Statistics Norway
Statistisk sentralbyrå

Nina Hagesæther og Li-Chun Zhang

Om estimeringsusikkerhet og utvalgsplan i AKU

Stabsavdeling/Seksjon for metoder og standarder

Innhold

1. Innledning

Utvalgsplanen for Arbeidskraftsundersøkelsen (AKU) 2005 kan karakteriseres som følger (Zhang og Vedø, 2004):

- Stratifisering etter alle 19 fylker.
- Primære trekkenheter er familier i henhold til Det sentrale folkeregisteret (DSF).
- DSF-familiene innen hvert fylke trekkes systematisk med antatt tilfeldig sortering.
- Alle personer i hver familie mellom 16 og 74 år inkluderes i utvalget.

Et spørsmål som lenge har blitt diskutert er valget av utvalgsenhet. Det er kjent at utvalgsvariansen for estimering av sysselsetting kan reduseres ved å trekke personer direkte istedenfor familier som nå er tilfellet, samtidig som kostnaden ved AKU vil øke ved en slik omlegging. For arbeidsledighet har valget av enhet liten betydning for utvalgsvariansen, vist ved simulering av Vedø og Rafat (2003). For sysselsetting er imidlertid størrelsen på variansøkning ved et familieutvalg uklar, da beregningene deres er lagt på en annen skala enn det AKU i realiteten er. Zhang og Vedø (2004) viste at et stratifisert familieutvalg har omtrent samme utvalgsvarians som et enkelt tilfeldig personutvalg, samtidig som et stratifisert personutvalg har enda mindre utvalgsvarians.

I den første delen av dette notatet studerer vi problemet nærmere ved å se på hvordan variansen til en kalibreringsestimator endres med hensyn til utvalgs- og estimeringsenhet, i samspill med tilleggsvariabler og trekksannsynligheter. Den mest interessante konklusjonen er som følger:

- Variansøkning ved å bruke familier som utvalgsenheten kan tas bort ved å la person være estimeringsenhet i stedet for familie. Et klyngeutvalg av familier, der person er estimeringsenheten, har omtrent like stor utvalgsvarians som et direkte personutvalg.

For både sysselsetting og arbeidsledighet er det derfor unødvendig å legge om AKU til et personutvalg. Konklusjonen er utelukkende basert på betraktninger av utvalgsvarians. Så vidt vi kjenner til er det første gang konklusjonen er fremsatt på denne måten.

Det faktiske estimeringsopplegget i AKU er mer komplisert enn enkel kalibrering. Vektene justeres tre ganger med blant annet fylkesvis kalibrering etter en etterstratifisering for hele landet (Heldal, 2000). Det er vanlig at man i variansberegning enten forenkler situasjonen til etterstratifisering eller kalibrering. I den andre delen av notatet beregner vi variansen i AKU under det faktiske opplegget med multiple justeringer av vektene. To metoder, linearisering (Zhang, 2006) og bootstrap (Shao, 1996), blir brukt og sammenlignet. Variansestimering med lineariseringsmetoden er enkel å implementere, og gir nesten samme resultater som bootstrap. Metoden bør kunne tas i bruk for kvalitetsrapportering.

2. Om utvalgs- og estimeringsenhet

2.1 Problemstilling

Vi ser bort fra frafall i dette avsnittet. Betrakt tre ulike scenarioer for utvalgs- og estimeringsenheter:

- Personutvalg og person som estimeringsenhet
- Familieutvalg og person som estimeringsenhet
- Familieutvalg og familie som estimeringsenhet

Vi ser bort fra personutvalg og familie som estimeringsenhet. Effekten av enhetsvalg kan tenkes å være påvirket av estimeringsmetode, spesielt hvilke tilleggsvariabler som brukes. Vi ser på to situasjoner:

- Kun informasjon om populasjonsstørrelse benyttes, enten antall familier eller antall personer
- Tilleggsopplysinger om kjønn, alder (12 grupper) og register sysselsettingsstatus (4-deling) benyttes

Også trekksannsynlighet kan påvirke utvalgsvariansen. Følgende valg er aktuelle:

- Trekking med lik sannsynlighet eller, for familieutvalg, en sannsynlighet proporsjonal med familiestørrelsen. Det siste er omtrent tilfellet når person er trekkenhet, men hele familien til de uttrukne personene inkluderes i utvalget.
- Stratifisering

Enhver estimator vi skal se på i dette avsnittet kan formuleres som en generalisert regresjonsestimator (GREG, Särndal *et al*., 1992), som er den vanligste kalibreringsesimatoren.

2.2 Variansestimering for GREG

Først skal vi kort oppsummere GREG estimatoren og hovedideen i variansestimering. La $U = \{1,...,N\}$ betegne en populasjon med enheter $i = 1,...,N$, og $s = \{1,...,n\}$ betegne utvalget med enheter $i = 1,...,n$. La a_i betegne utvalgsvekten gitt ved invers av trekksannsynligheten. GREG estimatoren er et spesielt tilfelle av kalibreringsestimering. La w betegne den kalibrerte vekten, som er gitt ved

$$w_{nx1} = a_{nx1} \otimes g_{nx1}$$

der $\otimes$ betyr at vektorene multipliseres element for element, og n er antallet enheter i utvalget, og

$$g_{nx1} = 1_{nx1} + [(X_{Kx1} - \hat{X}_{Kx1})^T \times \hat{A}^{-1}_{KxK} \times Xdesign_{nxK}{}^T]^T$$

Her er matrisemultiplikasjon angitt med $\times$. X er en vektor som inneholder K antall kalibreringstotaler. Xdesign er en designmatrise med en linje for hver enhet i utvalget og en kolonne for hver komponent i X. A er definert ved $Xpop_{N\times K}{}^T \times Xpop_{N\times K}$, der Xpop er designmatrisen for hele populasjon, satt opp på samme måten som Xdesign. $\hat{X}$ er estimatet for X basert på a, dvs. $\hat{X} = Xdesign^T \times a$. $\hat{A}$ er estimatet for A basert på a, dvs. $\hat{A} = (Xdesign^T \otimes a) \times Xdesign$, der $\otimes$ betyr at vektoren a multipliseres elementvis med hver kolonne i Xdesign. $\hat{A}^{-1}$ er invers til $\hat{A}$.

