Research and Development – Methodology reports from Statistics Sweden 2008:1

Statistiska centralbyrån Statistics Sweden

Optimalt antal kontaktförsök i en telefonundersökning

Annica Isaksson
Peter Lundquist
Daniel Thorburn

Rapportserien "**Research and Development – Methodology Reports from Statistics Sweden**" publicerar rapporter med resultat från SCB:s forsknings- och utvecklingsverksamhet. Fokus i rapportserien ligger på metodutveckling för offentlig statistikproduktion. Rapportserien publicerar bidrag från alla avdelningar inom SCB och är öppen för bidrag som behandlar en vid mängd av olika metodologiska problem.

Utgivna publikationer i serien:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

2007:2 Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator.

2007:3 Kartläggning av felkällor för bättre aktualitet

.

Research and Development – Methodology reports from Statistics Sweden 2008:1

Optimalt antal kontaktförsök i en telefonundersökning

Annica Isaksson
Peter Lundquist
Daniel Thorburn

Statistiska centralbyrån
2008

Research and Development – Methodology reports from Statistics Sweden 2008:1

On the optimum number of call attempts in a telephone survey

Statistics Sweden
2008

Producent
Producer
SCB, utvecklingsavdelningen
Statistics Sweden, Research and Development Department
SE-701 89 ÖREBRO
+ 46 19 17 60 00

Förfrågningar
Inquiries
Peter Lundquist, +46 8 5069 49 18
peter.lundquist@scb.se

Omslag/Cover Ateljén, SCB

ISSN 1653-7149
URN:NBN:SE:SCB-2008-X103BR0801_pdf (pdf)
Denna publikation finns enbart i elektronisk form på www.scb.se

Förord

Att genomföra en statistisk undersökning är ett komplicerat företag och involverar många beslut kring praktiska problem. Ett sådant problem är att bestämma antalet kontaktförsök med personer i urvalet, personer som inte besvarar brevenkäter eller telefonsamtal trots upprepade försök. Uteblivna svar innebär bortfall som riskerar att snedvrida undersökningens resultat. De ansvariga för undersökningen har därmed ett incitament till att genomföra förnyade kontaktförsök. Nya kontaktförsök innebär dock andra konsekvenser som kan motverka de positiva effekterna av en högre svarsfrekvens. Exempel är att undersökningen tar längre tid och kostnaderna ökar. I föreliggande rapport behandlas därtill effekter av intervjuarvarians och mätfel. Författarna föreslår en ansats till att väga för- och nackdelar med förnyade kontaktförsök och presenterar en metod för beräkning av optimalt antal kontaktförsök. Författarnas systematiska hantering av problemet och förslag till lösning är användbar i praktiska sammanhang. Den som är ansvarig för en undersökning får ett verktyg som kan användas för bestämning av antalet kontaktförsök. Resultaten lyfter också upp själva problematiken och poängterar behovet av speciella insatser när antalet kontaktförsök ökar.

Statistiska centralbyrån i februari 2008

Folke Carlsson

Innehåll

A separate text in English is provided at the end of the publication, on page 59.

Sammanfattning

I en telefonintervjuundersökning måste man ofta ringa utvalda individer flera gånger innan man får tag på dem. Vi behandlar problemet att bestämma hur många uppringningsförsök som ska göras innan man klassificerar en individ som ej anträffad. Utgångspunkten är att man vill uppskatta en total för en population av individer på basis av en urvalsundersökning. Data samlas in med hjälp av telefonintervjuer. Det bortfall som uppstår i samband med datainsamlingen hanteras i estimationen genom att man delar in urvalet i svarshomogenitetsgrupper och använder den så kallade direktvägningsestimatorn. Vi presenterar en metod för optimal bestämning av den övre gränsen för antalet kontaktförsök. Metoden tar hänsyn till varierande urvalssannolikheter, svarsvägran, intervjuarvarians och mätfel och till olika kostnader förknippade med datainsamlingen. Våra modeller för bortfall och mätfel medger att felkällornas inverkan på estimatorn förändras över antalet kontaktförsök. Antalet kontaktförsök bestäms genom en jämförelse av storleken på estimatorns standardfel i undersökningar med olika antal kontaktförsök – undersökningar som har dimensionerats så att de har samma kostnad. Som avslutning ges ett enkelt men realistiskt räkneexempel på bestämning av antalet kontaktförsök i en specifik undersökningssituation.

1 Inledning

Utformningen av en statistisk urvalsundersökning fordrar en mängd beslut, förknippade med olika faser av undersökningen. Vi fokuserar här på undersökningar där observationsobjekten är individer. Vårt forskningsproblem är hur många försök man ska göra att få tag på varje utvald individ innan han eller hon betraktas som bortfall. Den här beslutssituationen uppstår endast om vald datainsamlingsmetod fordrar kontakt med utvalda individer. Till datainsamlingsmetoder av denna typ hör exempelvis postala enkäter, besöks- och telefonintervjuer.

I en postal enkät utgör det första utskicket av frågeformulär med följebrev det första kontaktförsöket. De påminnelser som sänds till dem som inte svarat utgör efterföljande kontaktförsök. För enkelhetens skull görs utskicken ofta samtidigt till alla som ännu inte svarat, även om man teoretiskt skulle kunna skräddarsy strategier för olika delgrupper. I fallet postala enkäter handlar bestämning av antalet kontaktförsök om att avgöra när och hur många påminnelser som ska skickas ut och när datainsamlingen ska avbrytas. Beslutssituationen kompliceras om exempelvis aviseringsbrev eller kombinerade tack- och påminnelsebrev används.

Vi är främst intresserade av telefonintervjuundersökningar. I sådana undersökningar består det första kontaktförsöket i allmänhet av att man försöker ringa alla utvalda individer (eventuellt med stöd av en dator för själva uppringningen). Av praktiska skäl kan man inte, till skillnad från i postala enkäter, kontakta alla i stickprovet samtidigt: det finns bara ett begränsat antal intervjuare att tillgå. Dessutom kan det finnas önskemål om att sprida intervjuerna jämnt över mätperioden. Om man vet något om den utvalda personen kan det också vara motiverat att låta antalet kontaktförsök vara beroende på dennes egenskaper. Erfarenhetsmässigt vet man exempelvis att yngre ensamstående män har andra levnadsvanor än barnlediga mammor. Det typiska vid telefonintervjuer är att individernas ”kontaktstatus” (hur många gånger de har hunnit kontaktas) varierar ganska kraftigt mellan olika individer i stickprovet vid en given tidpunkt under mätperioden.

Vi förutsätter för enkelhetens skull att individurvalet görs från en bra ram över målpopulationen (som också den består av individer).

Följaktligen behandlar vi inte telefonnummerurval, som har sina speciella problem. En annan avgränsning är att vi endast behandlar engångsundersökningar. I panelundersökningar med roterande urval kan kontakt vid en undersökningsomgång underlätta etablerandet av kontakt i nästa omgång – ett fenomen som vi inte alls beaktar här.

I SCB:s undersökningar brukar man kräva att man försökt kontakta samtliga utvalda individer minst en gång före en given tidpunkt. Efterföljande kontaktförsök består av nya uppringningar vid senare tidpunkter. Vi tänker oss en situation där man på förhand bestämmer antalet gånger, A, som man ska försöka kontakta utvalda objekt som inte redan hunnit svara. På SCB används en sådan regel exempelvis i Arbetskraftsundersökningarna, där gränsen är satt till A=12.

Det är inte alldeles självklart hur ett kontaktförsök ska definieras. Har ett kontaktförsök ägt rum om en intervjuare ringt ett nummer där ingen svarat? Har ett kontaktförsök ägt rum om intervjuaren talat med någon annan person i hushållet än den som är utvald? Om intervjuaren ringt ett nummer och fått beskedet att abonnemanget har upphört? Vi förutsätter här att man använder sig av en teoretiskt rimlig, och praktiskt användbar, definition av kontaktförsök.

Det är komplicerat att planera och styra datainsamlingen i telefonintervjuundersökningar. Vi fokuserar på problemet att bestämma antalet kontaktförsök dels eftersom denna fråga behandlats mycket sparsamt i litteraturen, dels eftersom studier tyder på att antalet kontaktförsök kan ha stor betydelse för skattningarna (exempelvis visar figurerna 4b-c i Japec, 2005, Paper IV, hur bias och medelkvadratfel för en skattning av andel arbetslösa icke-svenska nordiska medborgare påverkas kraftigt av antalet kontaktförsök.) Vi avstår ifrån att behandla andra intressanta och närbesläktade frågor, såsom när på dagen eller veckan kontakterna ska tas med individerna i urvalet – det som brukar kallas för uppringningsalgoritmer eller kontaktstrategier (på engelska call scheduling algorithms). Kontaktstrategier för telefonintervjuer behandlas exempelvis i Japec och Lundquist (2000) och Weeks, Kulka och Pierson (1987).

Det är inte självklart att man bör göra så många kontaktförsök som möjligt. Ju fler kontaktförsök man gör, desto mer data kan man visserligen förvänta sig att få in. Men kostnaden för ytterligare kontaktförsök kan vara hög i förhållande till sannolikheten att få

kontakt och dessutom svar. I allmänhet har urvalsstorleken valts för att en viss precision i skattningarna ska erhållas. För att uppnå önskat antal svar kan det vara mera kostnadseffektivt att dra ett stort initialt urval och göra få kontaktförsök, än att göra många försök att kontakta individerna i ett mindre initialt urval. Inte heller ur biassynpunkt är det självklart att man ska göra så många kontaktförsök som möjligt. Visserligen bör bortfallsbiasen avta om man får in flera svar, men för vissa undersökningsvariabler riskerar man istället att introducera en växande mätfelsbias (till följd av exempelvis minnesfel). Ytterligare en aspekt på att göra många kontaktförsök är att produktionstiden förlängs, vilket försämrar statistikens aktualitet. Vår metod för bestämning av antalet kontaktförsök beaktar inte värdet av hög aktualitet, men vi är medvetna om att det är en viktig aspekt.

I en serie rapporter (Tångdahl, 2004, 2005, 2006) utreder Tångdahl hur olika estimatorers bias och varians förändras över fältarbetsperiodens gång, och föreslår en procedur för löpande avstämningar i syfte att avgöra när i tiden datainsamlingen ska avbrytas. Tångdahls arbete är i första hand anpassat till postala enkäter. De källor till bias och varians som hon beaktar är urval och svarsbortfall. Vårt arbete är besläktat med Tångdahls, men till skillnad från henne fokuserar vi på telefonintervjuundersökningar, och på problemet att bestämma hur många gånger man ska kontakta utvalda individer innan datainsamlingen avbryts. Utöver urval och bortfall tar vi hänsyn till mätfel. Vi intresserar oss för problemet att *på förhand* besluta hur många kontaktförsök som ska göras. I praktiken efterfrågas ofta ett sådant beslut, eftersom man vill bedöma den totala kostnaden för undersökningen, hur många intervjuare som kommer att behövas etc.

Vår metod för att på förhand fatta beslut om antalet kontaktförsök vilar på explicita antaganden. Detta möjliggör att metoden kan omprövas när kunskapen växer eller omständigheterna förändras. I beslutet kan inverkan av urval, bortfall och mätfel på skattningarna vägas in, liksom kostnaderna för datainsamling. Rent konkret innebär metoden att de undersökningskostnader som beror av stickprovsstorleken multipliceras med ett uttryck för aktuell skattnings varians (med avseende på urvalsdesignen och modeller för kontakt, svar och mätfel), varigenom ett approximativt uttryck för optimalt antal kontaktförsök erhålles.

I kapitel 2 placerar vi in vårt arbete i teorin för urval ur ändliga populationer. Notera dock att vårt problem även finns vid oändliga populationer, där man kan bortse ifrån ändlighetskorrektionen. I kapitel 3 uppmärksammar vi bortfall och mätfel som potentiella felkällor. Vi introducerar direktvägningsestimatorn, och en modell för mätfelens inverkan på observationsdata. Som ett alternativ till direktvägningsestimatorn när det finns hjälpinformation om alla individer i populationen föreslår vi den separata kvotestimatorn. De statistiska egenskaperna (väntevärde och varians) hos nämnda estimatorer utreds i kapitel 4. En enkel kostnadsmodell formuleras i kapitel 5. I kapitel 6 beskriver vi hur kostnadsmodellen, tillsammans med variansuttrycket från kapitel 4, kan användas för att bestämma antalet kontaktförsök. Kapitel 7 ger en illustration av hur den föreslagna metoden kan användas i praktiken. Våra räkneexempel i kapitel 7 baseras delvis på verkliga data. I kapitel 8, slutligen, summerar och diskuterar vi våra resultat.

2 Utgångspunkt

Betrakta en ändlig population U av individer, indexerade $k = 1,...,N$. Vi låter individerna i populationen representeras av sina index och betecknar populationen som $U = \{1,...,k,...,N\}$. Populationens storlek, *N*, antas vara känd. Vi är intresserade av att uppskatta populationstotalen för undersökningsvariabeln μ,

$$t_\mu = \sum_U \mu_k \tag{1}$$

där μ_k betecknar det fixa värdet på μ för individ $k \in U$. (I formel (1) är $\sum_U \mu_k$ en kortform av $\sum_{k \in U} \mu_k$. Vi använder genomgående detta förkortade skrivsätt. Om *D* är en mängd populationselement sådana att $D \subseteq U$ så skriver vi alltså $\sum_D$ för $\sum_{k \in D}$.)

