Stockholms universitet

# Inferens på sociala nätverk i "gömda" populationer

Anni Pilbacka

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:2
Matematisk statistik
April 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:2**,
http://www.math.su.se/matstat

# Inferens på sociala nätverk i ”gömda” populationer

Anni Pilbacka*

April 2010

**Sammanfattning**

Populationer där man inte känner till sannolikheten för att välja en viss individ till stickprovet, och man därmed inte kan tillämpa vanlig stickprovs- och skattningsteknik, kallas för gömda populationer. I den här uppsatsen ska vi simulera ett socialt nätverk av individer och göra ett stickprov på nätverket som vi ska tillämpa en relativt ny skattningsmetodik på - Respondent Driven Sampling (RDS). Metoden går ut på att individerna, som ingår i den population man är intresserad av att studera, väljs genom vänskapsbanden mellan de redan existerande medlemmarna i stickprovet. Vi skattar populationsandelar med RDS för ett antal olika modeller. Syftet är att se om skattningarna närmar sig den sanna andelen i populationen samt att se hur skattningens precision förbättras i takt med att urvalet växer. Vi kommer att visa att RDS fungerar bra när vänskapsbanden är ömsesidiga samt när endast enstaka vänskapsband som inte är ömsesidiga finns med. Vi kommer fram till att RDS fungerar mindre bra när merparten av vänskapsbanden inte är ömsesidiga.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: anni_pilbacka@hotmail.com. Handledare: Tom Britton.

# Inference on social networks in hidden populations


## Abstract

Populations where the probability of choosing a certain individual to a sample is unknown and where you thus cannot apply regular sampling and estimation technique are called hidden populations. In this paper we will simulate a social network of individuals and make a sample of the network that we will apply a relatively new estimation methodology on – Respondent-Driven Sampling (RDS). The method is based on that the individuals in the population that we are interested in studying recruit their friends to be included in the study. We will estimate the population proportions with RDS for a number of different models. The aim is to see if the estimates are approaching the true proportion of the population, and to see how the precision is improved as the sample is growing. We will show that RDS works well when the friendships are reciprocal and in the case when there are only a few friendships that are not reciprocal. We conclude that RDS works less good when the majority of friendships are not reciprocal.

## Förord

Denna uppsats utgör ett examensarbete om 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik vid Matematiska institutionen, Stockholms Universitet.

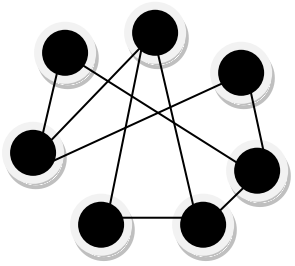
Ett stort tack riktas till professor Tom Britton, min handledare på Matematiska institutionen vid Stockholms Universitet, för vägledning och goda råd under arbetets gång.

# Innehåll

# 1 Introduktion

## 1.1 Inledning

Ett socialt nätverk beskriver relationer mellan människor som till exempel uppkommit genom släktskap, vänskap, bekantskap eller sexuella relationer. Man kan representera ett socialt nätverk med en graf där varje position i grafen symboliserar en plats i nätverket, vilket illustreras i figur 1.1.



**Figur 1.1.** Graf som symboliserar ett socialt nätverk.

Grafen kan representeras av en matris med nollor och ettor där en etta indikerar vänskapsband mellan individ *i* och individ *j*. Raderna (kolumnerna) representerar en individ och dess relationer till de andra individerna i nätverket. För att kunna göra en graf över ett socialt nätverk behöver man kartlägga vem som har kontakt med vem. Det kan vara av intresse att studera ett socialt nätverk av olika anledningar, ofta för att studera olika egenskaper hos individerna i den specifika population man är intresserad av.

Det är vanligt att man vill ta reda på andelen individer i populationen som har en viss egenskap, denna egenskap kan vara av olika karaktär. Vi ska i denna uppsats simulera ett socialt nätverk som vi ska ponera är drogmissbrukare. Vi är intresserade av att skatta andelen HIV-positiva i detta nätverk som då symboliserar till exempel missbrukare av injektionsdroger i en storstad. Vi antar att medlemmarna i denna population har sannolikheten $p_{HIV}$ att vara HIV-positiva. Vi ska slumpa ut denna sannolikhet på två olika sätt. Vi antar dels att individer är HIV-positiva oberoende och lika fördelat över hela populationen, dels antar vi att sannolikheten att en individ är HIV-positiv stiger i proportion med antalet vänner den har. Vi antar sedan att vi vill skatta andelen HIV-positiva.

Ur nätverket som vi skapat kommer vi att göra ett stickprov som vi ska tillämpa en relativt ny skattningsmetodik på, nämligen Respondent Driven Sampling (Salganik & Heckathorn, 2004). Respondent Driven Sampling (RDS) används för gömda populationer, det vill säga sådana populationer som inte har en urvalsram och man därmed inte kan dra ett vanligt

stickprov. Metoden grundas i att man tar tillvara på den information som man kan få ur ett socialt nätverk och det faktum att nätverket består av verkliga människor som på något sätt har ett samband med varandra.

I april 2009 fanns det totalt 128 pågående samt avslutade studier med Respondent Driven Sampling i 30 olika länder utanför staterna och i staterna har ett flertal studier utförts (Gile & Handcock, 2009). RDS används när man vill studera populationer som är svåra att urskilja ur befolkningen, exempelvis injektionsdrogmissbrukare, män som är gay och prostituerade (*What is Respondent Driven Sampling?,* Respondent Driven Sampling).

### 1.2. Bakgrund och Syfte

Ibland uppstår problem med att samla in korrekt information om ett socialt nätverk, detta sker då man har med så kallade gömda populationer att göra. När man tar ett traditionellt stickprov från en population är det en nödvändighet att man känner till sannolikheten för att välja en viss individ till stickprovet, annars kan man inte använda vanliga skattningstekniker som används vid standardiserat urval. Detta medför att en forskare behöver ha en lista på alla medlemmar i populationen, en så kallad ram, som man är intresserad av att studera. En sådan ram finns inte för gömda populationer, vilket är problemet. En forskare skulle kunna skapa en sådan ram själv, men för många populationer skulle ramen endast bli opraktisk eller så skulle det vara omöjligt att utföra på grund av att det är svårt att hitta individer som räknas till populationen.

När man vill göra en studie på en gömd population uppstår det ofta problematik i form av att det blir en kostnadsfråga eller att individerna i stickprovet inte blir slumpmässigt utvalda. Det som utmärker just dessa gömda populationer är att de består av ett socialt nätverk av människor. Det grundläggande tillvägagångssättet i RDS går ut på att individerna väljs genom vänskapsbanden mellan de redan existerande medlemmarna i stickprovet, alltså inte från en stickprovsram.

Det statistiska urvalet börjar med att man väljer ut ett litet antal populationsmedlemmar, som då blir de första att vara med i studien. Dessa individer rekryterar sedan sina vänner, som tillhör den population som är av intresse, till att vara med i studien. Dessa nya medlemmar gör sedan samma sak, det vill säga de rekryterar sina vänner. Processen fortsätter tills man uppnått ett satisfierande antal populationsmedlemmar. Notera att innebörden i ordet *vän* innefattar och avser någon slags förbindelse som man definierat i studien från början.

Problemet med den här typen av stickprov är att det är långt ifrån slumpmässigt. Individerna i populationen har inte alla samma sannolikhet att bli utvalda. Eftersom individer rekryterar sina vänner så har de med många vänner större sannolikhet att vara med i studien än de som är socialt isolerade. Detta faktum är grunden för den relativt nya metod som vi ska studera i denna uppsats – Respondent Driven Sampling (RDS). RDS är en modifikation av ”snowball sampling” som utarbetades av Goodman, 1961 (Heckathorn, 1997). Namnet kommer från att stickprovet växer i takt med en rullande snöboll. RDS utvecklades av Douglas Heckathorn 1997 som en del av ett nationellt projekt i Connecticut för att förhindra HIV-spridning och riktades in på injektionsdrogmissbrukare (*What is Respondent Driven Sampling?,* Respondent Driven Sampling).

Respondent Driven Sampling kan tillämpas på gömda populationer och från de här typerna av urval ska det med denna metod vara möjligt att göra väntevärdesriktiga skattningar av gömda populationer. Skattningarna ska också, enligt Heckathorn och Salganik (2004), vara asymptotiskt väntevärdesriktiga hur än de första individerna till stickprovet är valda.

Vi kommer att skatta populationsandelar med RDS för fyra olika modeller. Syftet är att undersöka om skattningen med RDS närmar sig den sanna andelen i populationen och om denna skattningsteknik verkar vara asymptotiskt väntevärdesriktig som Heckathorn och Salganik (2004) påstår. Vi kommer att undersöka under vilka omständigheter som RDS fungerar bra respektive mindre bra och vad som händer om vissa antaganden som man gör när man tillämpar RDS inte är uppfyllda.

### 1.3. Metod

Uppsatsen inleds med teori om RDS samt hur man går tillväga för att samla in stickprovet. Vi går igenom ett antal olika antaganden man gör när man tillämpar RDS och följer sedan upp med hur vi gör för att skatta populationsandelar med RDS. Vi kan, när vi har fått en bild av hur urvalet går till samt hur skattningar med RDS utförs, efterlikna detta på bästa sätt i våra simuleringar.

Vi ska generera ett antal slumpmässiga grafer som kommer att representera en population och dess sociala nätverk och låta individer vara HIV-positiva respektive HIV-negativa enligt olika modeller. Alla simuleringar kommer att göras i Matlab. För att undersöka om skattningarna är asymptotiskt väntevärdesriktiga kommer vi ur detta sociala nätverk att göra ett stickprov. På detta stickprov ska vi sedan göra skattningar med RDS. Resultatet kommer att illustreras i

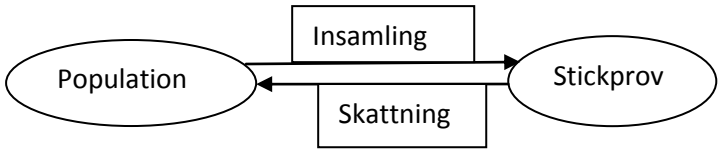
grafer och simuleringarna kommer att genomföras med olika värden på populationsstorleken, genomsnittliga antalet vänner samt den antagna andelen HIV-positiva i populationen.

# 2 Respondent Driven Sampling

## 2.1 Användandet av sociala nätverk för att skatta populationsandelar

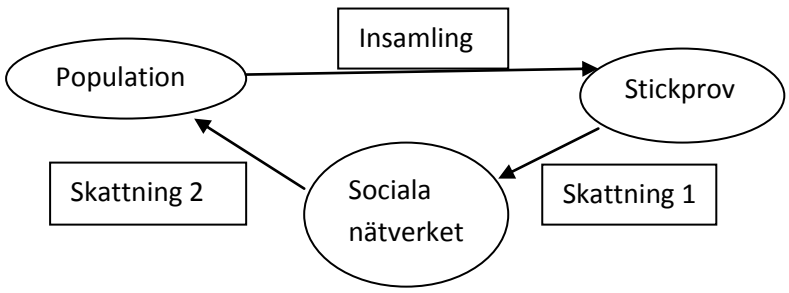
I Respondent Driven Sampling väljs individerna till stickprovet genom det sociala nätverket av de redan existerande medlemmarna i stickprovet. Först används stickprovet till att göra skattningar angående det sociala nätverket som sammanbinder populationen. Sedan använder man information om det sociala nätverket till att skatta populationsandelen med en viss egenskap, till exempel HIV-positiva. I denna rapport ska vi tänka oss en population med drogmissbrukare och att egenskapen av intresse är HIV-status.

Den grundläggande idén bakom skattningstekniken är att skattningarna inte härrör från proportionen HIV-smittade i stickprovsurvalet. I vanliga skattningstekniker förutsätter man att stickprovet kan ”symbolisera” populationen och skattar andelen HIV-smittade direkt ur stickprovet, som illustreras i figur 2.1.




**Figur 2.1.** Illustration av skattningsförfarandet i ett vanligt stickprov.

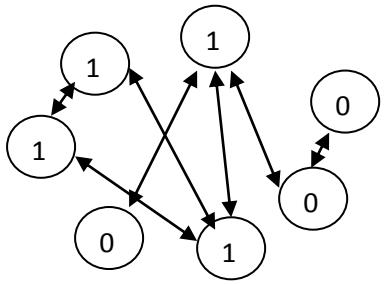
I RDS använder man stickprovsurvalet för att göra skattningar angående det sociala nätverket som binder samman populationen. Genom att använda informationen om det sociala nätverket kan man därefter skatta proportionen HIV-positiva respektive HIV-negativa i populationen. Detta illustreras i figur 2.2.




**Figur 2.2.** Illustration av skattningsförfarandet för ett RDS-stickprov.

Vi betraktar en population av drogmissbrukare där vissa av individerna i populationen är anknutna till varandra genom någon slags definition, till exempel vänskap. Populationen skapar då ett slags nätverk med noder (individer) och kanter (vänner), ett så kallat socialt nätverk.

Vi delar in individerna i två grupper, en grupp med individer som är HIV-positiva och en grupp med individer som är HIV-negativa. Vi är intresserade av att studera andelen HIV-positiva i denna population. Vi tilldelar de som är HIV-positiva beteckningen 1 och de som är HIV-negativa får beteckningen 0. I figur 2.3 illustrerar vi hur det sociala nätverket med HIV-positiva respektive HIV-negativa kan se ut, även om de verkliga nätverken av intresse givetvis är mycket större.




**Figur 2.3.** Graf över socialt nätverk med HIV-positiva respektive HIV-negativa individer.

Vi vill skatta andelen HIV-positiva i populationen. Vi repeterar att de som är HIV-positiva bemärks med en etta (1) och de som är HIV-negativa med en nolla (0). Vi betraktar endast ömsesidiga vänskapsband och man inser då att antalet vänskapsband från (1) till (0) är lika många som de från (0) till (1). Det här triviala konstaterandet kommer att bli väldigt användbart.

Först behöver man hitta ett fåtal individer som blir de första att delta i studien, så kallade frön. Man ber dessa individer att fylla i ett frågeformulär anonymt. Man ber dem också att skicka vidare frågeformuläret till några vänner som också hör till den populationen man är intresserad av att studera, i vårt fall drogmissbrukare. Sedan upprepas denna procedur genom att dessa nya individer i stickprovet gör samma sak. Det är viktigt att respondenterna känner sig anonyma och belöning är nödvändigt för att skapa incitament för individen att vara med i studien och för att uppnå hög svarsfrekvens.

Det kan anses möjligt att de noder som symboliserar HIV-positiva har fler grannar i jämförelse med de som symboliserar HIV-negativa, till exempel om man delar

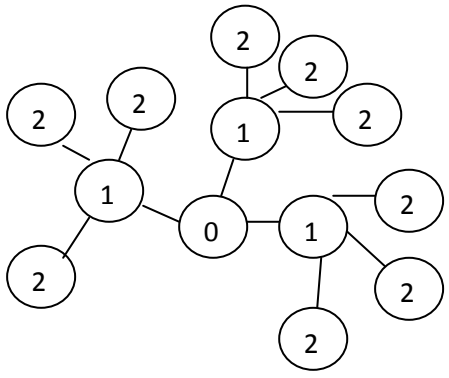
injektionsnålar. Detta leder till att de noder som symboliserar HIV-positiva blir överrepresenterade i stickprovet eftersom de individer som har fler vänner har en större sannolikhet att bli dragna till urvalet. Med andra ord så överskattas andelen HIV-positiva. Det är denna problematik RDS försöker lösa.

### 2.2. Insamling av ett RDS-stickprov

När man ska samla in urvalet som ska användas till att skatta nätverksinformation börjar man med att samla in ett antal individer som blir de första att medverka. I denna uppsats kommer vi dock att börja med endast en individ i våra simuleringar. Det första urvalet baseras på att individerna, som givetvis tillhör populationen man vill studera, är sådana som man varit i kontakt med förut. Dessa individer får ofta en ersättning för deras medverkan och dessa bildar våg 0 av stickprovet.

Varje individ i våg 0 överlämnas med $c$ unika rekryteringskuponger (som även de kan ha olika utseenden). Individerna i våg 0 bes att ge dessa kuponger till människor de känner som också tillhör den population man vill studera (drogmissbrukare). Eftersom varje kupong är unik kan dessa användas till att spåra rekryteringsmönstret i populationen. När en ny medlem av populationen man vill studera medverkar i studien, får rekryteraren till denna individ en extra belöning. På detta sätt skapar man incitament för individerna att vara med i studien.

Rekryteringarna som utförts av dem i våg 0 formar våg 1. När man är klar med den första fasen påbörjar man samma procedur igen, nu med dem som innefattas i våg 1. De här individerna kommer hjälpa till att forma våg 2. Urvalet fortsätter på samma sätt tills den önskade storleken på stickprovet uppnåtts. I figur 2.4 illustreras våg 0, våg 1 och våg 2 i ett stickprov.




**Figur 2.4.** Nätverk med de 2 första vågorna. Antalet kuponger per individ, *c*, är 3.

En viktig egenskap hos en individ är antalet vänner den har. För att skattningarna ska kunna genomföras måste man först veta hur många vänner varje individ som är med i stickprovet har. Detta medför att vi inför en definition, som är en av huvudstenarna i denna uppsats. Antag att vi har ett nätverk med N individer. Den $N \times N$-matris som symboliserar grafen över nätverket kallar vi S. Om vi vidare antar att individ $i$ och individ $j$ är vänner så är $s_{ij} = 1$, och på motsvarande sätt är $s_{ij} = 0$ om de inte är vänner. Vi definierar graden (eng. degree) som antalet vänner individ $i$ har, vilket också kan uttryckas som antalet kanter från nod $i$:

$$d_i = \sum_j s_{ij} \,. \tag{1}$$

De insamlade kupongerna används för att dokumentera rekryteringsmönstret så att varje stickprovsmedlem kan kopplas samman med den person som den blev rekryterad av. De här två delarna av information, graden för varje individ och det observerade rekryteringsmönstret, är av stor betydelse för skattningsprocessen.

Det är också viktigt att se till att varje person endast är med i studien en gång, det vill säga att en individ inte låtsas vara två olika individer, samt att kontrollera att individerna som är med i stickprovet tillhör populationen som man vill studera (Heckathorn, 1997). Informationen som samlats in om det sociala nätverket används för att göra skattningar över hela den population som man är intresserad av.

### 2.3. Antaganden

Vi ska nu presentera RDS. För detaljer hänvisar vi till Salganik och Heckathorn (2004).

För att kunna göra skattningar från stickprovet, måste vi först göra några antaganden angående populationen som vi studerar samt rekryteringsförfarandet.

Vi kan tänka oss att urvalsprocessen av stickprovet är en process som växlar mellan urvalet av noder och urvalet av kanter. Noder är först ritade för att forma våg 0 av stickprovet. Sedan väljer noderna kanter som definierar rekryteringen period 1 och som leder till att noderna i våg 1 formas. Processen fortsätter tills vi uppnått den storlek på stickprovet som vi var ute efter. Här nedan följer de antaganden som vi gör när vi gör skattningar med Respondent Driven Sampling.

1. Vänskapsbanden är ömsesidiga.
2. Vi antar att de som är med i studien rapporterar in rätt grad som de har i nätverket.