Hovedideen i lineariseringsmetoden for variansestimering ligger i følgende betraktning. For enhver gitt konstant vektor B_{Kx1}, kan interessevariabelen y_i transformeres til $\varepsilon_i = y_i - x_i^T B$, der x_i^T er raden i Xpop som svarer til enheten i. Vi har dermed at estimatet for totalen $Y = \sum_{i\in U} y_i$ er gitt ved

$$\hat{Y} = \sum_{i\in s} w_i y_i = \sum_{i\in s} w_i (x_i^T B + \varepsilon_i) = (\sum_{i\in s} w_i x_i)^T B + \sum_{i\in s} a_i g_i \varepsilon_i = X^T B + \sum_{i\in s} a_i g_i \varepsilon_i$$

siden vektene w_{nx1} er kalibrert med hensyn til X. Designforventningen til g_i er lik 1, slik at en linear tilnærming til $\hat{Y}$ og dens varians er ganske enkelt gitt ved

$$\hat{Y} \approx X^T B + \sum_{i\in s} a_i \varepsilon_i \qquad V(\hat{Y}) \approx V(\sum_{i\in s} a_i \varepsilon_i)$$

En tilnærmet varians til $\hat{Y}$ følger nå den generelle variansformel for en Horvitz-Thompson estimator, det vil si $\sum_{i\in s} a_i\varepsilon_i$. I praksis har vi som regel ikke B på forhånd, og dermed heller ikke $\varepsilon_{n\times l}$. $\hat{B}$ beregnes under en lineær regresjonsmodell som svarer til transformasjonen fra $y_{n\times l}$ til $\varepsilon_{n\times l}$, og settes inn for B. En variansestimator fås derfor ved å erstatte $\varepsilon_{n\times l}$ med et estimat:

$$e_{nx1} = y_{nx1} - \hat{y}_{nx1} = y - Xdesign \times \hat{A}^{-1} \times (Xdesign^T \times a \otimes y)$$

Imidlertid anbefaler Särndal *et al.* (1992) å bruke $g_i e_i$ istedenfor bare e_i . En grunn er at det gir en variansestimator som både er konsistent med hensyn utvalgtrekking og tilnærmet forventningsrett med hensyn til regresjonsmodellen. En annen grunn er at uten g_i vil man overse usikkerheten i estimeringen av ε_i .

2.2.1 Personutvalg og person som estimeringsenhet

Anta at inngangsvekten $a_i = N/n$ er lik for hver person i utvalget, det vil si enkel tilfeldig trekking. N er totalt antall personer i populasjonen U. I tilfellet kalibrering mot populasjonstotalet har vi $X = N$ og $x_i = 1$. I tilfellet kalibrering for kjønn, alder og sysselsettingsstatus er X en vektor med marginale tall for kjønn (to klasser), alder (11 klasser) og sysselsettingsstatus (tre klasser) hentet fra register. En aldersklasse og en klasse for sysselsettingsstatus er slettet for å unngå singularitet. $\hat{X}$ er nå de marginale tallene estimert ved utvalget basert på a, det vil si $Xdesign^T \times a$. De to første komponentene i x_i svarer til kjønn, de 11 neste alder og de tre siste sysselsettingsstatus. For eksempel, for en kvinne på 40 år sysselsatt innen primærnæringen, vil den tilsvarende raden i designmatrisen få verdiene {0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0}, mens en 65 år gammel mann sysselsatt i sekundærnæringen vil få verdiene {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0}.

En estimator for Y, der Y kan være totalt antall sysselsatte eller arbeidsledige i den endelige populasjonen U, er $\hat{Y} = \sum_{i\in s} w_i y_i$. Variansestimatoren til $\hat{Y}$ er gitt ved

$$\hat{V}(\hat{Y}) = \sum\sum_{i,j\in s} (a_i a_j - a_{i,j})(g_i e_i)(g_j e_j) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{n}\frac{(g_i e_i)^2}{n-1}$$

$$\approx \frac{N^2}{n^2}\sum_{i=1}^{n}(g_i e_i)^2 = \sum_{i=1}^{n}(w_i e_i)^2$$

der $a_{i,j} = \frac{N(N-1)}{n(n-1)}$ er invers av andreordens trekksannsynlighet. Den første tilnærming følger av $\frac{1}{N} \approx 0$ og $\frac{1}{n-1} \approx \frac{1}{n}$. Det siste følger av $w_i = a_i g_i = \frac{N}{n} g_i$.

Vi kan komme frem til en variansestimator tilsvarende det siste uttrykket med en modellbasert tankegang. Modellen er $y = Xdesign \times \beta + \varepsilon$, der ε_i har forventning 0 og varians σ_i^2 , og $\hat{\varepsilon}_i^2 = e_i^2 = \hat{\sigma}_i^2$. Vi betinger da $\hat{Y}$ på designmatrisen Xdesign og utvalget s, slik at den eneste variasjonen i $\hat{Y}$ er variasjonen i y_i . Samtidig antar vi at y_i er uavhengige av hverandre. Under disse forutsetningene er variansen til $\hat{Y}$ gitt ved $V(\hat{Y}) = \sum_{i\in s} w_i^2\sigma_i^2$, og en variansestimator blir da $\hat{V}(\hat{Y}) = \sum_{i\in s} w_i^2 e_i^2$.

2.2.2 Familieutvalg og person som estimeringsenhet

Anta nå et enkelt tilfeldig ettrinns klyngeutvalg. Først trekkes familier som primære utvalgsenheter, deretter intervjues alle personer som tilhører familien. Inngangsvekten er $a_i = M/m$, der M og m er henholdsvis antall familier i populasjonen og utvalget. En tilnærmet varians for $\hat{Y}$ i tilfellet totrinns klyngeutvalg med element som estimeringsenhet (Särndal *et al*., 1992) er

$$AV(\hat{Y}) = AV_{PSU} + AV_{SSU}$$

der PSU er primær og SSU sekundær utvalgsenhet, det vil si variansen ved henholdsvis første og andre trinnet. Siden alle personer i en trukket familie inkluderes i utvalget, har vi ikke variansen i det andre trinnet. La $z_k = \sum_{i \in k} g_i e_i$ være justerte residualer beregnet på personer aggregert til familienivå. La $a_{k,l} = \frac{M(M-1)}{m(m-1)}$. Variansestimatoren for AV_{PSU} kan forenkles til følgene uttrykk:

$$\hat{V}(\hat{Y}) = \sum\sum_{k,l \in S} (a_k a_l - a_{k,l}) z_k z_l = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) \sum_{k=1}^{m} \frac{z_k{}^2}{m-1} \approx \frac{M^2}{m^2} \sum_{k=1}^{m} z_k{}^2$$

2.2.3 Familieutvalg og familie som estimeringsenhet

Anta familieutvalg og familie som estimeringsenhet. Både $\hat{Y}$ og dens varians fås på samme måten som i avsnitt 2.2.1, men vi må først aggregere y_i til familienivå, dvs. $y_k = \sum_{i \in k} y_i$. Inngangsvekten er $a_k = M/m$. I tilfellet kalibrering mot populasjonstotalet har vi nå $X = M$ og $x_k = 1$. I tilfellet kalibrering for kjønn, alder og sysselsettingsstatus er X som definert ovenfor, og $x_k = \sum_{i \in k} x_i$.

