

Om justering för undertäckning vid undersökningar med urval i "rum och tid"

Bengt Rosén



R&D Report
Statistics Sweden
Research - Methods - Development
1990:3

Från trycket	Mars 1990
Producent	Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare	Åke Lönnqvist
Förfrågningar	Bengt Rosén, tel. 08-783 44 90

© 1990, Statistiska centralbyrån
ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1990

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1990:3. Om justering för undertäckning vid undersökningar med urval i "rum och tid" / Bengt Rosén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

Om justering för undertäckning vid undersökningar med urval i "rum och tid"

Bengt Rosén

On Adjustment for Non-coverage in Surveys with Samples in "Space and Time".

Abstract. We consider surveys with sampling in "space and time", a typical example being a transportation survey where lorries are selected randomly from a list and where measurement periods are allotted at random to the selected lorries.

If the population under consideration is changing rapidly (in the example; by "births" of new lorries and by "deaths" of existing ones), the "traditional" estimators for aggregated values may be biased, and the "births" lead to the biggest complications. We discuss methods to adjust for non-coverage bias in this type of context. The success of the adjustment procedures depend of course, as any adjustment procedure for non-coverage inevitably does, on the degree of fulfillment of certain assumptions.

To illustrate the general results as well as for its intrinsic interest we apply the results to Statistics Sweden's lorry transportation survey.

FÖRORD.

I rapporten presenteras en uppsättning förhållandevis allmänna resultat angående justering för skevhet på grund av undertäckning i urvalsramen. Resultaten har tagits fram i anslutning till planeringen av den UVAV-undersökning som genomförs under 1990.

Det skulle förvåna mig om de allmänna resultaten inte, åtminstone delvis, står att finna i litteraturen men jag har inte stött på dem (men heller inte gjort stora sökansträngningar). I vilket fall som helst tror jag att resultaten är värda att bevara i en SCB-rapport, eftersom problemställningen bör finnas på flera kanter inom SCB.

I mars 1990

Bengt Rosén

SCB, U/PLAN
Bengt Rosén

1990-03-06

OM JUSTERING FÖR UNDERTÄCKNING VID UNDERSÖKNINGAR MED URVAL I "RUM OCH TID".

1. Inledning och "läsanvisningar".

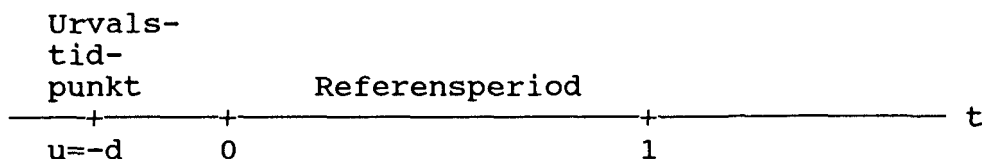
I det följande diskuteras hur man för undersökningar med "urval i tid och rum" ibland kan justera för skattningsskevheter beroende på undertäckning i urvalsramen.

Resultaten kommer att formuleras i termer av väntevärdesegenskaper för olika estimatorer. Härledningarna kommer dock att vara långt ifrån rigorösa, framför allt av det skälet att framställningen skulle bli synnerligen omständlig om man verkligen ginge in och gjorde stringenta feluppskattningar i alla de approximationer som görs. Istället tillåter vi oss att saxa oss fram dels med resonemang som enligt "de stora talens lag" ger formler som vore korrekta om det handlade om mycket stora urval från mycket stora populationer (men urvalsfraktionen får vara liten) dels med hjälp av följande "princip"; Resultat som enligt stora talens lag gäller med små relativa fel, ger för situationer med "måttligt" stora populations- och urvalsstorlekar approximativt korrekta formler för förväntade värden.

Vidare håller vi oss genomgående till "första ordningens approximationer", varmed avses att vi använder "räkneregler" av följande typ. När a och b är "små" jämfört med 1 har vi t ex $(1+a) \cdot (1+b) \approx 1+a+b$ och $(1+a)/(1+b) \approx 1+a-b$. Termer som försummas vid användande av den slags approximationer åberopar vi som "andra ordningens termer".

2. Situation av intresse.

Intresset gäller den produktionstotal, med avseende på en föreskriven (produktions)variabel x , som genereras av ett specificerat kollektiv av objekt under en föreskriven referensperiod. Tiden betraktas som kontinuerlig, och för enkelhets skull väljer vi tidsskala och tidsorigo så att referensperiodens längd är tidsenhet och referensperioden börjar vid tidpunkt 0. Vidare anges en tidpunkt $u=-d$ ($d \geq 0$) som urvalstidpunkt. (Det kommer att framgå varför vi väljer den benämningen.) Situationen illustreras i nedanstående figur.



Objekten av intresse kan såväl "dö" som "födas" (inklusive "återuppstå från de döda") och vi delar in dem i följande kategorier, där kategoriseringen görs relativt referensperioden och urvalstidpunkten;

- persistenter = objekt som lever vid $t=-d$ och som håller sig vid liv till $t=1$,
- (enkla) dödsobjekt = objekt som lever vid $t=-d$ men som dör före $t=1$ sedan förblir döda till $t=1$,
- (enkla) föddaobjekt = objekt som inte levde vid $t=-d$ men som föds före $t=1$ och sedan håller sig vid liv till $t=1$.
- transienter = objekt som hinner med att ändra tillstånd två eller fler gånger mellan $t=-d$ och $t=1$.

Låt

U = kollektivet av objekt som varit vid liv under någon del av referensperioden. Bokstaven i betecknar fortsättningsvis ett allmänt objekt i U . (2.1)

Q = antalet objekt i U , (2.2)

N = antalet objekt som är vid liv vid referensperiodens början, dvs vid $t=0$. (2.3)

Notera att Q och N i det allmänna fallet inte är lika, utan normalt är Q större än N .

