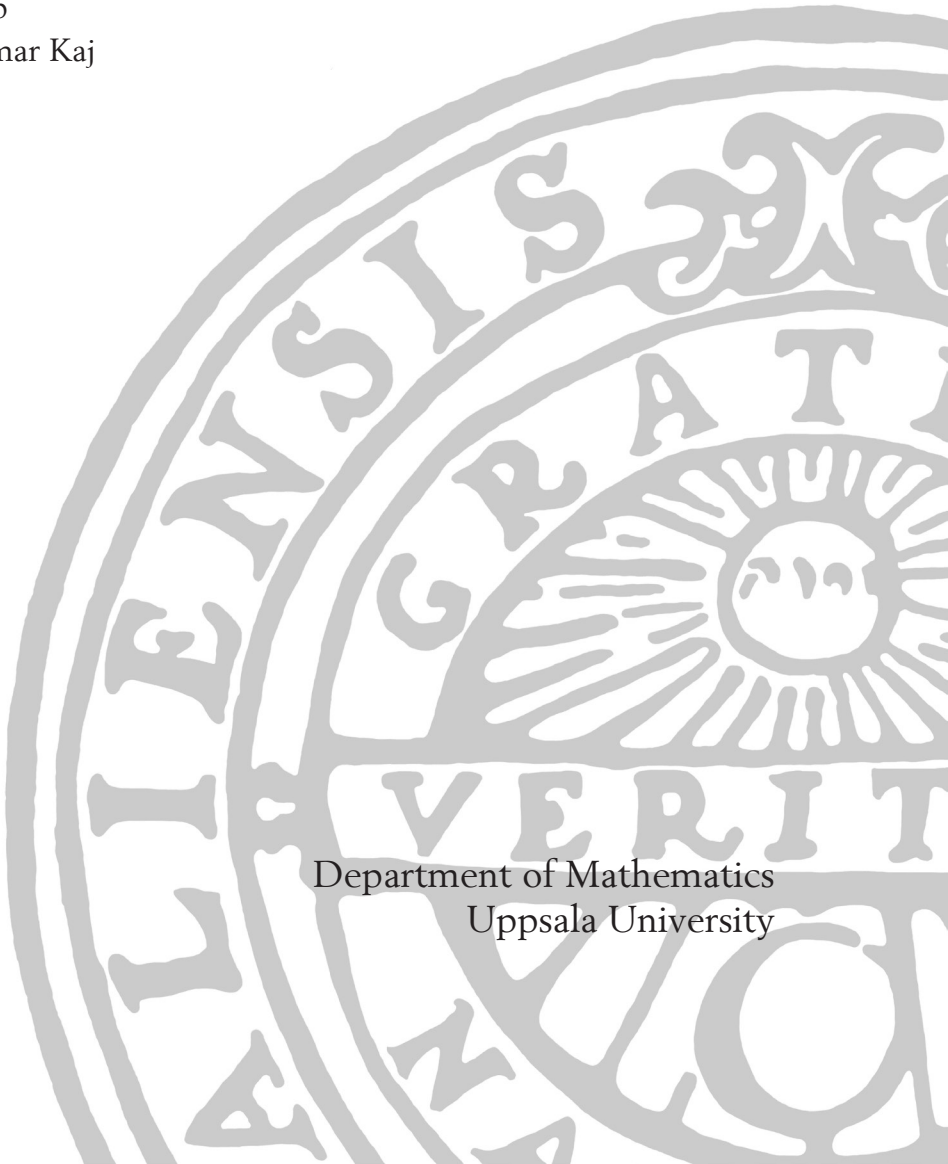
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2015:2

Numerisk analysmetod för oddskvot i en stratifierad modell

Mikael Jederström

Examensarbete i matematik, 30 hp
Handledare och examinator: Ingemar Kaj
Maj 2015



Department of Mathematics
Uppsala University

Titel
Numerisk analysmetod för oddskvot i en stratifierad modell

Uppsatsförfattare
Mikael Jederström

Sammanfattning av uppsatsförfattaren
Den i uppsatsen i detalj preciserade hypotesen om ett samband som antagits kunna generera en överrepresentation av huvudskador bland europeiska forwardspelare i NHL fanns det inte stöd för i data som analyserats i uppsatsen. Snarare det viktigaste analyssvaret pekade i en motsatt riktning. Sidoresultat hittades i datamaterialet som ger indikation i favör för förutsatt hypotes vilket möjligen kan uppmuntra till att vidare undersöka saken med kompletterande datakällor.

I examensarbetets har en metod arbetats fram ”Numerisk analysmetod för oddskvot i stratifierad modell” och metoden finns presenterad i uppsatsen, av uppsatsförfattare, som en statistisk analysmetod. Matematiska institutionen står inte som ansvarig för innehållet i något annat hänseende än att godta uppsatsen som del i ett examensarbete och betraktar uträkningar i uppsatsen som enskilda och icke validerade räkneexempel. Matematiska institutionen står inte heller som garant för användning av uppsatsens metoder i sammanhang utanför ramen av detta examensarbete.

Tack Ingemar Kaj, för synpunkter och all tillmötesgående

Innehållsförteckning

1. Inledning

Jag följer ishockeyn troget samtidigt finner jag det störande med dessa miljoner av ohälsosamma hjärnskakningar inom sporten. Jag ville undersöka om det finns möjlighet att hitta data och en analysmetod som kan ge stöd för att vissa grupper av hockeyspelare är särskilt utsatta för spelsituationer som resulterar i hjärnskakningar.

