Stockholms universitet

Respondent-driven sampling med viktade vänskapsband - en simuleringsstudie

Tuure Antonangeli

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2013:5
Matematisk statistik
Juni 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2013:5**,
http://www.math.su.se/matstat

Respondent-driven sampling med viktade vänskapsband - en simuleringsstudie

Tuure Antonangeli *

Juni 2013

Sammanfattning

En metod för att göra urval från populationer som saknar urvalsram är snöbollsurval. Metoden går ut på att deltagare rekryterar individer från sitt sociala nätverk för att hitta nya individer i målgruppen, och den appliceras främst på populationer som är svåra att nå genom traditionella urvalsmetoder. Det anses dock allmänt ej vara möjligt att att generera väntevärdesriktiga skattningar från ett stickprov taget med snöbollsurval, vilket exempelvis beror på att individer med stor social krets blir överrepresenterade i ett sådant stickprov. Respondent-driven sampling, RDS, är en utvidgning av snöbollsurval som kombinerar en effektiviserad och fördjupad samplingsmetodik med en matematisk modell som genom ett antal antaganden om samplingsprocessen möjliggör väntevärdesriktiga skattningar av populationsegenskaper. Ett av de antaganden som görs är att en deltagare i studien rekryterar sina bekanta i populationen med lika stor sannolikhet, vilket inte är troligt i de flesta undersökningar. Avsikten med denna studie är att undersöka hur RDS-skattningarna påverkas om en deltagare får rekrytera sina bekanta med olika stor sannolikhet, det vill säga om man tillåter relationer ha differentierade vikter. Studien utförs genom datorsimuleringar där vi undersöker andelen HIV-positiva i en fiktiv population som saknar urvalsram och har ett socialt nätverk med viktade kanter. Resultaten tyder på att RDS-skattningarna överskattar andelen HIV-positiva då vänskapsbanden är viktade och sannolikheten att vara HIV-positiv beror på summan av vikterna på ens sociala kontakter. Om sannolikheten att vara HIV-positiv däremot beror på antalet kontakter kan man inte observera några större avvikelser i RDS-skattningarna jämfört med det oviktade fallet.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: tuure.antonangeli@hotmail.com. Handledare: Jens Malmros.

Abstract

Snowball sampling is a method for collecting samples from a population that lacks a sampling frame, where in order to find new individuals in the target-group, participants recruite people into the study from their own social network. The method is mainly used for populations that are hard to reach with traditional sampling methods. It is however considererd impossible to generate unbiased estimates from samples obtained by snowball sampling, e.g. because individuals with a large number of people in their social network are overrepresented in the study.

Respondent-Driven Sampling, RDS, is an extension of snowball sampling which combines a more efficient and advanced sampling methodology with a mathematical model, which through some assumptions about the sampling process can produce unbiased population estimates. One of the assumpions is that a participant recruits randomly from his peers in the population. This assumption is, however, implausible. The purpose of this study is to investigate how the RDS estimates are affected if a participant recruits from the social network with unequal probabilities, i.e. if the relationships have different weights.

The study is performed through computer simulation where we estimate the proportion that is HIV-positive in a fictitious population that lacks a sampling frame and has a social network with weighted edges. The results imply that RDS overestimates the proportion of HIV-positive individuals in the case where the relationships are weighted and the probability of being HIV-positive depends on the sum of the weights of the individuals social contacts. On the other hand, if the probability of being HIV-positive depends on the number of contacts, no notable difference can be observed in the RDS estimates when compared with the unweighted case.

Förord

Denna uppsats utgör ett självständigt arbete i matematisk statistik om 15hp. Ett stort tack riktas till min handledare Jens Malmros och alla mina vänner på matematiska institutionen för värdefull hjälp och goda råd.

Innehåll

1 Introduktion

1.1 Bakgrund

Ett vanligt sätt att skatta egenskaper hos en stor population är genom att undersöka ett slumpmässigt urval av enheter. Slumpmässiga urval kräver dock en urvalsram – uppgifter om mängden enheter från vilka man drar urvalet. Detta innebär i praktiken en lista över de enheter som man vill undersöka, som innehåller all nödvändig information om målpopulationen som behövs för att kunna utföra urvalet korrekt samt skatta egenskaperna man är ute efter [7]. Exempel på sådan information är bland annat storleken hos populationen och kontaktuppgifter [7].

Det finns målpopulationer som inte går att undersöka med hjälp av traditionella slumpmässiga urval, eftersom en lämplig urvalsram saknas. Dessa målgrupper kallas för dolda populationer (eng. *hidden populations*) [4]. Dolda populationer är ofta utsatta delar av samhället såsom prostituerade, drogmissbrukare eller hemlösa, på vilka traditionella undersökningsmetoder (exempelvis telefon- eller enkätundersökningar) inte kan appliceras. Ytterligare problematik skapas ofta av att individerna i målgruppen är motvilliga att samarbeta på grund av att deras beteende är juridiskt och/eller socialt oacceptabelt [4].

En av metoderna som tidigare användes för att undersöka dolda populationer är snöbollsurval (eng. *snowball sampling, chain-referral sampling*), där individerna som som deltar i undersökningen refererar vidare i sitt sociala nätverk för att hitta nya personer som tillhör målgruppen [4]. Med andra ord, en individ som deltar i undersökningen rekryterar ett antal vänner eller bekanta för att medverka i undersökningen. Metoden utvecklades från 1940-talet och framåt och ansågs länge vara den bästa tekniken för att utforska dolda populationer [4] [3].

Urvalet av individerna i en snöbollsurval sker dock inte slumpmässigt, och den traditionella snöbollssamplingsmetoden leder till väldokumenterade systematiska avvikelser [4]. Källor till systematiska avvikelser är till exempel att människor tenderar att rekrytera andra som tillhör samma etnicitet, inkomstgrupp eller dylikt [9]. Som ett illustrerande exempel kan man betrakta fallet där man vill undersöka andelen HIV-positiva bland sprutmissbrukande kvinnor. Om en deltagare förutom att vara en sprutmissbrukande kvinna, också är prostituerad, kan det befaras att personen väljer att rekrytera andra kvinnor som är både sprutmissbrukare *och* prostituerade. Tendensen att deltagarna väljer kvinnor som också är prostituerade kan göra att det skapas systematiska fel som man inte har kontroll över.

Ett annat problem med snöbollsurvalsbaserade undersökningar är att sannolikheten att en individ ur populationen deltar i urvalet är beroende av antalet individer som man har i sitt sociala krets [4]. Människor med en omfattande social krets är överrepresenterade i urvalet, eftersom det finns flera möjliga vägar att nå en sådan individ [4]. Ifall även egenskapen som undersöks beror på storleken på den sociala kretsen blir resultatet ej väntevärdesriktigt. Ett exempel där storleken på den sociala kretsen kan påverka resultaten är om man vill skatta andelen HIV-positiva bland sprutmissbrukare: Man kan förmoda att det är

mer sannolikt att en individ är HIV-positiv om individen har många sprutmissbrukande vänner, eftersom sprutmissbrukare har benägenhet att dela använda sprutor med varandra.

1.2 Respondent driven sampling – RDS

Undersökningar baserade på snöbollsurval ger alltså i allmänhet systematiska fel i de alstrade skattningarna [4]. Det finns dock en variant av snöbollsurval, Respondent-driven sampling, där man med ett antal antaganden skapar en matematisk modell av händelseförloppet och på så sätt genererar väntevärdesriktiga estimatorer för egenskaperna hos populationen. Respondent driven sampling är en teknik som utvecklades av Douglas Heckathorn 1997 i samband med en forskningsprojekt om HIV-spridning bland drogmissbrukare [9]. Metoden använder sig bland annat av teorin för Markovkedjor, och det har visats att RDS-baserade skattningar minskar de systematiska avvikelserna jämfört med skattningar som baseras på traditionell snöbollsurval [4].

Såsom snöbollsurvalsmetoden är RDS ett effektivt sätt att penetrera populationer som i allmänhet är svåra för en forskare att nå. Samplingsprocessen startar med ett eller flera "frö" [4]. Frön är de första deltagarna i undersökningen som vanligen inte bestäms slumpmässigt, utan är i praktiken lättillgängliga personer som man vet tillhör målgruppen [4]. Ett frö kan således kontaktas till exempel på ett stödcenter, vårdcentral eller i fallet sprutmissbrukare, på en sprutbytescentral. Fröet tilldelas efter intervjun ett bestämt antal kuponger som han eller hon skall utdela till sina bekanta som också tillhör målgruppen. De personer som tilldelats en kupong får möjligheten att bli intervjuade mot en belöning, som ges även till kupongutdelaren. Efter intervjun tilldelas de intervjuade individerna nya kuponger och upprepar rekryteringsprocessen bland sina bekanta i populationen. Antalet kuponger som en deltagare ges för att sedan dela ut är alltid begränsat [4].

Viktiga skillnader mellan "vanligt" snöbollsurval och Respondent-driven sampling är:

1. RDS motiverar en deltagare att rekrytera nya individer till studien genom att erbjuda ytterligare en belöning när dessa deltar i studien [4].
2. En deltagare behöver inte namnge de individer som skall rekryteras – en viktig egenskap när man undersöker människor som av diverse anledningar vill förbli anonyma [4].
3. Kupongerna som delas ut vid intervjutillfället har ett specifikt serienummer för att övervaka vem som har rekryterat vem. Antalet kuponger är också alltid begränsat. På så sätt är det möjligt att erhålla en helhetsbild över rekryteringsprocessen [8].

2 Teori

Innan studien kan fortsätta vidare tar vi upp bakgrund om grafteori samt något om Markovkedjor i diskret tid, samt hur dessa är relaterade till rekryterings-

processen. I detta Avsnitt härleds även formeln för RDS-skattningen som är relevant för denna studie.

2.1 Något om grafer och deras relation till RDS-studier

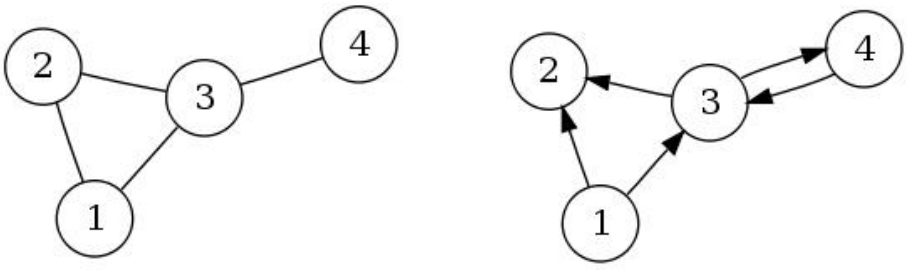
I detta Avsnitt kommer vi att gå igenom de nödvändiga begreppen inom grafteorin som behövs för denna studie. För en mera omfattande introduktion till grafteorin hänvisas till boken *Discrete and Combinatorial Mathematics* av Ralph P. Grimaldi, Avsnitt 11.

Definition 2.1. Graf: *Låt $V = \{v_1, v_2, ...\}$ vara en ändlig och icke-tom mängd, och låt $E \subseteq V \times V$. Paret (V, E) kallas en graf på V, där V är mängden noder (eng.* vertices*) och $E = \{e_1, e_2, ...\}$ är dess mängd riktade eller oriktade kanter (eng.* edges*). En sådan graf betecknas $G = (V, E)$.* [1]

En graf är således en mängd punkter – *noder* – som kan bindas samman med riktade eller oriktade linjer – *kanter*, se Figur 1. Om $u, v \in V$ och kanten mellan noderna u och v är riktad från u till v, innebär det att man kan gå från u till v men inte från v till u. Ett enkelt exempel är att noderna representerar städer, och kanterna motsvarar ett nätverk av vägar mellan städerna [1]. Om kanterna är riktade kan man tänka sig att vägen mellan två städer är enkelriktad. I denna studie kommer vi inte att betrakta grafer som innehåller så kallade loopar (eng. *loops*), som är kanter som leder tillbaka till samma nod [1].

Till varje kant e_i i en graf kan tilldelas ett positivt reellt tal som kallas för *vikten* för kanten och i denna studie benämns som $\mathrm{v}(e_i)$. Ifall en graf har egenskapen att den innehåller viktade kanter, kallas grafen för en viktad graf (eng. *weighted graph*). [1]

Ett socialt nätverk kan beskrivas som en graf där noderna i grafen representerar individerna i det sociala nätverket. Om det finns en kant mellan noderna, förekommer det någon form av relation mellan dessa individer. En relation innebär bekantskap, vänskap, sexuell kontakt eller dylikt. I denna studie innebär en kant mellan två personer att det finns möjlighet till rekrytering.