3. Vi ser endast till stickprovsurval med återläggning.
4. Det sociala nätverket av den gömda populationen ska forma en sammanhängande komponent.
5. Vi antar även att alla individer i urvalet tar emot och använder en kupong och att svaranden rekryterar slumpmässigt från alla kanter som de har (alla vänner de har).
6. Slutligen gör vi antagandet att de noder (individer) som väljs först, det vill säga våg 0, är dragna med en sannolikhet som är proportionell till deras grad. Med andra ord, en person med 10 vänner har dubbelt så stor chans att bli vald som en person med 5 vänner. Sannolikheten att individ *j* blir dragen i våg 0 där $d_j$ är antalet vänner person *j* har och $\sum_{i \in N} d_i$ är totala antalet vänskapsband i nätverket kan skrivas som

$$\mathrm{P}(\text{individ j blir vald i våg 0}) = \frac{d_j}{\sum_{i \in N} d_i}.$$

Vi gör det här antagandet eftersom de som blir valda som de första att vara med i studien ofta är de som är mest bekanta för de forskare som samlar in urvalet. De här välkända människorna tenderar att ha fler vänner än genomsnittet. Därför är det rimligt att anta att en individs chans att bli dragen som den första personen ökar med hans eller hennes grad.

### 2.4. Följder av antaganden

Eftersom vi har antagit att de första individerna är dragna med en sannolikhet som är proportionell mot graden, kan vi dra slutsatsen att sannolikheten för att en kant kommer dras i rekryteringen i period 1 är

$$\frac{d_j}{\sum_{i \in N} d_i} \cdot \frac{1}{d_j} = \frac{1}{\sum_{i \in N} d_i}.$$

Ekvationen indikerar att om noderna i våg 0 är dragna med en sannolikhet som är proportionell mot antalet vänner, så har varje kant samma sannolikhet att bli dragen i period 1. Sannolikheten att nod *j* kommer dras i våg 1 är då ekvivalent med summan av de $d_j$ kanterna som leder till denna nod kommer bli dragna i rekryteringsperiod 1:

$$\mathrm{P}(\text{individ j blir dragen till våg 1}) = \sum_{d_j} \frac{1}{\sum_{i \in N} d_i} = \frac{d_j}{\sum_{i \in N} d_i}.$$

Ekvationen visar att om fröna (våg 0) är dragna med en sannolikhet proportionell mot graden, så kommer noderna i våg 1 också bli dragna med en sannolikhet som är proportionell mot

graden. Om vi upprepar detta argument iterativt visar det sig att om noderna i våg $0$ är dragna med en sannolikhet som är proportionell till graden, så kommer noderna i alla efterkommande vågor bli dragna med en sannolikhet som är proportionell mot graden.

Argument kan också upprepas iterativt för att visa att om de första individerna som väljs till stickprovet är dragna med en sannolikhet som är proportionell mot graden, så kommer sannolikheten att en viss kant blir ritad i rekryteringen i en period *x* vara konstant och lika för alla kanter (Salganik & Heckathorn, 2004).

### 2.5. Beräkningar och beteckningar

Antalet vänner en individ har är huvudegenskapen hos en individ. Vi repriserar definitionen av graden hos en individ *i,* ekvation $(1)$:

$$d_i = \sum_j s_{ij}\,.$$

Totala antalet vänner som HIV-positiva har, det vill säga grupp $(1)$, är summan av graden av alla individer i grupp $(1)$ och vi skriver det som

$$N_{1\to1} + N_{1\to0} = \sum_{i\,\in 1} d_i = N_1 \cdot E(D_1).$$

Beteckningen $N_{1\to1}$ är antalet kanter från en HIV-positiv individ till en annan HIV-positiv individ och $N_{1\to0}$ är antalet kanter från en HIV-positiv individ till en HIV-negativ individ. Fortsättningsvis är $N_1$ totala antalet individer i som är HIV-positiva i populationen och $E(D_1)$ är genomsnittliga graden för HIV-positiva. Den genomsnittliga graden för en grupp skattar vi i avsnitt 2.6.

Vi får på samma sätt

$$N_{0\to0} + N_{0\to1} = \sum_{i\,\in 0} d_i = N_0 \cdot E(D_0).$$

Beteckningen $N_{0\to0}$ är antalet vänskapsband mellan HIV-negativa individer och $N_{0\to1}$ är antalet kanter från en HIV-negativ till en HIV-positiv individ. Vidare är $N_0$ totala antalet HIV-negativa i populationen och $E(D_0)$ är den genomsnittliga graden för HIV-negativa individer.

Vi vill använda informationen från stickprovet till att skatta andelen HIV-positiva i det sociala nätverket som binder samman populationen. För att skatta sannolikheten att man går från en nod i grupp $(1)$ till en nod som tillhör grupp $(0)$ kan vi använda oss av den information som vi har samlat in om nätverket. Det första vi vill skatta är sannolikheten att om vi följer en utvald individ i grupp $(1)$ att vi hamnar hos en individ i grupp $(0)$. Ett sätt att skatta sistnämnda sannolikhet är att fråga svaranden vilken procentandel av deras vänskapskrets som tillhör respektive grupp. Det är dock en omöjlighet eftersom det är svårt för respondenterna att veta om deras vänner är HIV-positiva eller inte.

Vi tittar istället på det faktiska beteendet. När en respondent rekryterar en annan, kan denna beteendelänk representera en länk i det sociala nätverket. Man kan verifiera länken genom att fråga den rekryterade om förhållandet till rekryteraren. Den rekryterade ska verifiera förhållandet till rekryteraren som en vän och inte som en främling. Endast de här verifierade länkarna kan användas för skattning.

Eftersom varje kant $j$ till $k$ har lika stor sannolikhet att bli dragen i varje rekryteringsperiod så är det visat att rekryteringarna som vi observerar är ett slumpmässigt urval av alla möjliga rekryteringar. Vi får då fyra möjliga rekryteringsslag.

Beakta nu, för ett givet nätverk $S$, att vi följer ett slumpmässigt valt vänskapsband med början i en person i grupp $(1)$. Eftersom alla observerade rekryteringar är ett slumpmässigt urval från alla kanter så är sannolikheten att vi hamnar hos en vän som också räknas till grupp $(1)$

$$p_{1\to1} = \frac{N_{1\to1}}{N_{1\to1} + N_{1\to0}}. \qquad (2)$$

Sannolikheten, givet att vi börjar hos en person i grupp $(1)$, att vi hamnar hos en person i grupp $(0)$ är då:

$$p_{1\to0} = 1 - p_{1\to1}.$$

På samma sätt fås $p_{0\to0}$ och $p_{0\to1} = 1 - p_{0\to0}$.

Eftersom vi endast betraktar ömsesidiga vänskapsband så vet vi att det är samma antal kanter från grupp $(1)$ till $(0)$ som det är från $(0)$ till $(1)$. Vi får alltså

$$N_1 \cdot E(D_1) \cdot p_{1\to0} = N_0 \cdot E(D_0) \cdot p_{0\to1} = (N - N_1) \cdot E(D_0) \cdot p_{0\to1}, \qquad (3)$$

där $N$ är storleken på den totala populationen. Notera att ekvation (3) innehåller både information om karaktären på noden (HIV-positiv eller HIV-negativ) och karaktären på det sociala nätverket (antalet HIV-positiva, genomsnittliga antalet vänner och sannolikheten för vänskapsband mellan 0 och 1).

Om vi har fullständig information om det sociala nätverket, det vill säga om vi känner till $E(D_1), E(D_0), p_{1\to0}$ och $p_{0\to1}$, så saknar vi fortfarande information om $N_1$ och $N_0$, det vill säga storleken på populationen i grupp (1) och grupp (0). Låt $p_1$ vara andelen HIV-positiva i populationen och $p_0$ vara andelen HIV-negativa i populationen. Om vi dividerar båda sidor av ekvation (3) med $N$ kan vi skriva ekvationen som

$$p_1 \cdot E(D_1) \cdot p_{1\to0} = p_0 \cdot E(D_0) \cdot p_{0\to1}.$$

där

$$p_1 = \frac{N_1}{N} \text{ och } p_0 = \frac{N_0}{N}.$$

Det leder till att vi inför en restriktion; summan av populationsandelarna måste vara 1.

$$p_1 + p_0 = 1.$$

Nu kan vi lösa ekvationssystemet

$$\begin{cases} p_1 \cdot E(D_1) \cdot p_{1\to0} = p_0 \cdot E(D_0) \cdot p_{0\to1} \\ p_1 + p_0 = 1 \end{cases}$$

och vi får

$$p_1 = \frac{E(D_0) \cdot p_{0\to1}}{E(D_1) \cdot p_{1\to0} + E(D_0) \cdot p_{0\to1}} \tag{4}$$

samt

$$p_0 = \frac{E(D_1) \cdot p_{1\to0}}{E(D_0) \cdot p_{0\to1} + E(D_1) \cdot p_{1\to0}}. \tag{5}$$

Genom att studera ekvationerna (4) och (5) ovan kan vi dra slutsatsen att det räcker med att ha information om strukturen på det sociala nätverket för att erhålla populationsandelarna för de olika grupperna.

## 2.6. Skattningar

Vi betecknar det observerade värdet av antalet hopp från en HIV-positiv individ till en annan HIV-positiv individ som $n_{1\to1}$. På liknande sätt får vi för de tre andra möjliga hoppen de observerade värdena $n_{1\to0}, n_{0\to0}$ och $n_{0\to1}$. Vi kan då, på samma sätt som i ekvation (2), givet med början i en person i grupp (1), skatta sannolikheten att vi hamnar hos en vän som också räknas till grupp (1):

$$\hat{p}_{1\to1} = \frac{n_{1\to1}}{n_{1\to1} + n_{1\to0}}.$$

Vi får då att $\hat{p}_{1\to0} = 1 - \hat{p}_{1\to1}$. På samma sätt får vi skattningarna för de andra två övergångssannolikheterna:

$$\hat{p}_{0\to0} = \frac{n_{0\to0}}{n_{0\to1} + n_{0\to0}}$$

och $\hat{p}_{0\to1} = 1 - \hat{p}_{0\to0}$.

Nu återstår skattningarna av $E(D_1)$ och $E(D_0)$. Vi har under rekryteringsprocessen, det vill säga insamlandet av stickprovet, samlat in graden hos varje person. Om vi skulle försöka att skatta väntevärdet av graden i varje grupp genom att ta väntevärdet av graden av personerna i stickprovet skulle vår skattning bli för hög. I och med utförandeformen av insamlingen av stickprovet så kommer personer med hög grad överrepresenteras (Erickson, 1979).

Eftersom väntevärdet på stickprovet inte är en bra skattning måste vi ta data från stickprovet och justera det så att det ger riktig information om populationen.

Ett sätt att skapa en skattning för den genomsnittliga graden, vilket Salganik och Heckathorn (2004) kallar ”degree distribution approach” är genom fördelningen av graden hos stickprovet och populationen.

Sannolikheten att man blir dragen till stickprovet är proportionell med antalet vänner man har. Vi illustrerar detta i figur 2.5.

**Figur 2.5.** En individ med 0 vänner har en sannolikhet 0 att bli dragen till stickprovet. En individ med 4 vänner har 4 gånger så stor sannolikhet att bli dragen som en individ med 1 vän.

Fördelningen för graden i stickprovet är (Salganik & Heckatorn, 2004)

$$\tilde{p}_j^{\,(1)} = \frac{jp_j^{(1)}}{\sum_{j=1}^{maxj} jp_j^{(1)}}. \tag{6}$$

Där $p_j^{(1)}$är andelen HIV-positiva i populationen med j vänner och $\tilde{p}_j^{\,(1)}$ är andelen HIV-positiva i stickprovet med j vänner, och där

$$\sum_{j=1}^{maxj} jp_j^{(1)} = 1 * P(\text{hivpositiv med 1 vän}) + 2 * P(\text{hivpositiv med 2 vänner}) + \cdots +$$

$$+ \max(j) * P(\text{hivpositiv med } \max(j) \text{ vänner}) \tag{7}$$

är en normeringskonstant som gör att $\sum_{j=1}^{maxj} \tilde{p}_j^{\,(1)}$ summerar till 1.

Ekvation (6) tillåter oss att prediktera fördelningen i stickprovet $\tilde{p}_j^{\,(1)}$, givet fördelningen för populationen $p_j^{(1)}$. Under våra antaganden är sannolikheten för en nod att bli dragen proportionell med deras grad. Det här faktumet samt kännedomen om fördelningen av graden i stickprovet $\tilde{p}_j^{\,(1)}$gör att vi kan prediktera fördelningen för populationen $p_j^{(1)}$. Genom utnyttjandet av att $\tilde{p}_j^{\,(1)}$ är proportionell med $jp_j^{(1)}$, får vi

$$\tilde{p}_j^{\,(1)} \propto jp_j.$$

Detta medför att

$$p_j^{(1)} \propto \frac{\tilde{p}_j^{\,(1)}}{j}.$$

Om fördelningen för stickprovet är $\tilde{p}_j^{\,(1)}$så kan fördelningen för populationen $p_j^{(1)}$ bli skattad som

$$p_j^{(1)} = \frac{\frac{1}{j}\tilde{p}_j{}^{(1)}}{\sum_{j=1}^{maxj}\frac{1}{j}\tilde{p}_j{}^{(1)}},$$

där $\sum_{j=1}^{maxj}\frac{1}{j}\tilde{p}_j{}^{(1)}$ på samma sätt som likheten (7) är en normeringskonstant.

Nu har vi fördelningen för skattningen av andelen HIV-positiva i populationen, $p_j^{(1)}$, och kan således skatta den genomsnittliga graden i populationen. Väntevärdet av en diskret sannolikhetsfunktion $p(x)$ är $\sum_{x=0}^{\infty} x \cdot p(x)$. Vi får då att

$$E(D_1) = \sum_{j=1}^{maxj} j\, p_j^{(1)} = \sum_{j=1}^{maxj} j \frac{\frac{1}{j}\tilde{p}_j{}^{(1)}}{\sum_{j=1}^{maxj}\frac{1}{j}\tilde{p}_j{}^{(1)}}.$$

Vilket ger

$$\sum_{j=1}^{maxj}\frac{1}{j}\tilde{p}_j{}^{(1)} = \sum_{j=1}^{maxj}\frac{1}{j}\frac{jp_j^{(1)}}{E(D_1)} = \frac{1}{E(D_1)}.$$

Vi får då

$$E(D_1) = \frac{1}{\sum_j \frac{1}{j}\tilde{p}_j{}^{(1)}}. \qquad (8)$$

Vi kan skatta $E(D_1)$ med att ersätta $\tilde{p}_j{}^{(1)}$ i ekvation (8) med stickprovets frekvenser. Vi skattar $\tilde{p}_j{}^{(1)}$ med $\widehat{\tilde{p}}_j{}^{(1)}$ som är RDS-stickprovets empiriska gradfördelning och får skattningen för den genomsnittliga graden:

$$\widehat{E(D_1)} = \frac{1}{\sum_j \frac{1}{j}\widehat{\tilde{p}}_j{}^{(1)}}. \qquad (9)$$

Skattningen $\widehat{E(D_1)}$ är enligt Salganik och Heckathorn (2004) asymptotiskt väntevärdesriktig.

På samma sätt får vi fram

$$\widehat{E(D_0)} = \frac{1}{\sum_j \frac{1}{j}\widehat{p}_j^{(0)}}.$$

Med kännedom om den genomsnittliga graden i grupp (1) och grupp (0) kan vi slutligen erhålla skattningen för andelen HIV-positiva med RDS:

$$\hat{p}_{RDS} = \hat{p}_1 = \frac{\widehat{E[D_0]}\hat{p}_{0\to 1}}{\widehat{E[D_1]}\hat{p}_{1\to 0} + \widehat{E[D_0]}\hat{p}_{0\to 1}}. \quad (10)$$

# 3 Modeller och simuleringar

Vi ska nu simulera ett socialt nätverk med hjälp av Matlab. Vi ska ur detta nätverk göra ett RDS-stickprov som vi ska göra skattningarna som vi gick igenom i avsnitt 2.6 på. Dessa skattningar ska vi utvärdera mot sanningen, alltså mot den riktiga andelen HIV-positiva i det sociala nätverket.

## 3.1. Den sociala grafen

Till att börja med behöver vi skapa ett socialt nätverk, det vill säga en matris som förklarar vilka som är vänner med varandra och hur individerna är kopplade till varandra. Matrisen kommer vi tillge benämningen den *sociala grafen*, vilken kommer betecknas med ett $S$. $S$ är en matris som representerar den sociala grafen och kommer att ha utseendet $s_{ij} = 1$ om person $i$ och person $j$ är vänner och $s_{ij} = 0$ om de inte är vänner. Vänskapsbanden mellan individer nämns för $kanter$ och varje individ kallas för en $nod$. Vi har en matris $S$ med $N\ noder$ och $\binom{N}{2}\ kanter$.

Till en början kommer vi endast titta på ömsesidiga vänskapsband, det vill säga om person $i$ är vän med person $j$ så är också person $j$ vän med person $i$. S kommer vara en $N \times N$-matris, där $N$ är antalet individer i populationen, och kan exempelvis se ut på följande vis:

$$S = \begin{bmatrix} 0 & 1 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ \vdots & 1 & \ddots & 0 & \vdots \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

Diagonalen kommer alltid att vara 0, man kan inte vara vän med sig själv. Matrisen kommer att vara symmetrisk, det vill säga $s_{ij} = s_{ji}$.

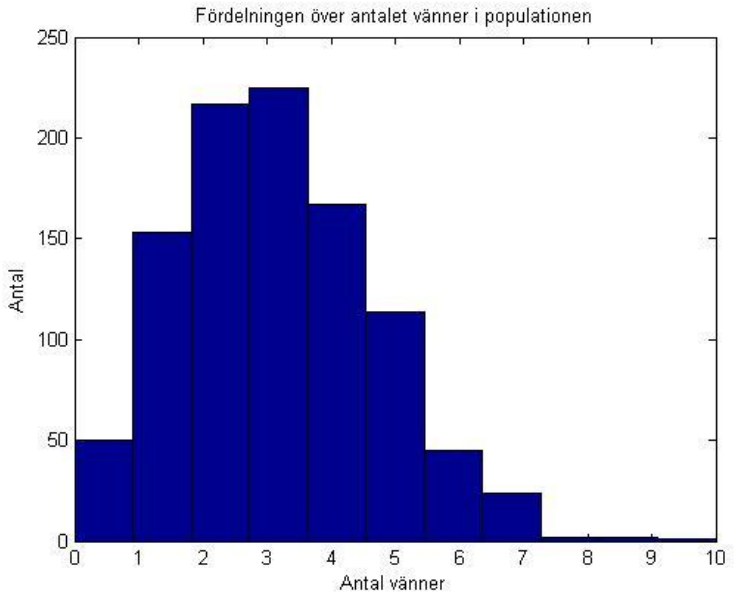
Första ansatsen är att simulera en slumpgraf, en så kallad Erdös-Rényi-graf där vi slumpar ut ettor och nollor oberoende och med samma sannolikhet. Grafen ska symbolisera det sociala nätverket i en population. Vi får fram grafen genom att slumpa 0:or och 1:or i varje position $(i, j)$. Varje rad (kolumn) symboliserar en individ $i$. Rad $i$ kommer alltså se likadan ut som kolumn $i$. En 1:a står för att det finns en kant mellan individ $i$ och individ $j$, en 0:a står för att det inte finns någon kant mellan $i$ och $j$.