I syfte att skatta t_μ dras ett stickprov *s* av storlek *n* från U. Variabeln μ_k observeras för alla $k \in s$. Dessutom är värdet på en hjälpvektor **x** känt. Vissa komponenter i hjälpvektorn är kända för alla individer i U och kan användas till att bestämma urvalsplanen och därmed inklusionssannolikheterna π_k. Det kan handla om registervariabler av typen ålder, kön och bostadsort. Övriga komponenter i hjälpvektorn är enbart kända för urvalet *s*. Detta gäller exempelvis ”tilldelningen” av intervjuare; alltså vilka intervjuare som utses att kontakta vilka utvalda individer. Det finns ingen anledning att utse intervjuare till individer som inte ingår i urvalet. Samtliga komponenter i hjälpvektorn kan användas för att modellera utvalda individers svarsbenägenhet (se avsnitt 3.1).

3 Bortfall och mätfel

Här introducerar vi modeller för uppkomsten av svarsbortfall och mätfel.

3.1 Bortfall

I praktiken är den fältarbetsperiod och de resurser som är tillgängliga för en undersökning alltid begränsade. Vid någon tidpunkt måste datainsamlingen avbrytas även om man endast fått in svar från en delmängd av stickprovet. Vi utgår ifrån att insamlingen avbryts när precis A stycken försök gjorts att kontakta de individer i stickprovet som inte redan hunnit svara. Den delmängd av stickprovet som man lyckats etablera kontakt med efter A kontaktförsök (eller tidigare) kallar vi *kontaktmängden* $s^{(A)}$ av storlek $n^{(A)}$. Vi definierar också de successiva delmängderna $s^{(1)} \subseteq ... \subseteq s^{(A)} \subseteq s$. I $s^{(A)}$ ingår även de som har svarat med att avböja medverkan.

Om vissa av de individer som man lyckas etablera kontakt med avböjer att medverka i undersökningen uppstår ett svarsbortfall. Låt $r^{(A)}$ beteckna mängden *verkligen* svarande $\left(r^{(A)} \subseteq s^{(A)}\right)$ av storlek $m^{(A)}$.

Vi förutsätter att man i estimationen försöker ta hänsyn till saknade data genom att dela in urvalet i svarshomogenitetsgrupper (på engelska response homogeneity groups, RHG). Vi definierar en RHG-modell genom att s delas in i H_s grupper $s_1,...,s_h,...,s_{H_s}$ (av storlek $n_1,...,n_h,...,n_{H_s}$) sådana att individerna inom en och samma grupp antas ha samma svarsmönster, dvs. samma sannolikhet att ingå i $r^{(a)}; 1 \le a \le A$ – sannolikheten behöver dock inte vara densamma i *olika* grupper. Indelningen görs med hjälp av hjälpvariabeln $\mathbf{x}_k$. Definiera också mängderna $r_h^{(A)} = r^{(A)} \cap s_h$ av storlek $m_h^{(A)}$ för $h = 1,...,H_s$. Det finns metoder att justera skattningar för bortfall, men vi fokuserar på denna enkla modell. Formellt har den följande utseende:

RHG-modell

$$\Pr\left(k \in r^{(A)} | s\right) = \theta_{k|s}^{(A)} = \theta_{hs}^{(A)} > 0 \quad \text{för alla } k \in s_h \text{ och alla } A \qquad (2)$$

$$\Pr\left(k \,\&\, l \in r^{(A)} | s\right) = \theta_{k|s}^{(A)} \theta_{l|s}^{(A)} \quad \text{för alla } k \neq l \in s \text{ för } h = 1,\ldots,H_s \qquad (3)$$

samt

$$\Pr\left(k \in r^{(A)} \,\&\, l \in r^{(B)} | s\right) = \theta_{k|s}^{(A)} \theta_{l|s}^{(B)} \text{ för alla } k \neq l \in s \text{ och alla } A, B \qquad (4)$$

och motsvarande antaganden om oberoende mellan fler än två individer.

Notera att gruppindelningen tillåts bero av vilket stickprov man fått. Detta leder till att estimatorer baserade på RHG får icke-explicita variansuttryck – en välkänd nackdel med att använda sig av en sådan här typ av modell.

Antagande (4) i RHG-modellen används framför allt i analysen av vår mätfelsmodell (se avsnitt 3.2).

Vår RHG-modell sammanfaller nästan helt med motsvarande modell i Tångdahl (2004, s. 9). En viktig skillnad är dock att vi låter *a* beteckna ett givet *antal kontaktförsök*, medan Tångdahl definierar *a* som en given *tidpunkt* under fältarbetsperioden. Tångdahls modell är i sin tur en lätt modifierad version av RHG-modellen i Särndal, Swensson och Wretman (1992, formel (15.6.6)).

I kapitel 5 modellerar vi, utöver svarsbortfall, även svarsvägran och behöver då kontaktsannolikheterna

$$\Pr\left(k \in s^{(A)} | s\right) = \omega_{k|s}^{(A)} = \omega_{hs}^{(A)} \qquad (5)$$

Kontaktsannolikheterna antas vara konstanta inom svarshomogenitetsgrupperna, och antas även uppfylla andra ordningens villkor, motsvarande antagandena (3) och (4) i RHG-modellen.

Det finns många estimatorer av t_μ som på något sätt justerar för bortfall (se Särndal m.fl. 1992, kap. 15). Vi väljer i denna rapport att fokusera på *direktvägningsestimatorn*, som formelmässigt är ganska enkel att hantera. Den har också fördelen att vara väntevärdesriktig för t_μ (givet att alla $m_h^{(A)} > 0$) om RHG-modellen är sann.

Estimatorn har dock i allmänhet större varians än exempelvis estimatorer av kvottyp. Direktvägningsestimatorn ges som

$$\hat{t}_{\mu c\pi^*}^{(A)} = \sum_{r^{(A)}} \frac{\mu_k}{\pi_k \hat{\theta}_{k|s}^{(A)}} = \sum_{h=1}^{H_s} \frac{n_h}{m_h^{(A)}} \sum_{r_h^{(A)}} \breve{\mu}_k \qquad (6)$$

(Tångdahl, 2004, formel (1); Särndal m.fl., 1992, formel (15.6.8)) där $\breve{\mu}_k = \mu_k / \pi_k$ och π_k betecknar första ordningens inklusionssannolikhet för individ *k*.

Om hjälpvariablerna $\mathbf{x}_k$ är kända för alla individer i U kan man stratifiera (eller poststratifiera) urvalet efter RHG. I så fall är den *separata kvotestimatorn* av t_μ att föredra – den har både mindre varians och en enklare variansskattningsformel än direktvägningsestimatorn. Den separata kvotestimatorn ges som

$$\hat{t}_{\mu r}^{(A)} = \sum_{h=1}^{H_s} N_h \frac{\sum_{r_h^{(A)}} \mu_k / \pi_k}{\sum_{r_h^{(A)}} 1/\pi_k} = \sum_{h=1}^{H_s} \frac{N_h}{\hat{t}_{M_h}^{(A)}} \sum_{r_h^{(A)}} \breve{\mu}_k \qquad (7)$$

(Särndal m.fl., 1992, formel (7.7.1)), där N_h är storleken på (post)stratum *h* och $\hat{t}_{M_h}^{(A)} = \sum_{r_h^{(A)}} 1/\pi_k$ kan betraktas som en skattning av antalet svarande efter *A* kontaktförsök i en totalundersökning av U. I föreliggande rapport förekommer $\hat{t}_{M_h}^{(A)}$ i härledningen av mätfelsvariansen för $\hat{t}_{\mu r}^{(A)}$ (se appendix B).

Andra tänkbara sätt att utnyttja hjälpvariablerna för att förbättra skattningarna inkluderar andra typer av regressionsestimatorer, imputering och propensity scoring (se anmärkning 3.3).

Anmärkning 3.1
Det är troligt att intervjuarna påverkar $\theta_{k|s}^{(A)}$. Vi antar att varje individ i *s* tilldelas en och endast en intervjuare. Informationen om vilken intervjuare detta är ingår i hjälpvektorn $\mathbf{x}_k$. När vi härleder varianserna för $\hat{t}_{\mu c\pi^*}^{(A)}$ och $\hat{t}_{\mu r}^{(A)}$ i kapitel 4 antar vi emellertid att $\mathbf{x}_k$ inte beror på *s*. Vi utgår från att urvalet *s* fördelas slumpmässigt över intervjuarna så att alla intervjuare tilldelas samma antal utvalda individer. Vi utgår också ifrån att intervjuartilldelningen och indelningen i svarshomogenitetsgrupper sker oberoende. Om intervjuartilldelningen sker på något annat sätt än vad som just beskrivits kan formlerna i kapitel 4 behöva justeras.

Anmärkning 3.2
Vår RHG-modell skulle kunna formuleras som att data är Missing at Random (MAR) givet gruppindelningen. Till standardreferenserna om MAR hör Rubin (1983, 1987) och Little och Rubin (2002).

Anmärkning 3.3
Ett alternativ till RHG-modellen är att använda sig av propensity score-teknik. Tekniken går ut på att man gör indelningen i svarshomogena grupper med hjälp av faktiskt observerat bortfall (Rosenbaum och Rubin, 1984). Först uppskattas svarssannolikheten som funktion av hjälpvektorn utifrån en lämplig modell. Stickprovet delas därefter in i grupper sådana att individer som har nästan samma skattade svarssannolikheter förs till samma grupp. Till skillnad mot i en RHG-modell antas svarssannolikheterna inom grupp bara vara ungefär lika. Ur biassynpunkt är propensity score-teknik ofta att föredra framför en RHG-modell. Det är dock inte självklart hur variansen för en propensity score-justerad skattning ska beräknas. En approximativ variansformel finns i Isaksson, Danielsson och Forsman (2004).

3.2 Mätfel

När data samlas in finns alltid risken att man observerar värden som avviker från de sanna värdena: det uppstår mätfel. Mätfelens storlek och riktning kan bland annat påverkas av de olika aktörer som är inblandade i datainsamlingen –i första hand intervjuare och respondenter. För att beskriva hur dessa effekter påverkar det sanna värdet formuleras en mätfelsmodell,

Observerat värde = sant värde + intervjuareffekt + slumpeffekt

där intervjuareffekten antas vara densamma för alla intervjuer genomförda av samma intervjuare, men kan variera mellan intervjuare. Det är en systematisk komponent där intervjuaren påverkar samtliga intervjuade respondenter på samma sätt. Däremot varierar slumpeffekten mellan intervjuerna även om det är samma intervjuare som har gjort intervjun. Detta är den idén som beskrivs i t.ex. Biemer och Trewin (1997).

Det finns flera orsaker till en systematiska intervjuareffekt. Till exempel om frågeformulärets instruktioner eller definitioner är oklara, de kan då tolkas på olika sätt av olika intervjuare. Varje intervjuare kan dessutom ha ett eget sätt att tolka otydliga svar på frågorna. Man brukar ibland också prata om att intervjuarnas egen-

skaper som ålder, kön, utbildning o.s.v. kan ha betydelse vid intervjun. Det finns även externa faktorer som kan förklara avvikelserna från det sanna värdet. Till exempel: blankett, intervjuutbildning, arbetsledning och genomlysning av intervjuarrutiner (se vidare Biemer och Lyberg 2003).

I vår studie tänker vi oss dessutom att mätfelen kan se olika ut för samma aktörer om observationerna gjorts efter olika många kontaktförsök:

- Under fältarbetsperioden utvecklar intervjuarna strategier för att genomföra intervjuerna. Denna inlärningsprocess kan innebära att intervjuarna mot slutet av fältarbetsperioden intervjuar på ett annat (effektivare?) sätt än i början av perioden – något som kan påverka variationen i data.
- Vissa frågor med tidsanknytning (”När besökte du senast en tandläkare?”, ”Hur många timmar arbetade du vecka 17?”) riskerar olika typer av minneseffekter att träda in (se exempelvis Christiansson, 1983; Tourangeau, Rips och Rasinski, 2000, kap. 4). Om minneseffekterna blir starkare över tiden, vilket verkar troligt, kan man förvänta sig större mätfel i svaren mot slutet av fältarbetsperioden.

Om insamlade data är behäftade med mätfel har man i estimationen inte tillgång till de sanna värdena μ_k utan bara till de observerade värdena. De senare betecknar vi y_k. Direktvägningsestimatorn i formel (6) modifieras till

$$\hat{t}_{yc\pi^*}^{(A)} = \sum_{h=1}^{H_s} \frac{n_h}{m_h^{(A)}} \sum\nolimits_{r_h^{(A)}} \breve{y}_k \tag{8}$$

där $\breve{y}_k = y_k / \pi_k$. Den separata kvotestimatorn i formel (7) modifieras på motsvarande sätt till

$$\hat{t}_{yr}^{(A)} = \sum_{h=1}^{H_s} \frac{N_h}{\hat{t}_{M_h}^{(A)}} \sum\nolimits_{r_h^{(A)}} \breve{y}_k \tag{9}$$

Vi antar att mätfelet i svarsdata består av ett intervjuareffekt och ett svarsfel. Både intervjuareffekten och svarsfelet varierar över antal kontaktförsök. Vi formulerar följande mätfelsmodell för y_k.