(a) Oriktad graf (b) Riktad graf

Figur 1: En illustration av en graf med fyra noder, riktad och oriktad version. Hämtad från *http://pl.wikipedia.org/wiki/Graf_%28matematyka%29* 5.4.2013

Definition 2.2. Grad: *Låt G vara en oriktad graf. För alla noder v i G definieras graden av v som antalet kanter som leder till v.* [1]

Med andra ord, graden för nod v är antalet noder som kan nås genom att endast åka längs *en* kant, se Figur 2. I en RDS-studie motsvarar graden för en person det teoretiska antalet individer som personen kan rekrytera [8]. En deltagare anger sin grad vid intervjutillfället.

Figur 2: Illustration av noder med grad 0-4. Figuren hämtad från [5].

2.2 Något om Markovkedjor i diskret tid

I detta delavsnitt utgår vi från boken *Introduction to Probability Models,* 9^{th} *edition* av Sheldon M. Ross.

Låt $\{X_n;\ n = 0, 1, 2, ...\}$ vara en sekvens av stokastiska variabler som tar värden i en uppräknelig mängd S. Om

$$\begin{aligned} &P(X_{n+1} = j \mid X_n = i,\ X_{n-1} = i_{n-1}, ..., X_1 = i_1,\ X_0 = i_0) \\ &\quad = P(X_{n+1} = j \mid X_n = i) = P_{ij} \in [0, 1] \end{aligned}$$

sägs $\{X_n;\ \ n = 0, 1, 2, ...\}$ vara en *Markovkedja.*

Om Markovkedjan befinner sig i tillstånd i, finns det en given sannolikhet P_{ij} för övergång till tillstånd j; $i, j \in S$. En viktig egenskap hos Markovkedjor är att sannolikhetsfördelningen för X_{n+1} *endast* beror på det nuvarande tillståndet X_n, och är oberoende av vilka tillstånd som besökts innan, det vill säga oberoende av $X_{n-1}, X_{n-2}, ..., X_1, X_0$.

Ett vanligt sätt att illustrera Markovkedjor är riktade grafer (Figur 1b), där noderna representerar tillstånden som Markovkedjan kan befinna sig i. Om det finns en pil från tillstånd i till tillstånd j, innebär det att en sådan övergång är möjlig med sannolikhet P_{ij}. Om P_{ij} är 0 är övergången ej möjlig och ingen pil ritas mellan tillstånden.

Övergångssannolikheterna för en Markovkedja brukar beskrivas med hjälp av en så kallad övergångsmatris $\mathbf{P}$ som består av elementen P_{ij}, så för $S \equiv \mathbb{N}$ och $i, j = 0, 1, 2, ...$ har vi

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & \cdots & P_{0j} & \cdots \\ P_{10} & P_{11} & \cdots & P_{1j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \cdots \\ P_{i0} & P_{i1} & \cdots & P_{ij} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Notera att en övergång alltid måste ske i varje tidssteg, även om övergången sker tillbaka till samma tillstånd, och att elementena i varje rad i $\mathbf{P}$ summeras till 1.

Exempel 2.1. Låt oss exempelvis betrakta övergångsmatrisen för den oriktade grafen i Figur 1a, och antag att sannolikheten för en övergång till ett annat tillstånd är lika stor för alla möjliga övergångar. Då gäller att

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Exempel 2.2. Låt oss istället betrakta övergångsmatrisen för den riktade grafen i Figur 1b, och antag att sannolikheten för en övergång till ett annat tillstånd är lika stor för alla möjliga övergångar. Då gäller att

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Antag nu att systemet befinner sig i tillstånd i. Om det är möjligt att nå tillstånd j i $n \in \mathbb{N}$ steg, sägs tillstånd j vara accessibelt (eng. *accessible*) från tillstånd i och vi skriver $i \longrightarrow j$. Om $i \longrightarrow j$ och $i \longleftarrow j$ sägs det att tillstånden *kommunicerar* och vi skriver $i \longleftrightarrow j$. Denna relation är en ekvivalensrelation och alltså har följande egenskaper:

- tillstånd i kommunicerar med tillstånd i
- om tillstånd i kommunicerar med tillstånd j, så kommunicerar tillstånd j med tillstånd i
- Om tillstånd i kommunicerar med tillstånd j och tillstånd j kommunicerar med tillstånd k, så kommunicerar tillstånd i med tillstånd k.

Tillstånd som kommunicerar med varandra sägs tillhöra samma *klass*. Till exempel har Markovkedjan som illustreras i Figur 1b klasserna $\{1\}$,$\{2\}$ och $\{3, 4\}$. Om det endast finns *en* klass, sägs Markovkedjan vara *irreducibel*.

För godtyckligt tillstånd i, låter vi f_i vara sannolikheten att återvända till tillstånd i givet att vi startar processen i tillstånd i. Om $f_i = 1$ sägs tillstånd i

vara *rekurrent*. Ifall $f_i < 1$ sägs tillstånd i vara *transient*.

Tillstånd i sägs ha perioden d om sannolikheten att återkomma till tillstånd i i n steg är noll, om n inte är delbart med d, och d är det minsta heltalet med denna egenskap. Om tillstånd i har perioden 1, sägs tillståndet vara *aperiodiskt*.

Markovkedjor med ändligt många tillstånd, som är irreducibla och där alla tillstånden är aperiodiska är *ergodiska*. Sådana markovkedjor är av intresse i denna studie eftersom det för en irreducibel och ergodisk Markovkedja med $N < \infty$ tillstånd existerar en unik asymptotisk fördelningsvektor $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_N)$. Med andra ord, om vi låter Markovkedjan pågå tillräckligt länge, konvergerar den genomsnittliga andelen tid som vi tillbringar i tillstånd i mot π_i.

2.3 RDS-process som en Markovkedja i diskret tid

Vi fortsätter med att beskriva den bakomliggande teorin som möjliggör RDS-skattningarna. Följande antaganden måste vara uppfyllda för RDS-baserade studier [8]:

1. *Reciprocitet* Reciprocitet innebär att ett eventuellt förhållande mellan två personer är ömsesidig; om individ A anser att individ B är dess vän, så anser individ B att individ A är dess vän. En eventuell rekrytering kan alltså ske åt båda hållen. I grafteoretiska termer innebär detta att det sociala nätverket beskrivs av oriktade kanter. [1]

2. En individ korrekt anger sin grad i nätverket. Detta är av stor vikt eftersom den genomsnittliga graden hos en population U, och den genomsnittliga graden hos en delpopulation $A \subset U$ kommer att vara nödvändiga för att göra en RDS-skattning av andelen av befolkningen som tillhör typ A.

3. När en deltagare rekryterar en annan individ har alla närliggande noder samma sannolikhet att bli rekryterade. Med andra ord, om graden för person A är n, är sannolikheten att rekrytera person B (som A har en relation till) lika med $1/n$.

För att förenkla modelleringen i denna studie antas även att processen startar med endast ett frö, samt att data som undersöks är av binär karaktär (Ja, Nej), (+,-), (0,1). Dessutom antar vi att en deltagare endast får *en* kupong vid intervjutillfället. Detta innebär att en individ endast kan rekrytera en person i det sociala nätverket som han eller hon har någon form av förhållande till. Dessa antaganden är dock inga krav för att RDS generellt skall fungera.

Det faktum att en individ endast kan rekrytera *en* annan individ gör att rekryteringsprocessen enkelt kan beskrivas som en slumpvandring på det sociala nätverket. Slumpvandringen modelleras bäst med en Markovkedja i diskret tid med tillståndsrum $\{1, 2, 3, ...\}$, där tillståndet som Markovkedjan befinner sig i motsvarar en individ som "blir intervjuad" i studien. Om det finns en kant

[1] RDS på delvis riktade nätverk kan studeras i [5].

mellan två individer, innebär det att en övergång mellan dessa två "tillstånd" kan ske. Om antalet individer i hela målpopulationen är N, har Markovkedjan således en motsvarande övergångsmatris av dimension $N \times N$. De tillstånd som Markovkedjan besöker utgör urvalet av individer.

Övergångsmatrisen för slumpvandringen antas vara irreducibel, det vill säga ifall vi startar från ett givet tillstånd, kan alla andra tillstånd nås efter ett ändligt antal övergångar. Notera att en individ enligt modellen kan rekryteras mer än en gång. Trots detta antas (givetvis) att en individ inte rekryterar sig själv, vilket innebär att diagonalelementena i övergångsmatrisen är noll.

I praktiken betyder irreducibiliteten att givet ett godtyckligt frö, sträcker sig det sociala nätverket över hela målpopulationen. Detta är naturligtvis ett idealiserat scenario, men faktum är att felen som uppstår med detta antagande visar sig vara små [8]. Inom nätverksteorin är det känt att de flesta nätverk innehar en sammanhängande delmängd noder som kallas för " den stora komponenten" (eng. *giant component*) [8]. Ifall den genomsnittliga graden för noderna är 5, så visar det sig att den stora komponenten utgör 99% av hela populationen [8]. Alltså kommer endast isolerade noder eller små grupperingar hamna utanför komponenten [5]. För att minimera risken att hamna utanför den stora komponenten då begynnelsedeltagaren väljs, kommer vi i denna studie endast betrakta fall där den genomsnittliga graden för noderna är 10 eller större.

Eftersom processen beskrivs av en ergodisk Markovkedja vet vi att det finns en unik asymptotisk fördelning för tillstånden [6]. Detta medför att efter att processen har fortgått under en tillräckligt lång tid så kan vi räkna ut sannolikheterna att befinna oss i vart och ett av tillstånden, oberoende av begynnelsetillståndet (fröet). Alltså finns det en *konvergens* i förloppet.

2.4 Härledning av RDS-skattaren

I detta delavsnitt utgår vi från artikeln *Probability Based Estimation Theory for Respondent Driven Sampling, (2008)* av Erik Volz och Douglas D. Heckathorn. Därför anses det lämpligt att införa liknande beteckningar som i denna artikel. Notationen finns listad i Tabell 1.

Tabell 1: Notation som används i Avsnitt 2.4

U är mängden av alla individer i populationen
S är mängden av alla individer i urvalet
$A, B, ...$ är disjunkta mängder av individer
N är antalet individer i populationen
n är antalet individer i urvalet
$P_A, P_B, ...$ är andelen av populationen som tillhör A, B,...
σ_{AB} är den skattade sannolikheten att en individ i A rekryterar någon i B
δ_i är graden hos individ i
δ_X är den genomsnittliga graden hos individerna i en mängd X

Låt oss ha ett socialt nätverk där antalet individer är N. Om det finns en

kant mellan noderna i och j; $i, j = 1, ..., N$, och δ_i är graden hos individ i så är övergångssannolikheten från nod i till nod j lika med $1/\delta_i$. Denna övergångssannolikhet betecknas σ_{ij}. Låt $\boldsymbol{\sigma}$ vara övergångsmatrisen. Matrisen $\boldsymbol{\sigma}$ kommer alltså att vara en matris av storlek $N \times N$ med $\sigma_{ij} = 0$ om en övergång inte är möjlig och $\sigma_{ij} = 1/\delta_i$ om en övergång är möjlig. Det är möjligt att härleda att vektorn $\mathbf{x}^*$, där elementena

$$x_i^* = \frac{\delta_i}{\sum_{j=1}^{N} \delta_j}, \tag{1}$$

är den asymptotiska fördelningsvektorn för tillstånden [8]. Ekvation (1) kan skrivas om för att beräkna sannolikheten p_i att en specifik individ i i det sociala nätverket blir utvald för intervju:

$$p_i = \frac{\delta_i}{N \cdot \delta_U}$$

som skattas med

$$\hat{p}_i = \frac{\delta_i}{N \cdot \hat{\delta}_U}$$

där $\hat{\delta}_U$ är den skattade genomsnittliga graden i det sociala nätverket. Den skattas med en så kallad Hansen & Hurwitz estimator – standardestimator i fallet slumpmässigt urval med återläggning från en population där enheterna har olika sannolikheter att bli utvalda:

$$\hat{\delta}_U = \frac{\sum_S \delta_i / np_i}{\sum_S 1/np_i} = \frac{n}{\sum_S \delta_i^{-1}}. \tag{2}$$

Eftersom vi i denna studie begränsar oss till binära utfall, kan vi dela upp målgruppen i två delmängder; typ A och typ B. Typ A och typ B kan exempelvis betyda HIV-positiv respektive HIV-negativ. Om vi vill skatta den genomsnittliga graden hos en delmängd $A \subset U$ kan man använda Ekvation (2) direkt, så att

$$\hat{\delta}_A = \frac{n_A}{\sum_{A \cap S} \delta_i^{-1}}$$

där $A \cap S$ är mängden individer i urvalet som tillhör typ A och n_A är antalet individer i urvalet som tillhör typ A.