Eftersom $S$ skall vara en symmetrisk matris kommer vi endast att slumpa i övre triangeln av matrisen och sedan spegla ned den övre halvan i den nedre halvan av matrisen. På så sätt får vi ett socialt nätverk av individer där endast ömsesidiga vänskapsband ingår. Vi noterar också att diagonalen endast kommer att bestå av $0: or$; vilket betyder att man, i detta sammanhang, inte kan vara vän med sig själv. Antalet vänner som person $i$ har kommer alltså vara summan av rad $i$, $d_i = \sum_j s_{ij}$.

En parameter $\lambda$ kommer vara väntevärdet för antalet vänner. Vi slänger ut kanter oberoende mellan varje par av individer med en sannolikhet $p = \frac{\lambda}{N}$. Antalet vänner som en individ har blir binomialfördelat med parametrarna $N - 1$ och $p = \frac{\lambda}{N}$, det vill säga $d_i \sim Bin(N - 1, \frac{\lambda}{N})$. Sannolikheten för nod *i* att ha j grannar ges alltså av

$$P(d_i = j) = \binom{N-1}{j} p^j (1-p)^{N-1-j}, \qquad j = 1, \dots, N-1.$$

När $N$ blir tillräckligt stort så kommer alltså $d_i \xrightarrow{d} Po(\lambda)$. I figur 3.1 illustreras fördelningen av antalet vänner för hela populationen, där $\lambda = 3$ och det verkar följa en Poissonfördelning.




**Figur 3.1.** Fördelningen av antalet vänner i populationen.

Nästa steg är att bestämma HIV-status för alla individer i populationen. Detta gör vi på två olika sätt. För att reprisera kommer de som är HIV-postiva att bemärkas med en etta $(1)$ och de som är HIV-negativa med en nolla $(0)$. Vi kommer att skapa en (1×N)-vektor som består av $1: or$ och $0: or$.

Till att börja med ska vi oberoende och lika fördelat över alla individer slumpa ut HIV-status, det vill säga $1: or$ och $0: or$. Antalet HIV-positiva, X, blir då binomialfördelat, $X \sim Bin(N, p_{HIV})$, och vi får då sannolikhetsfunktionen

$$P(X = k) = \binom{N}{k} {p_{HIV}}^k (1 - p_{HIV})^{N-k}, \qquad k = 0, \dots, N.$$

Där $p_{HIV}$ är den antagna sannolikheten att en person är HIV-positiv. Egentligen är vi inte så intresserade av just det här sättet att slumpa ut HIV-status bland individerna. Metoden som vi samlar in stickprovet med innebär att de med många vänner kommer att vara överrepresenterade i stickprovet. Som nämnt tidigare anses det möjligt att de noder som symboliserar HIV-positiva har fler grannar än de som symboliserar HIV-negativa och andelen HIV-positiva i populationen kommer därför att överskattas. Det är dock intressant att se hur pass bra skattningarna blir i detta fall.

Efter detta så slumpar vi ut HIV-status med en sannolikhet som är proportionell mot antalet vänner man har. Det kommer med andra ord vara så att de som har många vänner kommer ha

större sannolikhet att vara HIV-positiva. Vi betecknar även här den antagna sannolikheten att en person är HIV-positiv med $p_{HIV}$, alltså oberoende av hur många vänner man har. Sannolikheten att individ *i* är HIV-positiv kan följaktligen skrivas som

$$p_{ny}(i) = \frac{p_{HIV} * d_i}{\lambda} \wedge 1. \qquad (11)$$

Notera att sannolikheten i ekvation (11) kan bli större än 1, därav restriktionen $\wedge 1$, som betyder att alla sannolikheter över 1 ska sättas till 1. Sannolikheten att en individ är HIV-positiv är således beroende av hur pass social personen ifråga är.

Vi har nu en graf med ett socialt nätverk över en population med $N$ individer och en vektor som indikerar varje individs HIV-status i denna population, där det senare är gjort på två olika sätt.

### 3.2. Den stora komponenten

Vi har nu ett nätverk med kanter och individer som är antingen HIV-positiva eller HIV-negativa. Vi ska nu genomföra RDS och undersöka hur skattningen $\hat{p}_1$ beter sig.

I simuleringarna så genererar vi en population och gör sedan ett RDS-stickprov från populationen. Varje individ som väljs till stickprovet kommer kunna bli vald igen, med andra ord gör vi ett stickprov med återläggning. För att gå tillbaka till vårt antagande om att det sociala nätverket av den gömda populationen ska forma en enda sammanhängande komponent så ska vi först endast ta med de individer som ingår i den så kallade stora komponenten som vi betecknar med $SK$. Den stora komponenten är den största sammanhängande gruppen noder.

Enligt teorin för Erdös-Rényi-grafer så gäller det att om $\lambda > 1$ och om $N$ går mot oändligheten så kommer det helt säkert finnas en jättekomponent i grafen som binder samman en stor grupp individer. Enstaka individer samt små par/grupperingar hamnar utanför komponenten. I Jiong Caos kandidatarbete om Erdös-Rényi-grafer (2009:1) kan den intresserade läsaren läsa vidare om fasövergången för graferna.

Det är en stor sannolikhet att en individ som har många vänner också ingår i den stora komponenten. Vi vill nu hitta den stora komponenten i vår genererade graf. För att få ut $SK$ börjar vi därför med att välja en av de personer som har flest vänner i populationen. Därefter skapar vi en vektor $K$ som också kommer bestå av $1{:}\,or$ och $0{:}\,or$. Till en början indikerar vi den första individen med en $1{:}\,a$. Sedan går vi in i den sociala grafen och letar upp den

sistnämnde individens vänner. Vi markerar i sin tur vännerna med en $1{:}a$ i $K$-vektorn. Därefter upprepar vi proceduren med alla individer som har en $1{:}a$ i $K$-vektorn: Vi går in i den sociala grafen och letar upp de senast omnämnda personernas vänner och markerar dessa med en $1{:}a$ i $K$-vektorn. Vi upprepar förfarandet, med vad vi anser vara tillräckligt många gånger, för att vara helt säkra på att vi får med alla personer som ingår i den stora komponenten. De som är markerade med en 1:a i $K$-vektorn är alltså de individer som ingår i den stora komponenten.

Det är andelen HIV-positiva i den stora komponenten som vi först och främst är intresserade av att skatta och jämföra med den skattning som vi kommer få med RDS. Därför beräknar vi antalet individer som både ingår i den stora komponenten och som är HIV-positiva. Dividerar vi nu detta värde med det totala antalet individer i den stora komponenten får vi andelen HIV-positiva i den stora komponenten, som vi kallar $\hat{p}_{SK}$. Vi betecknar antalet HIV-positiva i den stora komponenten med $X_{SK}$, och antalet individer i den stora komponenten med $N_{SK}$. Vi får då

$$\hat{p}_{SK} = \frac{X_{SK}}{N_{SK}}. \tag{12}$$

### 3.3. Insamling av RDS-stickprovet

Hittills i våra beräkningar så har vi antagit att de första individerna är dragna med en sannolikhet som är proportionell mot graden. När en studie utförs så väljer man ofta individer som är välkända för en att bli de första individerna att starta stickprovet. Dessa individer är långt ifrån representativa och har ofta mycket högre grad än andra individer i gömda populationer. Skattningarna i RDS är väntevärdesriktiga om de första individerna är dragna med en sannolikhet som är proportionell mot graden. Det är dock inte så troligt att fröna blir dragna med exakt proportionalitet till graden. Det visar sig som tur är, genom att använda teori för Markovkedjor, att vi kan visa att skattningarna är asymptotiskt väntevärdesriktiga hur de första individerna än väljs (även om inte de väljs med en sannolikhet som är proportionell mot graden). Detta bevisas i Salganik och Heckathorn (2004).

Vi börjar samla in stickprovet och väljer ut en individ som blir själva startpunkten för vårt urval. Vi ska samla in stickprovet genom att endast välja en individ i taget, med andra ord symboliserar detta att individerna som är med i stickprovet endast får rekrytera en vän i taget.

Vi börjar med att slumpa en av de individer som har det maximala antalet vänner i populationen. Med tanke på det vi berört ovan kommer detta inte spela någon större roll, skattningarna är asymptotiskt väntevärdesriktiga hur de första individerna än väljs. Vi kan med största sannolikhet vara säkra på att denna individ ingår i den så kallade stora komponenten. Däremot ska den första individen inte ingå i det stickprov som vi ska använda för skattningen av RDS. Salganik och Heckathorn (2004) rekommenderar att man inte tar med fröet när man skattar den genomsnittliga graden eftersom de är dragna på ett annorlunda sätt.

Vi väljer ut en av den första individens vänner slumpmässigt, detta kommer bli den första person att bli vald till stickprovet som vi kommer använda för skattningen av RDS. Därefter noterar vi antalet vänner denna slumpmässigt valda vän har samt noterar vännens HIV-status. Därefter så upprepar vi samma procedur. Vi väljer ut en av den slumpmässigt valda vännens vänner på samma sätt som ovan, slumpmässigt. Vi noterar den nya individens HIV-status och antalet vänner individen har. Detta upprepar vi totalt $k = 500$ gånger. Vi sätter in HIV-status samt antalet vänner som individerna har i två nya vektorer. Vi har nu skapat vårt RDS-stickprov.

Vi har nu konstruerat två nya vektorer med dimensionen $(1 \times 500)$. De individer som ingår i de nya vektorerna tillhör stickprovet som vi nu ska skatta $\hat{p}_{RDS}$ för. Observera att samma individ kan ingå i dessa vektorer flera gånger, eftersom vi slumpar vänner med återläggning. När vi skattar $\hat{p}_{RDS}$ kommer vi göra detta för olika värden på $k$, nämligen $k = 1, 2, .., 500$. På detta sätt kan vi se hur skattningens precision förbättras i takt med att urvalet växer.

### 3.4. Bestämmande av skattningar

Genom att tillämpa de olika skattningar som vi gått igenom kan vi simulera detta i Matlab. Vi räknar ut antalet individer i stickprovet som är HIV-positiva och har *j* vänner,

$$j = 1, \dots, \max(d_i).$$

Vi erhåller en vektor $b$ med dimensionen $(1\ x\ j)$. Därefter räknar vi ut totala antalet HIV-positiva i stickprovet, $n_1$. Vi dividerar $b$-vektorn med $n_1$ och får det vi kallar $\widehat{\bar{p}_j}$, det vill säga en vektor med andelen HIV-positiva som har *j* vänner i stickprovet.

Vi räknar sedan ut det förväntade antalet vänner HIV-positiva i populationen har med ekvation (9), $\widehat{E(D_1)} = \frac{1}{\sum_j \frac{1}{j} \widehat{\bar{p}_j}^{(1)}}$.

Vi kan nu få ut antalet hopp från (1) till (1) och sedan sannolikheten $p_{1\to1}$ och gör likadant för de tre andra möjliga hoppen. Vi gör samma procedur för de som är HIV-negativa.

Till slut kan vi skatta $\hat{p}_{RDS}$ enligt ekvation (10), $\hat{p}_{RDS} = \frac{\widehat{E[D_0]}\hat{p}_{0\to1}}{\widehat{E[D_1]}\hat{p}_{1\to0}+\widehat{E[D_0]}\hat{p}_{0\to1}}$.

När stickprovsstorleken, k, är litet har vi ibland inte både HIV-negativa respektive HIV-positiva med i stickprovet. Om stickprovet endast består av HIV-negativa sätter vi skattningen av andelen HIV-positiva till 0. Om stickprovet endast består av HIV-positiva sätter vi skattnigen $\hat{p}_{RDS}$ till 1.

### 3.5. Modeller

I modellerandet som vi studerat ovan så skall vi variera värdena på $n, p$ och $\lambda$. Vi kommer att simulera ett socialt nätverk och slumpa ut sannolikheten för att två personer är vänner med ovan nämnda $\lambda$. Sannolikheten för att en person är HIV-positiv kommer vi slumpa ut i stigande proportion till antalet vänner man har, enligt ekvation (11), vi benämner denna modell för modell 1. Sedan slumpar vi ut HIV-positiva oberoende och lika fördelat över hela populationen, $p_{HIV}$, som blir vår modell 2.

Vi kommer att testa att simulera en population där det ingår fler ”trianglar”, vilket bör fungera enligt teorin. Vi utgår från den sociala grafen som vi skapade ovan. Om en person har två vänner så antar vi att det finns en ökad sannolikhet att också de två vännerna är vänner med varandra. Det vill säga om individ $i$ är vän med individ $j$ och individ $k$ så finns det en förhöjd sannolikhet $\alpha$ att också individ $j$ och $k$ är vänner. Vi simulerar detta genom att gå in i varje plats i $S$-matrisen och för varje par av vänner $(i, j)$ och $(i, k)$ slumpa ut en sannolikhet $\alpha$ att också $(j, k)$ är vänner. Om vi vill ha ungefär samma antal kanter som i den första sociala grafen så måste vi minska $\lambda$. Vi benämner denna modell för modell 3.

Enligt teorin för RDS gäller skattningarna bara på oriktade nätverk, men vi ska även testa RDS på riktade nätverk. Vi slumpar nu istället en hel matris med $1: or$ och $0: or$, där vi nu inte kommer erhålla att rad $i$ ser ut på samma sätt som kolumn $i$. Däremot kan det vara rimligt att anta att om person $i$ är vän med person $j$ så finns det en förhöjd sannolikhet att också $j$ är vän med individ $i$. Vi betecknar denna sannolikhet med $\beta$ och benämner denna modell för modell 4.

# 4 RDS för de olika modellerna

Vi ska nu presentera resultaten av våra simuleringar. Vi börjar med den graf där endast ömsesidiga vänskapsband ingår och där kanterna endast slumpats ut mellan par av individer och sannolikheten att en individ är HIV-positiv ökar med antalet vänner, det vill säga modell 1. Sedan gör vi det fallet där individer är HIV-positiva oberoende och lika fördelat över hela populationen, den modell vi kallar modell 2.

Vi går vidare med att presentera fallen där fler trianglar ingår i grafen, modell 3. Till sist presenterar vi resultaten för det fallet där det är tillåtet för kanterna att vara riktade, modell 4.

För att få en bra bild av hur bra RDS-skattningarna blir så plottar vi varje modell flera gånger. Vi presenterar dock endast en figur per modell här nedan, för fler figurer hänvisar vi till appendix.

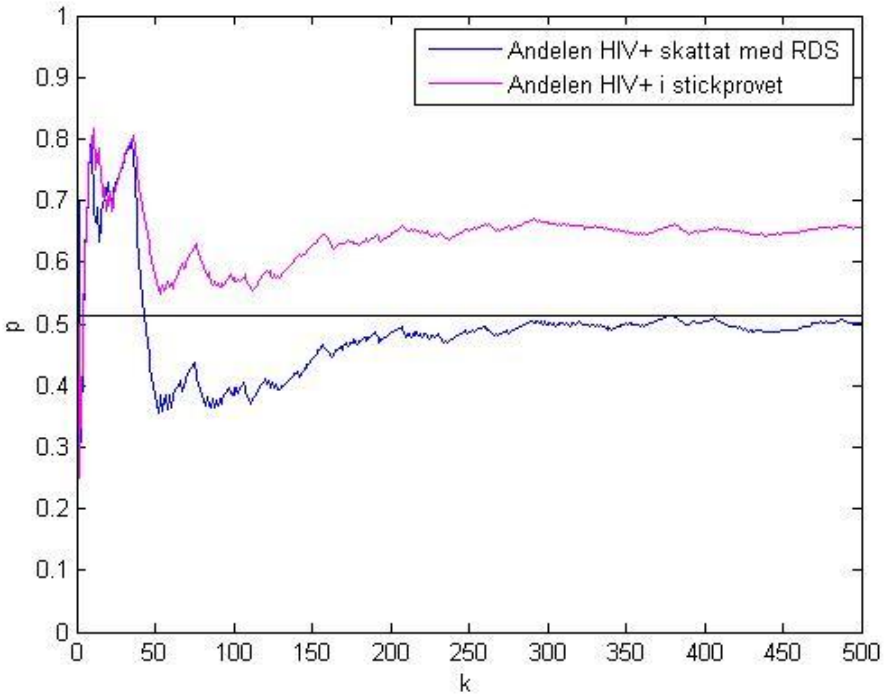
## 4.1. RDS på nätverk med ömsesidiga vänskapsband (modell 1)

De olika värden som testas för de olika parametrarna för den modell där HIV-status slumpas ut i proportion mot antalet vänner en individ har är $N = 1000\ och\ N = 5000,\ p_{HIV} = 0.5\ och\ p_{HIV} = 0.1, \lambda = 3\ och\ \lambda = 10.$

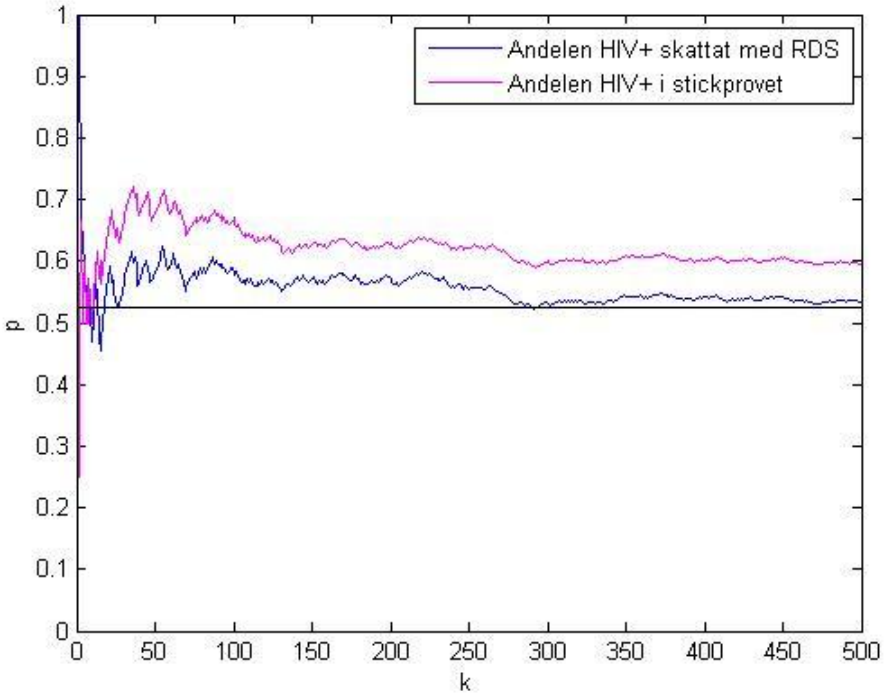
I figur 4.1 illustrerar den undre grafen andelen HIV-positiva skattat med RDS, $\hat{p}_{RDS}$, det vill säga skattat enligt ekvation (10), $\hat{p}_{RDS} = \frac{\widehat{E[D_0]}\hat{p}_{0\to1}}{\widehat{E[D_1]}\hat{p}_{1\to0}+\widehat{E[D_0]}\hat{p}_{0\to1}}$.

Skattningen verkar gå mot den sanna andelen HIV-positiva i den stora komponenten, värdet $\hat{p}_{SK} = 0.51$, enligt ekvation (12). Det är detta värde som skattningen för med $\hat{p}_{RDS}$ hela tiden jämförs med, eftersom vårt stickprov kommer från den stora komponenten. Den övre grafen visar andelen HIV-positiva i stickprovet. I figur 4.1 bekräftas det att andelen HIV-positiva överrepresenteras i stickprovet, jämfört med den egentliga andelen i den stora komponenten. Figurerna A1a-d i appendix kan studeras för att se fler grafer med samma värden på parametrarna.