Mätfelsmodell

Antag att det finns *I* intervjuare, betecknade $1,...,i,...,I$, tillgängliga. För individ $k \in r_h^{(A)}$, betingat *s* och att individen svarar vid kontaktförsök *a*, gäller att

$$y_k = \mu_k + \sum_{a=1}^{A} v_{k,a} \cdot (b_{i(k),a} + \varepsilon_{k,a}) \qquad (10)$$

där

- $i(k)$ anger vilken intervjuare som genomfört intervjun med individ k
- $v_{k,a}$ är en indikatorvariabel för antalet kontaktförsök som behövs för att få svar från individ *k*:

$$v_{k,a} = \begin{cases} 1 & \text{om individ } k \text{ svarar vid } a = a_k \text{ kontaktförsök} \\ 0 & \text{annars} \end{cases}$$

  för $a = 1,..., A$.

- $b_{i(k),a}$ är en slumpmässig intervjuareffekt av intervjuare *i* om individen svarar vid kontaktförsök *a*. Effektens väntevärde är 0, dess varians $\sigma_b^2 a^{\gamma_b}$ vid intervju $a = a_k$ och kovariansen mellan två personer, $k \neq l$, som har intervjuats av samma intervjuare, *i*(*k*)=*i*(*l*), är $\sigma_b^2 (a_k a_l)^{\gamma_b/2}$. Faktorn γ_b anger om variationen ökar ($\gamma_b > 0$), minskar ($\gamma_b < 0$), eller är konstant ($\gamma_b = 0$) när antalet kontaktförsök ökar.
- $\varepsilon_{k,a}$ är ett slumpmässigt svarsfel som beror på minneseffekten om individen svarar vid kontaktförsök *a*. Felets väntevärde är 0 och dess varians $\sigma_\varepsilon^2 a^{\gamma_\varepsilon}$ om svar vid $a = a_k$. Faktorn γ_ε anger om variationen ökar, minskar eller är konstant (på samma sätt som för γ_b).

Alla slumpmässiga effekter $b_1, \ldots, b_I$ och ε_k för $k \in r_h^{(A)}$, $h = 1,..., H_s$, antas vara oberoende av varandra.

Med avseende på mätfelsmodellen, som vi betecknar *m*, ges väntevärde och varians för y_k, $k \in r_h^{(A)}$, då individ *k* svarar vid kontaktförsök $a = a_k$, som

$$E_m\left(y_k \middle| s; v_{k,a_k} = 1\right) = E_m[\mu_k + 1 \cdot (b_{i(k),a_k} + \varepsilon_{k,a_k})] = \mu_k \tag{11}$$

$$V_m\left(y_k \middle| s; v_{k,a_k} = 1\right) = \sigma_b^2 a_k^{\gamma_b} + \sigma_\varepsilon^2 a_k^{\gamma_\varepsilon} \tag{12}$$

Mätfelen för två personer är oberoende om de har olika intervjuare. Om de däremot har samma intervjuare uppstår ett beroende. Vi modellerar detta genom formeln

$$Kov_m\left(y_k, y_l \middle| s; v_{k,a_k} = 1; v_{l,a_l} = 1\right) = \begin{cases} \sigma_b^2 (a_k a_l)^{\gamma_b/2} & k \neq l;\ i(k) = i(l) \\ 0 & i(k) \neq i(l) \end{cases} \tag{13}$$

Betingat att man får svar kan nu (12) vägas samman för de olika kontaktförsöken med sannolikheten $\theta_{hs}^{(A)}$. Sannolikheten att få svar i grupp *h* efter exakt *a* kontaktförsök är $\theta_{hs}^{(a)} - \theta_{hs}^{(a-1)}$. Av detta följer att

$$V_m\left(y_k \middle| s; k \in r_h^{(A)}\right) = \frac{1}{\theta_{hs}^{(A)}} \sum\nolimits_a (\theta_{hs}^{(a)} - \theta_{hs}^{(a-1)})(\sigma_b^2 a^{\gamma_b} + \sigma_\varepsilon^2 a^{\gamma_\varepsilon}) \tag{14}$$

På samma sätt gäller att om två individer har olika intervjuare är deras svar oberoende, men om de har samma intervjuare blir kovariansen

$$\begin{aligned} &Kov_m\left(y_k, y_l \middle| s; k \in r_g^{(A)}; l \in r_h^{(A)}; i(k) = i(l)\right) \\ &\quad = \sum\nolimits_{a_k} \sum\nolimits_{a_l} \frac{(\theta_{gs}^{(a_k)} - \theta_{gs}^{(a_k-1)})}{\theta_{gs}^{(A)}} \frac{(\theta_{hs}^{(a_l)} - \theta_{hs}^{(a_l-1)})}{\theta_{hs}^{(A)}} \sigma_b^2 (a_k a_l)^{\gamma_b/2} \end{aligned} \tag{15}$$

Om vi slutligen tar hänsyn till intervjuarfördelningen får vi, då $k \neq l$, att

$$\begin{aligned} &Kov_m\left(y_k, y_l \middle| s; k \in r_g^{(A)}; l \in r_h^{(A)}\right) \\ &= P(i(k) = i(l) \middle| s; k \in r_g^{(A)}; l \in r_h^{(A)}) \times \sum\nolimits_{a_k} \sum\nolimits_{a_l} \frac{(\theta_{gs}^{(a_k)} - \theta_{gs}^{(a_k-1)})}{\theta_{gs}^{(A)}} \frac{(\theta_{hs}^{(a_l)} - \theta_{hs}^{(a_l-1)})}{\theta_{hs}^{(A)}} \sigma_b^2 (a_k a_l)^{\gamma_b/2} \end{aligned} \tag{16}$$

Anmärkning 3.4
I vår enkla mätfelsmodell finns ingen bias i mätfelen. Det enda som påverkas av antalet kontaktförsök är variansen.

Anmärkning 3.5
Mätfelsmodellen i sig är inte beroende av svarsmodellen men däremot av vid vilket kontaktförsök man får svar.

Anmärkning 3.6
Vår mätfelsmodell kan ses som en utvidgning av den enkla modell som vanligen används, se t.ex. Biemer och Trewin (1997, formel (27.3)). För konstanta intervjuar- och minneseffekter ($\gamma_b = 0$ och $\gamma_\varepsilon = 0$) är det bara indikatorvariabeln $v_{k,a}$ som skiljer vår modell från modellen i Biemer och Trewin.

4 Estimatorernas egenskaper

Vi härleder väntevärde och varians för de två estimatorerna av intresse: direktvägningsestimatorn och den separata kvotestimatorn.

4.1 Utgångspunkt

Vi betraktar data som genererade från en trestegsprocess med slumpmässighet involverad i varje steg:

1) Ett stickprov *s* dras från populationen U. Alla individer i urvalet tilldelas en intervjuare.
2) Maximalt *A* försök görs att kontakta varje individ $k \in s$. Delmängden $r^{(A)}$ innehåller alla individer som har svarat inom *A* kontaktförsök. Svaren genereras i enlighet med RHG-modellen i avsnitt 3.1.
3) De observerade värdena y_k för $k \in r^{(A)}$ innehåller mätfel i enlighet med mätfelsmodellen i avsnitt 3.2.1.

Slumpmässigheten i första steget härrör från urvalsdesignen, här betecknad *p*. Slumpmässigheten i andra steget beror på om – och i så fall när – man lyckas etablera kontakt, och om man får svar. Denna källa till slumpmässighet betecknar vi *RD* (för Response Distribution). Slumpmässigheten i tredje steget styrs av vår mätfelsmodell *m*. Den simultana fördelningen för $\hat{t}_{yc\pi^*}^{(A)}$ med avseende på hela processen kallar vi *pRDm*-fördelningen.

4.2 Direktvägningsestimatorns egenskaper

Resultat 4.1 (Väntevärde för direktvägningsestimatorn)
Antag att $m_h^{(A)} \neq 0$ med sannolikheten 1 då $n_h \neq 0$; $h = 1,...,H_s$.
Under *pRDm*-fördelningen ges väntevärdet för $\hat{t}_{yc\pi^*}^{(A)}$ i formel (8) som

$$E(\hat{t}_{yc\pi^*}^{(A)}) = t_\mu \qquad (17)$$

•

Bevis. Definitionsmässigt gäller att vänsterledet i formel (17) är ekvivalent med

$$E(\hat{t}_{yc\pi^*}^{(A)}) = E_p[E_{RD}(E_m(\hat{t}_{yc\pi^*}^{(A)}\big|s;r^{(A)})|s)] \tag{18}$$

Eftersom mätfelen saknar bias är (18) detsamma som

$$E_p\left[E_{RD}\left(\sum_{h=1}^{H_s}\frac{n_h}{m_h^{(A)}}\sum\nolimits_{r_h^{(A)}}\frac{\mu_k}{\pi_k}\,|s\right)\right] \tag{19}$$

För att förenkla beteckningarna inför vi indikatorn J_k som är 1 om individ *k* svarar (0 annars). Väntevärdet i (19) kan nu skrivas som

$$E_p\left[E_{RD}\left(\sum_{h=1}^{H_s}\frac{n_h}{m_h^{(A)}}\sum\nolimits_{s_h}J_k\frac{\mu_k}{\pi_k}\,|s\right)\right] \tag{20}$$

Vi flyttar nu in väntevärdet betingat av $m_h^{(A)}$, vilket är möjligt eftersom varken summationen n_h eller $m_h^{(A)}$ beror av *RD* givet $m_h^{(A)}$.

$$E_p\left\{\sum_{h=1}^{H_s}E_{m_h^{(A)}}\left[\frac{n_h}{m_h^{(A)}}\sum\nolimits_{s_h}E_{RD}\left(J_k\big|s,m_h^{(A)}\right)\frac{\mu_k}{\pi_k}|s\right]\right\} \tag{21}$$

Uttrycket $E_{RD}(J_k\big|s,m_h^{(A)})$ har väntevärdet $m_h^{(A)}/n_h$ eftersom antalet av de n_h indikatorerna i RHG-gruppen *h*, som är skild från noll, är $m_h^{(A)}$ och våra antaganden i RHG-modellen innefattar att bortfallet för individerna i RHG-grupperna är oberoende och utbytbara givet $m_h^{(A)}$.Vi kan alltså ersätta väntevärdet med

$$E_p\left\{\sum_{h=1}^{H_s}E_{m_h^{(A)}}\left[\frac{n_h}{m_h^{(A)}}\sum\nolimits_{s_h}\frac{m_h^{(A)}\mu_k}{n_h\pi_k}|s\right]\right\} = E_p\left[\sum\nolimits_s\frac{\mu_k}{\pi_k}\right] = \sum\nolimits_U\mu_k = t_\mu \tag{22}$$

Resultat 4.1 är därmed bevisat.

Resultat 4.2 (Varians för direktvägningsestimatorn)
Antag att $m_h^{(A)} \neq 0$ med sannolikheten 1 då $n_h \neq 0$; $h = 1,...,H_s$, att intervjuartilldelningen har skett slumpmässigt så att varje intervjuare har fått en kvot om n/I objekt att intervjua, och att intervjuartilldelningen har gjorts oberoende av indelningen i RHG-

grupper. Under *pRDm*-fördelningen ges variansen för $\hat{t}_{yc\pi^*}^{(A)}$ i formel (8) som

$$V(\hat{t}_{yc\pi^*}^{(A)}) = V_I + V_{II} + V_{III} \tag{23}$$

där

$$V_I \approx V_p(\hat{t}_{\mu s}) = \sum\sum_U \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \mu_k \mu_l \tag{24}$$

med $\hat{t}_{\mu s} = \sum_s \breve{\mu}_k$ och

$$\begin{aligned} V_{II} &= E_p[V_{RD}(\hat{t}_{\mu c\pi^*}^{(A)} | s)] \\ &\approx E_p\left[\sum_{h=1}^{H_s} \frac{1-\theta_{hs}^{(A)}}{\theta_{hs}^{(A)}} \sum_{s_h} \left(\breve{\mu}_k - \frac{\sum_{s_h} \breve{\mu}_k}{n_h}\right)^2\right] \end{aligned} \tag{25}$$

Den sista termen i formel (23), *mätfelsvariansen* för $\hat{t}_{yc\pi^*}^{(A)}$, kan skrivas som

$$\begin{aligned} V_{III} &= E_p\left\{E_{RD}[V_m(\hat{t}_{yc\pi^*}^{(A)} \,|\, s; r^{(A)}) | s]\right\} \\ &\approx E_p\left[\sum_{h=1}^{H_s}\left(V_m \sum_{s_h} \frac{1}{\theta_{hs}^{(A)} \pi_k^2} + C_m \sum_{k\neq l}\sum_{s_h} \frac{1}{\pi_k \pi_l}\right)\right. \\ &\quad \left. + \sum_{g\neq h}\sum C_m \sum_{s_g} \frac{1}{\pi_k} \sum_{s_h} \frac{1}{\pi_l}\right] \end{aligned} \tag{26}$$

där

$$V_m = V_m(y_k \,|\, s; k \in r_h^{(A)}) = \sum_{a=1}^{A} \frac{\theta_{hs}^{(a)} - \theta_{hs}^{(a-1)}}{\theta_{hs}^{(A)}} (\sigma_b^2 a^{\gamma_b} + \sigma_\varepsilon^2 a^{\gamma_\varepsilon}) \tag{27}$$

$$\begin{aligned} C_m &= Kov_m(y_k, y_l \,|\, s; k \in r_g^{(A)}; l \in r_h^{(A)}) \\ &= \frac{n-I}{I(n-1)} \sum_{a_k=1}^{A}\sum_{a_l=1}^{A} \frac{\theta_{gs}^{(a_k)} - \theta_{gs}^{(a_k-1)}}{\theta_{gs}^{(A)}} \frac{\theta_{hs}^{(a_l)} - \theta_{hs}^{(a_l-1)}}{\theta_{hs}^{(A)}} \sigma_b^2 (a_k a_l)^{\gamma_b/2} \end{aligned} \tag{28}$$

•

Bevis. Formel (24)–(25) är hämtade från Tångdahl (2005, avsnitt 5.1). Formel (26) härleds i appendix A och formel (27) och (28) är omskrivningar av mätfelsvarians och kovarians i formel (14) och (16).