Låt nu y_i vara en egenskap hos individ i som man är intresserad av att undersöka, och låt T_y vara det totala värdet av y i hela populationen U, det vill säga $T_y = \sum_U y_i$. Även T_y kan skattas med en Hansen & Hurwitz – estimator, så att

$$\hat{T}_y = \frac{1}{n} \sum_S \frac{y_i}{\hat{p}_i} = \frac{1}{n} \sum_S \frac{\hat{\delta}_U N}{\delta_i} = \frac{\hat{\delta}_U N}{n} \sum_S \delta_i^{-1} y_i. \tag{3}$$

Vi ser att Ekvation (3) innehåller N – storleken på målgruppen. Denna storhet antas dock vara okänd eftersom vi har med dolda populationer att göra. Om vi däremot vill skatta medelvärdet av y i populationen faller N bort från Ekvation (3), eftersom medelvärdet av y beräknas som $\frac{T_y}{N}$ – totala värdet av y delat med storleken på populationen. Alltså

$$\hat{\bar{y}} = \frac{\hat{\delta}_U}{n} \sum_S \delta_i^{-1} y_i = \frac{\sum_S \delta_i^{-1} y_i}{\sum_S \delta_i^{-1}}$$

där vi i den andra olikheten har utnyttjat Ekvation (2).

Låt nu y vara en binär variabel som beskrivs av indikatorfunktionen

$$y_i = \begin{cases} 1 & \text{om individ } i \text{ tillhör grupp } A, \\ 0 & \text{om individ } i \text{ inte tillhör grupp } A. \end{cases}$$

Därmed kan vi göra omskrivningen

$$\sum_S \delta_i^{-1} y_i = \sum_{A \cap S} \delta_i^{-1}.$$

Låt P_{A} vara andelen av målgruppen som tillhör typ A. En väntevärdesriktig skattning $\hat{P}_{\text{A}}$ ges således av

$$\hat{P}_{\text{A}} = \frac{\sum_{A \cap S} \delta_i^{-1}}{\sum_S \delta_i^{-1}}. \tag{4}$$

3 Syfte

I början av Avsnitt 2.3 presenterades 3 villkor som måste uppfyllas för att RDS skall kunna generera väntevärdesriktiga skattningar. Syftet med denna studie är att med hjälp av datorsimuleringar undersöka hur RDS-skattningarna påverkas om antagande nummer 3 inte uppfylls. Med andra ord, *vad händer om det inte är samma sannolikhet för alla närliggande noder att bli rekryterade?* Givet att en individ har n stycken personer i sin sociala krets, är det inte troligt att alla har samma sannolikhet att bli rekryterade, utan troligtvis beror sannolikheten att bli rekryterad på relationen. En person tenderar förmodligen att rekrytera en god vän snarare än en bekant. I denna uppsats inför vi därför *viktade vänskapsband*, där vikten för vänskapsbandet anger nivån på relationen. Till exempel, om individ A har två kontakter varav en med vikten 2 och en med vikten 1, är det dubbelt så sannolikt att individ A rekryterar kontakten med vikten 2.

4 Metod

Vi antar för enkelhetens skull den fiktiva situationen att vi undersöker andelen HIV-positiva individer i en dold population. Således blir den så kallade grupp A som beskrevs i Avsnitt 2.4 delmängden av individer i populationen som är HIV-positiva. Andelen HIV-positiva individer betecknas därmed som P_{HIV}.

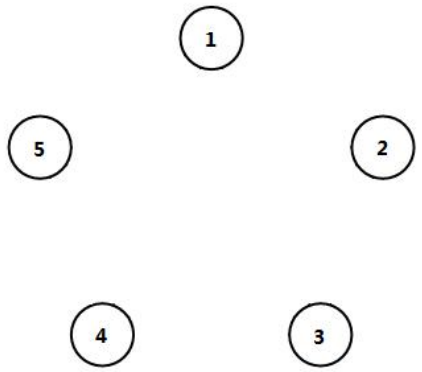
Studien utförs med hjälp av data som genereras i en Matlab-simulering. Analysen av det erhållna datamaterialet sker i första hand genom grafisk undersökning av låddiagram, men resultaten kommer även att presenteras i tabeller.

I Avsnitten 4.1 – 4.3 presenteras en schematisk beskrivning av hur simuleringsprocessen utförs. I samband med beskrivningen presenteras även ett illustrerande exempel.

4.1 Generering av nätverket

1. Vi skapar N stycken noder som motsvarar individerna i målpopulationen.

   Exempel 4.1. För att få ett hanterbart exempel väljer vi $N = 5$. Dessa illustreras i Figur 3.



Figur 3: 5 stycken noder som motsvarar en population av 5 individer.

2. Nu bestäms vilka av dessa N stycken individer som har kontakt med varandra. Låt oss bilda en matris $\mathbf{A}$ av storlek $N \times N$ som har egenskapen att $A_{ij} = A_{ji} \sim \text{Ber}(\lambda/N)$[2] och $A_{ii} = 0;\ 1 < \lambda < N;\ i, j = 1, 2, ..., N$. Denna matris tilldelar slumpmässigt kanter mellan noderna. Kanterna som genereras representerar vänskapsband mellan individerna i studien. Ifall $A_{ij} = A_{ji} = 1$ finns det en kant mellan noderna i och j, ifall $A_{ij} = A_{ji} = 0$ finns det ingen kant. Summan av elementena på rad i i $\mathbf{A}$ är således antalet kontakter som individ i har i det sociala nätverket, det vill säga *graden* för individ i. Låt D_i vara graden för individ i. Notera att D_i blir således summan av $N - 1$ (element $A_{ii} = 0\ \forall i = 1, 2, ..., N$) stycken $\text{Ber}(\lambda/N)$-fördelade stokastiska variabler. Detta innebär att $D_i \sim \text{Bin}(N - 1, \lambda/N)$.

   Eftersom N kan anses vara stort kan vi utnyttja oss av att $D_i \xrightarrow{d} Y$ där $Y \sim \text{Po}(\lambda)$. Detta kan verifieras genom att beräkna den sannolikhetsgenererande funktionen för D_i:

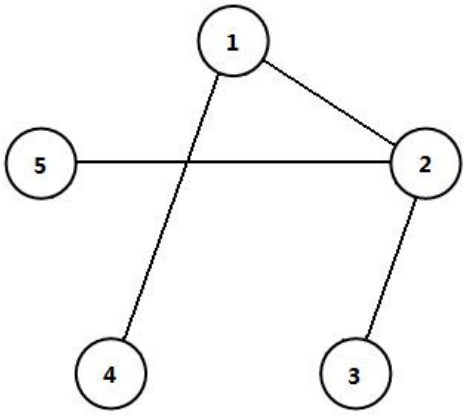
$$\begin{aligned} g_{D_i}(t) &= \left(1 - \frac{\lambda}{N} + \frac{\lambda}{N}t\right)^{N-1} = \left(1 + \frac{\lambda(t-1)}{N}\right)^{N-1} \\ &= \left(1 + \frac{\lambda(t-1)}{N}\right)^{N} \cdot \left(1 + \frac{\lambda(t-1)}{N}\right)^{-1} \\ &\longrightarrow e^{\lambda(t-1)} \cdot 1 = e^{\lambda(t-1)} = g_Y(t) \text{ då } N \longrightarrow \infty, \end{aligned}$$

[2]Detaljer om fördelningarna som används i denna studie återfinns i Bilaga B.

där $Y \sim \text{Po}(\lambda)$ [2].

Exempel 4.1 (fortsättning). Om vi väljer $\lambda = 2$ kan $\mathbf{A}$ i vårt exempel med $N = 5$ individer se ut som följande:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$



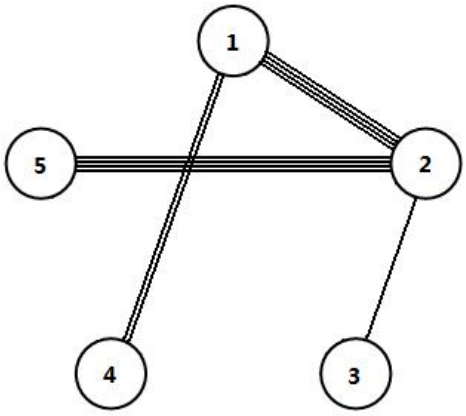
Figur 4: Grafen som genereras av matrisen $\mathbf{A}$ i Exempel 4.1. Kanterna motsvarar vänskapsband mellan individerna i populationen.

3. Vi skapar en matris $\mathbf{B}$ av storlek $N \times N$ sådan att om $A_{ij} = A_{ji} = 1$, så är $B_{ij} = B_{ji}$ ett utfall av en stokastisk variabel V. Elementena i $\mathbf{B}$ motsvarar *vikten* av vänskapsbandet – där djupet på relationen ökar med ökande vikt. Notera att $B_{ij} = B_{ji}$ inte kan tilldelas värdet noll, eftersom kontakten i så fall stryks helt och hållet och då skulle inte graden för individ i vara densamma som i $\mathbf{A}$.

   Man kan nu tänka sig att det istället för en enkel kant mellan två noder finns en kant med en stokastisk vikt – motsvarade vikten för relationen, och att det är större sannolikhet för övergång till en nod som förbinds med en kant med högre vikt.

   Exempel 4.1 (fortsättning). Om vi låter $B_{ij} = B_{ji} \sim \text{U}(4)$ kan matrisen $\mathbf{B}$ se ut som följande:

$$\mathbf{B} = \begin{pmatrix} 0 & 4 & 0 & 2 & 0 \\ 4 & 0 & 1 & 0 & 4 \\ 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \end{pmatrix}$$



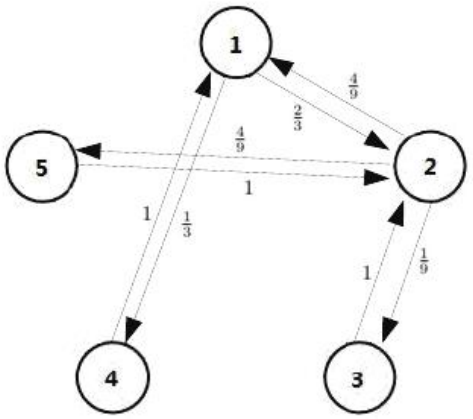
Figur 5: Den viktade grafen som skapas av matrisen **B** i Exempel 4.1. Enligt modellen känner individerna 2 och 1 med vikten 4 i deras vänskapsband, varandra bättre än individerna 2 och 3, som endast har vikten 1 i vänskapsbandet.

4. Matrisen **B** normaliseras så att varje vikt mellan två individer divideras med totala viktsumman för individen. Med andra ord, varje element i **B** delas med radsumman för raden där elementet befinner sig i. På så sätt har vi konstruerat en övergångsmatris **C** för en Markovkedja i diskret tid där övergångssannolikheten beror på djupet för relationen.

   Exempel 4.1 (fortsättning). Genom att normalisera **B** erhålls matrisen

$$\mathbf{C} = \begin{pmatrix} 0 & \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ \frac{4}{9} & 0 & \frac{1}{9} & 0 & \frac{4}{9} \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

   Nätverket som matrisen **C** genererar illustreras i Figur 6.



Figur 6: Grafen med övergångssannolikheterna som fås ur **C** i Exempel 4.1. Till exempel, sannolikheten att individ 2 rekryterar individ 3 är 1/9 medan sannolikheten att individ 2 rekryterar individerna 1 eller 5 är 4/9.

4.2 Allokering av HIV-status

4.2.1 Allokering proportionellt mot grad

Låt D_i vara graden för individ i och låt $\overrightarrow{D} = (D_1, D_2, ..., D_N)$. Som vi redan konstaterat i delavsnitt 4.1 så är D_i Po(λ) – fördelad för stora värden på N. Låt vidare H_i vara en stokastisk variabel som indikerar om individ i i populationen är HIV-positiv eller ej, så att

$$H_i = \begin{cases} 1 & \text{om individ } i \text{ är smittad} \\ 0 & \text{om individ } i \text{ inte är smittad} \end{cases} \tag{5}$$

och $\overrightarrow{H} = (H_1, H_2, ..., H_N)$. För att få en positiv korrelation mellan vektorerna $\overrightarrow{D}$ och $\overrightarrow{H}$ gör vi i enlighet med [5] följande ansats:

$$P(H_i = 1) = \min\left(1, \frac{p_0 \cdot D_i}{\mathrm{E}[D_i]}\right) = \min\left(1, \frac{p_0 \cdot D_i}{\lambda}\right) \tag{6}$$

där p_0 är en parameter som vi kan manipulera för att bestämma en sann proportion av populationen som är HIV-positiv. Ekvation (6) innebär att sannolikheten att individ i är HIV-positiv ökar linjärt med ökande grad, dock med villkoret att sannolikheten alltid är mindre än eller lika med 1. Givetvis gäller att $P(H_i = 0) = 1 - P(H_i = 1)$.