2.3 Empiriske resultater

La P-P betegne personutvalg og person som estimeringsenhet, C-P familieutvalg med person som estimeringsenhet, og C-C familieutvalg med familie som estimeringsenhet. Betegn med AKU kalibrering for kalibrering mot kjønn, alder og sysselsettingsstatus i register, og enhetskalibrering for kalibrering kun mot populasjonsstørrelse (N eller M).

Data er hentet fra første kvartal 2005, der N = 3 312 011 og n = 21 525 personer. I beregningen antar vi for enkelhetsskyld $M/m = N/n$. Tabell 1 gir en oversikt over standardfeilene både for arbeidsledighet og sysselsetting. Forskjeller i en bestemt kolonne i tabellen viser effekten av enheter, mens forskjeller i en rad viser effekten av tilleggsvariabler.

Tabell 1: Oversikt over standardavvik for sysselsetting og arbeidsledighet. P-P = personutvalg og person som estimeringsenhet, C-P = familieutvalg og person som estimeringsenhet, C-C = familieutvalg og familie som estimeringsenhet. Enhetskalibrering = kalibrering mot populasjonsstørrelse, AKU kalibrering = kalibrering mot kjønn, alder og register sysselsettingsstatus.

	Enhetskalibrering		AKU kalibrering	
	Sysselsatt	Arbeidsledig	Sysselsatt	Arbeidsledig
C-C	15 791	3 958	7 094	4 192
C-P	10 859	3 931	7 071	4 152
P-P	10 341	3 856	6 948	4 077

For sysselsetting ser vi at:

- En reduksjon i standardavviket fra 15 791 (C-C og enhetskalibrering) til 6 948 (P-P og AKU kalibrering) kan dekomponeres i tre steg. Først reduseres standardavviket til 10 859 ved å bruke personer som estimeringsenhet til tross for at familier er utvalgsenhet (det vil si C-P og enhetskalibrering); deretter reduseres standardavviket ytterligere til 7 071 ved hjelp av tilleggsvariablene (det vil si C-P og AKU kalibrering); til slutt får vi en meget liten redusering ved å trekke personer direkte i utgangspunktet (det vil si P-P og AKU kalibrering).
- Alternativt kan man gå først fra 15 791 (C-C og enhetskalibrering) til 10 859 (C-P og enhetskalibrering), deretter til 10 341 (P-P og enhetskalibrering), og til slutt 6948 (P-P og AKU kalibrering). Igjen reduseres variansen først og fremst på grunn av endringen i estimeringsenhet og bruken av tilleggsvariabler. Valget av utvalgsenhet har nesten ingen effekt i tillegg.
- Et tredje alternativ er å gå først fra 15 791 (C-C og enhetskalibrering) til 7 094 (C-C og AKU kalibrering). Ved å endre utvalget fra et familieutvalg til et personutvalg (fra C-P til P-P, begge med AKU kalibrering) får vi kun en liten endring i variansen (til 6 948).

For arbeidsledighet påvirkes ikke standardavviket i særlig stor grad, uansett utvalgs- og estimeringsenhet og tilleggsvariabler.

2.4 Dekomponering av varians

Det kanskje mest interessante resultatet i Tabell 1 dreier seg om standardavvik for estimatoren ved enhetskalibrering i tilfellene C-C, C-P og P-P. Egentlig er alle parvise forhold velkjente.

- Forskjellen mellom C-C og P-P er kjent som designeffekten av klyngeutvalg.
- Forskjellen mellom C-C og C-P er blitt diskutert i forbindelse med valget mellom Horvitz-Thompson og ratio-to-size estimator (Cochran, 1977, Kapittel 9A).
- Forskjellen mellom C-P og P-P har sammenheng med "designeffektformelen" (Kish, 1987; Gabler, *et al.*, 1999) i tilfellet flertrinnsutvalg og vekteklasser. Også Vedø og Rafat (2003) inneholder sammenligning av lignende scenarioer.

Følgende konklusjon oppstår i det de tre scenarioene settes under ett, nemlig at man ikke taper effisiens ved å trekke et klyngeutvalg så lenge man holder seg til element som estimeringsenhet.

Til tross for at eksakt eller tilnærmet variansformel finnes i alle tre situasjonene, er det ikke uten videre lett å se resultatet generelt. I det følgende skal vi benytte en annen fremgangsmåte. Først setter vi opp en populasjonsmodell med parametere som har klare tolkninger. Deretter finner vi en varians under modellen som er tilnærmet lik utvalgsvarians under hvert av de 3 scenarioene. Dette gjør det mulig å tolke forskjellen i utvalgvarianser ved hjelp av parametere i modellen. Resultatet blir en dekomponering av variansen i tilfellet C-C, som klart viser hvordan den reduseres først til tilfellet C-P og deretter P-P.

2.4.1 Modell

La (ki) betegne element *i* fra klynge *k*. Anta følgende modell på elementnivå:

$$y_{ki} = \mu + \varepsilon_{ki}$$

der $E_M(\varepsilon_{ki}) = 0$, $V_M(\varepsilon_{ki}) = \sigma^2$, $\mathrm{cov}_M(\varepsilon_{ki}, \varepsilon_{kj}) = \rho\sigma^2$ for to elementer i samme klynge og $\mathrm{cov}_M(\varepsilon_{ki}, \varepsilon_{lj}) = 0$ for to elementer i forskjellig klynge. Alle forventingene her er tatt med hensyn til modellen. Dette er den såkalte "intracluster correlation model" som er vanlig i studier av klyngeeffekt (se for eksempel Gabler *et al*., 1999), der korrelasjonen i interessevariabelen mellom to elementer innen samme klynge angis med ρ. På klyngenivået har vi da

$$y_k = \sum_i y_{ki} = \sum_i (\mu + \varepsilon_{ki}) = N_k\mu + \sum_i \varepsilon_{ki} = N_k\mu + \varepsilon_k$$

der N_k er antallet elementer i klynge. Vi antar at $V_M(N_k) = \tau^2$ og $\mathrm{cov}_M(N_k, \varepsilon_k) = 0$.