I figur 4.2 har det förväntade antalet vänner, $\lambda$, höjts till $\lambda = 10$. Genom att jämföra figur 4.1 och figur 4.2 samt figurerna A1a-d samt A2a-d i appendix så tycks variansen av skattningen av andelen HIV-positiva med RDS vara mindre i fallet $\lambda = 10$ än i fallet $\lambda = 3$.




**Figur 4.1.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.51$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.5$ och $\lambda = 3$.

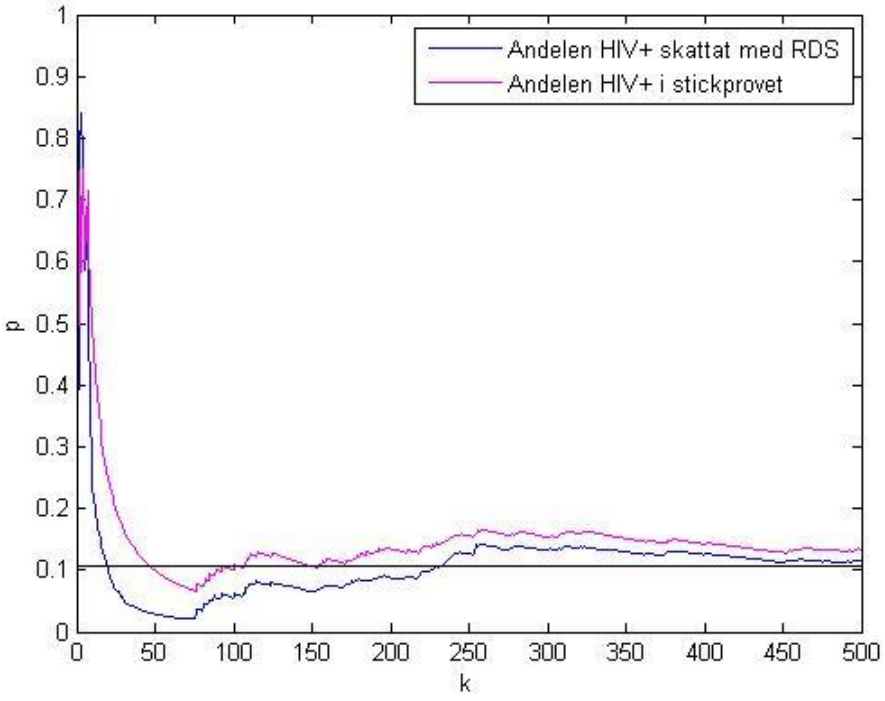



**Figur 4.2.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.53$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.5$ och $\lambda = 10$.

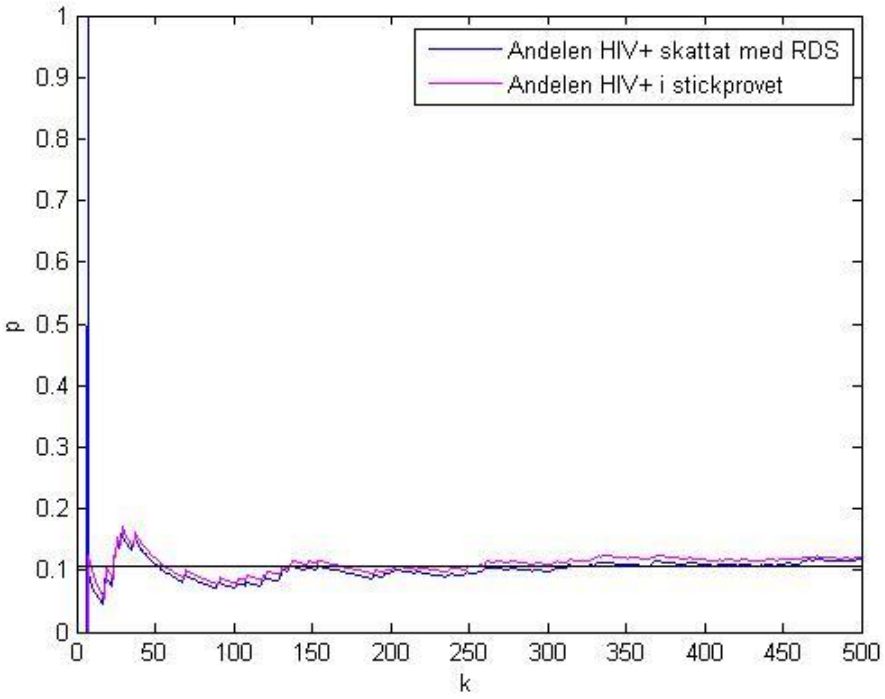
I figur 4.3 är istället $p_{HIV} = 0.1$. Med en lägre antagen andel HIV-positiva tycks skattningen även här närma sig det sanna värdet av andelen HIV-positiva i den stora komponenten. Jämförelse med figur 4.4, där vi har ett högre antaget förväntat antal vänner, $\lambda$, så verkar det även här som skattningen med större antaget förväntat antal vänner närmare sig den sanna andelen HIV-positiva med större precision. I appendix finnes fler figurer att jämföra, A3a-d respektive A4a-d.

Det verkar som att skattningen blir bättre när individerna har fler vänner. Jämförelse med figur 4.3 och figur 4.4 så har skattningen mindre variation i figur 4.4 när k är litet.

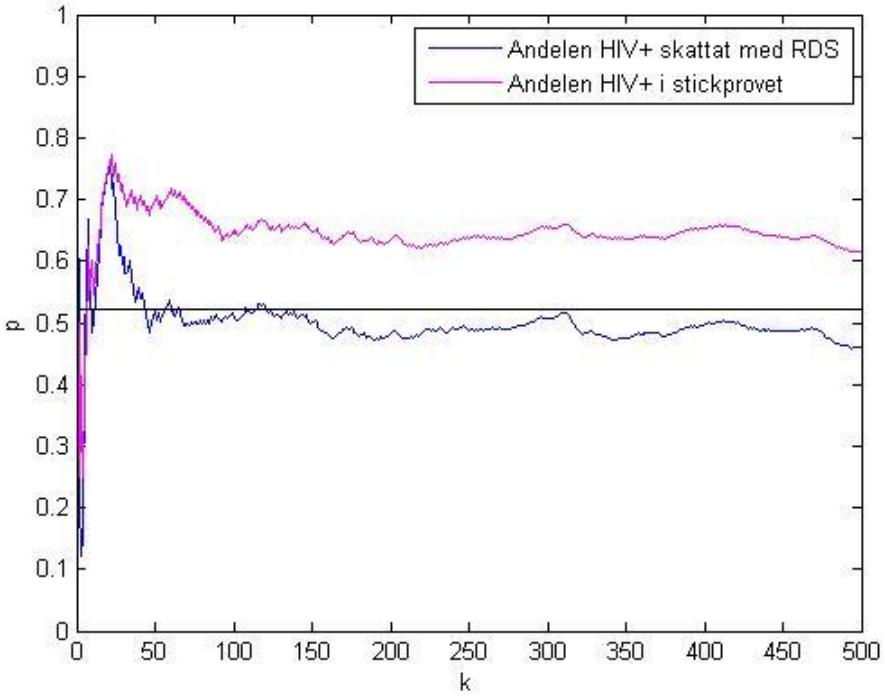
Skattningen för andelen i stickprovet verkar vara bättre. Om $\lambda$ är stor har de flesta individerna många vänner, så skillnaden mellan stickprovets andel HIV-positiva samt den sanna andelen i den stora komponenten blir mindre, därför blir även skattningen för andelen HIV-positiva i stickprovet bättre.




**Figur 4.3.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.11$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.1$ och $\lambda = 3$.




**Figur 4.4.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.11$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.1$ och $\lambda = 10$.

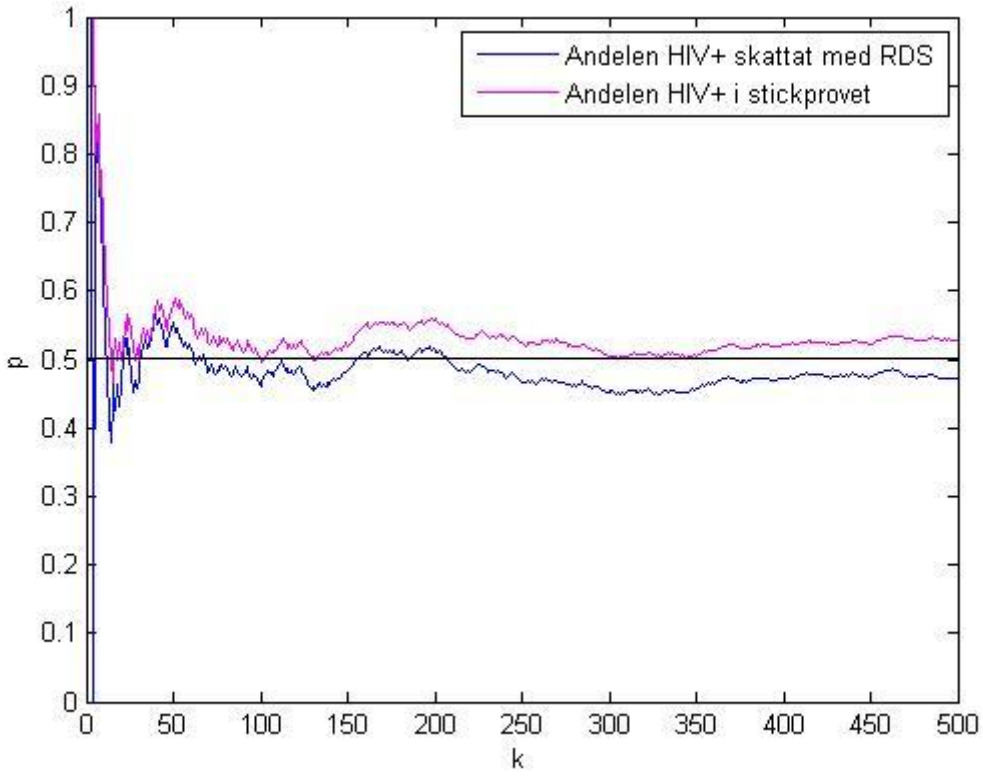



**Figur 4.5.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.52$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 5000, p_{HIV} = 0.5$ och $\lambda = 3$.

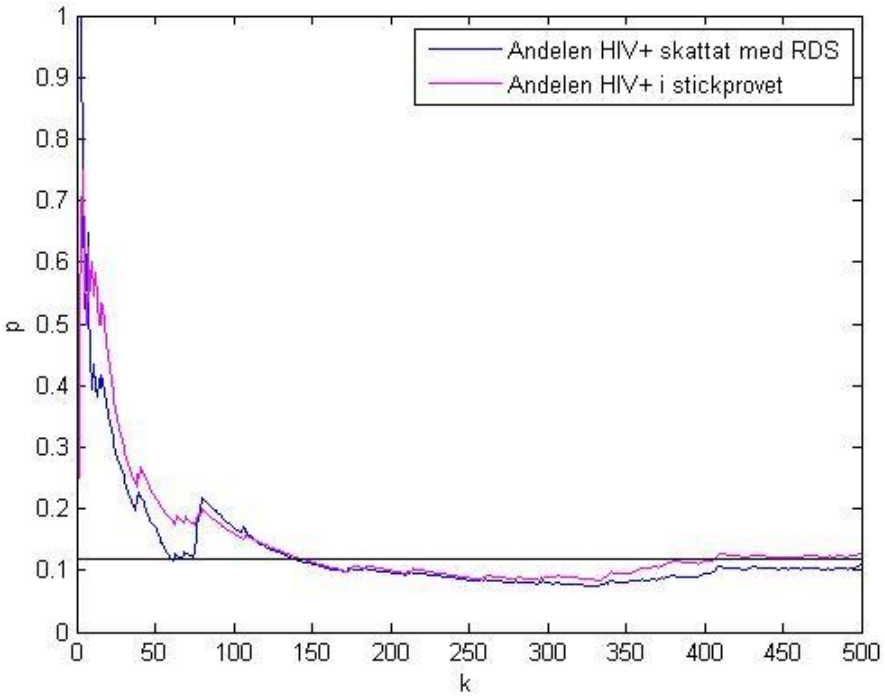
Samma värden på parametrarna används nu men storleken på populationen höjs till $N = 5000$. Jämförelse av figurerna 4.5 samt A5a-d I appendix med figurerna 4.1 samt A1a-d i appendix verkar det som att det är mer variation i skattningarna när k är litet när vi ökat populationsstorleken.

Med ett högre förväntat antal vänner, $\lambda$, så kan inte någon skillnad mellan en populationsstorlek på 1000 respektive 5000 individer uppfattas i plottarna. Genom att jämföra figur 4.2 samt A2a-d samt figurerna 4.6 samt A6a-b i appendix så verkar det som de närmar sig det sanna värdet med ungefär samma precision. Om figurerna A6c-d i appendix studeras avviker skattningen mer från den sanna andelen när k är litet jämfört med modell 1.

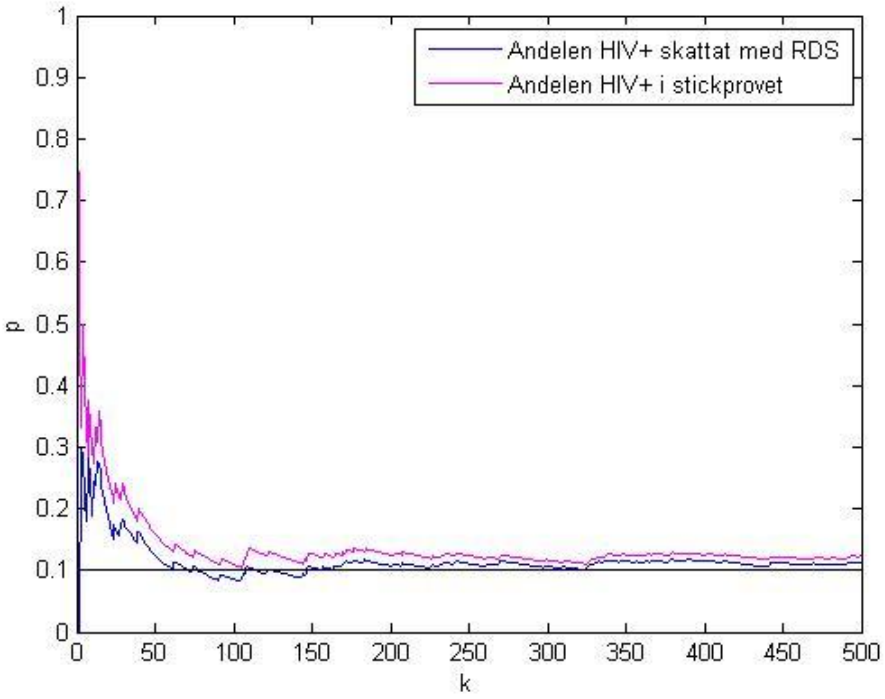
Den antagna andelen HIV-positiva i befolkningen minskas till $p_{HIV} = 0.1$. Genom att studera figurerna 4.7 samt A7a-d i appendix och 4.8 samt A9a-d i appendix och jämföra med motsvarande figurer för populationen på $N = 1000$ skådas att när $\lambda = 3$ blir skattningen med RDS inte lika god för $N = 5000$. När $\lambda = 10$ är variansen för fallet $N = 5000$ större när k är litet, jämfört med $N = 1000$.




**Figur 4.6.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.50$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 5000, p_{HIV} = 0.5$ och $\lambda = 10$.



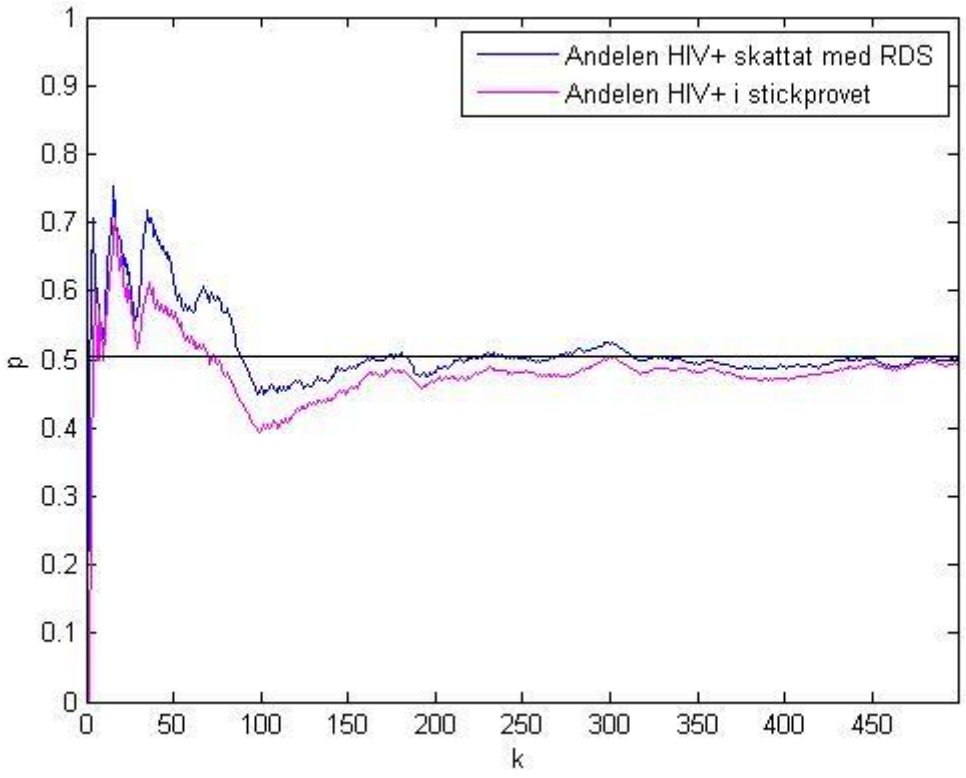

**Figur 4.7.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 5000, p_{HIV} = 0.1$ och $\lambda = 3$.




**Figur 4.8.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 5000, p_{HIV} = 0.1$ och $\lambda = 10$.

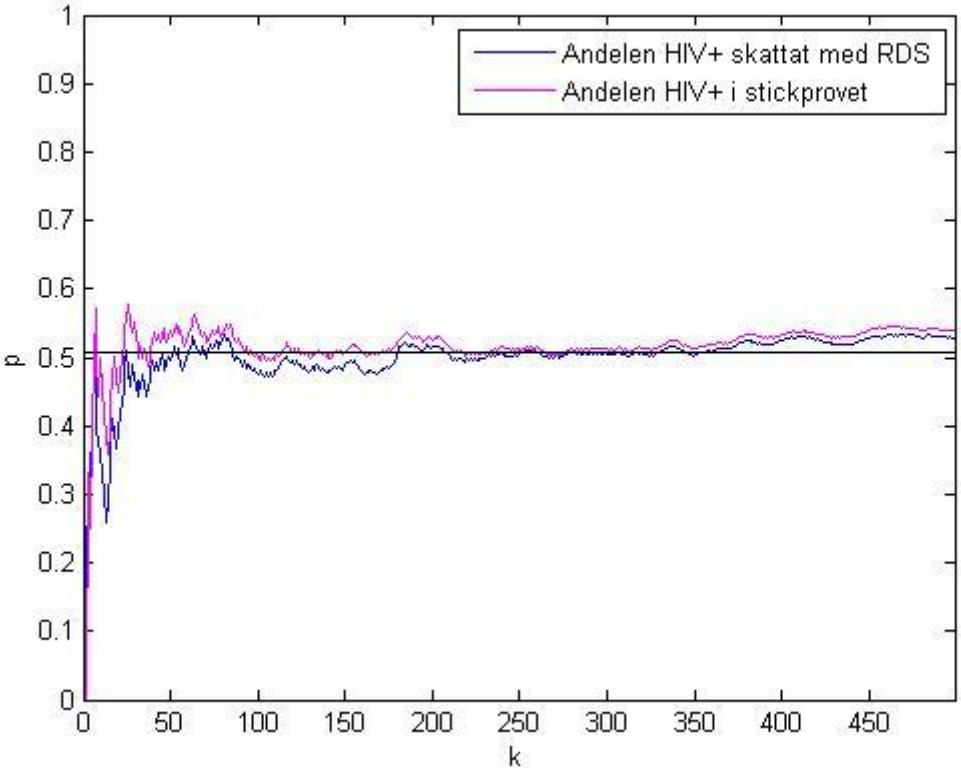
### 4.2. Oberoende och lika fördelat över populationen (modell 2)

Låt nu individer vara HIV-positiva oberoende och lika fördelat över hela populationen. Precis som i avsnitt 4.1. så kommer de med många vänner överrepresentera stickprovet. Däremot så ska det inte ha inverkan på fördelningen av HIV-positiva nu när vi slumpat ut HIV-positiva oberoende och lika fördelat över hela populationen. Eftersom de med fler vänner då inte har en större sannolikhet för att vara HIV-positiva. Denna modell testas på en populationsstorlek $N = 1000$ och de olika värden vi som testas för de olika parametrarna är $p_{HIV} = 0.5\ och\ p_{HIV} = 0.1, \lambda = 3\ och\ \lambda = 10$.