För att bestämma vilka individer i målpopulationen som är HIV-positiva genereras en vektor $\overrightarrow{H} = (H_1, H_2, ..., H_N)$ där $H_i,\ i = 1, ..., N$ har egenskaperna som beskrivs i ekvationerna (5) och (6). För varje $H_i,\ i = 1, ..., N$ genereras ett motsvarande slumptal r_i som är ett utfall av en U(0, 1)-fördelning. Ifall $r_i < \min(1, \frac{p_0 \cdot D_i}{\lambda})$ sätts $H_i = 1$, annars sätts $H_i = 0$.

Exempel 4.1 (fortsättning). Låt oss i vårt exempel med $N = 5$ och $\lambda = 2$ välja $p_0 = 0.4$. En datorsimulering ger följande värden:

Tabell 2: Allokering av HIV bland individerna i exempel 4.1. Allokeringen görs enligt Ekvation (6) med $N = 5$, $\lambda = 2$ och $p_0 = 0.4$. $r_i \sim \mathrm{U}(0, 1)$ genereras i Matlab.

i	D_i	$P(H_i = 1)$	r_i	H_i
1	2	0.40	0.19	1
2	3	0.60	0.46	1
3	1	0.20	0.82	0
4	1	0.20	0.52	0
5	1	0.20	0.03	1

Detta innebär att i det sociala nätverket som illustreras i Figur 4 är individerna 1,2 och 5 HIV-positiva.

4.2.2 Allokering proportionellt mot styrka

Som redan bekrivits i Avsnitt 4.1 tilldelas vänskapen mellan individerna i och j en stokastisk vikt. I denna studie låter vi vikten för relationen mellan två individer vara ett utfall av en stokastisk variabel V. Vi definierar *styrkan* för individ i som viktsumman, det vill säga $S_i = \sum_{j=1}^{D_i} V_j$, där V_j är vikten för individ i:s j:te kant och där $V_1, V_2, ...$ är oberoende och lika fördelade. Notera att S_i är således en summa av ett stokastiskt antal stokastiska variabler.

Låt H_i definieras av Ekvation (5) och låt $\overrightarrow{S} = (S_1, S_2, ..., S_N)$ där S_i är styrkan för individ i. För att få en positiv korrelation mellan vektorerna $\overrightarrow{H} = (H_1, H_2, ..., H_N)$ och $\overrightarrow{S} = (S_1, S_2, ..., S_N)$ låter vi analogt med Ekvation (6)

$$P(H_i = 1) = \min\left(1, \frac{p_0 \cdot S_i}{\mathrm{E}[S_i]}\right) \tag{7}$$

där vi enligt [2] har $\mathrm{E}[S_i] = \mathrm{E}[D_i] \cdot \mathrm{E}[V]$. Vektorn $\overrightarrow{H}$ genereras på samma sätt som i Avsnitt 4.2.1.

Exempel 4.1 (fortsättning). Tabell 3 illustrerar en generering av $\overrightarrow{H}$ - givet att sannolikheten att vara HIV-positiv beror på styrkan.

Tabell 3: Allokering av HIV bland individerna i exemplet. Allokeringen görs enligt Ekvation (7) med $V \sim \mathrm{U}(4)$, $N = 5$, $\lambda = 2$ och $p_0 = 0.4$.

i	S_i	$P(H_i = 1)$	r_i	H_i
1	6	0.48	0.86	0
2	9	0.72	0.64	1
3	1	0.08	0.10	0
4	2	0.16	0.11	1
5	4	0.32	0.33	0

4.3 Slumpvandringen och RDS-skattningen

Vi utgår från övergångsmatrisen $\mathbf{C}$ som genereras enligt beskrivning i Avsnitt 4.1. Vi skapar en tillståndsvektor $\overrightarrow{T} = (T_1, T_2, ..., T_n)$ som anger vilka tillstånd som besöks i Markovkedjan, och där n är storleken på urvalet. T_k motsvarar den k:te deltagaren i urvalet, och alltså motsvarar T_1 fröet i undersökningen. Fröet väljs slumpmässigt ur hela populationen.

Låt fröet vara individ $i \in \{1, ..., N\}$ i undersökningen. Detta innebär att vi befinner oss på rad i i övergångsmatrisen $\mathbf{C}$. Låt $\mathbf{C}_{i\cdot}$ vara rad i i $\mathbf{C}$ och låt $\mathbf{C}^*_{i\cdot}$ vara vektorn vars element är den kumulativa summan av elementena i $\mathbf{C}_{i\cdot}$. Vi påminner om att element $\mathbf{C}_{ij}$ är sannolikheten för övergång från tillstånd i till tillstånd j, såsom beskrevs i Avsnitt 2.2. För att bestämma tillstånd T_2 genereras ett slumptal $r \in$ U(0,1). Beroende på vilket intervall i den kumulativa summan av rad i i $\mathbf{C}$ r hamnar bestäms T_2. Om r hamnar mellan det $j-1$:e och det j:e elementet i $\mathbf{C}^*_{i\cdot}$, är $T_2 = j$. Processen upprepas för att bestämma T_3, T_4, till och med T_n, tills urvalet är klart.

För varje tillstånd $T_1, T_2, ..., T_n$ som besöks under Markovprocessen är det möjligt att kontrollera dess grad, dess styrka samt om tillståndet motsvarar en individ som är HIV-positiv eller ej. En RDS-skattning av andelen HIV-positiva i populationen görs enligt Ekvation (4).

5 Modeller och resultat

I detta avsnitt presenteras resultaten av simuleringarna. Vi kommer att utföra simuleringarna med varierande värden på parametern λ – den genomsnittliga graden för individerna i populationen, samt varierande värden på p_0 (se ekvationerna (6) och (7)). Dock kommer parametervärdena $N = 1000$ och $n = 500$ användas genomgående. Olika fördelningar för vikterna kommer också att testas. Tabell 4 illustrerar en sammanfattning av modellerna som används i studien.

Tabell 4: Sammanfattning av modellerna

Modell	Viktfördelning	Allokering av HIV	Parametervariationer
1	U(m)	Proportionellt mot grad	$m = 1, 2, 5, 10$ $\lambda = 10, 15$ $p_0 = 0.15, 0.50$
2	U(m)	Proportionellt mot styrka	
3	Exp(μ)	Proportionellt mot grad	$\mu = 1/2, 1$ $\lambda = 10, 15$ $p_0 = 0.15, 0.50$
4	Exp(μ)	Proportionellt mot styrka	
5	U(0,1)	Proportionellt mot grad	$\lambda = 10, 15$ $p_0 = 0.15, 0.50$
6	U(0,1)	Proportionellt mot styrka	

Figurerna 7-12 illustrerar i form av låddiagram avvikelsen för RDS-skattningarna från den sanna andelen HIV-positiva i populationen. Mera omfattande låddiagram återfinns i Bilaga A.

5.1 Modell 1

I den första modellen antar vi att vikterna är utfall av en U(m)-fördelning. Dessutom antar vi att sannolikheten att vara HIV-positiv beror på graden hos individen enligt Ekvation (6). Parametervärdena som vi använder i simuleringarna är $m = 1$, 2, 5 och 10; $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50. Notera att parametervärdet $m = 1$ motsvarar det oviktade fallet. Vi utför 100 stycken simuleringar av rekryteringsprocessen för alla kombinationer av m, λ och p_0, där varje enskild simulering kommer att generera ett socialt nätverk av 1000 individer med en stokastisk sann andel HIV-positiva individer. I Tabell 5 presenteras den genomsnittliga avvikelsen från det sanna värdet, det vill säga

$$\frac{1}{100}\sum_{i=1}^{100}\left(\hat{P}_{\text{HIV}}^{(i)} - P_{\text{HIV}}^{(i)}\right) \tag{8}$$

där $\hat{P}_{\text{HIV}}^{(i)}$ och $P_{\text{HIV}}^{(i)}$ är det i:te elementet i vektorn bestående av de 100 RDS-skattningarna respektive de 100 sanna värden. I Tabell 6 presenteras MSE (*Mean Square Error*) av $\hat{P}_{\text{HIV}}$, vilket beräknas som

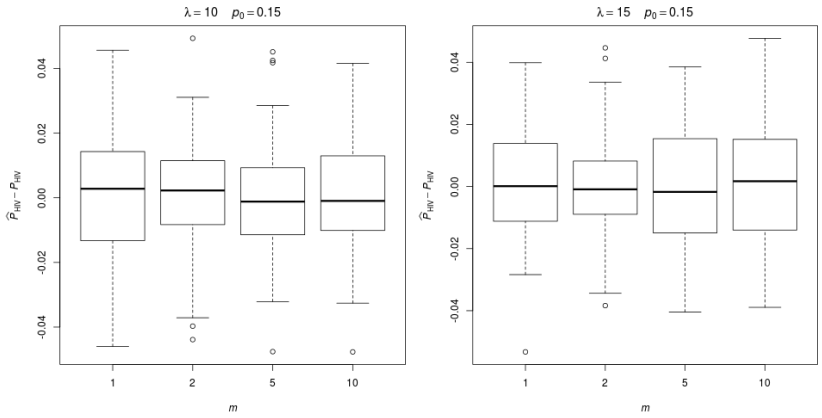
$$\frac{1}{100}\sum_{i=1}^{100}\left(\hat{P}_{\text{HIV}}^{(i)} - P_{\text{HIV}}^{(i)}\right)^2. \tag{9}$$

Tabell 5: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet för Modell 1.

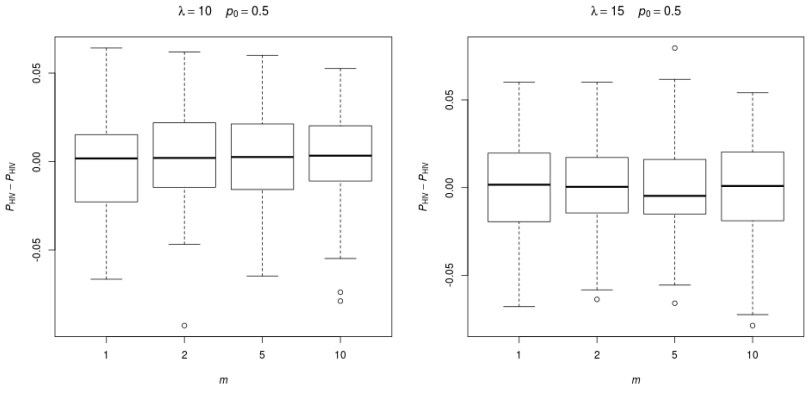
		m			
p_0	λ	1	2	5	10
0.15	10	$12.39 \cdot 10^{-4}$	$4.19 \cdot 10^{-4}$	$-9.92 \cdot 10^{-4}$	$3.35 \cdot 10^{-4}$
0.15	15	$11.04 \cdot 10^{-4}$	$2.20 \cdot 10^{-4}$	$-6.07 \cdot 10^{-4}$	$19.90 \cdot 10^{-4}$
0.50	10	$-17.57 \cdot 10^{-4}$	$31.78 \cdot 10^{-4}$	$21.44 \cdot 10^{-4}$	$22.36 \cdot 10^{-4}$
0.50	15	$3.98 \cdot 10^{-4}$	$17.14 \cdot 10^{-4}$	$-17.63 \cdot 10^{-4}$	$-15.94 \cdot 10^{-4}$

Tabell 6: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 1

		m			
p_0	λ	1	2	5	10
0.15	10	$3.26 \cdot 10^{-4}$	$2.96 \cdot 10^{-4}$	$2.87 \cdot 10^{-4}$	$2.95 \cdot 10^{-4}$
0.15	15	$2.80 \cdot 10^{-4}$	$2.38 \cdot 10^{-4}$	$3.20 \cdot 10^{-4}$	$3.94 \cdot 10^{-4}$
0.50	10	$6.58 \cdot 10^{-4}$	$7.91 \cdot 10^{-4}$	$7.67 \cdot 10^{-4}$	$6.63 \cdot 10^{-4}$
0.50	15	$6.87 \cdot 10^{-4}$	$6.52 \cdot 10^{-4}$	$6.26 \cdot 10^{-4}$	$7.86 \cdot 10^{-4}$



(a) $\lambda = 10 \quad p_0 = 0.15$ (b) $\lambda = 15 \quad p_0 = 0.15$



(c) $\lambda = 10 \quad p_0 = 0.50$ (d) $\lambda = 15 \quad p_0 = 0.50$

Figur 7: Låddiagram över $\hat{P}_{\text{HIV}} - P_{\text{HIV}}$, Modell 1.