Jag valde en modell med endast en parameter nämligen parameter för oddskvot. Ett analysresultat som talar för ett samband riskerar att bli något intetsägande om inte analysen korrigerar för oväsentliga faktorer. Gruppindelningar skapar just förutsättningar för att kunna korrigera och minska effekten ifrån oväsentliga faktorer. Den valda modellen möjliggör också gruppindelning av spelare och utan att därmed också behöva införa ytterligare parametrar. Det hela utmynnade i en analysmetod som kan användas på en liten datamängd, på en population som är indelad i grupper, med ett analysförfarande som kan ge indikation på samband som genererar överrepresentation och som levererar skattningar av oddskvot inom exakta intervall. Analysmetoden, vår exakta metod, gavs namnet ”Numerisk analysmetod för oddskvot i en stratifierad modell”. För att utveckla och testa metoden har vi valt ett datamaterial och designat en datastudie för detta ändamål. Datastudien hämtar data ifrån NHL och säsongen 2013-14 där 42 st. ishockeyspelare hade frånvaro ifrån spel med diagnosen ”concussion”. Frågeställning för datastudien är om det finns ett samband mellan att spelare under spel blir utsatta för huvudskada och om spelare har amerikansk eller europeisk hockeybakgrund.

För att finna jämförelser så används fyra olika analysmetoder för oddskvot.

Metod ett till metod fyra syftar på:

Metod 1.	Vår exakta metod, ostratifierad data
Metod 2.	Metod för oddskvot i Profile likelihood, ostratifierad data
Metod 3.	Vår exakta metod, stratifierad data
Metod 4.	Mantel-Haenszels metod, stratifierad data

Båda metoderna ”Profile likelihood” (metod 2) och ”Mantel-Haenszels” (metod 4) är metoder som kan betecknas som approximativa. I båda de approximativa metoderna blir observerad data (data i form av skattad oddskvot-parameter) i log skala betraktad som ett utfall ifrån en normalfördelad sannolikhetsmodell. Att notera angående oddskvot, ofta betraktar vi oddskvotparametern i log skala för att få mer hanterbara siffror att arbeta med. Skillnader mellan exakta och approximativa analysmetoderna kan i första hand beskrivas av att de approximativa metoderna analyserar data via en modell som genererar ett normalfördelat datautfall vilket ju då bildar ett kontinuerligt och symetriskt fördelat datautfall. Motsvarande dataanalys utförd med exakt metod analyserar oddskvot utgående ifrån en modell som bildar en diskret och dessutom osymmetrisk sannolikhetsfördelning för utfallen. Med samma datauppsättningar kan analys med exakt metod alltså jämföras med analys med en approximativ metod. Eftersom skillnader i resultat mellan metoderna inte alltid förväntas att vara försumbara, särskilt när

data mängden är liten, är det angeläget att tydligt underbygga och beskriva vår exakta metod och ett särskilt kapitel ägnas åt att presentera metoden. Skillnaderna mellan de exakta och approximativa metoderna förväntas bli utjämnade när datamängden ökar. Det är därför intressant att med olika dataunderlag, i första hand genom vår datastudie, jämföra analysresultaten avseende oddskvot och granska om exakta metoder och approximativa metoder levererar likbetydande resultat.

2. Bakgrund och metodik

2.1. Bakgrund

Ett syfte med datastudien är att analysera samband i en situation där många hockeyspelare har fått sin hälsa degraderade och karriärer ödelagda av hjärnskakning. Svenska hockeyrinkar under 2015 kan liknas vid ett slagfält. I NHL upplever jag tvärtom att spelet har blivit mycket renare efter att 100 000$ böter ges repressivt för huvudtacklings förseelser. Det kan vara intressant att undersöka möjliga mekanismer eller faktorer som påverkar förekomsten av hjärnskakning inom sporten.

I datastudien undersöks en hypotes om att ett samband existerar som tenderar att ge en överrepresentation för europeiska spelare att bli utsatta för skadehändelser som leder till hjärnskakning. En preferens har varit att i första hand undersöka redovisningsgruppen forwardsspelare i åldern upp till 32 år. De eventuella fackkunskaper inom hockey som finns att tillgå kan anses vara begränsad till den gruppen. Datauppgifter hämtas ifrån säsong 2013-14 i NHL och samtliga spelare med istid under säsongen bidrar till datastudiens resultat. Inget urval av spelare utförs. Information om spelares istider under säsong 2013-14 används inte. Endast två datauppgifter ifrån säsong 2013-14 beaktas, för varje spelare, antingen hade spelaren diagnosen ’concussion’ eller inte och antingen var spelaren amerikan eller europé. Spelare kan indelas i olika grupper, det förekommer en finare indelning av spelare i strata och även indelningar av spelarna i olika redovisningsgrupper (exempelvis forwards och backar). Datauppgifter för spelare som avgör spelarnas gruppindelning (i strata) är uteslutande hämtade ifrån säsong 2012-13.

2.2. Urval

Inget urval görs. Man kan tänka sig att reglera modellen genom att välja att inte ta med en andel icke-skadade spelare genom att slumpmässigt välja ut ett antal spelare (retrospektivt) bland de icke-skadade spelarna. Fall- och kontroll studier innehåller ett moment med urval av kontroller (icke-skadade spelare i vår datastudie). Men datastudien är inte en fall- och kontrollstudie och inget (retrospektivt) urval av kontroller behöver genomföras.

2.3. Oddskvot

Resultaten som tas fram i datastudien ges i from av oddskvoter, oddskvoten ger ett storleksmått på samband och bakomliggande tendens. Skattade oddskvoter kan också användas för att ge ett numerärt värde för någon typ av observerad överrepresentation. 42 stycken diagnoser ’concussion’ är uppdelade på europeiska och amerikanska hockeyspelare sammanlagt 886 st. Med uppdelningen av spelare i fyra kategorier, med eller utan diagnosen ’concussion’ och amerikanska eller europeiska spelare kan en skattning av oddskvot beräknas.
Definitionen av oddskvot med parameter notation:

π_1		I vårt exempel: Risk för fall (huvudskada) för europeisk spelare.
π_0		I vårt exempel: Risk för fall (huvudskada) för amerikanska spelare.
Ω	$\frac{\pi}{(1-\pi)}$	Odds. (Mått mellan sannolikheter för två olika möjliga utgångar antingen inträffar den ena eller andra utgången.)
θ	$\theta = \frac{\Omega_1}{\Omega_0}$	Oddskvot.