Låt oss nu betrakta ett enkelt specialfall av förutsättningarna i avsnitt 4.1.

Specialfall

- Stickprovet s dras med obundet slumpmässigt urval (OSU) från U.
- Alla intervjuare tilldelas lika många (n/I) utvalda individer som de ska försöka kontakta och intervjua.
- Alla individer i stickprovet tillhör samma svarshomogenitetsgrupp.
- Kontakt- och svarssannolikheterna antas vara oberoende av stickprovet. Vi antar alltså att $\omega_{k|s}^{(A)} = \omega^{(A)}$ och $\theta_{k|s}^{(A)} = \theta^{(A)}$ för alla $k \in s$.

I detta specialfall förenklas direktvägningsestimatorn $\hat{t}_{yc\pi*}^{(A)}$ till *expansionsestimatorn* (på engelska (straight) expansion estimator). Dess egenskaper ges av resultat 4.3.

Resultat 4.3 (Väntevärde och varians för expansionsestimatorn) Antag att $m_h^{(A)} \neq 0$ med sannolikheten 1 då $n_h \neq 0$; $h = 1,..., H_s$. Under *pRDm*-fördelningen och specialfallet ovan är $\hat{t}_{yc\pi*}^{(A)}$ approximativt väntevärdesriktig för t_μ. Ett approximativt uttryck för variansen för $\hat{t}_{yc\pi*}^{(A)}$ ges som

$$V_{\text{OSU}}(\hat{t}_{yc\pi*}^{(A)}) \approx N^2 \frac{S_{\mu U}^2}{n\theta^{(A)}} + N^2\left(\frac{V_m}{n\theta^{(A)}} + C_m\right) \qquad (29)$$

•

Resultat 4.3, som följer av resultat 4.1 och 4.2, härleds i appendix C.

4.3 Den separata kvotestimatorns egenskaper

Om hjälpinformationen som används till svarshomogenitetsgrupperingen är känd för hela U rekommenderade vi i kapitel 3 den separata kvotestimatorn $\hat{t}_{yr}^{(A)}$ i formel (9). Dess egenskaper utreds i resultat 4.4.

Resultat 4.4 (Väntevärde och varians för den separata kvotestimatorn)
Antag att $m_h^{(A)} \neq 0$ med sannolikheten 1 då $n_h \neq 0$; $h = 1,...,H_s$. Under *pRDm*-fördelningen är $\hat{t}_{yr}^{(A)}$ då en approximativt väntevärdesriktig estimator av t_μ. Variansen för $\hat{t}_{yr}^{(A)}$ ges approximativt som

$$
\begin{aligned}
V(\hat{t}_{yr}^{(A)}) &= V_I + V_{II} + V_{III} \\
&\approx \sum\sum_U \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} E_k E_l \\
&+ E_p \left\{ \sum_{h=1}^{H_s} \left(\frac{N_h}{\hat{t}_{N_h}} \right)^2 \frac{(1-\theta_{hs}^{(A)})}{\theta_{hs}^{(A)}} \sum_{s_h} \frac{(\mu_k - \hat{t}_{\mu r^*})^2}{\pi_k^2} \right\} \qquad (30) \\
&+ E_p \left\{ \sum_{h=1}^{H_s} \left(\frac{N_h}{\hat{t}_{N_h}} \right)^2 \left(\sum_{s_h} \frac{V_m}{\theta_{hs}^{(A)} \pi_k^2} + \sum\sum_{k \neq l} \sum_{s_h} \frac{C_m}{\pi_k \pi_l} \right) + \sum\sum_{g \neq h} N_g N_h C_m \right\}
\end{aligned}
$$

•

Variansen i formel (30) härleds i appendix B.

Anmärkning 4.1
I praktiken är det sällan uppfyllt att alla $m_h^{(A)} \neq 0$ med sannolikheten 1 då $n_h \neq 0$. Resultat 4.1-4.4 gäller ändå approximativt om $m_h^{(A)} \neq 0$ med stor sannolikhet då $n_h \neq 0$; $h = 1,...,H_s$, och man substituerar värden (exempelvis 0) för de okända μ_k.

5 En kostnadsmodell

Man kan formulera många tänkbara modeller för en undersöknings kostnader beroende på vad modellen ska användas till. En detaljerad kostnadsmodell för en centraliserad telefonintervjuundersökning finns exempelvis i Groves (1989, kap. 11), men Groves modell är utformad för andra syften än våra. Vi formulerar här en mycket enklare modell, liknande den i Lindström (1991, kap. 6) och Thorburn (2004), där fokus ligger på antalet kontaktförsök.

Vi identifierar följande kostnader förknippade med datainsamlingen i en telefonintervjuundersökning. Dels uppstår en overheadkostnad, C_0, som är oberoende av urvalsstorleken. Härutöver uppstår ett antal marginalkostnader: för ytterligare en individ, för ytterligare ett kontaktförsök och för ytterligare en intervju. För $k \in s_h, h = 1,..., H_s$, låt

$C_{\text{start},h}$ kostnad fram till första kontaktförsöket (här ingår bl.a. spårning och introduktionsbrev)

$C^{(a)}_{\text{kontakt},h}$ kostnad för ytterligare ett kontaktförsök om *a*-1 kontaktförsök redan gjorts

$C^{(a)}_{\text{intervju},h}$ kostnad för att genomföra en intervju (här ingår kostnader inte bara för själva intervjun, utan även för efterarbete, mikrogranskning, editering etc). Kostnaden antas oberoende av antalet misslyckade kontaktförsök

För enkelhetens skull antar vi att $C_{\text{start},h}$, $C^{(a)}_{\text{kontakt,h}}$ och $C^{(a)}_{\text{intervju},h}$ är lika för alla svarshomogenitetsgrupper och utelämnar index *h*. Vi antar även att både kontaktkostnaden och kostnaden för intervjuer är lika över kontakter, utelämnar (*a*) och sätter t.ex. $C^{(a)}_{\text{kontakt}} \equiv C_{\text{kontakt}}$. I en verklig situation kan speciellt intervjukostnaden variera med *a*, eftersom inställningen till att delta i surveyer och de svarandes egenskaper förmodligen varierar med *a*. Intervjuarna blir förmodligen också mera vana med tiden, och kostnaden kan av det skälet väntas minska med antalet intervjuer en intervjuare hinner göra. Thorburn (2004) låter exempelvis kostnaden variera med hur många kontaktförsök man behöver innan man lyckas genomföra en intervju.

När vi formulerar ett uttryck för undersökningens kostnadsfunktion resonerar vi på följande sätt:

(i) Varje individ $k \in s$ kontaktas minst en gång. Undersökningen har alltså en total startkostnad om $C_0 + C_{\text{start}} n$.

(ii) Det första kontaktförsöket genererar kostnaden $C_{\text{kontakt}} n$; det andra försöket genererar kostnaden $C_{\text{kontakt}}(n - n^{(1)})$; det tredje försöket kostnaden $C_{\text{kontakt}}(n - n^{(2)})$ osv. Låt $n^{(0)} = 0$. Den samlade kostnaden för kontakt ges nu som $C_{\text{kontakt}} \sum_{a=1}^{A} (n - n^{(a-1)})$.

(iii) Varje individ som intervjuas (varje $k \in r^{(A)}$) genererar intervjukostnaden C_{intervju}. Undersökningens samlade intervjukostnad är således $m^{(A)} C_{\text{intervju}}$.

Av punkt (i)-(iii) följer att undersökningens kostnadsfunktion ges som

$$C = C_0 + nC_{\text{start}} + C_{\text{kontakt}} \cdot \sum_{a=1}^{A} (n - n^{(a-1)}) + m^{(A)} C_{\text{intervju}} \tag{31}$$

Kostnaderna förknippade med urvalsstorleken, *n*, framträder tydligare i följande enkla omskrivning av formel (31):

$$C^* = C - C_0 = nC_{\text{start}} + C_{\text{kontakt}} \sum_{a=1}^{A} (n - n^{(a-1)}) + m^{(A)} C_{\text{intervju}} \tag{32}$$

Vid bestämning av optimalt antal kontaktförsök använder vi oss av väntevärdet av *C**:

$$E(C^*) = nC_{\text{start}} + C_{\text{kontakt}} \sum_{a=1}^{A} [n - E(n^{(a-1)})] + E(m^{(A)}) C_{\text{intervju}} \tag{33}$$

Formel (33) är dock inte användbar så länge den saknar explicita uttryck för väntevärdena för $m^{(A)}$ och $n^{(a-1)}$. Väntevärdet för $m^{(A)}$ ges som

$$E(m^{(A)}) = E_p[\sum_{h=1}^{H_s} E_{RD}(m_h^{(A)}|s)] = E_p[\sum_{h=1}^{H_s} n_h \theta_{hs}^{(A)}] \tag{34}$$

Vi antog i avsnitt 3.1 att RHG-grupperna är homogena även med avseende på kontaktsannolikheterna $\omega_{k|s}^{(a)} = \omega_{hs}^{(a)}$. Härav följer att

$$E(n^{(a)}) = E_p[\sum_{h=1}^{H_s} E_{RD}(n_h^{(a)}|s)] = E_p[\sum_{h=1}^{H_s} n_h \omega_{hs}^{(a)}] \tag{35}$$

Notera att för $a = 0$ gäller att $E(n^{(0)}) = 0$ och att $\omega_{hs}^{(0)} = 0$.

Insättning av väntevärdena (34) och (35) i formel (33) ger

$$E(C^*) = nC_{\text{start}} + C_{\text{kontakt}} \sum_{a=1}^{A} [n - E_p(\sum_{h=1}^{H_s} n_h \omega_{hs}^{(a-1)})] + C_{\text{intervju}} [E_p(\sum_{h=1}^{H_s} n_h \theta_{hs}^{(A)})] \qquad (36)$$

6 Optimalt antal kontaktförsök

6.1 Generell ansats

I en undersökning vill man använda direktvägningsestimatorn $\hat{t}_{yc\pi^*}^{(A)}$ av t_μ. I planeringsskedet av undersökningen vill man fatta beslut om hur många kontaktförsök *A* man högst ska göra. Beslutet påverkar både estimatorns varians och den förväntade kostnaden för undersökningen eftersom båda är funktioner av *A*. Av olika möjliga antal kontaktförsök definierar vi det optimala antalet som det som ger den lägsta ”kostnaden” för undersökningen – mätt antingen i monetära enheter eller i precision.

Resultat 6.1 (Optimalt antal kontaktförsök)
I en urvalsundersökning skattas totalen t_μ med direktvägningsestimatorn $\hat{t}_{yc\pi^*}^{(A)}$. Den förväntade kostnaden för undersökningen, $C^{(A)}$, ges i formel (36). Låt A_1, A_2, … vara en uppsättning möjliga antal kontaktförsök. Av dessa är det optimala antalet kontaktförsök det *A* som ger det lägsta värdet på produkten

$$OPT(\hat{t}_{yc\pi^*}^{(A)}) = V(\hat{t}_{yc\pi^*}^{(A)})C^{(A)} \qquad (37)$$

där variansen för $\hat{t}_{yc\pi^*}^{(A)}$ med avseende på *pRDm*-fördelningen, $V(\hat{t}_{yc\pi^*}^{(A)})$, ges i Resultat 4.2.

•

Resultat 6.1 vilar på följande resonemang. Antag att vi vill jämföra två olika val av *A*: A_1 och A_2, $A_1<A_2$. Om man väljer A_1 blir variansen V_1 och undersökningens förväntade kostnad C_1; om man väljer A_2 blir variansen V_2 och undersökningens förväntade kostnad $C_2> C_1$. Om vi adderar C_2/C_1 kontaktförsök till A_1 blir undersökningskostnaden C_2 och variansen V_1C_1/C_2. Den variansen är mindre än V_2 om $V_1C_1< V_2C_2$. I så fall är alltså A_1 ett bättre (billigare) val än A_2 om man översätter till samma kostnad. På samma sätt kan man visa att A_1 blir billigare än A_2 om man översätter till samma varians. På det här sättet kan vi alltså jämföra olika möjliga antal kontaktförsök

genom att multiplicera de beräknade varianserna med de förväntade kostnaderna och titta på produkterna.