De radvisa värdena i Tabell 5 håller sig i samma storleksordning för varierande värden på m. Man kan även i Tabell 6 observera att MSE varierar extremt lite och osystematiskt för samtliga permutationer av λ och p_0. Ur tabellerna 5 och 6 anses man inte kunna urskilja något systematisk fel i de fall där $m \neq 1$. Detta bekräftas även genom att observera låddiagrammen för Modell 1 i Figur 7. Boxplottarna för RDS-skattningarna håller sig till synes på samma nivå, kring värdet noll, oberoende av parametern m.

5.2 Modell 2

Även i den andra modellen antas vikterna vara utfall av en U(m)-fördelning. Däremot poneras att sannolikheten att vara HIV-positiv beror på styrkan hos individen enligt Ekvation (7), där vi således får att

$$\mathrm{E}[S_i] = \mathrm{E}[D_i] \cdot \mathrm{E}[V] = \frac{\lambda(m+1)}{2}.$$

På samma sätt som i Modell 1 används parametervärdena $m = 1$, 2, 5 och 10; $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50, och även här motsvarar parametervärdet $m = 1$ det oviktade fallet. Vi utför 100 stycken simuleringar av rekryteringsprocessen för alla kombinationer av m, λ och p_0. I Tabell 7 presenteras den genomsnittliga avvikelsen från det sanna värdet som beräknas enligt Ekvation (8), medan i Tabell 8 presenteras MSE av $\hat{P}_{\text{HIV}}$, som beräknas enligt Ekvation (9).

Tabell 7: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet i Modell 2.

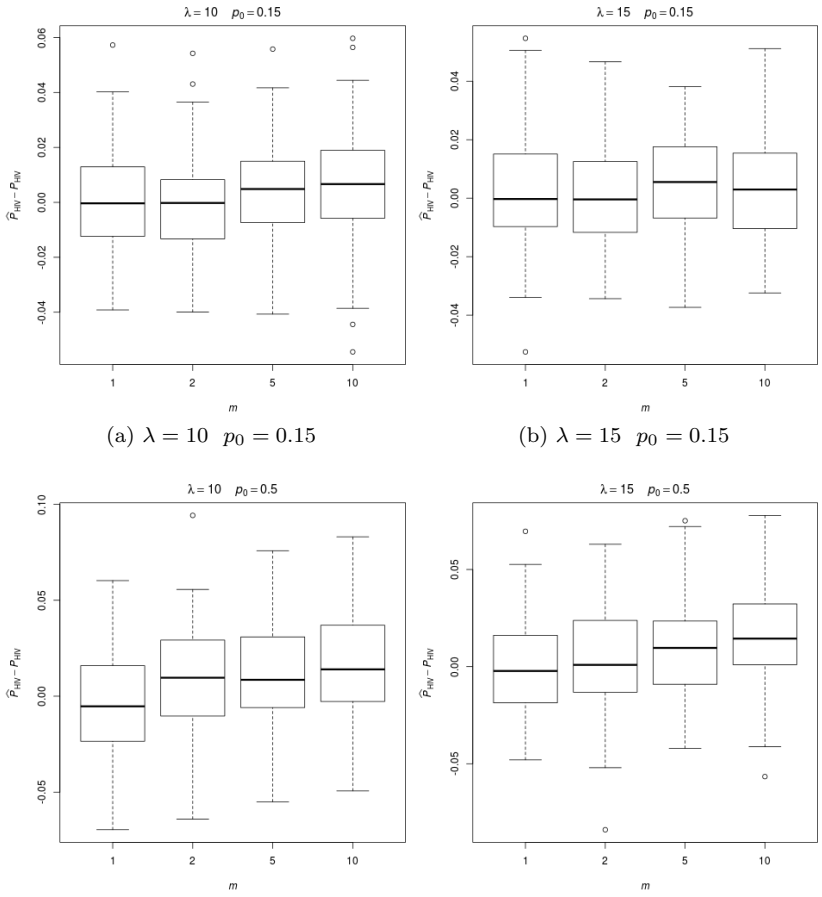
		m			
p_0	λ	1	2	5	10
0.15	10	$6.01 \cdot 10^{-4}$	$-19.90 \cdot 10^{-4}$	$40.91 \cdot 10^{-4}$	$59.16 \cdot 10^{-4}$
0.15	15	$20.80 \cdot 10^{-4}$	$6.99 \cdot 10^{-4}$	$51.07 \cdot 10^{-4}$	$28.95 \cdot 10^{-4}$
0.50	10	$-38.99 \cdot 10^{-4}$	$89.29 \cdot 10^{-4}$	$108.51 \cdot 10^{-4}$	$165.55 \cdot 10^{-4}$
0.50	15	$-8.58 \cdot 10^{-4}$	$4.47 \cdot 10^{-4}$	$64.39 \cdot 10^{-4}$	$140.75 \cdot 10^{-4}$

Tabell 8: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 2.

		m			
p_0	λ	1	2	5	10
0.15	10	$3.46 \cdot 10^{-4}$	$3.32 \cdot 10^{-4}$	$3.30 \cdot 10^{-4}$	$4.29 \cdot 10^{-4}$
0.15	15	$3.35 \cdot 10^{-4}$	$3.09 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$	$3.08 \cdot 10^{-4}$
0.50	10	$7.60 \cdot 10^{-4}$	$7.67 \cdot 10^{-4}$	$7.62 \cdot 10^{-4}$	$10.27 \cdot 10^{-4}$
0.50	15	$5.79 \cdot 10^{-4}$	$5.69 \cdot 10^{-4}$	$7.50 \cdot 10^{-4}$	$8.34 \cdot 10^{-4}$

I de fall där $p_0 = 0.50$ (figurerna 8c och 8d), kan en måttlig ökning av $\hat{P}_{\text{HIV}} - P_{\text{HIV}}$ observeras för ökande värde på m. Detta är ett tecken på att individer med högre styrka – som även har högre sannolikhet att vara HIV-positiva – är överrepresenterade i urvalet, och på så sätt har skapat ett systematiskt fel i RDS-skattningen. Detta stärks av det som kan observeras i Tabell 7, där den genomsnittliga avvikelsen mellan RDS-skattningen och det sanna värdet ökar för ökande värde på m i de fall där $p_0 = 0.50$. Detta skulle innebära att RDS överskattar andelen HIV-positiva i populationen givet Modell 2.

Om man istället observerar figurerna 8a och 8b kan man inte se denna effekt lika tydligt då $p_0 = 0.15$. Man kan ur Tabell 7 avläsa att den genomsnittliga skillnaden mellan RDS-skattningen och det sanna värdet möjligtvis skulle kunna öka för ökande värde på m, men inget definitivt uttalande kan göras. En kvalificerad gissning är att RDS-skattningen överskattar andelen HIV-positiva i populationen även i de fall där $p_0 = 0.15$, fastän effekten inte syns så tydligt i dessa fall.



(a) $\lambda = 10 \quad p_0 = 0.15$ (b) $\lambda = 15 \quad p_0 = 0.15$

(c) $\lambda = 10 \quad p_0 = 0.50$ (d) $\lambda = 15 \quad p_0 = 0.50$

Figur 8: Låddiagram över RDS-skattningen minus det sanna värdet, Modell 2.

5.3 Modell 3

I denna modell antar vi att $V \sim \text{Exp}(\mu)$, samt att sannolikheten att vara HIV-positiv beror på graden hos individen enligt Ekvation (6). Vi undersöker hur RDS-skattningen påverkas med parametervärdena $\mu = 1/2$ och 1; $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50. Vi utför 100 stycken simuleringar av rekryteringsprocessen för alla kombinationer av μ, λ och p_0. I Tabell 9 presenteras den genomsnittliga avvikelsen från det sanna värdet som beräknas enligt Ekvation (8), medan i Tabell 10 presenteras MSE av $\hat{P}_{\text{HIV}}$, som beräknas enligt Ekvation (9).

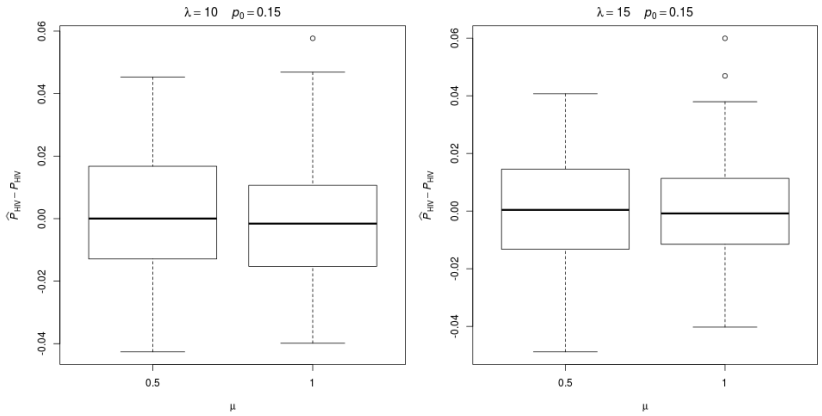
Tabell 9: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet i Modell 3.

		μ	
p_0	λ	$1/2$	1
0.15	10	$16.47 \cdot 10^{-4}$	$-20.79 \cdot 10^{-4}$
0.15	15	$-9.71 \cdot 10^{-4}$	$3.72 \cdot 10^{-4}$
0.50	10	$-1.39 \cdot 10^{-4}$	$-4.00 \cdot 10^{-4}$
0.50	15	$24.76 \cdot 10^{-4}$	$23.06 \cdot 10^{-4}$

Tabell 10: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 3.

		μ	
p_0	λ	$1/2$	1
0.15	10	$4.05 \cdot 10^{-4}$	$3.35 \cdot 10^{-4}$
0.15	15	$3.71 \cdot 10^{-4}$	$3.17 \cdot 10^{-4}$
0.50	10	$9.32 \cdot 10^{-4}$	$9.68 \cdot 10^{-4}$
0.50	15	$7.69 \cdot 10^{-4}$	$5.43 \cdot 10^{-4}$

Enligt de erhållna resultaten i Modell 3 verkar RDS-skattningarna inte rubbas avsevärt. Ur samtliga låddiagram (Figur 9) kan man se att mätvärdena samlas kring noll. Man kan dessutom ur Tabell 9 observera att den genomsnittliga skillnaden mellan RDS-skattningen och det sanna värdet är liten för alla kombinationer av λ, μ och p_0. Man kan i detta fall alltså dra samma slutsatser som för Modell 1 – nämligen att de viktade vänskapsbandens effekt är tämligen liten.



(a) $\lambda = 10 \quad p_0 = 0.15$ (b) $\lambda = 15 \quad p_0 = 0.15$

(c) $\lambda = 10 \quad p_0 = 0.50$ (d) $\lambda = 15 \quad p_0 = 0.50$

Figur 9: Låddiagram över RDS-skattningen minus det sanna värdet, Modell 3.

5.4 Modell 4

Vi antar en identisk modell som Modell 3, förutom att sannolikheten att vara HIV-positiv beror på styrkan hos individen enligt Ekvation (7). $E[S_i]$ i Ekvation (7) blir således

$$\mathrm{E}[S_i] = \mathrm{E}[D_i] \cdot \mathrm{E}[V] = \lambda \cdot \frac{1}{\mu} = \frac{\lambda}{\mu}.$$

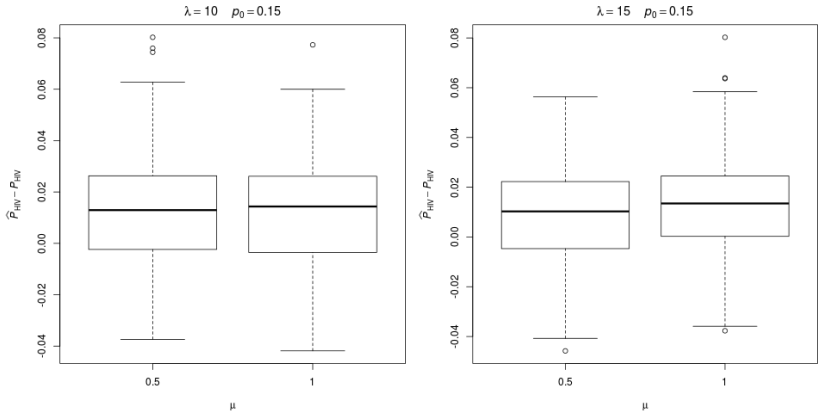
På samma sätt som i Modell 3 undersöker vi för parametervärdena $\mu = 1/2$ och 1; $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50. Den genomsnittliga avvikelsen av RDS-skattningen från det sanna värdet samt MSE av $\hat{P}_{\mathrm{HIV}}$ presenteras i tabellerna 11 respektive 12.

Tabell 11: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet i Modell 4.