**Figur 4.9.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.50$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.5$ och $\lambda = 3$.

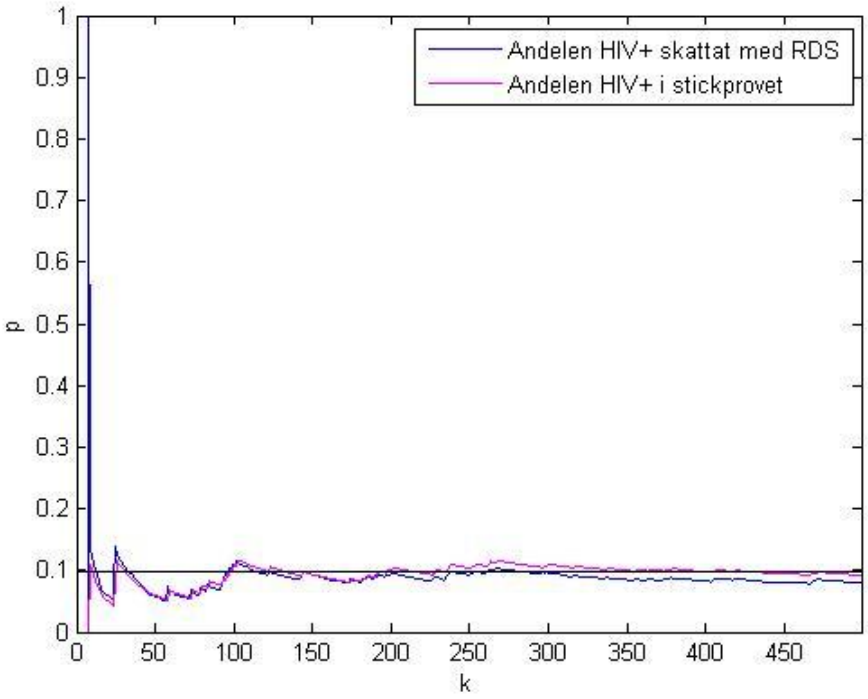
I figur 4.9 ses att andelen i stickprovet närmar sig den sanna andelen i den stora komponenten, likaså skattningen med RDS. I figur 4.10 plottas grafen när det förväntade antalet vänner en individ har höjts. På samma sätt som i figur 4.9 fås bra skattningar med båda metoderna. Det kan urskiljas att även här fås skattningar med bättre precision när det förväntade antalet vänner, $\lambda$, höjts.



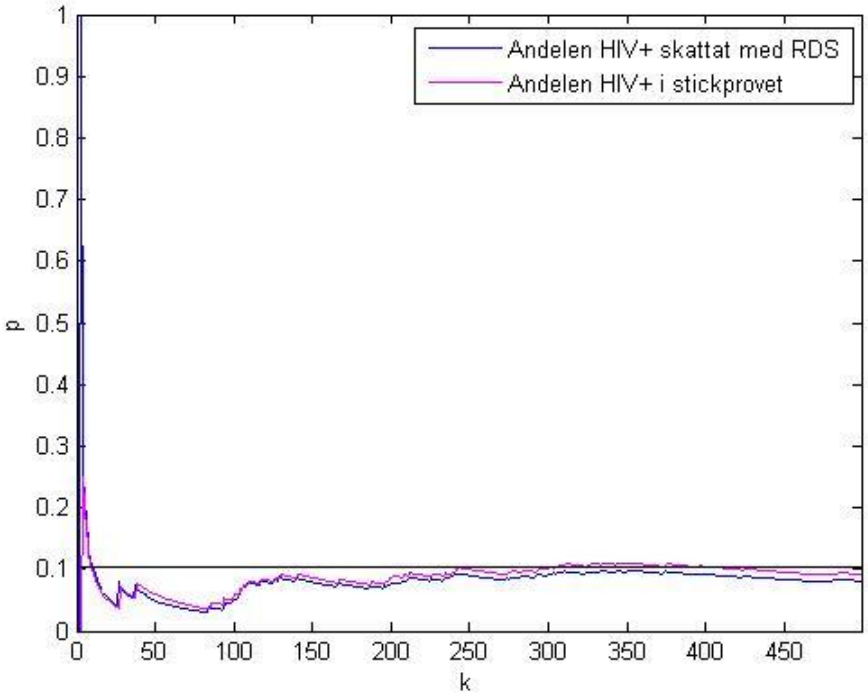

**Figur 4.10.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.51$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.5$ och $\lambda = 10$.

I figur 4.11 samt 4.12 ses hur skattningen närmar sig den sanna andelen HIV-positiva i den stora komponenten, $p_{HIV} = 0.1$.

Andelen HIV-positiva i stickprovet närmar sig andelen HIV-positiva i den stora komponenten, även för RDS. Som vi tidigare konstaterat har de med många vänner en större sannolikhet att bli dragna till stickprovet. I förra avsnittet slumpades HIV-status ut med stigande sannolikhet mot antalet vänner en individ har. När vi slumpar ut HIV-status lika fördelat över hela populationen så kommer stickprovet inte överrepresenteras av HIV-positiva, utan endast av de med många vänner. Stickprovet innehåller fortfarande en majoritet av sociala individer men kan nu representera hela befolkningen när man ser till utbredningen av HIV-positiva. Stickprovsandelen är en bra skattning när sannolikheten för HIV är oberoende av antalet vänner en individ har, detta är dock sällan fallet i verkligheten.



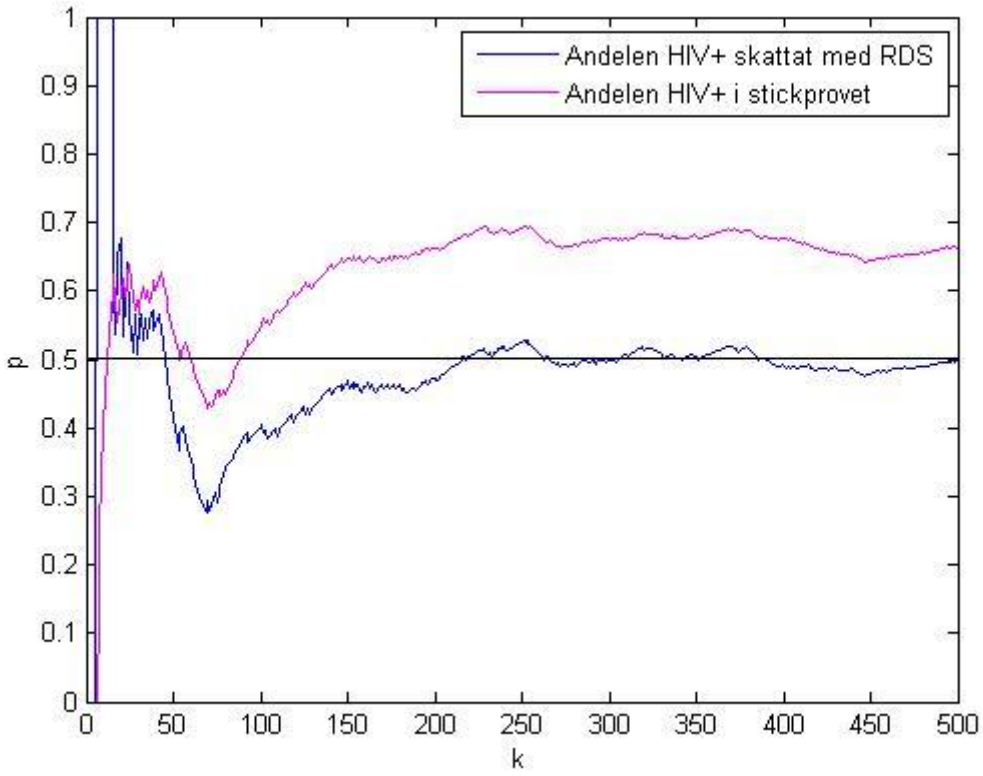

**Figur 4.11.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.1$ och $\lambda = 3$.




**Figur 4.12.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.1$ och $\lambda = 10$.

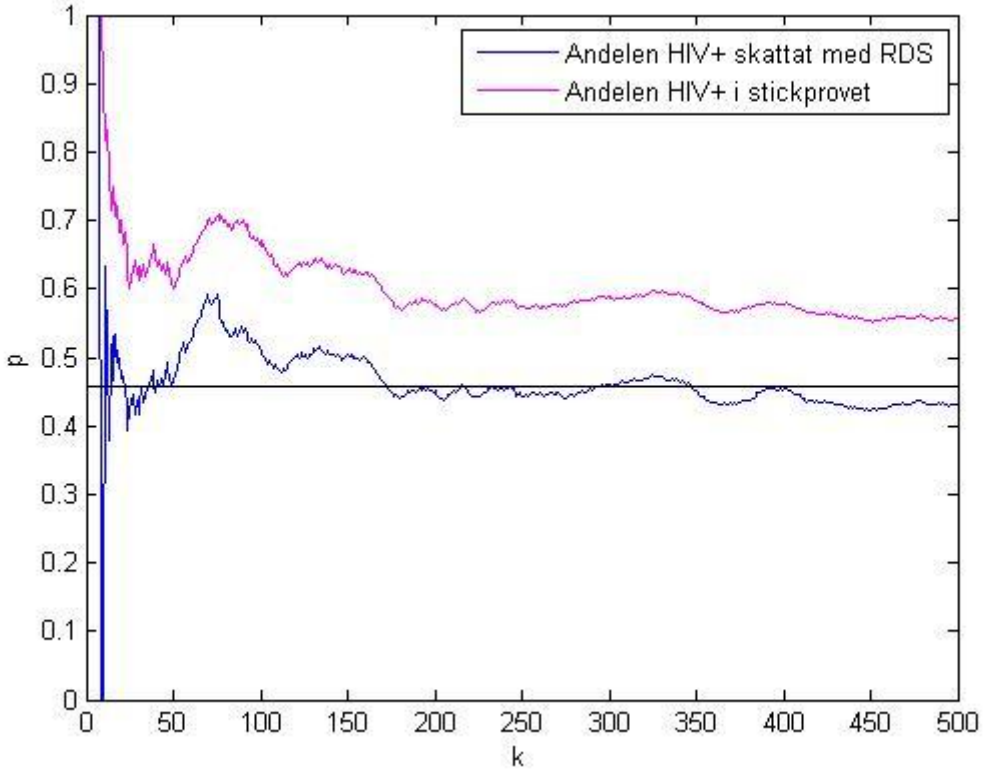
### 4.3. RDS på grafer med fler trianglar (modell 3)

Om individ $i$ är vän med individ $j$ och individ $k$ så finns det en förhöjd sannolikhet $\alpha$ att också individ $j$ och $k$ är vänner. Nu slumpas endast HIV-status ut med en ökande sannolikhet till antalet vänner man har. Eftersom sannolikheten för att en individ är HIV-positiv ökar med antalet vänner den har och på grund av att det finns fler vänskapsband där tripletter ingår så måste $\lambda$ och $p$ anpassas för att få en inte alltför hög andel HIV-positiva i populationen. Vi kommer att testa på en populationsstorlek på $N = 1000$ och använda oss av $\alpha = 0.3$ och $\lambda = 2$ samt $\alpha = 0.2, \lambda = 3$ och $p_{HIV} = 0.35$ och $p_{HIV} = 0.07$.




**Figur 4.13.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.50$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.35$ och $\lambda = 2$ och $\alpha = 0.3$.

I figur 4.13 så verkar skattningen med RDS fungera bra på grafer där fler trianglar ingår. Det verkar som att variansen för skattningarna är större om vi jämför med Modell1. Vi ändrar nu värdena på $\lambda$ och $\alpha$ till $\lambda = 3$ och $\alpha = 0.2$. Jämförelse med figur 4.13 och figur 4.14 samt figurerna A9a-d samt figurerna A10a-d i appendix så kan ingen stor skillnad mellan dessa urskiljas. Skattningen med RDS verkar dock närma sig det sanna värdet, men variationen är större om man jämför med skattningarna i modell 1.

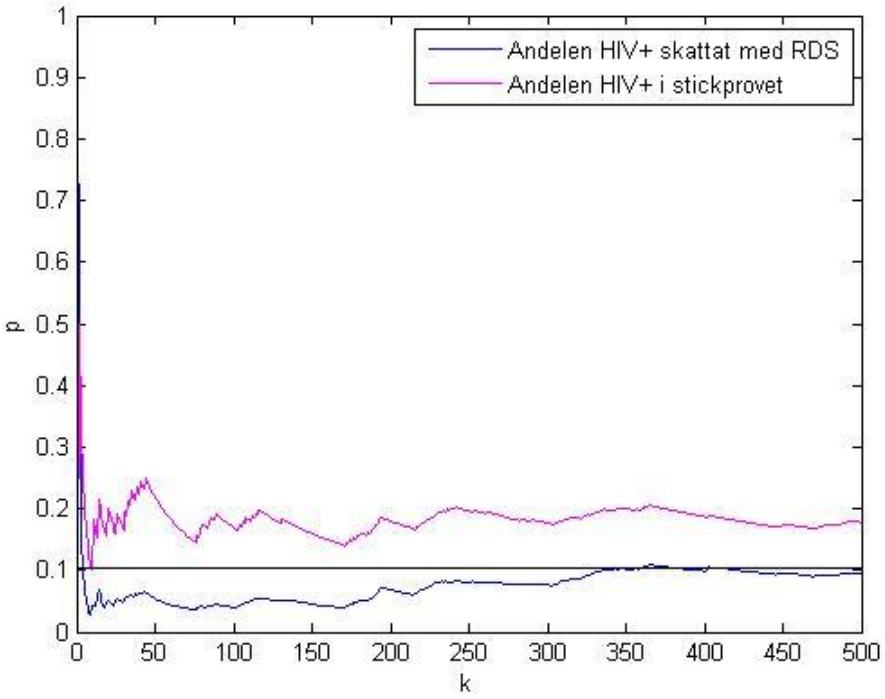



**Figur 4.14.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.47$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.35$ och $\lambda = 3$ och $\alpha = 0.2$.

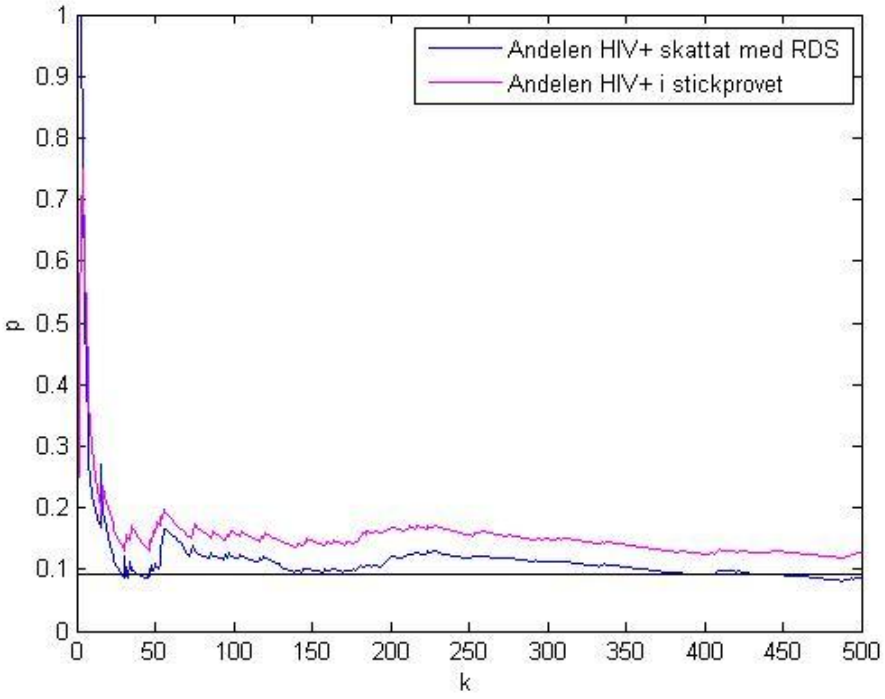
I figurerna 4.15 och 4.16 samt figurerna A11a-d och A12a-d i appendix skådas hur skattningen beter sig när vi ändrar den antagna sannolikheten för HIV till $p_{HIV} = 0.07$. Skattningens varians verkar vara större när k är litet om man jämför med modell 1.

När fler trianglar finns med i nätverket verkar skattningen av andelen HIV-positiva med RDS verkar konvergera mot andelen HIV-positiva i den stora komponenten i långsammare takt. Variationen i skattningen med RDS är större här när k är litet jämfört med i modell 1.

För att bäst se vad fler trianglar i grafen får för effekt så måste $\alpha$ höjas. Vi vill dock inte att individerna ska ha allt för många vänner och genom testkörningar så kan vi komma fram till att ovan valda värden på $\alpha$ och $\lambda$ är bra. Vi vill inte välja $\lambda$ för lågt eftersom då kommer många ha 0 vänner i populationen och i stickprovet kommer många endast ha en vän. Skattningarna för nätverk med fler trianglar verkar konvergera mot andelen HIV-positiva i den stora komponenten i långsammare takt.



