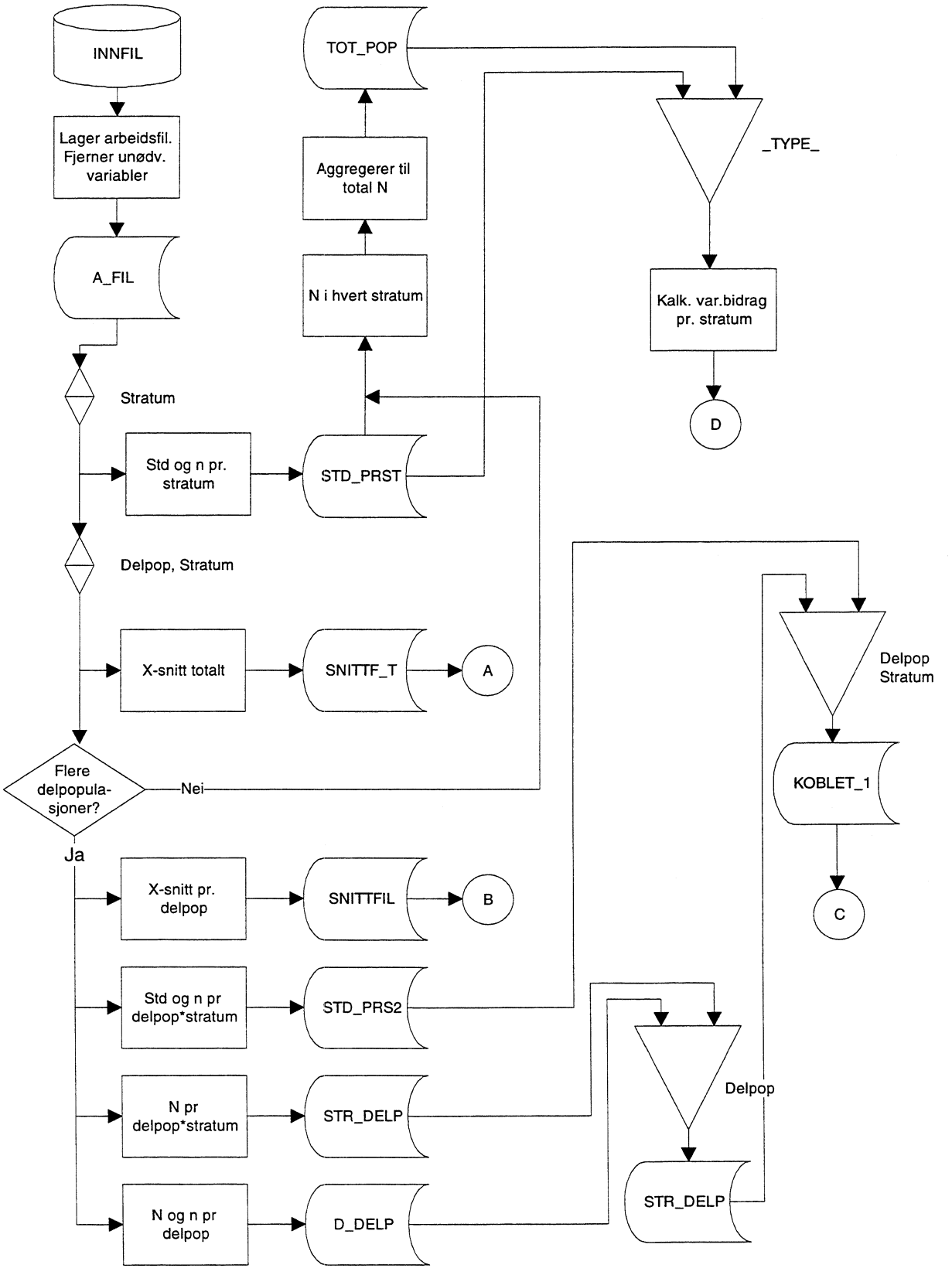
Dataflytdiagram - Trek_sys (Trekk utvalget)



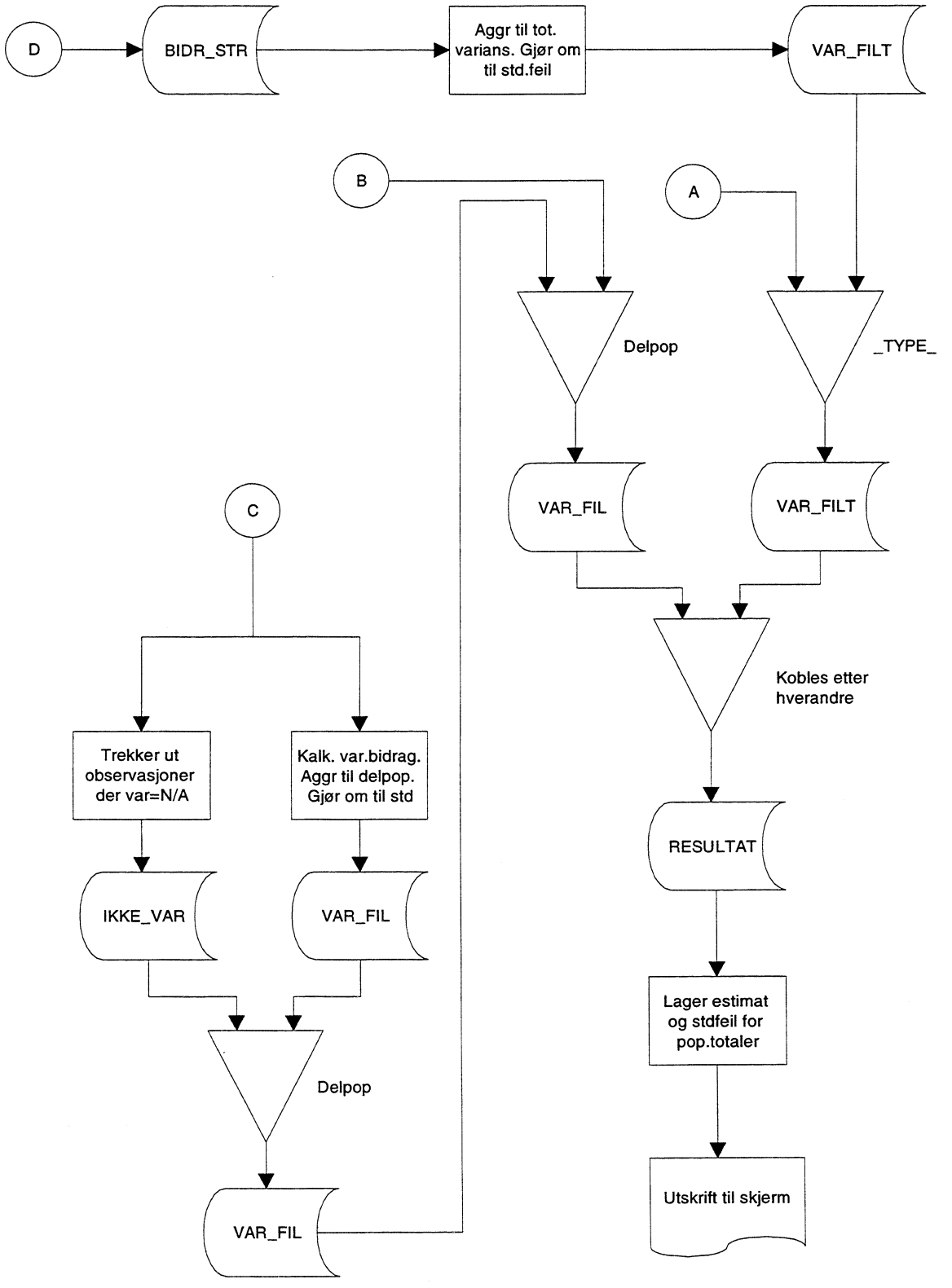
Dataflytdiagram - Trek_pps (Trekk utvalget)



Flytdiagram - Des_var.sas (side 1)



Flytdiagram - Des_var.sas (side 2)



```
************************************************************ ;
*  Makro:   Trek_sys.                                      ;
*  ********************************************************* ;
*  Laget av      : Egil Heilund                            ;
*  Sist oppdatert: 30.10.95                                ;
*  ********************************************************* ;
*                                                          ;
*  Makroen utfører et systematisk trekk. En kan velge om   ;
*  utvalgsstørrelsene skal oppgis i prosent eller i        ;
*  absolutt antall. Følgende forutsetninger er nødvendig for ;
*  å kjøre makroen:                                        ;
*                                                          ;
*  -  Innfilen (trekkgrunnlaget) må ha en unik variabel    ;
*     (identifikasjonsvariabel).                           ;
*                                                          ;
*     fil_i    Navn på innfil (trekkgrunnlaget).           ;
*     fil_u    Navn på utfil                               ;
*     ident    Navn på en unik variabel i innfilen         ;
*     j        Navn på variabelen som angir stratum        ;
*              Oppgi 0 dersom stratumvariabel ikke eksisterer ;
*     gruppe   Navn på 'ekstra' variabel som ønskes        ;
*              representativitet med hensyn på. Trekkgrunnlaget;
*              Trekkgrunnlaget vil da bli sorter på denne  ;
*              variabelen før trekking. Dersom slik variabel ;
*              ikke eksisterer, oppgi 0                    ;
*     fast     Indikerer hvorvidt utvalgsstørrelsene skal  ;
*              oppgis som prosent (0) eller absolutt antall (1);
*     z        Startverdi. Dersom z=0, vil startverdien bli ;
*              bestemt av klokken i maskinen.              ;
*              OBS: Dersom du ønsker å rekonstruere utvalget ;
*              senere, må du ikke sette z lik 0.           ;
*                                                          ;
*     En 'kjørerapport' vil bli skrevet til SAS-logen      ;
************************************************************ ;

%macro Trek_sys (fil_i, fil_u, ident, j, gruppe, fast, z);


*;
*  Sletter eventuell gammel utfil;
*;

 proc datasets library=work;
    delete UTV_TMP;


*;
*  Lager arbeidsfil.;
*;

data ARB_FIL;
   set &FIL_I;
run;


*;
*  Dersom variabel for stratum ikke eksisterer, opprettes denne;
*;

%if &j=0 %then %do;

   data ARB_FIL;
      set ARB_FIL;
      stratum=0;
      output;
   run;

   %let j=stratum;