Levande objekt producerar " x -mängd" som kan aggregeras över tiden, medan döda och ofödda objekt inte producerar någonting. För ett delintervall Ω av $[0,1]$ sätter vi

$\underline{x}(\Omega)_i$ = den x -mängd som objekt i producerar under tidsintervallet Ω , $i=1,2,\dots,Q$. (2.4)

Intresset gäller den (globala) x -total $\theta(\underline{x})$ som erhålls genom aggregering av x -produktionen över hela referensperioden och över alla objekt i U , dvs

$$\theta(\underline{x}) = \sum_{i \in U} x([0,1])_i. \quad (2.5)$$

Referensperioden $[0,1]$ delas upp i A stycken lika stora delintervall, vart och ett alltså av längden $1/A$. Dessa delintervall skall spela rollen av mätintervall. Vi inför det förkortade beteckningssättet

$$x(d)_i = x([(d-1)/A, d/A])_i, \quad i=1,2,\dots,Q, \quad d=1,2,\dots,A. \quad (2.6)$$

Storheten i (2.6) är objekt i:s $\underline{x}$ -produktion under det d:te mätintervallet. Vi kan då skriva totalen i (2.5) på följande sätt;

$$\Theta(\underline{x}) = \sum_{d=1}^A \sum_{i \in U} x(d)_i. \quad (2.7)$$

Vi gör följande antaganden om födelse och död i populationen.

Objekt föds och dör i "jämn och stadig takt" med födelseintensitet α och dödsintensitet β . (2.8)

Intensiteterna α och β antas båda vara små i förhållande till 1. (2.9)

Mer precist lägger vi in följande i (2.8), och vi försummar då tänkbara "andra ordningens effekter".

Antalet objekt som föds [dör] under ett tidsintervall Ω är (åtminstone approximativt) $N \cdot \alpha \cdot l(\Omega)$ [$N \cdot \beta \cdot l(\Omega)$] (där $l(\Omega)$ står för längden av intervallet Ω). (2.10)

Fortsättningsvis resonerar vi på ett överslagsmässigt sätt. Dels tar vi (2.9) som intäkt för att försumma andra ordningens α - och β -termer dels kommer vi att resonera på basis av bl a följande premisser;

Varje levande objekt har "genomsnittlig" produktivitet, och denna genomsnittliga produktivitet är konstant över referensperioden. (2.11)

Transienternas bidrag till totalen $\Theta(\underline{x})$ kan försummas. (2.12)

Låt

q = genomsnittlig $\underline{x}$ -produktion för ett objekt under ett mätintervall (givet att objektet lever hela mätintervallet). (2.13)

Med (2.8) och (2.10)-(2.13) som bakgrund skriver vi ner följande uttryck för $\underline{x}$ -totalen (och lämnar en del "ifyllande" åt läsaren),

$$\Theta(\underline{x}) \approx N \cdot A \cdot q \cdot [1 + \alpha/2 - \beta/2]. \quad (2.14)$$

3. Urval och observationer.

Vi antar att det vid urvalstidpunkten föreligger en urvalsram som omfattar (precis) de levande objekten vid $t=-d$, vilka säges utgöra rampopulationen. Från den ramen dras ett OSU-urval utan återläggning. Låt

N_u och n = antalet objekt i urvalsramen respektive i urvalet. (3.1)

Vidare;

De urvalda objekten fördelas (t ex genom lottning) jämnt på de A mätintervallen, varvid ett och endast ett mätintervall tilldelas varje urvalt objekt. (3.2)

Vi tillåter möjligheten att alla objekt i populationen tas med i urvalet (dvs att $n = N_u$), varvid ordet "urval" kanske är litet inadekvat. Dock, även i det fallet finns en slumpurvalsaspekt med i bilden, nämligen den som kommer in via den slumpmässiga tillottningen av mätintervall. Det är den sistnämnda aspekten vi har i tankarna när vi, som i rubriken, talar om "urval i tiden". Aspekten "urval i rummet" hänför sig till urvalet av objekt.

Avsikten är att de urvalda objektens x -produktion skall observeras under tillottat mätintervall. Dock, när ett objekt blir aktuellt för mätning kan det redan ha dött, och i så fall klassas det som ett övertäckningsobjekt, och övertäckningsobjekten är inte aktuella för observation (eftersom man vet att de inte producerar någonting). Är objektet vid liv vid mättillfället säges det vara ett intressant objekt, och det är för dessa objekt man vill göra mätning av x -produktionen under mätintervallet. Sätt

$$n_I = \text{antalet intressanta objekt i urvalet,} \quad (3.3)$$

$$n_O = \text{antalet övertäckningsobjekt i urvalet.} \quad (3.4)$$

I den praktiska situationen råkar man normalt också ut för svarsbortfall, och vi skall betrakta fall där svarsbortfall kan inträffa för såväl intressanta objekt som för övertäckningsobjekt. I det senare fallet medför svarsbortfall att ett objekt inte blir "identifierat" som övertäckningsobjekt. Sätt

$$n_S = \text{antalet svarande intressanta objekt,} \quad (3.5)$$

$$\tilde{n}_O = \text{antalet svarande övertäckningsobjekt.} \quad (3.6)$$

Vi ger härnäst några "genomsnittsformler" för de ovan införda n -värdena, och vi delar då upp i fallen med och utan svarsbortfall.

Inget svarsbortfall.

Om inget svarsbortfall förekommer är $n_S = n_I$ och under antagandet (2.8) om "jämn födelse- och dödskraft" gäller;

$$n_S = n_I \approx n \cdot [1 - \beta(d+1/2)] \quad \text{och} \quad n_O \approx n \cdot \beta \cdot (d+1/2). \quad (3.7)$$

Svarsbortfall förekommer.

Fortsättningsvis håller vi oss inom ramen för följande modell för bortfallsskeendet.