Tab 2.1 Oddskvot, fiktiva räkneexempel på oddskvoter

	Risk för grupp "1" π_1	Risk för grupp "0" π_0	Oddskvot $\theta = \frac{\pi_1 \cdot (1-\pi_0)}{\pi_0 \cdot (1-\pi_1)}$
Dubbel risk	0.1 π_1	0.05 π_0	2.11 θ
Dubbel risk	0.05 π_1	0.025 π_0	2.05 θ
3- Faldig riskökning	0.15 π_1	0.05 π_0	3.35 θ
3- Faldig riskökning	0.075 π_1	0.025 π_0	3.16 θ
4- Faldig riskökning	0.12 π_1	0.03 π_0	4.41 θ
4- Faldig riskökning	0.06 π_1	0.015 π_0	4.19 θ
Lika stor risk	0.1 π_1	0.1 π_0	1 θ

2.4. Modell antagande om en konstant oddskvot

Med gruppindelning av spelarna i strata följer att risknivåer tillåts att variera men vad som antas vara konstant över indelningarna är att en förhärskande tendens som genererar överrepresentation eller att ett bakomliggande samband är konstant fast risknivåernas storlek mycket väl kan variera mellan strata.

2.5. Notation för variabler

Tab 2.2 *Notation för variabler (med värden)*

		Antal skadade		Fria ifrån diagnos			
Samtliga spelare 2013-14	**Européer**	14	D_1	180	H_1	194	N_1
	Amerikaner	28	D_0	664	H_0	692	N_0
	Alla spelare	42	D	844	H	886	N

Form för datarepresentation som kontingents tabell.

Tab 2.3 *Ekvivalent datainformation som i tab 2.2*

Stratum	**Européer (antal)**		**Amerikaner (antal)**		**Antal skadade**		**Observerat antal skadade européer**	
Samtliga spelare 2013-14	194	N_1	692	N_0	42	D	14	D_1

2.6. Notation för variabler vid gruppindelning

Tab 2.4 *Notation för variabler när en andel av populationen är ytterligare indelad i grupper*

Stratum		Européer (antal)		Amerikaner (antal)		Antal skadade		Observerat antal skadade européer	
'Status'(+) 'Fysisk stil'(-)	1 (s)	33	N_1^1	69	N_0^1	7	D^1	4	D_1^1
'Status'(+) 'Fysisk stil'(+)	2 (s)	20	N_1^2	81	N_0^2	9	D^2	3	D_1^2
Summa		53	N_1	150	N_0	16	D	7	D_1

På just den här kombinationen av data utförs ingen analys. Detta exempel på indelning utgörs av alla spelare med ”hög status” som i sin tur är indelad i 2 grupper (2 strata).

Tab 2.5 *Ekvivalent data- information som i 2.4*

		Antal skadade		Fria ifrån diagnos			
1 'Status'(+) 'Fysisk stil'(-)	**Européer**	4	D_1^1	29	H_1^1	33	N_1^1
	Amerikaner	3	D_0^1	66	H_0^1	69	N_0^1
	Alla spelare	7	D^1	95	H^1	102	N^1
		Antal skadade		**Fria ifrån diagnos**			
2 'Status'(+) 'Fysisk stil'(+)	**Européer**	3	D_1^2	17	H_1^2	20	N_1^2
	Amerikaner	6	D_0^2	75	H_0^2	81	N_0^2
	Alla spelare	9	D^2	92	H^2	101	N^2

2.7. Fyra analysmetoder och matematisk modell

Som redan nämnts i inledningen presenteras och används fyra olika metoder för analys av oddskvot:

Metod 1.	Vår exakta metod, ostratifierad data
Metod 2.	Metod för oddskvot i Profile likelihood, ostratifierad data
Metod 3.	Vår exakta metod, stratifierad data
Metod 4.	Mantel-Haenszels metod, stratifierad data

Metod ett, tre (vår exakta metod) och metod fyra (Mantel-Haenszels metod) utgår alla tre ifrån en matematisk modell med hypergeometrisk likelihood-funktion.

Hypergeometrisk likelihood-funktion $L^T(\theta)$ skrivs enligt [1 Clay]

$$L^T(\theta) = \prod_{t=1}^{T(\text{Alla strata})} \frac{1}{K^t(\theta)} \bullet \frac{\theta^{D_1^t}}{D_1^t!D_0^t!H_1^t!H_0^t!} \ , \ K^t(\theta) = \sum_{\text{alla möjliga } D_1^t} \frac{\theta^{D_1^t}}{D_1^t!D_0^t!H_1^t!H_0^t!}$$

(2.1).

En användbar ansats och god start när uttrycket för $L^T(\theta)$ ska härledas är att fixera variabelvärdena N_1^t, N_0^t och D^t, se tab 2.4-2.5. Fortsättning av härledning av $L^T(\theta)$ utelämnas.

I denna matematiska modell blir ”marginalerna” i kontingentstabellerna betraktade som fixerade. (Marginalerna i kontingentstabellerna är variablerna i de vita fälten i tab 2.2 och 2.5.) (Eller motsvarande i tabellerna tab 2.3 och tab 2.4 där de gröna fälten samlar variablerna som representerar variabler i modellen som betraktas som fixerade.) Eftersom vår exakta metod och Mantel-Haenszels metod utgår ifrån en hypergeometrisk likelihood-funktion som förutsätter den här betingningen av variabler blir det mer lämpligt att representera data med enstaka rader (som i tab 2.3 och 2.4) och inte tre rader som i kontingentstabell (tab 2.2 och 2.5). Variablerna D_1^t ”observerat antal skadade européer” (de vita fälten i tab 2.3 och 2.4) är slumpvariablerna i den hypergeometriska modellen. Slumpvariablerna, betecknade D_1^t, kan representera möjliga utfall eller det faktiska observerade utfallet i stratum t. Det siffervärde som står jämte symbolen D_1^t (i tab 2.3 och 2.4) är det observerade utfallet för antalet skadade européer i stratum ”t”. Mantel-Haenszels (metod fyra) approximativa metod för stratifierad dataanalys presenteras i ett följande avsnitt. Vår exakta metod (vilken syftat båda på metod ett och metod tre) beskrivs i ett särskilt kapitel ”4 Metodbeskrivning”.