%end;

*;
*  Fjerner unødvendige variable;
*;
```

```
%if &gruppe=0 %then %do;
   data ARB_FIL;
      set ARB_FIL (keep=&ident &j);
%end;

%else %do;
   data ARB_FIL;
      set ARB_FIL (keep=&ident &j &gruppe);
%end;


*;
*  Sorterer etter stratum, eventuelt også etter grupperingsvariabel;
*;

  %if &gruppe=0 %then %do;
     proc sort data=ARB_FIL; by &j;
  %end;
  %else %do;
     proc sort data=ARB_FIL; by &j &gruppe;
  %end;


*;
*  Lager fil over alle strata (en observasjon for hvert stratum);
*  Variabelen &n angir antall strata i populasjonen.            ;
*;

data STRA_FIL;
  set ARB_FIL;
  by &j;
  if first.&j then do;
     k+1;
     output;
     end;
  call symput ('n',k);
  keep &j;
run;


*;
*  Løkken for trekkingen starter her, en runde for hvert stratum;
*;

%do s=1 %to &n;

*;
*  Variabelen stra får i den 1. runde navnet til 1. stratum, etc;
*;

data _NULL_;
  peker=symget('s');
  pek=int(peker);
  set STRA_FIL point=pek;
    call symput ('stra', &j);
  stop;
run;

   %let p= ;    * Initierer p;

*;
*  Nå kalles opp et vindu som spør om trekksannsynlighet (fast=0);
*  eller utvalgsstørrelse (fast=1) i det aktuelle stratumet;
*;

   %window v_pros
    #10  @ 10 "Stratum                        :      &stra"
    #12  @ 10 'Trekksannsynlighet (prosent) :'
         @ 50 p    4 attr=highlight required=yes;

   %window v_fast
    #10  @ 10 "Stratum            :      &stra"
    #12  @ 10 'Utvalgsstørrelse   :'
         @ 40 p    5 attr=highlight required=yes;

  %if &fast=0 %then %display v_pros;
```

```
      %else %display v_fast;

  %if &p=0 %then %goto null_utv;


*;
*  Trekker ut alle observasjoner i aktuelt stratum;
*  Kun det ene av de to datastegene vil bli utført, avhengig;
*  av om stratumvariabelen er numerisk eller alfanumerisk;
*;

  data OBS_STR;
     set ARB_FIL;
     where &j="&stra";
  run;

  data OBS_STR;
     set ARB_FIL;
     where &j=&stra;
  run;


*;
*  Dersom utvalgsstørrelsen er oppgitt i prosent, omgjøres denne;
*  til antall;
*;

%if &fast=0 %then
  %do;

     data OBS_STR;
        set OBS_STR nobs=N;
        p=round((&p/100)*N);
        call symput ('p',p);
   %end;


*;
*  Selve trekket foregår her. De uttrukne observasjonene legges;
*  ut i filen UTV_STRA;
*;

  data UTV_STRA;

   interv=N/&p;
   v=ceil(interv * ranuni(N*&z));
   do i=1 to &p;
      x=v+(i-1)*interv;
      set OBS_STR nobs=N point=x;
      output;
      keep &ident;
   end;
   stop;

  run;


*;
*  Utvalgsenhetene i aktuelt stratum legges til den totale;
*  utvalgsfilen;
*;

proc append base=UTV_TMP data=UTV_STRA; run;

%null_utv:

%end;

*;
*  Utvalgsfilen kobles mot den orginale inn-filen;
*  og legges på utfil.;
*;

proc sort data=UTV_TMP; by &ident;
proc sort data=&fil_i ; by &ident;
```

```