- Ett övertäckningsobjekts karaktär som övertäckning upptäcks endast genom "svar". (3.8)

- Slumpen avgör om ett urvalt objekt svarar eller ej. Härvid agerar objekten oberoende av varandra. Intressanta objekt har en och samma svarandebenägenhet, ϕ_I , och över-täckningsobjekt har likaså en och samma svarandebenägenhet, $\phi_{\bar{O}}$. (3.9)

Med genomsnittsresonemang leds vi till;

$$n_S \approx \phi_I \cdot n_I \quad \text{och} \quad \tilde{n}_{\bar{O}} \approx \phi_{\bar{O}} \cdot n_{\bar{O}}. \quad (3.10)$$

4. Estimation.

4.1 Inget svarsbortfall.

I första omgången antar vi att inget svarsbortfall förekommer.

Låt $W(\underline{x})$ beteckna den observerade urvalssumman för $\underline{x}$ -produktionen under mätintervallen, dvs

$$W(\underline{x}) = \sum_{i \in S} x(D_i) i, \quad (4.1)$$

där S står för urvalet och D_i för det mätintervall som lottats till (det urvalda) objektet i . Inför statistikan

$$T(\underline{x}) = [N_U/n] \cdot A \cdot W(\underline{x}). \quad (4.2)$$

Följande resultat betraktas som välkänt.

- Den med ovanstående faktor $[N_U/n] \cdot A$ "inflaterade" urvalssumman ger väntevärdesriktig skattning av den produktions-total som genereras av objekten i rampopulationen. (4.3)

Den produktionstotal som genereras av objekten i rampopulationen är $\theta(\underline{x})$ minus produktionsbidraget från de objekt som föds under tidsintervallet $[-d, 1] \approx$ (med (2.8) och (2.11)-(2.13) som bakgrund, och med försummande av andra ordningens termer)

$$\begin{aligned} &\approx \theta(\underline{x}) - q \cdot A \cdot N \cdot \alpha \cdot (d+1/2) \approx (\text{enligt (2.14)}) \approx \\ &\approx \theta(\underline{x}) \cdot [1 - \alpha \cdot (d+1/2)]. \end{aligned} \quad (4.4)$$

(4.3) och (4.4) leder till nedanstående formel (4.5) i vilken, och fortsättningsvis, E står för väntevärde.

$$E[T(\underline{x})] \approx \theta(\underline{x}) \cdot [1 - \alpha \cdot (d+1/2)]. \quad (4.5)$$

Det är $\theta(\underline{x})$ vi är intresserade av att skatta. Från (4.5) ser vi att $T(\underline{x})$ ger skev skattning, och att skevheten emanerar från de födda objekten, vilka ju inte finns med i urvalsramen. I fortsättningen skall vi diskutera några möjligheter att justera för skevheten ifråga. De justeringar som vi skall betrakta kommer att vara av typen; Modifikation av "stora N " och/eller "lilla n " i (4.2). Vi börjar med ett resultat av mer allmän karaktär, och vi inför storheterna

$$N(t) = N \cdot [1 + t(\alpha - \beta)], \quad (4.6)$$

$$n(\delta) = n_I + (1 - \delta) \cdot n_O. \quad (4.7)$$

Storheten $n(\delta)$ är enkelt relaterad till observationerna, och faktorn δ kallar vi undertryckningsgraden av övertäckningen. Storheten $N(t)$ kommer att konkretiseras snart. Vi inför estimatorn

$$T(\underline{x}; t, \delta) = [N(t)/n(\delta)] \cdot A \cdot W(\underline{x}). \quad (4.8)$$

Vad är väntevärdet för $T(\underline{x}; t, \delta)$? För att besvara den frågan noterar vi först nedanstående relation (där $T(\underline{x})$ är enligt (4.2)),

$$T(\underline{x}; t, \delta) = T(\underline{x}) \cdot [N(t)/N_U] \cdot [n/n(\delta)]. \quad (4.9)$$

Som konsekvens av (2.10) har vi, återigen med försummande av andra ordningens termer,

$$N_U \approx N \cdot [1 - d \cdot (\alpha - \beta)], \quad (4.10)$$

som tillsammans med (4.6) ger,

$$N(t)/N_U \approx 1 + (t + d) \cdot (\alpha - \beta). \quad (4.11)$$

Från (4.7) och (3.7) följer, till första ordningens approximation,

$$n/n(\delta) \approx 1 + \delta \cdot \beta \cdot (d + 1/2). \quad (4.12)$$

Genom att kombinera formlerna (4.9), (4.5), (4.11) och (4.12) och försumma andra ordningens termer, kommer man till följande approximationsresultat.

LEMMA 1: Under gjorda antaganden gäller;

$$E[T(\underline{x}; t, \delta)] \approx \theta(\underline{x}) \cdot [1 - \alpha \cdot (1/2 - t) + \beta \cdot \{\delta \cdot (d + 1/2) - t - d\}]. \quad (4.13)$$

Speciellt gäller följande; Som estimator för $\theta(\underline{x})$ har $T(\underline{x}; t, \delta)$ nedanstående (approximativa) skevhet;

Relativ underskattning $\approx$

$$\approx \alpha \cdot [1/2 - t] - \beta \cdot [\delta \cdot (d + 1/2) - t - d]. \quad (4.14)$$

Ovanstående Lemma 1 kommer att vara utgångspunkt för mer "praktisknära" betraktningar, där vi dels inkluderar möjligheten att storleken på den aktuella undersökningspopulationen räknas vid andra tidpunkter än bara vid urvalstillfället dels tillåter att svarsbortfall förekommer.

4.2 Flera räkningar av populationens storlek.

Vi börjar med att föra in litet terminologi och beteckningar.
Sätt

$$N_t = \text{populationens storlek (= antalet levande objekt i den) vid tidpunkt } t. \quad (4.15)$$

Vi antar att räkning av populationens storlek görs vid räknetidpunkterna $t=(t_1, t_2, \dots, t_k)$. Vi säger att N^* är den vägda populationsstorleken och t^* den vägda räknetidpunkten med vikterna $\pi=(\pi_1, \pi_2, \dots, \pi_k)$, när

$$N^* = \pi_1 \cdot N_{t_1} + \pi_2 \cdot N_{t_2} + \dots + \pi_k \cdot N_{t_k}, \quad (4.16)$$

$$t^* = \pi_1 \cdot t_1 + \pi_2 \cdot t_2 + \dots + \pi_k \cdot t_k. \quad (4.17)$$

Vikterna $\pi=(\pi_1, \pi_2, \dots, \pi_k)$ förutsätts satisfiera

$$\pi_1 + \pi_2 + \dots + \pi_k = 1, \quad (4.18)$$

men vi ställer inte krav om att vikterna nödvändigtvis skall vara icke-negativa.