		μ	
p_0	λ	$1/2$	1
0.15	10	$1.29 \cdot 10^{-2}$	$1.29 \cdot 10^{-2}$
0.15	15	$0.81 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$
0.50	10	$4.55 \cdot 10^{-2}$	$4.80 \cdot 10^{-2}$
0.50	15	$3.55 \cdot 10^{-2}$	$3.15 \cdot 10^{-2}$

Tabell 12: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 4.

		μ	
p_0	λ	$1/2$	1
0.15	10	$6.71 \cdot 10^{-4}$	$6.74 \cdot 10^{-4}$
0.15	15	$4.45 \cdot 10^{-4}$	$6.09 \cdot 10^{-4}$
0.50	10	$29.19 \cdot 10^{-4}$	$34.43 \cdot 10^{-4}$
0.50	15	$19.03 \cdot 10^{-4}$	$18.47 \cdot 10^{-4}$



(a) $\lambda = 10 \quad p_0 = 0.15$ (b) $\lambda = 15 \quad p_0 = 0.15$

(c) $\lambda = 10 \quad p_0 = 0.50$ (d) $\lambda = 15 \quad p_0 = 0.50$

Figur 10: Låddiagram över RDS-skattningen minus det sanna värdet, Modell 4.

I Modell 4 kan man i låddiagrammen se att RDS-skattningarna överskattar andelen HIV-positiva i de flesta av de 100 simuleringarna. Detta syns tydligast i de fall där p_0 är satt till 0.5 (figurerna 10c och 10d), där vi kan observera att samtliga mätvärden mellan den första och den tredje kvartilen i boxplottarna ligger ovanför värdet noll. Även om effekten inte är så stark då $p_0 = 0.15$ (figurerna 10a och 10b) ser vi även i dessa fall att den absolut största delen av mätvärdena ligger ovanför värdet noll. Våra observationer stämmer också överens med den genomsnittliga skillnaden mellan RDS-skattningarna och det sanna värdet (Tabell 11), där samtliga värden är större än noll – och dessutom till beloppet större än i simuleringarna där väskapsbanden är oviktade (tabellerna 5 och 7, $m = 1$).

Hur mycket RDS-skattningarna påverkas av viktade vänskapsband tycks bero på parametrarna λ, p_0 och μ, samt deras eventuella samverkan. En djupare analys av hur valet av parametrarna påverkar resultaten utelämnas från denna studie, men slutsatsen som dras är att RDS överskattar andelen HIV-positiva individer i populationen givet Modell 4.

5.5 Modell 5

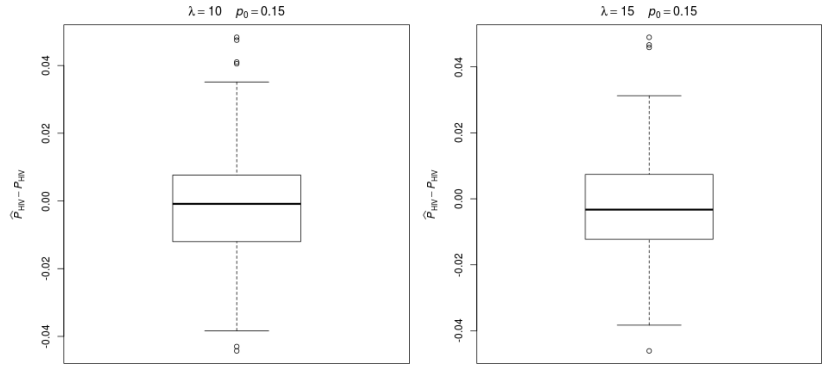
I den femte modellen sätter vi vikterna till utfall av en U(0,1)-fördelning. Vi ponerar att sannolikheten att vara HIV-positiv beror på graden hos individen enligt Ekvation (6). Som vanligt undersöker vi hur RDS-skattningen påverkas med parametervärdena $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50. Vi utför 100 stycken simuleringar av rekryteringsprocessen för alla kombinationer av λ och p_0. I Tabell 13 presenteras den genomsnittliga avvikelsen från det sanna värdet som beräknas enligt Ekvation (8), medan i Tabell 14 presenteras MSE av $\hat{P}_{\text{HIV}}$, som beräknas enligt Ekvation (9).

Tabell 13: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet i Modell 5.

p_0	λ	$\overline{\hat{P}_{\text{HIV}} - P_{\text{HIV}}}$
0.15	10	$-8.89 \cdot 10^{-4}$
0.15	15	$-19.38 \cdot 10^{-4}$
0.50	10	$35.18 \cdot 10^{-4}$
0.50	15	$9.16 \cdot 10^{-4}$

Tabell 14: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 5.

p_0	λ	$\text{MSE}(\hat{P}_{\text{HIV}})$
0.15	10	$3.43 \cdot 10^{-4}$
0.15	15	$2.82 \cdot 10^{-4}$
0.50	10	$6.42 \cdot 10^{-4}$
0.50	15	$6.39 \cdot 10^{-4}$



(a) $\lambda = 10 \quad p_0 = 0.15$ (b) $\lambda = 15 \quad p_0 = 0.15$

(c) $\lambda = 10 \quad p_0 = 0.50$ (d) $\lambda = 15 \quad p_0 = 0.50$

Figur 11: Låddiagram över RDS-skattningen minus det sanna värdet, Modell 5.

Såsom för de tidigare nämnda modeller där sannolikheten att vara HIV-positiv beror på graden hos individen enligt Ekvation (6), kan man inte observera stora avvikelser för RDS-skattningarna. Samtliga mätvärden (Figur 11) verkar vara symmetriskt samlade kring noll, och inga större avvikelser kan noteras i tabellerna 13 och 14.

5.6 Modell 6

I den sjätte och sista modellen sätter vi vikterna till utfall av en U(0,1)-fördelning. Vi ansätter sannolikheten att vara HIV-positiv beror på styrkan hos individen enligt Ekvation (7), där vi i detta fall får

$$\mathrm{E}[S_i] = \mathrm{E}[D_i] \cdot \mathrm{E}[V] = \lambda \cdot \frac{1}{2} = \frac{\lambda}{2}.$$

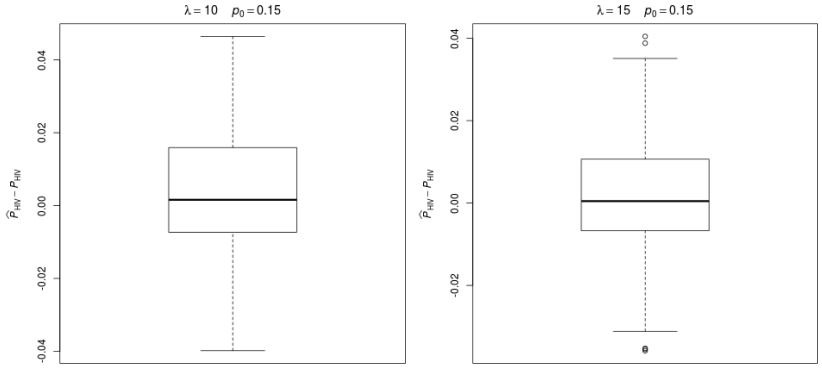
Som vanligt undersöker vi hur RDS-skattningen påverkas med parametervärdena $\lambda = 10$ och 15 samt $p_0 = 0.15$ och 0.50. Vi utför 100 stycken simuleringar av rekryteringsprocessen för alla kombinationer av λ och p_0. I Tabell 15 presenteras den genomsnittliga avvikelsen från det sanna värdet som beräknas enligt Ekvation (8), medan i Tabell 16 presenteras MSE av $\hat{P}_{\mathrm{HIV}}$, som beräknas enligt Ekvation (9).

Tabell 15: Genomsnittlig avvikelse av RDS-skattningen från det sanna värdet i Modell 6.

p_0	λ	$\overline{\hat{P}_{\text{HIV}} - P_{\text{HIV}}}$
0.15	10	$34.58 \cdot 10^{-4}$
0.15	15	$8.53 \cdot 10^{-4}$
0.50	10	$180.59 \cdot 10^{-4}$
0.50	15	$116.91 \cdot 10^{-4}$

Tabell 16: MSE av RDS-skattning $\hat{P}_{\text{HIV}}$ i Modell 6.

p_0	λ	$\text{MSE}(\hat{P}_{\text{HIV}})$
0.15	10	$3.56 \cdot 10^{-4}$
0.15	15	$2.60 \cdot 10^{-4}$
0.50	10	$10.76 \cdot 10^{-4}$
0.50	15	$7.64 \cdot 10^{-4}$



(a) $\lambda = 10 \;\; p_0 = 0.15$

(b) $\lambda = 15 \;\; p_0 = 0.15$

(c) $\lambda = 10 \;\; p_0 = 0.50$

(d) $\lambda = 15 \;\; p_0 = 0.50$

Figur 12: Låddiagram över RDS-skattningen minus det sanna värdet, Modell 6.

I Figur 12 kan vi se att RDS-skattningen tenderar att vara högre än det sanna värdet. Detta syns tydligast i de fall där p_0 är ansatt till 0.5, men även i en viss grad i de fall där $p_0 = 0.15$. Detta stärks av det man kan observera i Tabell 15, där samtliga värden är större än noll, och jämförelsevis stora i de fall där $p_0 = 0.50$. Notera speciellt hur mycket storleken för den genomsnittliga skillnaden varierar för varierande värden på λ och p_0.

Slutsatsen i Modell 6 är att RDS överskattar andelen HIV-positiva i populationen, och storleken på det systematiska felet är rätt så beroende på valet av parametrarna λ och p_0.

6 Slutsatser

En grafisk undersökning av resultaten av simuleringarna tyder på att RDS med viktade vänskapsband överskattar andelen HIV positiva i populationen givet att sannolikheten att vara HIV-positiv beror på styrkan hos individen enligt Ekvation (7). Hur mycket RDS-skattningen påverkas beror på den genomsnittliga graden hos individerna i populationen (λ), samt den sanna andelen HIV-smittade i populationen.

RDS med viktade vänskapsband verkar inte felskatta avsevärt andelen HIV-positiva ifall sannolikheten att vara HIV-positiv beror på graden enligt Ekvation (6).

7 Diskussion

I ett viktat nätverk är det större sannolikhet för en individ med hög styrka att delta i undersökningen. Detta motiveras med följande tankegång: Antag matrisen $\mathbf{B}$ som beskrevs i Avsnitt 4.1, och antag att summan av elementena på rad $k \in \{1, ..., N\}$ är mycket stor jämfört med summan av elementena på de andra raderna, trots att graden inte skiljer sig mycket åt. Således är styrkan för individ k stor medan graden är i samma storleksordning som för resten av individerna i populationen. Eftersom matrisen $\mathbf{B}$ är symmetrisk kring diagonalen (på grund av reciprocitetsantagandet, se Avsnitt 2.3), så är även elementena på kolumn k jämförelsevis stora. När man sedan normaliserar $\mathbf{B}$ genom att dela elementena med radsumman, försvinner de stora vikternas effekt på rad k. Däremot förblir elementena på kolumn k stora. Således är det större sannolikhet att övergå till tillstånd k från alla andra tillstånd.

Denna motivering kan naturligtvis tillämpas på fall där flera rader har i jämförelse stora viktsummor. Detta innebär att det är större sannolikhet att övergå till just dessa tillstånd – som alltså kommer att bli överrepresenterade i urvalet.

Om sannolikheten att vara HIV-positiv beror på graden hos individen, kommer det inte spela någon roll om den k:te raden har en jämförelsevis stor viktsumma. De individer som har stor viktsumma kommer ändå att tilldelas HIV proportionellt mot deras grad, som vi antog var i samma storleksordning som för resten av populationen. Det har således inte så stor betydelse om individer med stor

viktsumma blir översamplade, eftersom de ändå utgör ett representativt urval av hela populationen i detta fall. Detta leder till att RDS-skattningarna till synes inte rubbas av viktade vänskapsband i de fall där sannolikheten att vara HIV-positiv beror på graden.

Om man däremot ponerar att sannolikheten att en individ är HIV-positiv beror på styrkan så att ju högre styrka, desto högre sannolikhet att vara HIV-positiv, kommer det faktum av individer med hög styrka är överrepresenterade definitivt skapa ett systematiskt fel i RDS-skattningarna.

Denna tankegång förklarar givetvis inte hela verkligheten, utan är snarare förklaringen till en tendens. Det är i våra modeller inte så sannolikt att vissa individer har en hög styrka, trots att deras grad är i samma storleksordning som för resten av populationen. Därför kommer effekten i den ovannämnda tankegången att dämpas av att det finns en positiv korrelation mellan vikt och styrka – en individ med *hög styrka* har med stor sannolikhet även en *hög grad*.