data &fil_u;
   merge UTV_TMP (in=a) &fil_i (in=b);
   by &ident;
   if a;

*;
*   Finner antall observasjoner i populasjon og utvalg;
*;

data _NULL_;
  set &fil_u nobs=n;
  call symput ('ant_utv',n);
run;

data _NULL_;
  set ARB_FIL nobs=n;
  call symput ('ant_pop',n);
run;


  %put ************************************************************;
  %put Dato og klokkeslett                  &sysdate &systime      ;
  %put Innfil (trekkgrunnlag)               &fil_i                 ;
  %put Antall observasjoner i populasjonen  &ant_pop               ;
  %put Utfil                                &fil_u                 ;
  %put Antall observasjoner i utvalget      &ant_utv               ;
  %put Startverdi                           &z                     ;
  %put ************************************************************;

%mend Trek_sys;

run;
```

```
*************************************************************  ;
*  Makro:   Trekk_pps                                        ;
*  *********************************************************  ;
*  Laget av      : Egil Heilund                              ;
*  Sist oppdatert: 30.10.95                                  ;
*  *********************************************************  ;
*                                                            ;
*  Makroen utfører et trekk proporsjonalt med en angitt      ;
*  variabel. En kan velge hvorvidt utvalgsstørrelsene        ;
*  skal oppgis i prosent eller i                             ;
*  absolutt antall. Følgende forutsetninger er nødvendig for ;
*  å kjøre makroen:                                          ;
*                                                            ;
*  -  Innfilen (trekkgrunnlaget) må ha en unik variabel      ;
*     (identifikasjonsvariabel).                             ;
*  -  Innfilen må ha en numerisk variabel som utvalget skal  ;
*     trekkes proporsjonalt med hensyn på.                   ;
*  -  Variabelen som angir stratum må være numerisk. Dersom  ;
*     en slik variabel ikke eksisterer (kun ett stratum),    ;
*     oppgis parameteren til verdi 0.                        ;
*                                                            ;
*     fil_i    Navn på innfil (trekkgrunnlaget).             ;
*     fil_u    Navn på utfil                                 ;
*     ident    Navn på en unik variabel i innfilen           ;
*     j        Navn på variabelen som angir stratum          ;
*              Oppgi 0 dersom stratumvariabel ikke eksisterer;
*     prop     Variabelen som det skal trekkes proporsjonalt ;
*              med hensyn på.                                ;
*     fast     Indikerer hvorvidt utvalgsstørrelsene skal    ;
*              oppgis som prosent (0) eller absolutt antall (1);