Anmärkning 6.1
Resultat 6.1 kan även användas för att bestämma det antal kontaktförsök A som minimerar $V(\hat{t}_{yc\pi*}^{(A)})$ under en given förväntad kostnad $C^{(A)}$, alternativt att bestämma det A som minimerar $C^{(A)}$ under en given varians $V(\hat{t}_{yc\pi*}^{(A)})$.

Anmärkning 6.2
Man kan på motsvarande sätt optimera antalet kontaktförsök för estimatorer som inte är väntevärdesriktiga (exempelvis olika estimatorer av kvottyp). Det kommer då att visa sig att ju större stickprovet är, desto större betydelse får biasen vid valet av optimalt antal kontaktförsök.

Anmärkning 6.3
När man optimerar antalet konktaktförsök för estimatorer som inte är väntevärdesriktiga bör målet aldrig vara att enbart minimera biasen. För att minimera biasen ska man välja en enda observation och sedan satsa så mycket resurser på denna individ att man får ett svar.

Anmärkning 6.4
För optimal allokering i några andra situationer än den vi behandlar här, se Wolter och Pyne (1978), Biemer (1983) och Lepkowski och Groves (1986).

6.2 Specialfallet (fortsättning)

Om vi gör samma antagande för kontakt som för svar (rak uppräkning) och att $\omega_{k|s}^{(a)} = \omega^{(a)}$, med $\omega^{(0)} = 0$, för alla $k \in s$ har vi att den förväntade kostnaden (37) kan skrivas som

$$\begin{aligned} C^{(A)} &= nC_{\text{start}} + C_{\text{kontakt}}\{\sum_{a=1}^{A}[n - n\omega^{(a-1)}]\} + C_{\text{intervju}} n\theta^{(A)} \\ &= n\left[C_{\text{start}} + C_{\text{kontakt}}\{\sum_{a=1}^{A}[1 - \omega^{(a-1)}]\} + C_{\text{intervju}}\theta^{(A)} \right] \end{aligned} \qquad (38)$$

Det optimala antalet kontaktförsök är alltså enligt formel (37) det A som minimerar

$$\begin{aligned} OPT_{OSU}(\hat{t}^{(A)}_{yc\pi^*}) &= V_{OSU}(\hat{t}^{(A)}_{yc\pi^*})E(C^*) \\ &= N^2 \frac{S^2_{\mu I}}{n\theta^{(A)}} + N^2\left(\frac{V_m}{n\theta^{(A)}} + nC_m\right) \\ &\times \left[C_{\text{start}} + C_{\text{kontakt}}\{\sum_{a=1}^{A}[1-\omega^{(a-1)}]\} + C_{\text{intervju}}\theta^{(A)} \right] \end{aligned} \tag{39}$$

7 Bestämning av antal kontaktförsök – ett exempel

Låt oss nu titta på hur man kan bestämma optimalt antal kontaktförsök i praktiken. Vi utgår ifrån det specialfall som beskrivs i avsnitt 4.2. och 6.2

7.1 Förutsättningar

Parametrarna i detta exempel är valda så att de ska vara så realistiska som möjligt och kunna förekomma i en typisk SCB-undersökning. Vi antar följande:

- Populationens storlek är N = 7 000 000.
- Urvalet görs med OSU med stickprovsstorleken n = 2 000.
- Stickprovet bildar en enda RHG-grupp.
- Populationsvariansen, $S^2_{\mu I}$, antar värdet 0,5 eller 1. Dessa värden är valda utifrån bedömningar av undersökningens intraintervjuarkorrelation – se resonemang nedan.
- Mätfelsmodellens parametrar är:
  - $\sigma_b = \sigma_\varepsilon = 0{,}05;\ 0{,}15$ (vilket motsvarar ett ”normalt” och ett ”litet högre än normalt” värde)
  - $\gamma_b = -1;\ 0;\ 1$ (vilket motsvarar att intervjuareffekten minskar, är konstant eller ökar över tiden)
  - $\gamma_\varepsilon = 0$ (vilket motsvarar att vi inte förväntar oss några minneseffekter)
  - $I = 100$
- Kostnaderna är C_0=150 000 kr, C_{start}=25 kr, C_{kontakt}=20 kr och C_{intervju}=100 kr. Siffrorna bygger på uppskattningar från personer som planerar undersökningar vid SCB
- Parametrarna $\theta^{(a)}, \omega^{(a)}$ och a har uppskattats från observerat material (tabell 7.1). De baseras på data från det så kallade WinDATI-provet 1996 på SCB:s Partisympatiundersökning (se vidare Japec och Lundquist 2000, avsnitt 3).

Tabell 7.1 Uppskattade värden på $\theta^{(a)}$, $\omega^{(a)}$, $m^{(a)}$ och a

Kontaktförsök a	$\theta^{(a)}$	$\omega^{(a)}$	$m^{(a)}$	$m^{(a)} - m^{(a-1)}$
1	0,280	0,370	560	560
2	0,480	0,620	960	400
3	0,600	0,765	1200	240
4	0,660	0,840	1320	120
5	0,700	0,890	1400	80
6	0,720	0,915	1440	40
7	0,735	0,935	1470	30
8	0,745	0,950	1490	20
9	0,755	0,962	1510	20
10	0,765	0,975	1530	20
11	0,770	0,982	1540	10
12	0,775	0,988	1550	10

Våra val av populationsvarianser, $S^2_{\mu I} =$ 0,5 och 1, är kopplade till en tänkt intervjuar- och minneseffekt genom *intraintervjuarkorrelationskoefficienten*

$$\rho_y = \frac{\sigma_b^2}{S^2_{\mu I} + \sigma_b^2 + \sigma_\varepsilon^2}$$

(Biemer och Trewin, 1997, formel (27.16)). Internationella intervjuarvariansstudier visar låga värden på ρ_y (se t.ex. Groves 1989). Några svenska studier finns tyvärr inte att tillgå. Insättning av de värden vi antar på σ_b, σ_ε och $S^2_{\mu I}$ ger värden på ρ_y enligt tabell 7.2.

Tabell 7.2 Intra-intervjuarkorrelationer ρ_y

		$\sigma_b = \sigma_\varepsilon$	
		0,05	0,15
$S^2_{\mu I}$	0,5	0,005	0,020
	1	0,002	0,040

Variansen i formel (29) skattar vi med

$$\hat{V}_{OSU}(\hat{t}_{yc\pi^*}^{(A)}) = N^2 \frac{S_{\mu U}^2}{m^{(A)}} + \left(\frac{N}{m^{(A)}}\right)^2 \left\{ \sum_a (m^{(a)} - m^{(a-1)})(\sigma_b^2 a^{\gamma_b} + \sigma_\varepsilon^2 a^{\gamma_\varepsilon}) + \frac{\sigma_b^2 (n-I)}{I(n-1)} \left[\sum_a (m^{(a)} - m^{(a-1)}) a^{\gamma_b/2} \right]^2 \right\} \tag{40}$$

I figur 7.1 och 7.2 redovisas skattningar av standardfelen $[\hat{V}_{OSU}(\hat{t}_{yc\pi^*}^{(A)})]^{1/2}$. På motsvarande sätt skattas formel (39) med

$$O\hat{P}T_{OSU}(\hat{t}_{yc\pi^*}^{(A)}) = \hat{V}_{OSU}(\hat{t}_{yc\pi^*}^{(A)}) \times n \left[C_{\text{start}} + C_{\text{kontakt}} \{ \sum_{a=1}^{A} [1 - \omega^{(a-1)}] \} + C_{\text{intervju}} \theta^{(A)} \right] \tag{41}$$

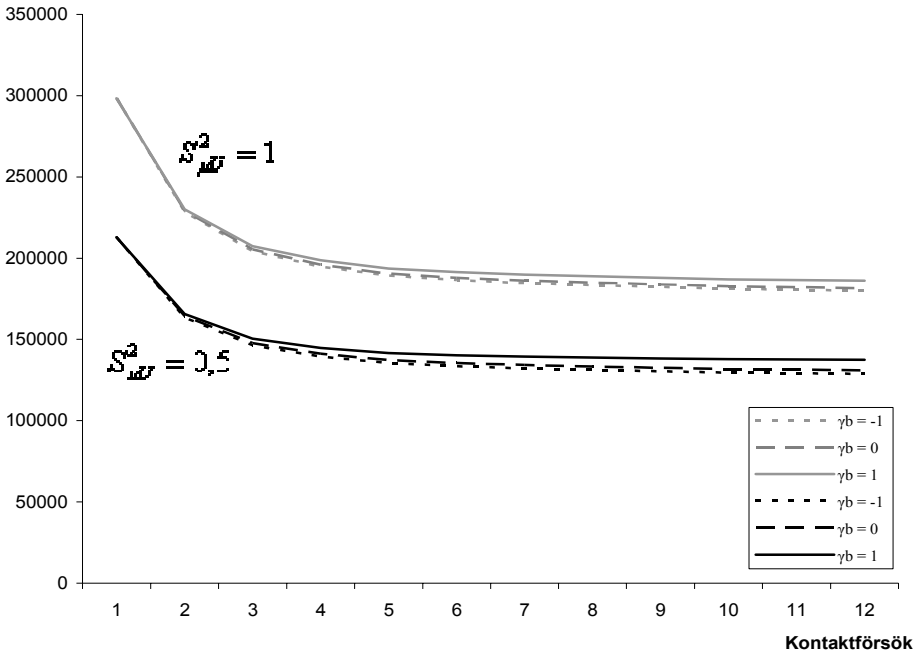
Skattningarna av formel (40) och (41) fås ur det observerade materialet i tabell 7.1.

7.2 Analys av standardfel

Om intervjuarvariansen är låg (figur 7.1) blir skillnaderna i mätfel små oavsett om intervjuareffekten minskar, är konstant eller ökar. De minsta standardfelen erhålles för 12 kontaktförsök. Populationsvariansen har ett stort genomslag i förhållande till mätfelet. Om däremot intervjuareffekten (figur 7.2) är stor och ökar med antalet kontaktförsök ($\gamma_b = 1$) så påverkas även standardfelen kraftigt. En slutsats är att färre kontaktförsök bör göras om intervjuareffekten ökar med antalet försök.

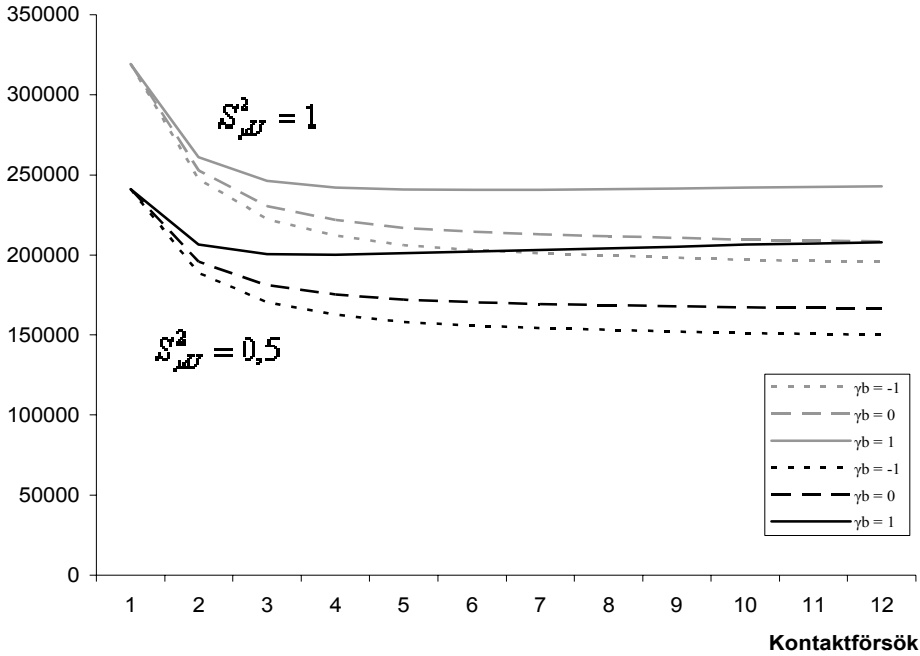
Figur 7.1 Skattning av standardfelet för $\hat{t}_{yc\pi^*}^{(A)}$ då $\sigma_b = \sigma_\varepsilon = 0{,}05$

$[\hat{V}_{OSU}(\hat{t}_{yc\pi^*}^{(A)})]^{1/2}$



Figur 7.2 Skattning av standardfelet för $\hat{t}_{yc\pi^*}^{(A)}$ då $\sigma_b = \sigma_\varepsilon = 0{,}15$

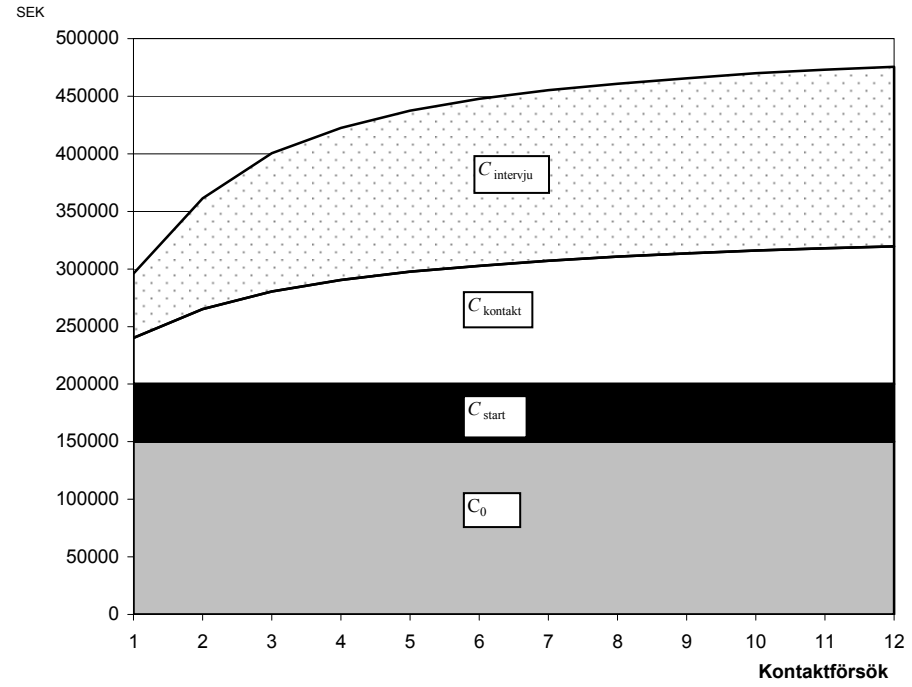
$[\hat{V}_{OSU}(\hat{t}_{yc\pi^*}^{(A)})]^{1/2}$



7.3 Analys av kostnader

Figur 7.3 beskriver hur kostnadsfunktion ser ut för en undersökning med de valda parametrarna från avsnitt 7.1.