2.4.2 Utvalgsvarians P-P ved enhetskalibrering

I tilfellet P-P med enkel tilfeldig trekking har vi

$$\hat{Y}_{P-P} = \sum_{(ki)\in s} a_{ki} y_{ki} = \sum_{(ki)\in s} \frac{N}{n} y_{ki}.$$

Vi ser bort fra den endelige populasjonsfaktoren (epf), slik at utvalgsvariansen er gitt ved

$$V_D(\hat{Y}_{P-P}) \approx \frac{N^2}{n} \frac{\sum_{(ki)\in U} (y_{ki} - \bar{Y})^2}{N-1} \approx N^2 \frac{\sigma^2}{n}$$

der vi antar at $\bar{Y} = \sum_{(ki)\in U} y_{ki} / N \approx \mu$ og $1/(N-1) \approx 1/N$, og V_D betegner utvalgsvariansen. Det siste uttrykket er lik modellvariansen til $\hat{Y}_{P-P}$, betinget på at det ikke finnes flere elementer fra samme klynge i utvalget, betegnet med $\lambda_{P-P} = 1$, dvs.

$$V_D(\hat{Y}_{P-P}) \approx N^2 \frac{\sigma^2}{n} = V_M(\hat{Y}_{P-P} \mid \lambda_{P-P} = 1) = \left(\frac{N}{n}\right)^2 \sum_{(ki)\in s} \sigma^2$$

2.4.3 Utvalgsvarians C-P ved enhetskalibrering

I tilfellet C-P med enkel tilfeldig trekking av klynger har vi

$$\hat{Y}_{C-P} = \sum_{(ki)\in s} \frac{N}{n} y_{ki} = \frac{N}{n} \sum_{k\in s} y_k$$

som ser likt ut som $\hat{Y}_{P-P}$, og er kjent som "ratio-to-size" estimatoren. Forskjellen fra P-P er at n er stokastisk. Legg merke til at vekten er kalibrert, og ikke gitt ved invers av trekkesannsynligheten. La $\bar{Y}$ være element-gjennomsnitt som før. Anta $\bar{Y} \approx \mu$. Uten epf er utvalgsvariansen gitt ved

$$\begin{aligned}
V_D(\hat{Y}_{C-P}) &\approx \frac{M^2}{m} \frac{\sum_{k\in U} (y_k - N_k\bar{Y})^2}{M-1} \\
&\approx \frac{M}{m} \sum_{k\in U} (y_k - N_k\mu)^2 = \frac{M}{m} \sum_{k\in U} \varepsilon_k^2 = \frac{M}{m} \sum_{k\in U} \left(\sum_i \varepsilon_{ki}^2 + \sum_{i\neq j} \varepsilon_{ki}\varepsilon_{kj} \right) \\
&\approx \frac{M}{m} \sum_{k\in U} \left(N_k\sigma^2 + N_k(N_k - 1)\rho\sigma^2 \right) \approx \frac{M\cdot N}{m} \frac{\sum_{k\in s} \left(N_k\sigma^2 + N_k(N_k-1)\rho\sigma^2 \right)}{n} \\
&= \frac{M\cdot N}{m\cdot n} \sum_{k\in s} \left(\sum_i V_M(\varepsilon_{ki}) + \sum_{i\neq j} \mathrm{cov}_M(\varepsilon_{ki,}\varepsilon_{kj}) \right) = \frac{M\cdot n}{m\cdot N} \left(\frac{N}{n}\right)^2 \sum_{k\in s} V_M(\varepsilon_k \mid N_k) \\
&= V_M(\hat{Y}_{C-P} \mid N_k, k\in s) = \frac{M\cdot n}{m\cdot N} V_M(\hat{Y}_{P-P} \mid \lambda_{P-P} = 1) + \frac{M\cdot N}{m\cdot n} \left(\frac{N}{n}\right)^2 \sum_k N_k(N_k - 1)\rho\sigma^2
\end{aligned}$$

siden

$$\{\sum_{k\in U} N_k + \sum_{k\in U} N_k(N_k - 1)\rho\} / \{\sum_{k\in s} N_k + \sum_{k\in s} N_k(N_k - 1)\rho\} \approx N/n$$

for stort utvalg i tilfellet enkel tilfeldig trekking av klynger. Hvis vi antar at $M/m \approx N/n$ kan variansen i tilfellet C-P skrives som variansen i tilfellet P-P, pluss et ekstra ledd som skyldes korrelasjon innen klynge og klyngestørrelse:

$$V_D(\hat{Y}_{C-P}) \approx V_M(\hat{Y}_{P-P} \mid \lambda_{P-P} = 1) + \left(\frac{N}{n}\right)^2 \sum_k N_k(N_k - 1)\rho\sigma^2$$

Det ekstra leddet er lite hvis ρ er liten, og det forsvinner hvis $\rho = 0$. Også for klynger bestående av et element vil leddet være eksakt 0.

Gabler *et al.* (1999) brukte den samme modellen for å studere Kishs "designeffektformel" i tilfellet klyngeutvalg og vekteklasser. I vårt tilfelle finnes det kun en vekteklasse. Deres resultat blir da

$$\frac{V_M(\hat{Y}_{C-P} \mid N_k, k \in s)}{V_M(\hat{Y}_{P-P} \mid \lambda_{P-P} = 1)} = 1 + \frac{\sum_{k \in s} N_k(N_k - 1)\rho\sigma^2}{n\sigma^2} = 1 + \rho\frac{\sum_k N_k^2 - n}{n}$$
$$= 1 + \rho(\bar{N}^* - 1)$$

der $\bar{N}^* = \sum_{k \in s} N_k^2 / n$ er en slags gjennomsnittlig klyngestørrelse i utvalget. Resultatet er noe lettere å bruke enn variansdekomponering ovenfor. I tilfellet AKU har vi $\bar{N} \approx 1.6$ og $\rho \approx 0.2$ for sysselsetting (Rafat, 2002). Vi har tilnærmet

$$1 + \rho(\bar{N} - 1) \approx 1 + 0.2 \cdot 0.6 = 1.12$$

Mens i Tabell 1 finner vi

$$\frac{\hat{V}_D(\hat{Y}_{C-P})}{\hat{V}_D(\hat{Y}_{P-P})} = \left(\frac{10859}{10341}\right)^2 = 1.05^2 = 1.10$$

Imidlertid er det formelt ikke riktig å betrakte $V_D(\hat{Y}_{C-P}) / V_D(\hat{Y}_{P-P})$ som designeffekt i tilfellet AKU, siden $\hat{Y}_{C-P}$ er kalibrert, og dermed inneholder også en estimeringseffekt. En bedre definisjon på designeffekt er variansforhold mellom to Horvitz-Thompson estimatorer, som alene avhenger av design.