*     z        Startverdi. Dersom z=0, vil startverdien bli  ;
*              bestemt av klokken i maskinen.                ;
*              OBS: Dersom du ønsker å rekonstruere utvalget ;
*              senere, må du ikke sette z lik 0.             ;
*                                                            ;
*     En 'kjørerapport' vil bli skrevet til SAS-logen        ;
*************************************************************  ;

%macro Trek_pps (fil_i, fil_u, ident, j, prop, fast, z);


*;
*  Sletter eventuell gammel utfil;
*;

 proc datasets library=work;
    delete UTV_TMP;


*;
*  Lager arbeidsfil.;
*;

data ARB_FIL;
   set &FIL_I;
run;

*;
*  Dersom variabel for stratum ikke eksisterer, opprettes denne;
*;

%if &j=0 %then %do;

   data ARB_FIL;
      set ARB_FIL;
      stratum=0;
      output;
   run;

   %let j=stratum;

%end;

*;
```

```
*  Fjerner unødvendige variable og sorterer etter stratum;
*;

   data ARB_FIL;
      set ARB_FIL (keep=&ident &j &prop);

     proc sort data=ARB_FIL; by &j;


*;
*  Lager fil over alle strata (en observasjon for hvert stratum);
*  Variabelen &n angir antall strata i populasjonen.           ;
*;

data STRA_FIL;
  set ARB_FIL;
  by &j;
  if first.&j then do;
     k+1;
     output;
     end;
  call symput ('n',k);
  keep &j;
run;


*;
*  Løkken for trekkingen starter her, en runde for hvert stratum;
*;

%do s=1 %to &n;

*;
*  Variabelen stra får i den 1. runde navnet til 1. stratum, etc;
*;

data _NULL_;
  peker=symget('s');
  pek=int(peker);
  set STRA_FIL point=pek;
    call symput ('stra', &j);
  stop;
run;

   %let p= ;   * Initierer p;

*;
*  Nå kalles opp et vindu som spør om trekksannsynlighet (fast=0);
*  eller utvalgsstørrelse (fast=1) i det aktuelle stratumet;
*;

   %window v_pros
    #10  @ 10 "Stratum                        :      &stra"
    #12  @ 10 'Trekksannsynlighet (prosent) :'
         @ 50 p    4 attr=highlight required=yes;

   %window v_fast
    #10  @ 10 "Stratum            :      &stra"
    #12  @ 10 'Utvalgsstørrelse   :'
         @ 40 p    5 attr=highlight required=yes;

  %if &fast=0 %then %display v_pros;
      %else %display v_fast;


  %if &p=0 %then %goto null_utv;


*;
*  Trekker ut alle observasjoner i aktuelt stratum;
*  Kun den ene av de to datastegene vil bli utført,;
*  avhengig av om stratumvariabelen er numerisk eller alfanumerisk;
*;
```

```
  data OBS_STR;
     set ARB_FIL;
     where &j="&stra";

  data OBS_STR;
     set ARB_FIL;
     where &j=&stra;


*;
*  Dersom utvalgsstørrelsen er oppgitt i prosent,;
*  omgjøres denne til antall.;
*;

  %if &fast=0 %then
  %do;
     data OBS_STR;
        set OBS_STR nobs=n;
        p=round((&p/100)*n);
        call symput ('p',p);
  %end;

*;