Metod två (Metod för oddskvot i Profile likelihood, ostratifierad data) skiljer sig ifrån metod ett, tre och fyra genom att den matematiska modellen för metoden inte förutsätter en betingning av variabler. Tyvärr går det inte att skriva ned ett matematiskt uttryck för Profile likelihood-funktion som metoden använder men detta är problem som går att hantera genom att approximera Profile log likelihood-funktionen med normalfördelningens log likelihood-funktion. Metoden presenteras sist i detta kapitel.

2.8. Mantel-Haenszels metod

Mantel-Haenszel metod (metod fyra) för oddskvot i stratifierad analys är baserade på hypergeometrisk likelihood- funktion och ger approximativa resultat. Formler som används i Mantel-Haenszel metod är framtagna genom att söka lämpliga normal log likelihood-approximation till hypergeometrisk log likelihood. Formler tas fram genom att approximera log likelihood-funktion för log oddskvot-parametern (givet ett observerat utfall ifrån den diskreta sannolikhetsmodellen). Hur just de följande approximationerna och formlerna (2.2-2.4) har tagits fram ligger utanför våra ramar men en kontroll hur bra och lämpliga approximationerna är för aktuell data kan faktiskt kontrolleras grafiskt (se sist i detta avsnitt). Mantel-Haenszel skattning av oddskvot θ ges enligt [1 Clay] av

$$Q = \sum_{t=1}^{T} \frac{D_1^t H_0^t}{N^t}, \quad R = \sum_{t=1}^{T} \frac{D_0^t H_1^t}{N^t},$$

$$\hat{\theta} = \frac{Q}{R} \qquad (2.2).$$

Ett approximativt sannolikhetsvärde för observerat utfall med nedanstående hypotes och mothypotes kan beräknas genom att bilda uttrycket (2.3), enligt [1 Clay].

Scoretest av $H_0 : \theta = 1$,

mothypotes $H_1 : \theta > 1$.

$$U = Q - R,$$

$$V = \sum_{t=1}^{T} \frac{D^t H^t N_0^t N_1^t}{N^t N^t (N^t - 1)},$$

$$\frac{U^2}{V} \sim approximativt\, \chi^2(1) \qquad (2.3).$$

Ett approximativt konfidensintervall för θ bildas genom att använda Wald-test statistika (2.4), enligt [1 Clay].

Wald test av $H_0 : \theta = \theta_\Theta$,

standardavvikelse,

$$s = \sqrt{\frac{V}{QR}},$$

$$\left(\frac{\log \hat{\theta} - \log \theta_\Theta}{s} \right)^2 \sim approximativt\, \chi^2(1) \qquad (2.4).$$

Med Mantel-Haenszel metod och formler (2.2-2.4) ges verktyg för en lättberäknad dataanalys för oddskvot. Hur bra approximationen i formel (2.4) är för redovisningsgrupperna i vår datastudie kan undersökas grafiskt genom att rita upp normal log likelihood-ratio kurva med vär-

dena i formel (2.4) och sedan jämföra med uppritad hypergeometisk log likelihood-ratio kurva, sådana kurvor har ritats upp i fig 5.1-5.2.

2.9. Metod för oddskvot i Profile likelihood

Analysmetoden (metod två) av oddkvot sker via en Profile likelihood-funtkion för oddskvot, metoden ger approximativa resultat och används i ostratifierad analys. Profile likelihood-funktioner grundar sig på ett uttryck som bildas ifrån Binomial likelihood-funktion.

Tab 2.2 åter igen. Kontingentstabell med notation för variabler

		Antal skadade		Fria ifrån diagnos			
Samtliga spelare 2013-14	**Européer**	14	D_1	180	H_1	194	N_1
	Amerikaner	28	D_0	664	H_0	692	N_0
	Alla spelare	42	D	844	H	886	N

Tab 2.2
Spelarna kan delas upp antingen i gruppen amerikanska spelare eller europeiska spelare. Och {1,0}-värden kan ges för geografisk exponering. Nedsänkt etta står för europeisk kategori. En spelares säsong kan också tilldelas ett {1,0}-värde och kategori {D, H}. Kategori variabel D räknar antalet spelare vars spelares säsong innehåller ett "fall" en skada 'concussion', variabel H räknar antalet spelare vars spelares säsonger är fria ifrån skada 'concussion'.

Av sammanlagt N försök observeras D händelser med utfallet 1 och (N-D) med utfallet 0. Likelihood för π_1 risk för europeiska spelare blir uttrycket

$$\pi_1^{D_1}(1-\pi_1)^{N_1-D_1},$$

likelihood för π_0 risk för amerikanska spelare blir uttrycket

$$\pi_0^{D_0}(1-\pi_0)^{N_0-D_0},$$

likelihood för π_0 och π_1 blir då

$$\pi_1^{D_1}(1-\pi_1)^{N_1-D_1} \cdot \pi_0^{D_0}(1-\pi_0)^{N_0-D_0} \qquad (2.5).$$

Efter variabelbyte och logaritmering av (2.5) som inte beskrivs här kan parametrarna π_1 och π_0 bytas ut mot parameter för oddskvot θ, men även parametern odds för amerikanska spelare Ω_0 tillkommer. Den totala log likelihood-funktionen för parametrarna θ (och Ω_0) blir efter omskrivningen av (2.5) enligt [1 Clay]

$$D_0 \log(\Omega_0) - N_0 \log(1+\Omega_0) + D_1 \log(\theta \cdot \Omega_0) - N_1 \log(1+\theta \cdot \Omega_0) \qquad (2.6).$$