Kommentar 4.1: Ett i praktiken rätt vanligt förekommande fall är att den aktuella undersökningen genomförs periodiskt. Efter en referensperiod kommer en ny referensperiod av samma längd med ny undersökning, och därefter ännu en referensperiod osv. Vidare förutsätter vi att urvalstidpunkten varje gång läggs lika långt före referensperioden. Urvalstidpunkterna för de successiva referensperioderna 1, 2, 3, ... är då $t = -d, 1-d, 2-d, \dots$. Denna typ av situation åberopar vi som periodisk undersökning. ☒

Vi inför estimatorn

$$T(\underline{x}; \underline{t}, \pi, \delta) = [N^*/n(\delta)] \cdot A \cdot W(\underline{x}). \quad (4.19)$$

Med (2.8) som bakgrund har vi,

$$N_t \approx N \cdot [1 + t \cdot (\alpha - \beta)], \quad (4.20)$$

som i sin tur leder till relationen

$$N^* \approx N \cdot [1 + t^* \cdot (\alpha - \beta)] = N(t^*). \quad (4.21)$$

Följande resultat är nu en i stort sett omgående konsekvens av Lemma 1.

LEMMA 2: Som estimator för $\theta(\underline{x})$ har $T(\underline{x}; \underline{t}, \pi, \delta)$ i (4.19) nedanstående (approximativa) skevhet

Relativ underskattning $\approx$

$$\approx \alpha \cdot [1/2 - t^*] - \beta \cdot [\delta \cdot (d + 1/2) - t^* - d]. \quad (4.22)$$

4.3 Bortfall förekommer.

I det följande tillåter vi även att svarsbortfall förekommer (vilket torde vara det i praktiken klart vanligaste fallet). För att indikera att observationerna kan ha utsatts för svarsbortfall modifierar vi beteckningar i enlighet med (3.6). Låt

$$\tilde{W}(\underline{x}) = \text{urvalssumman över de svarande intressanta objekten.} \quad (4.23)$$

$$\tilde{n}(\delta) = n_S + (1-\delta) \cdot \tilde{n}_O \quad (4.24)$$

och inför estimatorn

$$\tilde{T}(\underline{x}; \underline{t}, \underline{\pi}, \delta) = [N^* / \tilde{n}(\delta)] \cdot A \cdot \tilde{W}(\underline{x}). \quad (4.25)$$

I analogi med det tidigare skall vi ge ett approximativt uttryck för väntevärdet av nyss införda estimator, och vi börjar med att skriva ned några relationer som vi kommer att behöva framgent.

Med (2.11), (3.9) och (3.10) som bakgrund tror vi på följande relation;

$$\tilde{W}(\underline{x}) \approx [n_S / n_I] \cdot W(\underline{x}) \approx \phi_I \cdot W(\underline{x}). \quad (4.26)$$

Från (4.19), (4.25) och (4.26) får vi,

$$\tilde{T}(\underline{x}; \underline{t}, \underline{\pi}, \delta) \approx T(\underline{x}; \underline{t}, \underline{\pi}, \delta) \cdot \phi_I \cdot [n(\delta) / \tilde{n}(\delta)] . \quad (4.27)$$

Från (4.7), (4.24), (3.10) och (3.7) fås efter sedvanliga approximationer (och en del detaljer lämnas åt läsaren),

$$\phi_I \cdot [n(\delta) / \tilde{n}(\delta)] \approx 1 + (1-\delta) \cdot \beta \cdot (d+1/2) \cdot (1-\phi_O / \phi_I). \quad (4.28)$$

Formlerna (4.27) och (4.28) tillsammans med Lemma 2 leder nu efter sedvanliga kalkyler till följande resultat (och återigen lämnas detaljer åt läsaren).

LEMMA 3: Som estimator för $\theta(\underline{x})$ har $\tilde{T}(\underline{x}; \underline{t}, \underline{\pi}, \delta)$ i (4.25) nedanstående approximativa skevhet;

Relativ underskattning $\approx$

$$\approx \alpha \cdot [1/2 - t^*] - \beta \cdot [\{\delta \cdot (d+1/2) - t^* - d\} + (1-\delta) \cdot (d+1/2) \cdot (1-\phi_O / \phi_I)] . \quad (4.29)$$

Därmed har vi härlett det teoretiska huvudresultatet, och vi fortsätter med att ge några tillämpningar av det.

5. Tillämpningar.

Vi skall här formulera specialiseringar av Lemma 3 under några olika antaganden om alla de parametrar som finns med i bilden.

Intresset gäller fortsättningsvis egenskaper hos estimatorn i (4.25) som estimator för totalen $\theta(\underline{x})$.

Fall 1: Födelse- och dödsintensiteter är lika.

Från (4.29) framgår att under förutsättningen $\alpha=\beta$ gäller;

Om δ sätts till 1 (dvs "lilla n" väljs som n_S , se (4.24))

ger $\tilde{T}(\underline{x}; \underline{t}, \pi, \delta)$ approximativt väntevärdesriktig estimation av $\theta(\underline{x})$, och approximativ väntevärdesriktighet föreligger;

- (i) oavsett värdet på $\alpha=\beta$,
- (ii) oavsett värdet på d ,
- (iii) oavsett räknepunkterna $\underline{t}$ och vägningen π ,
- (iv) oavsett värdena på ϕ_I och ϕ_O . (5.1)

Estimatorn antar då formen

$$\tilde{T}(\underline{x}; \underline{t}, \pi) = [N^*/n_S] \cdot A \cdot \tilde{W}(\underline{x}). \quad (5.2)$$

Kommentar 5.1: Vid litet eftertanke inses att antagandet om att födelse- och dödsintensiteter är lika innebär ett antagande om att populationen (av levande objekt) är (stokastiskt) stationär. Mot den bakgrunden är det inte förvånande att räknepunkter och vägning inte spelar roll för estimatorns väntevärdesriktighet. ☒

Fall 2: Födelse- och dödsintensiteter får vara olika; flera populationsräkningar.