2.4.4 Utvalgsvarians C-C

I tilfellet C-C med enkel tilfeldig trekking av klynger har vi

$$\hat{Y}_{C-C} = \sum_{(ki) \in s} \frac{M}{m} y_{ki} = \frac{M}{m} \sum_{k \in s} y_k$$

Uten epf er utvalgsvariansen gitt ved

$$V_D(\hat{Y}_{C-C}) \approx \frac{M^2}{m} \frac{\sum_{k \in U} (y_k - \sum_{k \in U} y_k / M)^2}{M - 1}$$

Anta $\sum_{k \in U} y_k / M \approx E_M(y_k) = \mu E_M(N_k)$ og $M/m \approx N/n$. Vi har da

$$V_D(\hat{Y}_{C-C}) \approx V_M(\hat{Y}_{C-C}) \approx \left(\frac{N}{n}\right)^2 \sum_{k\in s} V_M(y_k) = \left(\frac{N}{n}\right)^2 \{\sum_{k\in s} V_M E_M(y_k \mid N_k) + \sum_{k\in s} E_M V_M(\varepsilon_k \mid N_k)\}$$

$$= \left(\frac{N}{n}\right)^2 \{\sum_{k\in s} \mu^2 V_M(N_k) + E_M(\sum_{k\in s} V_M(\varepsilon_k \mid N_k))\} \approx \left(\frac{N}{n}\right)^2 \mu^2 \sum_{k\in s} V_M(N_k) + V_M(\hat{Y}_{C-P} \mid N_k, k \in s)$$

Variansen til C-C er omtrent lik variansen til C-P pluss et ekstra ledd, som skyldes variasjon i klyngestørrelsen og gjennomsnitt per element. Tabell 1 viser at leddet er betydelig for sysselsetting, og nær 0 for arbeidsledighet. Siden variasjonen i klyngestørrelse er identisk uansett interessevariabel, må forskjellen ligge i μ^2. For arbeidsledighet er μ svært lav (ca. 0.03), og dermed bidrar leddet nesten ingenting i tillegg. Når det gjelder sysselsetting er μ mye høyere (ca. 0.7), og leddet er faktisk større enn variansen i tilfellet C-P, og utgjør mer enn halvparten av variansen i tilfellet C-C.

Variansforholdet mellom C-C og P-P er designeffekt i tilfellet AKU. Vi har vist at variansøkningen har to komponenter: den ene skyldes korrelasjon i interessevariabelen innen samme klynge og klyngestørrelser, og den andre skyldes variasjon i klyngestørrelsen og gjennomsnittet i interessevariabel per element. Den første komponenten finnes under scenario C-P, men ikke den andre. I tilfellet AKU er den andre komponenten betydelig for sysselsetting, men ikke den første. Slik forholder variansestimatene seg imellom i Tabell 1 under enhetskalibrering.

2.5 Andre effekter

For det første ser vi på effekten av tilleggsvariabler. Modellen er nå $y_{ki} = x_{ki}^T\beta + \varepsilon_{ki}$, med samme kovariansstruktur for ε_{ki} som før (kapittel 2.4.1). I formelen for utvalgsvarians erstattes y_{ki} med ε_{ki}, og y_k med ε_k. Modellbasert tilnærming av utvalgsvarians er den samme som før under P-P og C-P. I tilfellet C-C har vi $E_M(\varepsilon_k \mid N_k) = 0 = E_M(\varepsilon_k) \approx \sum_{k\in U} \varepsilon_k / M$, slik at det ekstra leddet i forhold til C-P forsvinner. Når det gjelder variansforskjellen mellom C-P og P-P, er det rimelig å tro at den er liten siden ρ må være veldig liten betinget på tilleggsvariabler i AKU. Med andre ord, variansøkningen, som skyldes at familie er utvalgsenhet, kan like godt tas bort ved bruken av gode tilleggsopplysinger om personer. Igjen er egentlig person enheten i estimering siden modellen er satt opp på personnivå.

For de andre ser vi på effekten av stratifisering. Dagens AKU-utvalgsplan bruker fylker som strata. Zhang og Vedø (2004) forslo et stratifisert familie utvalg spesielt mht. familiestørrelsen i tillegg. Designeffekten er omtrent lik 1 for både sysselsetting og arbeidsledighet. Men et personutvalg stratifisert etter for eksempel alder og kjønn i tillegg til fylke har enda mindre utvalgsvarians for sysselsetting. Effisiensen kan likevel gjenvinnes i estimeringen, siden alder og kjønn brukes som tilleggsvariabler der. Heller ikke er stratifiseringen av familier avgjørende for de endelige AKU estimatene. Det viktigste tiltaket i denne sammenhengen er å gå over til enkel tilfeldig trekking enn å fortsette med systematisk trekking, slik Zhang og Vedø (2004) har argumentert for.

Betrakt til slutt familieutvalg trukket via personer: først trekkes personer tilfeldig, deretter inkluderes alle personer fra samme familie/husholdning i utvalget. Dette svarer omtrent til å trekke familier med en sannsynlighet som er proporsjonell med (familie-) størrelsen (pps). Uten epf kan vi tilnærme utvalgsvariansen til Horvitz-Thompson estimatoren ved å anta at utvalget er trukket med tilbakelegging. Forholdet til en varians under modellen i avsnitt 2.4.1 følger deretter (Cochran, 1977, avsnitt 9A.7). Spesielt er pps-trekking av klynger omtrent like effisient som enkel tilfeldig trekking dersom $\rho \approx 0$, som er tilfellet i AKU når tilleggsvariabler brukes i estimering. Det synes derfor lite å velge mellom pps og enkelt tilfeldig trekking hva effisiensen angår.