Figur 7.3 Ackumulerade kostnader för antalet kontaktförsök



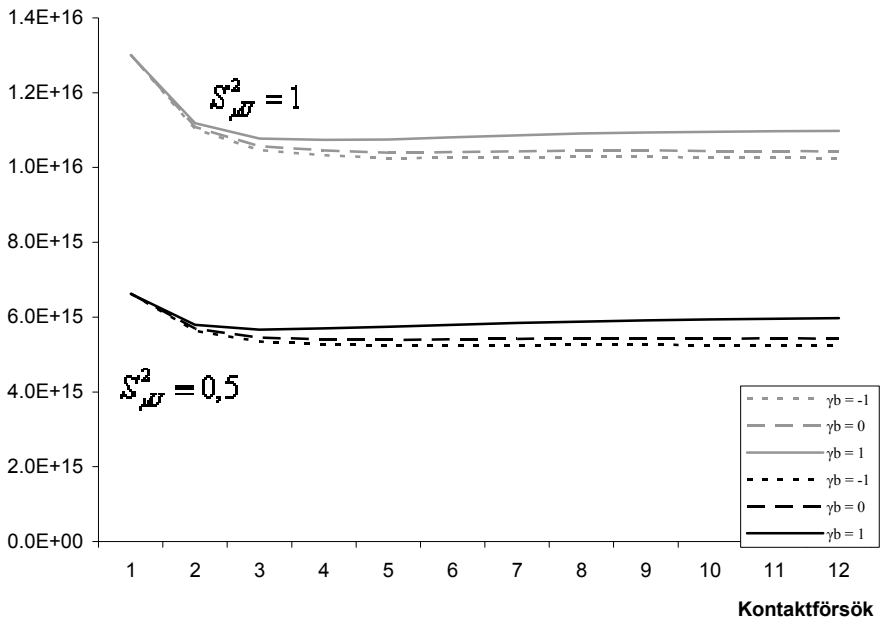
7.4 Avvägning mellan standardfel och kostnader

Formel (41) ger ett uttryck för $\hat{OPT}_{OSU}(\hat{t}^{(A)}_{yc\pi^*})$. Om standardfelen och kostnadsfunktionen från tidigare avsnitt sätts in får man figurerna 7.4 och 7.5, där optimalt A ges av kurvornas lägsta punkt. Resultaten indikerar att man inte behöver göra 12 kontaktförsök för de flesta av de valda kombinationerna. Om det finns en intervjuareffekt i variabeln påverkas $\hat{OPT}_{OSU}(\hat{t}^{(A)}_{yc\pi^*})$ kraftigare.

I mer än hälften av fallen ger kurvorna olika optimum. Om intervjuareffekten ökar över tiden bör man inte göra fler än fyra kontaktförsök; om effekten är konstant bör man inte göra fler än fem kontaktförsök, och om intervjuareffekten avtar över tiden ges optimum vid

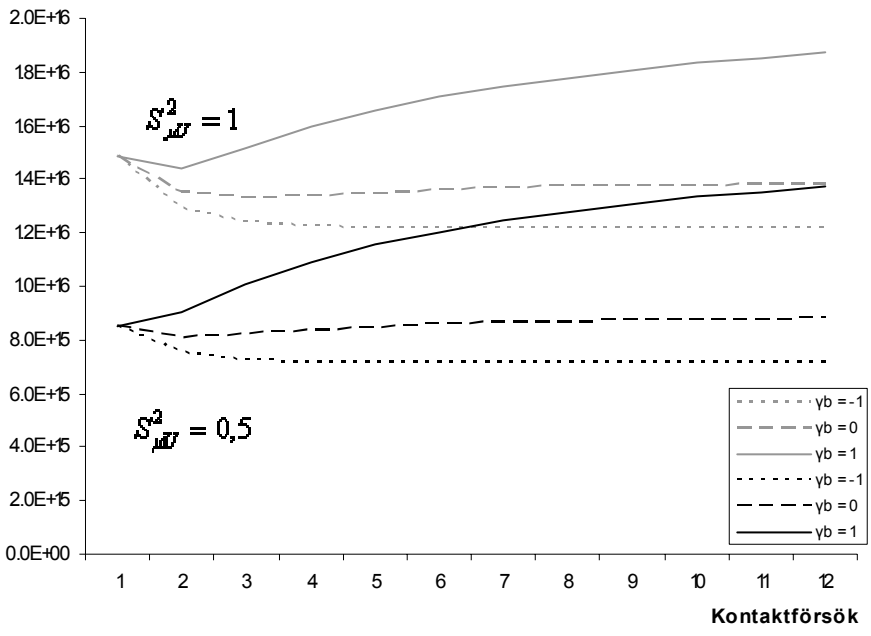
12 kontaktförsök (men minimat är flackt). Förutom i de fall då intervjuarpåverkan är stark eller växande med antalet kontaktförsök är effektivitetsförlusten alternativt vinsten vid 12 kontaktförsök försumbar. I så fall bör antalet kontaktförsök väljas av andra skäl, t.ex. för att korta produktionstiden.

Figur 7.4 Optimalt antal kontaktförsök då $\sigma_b = \sigma_\varepsilon = 0{,}05$



Figur 7.5 Optimalt antal kontaktförsök då $\sigma_b = \sigma_\varepsilon = 0{,}15$

$O\hat{P}T(\hat{t}^{(A)}_{yc\pi^*})$



8 Slutsatser och diskussion

Vi har visat att det är möjligt att teoretiskt beräkna en optimal övre gräns för antalet kontaktförsök fram till dess att den utvalda intervjupersonen ska betraktas som oanträffad. I vårt exempel är det optimala antalet försök lågt (mellan tre och fem försök) eftersom intervjuareffekten var stor och mätfelet störst för de individer som var svåra att få kontakt med. Detta skulle kunna inträffa i praktiken exempelvis om tidspressen på intervjuarna ökar med tiden, om minnesfelen ökar p.g.a. att referenstidpunkten förflyttas längre bak i tiden, eller om de som är svåra att få tag på också är mera negativa till att delta i undersökningar. Vi har dock inte jämfört med alternativen att öka antalet intervjuare eller välja en annan estimator än direktvägningsestimatorn. Vårt resultat gäller formellt bara för det enkla exempel som vi har tittat på. Vi tror dock att det är mera allmängiltigt än så. Om intervjuareffekten är försumbar och svarsfelen inte ökar med antalet misslyckade kontaktförsök finns ingen övre gräns för antalet kontaktförsök. I så fall kommer tidsaspekten att vara avgörande för när insamlingen ska brytas. För att hitta optimalt antal kontaktförsök i det fallet måste man bedöma kostnaden för minskad aktualitet, och jämföra den med kostnaden för att uppnå samma precision på kortare tid men med ett större urval.

Vi konstaterar att det går inte att bestämma den optimala brytpunkten utan att ha en bra kostnadsmodell för datainsamlingen. Det räcker inte att konstatera att data insamlade i slutet av insamlingsperioden har liten inverkan på resultatet. Det är en naturlig konsekvens av stora talens lag.

Vårt arbete har många tänkbara utvecklingsmöjligheter. De teoretiska härledningarna kan utvidgas till att omfatta andra tänkbara estimatorer – såsom olika typer av regressionsestimatorer – och till andra parametrar än totaler. Framförallt är det dock angeläget att studera fler exempel på användning av metoden. Exemplet i avsnitt 7 utgår ifrån att man bara har en enda svarshomogenitetsgrupp, som antas överensstämma med den sanna svarsfördelningen. Det vore värdefullt att se hur metoden fungerar vid flera grupper. En annan utvidgning vore att tillåta bortfallsbias. Det kan göras som i Thorburn (2004), som dock bara behandlar fallet OSU. Man kan också tänka sig att bygga in tidhållningen i kostnadsmodellen. Ett litet antal kontaktförsök innebär en tidigare publicering

och mera aktuell statistik. En fjärde utvecklingsmöjlighet är att låta antalet kontaktförsök variera mellan olika individer med olika egenskaper. Man kan också låta antalet kontaktförsök styras av hur insamlingsprocessen förlöper. Om man fått oväntat många svar från en grupp, och oväntat få från en annan, kan man kanske sänka gränsen för den förstnämnda gruppen och höja den för den sistnämnda.

Vidare bör metoden provas på en riktig undersökning. Detta kräver sannolikt en viss anpassning av modellerna – inte minst kostnadsmodellen – till den verkliga undersökningssituationen. Det fordrar också uppskattningar av olika modellparametrar. Den framtagna modellen visar rätt väl vilka grunduppgifter som behöver samlas in för att en optimering ska kunna genomföras. Det är alltså viktigt att utforma processövervakningssystem som stöder framtagning av dessa uppgifter. Vi ser det också som angeläget att kombinera detta arbete med studier av uppringningsalgoritmer där både tidpunkt för kontakt och antal kontaktförsök styrs av tillgänglig hjälpinformation.

9 Tack

Vi vill tacka Martin Axelson, som granskat en tidigare version av rapporten, för värdefulla synpunkter. Alla kvarvarande fel och brister är helt och hållet vårt eget ansvar.

Referenser

Biemer, P. (1983). Optimal dual frame design: Results of a simulation study. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 630–635.

Biemer, P. P. och Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. I L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz och D. Trewin (red.), *Survey Measurement and Process Quality*. New York: Wiley.

Biemer, P.P. och Lyberg, L.E. (2003). *Introduction to Survey Quality.* New York: Wiley.

Christiansson, A. (1983). Balancing Recall Bias against Sampling Errors in the Swedish TV Audience Surveys. *Statistical Review* 1983:5.

Groves, R. M. (1989). *Survey Errors and Survey Costs.* New York: Wiley.

Isaksson, A., Danielsson, S. och Forsman, G. (2004). On the variability of estimates based on propensity score weighted data from web panels. *2004 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association*, 3689–3696.

Japec, L. (2005). Quality Issues in Interview Surveys – Some Contributions. *PhD Thesis, Stockholm University.*

Japec, L. och Lundquist, P. (2000). Bortfallet – påverkas det av intervjuarnas attityder och strategier? *Statistiska centralbyrån.*

Lepkowski, J. M. och Groves, R. M. (1986). A mean squared error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930–937.

Lindström, H. L. (1991). Interacting Nonresponse and Response Errors. *R & D Report 1991:3, Statistics Sweden.*

Little, R. J. A. och Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* 2nd ed. New York: Wiley.

Rosenbaum och Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. B. (1983). Conceptual issues in the presence of nonresponse. I W. G. Madow, I. Olkin och D. B. Rubin (red.), *Incomplete Data in Sample Surveys,* vol. 2. New York: Academic Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Särndal, C.-E. och Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: Wiley.

Särndal, C.-E., Swensson, B. och Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Thorburn, D. (2004). Officiell statistik. *Kurskompendium, Department of Statistics, Stockholm University.*

Tourangeau, R., Rips, L. J. och Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tångdahl, S. (2004). Nonresponse bias for some common estimators and its change over time in the data collection process. *Working paper no. 13, ESI, Örebro University.*

Tångdahl, S. (2005). The variance of some common estimators and its components under nonresponse. *Working paper no. 9, ESI, Örebro University.*

Tångdahl, S. (2006). On the evaluation of the cost efficiency of nonresponse rate reduction efforts – some general considerations. *Working paper no. 5, ESI, Örebro University.*

Weeks, M. F., Kulka, F. A. och Pierson, S. A. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, 51, 540–549.

Wolter, K. M. och Pyne, D. A. (1978). Optimum sample allocation to different modes of enumeration. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 414–419.