Från (4.29) framgår följande;

Om δ sätts till 1, dvs om estimatorn väljs enligt (5.2), ger estimatorn approximativt väntevärdesriktig estimation av $\theta(\underline{x})$ om följande relation är uppfylld

$$t^* = 1/2. \quad (5.3)$$

Den approximativa väntevärdesriktigheten föreligger;

- (i) oavsett värdena på α och β ,
- (ii) oavsett värdet på d ,
- (ii) oavsett värdena på ϕ_I och ϕ_O .

Kommentar 5.2: Vi betraktar fallet när man vill beräkna N^* på basis av populationsräkningar vid två tidpunkter, vilka vi kallar t_1 och t_2 och vi antar att $t_1 < t_2$. Då är villkoret (5.3) uppfyllt om vikterna π_1 och π_2 väljs på följande sätt;

$$\pi_1 = (t_2 - 1/2)/(t_2 - t_1), \quad \pi_2 = (1/2 - t_1)/(t_2 - t_1), \quad (5.4)$$

vilket ger

$$N^* = [N_{t_1} \cdot (t_2 - 1/2) + N_{t_2} \cdot (1/2 - t_1)] / (t_2 - t_1). \quad (5.5)$$

I fallet när $t_1 < 1/2 < t_2$ är den naturliga representationen av N^* den som ges i (5.5). I fallet när $t_1 < t_2 < 1/2$ är det kanske naturligare att skriva N^* på följande sätt som en "extrapolation" av N_{t_2} ;

$$N^* = N_{t_2} + (N_{t_2} - N_{t_1}) \cdot (1/2 - t_2) / (t_2 - t_1). \quad (5.6)$$

Notera att i (5.6) blir π -vikten för N_{t_1} negativ.

I det periodiska undersökningsfallet (se Kommentar 4.1) är det naturligt att bestämma N^* enligt (5.5) med hjälp av de populationsräkningar vars tidpunkter ligger på ömse sidor om $1/2$.

Fall 3: Födelse- och dödsintensiteter får vara olika; endast urvalsramen räknad.

När populationen är räknad endast vid urvalstillfället gäller $t^* = -d$. (4.29) ger då följande approximativa uttryck för den relativa underskattningen;

$$(d + 1/2) \cdot [\alpha - \beta \cdot \{\delta + (1 - \delta) \cdot (1 - \phi_{\bar{O}} / \phi_I)\}]. \quad (5.7)$$

Villkor för approximativ väntevärdesriktighet fås genom att sätta uttrycket i (5.7) lika med 0, vilket leder till att den enda parameter som står till förfogande, δ , skall väljas enligt nedan

$$\delta = [\alpha / \beta + 1 - \phi_{\bar{O}} / \phi_I] / [\phi_{\bar{O}} / \phi_I]. \quad (5.8)$$

Under antagande om samma bortfallsbenägenhet för intressanta objekt och övertäkningsobjekt, dvs att $\phi_I = \phi_{\bar{O}}$, förenklar sig formeln (5.8) till

$$\delta = \alpha / \beta. \quad (5.9)$$

Kommentar 5.3: Om man förutom $\phi_I = \phi_{\bar{O}}$ antar att $\alpha = 0$, dvs att inga födelser förekommer, så ger (5.9) $\delta = 0$, vilket innebär att estimatoren är

$$\tilde{T}(\underline{x}) = [N_U / \tilde{n}] \cdot A \cdot \tilde{W}(\underline{x}). \quad (5.10)$$

Estimatoren i (5.10) innebär den gängse "raka uppräknningen". Jämför med utsagan i (4.3).

Kommentar 5.4: Hittills har vi antagit att urvalsförfarandet varit ett OSU-urval från hela populationen. Resultaten kan dock på naturligt sätt appliceras i fallet med stratifierat urval med OSU-urval inom stratan. Man genomför stratumvisa skattningar av stratumtotaler med hjälp av det som sagts ovan och adderar sedan de skattade stratumtotalerna för att få en skattning av totalen över hela populationen.

6. UVAV-undersökningen; tillämpning och numerisk belysning.

Bakgrunden för de tidigare presenterade allmänna resultaten har varit planeringen av UVAV-undersökningen för 1990. Vi ska här gå litet närmare in på den undersökningen, dels för att ge konkretisering av de allmänna begreppen och resultaten dels för att diskutera några frågeställningar i den undersökningen per se. Särskilt skall vi intressera oss för antagandet (2.8) om att födelse- och dödsintensiteter är förhållandevis stabila över tiden från urvalstidpunkt till referensperiodens slut.

6.1 Om UVAV-undersökningen.

I UVAV-undersökningen vill man skatta olika transportvariabel-aggregat för transporter med lastbilar som har maxlastvikt 2 ton eller mer. En "primär" UVAV-undersökning har referensperioden kvartal (mer precist en 13-veckors-period). Mätperioden är en vecka, under vilken bilarna i undersökningen för journal över sina körningar. Urvalsramen för en undersökning hämtas från Centrala Bilregistret (CBR) genom att i det selektera ut de lastbilar med maxlastvikt ≥ 2 ton som är s k aktivregistrerade vid urvalstillfället. Urvalet dras som ett stratifierat OSU-urval (jämför Kommentar 5.4). Urvalstidpunkten ligger cirka två månader före referensperiodens början.