Poängen är att Ω_0 finns kvar i uttrycket som extra parameter och stör. (I hypergeometriska likelihood för oddskvot θ används tekniken att införa restriktion på modellen, en betingning av variabler i modellen, vilket möjliggör att bilda likelihood-funktion för θ utan stör parametrar. Men här används en annan teknik som beskrivs här översiktligt.) Genom att punktvis i

funktionen (2.6) för varje värde på θ ersätta parametern Ω_0 med ett algebraiskt uttryck fås punktvis för varje värde på parameter θ en profile log likelihood-funktion som bara innehåller en parameter θ. Det finns inget algebraiskt utryck att skriva för profile log likelihood-funktionen som gäller för varje punktvärde på θ. Profile log likelihood-funktionen måste skrivas om för varje möjligt punktvärde på θ. Och för ett punktvärde på oddskvoten θ när en profile log likelihood har definierats så ger den profile log likelihood funktionen ett funktionsutryck för varje värde på θ. Punktskattningen av oddskvot är

$$\hat{\theta} = \frac{D_1/D_0}{H_1/H_0} \qquad (2.7).$$

I skattningen (2.7) används en kvot mellan maximum likelihhood-skattning av Ω_1 och en maximum likelihhood-skattning av Ω_0. Problemet med att inte kunna skriva ned en enhetlig Profile likelihood-funktion löses genom att approximera Profile log likelihood-funktionen med normal fördelningens log likelihood. Normalapproximationen leder fram till formlerna (2.8-2.9) och som ger medel för en lättberäknad dataanalys. Ett approximativt konfidensintervall för oddskvot θ kan bildas enligt [1 Clay] genom att bilda Wald-test statistika (2.8).

Wald test av $H_0 : \theta = \theta_\Theta$,

$$s = \sqrt{\frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{H_1} + \frac{1}{H_0}},$$

$$\left(\frac{\log\hat{\theta} - \log\theta_\Theta}{s} \right)^2 \sim approximativt\, \chi^2(1) \qquad (2.8)$$

Ett approximativt sannolikhetsvärde för observerat utfall med nedanstående hypotes och mothypotes kan beräknas genom att bilda uttrycket (2.9), enligt [1 Clay].

Scoretest av $H_0 : \theta = 1$,

mothypotes $H_1 : \theta > 1$,

$$U = D_1 - D \cdot \frac{N_1}{N},$$

$$V = DHN_0N_1/N^3,$$

$$\frac{U^2}{V} \sim approximativt\, \chi^2(1) \qquad (2.9).$$

Score-varians V säger något om krökningen hos aktuell log likelihood-funktionen för observerad data i punkten oddskvot=1. Notera uttrycket för V och observera skillnaden mellan hypergeometrisk modell med $N \cdot N \cdot (N-1)$ i nämnaren och här med modell baserad på en Profile log likelihood-funktion har vi istället N^3 i nämnaren i uttrycket för V. Följaktligen och särskilt när datamängden är liten så skiljer sig även dataanalysen i hypergeometrisk modell ifrån dataanalysen i modell baserad på Profile likelihood-funktion.

3. Datamaterial

Datamaterialet ur vilka alla identifierade huvudskador hämtas är en databas [DB01] med information som är fritt tillgängligt på Internet. Datamaterialet har vissa brister som underlag att bedöma det samband som uppsatsen syftar till men bättre underlag fanns inte att tillgå utan kostnad. Datamaterialet fungerar också och utmärkt som räkneexempel.

3.1. Redovisningsgrupper i datastudien

Datamaterialet innehåller 886 spelare ifrån säsong 2013-14 och med c:a 40 stycken spelare som har en diagnos hjärnskakning. En grupp spelare ’Forwards’ är utvald att utgöra huvudintresset för analysen av datamaterialet. Resultatet ifrån gruppen ’Forwards’ presenteras i resultaten som huvudresultatet för datastudien. Övriga resultat ifrån övriga spelaregrupper presenteras som sidoresultat.

Populationen delas in i 4 stycken redovisningsgrupper:
Forwards, backar, erfararna spelare, rookies samt grupperna i förening. Dessutom presenteras respektive redovisningsgrupp med eller utan ytterligare separata indelningar (strata). (Resultaten ifrån redovisningsgrupper ”backar”, ”erfararna spelare”, ”rookies” utelämnas.)

3.2. Stratifiering

Gruppindelning av hockeyspelarna har skett genom en procedur som av utrymmes skäl inte redogörs för här. Avsikten med gruppindelning är att minska betydelsen av möjliga riskfaktorer som kan förväntas störa analysen av ett samband mellan faktorer eller spelarkategorier och benägenhet hos spelare att ådra sig en huvudskada. Syftet med stratifiering är att ge de riskfaktorer som är utvalda att vara meningsfulla ett större utrymme och utgöra en dominerande förklaringsgrund till de samband som eventuellt går att påvisa i datamaterialet. De riskfaktorer som inte är av primärt intresse och som därför genom utförd gruppindelning ska ges en minskad betydelse i dataanalysen är

- spelstil,
- status (spelar värde i $),
- spelares storlek (kg).

Bakomliggande faktorer som valts ut vara intressanta och att hypotetiskt kunna vara av betydelse för ett samband samt spekulativt bestämts att kunna anses vara nära knutet till spelarens härkomst (geografiska tillhörighet antingen europeiskt eller amerikansk) är

- spelarens hockeyutbildning,
- olika tänkbara skäl till att inte bli betraktad som likvärdig av publik, medier, disciplin nämnd och övriga spelare.