Bilagor

A Härledning av variansen i resultat 4.2

Genom att skriva om variansuttrycken är det möjligt att utnyttja resultaten i Tångdahl (2005) tillsammans med vår mätfelsmodell. Vi har antagit att RGH-modellen gäller och får en mer förenklad varians

$$V(\hat{t}_{yc\pi*}^{(A)}) = V_p\{E_{RD}[E_m(\hat{t}_{yc\pi*}^{(A)}|s;r^{(A)})|s]\} + E_p[V_{RD}(\hat{t}_{\mu c\pi*}^{(A)}|s)] + E_p\{E_{RD}[V_m(\hat{t}_{yc\pi*}^{(A)}|s;r^{(A)})|s]\}$$
$$= V_I + V_{II} + V_{III}$$

där V_I står för *samplingvariansen*, V_{II} är *den betingade bortfallsvariansen* och *mätfelsvariansen* betecknas med V_{III}.

I härledningen av samplingvariansen och den betingade bortfallsvariansen använder vi att väntevärdet för $\hat{t}_{yc\pi*}^{(A)}$ under *m* och *RD*-fördelningen approximativt kan skrivas som

$$E_{RD}[E_m(\hat{t}_{yc\pi*}^{(A)}|s;r^{(A)})|s] = E_{RD}[\hat{t}_{\mu c\pi*}^{(A)}|s] \approx \hat{t}_{\mu s}$$

där $\hat{t}_{\mu s} = \sum_s \breve{\mu}_k$.

Samplingvariansen kan nu skrivas som

$$V_I \approx V_p(\hat{t}_{\mu s}) = \sum\sum_U \frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}\mu_k\mu_l$$

För att härleda V_{II} noterar vi dels att mätfelen saknar bias, vilket ger oss $E_m(\hat{t}_{yc\pi*}^{(A)}|s;r^{(A)}) = \hat{t}_{\mu c\pi*}^{(A)}$, dels att direktvägningsestimatorn $\hat{t}_{\mu c\pi*}^{(A)}$ alternativt kan framställas som

$$\hat{t}_{\mu c\pi*}^{(A)} = \sum_{h=1}^{H_s}\frac{n_h}{m_h^{(A)}}\sum_{r_h^{(A)}}\breve{\mu}_k = \sum_{h=1}^{H_s} n_h\frac{\sum_{r_h^{(A)}}\breve{\mu}_k}{\sum_{r_h^{(A)}}1} = \sum_{h=1}^{H_s} n_h\frac{\sum_{s_h}R_{k|s}^{(A)}\breve{\mu}_k}{\sum_{s_h}R_{k|s}^{(A)}} = \sum_{h=1}^{H_s} n_h\frac{\hat{t}_{\mu r_h}^{(A)}}{\hat{t}_{\theta r_h}^{(A)}}$$

I estimatorn är $R_{k|s}^{(A)}$ en indikatorvariabel, definierad som

$$R_{k|s}^{(A)} = \begin{cases} 1 & \text{om } k \in r^{(A)}|s \\ 0 & \text{f.ö.} \end{cases}$$

med väntevärde $\theta_{hs}^{(A)}$ för $k \in s_h$ och alla *A*.

Första ordningens Taylorapproximation av $\hat{t}_{\mu c\pi^*}^{(A)}$ kan skrivas

$$\hat{t}_{\mu c\pi^*}^{(A)} \approx \sum_{h=1}^{H_s} n_h \frac{\sum_{s_h} \theta_{hs}^{(A)} \breve{\mu}_k}{\sum_{s_h} \theta_{hs}^{(A)}} + \sum_{h=1}^{H_s} \frac{n_h}{\sum_{s_h} \theta_{hs}^{(A)}} \sum_{s_h} R_{k|s}^{(A)} \left[\breve{\mu}_k - \frac{\sum_{s_h} \theta_{hs}^{(A)} \breve{\mu}_k}{\sum_{s_h} \theta_{hs}^{(A)}} \right]$$

$$= \sum_s \breve{\mu}_k + \sum_{h=1}^{H_s} \frac{1}{\theta_{hs}^{(A)}} \sum_{s_h} R_{k|s}^{(A)} \left[\breve{\mu}_k - \frac{1}{n_h} \sum_{s_h} \breve{\mu}_k \right]$$

I härledningen av V_{II} används följande Taylorapproximation:

$$V_{II} = E_p[V_{RD}(\hat{t}_{\mu c\pi^*}^{(A)}|s)] \approx E_p \left\{ V_{RD} \left[\sum_s \breve{\mu}_k + \sum_{h=1}^{H_s} \frac{1}{\theta_{hs}^{(A)}} \sum_{s_h} R_{k|s}^{(A)} \left(\breve{\mu}_k - \frac{1}{n_h} \sum_{s_h} \breve{\mu}_k \right) | s \right] \right\}$$

$$= E_p \left\{ \sum_{h=1}^{H_s} \left(\frac{1}{\theta_{hs}^{(A)}} \right)^2 \sum_{s_h} V_{RD}(R_{k|s}^{(A)}|s) \left(\breve{\mu}_k - \frac{1}{n_h} \sum_{s_h} \breve{\mu}_k \right)^2 \right\}$$

$$= E_p \left\{ \sum_{h=1}^{H_s} \frac{(1-\theta_{hs}^{(A)})}{\theta_{hs}^{(A)}} \sum_{s_h} \left(\breve{\mu}_k - \frac{1}{n_h} \sum_{s_h} \breve{\mu}_k \right)^2 \right\}$$

där $V_{RD}(R_{k|s}^{(A)}|s) = \theta_{hs}^{(A)}(1-\theta_{hs}^{(A)})$.

För mätfelsvariansen konstaterar vi att mätfelsmodellen i formel (10) tar hänsyn till vilken intervjuare som genomfört intervjun. Om intervjuarna fördelas slumpmässigt över de *n* dragna objekten så att varje intervjuare får $q = n/I$ objekt (vi antar att *q* är ett heltal) blir t.ex. sannolikheten att objekt *k* och *l* tilldelas samma intervjuare *i*

$$\Pr[i(k) = i(l)] = \frac{\binom{I}{1}\binom{q}{2}}{\binom{n}{2}} = \frac{n-I}{I(n-1)} \qquad \text{(a.1)}$$

Formel (a.1) används i kovariansen i formel (16) och (29). Eftersom intervjuareffekterna är oberoende i den valda mätfelmodellen kan mätfelsvariansen skrivas

$$V_m(\hat{t}_{yc\pi*}^{(A)}\big|s;r^{(A)}) = V_m\left(\sum_{h=1}^{H_s}\frac{n_h}{m_h^{(A)}}\sum\nolimits_{r_h^{(A)}}\breve{y}_k\big|s;r_h^{(A)}\right)$$

$$= \sum_{h=1}^{H_s}\left(\frac{n_h}{m_h^{(A)}}\right)^2\left[\sum\nolimits_{r_h^{(A)}}\frac{V_m(y_k\big|s;r_h^{(A)})}{\pi_k^2} + \sum_{k\neq l}\sum\nolimits_{r_h^{(A)}}\frac{Kov_m(y_k,y_l\big|s;k,l\in r_h^{(A)})}{\pi_k\pi_l}\right]$$

$$+\sum_{g\neq h}\sum\frac{n_g n_h}{m_g^{(A)}m_h^{(A)}}\sum\nolimits_{r_g^{(A)}}\sum\nolimits_{r_h^{(A)}}\frac{Kov_m(y_k,y_l\big|s;k\in r_g^{(A)};l\in r_h^{(A)})}{\pi_k\pi_l}$$

Vi ser att variansen kan delas upp i tre termer: en varians som motsvarar formel (14) i avsnitt 3.2.2,

$$V_m = V_m\left(y_k\big|s;k\in r_h^{(A)}\right) = \frac{1}{\theta_{hs}^{(A)}}\sum\nolimits_a(\theta_{hs}^{(a)}-\theta_{hs}^{(a-1)})(\sigma_b^2 a^{\gamma_b}+\sigma_\varepsilon^2 a^{\gamma_\varepsilon})$$

och två kovarianstermer som ges av formel (16) i samma avsnitt,

$$C_m = Kov_m\left(y_k,y_l|s;k\in r_g^{(A)};l\in r_h^{(A)}\right)$$

$$= P(i(k)=i(l)|s;k\in r_g^{(A)};l\in r_h^{(A)})$$

$$\times\sum\nolimits_{a_k}\sum\nolimits_{a_l}\frac{(\theta_{gs}^{(a_k)}-\theta_{gs}^{(a_k-1)})}{\theta_{gs}^{(A)}}\frac{(\theta_{hs}^{(a_l)}-\theta_{hs}^{(a_l-1)})}{\theta_{hs}^{(A)}}\sigma_b^2(a_k a_l)^{\gamma_b/2}$$

$$= \frac{n-I}{I(n-1)}\sum\nolimits_{a_k}\sum\nolimits_{a_l}\frac{(\theta_{gs}^{(a_k)}-\theta_{gs}^{(a_k-1)})}{\theta_{gs}^{(A)}}\frac{(\theta_{hs}^{(a_l)}-\theta_{hs}^{(a_l-1)})}{\theta_{hs}^{(A)}}\sigma_b^2(a_k a_l)^{\gamma_b/2}$$

Vi har här satt in sannolikheten definierad i formel (a.1). Observera att kovariansen förenklas något i det första fallet då $k,l\in r_h^{(A)}$.

Genom att sätta in V_m och C_m i variansen har vi

$$V_m(\hat{t}_{yc\pi^*}^{(A)}\big|s;r^{(A)}) = \sum_{h=1}^{H_s}\left(\frac{n_h}{m_h^{(A)}}\right)^2\left[\sum_{r_h^{(A)}}\frac{V_m}{\pi_k^2}+\sum\sum_{k\neq l}\sum_{r_h^{(A)}}\frac{C_m}{\pi_k\pi_l}\right] + \sum\sum_{g\neq h}\frac{n_g n_h}{m_g^{(A)}m_h^{(A)}}\sum_{r_g^{(A)}}\sum_{r_h^{(A)}}\frac{C_m}{\pi_k\pi_l}$$

och det är nu möjligt att ta fram *RD*-väntevärdet för uttrycket. Approximationen beror på att det är en kvotskattning i *RD*-mening. Väntevärdet av kvoten approximeras med kvoten av väntevärdena

$$E_{RD}[V_m(\hat{t}_{yc\pi^*}^{(A)})] \approx \sum_{h=1}^{H_s} n_h^2\left[\frac{V_m\sum_{s_h}\frac{\theta_{hs}^{(A)}}{\pi_k^2}+C_m\sum\sum_{k\neq l}\sum_{s_h}\frac{\theta_{hs}^{(A)}\theta_{hs}^{(A)}}{\pi_k\pi_l}}{(\sum_{s_h}\theta_{hs}^{(A)})^2}\right]$$

$$+\sum\sum_{g\neq h} n_g n_h\cdot C_m\frac{\sum_{s_g}\sum_{s_h}\frac{\theta_{gs}^{(A)}\theta_{hs}^{(A)}}{\pi_k\pi_l}}{\sum_{s_g}\theta_{gs}^{(A)}\sum_{s_h}\theta_{hs}^{(A)}}$$

$$=\sum_{h=1}^{H_s}\left(V_m\sum_{s_h}\frac{1}{\theta_{hs}^{(A)}\pi_k^2}+C_m\sum\sum_{k\neq l}\sum_{s_h}\frac{1}{\pi_k\pi_l}\right)+\sum\sum_{g\neq h}C_m\sum_{s_g}\sum_{s_h}\frac{1}{\pi_k\pi_l}$$

Genom att ta designförväntan, $E_p(\cdot)$, erhåller vi ett uttryck för V_{III}.

B Approximativ varians för den separata kvot-estimatorn

När vi beräknar variansen för estimatorn $\hat{t}_{yr}^{(A)}$ använder vi samma tillvägagångssätt i härledningen som vi använde oss av i appendix A. Vi har att

$$V(\hat{t}_{yr}^{(A)}) = V_p\{E_{RD}[E_m(\hat{t}_{yr}^{(A)}\big|s;r^{(A)})|s]\}+E_p[V_{RD}(\hat{t}_{\mu r}^{(A)}|s)]+E_p\{E_{RD}[V_m(\hat{t}_{yr}^{(A)}\big|s;r^{(A)})|s]\}$$
$$= V_I + V_{II} + V_{III}$$

I härledningen av samplingvariansen och den betingade bortfallsvariansen använder vi att $E_{RD}[E_m(\hat{t}_{yr}^{(A)}\big|s;r^{(A)})|s] = E_{RD}[\hat{t}_{\mu r}^{(A)}|s] \approx \hat{t}_{\mu r^*}$ där

$$\hat{t}_{\mu r*} = \sum_{h=1}^{H_s} N_h \frac{\sum_{s_h} \mu_k / \pi_k}{\sum_{s_h} 1/\pi_k} = \sum_{h=1}^{H_s} N_h \frac{\hat{t}_{h\pi}}{\hat{t}_{N_h}}.$$

Genom att använda estimatorn $\hat{t}_{\mu r*}$ har vi ett approximativt uttryck för V_I:

$$V_I \approx V_p(\hat{t}_{\mu r*}) = \sum \sum_U \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} E_k E_l$$

där $E_k = \mu_k - \overline{\mu}_{U_h}$ för $k \in U_h$.