*  Selve trekket foregår her. De uttrukne observasjonene legges;
*  ut i filen UTV_STRA;
*;

  data OBS_STR;
     set OBS_STR nobs=N;
     r=ranuni(N*&z)/&prop;

  proc sort; by r;

  data UTV_STRA;
     set OBS_STR;
     i+1;
     if i le &p then output;


*;
*  Utvalgsenhetene i aktuelt stratum legges til den totale;
*  utvalgsfilen;
*;

proc append base=UTV_TMP data=UTV_STRA; run;

%null_utv:

%end;

*;
*  Utvalgsfilen kobles mot den orginale inn-filen;
*  og legges på utfil.;
*;

data UTV_TMP;
  set UTV_TMP (keep=&ident);

proc sort data=UTV_TMP; by &ident;
proc sort data=&fil_i ; by &ident;

data &fil_u;
   merge UTV_TMP (in=a) &fil_i (in=b);
   by &ident;
   if a;

*;
*   Finner antall observasjoner i populasjon og utvalg;
*;

data _NULL_;
  set &fil_u nobs=n;
  call symput ('ant_utv',n);
run;
```

```
data _NULL_;
  set ARB_FIL nobs=n;
  call symput ('ant_pop',n);
run;


  %put ************************************************************;
  %put Dato og klokkeslett                  &sysdate &systime      ;
  %put Innfil (trekkgrunnlag)               &fil_i                 ;
  %put Antall observasjoner i populasjonen  &ant_pop               ;
  %put Utfil                                &fil_u                 ;
  %put Antall observasjoner i utvalget      &ant_utv               ;
  %put Startverdi                           &z                     ;
  %put ************************************************************;

%mend Trek_pps;

run;
```

```
*****************************************************************;
*  Makro:   Des_var                                              ;
*                                                                ;
*  Programmet lager en makro som beregner designvarians.         ;
*                                                                ;
*  Parametre  fil        Navn på datafilen                       ;
*             str        Variabel med stratumidentifikasjon      ;
*             v          Variabel med vekt (oppblåsningsfaktor   ;
*             x          Variabel som skal analyseres (sum, snitt) ;
*             Delp       Grupperingsvariabel (f.eks. kjønn, geogr.);
*                                                                ;
* Forutsetninger:  Enkelt tilfeldig utvalg innenfor hvert stratum ;
*                  En oppblåsningsfaktor pr. stratum             ;
*                  Beregner varians for gjennomsnittsverdi.      ;
*                                                                ;
*  Estimat og standardfeil avrundes til nærmeste heltall.        ;
*****************************************************************;

%macro Des_var (fil, str, v, x, delp);


*;
*  Lager arbeidsfil, fjerner unødvendige variabler;
*  Dersom ikke delpopulasjonsvariabel (delp=0), lages denne;
*;