4. Metodbeskrivning

Detta kapitel är vikt enbart till beskrivning av vår exakta analysmetod för oddskvot. Metoden är egentligen endast en enda metod men eftersom metoden fungerar utan några ändringar i beräkningskod för båda stratifierad data och ostratifierad data kan den anses fungera som två metoder. Metoden som tidigare nämnts utgår ifrån en matematisk modell med hypergeometrisk likelihood-funktion. Modellantaganden och hur modellen ska kunna användas i ett sammanhang i verkligheten lämnas för tillfället, det behandlas i kapitel 2. Som räkneexempel för metoden används vår datastudie och beräkningar med data ifrån studien åskådliggörs med figurer i slutet av kapitlet. Utmärkande för metoden är att den inte använder någon tabellfördelning för att ange utfallsmöjligheter och sannolikheter för utfall. Istället så används hypergeometrisk sannolikhetsfunktion till att ange alla möjliga utfall i detalj samt att beräkna sannolikhet exakt för varje specificerat möjligt utfall. När data är stratifierat så växer antalet möjliga utfall och beräkningsmängden ökar snabbt. Det finns olika former att välja för representationen av enskilda datautfall. Ett enskilt datautfall kan dessutom antingen vara ett observerat datautfall eller ett möjligt utfall. Valda representationen av datautfallen i metoden är maximum likelihood-skattning av oddskvot. Bör noteras att det saknas algebraiskt uttryck och formel för maximum likelihood-skattning i hypergeometrisk likelihood-funktion för oddskvot. Maximum likelihood-skattning av oddskvot måste beräknas numeriskt genom att maximera likelihood-funktion för oddskvot med aktuellt data. Ett överraskande resultat för hypergeometrisk likelihood-funktion för oddskvot i stratifierad modell är att antalet möjliga olika siffervärden för utfallen som kan bildas med representationen maximum likelihood-skattning av oddskvot är ett begränsat antal. I följande texter utvecklas resonemanget ytterligare kring varför de möjliga utfallen bildar en diskret och begränsad utfallsmängd.

Utgående ifrån en datauppsättning skapas en statisk utfallsmängd av maximum likelihood-skattningar av oddskvot. Metoden bygger sedan alltså på att för den skapade oföränderliga utfallsmängden bilda och beräkna exakta sannolikhetsfördelningar. Sannolikhetsfördelningen för utfallen ändras beroende av det aktuella värdet på oddskvot-parametern. Sannolikhetsfördelningarna är beräkningsbetungande att räkna fram eftersom en sannolikhetsfördelning består av ett stort antal möjliga utfall som växer snabbt med datastorlek. I vår datastudie hittas en datauppsättning där antalet möjliga utfall är cirka 500 000. Och sannolikheterna för dem alla måste beräknas. Dessutom kan metoden kräva att beräkningen av sannolikheter kan behöva ske så många gånger som 80 gånger för samma datauppsättning. En ny sannolikhetsfördelning måste beräknas för varje värde på oddskvot-parametern. I metoden och med dataexemplet med 500 000 möjliga utfall består utfallsmängden av oddskvot skattningar endast av ett fåtal diskreta gemensamma värden. Antalet diskreta möjliga värden som utfallen i utfallsmängden kan anta i exemplet ovan är endast omkring 40 olika värden. Metoden använder som nämnts inte en tabellfördelning för dataanalysen. Det är dessa ovan beskrivna exakta diskreta sannolikhetsfördelningar som utgör grunden för all analys i metoden. Och för att sammanfatta, en datauppsättning i datastudien har en utfallsmängd av möjliga utfall som är statisk. Men datauppsättningens sannolikheter för möjliga utfall förändras och beror av värdet på parame-

tervärdet för oddskvoten, En diskret sannolikhetsfördelning måste alltså beräknas för varje värde på oddskvot-parametern och för varje datauppsättning.

I följande ges specifikationer för metoden, en stegvis beskrivning av metodiken för metodens genomförande och figurer presenteras för att tydliggöra metodiken i metoden. I framställningen hittas även en skissad bevisföring som säger att det finns ett gemensamt extremvärde på oddskvot-parametern i hypergeometrisk likelihood- funktion för oddskvot i stratifierad modell för datautfall med samma antal exponerade fall. Den gemensamma extrempunkten motiverar att datautfallen sammanfaller till en påfallande diskret utfallsmängd av skattade oddskvoter på det sätt som nämnts ovan. De beräkningstunga numeriska beräkningarna genomförs i programmeringsmiljön MATLAB och koden finns i uppsatsförfattarens ägo.

4.1. Fakta om metod

Resultat som levereras ifrån ”Numerisk analysmetod för oddskvot i stratifierad modell”:

Maximum Likelihood- skattning av gemensam oddskvot för stratifierad data.
95 % Konfidens intervall för oddskvot-parameter. (Exakta p-värden).
90 % Konfidens intervall för oddskvot-parameter. (Exakta p-värden).
Test och P-värde: Exakt sannolikhet för observerad data med en nollhypotes oddskvot=1. Test med signifikans nivå 0.05. P-värden angivna med en enkel sidig mothypotes.
Styrka: Sannolikhet (exakt) för att detektera signifikans med en oddskvot =2.2.
Kontroll av modellanpassning för observerad data.

Andra fakta:

Modell: Hypergeometrisk likelihood funktion för oddskvot.
Numeriska beräkningar för ML-skattning av oddskvot och exakta konfidensintervall.
Metoden är verifierad att fungera med upp till 42 exponerade fall och 1-7 strata och med antal kontroller upp till 800st.
Programkod är skriven i MATLAB.
Resultaten ges med tre siffrors noggrannhet.
Data input samt metodens metodiksteg för beräkningarna ges i metodbeskrivning.

4.2. Metodiksteg 1. Datainput till metoden

Följande variabler behöver laddas och är tillräckligt för att kunna genomföra alla beräkningar i metoden för aktuell datauppsättning.

$[N_1^1, N_1^2, ..., N_1^t, ..., N_1^T]$: Antal exponerade objekt i respektive stratum.

$[N_0^1, N_0^2, ..., N_0^t, ..., N_0^T]$: Antal icke exponerade objekt i respektive stratum.