Vi har i denna uppsats inte gått så djupt in på huruvida de nämnda modellerna är verklighetsbaserade, utan tanken med denna studie har varit att undersöka om RDS fungerar under extremare modeller än de som tidigare antagits (se Avsnitt 2.3). Förslag till framtida forskning om ämnet är hur viktade vänskapsband påverkar RDS-skattningarna ur ett teoretiskt perspektiv, med utgångspunkt i teorin för Markovkedjor i diskret tid och deras asymptotiska fördelningar, möjligtvis för att kunna teoretiskt erhålla väntevärdesriktiga skattare med färre antaganden än de som i dagsläget behövs.

Referenser

[1] Ralph P. Grimaldi (2004), Discrete and Combinatorial Mathematics – An applied introduction, fifth edition, *Pearson education (us)*

[2] Allan Gut, An Intermediate Course in Probability, Second Edition, *Springer*, 2009

[3] Mark S. Handcock, Krista J. Gile, (2011). *Comment: On the concept of "snowball sampling"*, Sociological Methodology, Vol. 41, Issue 1, pp. 367–371

[4] Heckathorn, D. D. (1997). *Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.* Social problems, pp. 174–199

[5] Anni Pilbacka (2010). *Inferens på sociala nätverk i "gömda" populationer*, Kandidatuppsats 2010:2, Matematisk statistik, Stockholms universitet

[6] Sheldon M. Ross (2007). Introduction to Probability Models, 9^{th} edition, *Academic Press Inc*

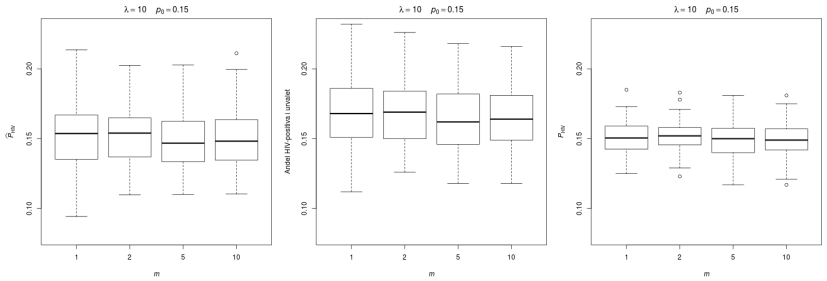
[7] Carl-Erik Särndal, Bengt Swensson, Jan Wretman (2003). *Model Assisted Survey Sampling*, Springer-Verlag New York Inc.

[8] Erik Volz and Douglas D. Heckathorn (2008). *Probability Based Estimation Theory for Respondent Driven Sampling*, Journal of Official Statistics, Vol 24, No. 1, pp. 79–97

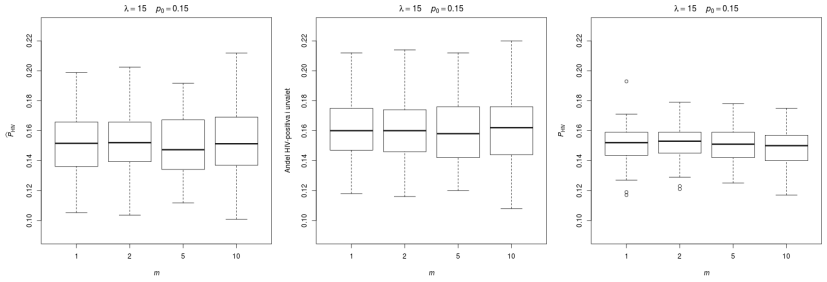
[9] *What is Respondent Driven Sampling?,* www.respondentdrivensampling.org 1.4.2013

A Låddiagram

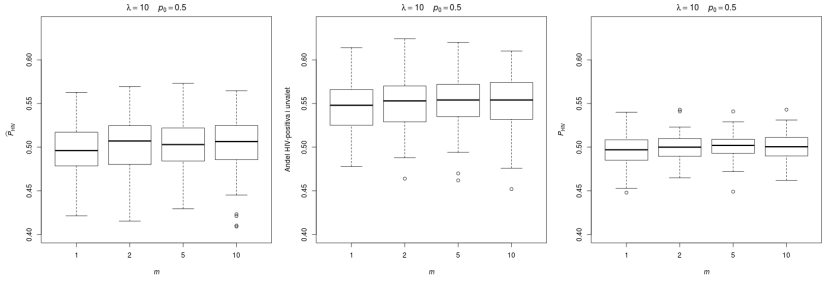
A.1 Modell 1



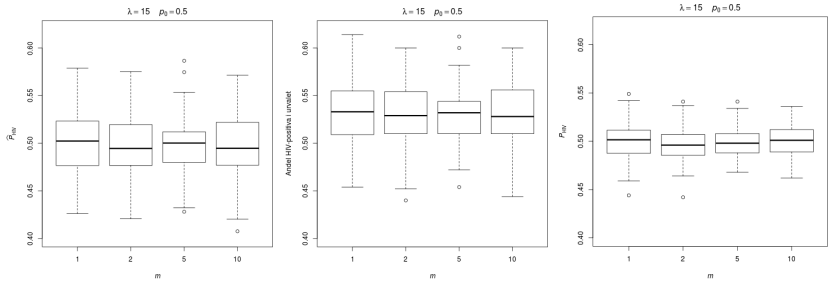
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



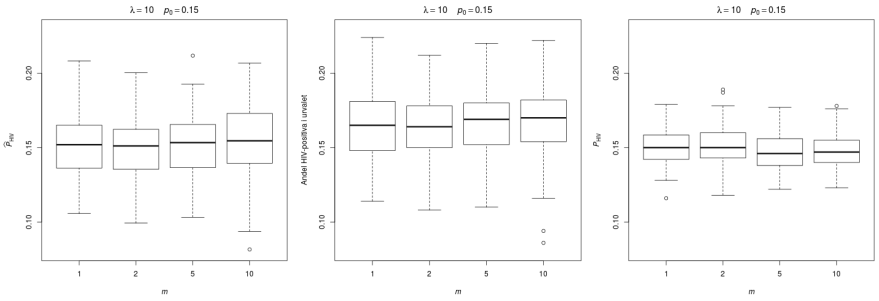
$\lambda = 10$ och $p_0 = 0.50$



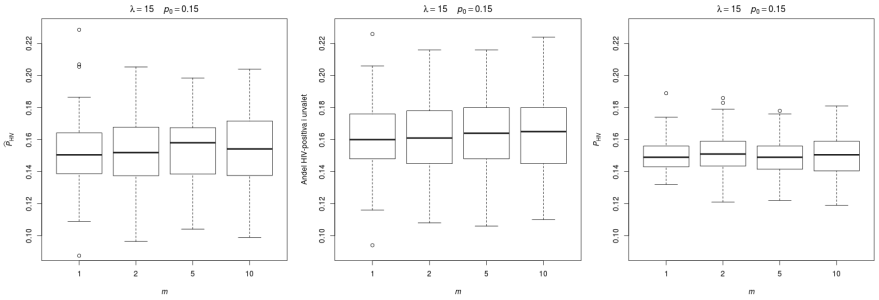
$\lambda = 15$ och $p_0 = 0.50$

Figur A1: Låddiagram över resultat av simuleringarna med Modell 1. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

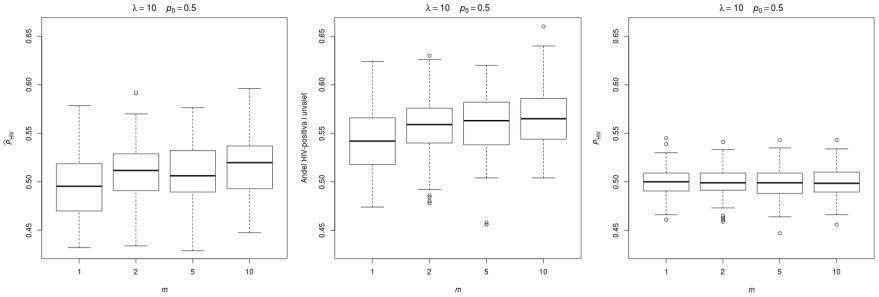
A.2 Modell 2



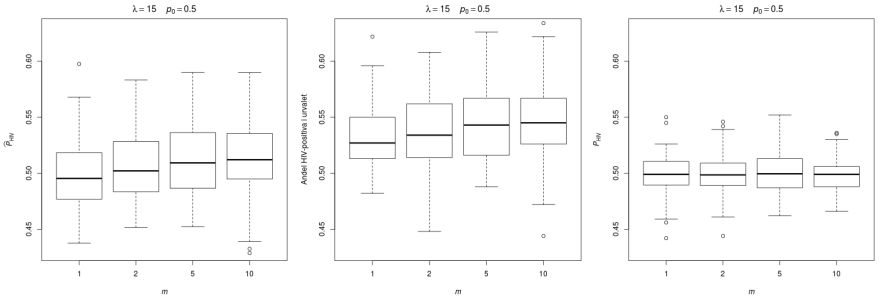
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



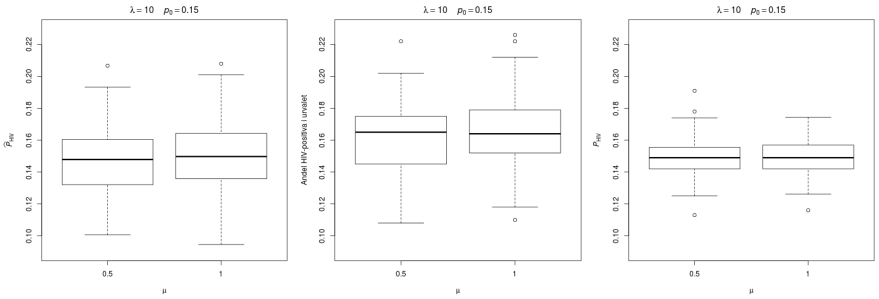
$\lambda = 10$ och $p_0 = 0.50$



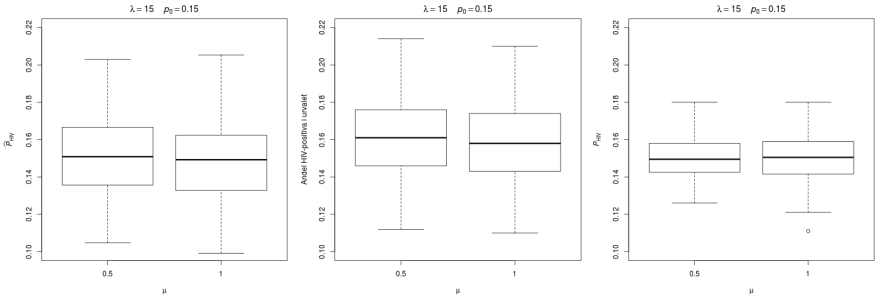
$\lambda = 15$ och $p_0 = 0.50$

Figur A2: Låddiagram över resultat av simuleringarna med Modell 2. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

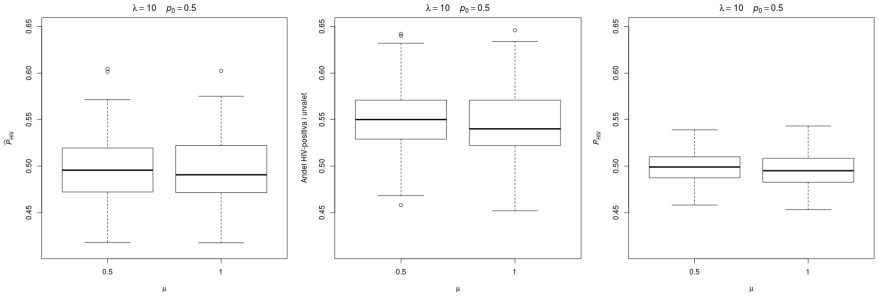
A.3 Modell 3



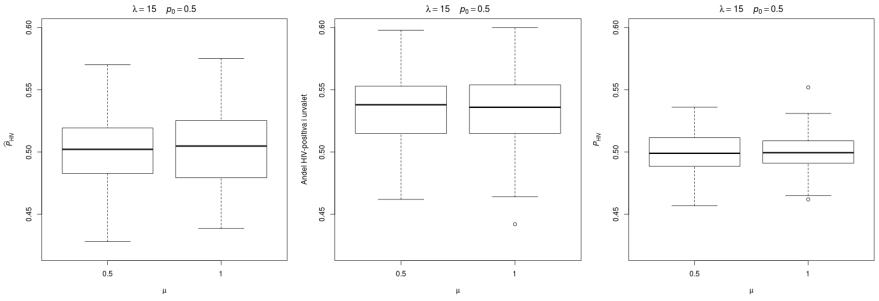
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