Under ett UVAV-undersökningsår genomförs nyss nämnda "primära" undersökning för vart och ett av årets fyra kvartal. De skattade kvartalstotalerna adderas sedan till en skattad årstotal. T o m 1987 genomfördes UVAV "löpande", dvs år efter år, och man befann sig då i den typ av situation som vi i Kommentar 4.1 kallade "periodisk undersökning". Sedan dess har UVAV gjorts intermittent, och genomförs nu vart tredje år, med 1990 som UVAV-undersökningsår. Det är UVAV90 vi skall intressera oss för i fortsättningen, och den undersökningen kan rubriceras som "semi-periodisk" (enligt klassificeringen i Kommentar 4.1). Nedan anges referensperioder och urvalstidpunkter för UVAV90.

Referensperiod	Urvalstidpunkter (genomförda och planerade)
1:a kvartalet (jan,feb,mar 90)	30 okt 89
2:a kvartalet (apr,maj,jun 90)	31 jan 90
3:e kvartalet (jul,aug,sep 90)	30 april 90
4:e kvartalet (okt,nov,dec 90)	juli/aug 90

Åtminstone en "extra" populationsräknetidpunkt kommer att att läggas in i slutet av året.

Med beteckningar och normering enligt Avsnitt 2 är alltså tidsenheten 3 månader och $d \approx 2/3$. Vidare är $A=13$.

I UVAV är objekten av intresse lastbilar med maxlastvikt om minst två ton. Härnäst några ord om den konkreta innebörden av begreppen "levande", "födas" och "dö" i UVAV-sammnhanget, åtminstone som de hittills använts. (Vi föreslår viss modifiering längre fram.)

En lastbil räknas som levande om den är registrerad som aktiv i CBR. En lastbil kan dö antingen genom att den avregistreras (t ex när bilen skrotas) eller genom att den avställes (en åtgärd som medför bl a att skatt inte behöver betalas för den period bilen är avställd). Lastbilar kan födas dels genom att genuint nya bilar förs in i beståndet, nyregistrering sker, eller genom att en avställd bil aktivregistreras.

6.2 Om födelse- och dödsintensiteter.

För att kunna bedöma riktigheten av antagandet (2.8) om födelse- och dödsintensiteternas (dvs α och β) någorlunda konstans i tiden skulle man idealt vilja ha tillgång till numeriska värden på α och β som hänför sig till tidsperioder som är kortare än referensperioden. I UVAV-fallet vore månadsvisa intensitetsvärden en lagom fin uppdelning. Tyvärr är dock möjligheten att komma över empiriska intensitetsvärden mycket begränsad, men ett par källor kan i alla fall ge viss information enligt vad som redovisas nedan; dels UVAV90s två hittillsvarande urvalsramar (per 31 okt 89 resp 31 jan 1990) dels den löpande statistiken över lastbilsbeståndet.

För de nyss nämnda urvalsramarna har räknats; dels antalet bilar som fanns med i båda urvalsramarna, dels antalet bilar som fanns med i den första men ej i den andra, dels antalet bilar som fanns med i den andra men ej i den första. Under antagande om att transienterna är så få att man kan bortse från dem kan resultatet beskrivas enligt nedan.

Populationens storlek per 89-10-30:	70 950	
Antal födda 89-10-30 - 90-01-31:	3 652	
Antal döda 89-10-30 - 90-01-31:	8 281	
Populationens storlek per 90-01-31:	66 321.	(6.1)

Från (6.1) följer att för perioden 89-10-30 - 90-01-31 gällde;

$$\alpha = 5.2\% , \quad \beta = 11.8\% . \quad (6.2)$$

Om inte annat så framgår från (6.2) att;

- Födelseintensiteten var så pass hög att man inte kan bortse från födelserna utan att påtaglig skevhet uppstår.
- Justering för undertäckning baseras naturligen på premisen att födelse- och dödsintensiteter kan vara olika stora.

Värdena i (6.2) är egentligen det enda riktigt bra empiriska intensitetsvärden som står till förfogande, men de ligger ju långt från det som ovan rubricerades som idealt; månadsvisa födelse- och dödsintensiteter. Dock, viss information åt det önskade hållet kan fås ur statistiken över totala lastbilsbeståndet enligt CBR, och utdrag ur den statistiken redovisas nedan i Tabell 6.1. Notera att det där handlar om totala lastbilsbeståndet, vilket är avsevärt större än UVAV-beståndet eftersom det omfattar även lastbilar med maxlastvikt mindre än 2 ton.

	Antal aktivregi- strerade lastbilar		Antal nyregistreade lastbilar	
	1988	1989	1988	1989
jan	245 585	267 974	2 278	3 472
febr	246 692	270 188	2 601	3 102
mars	250 231	275 170	3 109	3 825
april	256 007	281 619	2 789	4 075
maj	262 110	287 600	3 274	3 728
juni	266 477	291 831	3 683	4 021
juli	262 314	287 625	2 647	2 588
aug	269 537	294 784	2 771	3 143
sept	269 736	295 626	3 210	3 242
okt	269 138	294 339	3 329	3 482
nov	268 000	293 725	3 526	3 670
dec	267 281	294 901	4 170	6 453

Tabell 6.1. Utveckling av lastbilsbeståndet, enligt CBR.

Från värdena i Tabell 6.1 kan man beräkna födelseintensiteter vad gäller den del av födelserna som hänför sig till nyregistreringar (men inte vad gäller den del som avser aktivering av avställda bilar). Man kan också beräkna förändringsintensiteter vad gäller aktiv-registrerade bilar. Om man bortser från transienter, kan den senare intensiteten tolkas som $\alpha-\beta$. Nyss nämnda intensiteter redovisas i nedanstående Tabell 6.2. Notera att intensiteterna anges "per tidsenhet" vilket i UVAV-fallet innebär "per kvartal".