%if &delp=0 %then
   %do;

      %let kun_en=1;

      data A_FIL;
         set &fil;
         Gruppe='Totalt';
         keep &str &v &x Gruppe;
      run;

      %let delp=Gruppe;

   %end;

%else
   %do;

      %let kun_en=0;

      data A_FIL;
         set &fil;
         tmp=put(&delp,15.);
         keep &str &v &x tmp;

      proc datasets library=work;
         modify A_FIL;
         rename tmp=&delp;

   %end;


proc sort data=A_FIL; by &str &v;

*;
*  Finner standardavviket og antall observasjoner.               ;
*  i hvert stratum.                                              ;
*;


proc means data=A_FIL noprint;
  by &str &v;
  var &x;
  output out=STD_PRST std=std n=n_utv;
run;
```

```
*;
*  Finner veiet gjennomsnitt totalt;
*;

proc sort data=A_FIL; by &delp &str;

proc means data=A_FIL noprint;
  weight &v;
  var &x;
  output out=SNITTF_T mean=Snitt sumwgt=Ant_pop;


*;
*  Hopper over variansberegningene pr. delpopulasjon dersom      ;
*  dette ikke er aktuelt                                         ;
*;

%if &kun_en=1 %then %goto total;


*;
*  Finner veiet gjennomsnitt totalt;
*;

proc means data=A_FIL noprint;
  by &delp;
  weight &v;
  var &x;
  output out=SNITTFIL mean=Snitt sumwgt=Ant_pop;


*;
*  Finner standardavviket og antall observasjoner.               ;
*  i hver kombinasjon av delp*stratum.                           ;
*;

proc means data=A_FIL noprint;
  by &delp &str;
  var &x;
  output out=STD_PRS2 std=std n=n_utv;
run;


*;
*  Finner antall i populasjon pr (delpopulasjon * stratum)   ;
*;

proc means data=A_FIL noprint;
  by &delp &str;
  var &v;
  output out=STR_DELP sum=ds_popul;
run;


*;
*  Finner antall i utvalg og populasjon pr. delpopulasjon ;
*;

proc means data=A_FIL noprint;
  by &delp;
  var &v;
  output out=D_DELP sum=d_popul;
run;


*;
* Kobler populasjonsantall pr. delpopulasjon sammen med;
* populasjonsantall pr. delp*stratum;
*;


data STR_DELP;
  merge STR_DELP (keep=&delp &str ds_popul)
        D_DELP   (keep=&delp d_popul);
  by &delp;
```

```
*;
* Kobler på standardavvik og antall i utvalg pr. delp*stratum;
* Beregner ett variansbidrag pr. delp*stratum;
* Identifiserer ledd som ikke kan beregnes (err=1);
*;

data KOBLET_1;
   merge STD_PRS2 (in=a) STR_DELP (in=b);
   by &delp &str;

   bidrag=( (ds_popul/d_popul) * sqrt(1-n_utv/ds_popul) *
          std/sqrt(n_utv) ) ** 2 ;

   if bidrag='' then err=1;
   else err=0;


*;
* Aggregerer til bidrag pr. delpopulasjon;
*;

proc means data=KOBLET_1 noprint;
   by &delp;
   var bidrag n_utv;
   output out=VAR_FIL sum=Varians antall;


*;
*  Finner hvilke av delpopulasjonene ikke kan regnes ut variansen for;
*  Omgjør varians til standardfeil. Kobler på estimatene i hver;
*  delpopulasjon (SNITTFIL);
*;