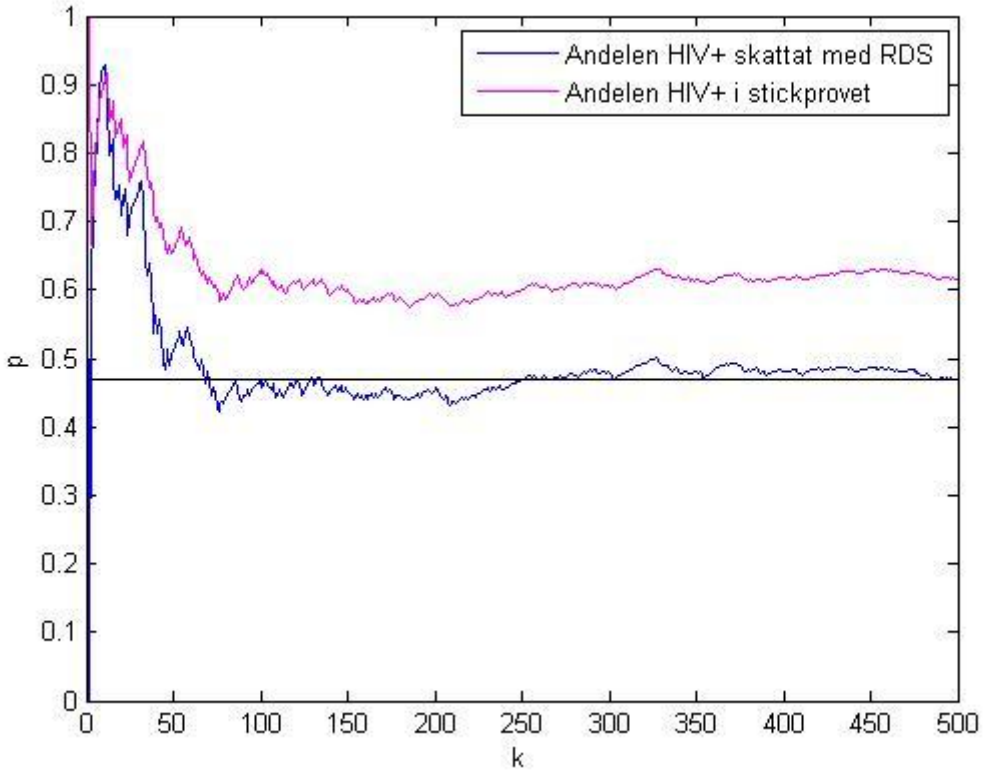

**Figur 4.15.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.07$ och $\lambda = 2$ och $\alpha = 0.3$.




**Figur 4.16.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.10$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, p_{HIV} = 0.07$ och $\lambda = 3$ och $\alpha = 0.2$.

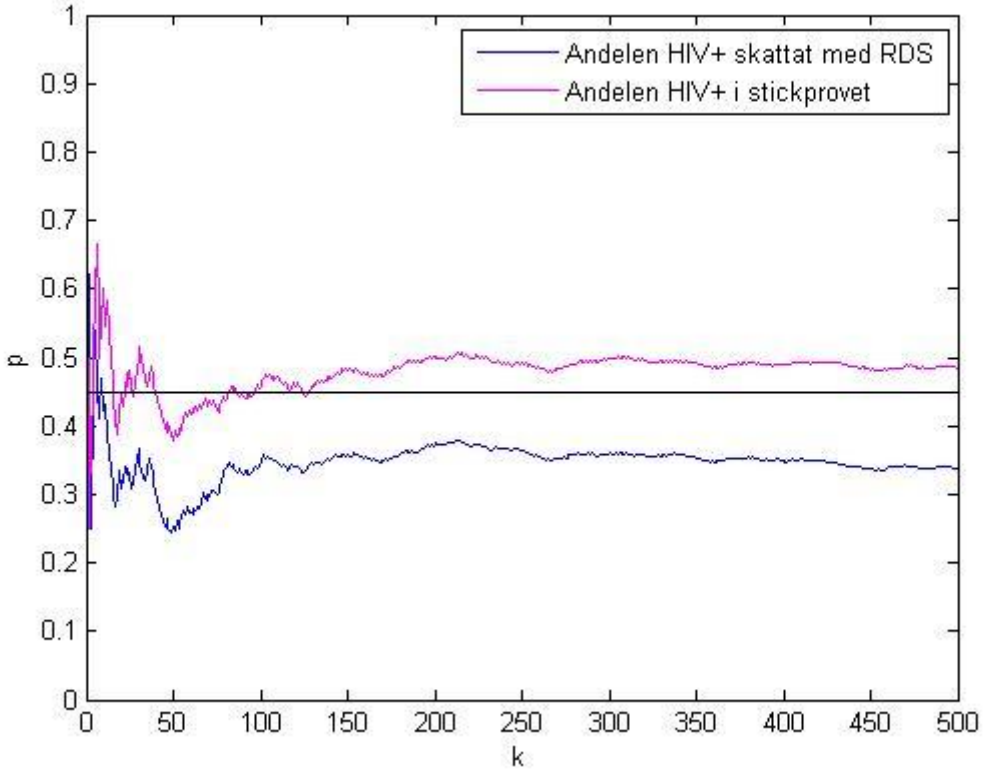
### 4.4. RDS på riktade grafer (modell 4)

Enligt teorin fungerar RDS endast om nätverket består av oriktade kanter. Vi ska nu undersöka hur bra RDS fungerar om vänskapsbanden inte enbart är ömsesidiga, det vill säga att kanterna kan vara riktade. Sannolikheten $\beta$ är den förhöjda sannolikheten för att kanten även är riktad åt andra hållet (vi får då ett ömsesidigt vänskapsband). Vi kommer variera värdena på $\beta$ för att se hur pass väl RDS fungerar i de olika fallen. De värden vi kommer testa på $\beta$ är $\beta = 0.3$ och $\beta = 0.8$.




**Figur 4.17.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.48$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, \beta = 0.3$, $p_{HIV} = 0.35$ och $\lambda = 3$.

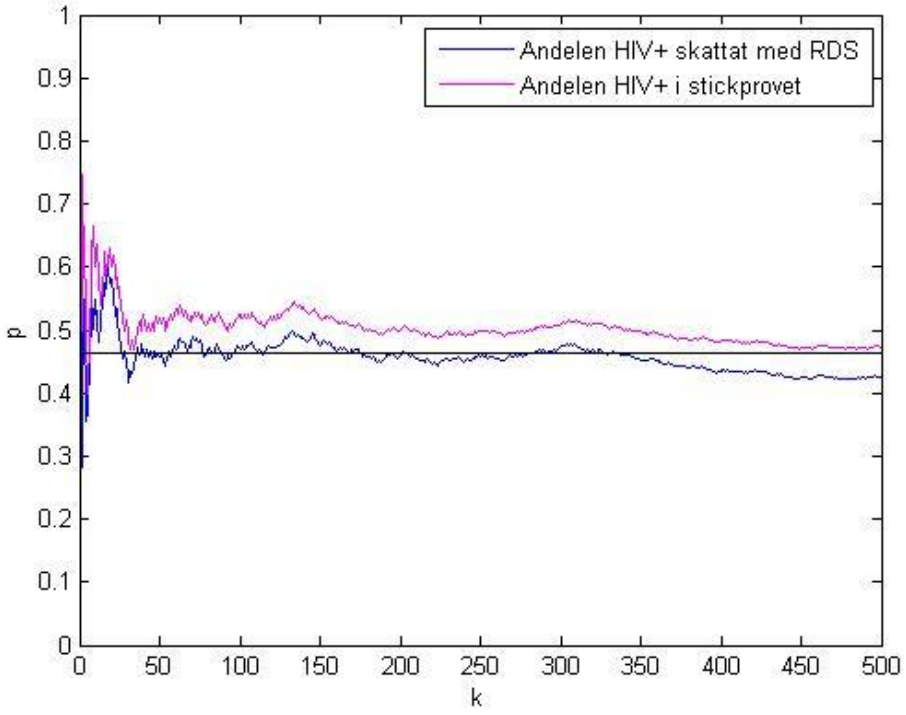
I figur 4.17 ser skattningen med RDS ut att fungera väl. I figur 4.18 kan en annan körning med samma parametrar som vi använde i figur 4.17 studeras. Skattningen blir inte lika bra. Studeras figurerna A13a-d i appendix verkar skattningen med RDS inte fungera bra i denna modell.



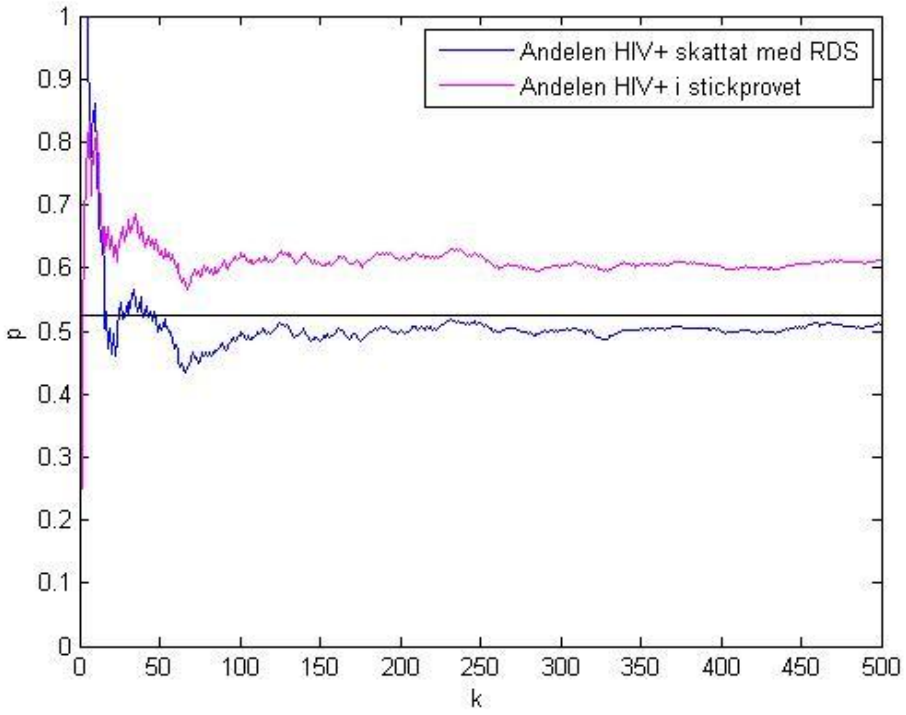

**Figur 4.18.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.45$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, \beta = 0.3$, $p_{HIV} = 0.35$ och $\lambda = 3$.

Det kan vara intressant att se vad som sker om vi ökar det förväntade antalet vänner en individ har, $\lambda$. I figur 4.19 samt figurerna A14a-d i appendix kan detta studeras. Skattningarna blir bättre när vi ökar $\lambda$ om vi jämför med föregående värden på parametrarna**.** Dock fungerar det inte lika bra som när endast ömsesidiga vänskapsband ingår.

Det kan vara intressant att se hur pass väl RDS fungerar när vi tillåter merparten av kanterna att vara oriktade men att även riktade kanter får ingå. Om vi sätter $\beta = 0.8$ och därmed tillåter färre riktade kanter kan man tänka sig att det endast är några riktade kanter som finns. Genom att studera figur 4.20 samt figurerna A15a-d i appendix kan det ses hur väl skattningen fungerar. Då är variansen stor när k är litet men närmar sig den sanna andelen HIV-positiva med lika god precision som i fallet där endast ömsesidiga vänskapsband ingår när k blir stort.




**Figur 4.19.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.46$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, \beta = 0.3$, $p_{HIV} = 0.35$ och $\lambda = 10$.




**Figur 4.20.** Sanna andelen HIV-positiva i den stora komponenten (horisontell linje), $\hat{p}_{SK} = 0.52$, och skattad andel HIV-positiva som funktion av stickprovsstorleken k=1, …, 500. Här är $N = 1000, \beta = 0.8$, $p_{HIV} = 0.3$ och $\lambda = 3$.

# 5 Slutsatser

RDS fungerar bra när vänskapsbanden mellan individer är ömsesidiga. Variationen i skattningarna verkar bli mindre när vi ökar det förväntade antalet vänner en individ har, det vill säga om vi ökar $\lambda$. Då vi ökar populationsstorleken från $N = 1000$ till $N = 5000$ går skattningen med RDS i lite långsammare takt mot sitt sanna värde när $\lambda = 3$.

När vi gör skattningen med RDS då fler trianglar ingår i grafen, har skattningen mer osäkerhet. Skattningen närmar sig sitt sanna värde men även här i långsammare takt än i förstnämnda fallet.

I grafer med merparten riktade kanter fungerar skattningen med RDS inte bra. Skattningen med RDS fungerar väl när vi låter färre riktade kanter finnas med. I det sistnämnda fallet fungerar skattningen med RDS lika bra som för modell 1.

# 6 Diskussion

RDS fungerar väl i de fall där endast ömsesidiga vänskapsband ingår i den gömda populationen. Antagandena säger också att detta måste vara fallet. Det visar sig i våra plottar att RDS verkar fungera även när det finns enstaka riktade kanter med i grafen. Ibland förekommer det ”fusk” rekryteringarna, det vill säga att en individ ger en kupong till någon den inte känner, vilket i så fall inte torde ha någon inverkan i och med vårt konstaterande.

Med högre väntevärde för antalet vänner är det till synes som att skattningen konvergerar fortare mot sitt väntevärde. I Salganik och Heckathorn (2004) kan vi läsa att det är bra om antalet kuponger per individ, *c*, är litet. Då går stickprovet genom många vågor innan den eftersträvade stickprovsstorleken är uppnådd. Med långa rekryteringskedjor kommer också alla medlemmar i populationen ha en chans att bli dragna till stickprovet. Om många har många vänner kommer de individerna i populationen som har få vänner ha en större sannolikhet att bli dragna än om alla individer i populationen har få vänner.

I graferna där fler trianglar ingår närmar sig skattningen det sanna värdet lite långsammare jämfört med modell 1. Eftersom det är många trianglar av vänner kan det vara fallet att det finns en ökad sannolikhet att man står och ”hoppar” bland en grupp individer.

En fördel med RDS är att informationen man kan tillgodose sig med från det sociala nätverket gör att man kan få en billig och riktig stickprovs- och skattningsteknik. Ett problem med

insamlandet av ett RDS stickprov är att det kan medföra att det blir en viss skevhet i stickprovet om inte rekryteringarna är slumpmässiga. Ett annat problem är att individer kan försöka vara med i studien flera gånger för ersättningens skull, vilket kan påverka data och därmed riktigheten i skattningarna med RDS. Det kan också vara ett problem att inhämta den korrekta graden för en individ samt att individer som inte tillhör den population man vill studera försöker vara med i studien.

# 7 Referenser


[1] Britton, T. (2009). *Networks and respondent driven sampling: a survey.* Kompendium, Stockholms Universitet.

[2] Cao, J. (2010). *Erdös-Rényi-grafer* (2010:1). Kandiduppsats, Stockholms Universitet. Hämtad från http://www2.math.su.se/matstat/reports/seriec/2010/rep1/report.pdf

[3] Erickson, B.H. (1979). *Some problems on inference from chain data*. Sociological Methodology, Vol. 10, 276-302

[4] Gile, K.J., Handcock, M.S. (2009). *Respondent-Driven Sampling: An Assessment of Current Methodology*, arXiv:0904.1855v1 [stat.AP].

[5] Gut, A. (1995). *An Intermediate Course in Probability.* Springer.

[6] Heckathorn, D. D. (1997). *Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations*. Social Problems, Vol. 44:2, 174-197.

[7] Salganik, Matthew J., Heckathorn D.D. (2004). *Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling*. Sociological Methodology.

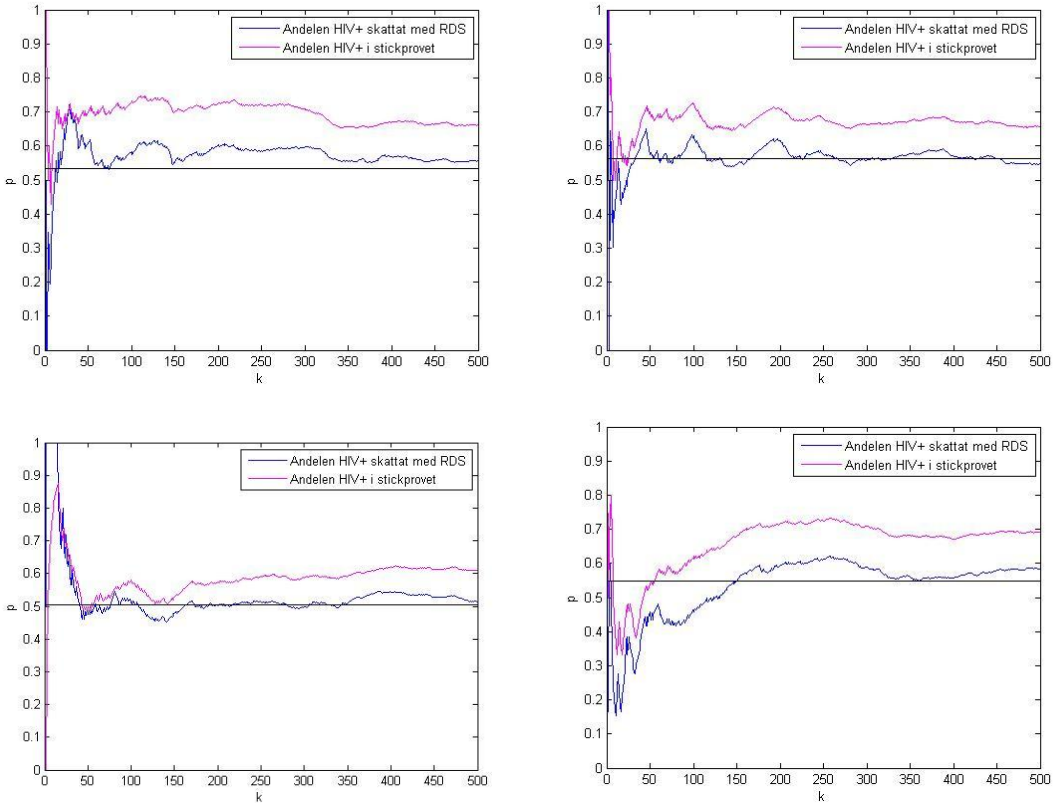
[8] Volz, E., Heckathorn D.D. (2008). *Probability Based Estimation Theory for Respondent Driven Sampling*. Journal of official Statistics, Vol. 24:1, 79-97.

[9] *What is Respondent Driven Sampling?*, Respondent Driven Sampling. Hämtad 17 mars 2010 från http://www.respondentdrivensampling.org.

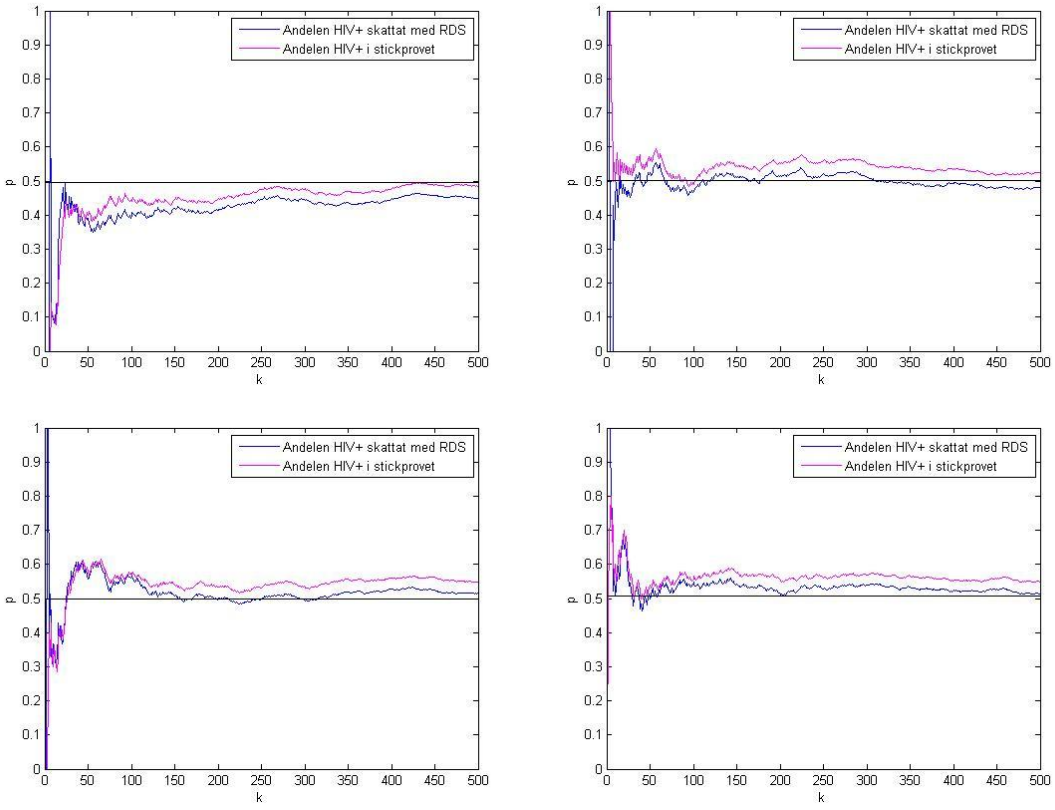
[10] http://en.wikipedia.org/wiki/Snowball_sampling. Hämtad januari 2010.