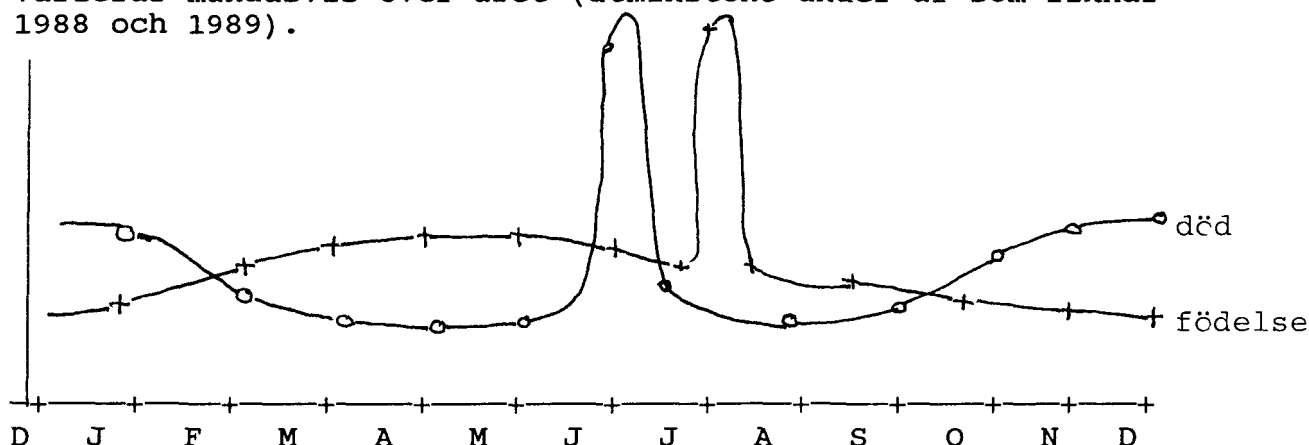
	Skillnad mellan födelse- och dödsintensitet ($\alpha-\beta$).		Nyregistreringsin- tensitet.	
	1988	1989	1988	1989
jan	-0.2%	0.8%	2.8%	3.9%
febr	1.4%	2.5%	3.2%	3.4%
mars	4.3%	5.5%	3.7%	4.2%
april	6.9%	7.0%	3.3%	4.3%
maj	7.2%	6.4%	3.7%	3.9%
juni	5.0%	4.4%	4.1%	4.1%
juli	-4.7%	-4.3%	3.0%	2.7%
aug	8.3%	7.5%	3.1%	3.2%
sept	0.2%	0.9%	3.6%	3.3%
okt	-0.7%	-1.3%	3.7%	3.5%
nov	-1.3%	-0.6%	3.9%	3.7%
dec	-0.8%	1.2%	4.7%	6.6%
året	2.1%	2.5%	3.5%	3.9%

Tabell 6.2. Vissa förändringsintensiteter för lastbilsbeståndet.

Från värdena i Tabell 6.2 drar vi följande tentativa slutsatser angående UVAV-populationens födelse och dödsintensiteter över året, slutsatser som också stöds av annan erfarenhet;

- Nyregistreringsintensiteten ligger tämligen konstant över året. Det finns anledning att förmoda att samma förhållande föreligger vad gäller "att ta bilar permanent ur drift".
- De förhållandevis stora fluktuationer över året som kan noteras för $(\alpha-\beta)$ hänger framför allt samman med att mer eller mindre kortvariga avställningar av bilar görs, vilka sedan följs av aktivregistreringar. Huvudmönstret ges härvidlag av nedanstående förhållanden;
 - Under vintern ställs vissa bilar av under jämförelsevis långa perioder.
 - Ett stort antal bilar ställs av under "semestern", och då är det fråga om klart kortvariga avställningar, 4 à 5 veckor. Avställningsperioderna ligger framför allt i juli.

Mot ovanstående bakgrund formulerar vi nedanstående förmodade bild av hur UVAV-populationens födelse- och dödsintensiteter varierar månadsvis över året (åtminstone under år som liknar 1988 och 1989).



Figur 6.1. Förmodat utseende av UVAV-populations födelse- och dödsintensiteter.

6.3 Rekommendationer för undertäcknings-justering.

Vi har redan pekat på att UVAV-populationens födelse- och dödsintensiteter skiljer sig åt rätt markant. Vidare, flera populationsräkningar kommer att göras under undersökningsåret. Det ligger därför närmast till hands att applicera resultaten under Fall 2 i Avsnitt 5. Det som dock vållar problem är den anhopning av kortvariga avställningar som förekommer under juli, med tillhörande kraftiga höjningar av döds- och födelseintensiteter i juli och augusti. För att dämpa effekten av denna komplikation föreslås följande modifieringar relativt det "periodiska fallet";

- Urvalsramen för det 4:e kvartalet tas ut den 15:e augusti (om det är praktiskt möjligt att göra urvalet så sent. (6.3)
- Den "extra" populationsräkningen görs den 15:e november. (6.4)
- Under del av tredje kvartalet används modifierade definitioner av svarande intressanta och övertäckning. (Mer om det nedan.) (6.5)

Vad gäller (6.3), är förmodan att per mitten av augusti har den övervägande delen av de semesteravställda bilarna återigen aktiverats.

I enligt med ovanstående förslag räknar vi fortsättningsvis med nedanstående populationsräknetidpunkter.

Räknepunkt	Beteckning för stratumstorlekar
89-10-30	N(okt89)
90-01-31	N(jan90)
90-04-30	N(apr90)
90-08-15	N(aug90)
90-11-15	N(nov90).

(6.6)

Vi skjuter på behandlingen av det litet komplicerade 3:e kvartalet. För de övriga kvartalen anser vi födelse- och dödsintensiteter vara så pass stabila att Fall 2 i Avsnitt 5 kan tillämpas. Nedan anges de värden på "lilla n" och "stora N" som ges av Fall 2 när "stora N" beräknas genom vägning enligt (5.5) av de N-värden som kommer från populationsräkningarna som ligger på ömse sidor om referensperiodens mittpunkt.