3. Varians i AKU estimering

3.1 Linearisering

I AKU estimeringen er personer estimeringsenheten, og vektene justeres tre ganger som følger:

- Inngangsvektene $a_i = M_h / m_h$ er like for alle personer innen et fylke, der M_h er antallet familier i fylke h, og m_h er antallet familier i nettoutvalget i samme fylke.
- W1 er en vekt som innen hvert fylke summerer seg til populasjonsantallet, og er definert ved $W1_{hi} = a_i \cdot b_h = N_h / n_h$, der $b_h = N_h / \tilde{N}_{hr}$, $\tilde{N}_{hr} = \sum_{(hi)\in s_h} a_i r_{hi}$, N_h er antall personer i fylket, n_h er antall personer i nettoutvalget, og $r_{hi} = 1$ for svar og 0 for frafall. Med tilde menes en størrelse beregnet på grunnlag av inngangsvekter.
- W2 er en vekt som innen hvert etterstratum summerer seg til populasjonsantallet, og er definert ved $W2_{hi} = W1_{hi} c_p$. Her er $c_p = N_p / (\sum_h \frac{N_h}{\tilde{N}_{hr}} \tilde{N}_{phr})$, der N_p er antall personer i etterstratum p, $\tilde{N}_{phr} = \sum_{(hi)\in s_{ph}} a_i r_{hi}$, og s_{ph} er alle personer fra stratum p og fylke h i utvalget.
- W3 tar høyde for kalibrering innen hvert fylke, med hensyn til kjønn, alder og en todelt register sysselsettingsstatus, og er definert ved $W3_{hi} = W2_{hi} g_{hi}$. Her er *g* en justering basert på kalibrering, beregnet på samme måte som i kapittel 2, men innen hvert fylke med en todeling av register sysselsettingsstatus. Det vil si at det konstrueres en designmatrise Xdesign basert på nettoutvalget og kalibreringsvariablene, og for hver person i hvert fylke beregnes $g_{hi} = 1 + (X_h - \hat{X}_h)^T \hat{A}_h^{-1} x_{hi}$. X_h er en vektor med marginale tall, $\hat{X}_h = \sum_{(hi)\in s} W2_{hi} x_{hi} = \sum_p b_h c_p \tilde{X}_{phr}$, der $\tilde{X}_{phr} = \sum_{(hi)\in s_{ph}} a_i r_{hi}$ (marginale tall estimert ved nettoutvalget og inngangsvektene) og $\hat{A}_h = \sum_{(hi)\in s} W2_{hi} (x_{hi} x_{hi}^T) = \sum_p b_h c_p \tilde{A}_{phr}$, der $\tilde{A}_{phr} = (\text{Xdesign}^T \otimes a) \times \text{Xdesign}$, og x_{hi}^T er raden i Xdesign som svarer til person i.

For et gitt fylke h er $\hat{Y}_h = \sum_{i\in s_h} W3_{hi} y_{hi}$ en estimator for Y_h, og $\hat{Y} = \sum_h \hat{Y}_h$.

Anta enkelt tilfeldig trekking av familieutvalget innen hvert fylke, i stedet for systematisk trekking. For enkelhets skyld ser vi nå bort fra klynger og antar at vi har et P-P scenario, siden det er liten forskjell mellom P-P og C-P, som vi har sett. Følgende empiriske variansestimator gir opphav til et variansestimat:

$$\hat{V}_h = \frac{n_h}{n_h - 1} \sum_{(hi)\in s_h} \left(W3_{hi} e_{hi} - \frac{\sum W3_{hi} e_{hi}}{n_h} \right)^2 \approx \sum_{(hi)\in s_h} (W3_{hi} e_{hi})^2 \qquad (3)$$

der e_i er residualet ved fylkesvis kalibrering. Det siste uttrykket følger av at summen av vektede residualer er lik 0. Det totale variansestimatet finner vi ved å summere over fylker; $\hat{V} = \sum_h \hat{V}_h$. Bak (3) ligger det en del asymptotiske antagelser (Zhang, 2006), der $\phi_h = N_h / N_{hr}$, $\kappa_p = N_p / (\sum_h \frac{N_h}{N_{hr}} N_{phr})$ og $\delta_{hi} = 1 + (X_h - \sum_p \phi_h \kappa_p X_{phr})^T (\sum_p \phi_h \kappa_p A_{phr})^{-1} x_{hi}$ tilnærmes med henholdsvis b_h, c_p og g_{hi}. Tabell 2 og 3 viser en oversikt over variansestimater, både for totalen og for hvert enkelt fylke.

Variansen i (3) kan dekomponeres i to ved å betrakte frafall som en tilleggsfase i trekkingen, gitt bruttoutvalget. Estimeringsopplegget i AKU impliserer *formelt* en frafallsmodell der frafallet er tilfeldig

innen fylker. La ω_h betegne $\sum_p \sum_{i \in s_{phr}} a_{hi}\phi_h \kappa_p \delta_{hi}\varepsilon_{hi}$. Variansen til ω_h vil være lik den asymptotiske variansen til $\hat{Y}$ (Zhang, 2006), og kan dekomponeres som

$$V(\omega_h) = V_s(E_r(\omega_h \mid s)) + E_s(V_r(\omega_h \mid s))$$

Den første delen består av varians forårsaket av utvalgstrekkingen. Den andre delen som skyldes frafall tilnærmes ved $V_r(\omega_h \mid s)$, som kan estimeres ved

$$\hat{V}_{h,Frafall} = \frac{b_h - 1}{b_h} \sum_i (W3_{hi} e_{hi})^2 \qquad (4)$$

Vi ser at forholdet mellom (4) og (3) blir en konstant, $(b_h - 1)/b_h$, som er tilnærmet lik den fylkesvise frafallsandelen i populasjonen, som igjen er tilnærmet lik den fylkesvise frafallsandelen i utvalget. Det følger at variansen som forårsakes av frafall er like stor for arbeidsledighet som for sysselsetting. Resultatene (Tabell 2 og 3) er noe urealistiske og skyldes den forenklede frafallsmodellen. Det er grunn til å tro at frafallsjustering ved estimeringsopplegget i AKU er mer informativt enn som så, siden estimatene for hele landet endres mer fra W1 til W2 enn fra inngangsvektene til W1. Problemet er at det formelt ikke lar seg formulere annerledes enn den fylkesvise homogene frafallsmodellen, så lenge W1 er i bildet.