# 8 Appendix

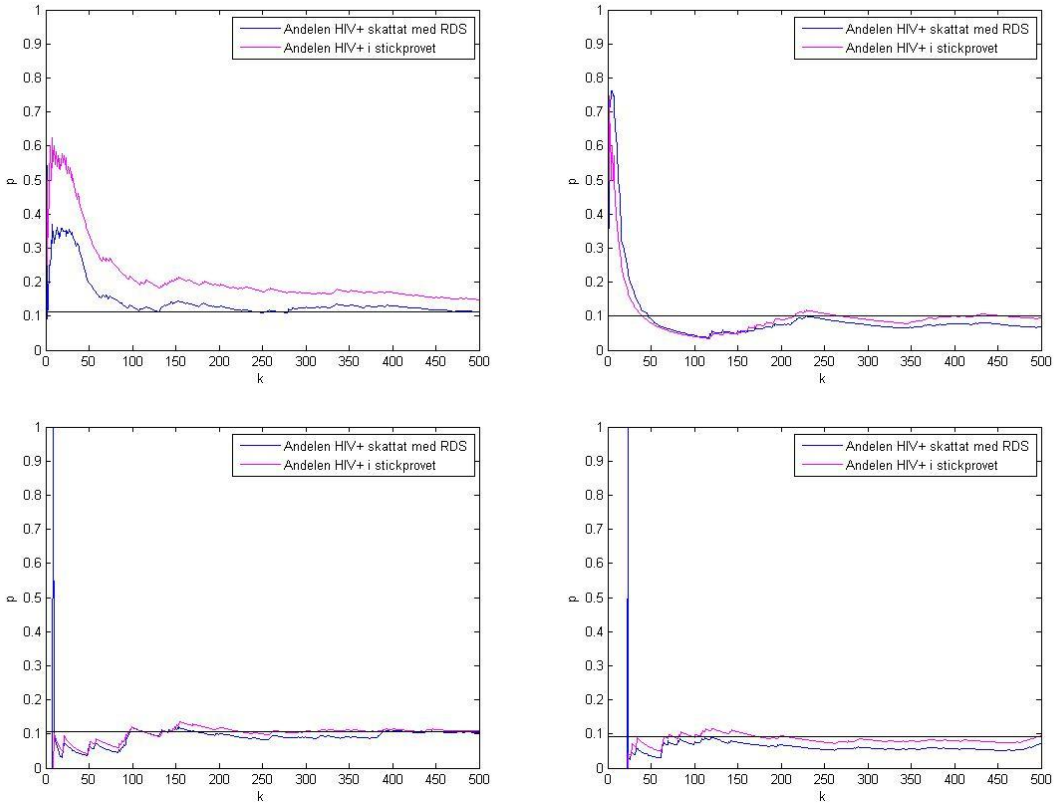
Här återfås figurer för modell 1, modell 3 samt modell 4. Figurerna visar den skattade andelen HIV-positiva som funktion av stickprovsstorleken k=1, …, 500 och läses i ordningen vänster till höger. Figurerna A1a-d till A8a-d är figurer för modell 1. Figurerna A9a-d till A12a-d är figurer för modell 3. Figurerna A13a-d till A15a-d är figurer för modell 4. Den horisontella linjen i figurerna är den sanna andelen HIV-positiva i den stora komponenten.