$[D^1, D^2, ..., D^t, \ldots, D^T]$: Antalet fall i respektive stratum.

$[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$: Antalet observerade exponerade fall i respektive stratum.

Den matematiska modellen som används beskrivs av hypergeometrisk likelihood-funktion. Variablerna $[D^1, D^2, ..., D^t, \ldots, D^T]$, $[N_0^1, N_0^2, ..., N_0^t, ..., N_0^T]$ och $[N_1^1, N_1^2, ..., N_1^t, ..., N_1^T]$ är fixerade av modellbetingelserna och betraktas som konstanter. Variablerna $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$ är de stokastiska variablerna i modellen.

4.3. Metodiksteg 2. Generera samtliga utfall

Alla möjliga utfall skrivna på formen $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$ och inte enbart det enskilda observerade utfallet genereras i ett datorprogram och samlas i en datastruktur. Antalet möjliga utfall växer exponentiellt och för en datauppsättning med 7 stratum med 886 spelare genererades exempelvis c:a 500 000 olika möjliga utfall.

Mängden C definieras att utgöras av samtliga möjliga utfall på formen

$[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$.

$y_i \in C$ och y_i betecknar ett utfall representerad på formen $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$.

$y_{observerad}$ betecknar observerat utfall och är ett utfall av alla möjliga utfall i C.

4.4. Metodiksteg 3. Välja en skattning av oddskvot

Varje möjligt utfall bestående av data ifrån T olika strata tranformeras till ett endimensionellt utfallsvärde genom att för varje utfall på formen $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$ beräkna en skattning av oddskvot $\hat{\theta}$. Maximum likelihood-skattning $\hat{\theta}_{ML} \mid y_i$ som skattning av oddskvot är ett naturligt alternativ att välja. För varje möjligt utfall y_i ($y_i \in C$) skattas en $\hat{\theta}_{ML}$ för y_i. Maximum likelihood-skattning $\hat{\theta}_{ML} \mid y_i$ skattas för alla y_i numeriskt genom att maximera hypergeometriska likelihood-funktion för oddskvot i stratifierad modell och aktuellt utfall y_i.

4.5. Metodiksteg 4. Ordna alla utfall efter storlek

För att kunna ordna alla möjliga utfall $y_i \in C$ visar det sig att summan D_1 är tillräcklig information för att ordna utfallen. Att ordna alla möjliga utfall efter summan D_1 är ekvivalent med att ordna alla $y_i \in C$ efter storleken på skattningen av oddskvot $\hat{\theta}_{ML} \mid y_i$ i hypergeometrisk likelihood-funktion för oddskvot och i stratifierad modell.

$$(D_1 = \sum_{t=1}^{T} D_1^t .)$$

Ytterligare förtydligande: Utfall $y_i \in C$ med samma summa D_1 har extrempunkter med gemensam punkt på parameteraxeln θ, extrempunkterna har samma värde på θ vilket är det väsentliga för skattningen av oddskvot $\hat{\theta}_{ML} \mid y_i$. Extrempunkterna har olika maximum värden på funktion $L^T(\theta)$, se formel (4.1) nedan, vilket för skattningarna av θ inte är väsentligt. Med hjälp av matematiskt uttryck nedan, se (4.2), skissas en bevisföring som säger att samtliga möjliga utfall med identisk summa D_1 har extrempunkter med en gemensam punkt på parameteraxeln.

4.6. Metodiksteg 5. Ordna utfallen kring observerat utfall

Observerad data utgörs av endast ett av de möjliga utfallen av typen $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$ med en skattad oddskvot $\hat{\theta}_{ML} \mid y_{observerad}$.
En illustration över när samtliga möjliga utfall på formen $\hat{\theta}_{ML} \mid y_i$ blivit ordnade framgår nedan i figur, se fig 4.1.

Samtliga utfall y_i, $y_i \in C$, skrivna på formen $\hat{\theta}_{ML} \mid y_i$ delas in i tre kategorier.
Kategori 1 - mera extrem än observerad data (gröna staplar i fig 4.1).
Kategori 2 - likvärdig med observerad data alltså lika stor som $\hat{\theta}_{ML} \mid y_{observerad}$ (svart stapel).
Kategori 3 - mindre extrem än observerad data (rosa staplar i fig 4.1).

4.7. Metodiksteg 6. Sannolikhet för varje utfall

Sannolikheten beräknas för varje utfall $y_i \in C$. Sannolikhetsfördelningar för utfallen y_i bildas genom att beräkna sannolikheterna med aktuellt värde på oddskvoten θ. Aktuellt värde på oddskvoten θ ändras alltefter när metodens konfidensintervall, p-värden och styrka ska beräknas. Sannolikheten för varje enskilt utfall av typen $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$ beräknas genom att använda sannolikhetsfunktionen $P(y_i \mid \theta)$ tillika likelihood-funktion $L^T(\theta)$. Likelihood $L^T(\theta)$ presenterades även i kap 2, se (2.1), och skrivs åter igen (4.1).

Hypergeometrisk likelihood-funktion (och sannolikhetsfunktion $P(y_i \mid \theta)$)

$$L^T(\theta) = \prod_{t=1}^{T(\text{Alla strata})} \frac{1}{K^t(\theta)} \cdot \frac{\theta^{D_1^t}}{D_1^t! D_0^t! H_1^t! H_0^t!}, \quad K^t(\theta) = \sum_{\text{alla möjliga } D_1^t} \frac{\theta^{D_1^t}}{D_1^t! D_0^t! H_1^t! H_0^t!} \tag{4.1}$$

OBS! Som tidigare antytts, se tab 2.3, Variablerna D_1^t, D_0^t, H_1^t och H_0^t innehåller ekvivalent information som den alternativa representationen för datainput i variablerna N_1^t, N_0^t, D^t och D_1^t.