3.2 Bootstrap

Bootstrap variansestimater er et alternativ til variansestimatene regnet ut ved (3). Under et P-P scenario trakk vi med tilbakelegging B nye datasett ved å for hvert datasett b trekke like mange observasjoner fra AKU datasettet som det opprinnelig er. AKU kalibreringen beskrevet i avsnitt 3.1 ble gjentatt for hvert datasett, og $\hat{Y}$ ble beregnet. Slik fikk vi like mange estimater for totalt antall arbeidsledige og sysselsatte som nye datasett vi trakk, det vil si $\hat{Y}^{*1}, \hat{Y}^{*2}, ..., \hat{Y}^{*B}$, der $\hat{Y}^{*b}$ betegner bootstrap-estimat b av $\hat{Y}$ og b = 1, ..., B. Den empiriske variansen til $\hat{Y}^*$ er et bootstrap estimat for variansen av $\hat{Y}$, hvis kvadratrot betegnes med σ^*. Vi benyttet to ulike metoder:

- Metode 1: Størrelsen til nettoutvalget og frafallsandelen antas fast for hvert fylke. Det vil si at for hver replikasjon inneholder både nettoutvalget og frafallsandelen like mange personer som i AKUs første kvartal 2005, for hvert enkelt fylke. Dermed vil a og b_h være lik som i avsnitt 3.1 for hver replikasjon, mens $\hat{X}$ og Xdesign vil endre seg for hvert bootstrap datasett.
- Metode 2: Nettoutvalget og frafallsandelen for fylkene varierer for hver replikasjon, mens bruttoutvalget for hvert fylke er fast. Det vil si at bruttoutvalget inneholder like mange personer som i AKUs første kvartal 2005, men ikke nettoutvalget. Dermed vil både b_h, $\hat{X}$ og Xdesign variere mellom bootstrap datasettene.

Tabell 2 og 3 viser bootstrap estimater σ^* basert på B = 5 000 replikasjoner. Tallet i parentes er standardavviket til σ^*, som er beregnet på følgende måte: Trekk tilfeldig og med tilbakelegging 5000 ganger blant $\hat{Y}^{*1}, \hat{Y}^{*2}, ..., \hat{Y}^{*5000}$, betegnet med $\hat{Y}_{(1)}{}^{*1}, \hat{Y}_{(1)}{}^{*2}, ..., \hat{Y}_{(1)}{}^{*5000}$. Den empiriske variansen blant $\hat{Y}_{(1)}{}^{*1}, \hat{Y}_{(1)}{}^{*2}, ..., \hat{Y}_{(1)}{}^{*5000}$ gir oss et nytt bootstrap estimat for variansen til $\hat{Y}$, og dermed dens standardavvik, betegnet med $\sigma^*_{(1)}$. Gjenta dette 5000 ganger, dvs. opptil $\hat{Y}_{(5000)}{}^{*1}, \hat{Y}_{(5000)}{}^{*2}, ..., \hat{Y}_{(5000)}{}^{*5000}$, slik at vi nå har $\sigma^*_{(1)}, ..., \sigma^*_{(5000)}$. Den empiriske variansen blant $\sigma^*_{(1)}, ..., \sigma^*_{(5000)}$ er et estimat for variansen til σ^*, og dens kvadratrot er et estimat for standardavviket til σ^*.

Fra Tabell 2 og 3 kan vi trekke følgende konklusjoner:

- Ved å la nettoutvalget og frafallsandelen variere mellom replikasjoner, forventes det at variansen for estimatet av antall arbeidsledige eller sysselsatte skal øke. Dermed skulle vi forvente at variansen basert på Metode 1 ville ligge under Metode 2. Imidlertid er det liten forskjell å se, spesielt for sysselsetting, men også for arbeidsledighet. Variasjon i nettoutvalgsstørrelsen har nesten ingen effekt på variansen. Det er uklart hvorvidt konklusjonen skyldes at frafallsmodellen er forenklet og urealistisk, som nevnt før.
- Det er liten forskjell mellom metoden basert på linearisering og bootstrap metodene, både for sysselsetting og arbeidsledighet. Bootstrap metoden er mer krevende, siden vi trenger mange replikasjoner for å få gode estimat. Det største problemet med bootstrap ligger likevel i tilrettelegging av data. Hvis vi ønsker å kjøre bootstrap med familier som enhet for å gjenskape klyngeutvalgsplanen, må vi identifisere familiemedlem blant frafallet i tillegg til nettoutvalget. I praksis krever dette mye arbeid, siden nettoutvalget, frafallet og DSF med familienummer ligger på forskjellige filer og steder. I mange andre land er koblingen som skal til rett og slett umulig. Derfor er det godt å vite at lineariseringsmetoden i Kapittel 3.1. nesten gir det samme resultatet. Å kjøre sistnevnte krever ikke mer enn aggregering over personer innen samme familie i nettoutvalget, som i Kapittel 2.2. Metoden er klart lettere å implementere, og bør kunne brukes blant annet for kvalitetsrapportering av AKU.

Tabell 2: Resultater for sysselsetting ved enkelt tilfeldig utvalg der personer er trekkenhet. 1 = Standardavvik (Std) og relativt standardavvik (RSE) ved lineariseringsmetoden. 2 = standardavvik og relativt standardavvik for bootstrap metode 1 (fast nettoutvalg og frafallsandel). 3 = standardavvik og relativt standardavvik for bootstrap metode 2 (varierende nettoutvalg og frafallsandel). 4 = andel av total varians som skyldes frafall.

	1. Std	1. RSE	2. Std	2. RSE	3. Std	3. RSE	4. Frafallsandel
Østfold	1 692	1,35	1 704 (17)	1,36	1 683 (17)	1,34	10,98
Akershus	2 287	0,92	2 313 (22)	0,93	2 295 (23)	0,92	11,29
Oslo	2 933	1,03	2 928 (31)	1,03	2 938 (29)	1,03	20,41
Hedmark	1 326	1,56	1 300 (13)	1,53	1 333 (13)	1,57	9,17
Oppland	1 371	1,54	1 368 (14)	1,54	1 380 (14)	1,55	11,18
Buskerud	1 551	1,26	1 539 (16)	1,25	1 550 (15)	1,26	10,88
Vestfold	1 557	1,48	1 561 (16)	1,49	1 546 (16)	1,47	12,85
Telemark	1 304	1,67	1 310 (13)	1,67	1 323 (14)	1,69	11,18
Aust-Agder	865	1,82	872 (9)	1,83	880 (9)	1,85	11,42
Vest-Agder	1 341	1,77	1 371 (14)	1,81	1 351 (14)	1,78	9,69
Rogaland	1 925	0,98	1 927 (19)	0,99	1 948 (20)	1,00	7,04
Hordaland	2 364	1,05	2 344 (23)	1,04	2 377 (23)	1,05	10,75
Møre og Romsdal	875	1,75	884 (9)	1,77	891 (9)	1,78	8,42
Sogn og Fjordane	1 748	1,47	1 775 (17)	1,49	1 768 (18)	1,48	12,10
Sør-Trøndelag	1 653	1,20	1 629 (16)	1,18	1 645 (16)	1,19	9,55
Nord-Trøndelag	983	1,76	1 016 (10)	1,82	980 (10)	1,76	7,21
Nordland	1 473	1,34	1 483 (14)	1,34	1 459 (14)	1,32	11,91
Troms	1 397	1,86	1 406 (14)	1,87	1 416 (14)	1,88	14,37
Finmark	703	2,06	711 (7)	2,08	720 (7)	2,11	13,59
Totalt	7 128	0,31	7 040 (69)	0,31	7 036 (70)	0,31	10,98

Tabell 3: Resultater for arbeidsledighet ved enkelt tilfeldig utvalg der personer er trekkenhet. 1 = Standardavvik (std) og relativt standardavvik (RSE) ved lineariseringsmetoden. 2 = standardavvik og relativt standardavvik for bootstrap metode 1 (fast nettoutvalg og frafallsandel). 3 = standardavvik og relativt standardavvik for bootstrap metode 2 (varierende nettoutvalg og frafallsandel). 4 = andel som skyldes frafall.