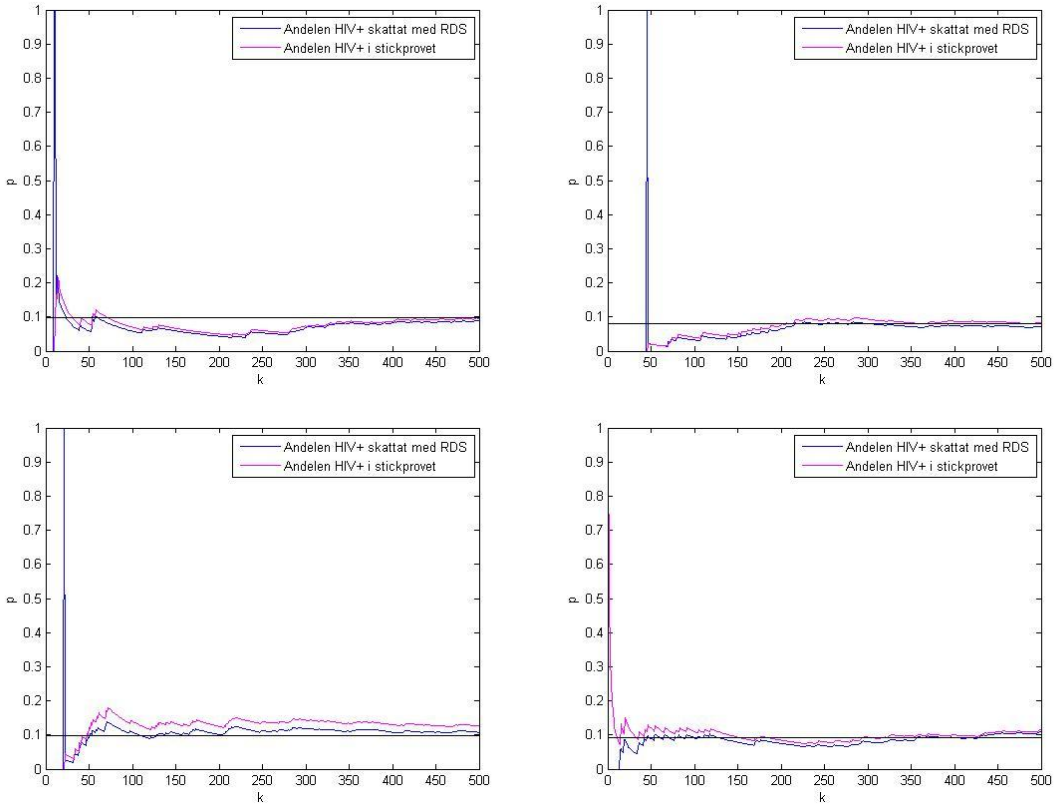

**Figur A1a-d.** Modell 1. Här är $N = 1000,\ p_{HIV} = 0.5$ och $\lambda = 3$.



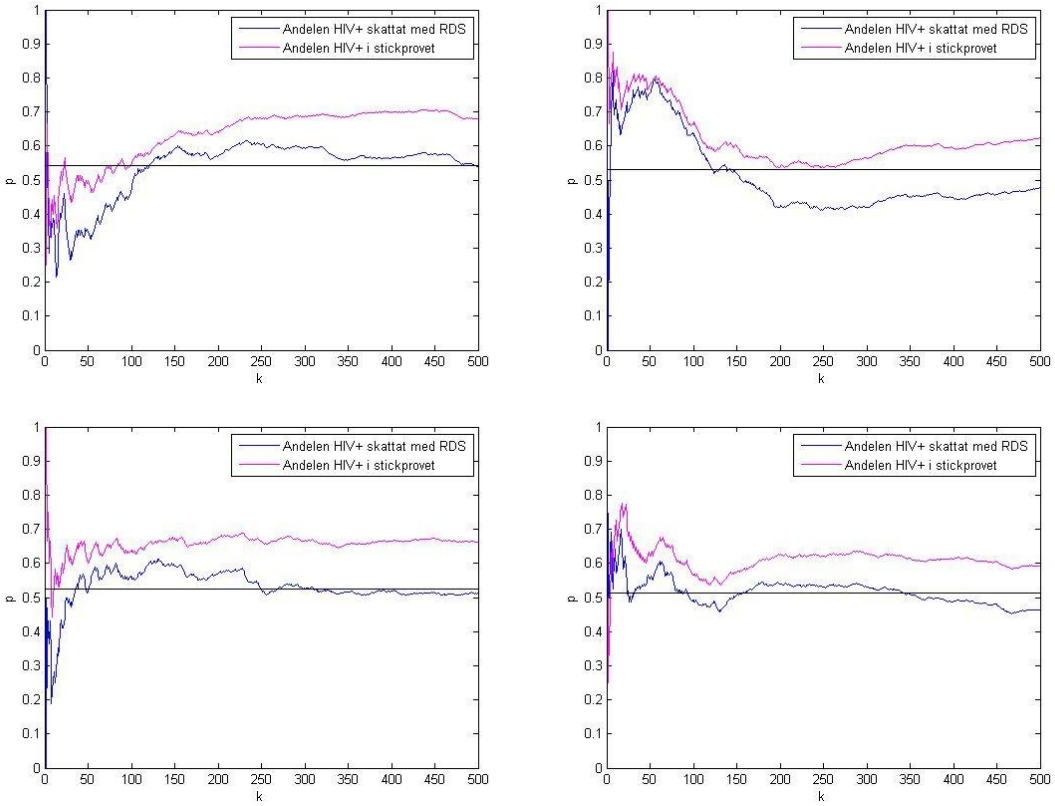

**Figur A2a-d.** Modell 1. Här är $n = 1000, p_{HIV} = 0.5$ och $\lambda = 10$.



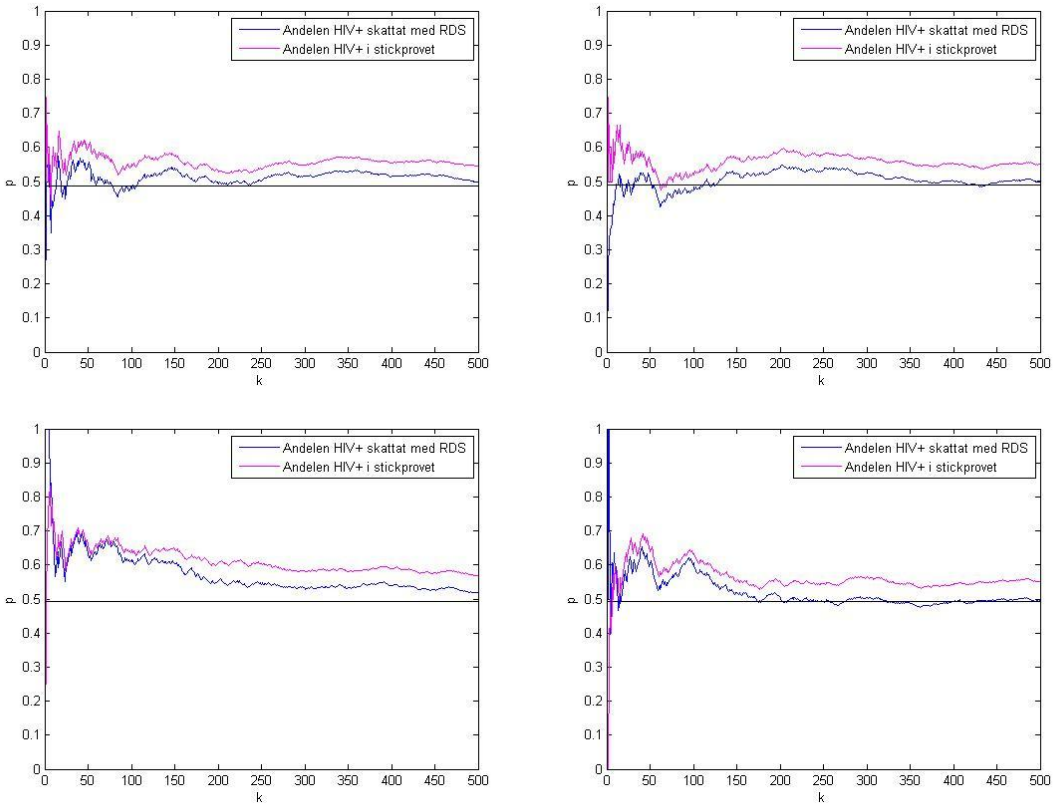

**Figur A3a-d.** Modell 1. Här är $N = 1000,\ p_{HIV} = 0.1$ och $\lambda = 3$.



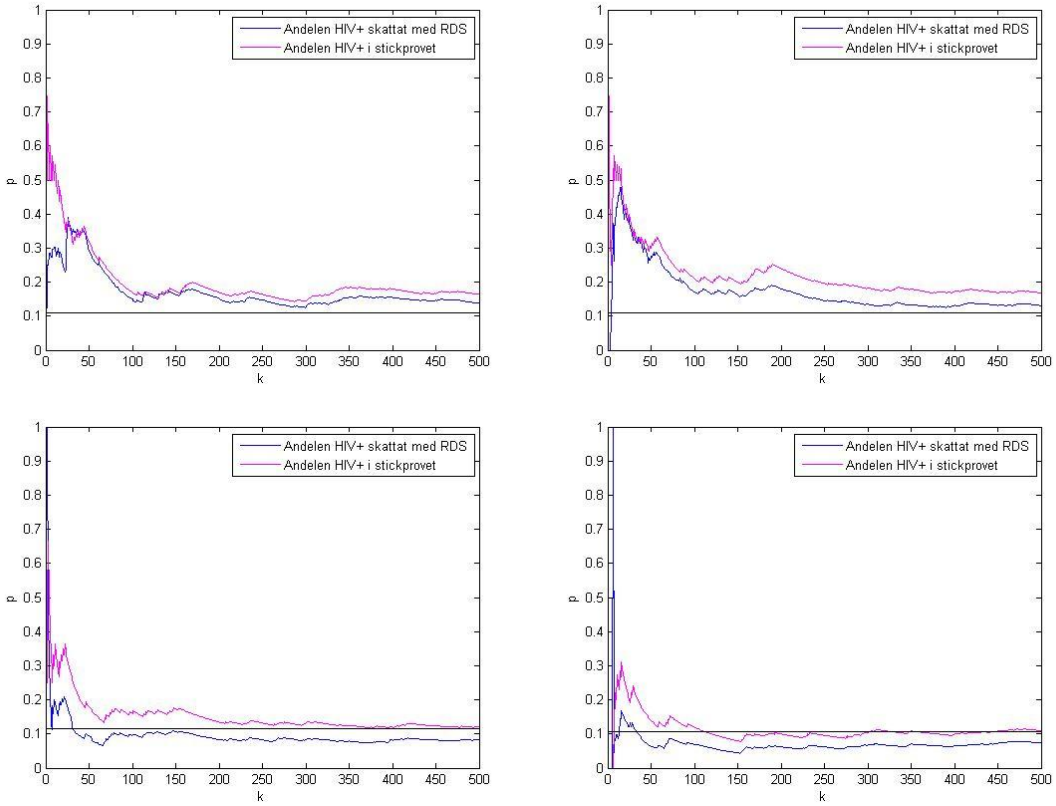

**Figur A4a-d.** Modell 1. Här är $N = 1000$, $p_{HIV} = 0.1$ och $\lambda = 10$.



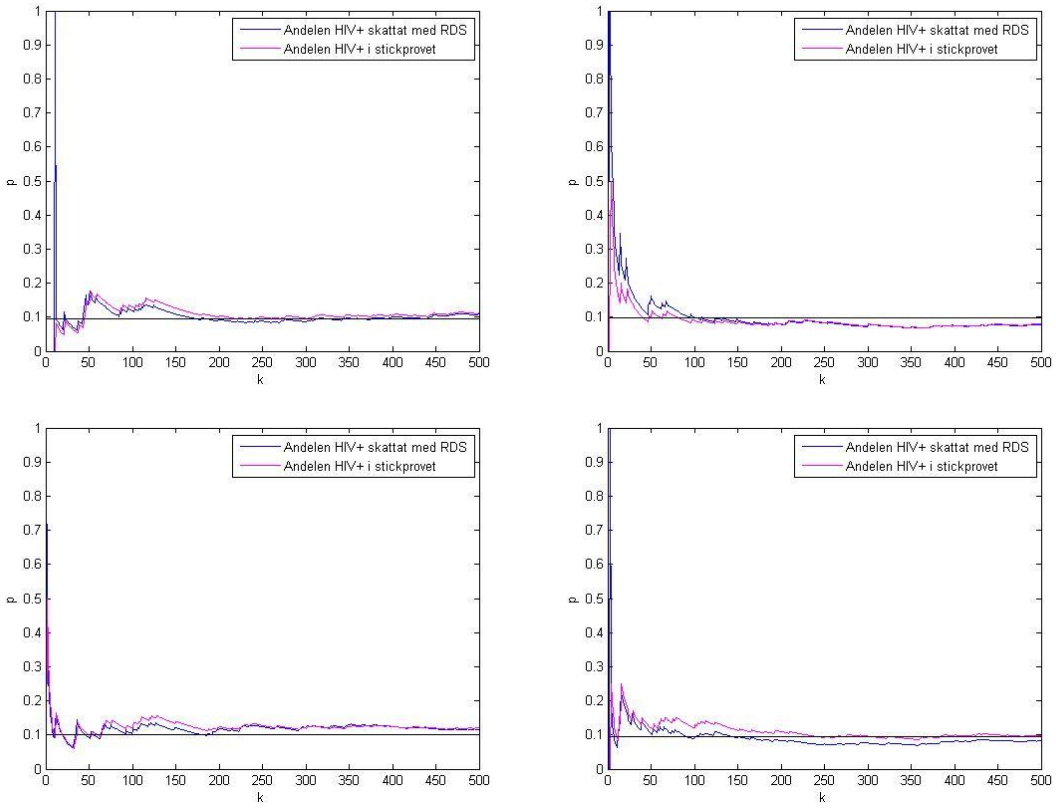

**Figur A5a-d.** Modell 1. Här är $N = 5000,\ p_{HIV} = 0.5$ och $\lambda = 3$.



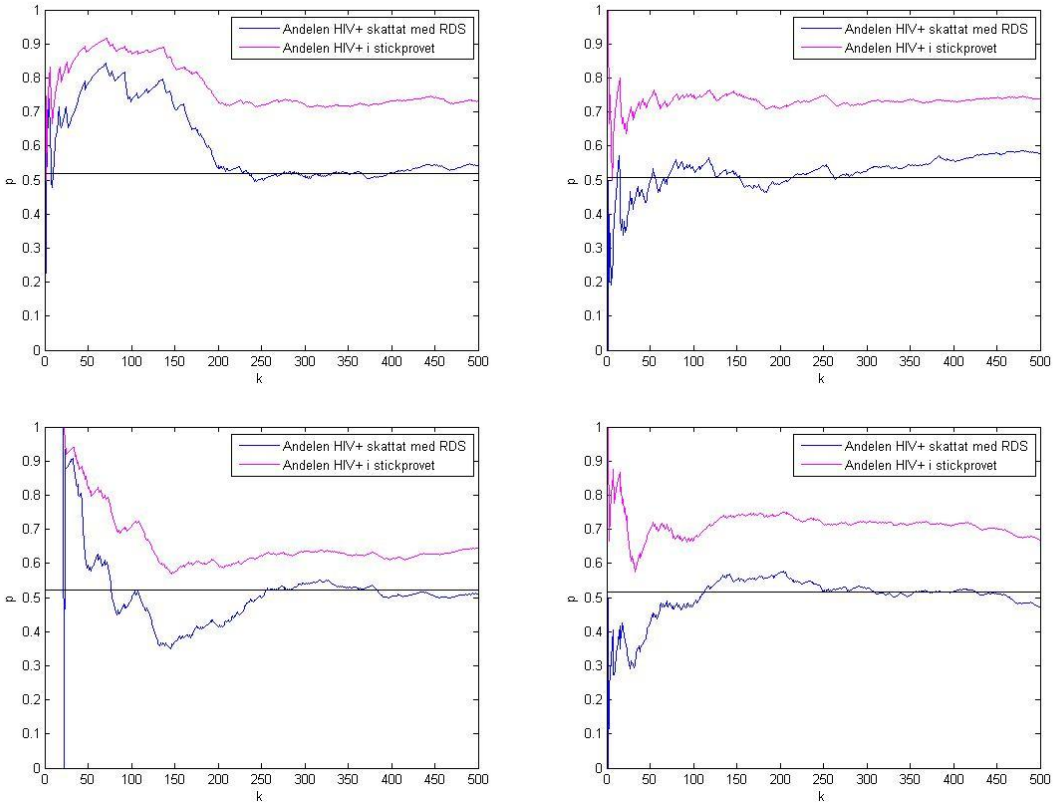

**Figur A6a-d.** Modell 1. Här är $N = 5000,\ p_{HIV} = 0.5$ och $\lambda = 10$.



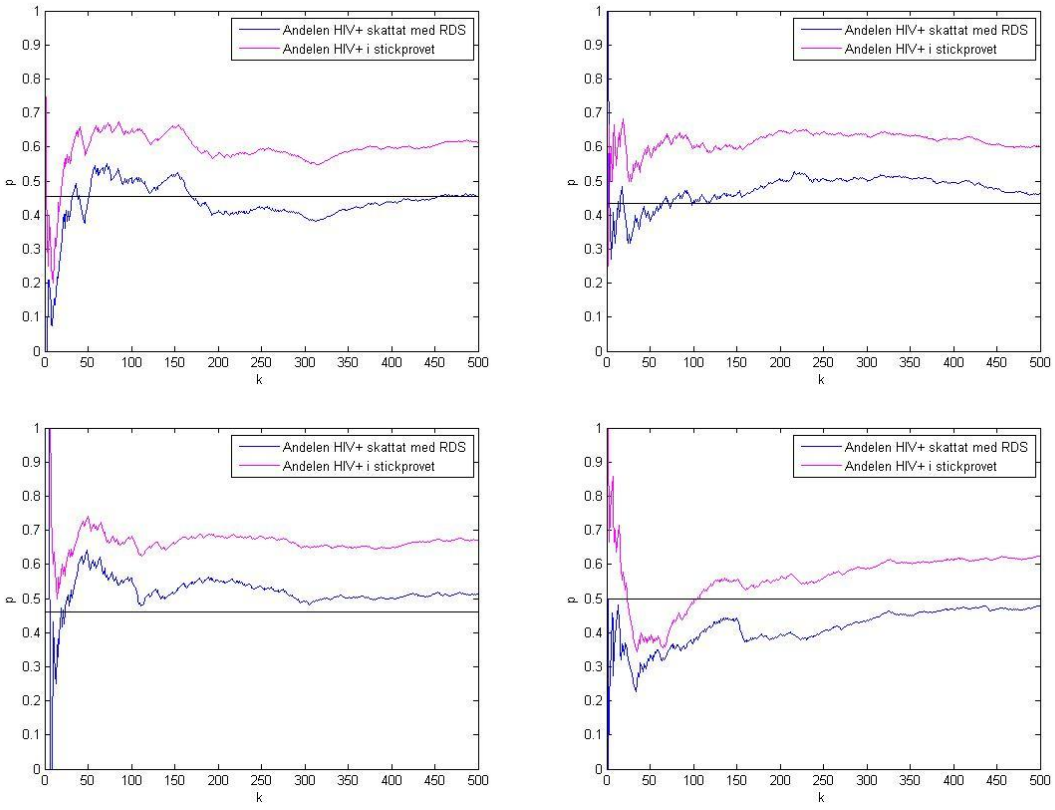

**Figur A7a-d.** Modell 1. Här är $N = 5000,\ p_{HIV} = 0.1$ och $\lambda = 3$.



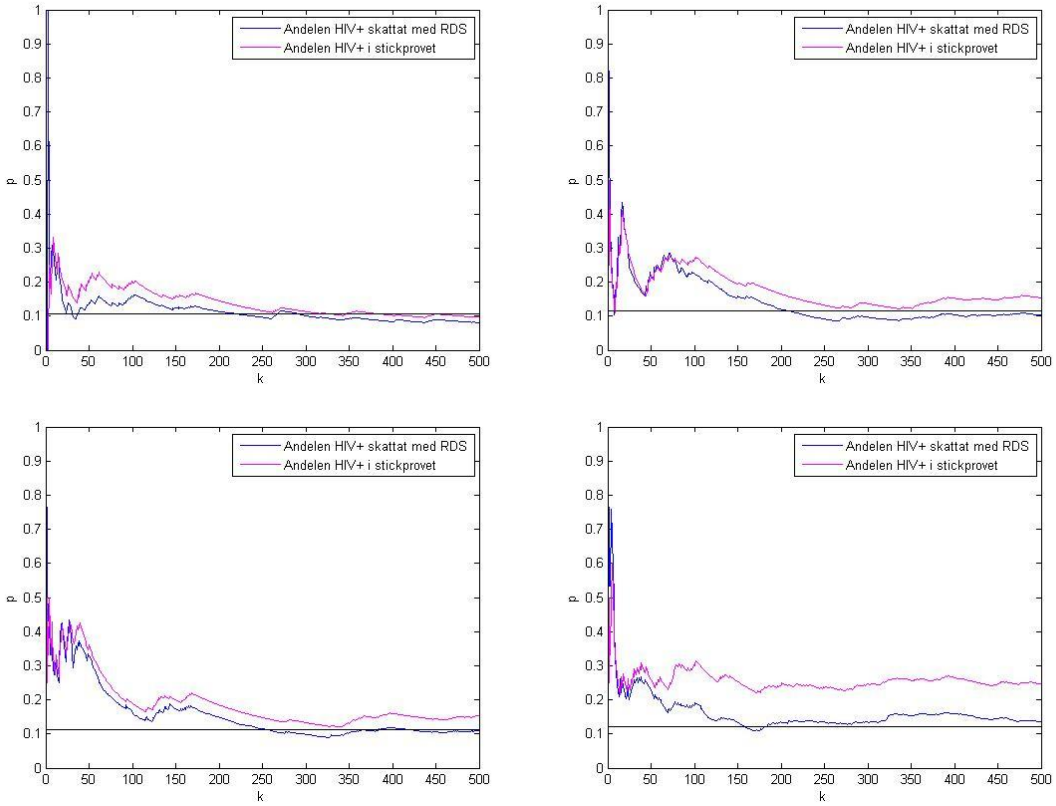

**Figur A8a-d.** Modell 1. Här är $N = 5000,\ p_{HIV} = 0.1$ och $\lambda = 10$.



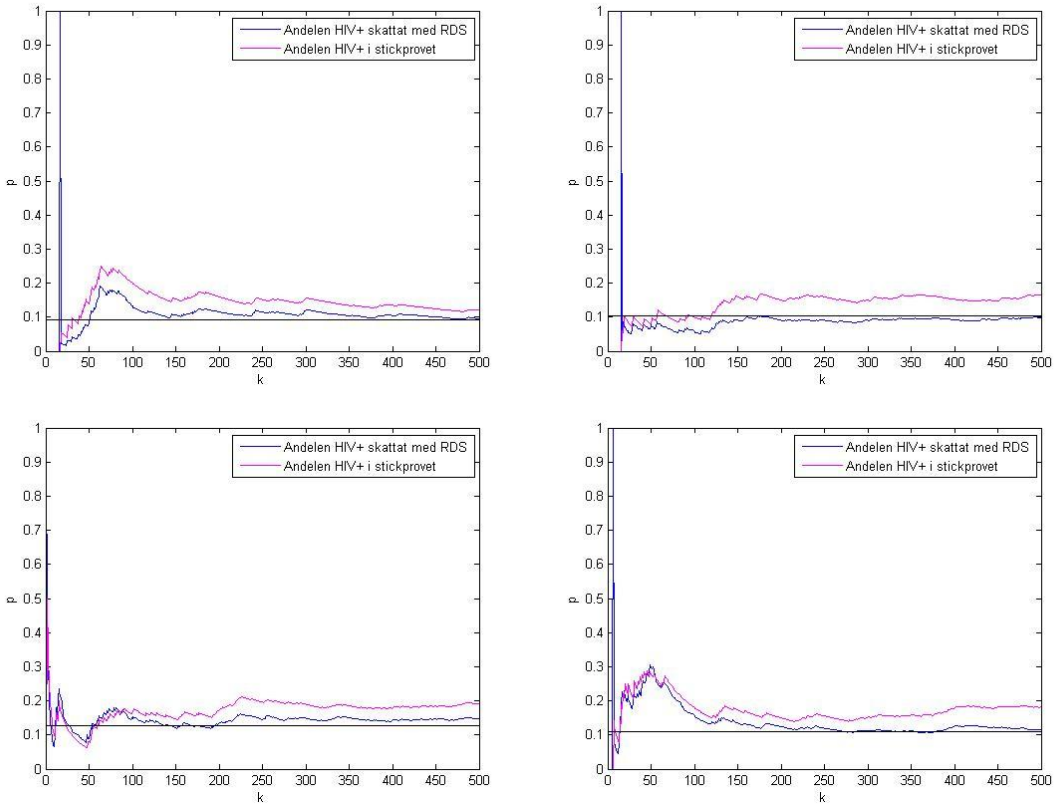

**Figur A9a-d.** Modell 3. Här är $N = 1000$, $p_{HIV} = 0.35$ $\lambda = 2$ och $\alpha = 0.3$.



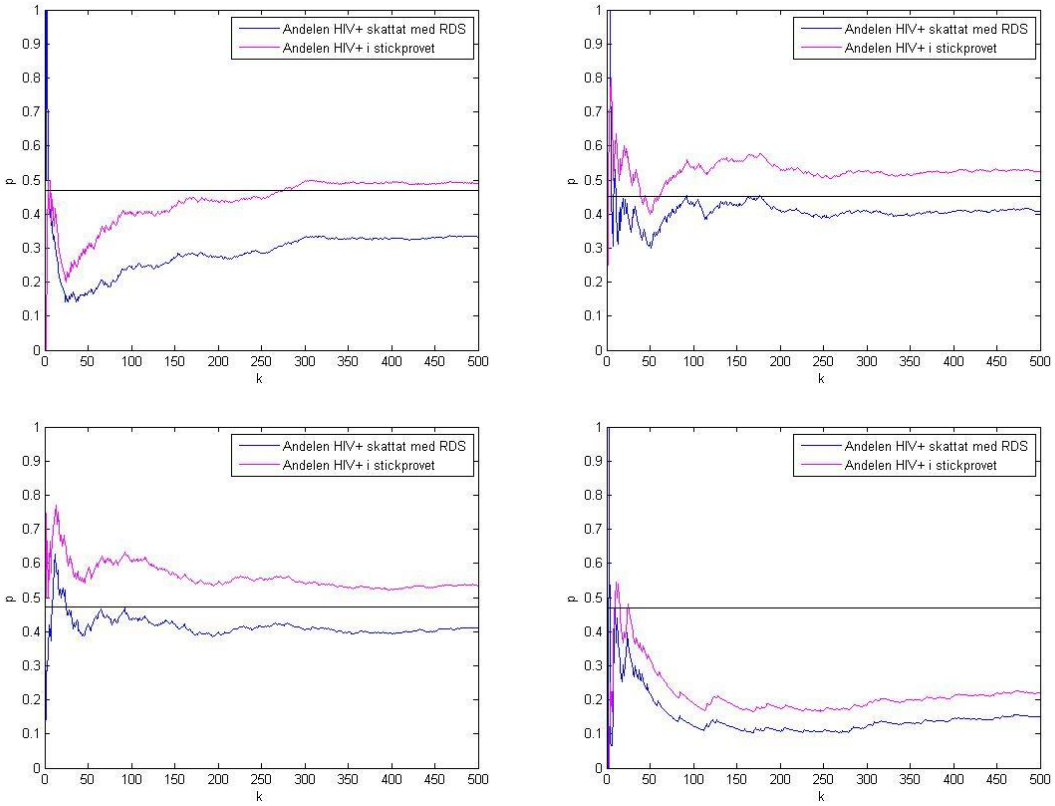

**Figur A10a-d.** Modell 3. Här är $N = 1000$, $p_{HIV} = 0.35$ $\lambda = 3$ och $\alpha = 0.2$.



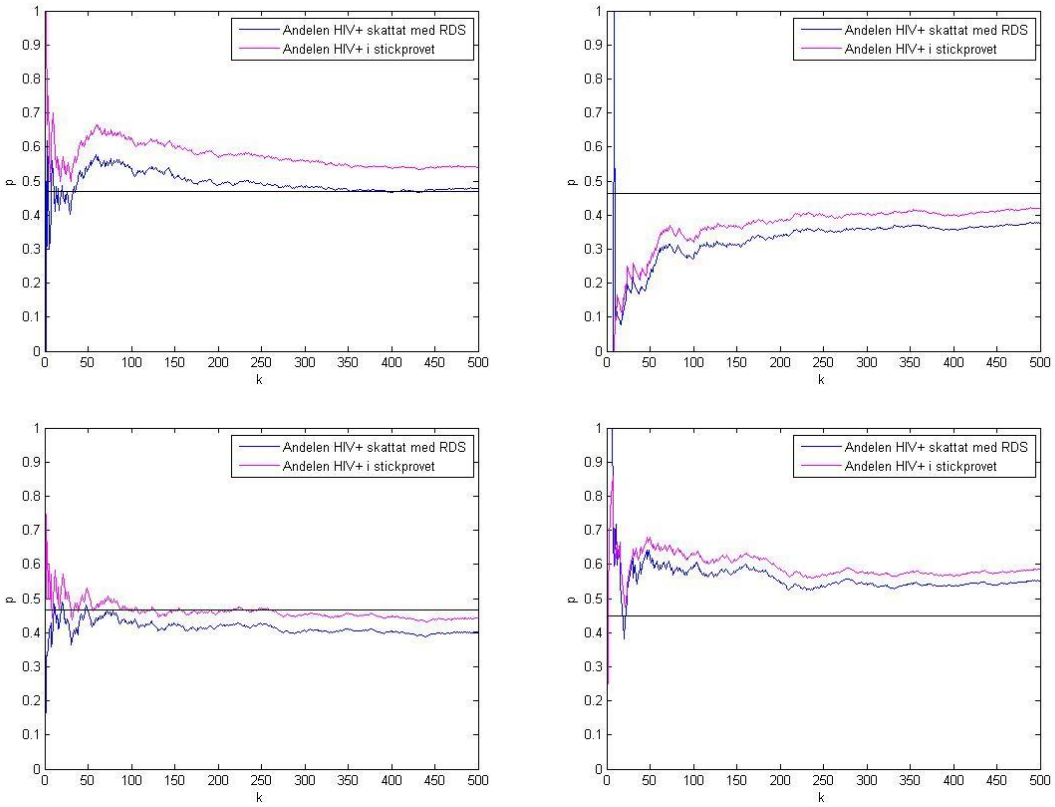

**Figur A11a-d.** Modell 3. Här är $N = 1000$, $p_{HIV} = 0.07$ $\lambda = 2$ och $\alpha = 0.3$.



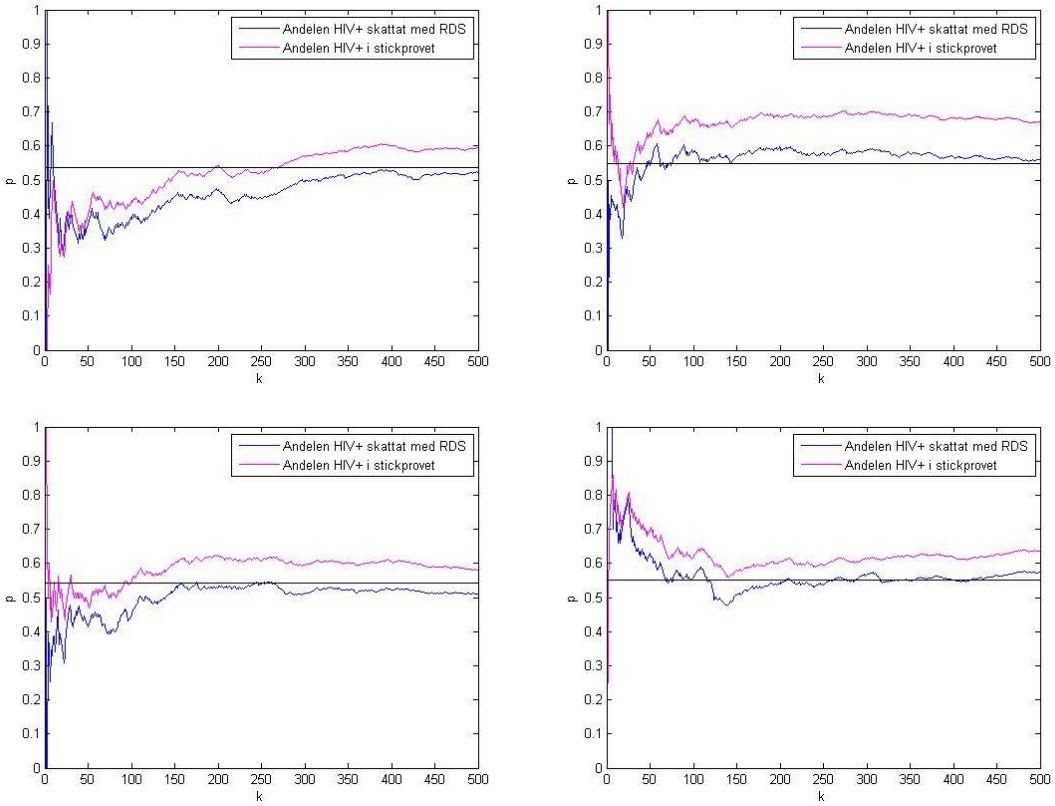

**Figur A12a-d.** Modell 3. Här är $N = 1000$, $p_{HIV} = 0.07$ $\lambda = 3$ och $\alpha = 0.2$.




**Figur A13a-d.** Modell 4. Här är $N = 1000$, $\beta = 0.3, p_{HIV} = 0.35$ och $\lambda = 3$.




**Figur A14a-d.** Modell 4. Här är $N = 1000$, $\beta = 0.3, p_{HIV} = 0.35$ och $\lambda = 10$.




**Figur A15a-d.** Modell 4. Här är $N = 1000$, $\beta = 0.8, p_{HIV} = 0.3$ och $\lambda = 3$.