För att härleda V_{II} noterar vi dels att mätfelen saknar bias, vilket ger oss $E_m(\hat{t}_{yr}^{(A)} \big| s; r^{(A)}) = \hat{t}_{\mu r}^{(A)}$, dels att den separata kvotestimatorn $\hat{t}_{\mu r}^{(A)}$ alternativt kan framställas som

$$\hat{t}_{\mu r}^{(A)} = \sum_{h=1}^{H_s} N_h \frac{\sum_{r_h^{(A)}} \mu_k / \pi_k}{\sum_{r_h^{(A)}} 1/\pi_k} = \sum_{h=1}^{H_s} N_h \frac{\sum_{s_h} R_{k|s}^{(A)} \mu_k / \pi_k}{\sum_{s_h} R_{k|s}^{(A)} / \pi_k} = \sum_{h=1}^{H_s} N_h \frac{\hat{t}_{\mu r_h}^{(A)}}{\hat{t}_{\theta \pi r_h}^{(A)}}$$

Första ordningens Taylorapproximation av $\hat{t}_{\mu r}^{(A)}$ ges som

$$\begin{aligned}\hat{t}_{\mu r}^{(A)} &\approx \sum_{h=1}^{H_s} N_h \frac{\sum_{s_h} \theta_{hs}^{(A)} \breve{\mu}_k}{\sum_{s_h} \theta_{hs}^{(A)} / \pi_k} + \sum_{h=1}^{H_s} \frac{N_h}{\sum_{s_h} \theta_{hs}^{(A)} / \pi_k} \sum_{s_h} \frac{R_{k|s}^{(A)}}{\pi_k} \left[\mu_k - \frac{\sum_{s_h} \theta_{hs}^{(A)} \breve{\mu}_k}{\sum_{s_h} \theta_{hs}^{(A)} / \pi_k} \right] \\ &= \sum_{h=1}^{H_s} N_h \hat{t}_{\mu r*} + \sum_{h=1}^{H_s} \frac{N_h}{\theta_{hs}^{(A)} \hat{t}_{N_h}} \sum_{s_h} \frac{R_{k|s}^{(A)}}{\pi_k} (\mu_k - \hat{t}_{\mu r*})\end{aligned}$$

Taylorapproximation används i härledningen av V_{II}:

$$
\begin{aligned}
V_{II} &= E_p[V_{RD}(\hat{t}_{\mu r}^{(A)}|s)] \\
&\approx E_p\left\{V_{RD}\left[\sum_{h=1}^{H_s} N_h \hat{t}_{\mu r^*} + \sum_{h=1}^{H_s} \frac{N_h}{\theta_{hs}^{(A)} \hat{t}_{N_h}} \sum_{s_h} \frac{R_{k|s}^{(A)}}{\pi_k}(\mu_k - \hat{t}_{\mu r^*}) \,\middle|\, s\right]\right\} \\
&= E_p\left\{\sum_{h=1}^{H_s}\left(\frac{N_h}{\theta_{hs}^{(A)} \hat{t}_{N_h}}\right)^2 \sum_{s_h} \frac{V_{RD}(R_{k|s}^{(A)}|s)}{\pi_k^2}(\mu_k - \hat{t}_{\mu r^*})^2\right\} \\
&= E_p\left\{\sum_{h=1}^{H_s}\left(\frac{N_h}{\hat{t}_{N_h}}\right)^2 \frac{(1-\theta_{hs}^{(A)})}{\theta_{hs}^{(A)}} \sum_{s_h} \frac{(\mu_k - \hat{t}_{\mu r^*})^2}{\pi_k^2}\right\}
\end{aligned}
$$

där $V_{RD}(R_{k|s}^{(A)}|s) = \theta_{hs}^{(A)}(1-\theta_{hs}^{(A)})$.

För mätfelsvariansen gäller samma förutsättningar som för direktvägningsestimatorn i appendix A. Vi har att

$$
\begin{aligned}
V_m(\hat{t}_{yr}^{(A)}|s; r^{(A)}) &= V_m\left(\sum_{h=1}^{H_s} \frac{N_h}{\hat{t}_{M_h}} \sum_{r_h^{(A)}} \breve{y}_k \,\middle|\, s; r_h^{(A)}\right) \\
= \sum_{h=1}^{H_s}\left(\frac{N_h}{\hat{t}_{M_h}}\right)^2 &\left[\sum_{r_h^{(A)}} \frac{V_m(y_k|s; r_h^{(A)})}{\pi_k^2} + \sum_{k\neq l}\sum_{r_h^{(A)}} \frac{Kov_m(y_k, y_l|s; k, l \in r_h^{(A)})}{\pi_k \pi_l}\right] \\
&+ \sum_{g\neq h}\sum \frac{N_g N_h}{\hat{t}_{M_g} \hat{t}_{M_h}} \sum_{r_g^{(A)}} \sum_{r_h^{(A)}} \frac{Kov_m(y_k, y_l|s; k \in r_g^{(A)}; l \in r_h^{(A)})}{\pi_k \pi_l}
\end{aligned}
$$

Genom att sätta in V_m och C_m i variansen har vi

$$
\begin{aligned}
V_m(\hat{t}_{yr}^{(A)}|s; r^{(A)}) = \sum_{h=1}^{H_s}\left(\frac{N_h}{\hat{t}_{M_h}}\right)^2 &\left[\sum_{r_h^{(A)}} \frac{V_m}{\pi_k^2} + \sum_{k\neq l}\sum_{r_h^{(A)}} \frac{C_m}{\pi_k \pi_l}\right] \\
&+ \sum_{g\neq h}\sum \frac{N_g N_h}{\hat{t}_{M_g} \hat{t}_{M_h}} \sum_{r_g^{(A)}} \sum_{r_h^{(A)}} \frac{C_m}{\pi_k \pi_l}
\end{aligned}
$$

och det är nu möjligt att ta fram *RD*-väntevärdet för uttrycket. Approximationen beror på att det är en kvotskattning i *RD*-mening. Väntevärdet av kvoten approximeras med kvoten av väntevärdena

$$E_{RD}[V_m(\hat{t}_{yr}^{(A)})] \approx \sum_{h=1}^{H_s} N_h^2 \left[\frac{V_m \sum_{s_h} \frac{\theta_{hs}^{(A)}}{\pi_k^2} + C_m \sum_{k\neq l} \sum_{s_h} \frac{\theta_{hs}^{(A)}\theta_{hs}^{(A)}}{\pi_k \pi_l}}{(\sum_{s_h} \theta_{hs}^{(A)} / \pi_k)^2} \right]$$

$$+ \sum_{g\neq h} \sum N_g N_h \cdot C_m \frac{\sum_{s_g} \sum_{s_h} \frac{\theta_{gs}^{(A)}\theta_{hs}^{(A)}}{\pi_k \pi_l}}{\sum_{s_g} \theta_{gs}^{(A)} / \pi_k \sum_{s_h} \theta_{hs}^{(A)} / \pi_l}$$

$$= \sum_{h=1}^{H_s} \left(\frac{N_h}{\hat{t}_{N_h}} \right)^2 \left(\sum_{s_h} \frac{V_m}{\theta_{hs}^{(A)} \pi_k^2} + \sum_{k\neq l} \sum_{s_h} \frac{C_m}{\pi_k \pi_l} \right) + \sum_{g\neq h} \sum N_g N_h C_m$$

Genom att ta designförväntan, $E_p(\cdot)$, erhåller vi ett uttryck för V_{III}.

Vi kan nu skriva variansen som

$$V(\hat{t}_{yr}^{(A)}) = V_I + V_{II} + V_{III}$$

$$\approx \sum\sum_U \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} E_k E_l$$

$$+ E_p \left\{ \sum_{h=1}^{H_s} \left(\frac{N_h}{\hat{t}_{N_h}} \right)^2 \frac{(1-\theta_{hs}^{(A)})}{\theta_{hs}^{(A)}} \sum_{s_h} \frac{(\mu_k - \hat{t}_{\mu r*})^2}{\pi_k^2} \right\}$$

$$+ E_p \left\{ \sum_{h=1}^{H_s} \left(\frac{N_h}{\hat{t}_{N_h}} \right)^2 \left(\sum_{s_h} \frac{V_m}{\theta_{hs}^{(A)} \pi_k^2} + \sum_{k\neq l} \sum_{s_h} \frac{C_m}{\pi_k \pi_l} \right) + \sum_{g\neq h} \sum N_g N_h C_m \right\}$$

C Härledning av resultat 4.3

Väntevärdet för $\hat{t}_{yc\pi^*}^{(A)}$ ges som

$$E_{OSU}(\hat{t}_{yc\pi^*}^{(A)}) \approx E_p \left(n \frac{\sum_s \theta^{(A)} \breve{\mu}_k}{\sum_s \theta^{(A)}} \right) = n \frac{E_p \left(\theta^{(A)} \sum_s \breve{\mu}_k \right)}{\theta^{(A)} n} = \frac{1}{\theta^{(A)}} \theta^{(A)} t_\mu = t_\mu$$

Variansen för $\hat{t}_{yc\pi^*}^{(A)}$ fås ur

$$V_{\text{OSU}}(\hat{t}_{yc\pi*}^{(A)}) = V_I + V_{II} + V_{III} \qquad \text{(b.1)}$$

Första komponenten i variansen för $\hat{t}^{(A)}_{yc\pi^*}$, V_I, erhålles enkelt genom insättning av aktuella inklusionssannolikheter under OSU:

$$V_I \approx N^2\left(1-\frac{n}{N}\right)\frac{S^2_{\mu U}}{n}$$

Komponenten V_{II} ges som

$$V_{II} \approx E_p\left[\sum\nolimits_s \frac{(1-\theta^{(A)})}{\theta^{(A)}}\left(\breve{\mu}_k - \frac{\sum_s \breve{\mu}_k}{n}\right)^2\right]$$

$$= \frac{1-\theta^{(A)}}{\theta^{(A)}}\left(\frac{N}{n}\right)^2(n-1)E_p\left(S^2_{\mu s}\right) = \frac{1-\theta^{(A)}}{\theta^{(A)}}\left(\frac{N}{n}\right)^2(n-1)S^2_{\mu U}$$

Mätfelsvariansen V_{III} ges slutligen som

$$V_{III} \approx E_p\left[V_m\sum\nolimits_s \frac{1}{\theta^{(A)}\pi_k^2} + C_m\sum_{k\neq l}\sum\nolimits_s \frac{1}{\pi_k\pi_l}\right] = \left(\frac{N}{n}\right)^2\left[V_m\frac{n}{\theta^{(A)}} + C_m n(n-1)\right]$$

Dessa varianskomponenter sätts nu in i (b.1):

$$V_{\mathrm{OSU}}(\hat{t}^{(A)}_{yc\pi^*}) \approx N^2\left(1-\frac{n}{N}\right)\frac{S^2_{\mu U}}{n} + \frac{1-\theta^{(A)}}{\theta^{(A)}}\left(\frac{N}{n}\right)^2(n-1)S^2_{\mu U}$$

$$+\left(\frac{N}{n}\right)^2\left[V_m\frac{n}{\theta^{(A)}} + C_m n(n-1)\right]$$

$$\approx \frac{1}{\theta^{(A)}}N^2\frac{S^2_{\mu U}}{n} + N^2\left(\frac{V_m}{n\theta^{(A)}} + C_m\right)$$

där vi i andra approximationen använder att $n \approx n-1$ och $1-n/N \approx 1$.

In English

Summary

In telephone surveys, several calls are often needed to establish contact with selected individuals. In this report, we consider the problem of choosing the maximum number of call attempts to be made to each selected individual before he or she is classified as unavailable. We start from the assumptions that the survey goal is to estimate a population total, that the sample can be divided into response homogeneity groups, and that the direct weighting estimator is to be used. We suggest a strategy for choosing the number of call attempts that takes the survey design, nonresponse, measurement errors and costs of data collection into account. The strategy relies on models for nonreponse and measurement errors. In our models, the impact of the error sources on the estimator may differ for different numbers of call attempts. According to our strategy, the number of call attempts is decided through a comparison of the estimator's standard error for different numbers of call attempts but the same cost. A simple but realistic example, in which the strategy is applied to a specific survey setting, is also given.

Optimalt antal kontaktförsök i en telefonundersökning

Hur många gånger ska man försöka ringa upp en urvalsperson i en telefonundersökning innan man ger upp? Detta är en viktig fråga som borde ställas i alla telefonenkäter. Många uppringningsförsök leder till höga kostnader och en lång produktionstid medan för få försök leder till för stora bortfallsfel.

Här presenteras en teoretisk metod för att bestämma antalet. Metoden går ut på att man jämför undersökningsupplägg med olika gränser men med samma kostnad. Det upplägg som ska väljas är det som ger lägst standardfel med avseende på alla osäkerhetskällor: urvalsfel, svarssannolikhet och mätfel. Metoden illustreras med ett räkne-exempel baserat på en realistisk undersökningssituation.

ISSN 1653-7149

Publikationstjänsten:
E-post: publ@scb.se, tfn: 019-17 68 00, fax: 019-17 64 44. Postadress: 701 89 Örebro.
Information och bibliotek: E-post: information@scb.se, tfn: 08-506 948 01, fax: 08-506 948 99.
Försäljning över disk, besöksadress: Biblioteket, Karlavägen 100, Stockholm.

Publication services:
E-mail: publ@scb.se, phone: +46 19 17 68 00, fax: +46 19 17 64 44. Address: SE-701 89 Örebro.
Information and Library: E-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99.
Over-the-counter sales: Statistics Sweden, Library, Karlavägen 100, Stockholm, Sweden.

www.scb.se