	1. Std	1. RSE	2. Std	2. RSE	3. Std	3. RSE	4. Frafallsandel
Østfold	989	15,85	998 (10)	16,00	989 (10)	15,85	10,98
Akershus	1312	12,89	1 308 (13)	12,88	1 299 (13)	12,80	11,29
Oslo	2041	10,00	2 021 (21)	9,89	2 038 (20)	9,96	20,41
Hedmark	667	21,11	676 (7)	21,34	674 (7)	21,22	9,17
Oppland	794	20,06	794 (8)	20,04	803 (8)	20,33	11,18
Buskerud	825	19,52	832 (9)	19,69	817 (8)	19,36	10,88
Vestfold	875	18,29	880 (8)	18,40	878 (9)	18,50	12,85
Telemark	858	16,54	869 (9)	16,80	858 (9)	16,56	11,18
Aust-Agder	528	19,70	540 (6)	20,11	530 (5)	19,92	11,42
Vest-Agder	819	18,62	839 (8)	19,10	810 (8)	18,40	9,69
Rogaland	1149	12,35	1 162 (12)	12,47	1 134 (11)	12,21	7,04
Hordaland	1413	12,76	1 407 (14)	12,71	1 415 (14)	12,80	10,75
Møre og Romsdal	469	21,20	464 (5)	20,93	483 (5)	21,76	8,42
Sogn og Fjordane	928	17,47	938 (9)	17,68	936 (9)	17,62	12,10
Sør-Trøndelag	717	20,73	707 (7)	20,34	723 (7)	20,94	9,55
Nord-Trøndelag	473	25,04	476 (5)	25,23	470 (5)	25,00	7,21
Nordland	1046	13,67	1 034 (11)	13,50	1 061 (11)	13,88	11,91
Troms	651	26,91	660 (7)	27,28	647 (6)	26,87	14,37
Finmark	484	16,83	495 (5)	17,09	494 (5)	17,10	13,59
Totalt	4 236	3,80	4 273 (42)	3,83	4 179 (41)	3,75	10,98

Appendiks: Om data

Datagrunnlaget for analysene bygger på netto og brutto AKU-tall for første kvartal i år 2005, noe som utgjør 24 377 personer. AKU inneholder en variabel som beskriver intervjuobjektets sysselsettingsstatus i A/A-registeret. Dersom stmnace = 1, 2, eller 3 er personen sysselsatt ifølge registeret, ellers er personen ikke sysselsatt. I nettoutvalget er det 21 565 personer, men 40 av disse mangler opplysning om stmnace, og inngår dermed i frafallsgruppen.

Stmnace deles i AKU produksjonen i to grupperinger:

- 4-deling: Sysselsatt i primær-, sekundær-, eller tertiærnæringen, og ikke-sysselsatt
- 2-deling: Sysselsatt og ikke-sysselsatt

Det finnes 36 personer i AKU 1. kvartal 2005 som er definert som både arbeidsledig i henhold til Arena og sysselsatt i henhold til A/A-registeret. Disse defineres som ikke-sysselsatte. Det er 328 personer som mangler familienummer, og som får tildelt et unikt familienummer hver. Det er verdt å merke seg at 90% av disse faller i kategorien ikke-sysselsatt i henhold til registeret. En person mangler informasjon om fylke, og tre tilhører fylke 21, som tilsvarer Svalbard. For å unngå å fjerne noen intervjuobjekter slik at summen av vektene ikke lenger stemmer med AKU produksjonen, allokeres disse til fylke 03 (Oslo) som er det fylket med flest tilhørende personer.

Alder deles i 12 klasser; 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69 og 70-74 år.

Kjønn, alder og en 4-deling av registersysselsetting gir grunnlag for 2 x 12 x 4 = 96 etterstrata. For å unngå tomme celler, brukes det i AKU produksjonen en 2-deling av registersysselsetting for den eldste aldersgruppen (Heldal, 2000).

Hver person har en oppblåsingsvekt (mndnetto) som er beregnet for AKU månedstall. For AKU kvartalstall skal vekten deles med 3. Et datasett ble konstruert:

- Kjønn (kjonn)
- Alder (basert på variabelen alder)
- Registerstatus for sysselsetting (basert på variabelen stmnace) - 2 grupperinger
- AKU sysselsettingsstatus (basert på variabelen sstat)
- Vekt (mndnetto delt med 3)
- Fylke (fylke)
- Familienummer (IOs_fam)

Populasjonstotaler for grupper defineres til summen av vekt for gruppen. Datasettet danner grunnlaget for påfølgende analyser.

Referanser

Cochran, W.G. (1977): Sampling Techniques (3rd ed.). Wiley.

Gabler, S., Haeder, S. og Lahiri, P. (1999): A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, **25**, 105-106.

Heldal, J. (2000): Kalibrering av AKU. Dokumentasjon av metode og program. Notat 2000/7.

Kish, L. (1987): Weighting in Deft^2 . *The Survey Statistician*, June 1987.

Rafat, D. (2002): Analyse av sammenheng mellom ektefellers sysselsetting i en familie. Notat 2002/35.

Särndal, C-E., Swensson, B. og Wretman, J. (1992): Model Assisted Survey Sampling. Springer Verlag.

Shao, J. (1996): Resampling methods in sample surveys (with discussion). *Statistics,* **27**, 203-254.

Vedø, A. og Rafat, D. (2003): Sammenligning av utvalgsplaner i AKU. Notat 2003/56.

Zhang, L-C. og Vedø, A. (2004): Omlegging av utvalgsplan for AKU. Notat 2004/86.

Zhang, L-C. (2006): A simplistic approach to variance estimation for calibrated estimators subjected to weighting adjustments. Upublisert notat.