1:a kvartalet: "Lilla n" och "stora N" väljs stratumvis enligt;

$$n = n_S, \quad (6.7)$$

$$N^* = (5/6) \cdot N(\text{jan90}) + (1/6) \cdot N(\text{apr90}). \quad (6.8)$$

2:a kvartalet: "Lilla n" och "stora N" väljs stratumvis enligt;

$$n = n_S, \quad (6.9)$$

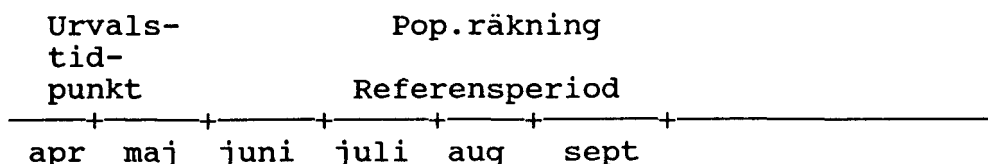
$$N^* = (6/7) \cdot N(\text{apr90}) + (1/7) \cdot N(\text{aug90}). \quad (6.10)$$

4:e kvartalet: "Lilla n" och "stora N" väljs stratumvis enligt;

$$n = n_S, \quad (6.11)$$

$$N^* = N(\text{nov90}). \quad (6.12)$$

3:e kvartalet vållar som sagt problem beroende på avställningarna i juli. För det kvartalet ser den "primära" undersökningen ut på följande sätt;



Figur 6.2: 3:e kvartalets undersökning.

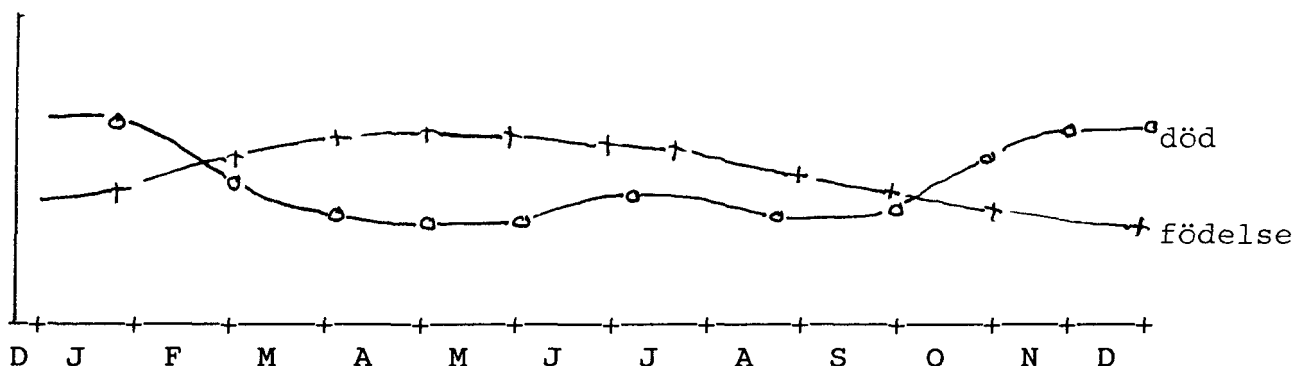
Notera att en semesteravställning i juli leder till det som kallades en transient. För 3:e kvartalet kommer inte antagandet (2.12) om transienternas försumbarhet att vara uppfyllt. Ett sätt att komma runt denna svårighet, och det kanske naturligaste sättet åtminstone om det inte föranleder alltför stora praktiska svårigheter, är att för bilar som har mätvecka under juli och första halvan av augusti ändra definitionerna av födelse och död enligt nedan.

Som död räknas avregistrering samt avställning som varar mer än 40 dagar. (6.13)

Som födelse räknas nyregistrering samt aktivering av bil som varit avställd mer än 40 dagar. (6.14)

En konsekvensen av ovan nämnda modifikation är att "svarande intressanta" och "övertäckning" ändrar innebörd.

Om födelse och död definieras enligt (6.13) och (6.14) kommer födelse- och dödsintensiteter att modifieras till någonting i nedanstående stil (åtminstone förmodligen);



Figur 6.3. Förmodade födelse- och dödsintensiteter med modifierade definitioner av födelse och död.

För att kunna identifiera övertäckning enligt den modifierade definitionen måste frågeställandet ändras något i förhållande till frågeformuläret, förslagsvis på följande sätt. De som svarar att bilen varit avställd under mätvecka i juli eller första halvan av augusti kontaktas för att inhämta svar på en uppföljningsfråga enligt nedan, och de får alltså två frågor angående avställning.

Var bilen avställd under mätveckan ? Ja Nej

Om svaret på ovanstående fråga är Ja;
Beräknas avställningen gälla i mer än
40 dagar? Ja Nej

Enligt hittillsvarande UVAV-rutiner räknas alla som svarar Ja på första frågan som övertäckning. Förslaget innebär att detta modifieras så att de som svarar Ja + Nej på ovanstående två frågor inte räknas som övertäckning utan som svarande intressanta objekt. För dessa objekt sätts alla transportvariabelvärden till 0. Objekten ifråga behandlas alltså på samma sätt som de som svarar; Bilen ej avställd, men utnyttjades inte under någon dag i mätveckan.

Därefter behandlas 3:e kvartalet på följande sätt.

3:e kvartalet: "Lilla n" och "stora N" väljs stratumvis enligt;

$$n = n_S \text{ (där antalet svarande beräknas i enlighet med ovanstående modifikation),} \quad (6.14)$$

$$N^* = N(\text{aug90}). \quad (6.15)$$

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metod-rapporter från SCB).

R & D Reports Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown /beige/ covers).

Report published earlier during 1990:

- | | |
|------------------|--|
| 1990:1
(grön) | Calibration Estimators and Generalized Raking Techniques in Survey Sampling (Jean-Claude Deville, Carl-Erik Särndal) |
| 1990:2
(grön) | Sampling, Nonresponse and Measurement Issues in the 1984/85 Swedish Time Budget Survey (Ingrid Lyberg) |
| 1990:3
(grön) | Om justering för undertäckning vid undersökningar med urval i "rum och tid" (Bengt Rosén) |

Kvarvarande beige och gröna exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 78.

Kvarvarande gula exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.