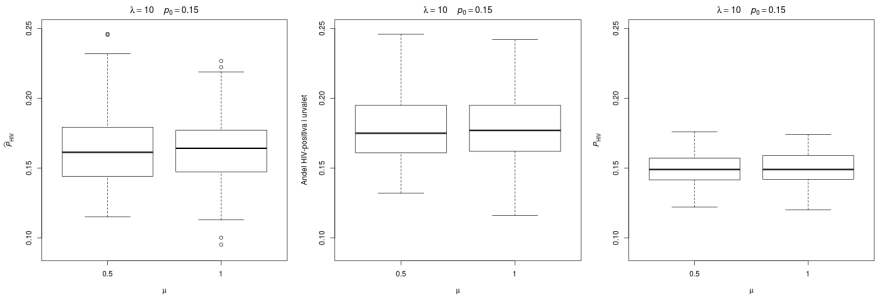
$\lambda = 10$ och $p_0 = 0.50$



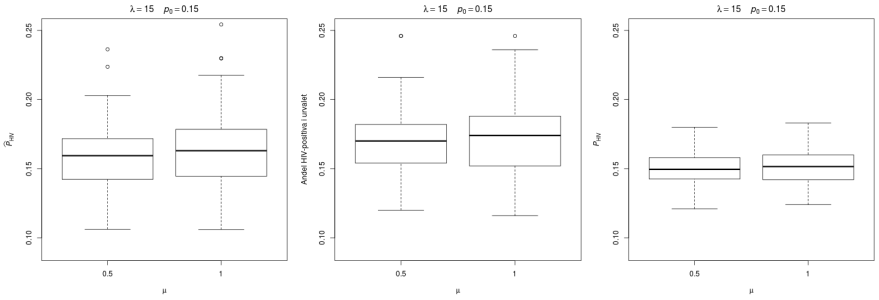
$\lambda = 15$ och $p_0 = 0.50$

Figur A3: Låddiagram över resultat av simuleringarna med Modell 3. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

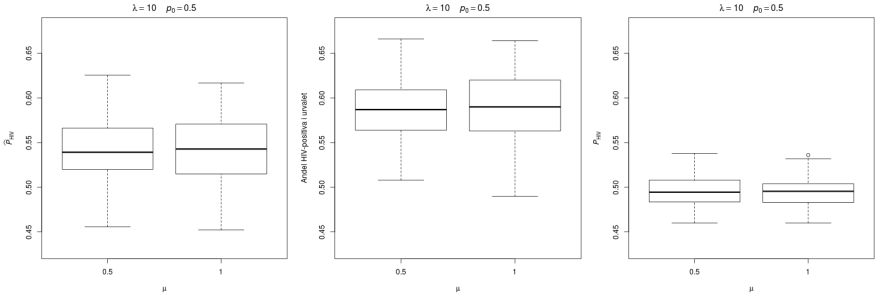
A.4 Modell 4



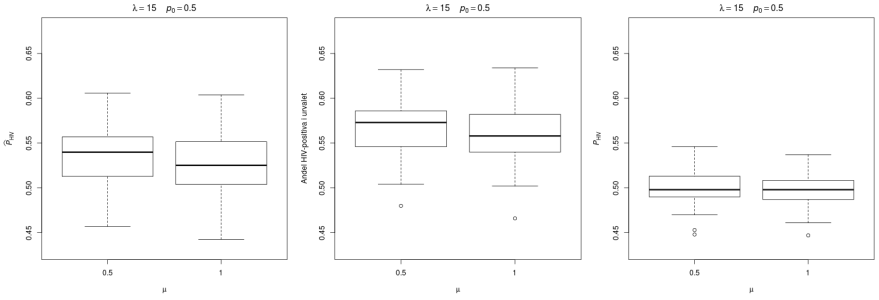
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



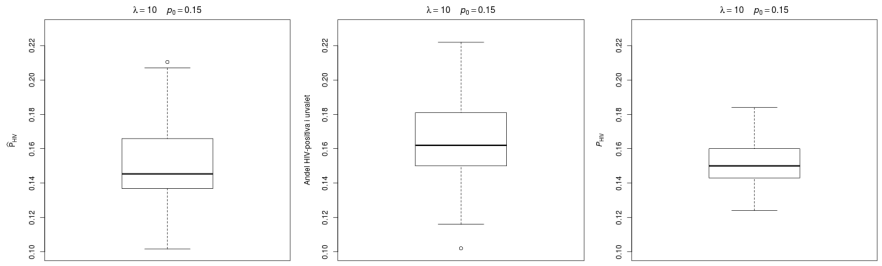
$\lambda = 10$ och $p_0 = 0.50$



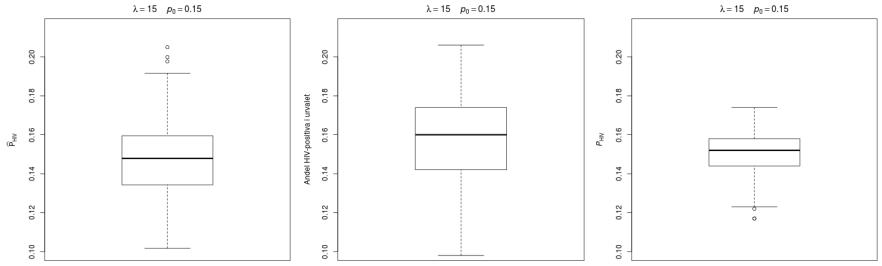
$\lambda = 15$ och $p_0 = 0.50$

Figur A4: Låddiagram över resultat av simuleringarna med Modell 4. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

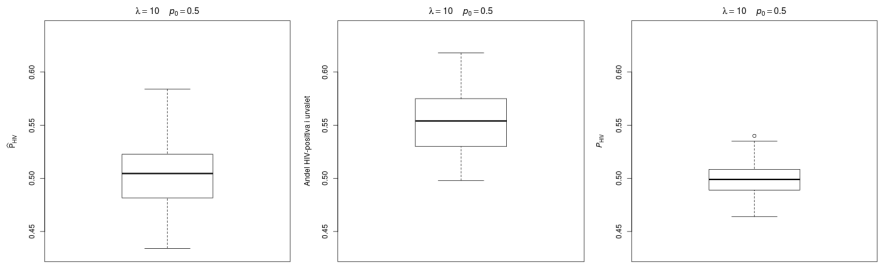
A.5 Modell 5



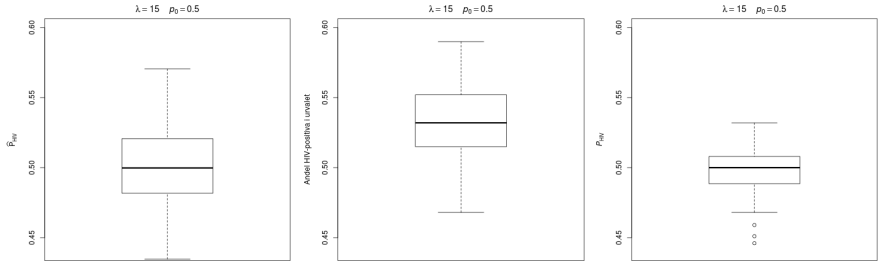
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



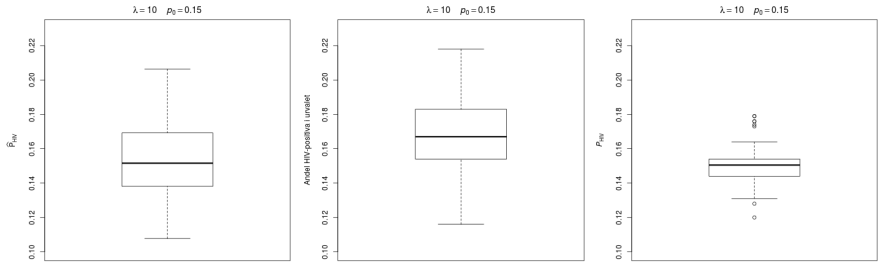
$\lambda = 10$ och $p_0 = 0.50$



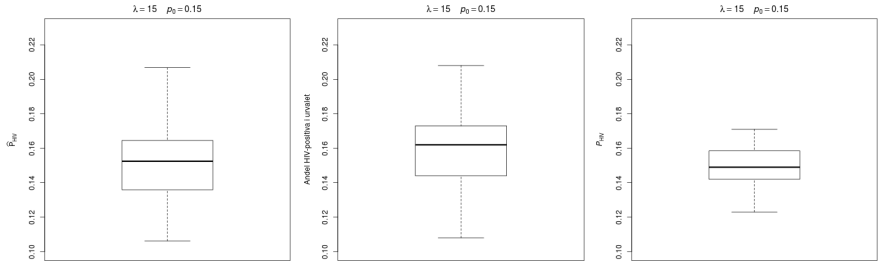
$\lambda = 15$ och $p_0 = 0.50$

Figur A5: Låddiagram över resultat av simuleringarna med Modell 5. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

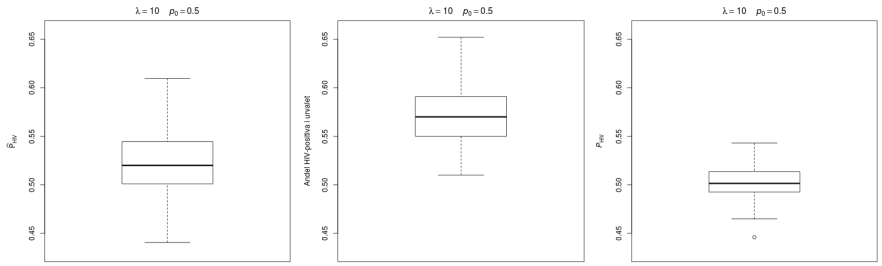
A.6 Modell 6



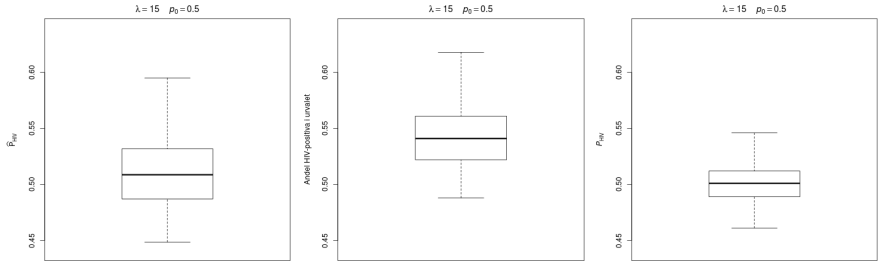
$\lambda = 10$ och $p_0 = 0.15$



$\lambda = 15$ och $p_0 = 0.15$



$\lambda = 10$ och $p_0 = 0.50$



$\lambda = 15$ och $p_0 = 0.50$

Figur A6: Låddiagram över resultat av simuleringarna med Modell 6. Den vänstra kolumnen visar RDS-skattningarna, den mittersta kolumnen visar andelen HIV-positiva i urvalet och den högra kolumnen visar sanna andelen HIV-positiva i populationen.

B Fördelningar

Tabell B1: Lista över fördelningar som används i studien. Definitionerna görs i enlighet med [2], förutom för den diskret likformiga fördelningen, där informationen är hämtad från
http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29 20.5.2013 samt
http://www.proofwiki.org/wiki/Probability_Generating_Function_of_Discrete_Uniform_Distribution 20.5.2013 (sannolikhetsgenererande funktion). En asterisk (*) innebär att en sannolikhetsgenererande funktion inte existerar.

Fördelning	Notation	Sannolikhetsfunktion/Densitet	E[X]	Var[X]	$g_X(t)$
Bernoulli	Be(p), $0 \leq p \leq 1$	$p(0) = q, \;\; p(1) = p,\ q = 1 - p$	p	pq	$q + pt$
Binomial	Bin(n, p), $n = 1, 2, ...;\ 0 \leq p \leq 1$	$p(k) = \binom{n}{k} p^k (q)^{n-k}, k = 0, 1, ...,\ q = 1 - p$	np	npq	$(q + pt)^n$
Poisson	Po(m), $m > 0$	$p(k) = e^{-m} \frac{m^k}{k!},\ k = 0, 1, 2, ...$	m	m	$e^{m(t-1)}$
Diskret likformig	U(m)	$p(k) = \frac{1}{m},\ k = 1, 2, ..., m$	$\frac{m+1}{2}$	$\frac{m^2-1}{12}$	$\frac{t}{m}\left(\frac{1-t^m}{1-t}\right)$
Likformig	U(a, b)	$f(x) = \frac{1}{b-a},\ a < x < b$	$\frac{1}{2}(a + b)$	$\frac{1}{12}(b - a)^2$	*
Exponential	Exp(a), $a > 0$	$f(x) = \frac{1}{a} e^{-x/a},\ x > 0$	a	a^2	*