proc means data=KOBLET_1 noprint;
   by &delp;
   var err;
   output out=IKKE_VAR sum=Mangler;


data VAR_FIL;
   merge VAR_FIL IKKE_VAR;
   by &delp;

   if mangler gt 0 then Varians=.;
   Std_sn=sqrt(Varians);

   merge SNITTFIL VAR_FIL;
   by &delp;


*****************************************************************;
*  Har nå beregnet standardfeil for hver delpop. Må nå beregne  ;
*  standardfeil for hele massen                                 ;
*****************************************************************;

%total:

*;
*  Beregner antall i populasjonen pr. stratum;
*;

data STD_PRST;
   set STD_PRST;
   n_pop=n_utv*&v;


*;
*  Summerer populasjonsantall i hvert stratum og trekker denne ut;
*  i en egen fil;
*;

proc means data=STD_PRST noprint;
   var n_pop;
   output out=TOT_POP sum=n_tot;
```

```
*;
*  Kalkulerer variansbidrag pr. stratum;
*;


data KOBLET_2;
   merge TOT_POP STD_PRST;
   by _type_;
   bidrag=( (n_pop/n_tot) * sqrt(1-n_utv/n_pop)
          * std/sqrt(n_utv) ) ** 2;


*;
*  Summerer variansbidragene;
*;

proc means data=KOBLET_2 noprint;
   var bidrag n_utv;
   output out=VAR_FILT sum=Varians antall;


*;
*  Omgjør varians til standardavvik;
*  Lager label for delpopulasjon (totalt);
*;

data VAR_FILT;
   set;
   Std_sn=sqrt(Varians);
   &delp='Totalt';

   merge SNITTF_T VAR_FILT;
   by _type_;


*;
*  Kobler sammen resultatene pr. delpopulasjon (VAR_FIL) og;
*  for totalen (VAR_FILT).;
*  Dersom uten delpopulasjon, benyttes kun VAR_FILT;
*  Lager estimat og std.feil for sum;
*;

%if &kun_en=0 %then
   %do;

      data RESULTAT;
         set VAR_FIL VAR_FILT;

   %end;

%else
   %do;

      data RESULTAT;
         set VAR_FILT;

   %end;

data RESULTAT;
   set RESULTAT;
   Sum     = Snitt*Ant_pop;
   Std_sum = Std_sn*Ant_pop;
   Rel_std = Std_sn/Snitt;


*;
*  Skriver ut;
*;

proc format;

   picture ssb (round)
       -999999999999.5 <-< 0 = '000 000 000 009' (prefix='-')
       0 <-< 999999999999.5  = '000 000 000 009';

proc print data=RESULTAT split='*';
```

```
  title "Estimat og standardfeil for snitt og sum, variabel &x.";
  label Antall  ='Antall*i utvalg      '
        Ant_pop ='Oppblåst*antall'
        Snitt   ='Estimert*gj.snitt'
        Std_sn  ='Estimert*std.feil'
        Sum     ='Estimert*sum'
        Std_sum ='Estimert*std.feil'
        Rel_std ='Relativ*std.feil';

  format Antall ssb. Ant_pop ssb. Snitt ssb.
         Std_sn ssb. Sum ssb. Std_sum ssb. Rel_std 6.3;
  var Antall Ant_pop Snitt Std_sn Rel_std Sum Std_sum;
  id &delp;
run;

%mend Des_var;
run;
```

Statistisk sentralbyrå

Oslo
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger
Postboks 1260
2201 Kongsvinger

Telefon: 62 88 50 00
Telefaks. 62 88 50 30

ISSN 0806-3745

Statistisk sentralbyrå
Statistics Norway