För att visa att samtliga möjliga utfall med samma antal exponerade fall D_1 också har ett gemensamt extremvärde på oddskvot-parametern i hypergeometrisk likelihood-funktion för oddskvot i stratifierad modell bildas uttrycket (4.2). En skissartad bevisföring ges här (på begäran kan en tydligare förklaring visas). Efter logaritmering av (4.1) och variabel byte och derivering fås uttrycket

$$\frac{\partial(\gamma^T(\beta))}{\partial\beta} = -\sum_{t=1}^{T} \frac{\partial(\log K_\gamma{}^t(\beta))}{\partial\beta} + \sum_{t=1}^{T} D_1^t \tag{4.2}$$

$$(D_1 = \sum_{t=1}^{T} D_1^t .)$$

För värden på β där uttrycket (4.2) antar värdet noll erhålls en maximum likelihood-skattning av oddskvot. Första termen i (4.2) till vänster om likhetstecknet är per definition identisk för varje möjligt utfall. Alla utfall med samma totalsumma antal exponerade fall (andra termen D_1) kommer också ha gemensamma maximum likelihood-skattningar $\widehat{\theta}_{ML}$ av oddskvot-parametern. Slutsatsen av bevisföringen utnyttjas till att motivera varför utfallen sammanfaller till en påfallande diskret utfallsmängd av skattade oddskvoter. Resultatet att utfall med samma summa D_1 har gemensam skattning på oddskvoten utnyttjas i metodiksteg ovan när utfallsmängden ordnas.

4.8. Konfidensintervall för oddskvot

Konfidensintervall för oddskvot beräknas genom att följa metodiksteg 1 till 6 enligt ovan. Metodiksteg 6 upprepas genom att en serie av olika värden på oddskvot-parametern prövas iterativt tills att summerade sannolikheter för enskilda utfall som är mindre extrem än observerat utfall bildar summan 0.95000. (Vilket är ekvivalent med att summan av rosa staplar blir 0.95000 i fig 4.2.) Ett parametervärde på oddskvoten och undre gräns i ett 90 % konfidensintervall för oddskvot har därmed beräknats kopplat till ett exakt p-värde =0.05.

Och generellt bildas övre (eller nedre) gränser i konfidensintervall iterativt genom en algoritm som gissar en serie av odds, en serie som konvergerar, mot en oddskvot som ger en summerad sannolikhetsmassa för samtliga möjliga utfall mera (eller mindre) extrem än observerad data ett givet sannolikhetsvärde (0.97500 eller 0.95000).

4.9. Test och sannolikhetsvärde för utfall. P- värde

Utfallsmängden av maximum likelihood-skattningar av oddskvot bildas på sätt som beskrivet ovan genom metodiksteg 1-5. När sannolikhetsfördelningen för samtliga utfall bildas, enligt metodiksteg 6, så väljs oddskvot =1. P-värdet beräknas genom att summera sannolikheter för möjliga utfall (i form av $\widehat{\theta}_{ML}$ skattningar) mera extrem eller lika stor som observerat utfall, se fig 4.4.
Testets hypotes och signifikansnivå är:

nollhypotes $H_0 : \theta = 1$,

mothypotes $H_1 : \theta > 1$,

signifikans nivå: 0.05.

4.10. Styrka

Metodens styrka eller prestanda på metoden beräknas genom att använda metodens test till att fastställa kritisk gräns för aktuell datauppsättning. Sedan beräknas exakta sannolikheten för att uppnå kritisk gräns med en förhärskande oddskvot på =2.2, se räkneexempel i fig 4.5.

4.11. Kontroll av modellanpassning hos data

Kontroll av modellanpassning hos data kan enkelt genomföras genom att granska varje observerat utfall D_1^t i strata t och i $[D_1^1, D_1^2, ..., D_1^t, ..., D_1^T]$.
och om något observerat utfall D_1^t (eller en serie av flera observerade D_1^t) har ett ”allt för” lågt sannolikhetsvärde kan antagandet om en konstant oddskvot vara felaktigt för observerad data.

4.12. Figurer förklarade

Fig 4.1 -fig 4.5 delar samma möjliga diskreta utfallsmängd (men har olika sannolikhetsfördelningar för utfallen eftersom oddskvotens värde varierar). De gemensamma utfallsmängderna har blivit uträknade utgående ifrån samma datainput (886 spelare i redovisningsgrupp ”Förenade, samtliga spelare”). Exempelvis svart stapel, i fig 4.1-4.4, som samlar en delmängd av utfallsmängden innehåller c:a 20 000 olika möjliga utfall vilket går att utläsa ur fig 4.1. Dessa möjliga 20 000 utfall i svart stapel delar ett gemensamt diskret värde på $\hat{\theta}_{ML} \mid y_i$. Utfall som har samma antal exponerade fall (samma antal europeiska spelare med huvudskada) har alla ett gemensamt värde på maximum likelihood skattningen av oddskvot, vilket har motiverats i ett ovanstående avsnitt. När sannolikheter ska beräknas, exempelvis för att vi ska erhålla ett utfall placerad i hopen av möjliga utfall i svarta stapeln så måste samtliga 20 000 separata sannolikheter för varje möjligt utfall i svarta stapeln beräknas separat och sedan adderas samman. Summan av samtliga antal utfall (Gröna, svarta och rosa staplar) består av c:a 500 000 stycken utfall. Separat uträknade sannolikheter för samtliga c:a 500 000 utfall beräknas och en sannolikhetsfördelning bildas med aktuellt värde på oddskvot-parametern. (I data koden kontrolleras också att summan av c:a 500 000 termer verkligen summeras exakt till =1.)

I fig 4.2, fig 4.3, 4.4 och fig 4.5 visar exempel på bildandet av exakta sannolikhetsfördelningar för utfallen med representationen för utfallen i form av maximum likelihood-skattning av oddskvot. Med olika värden på oddskvot-parametern har sannolikheten för samtliga möjliga utfall beräknats genom att använda hypergeometrisk likelihood-funktion (4.1) för oddskvoten. De olika värdena på oddskvot-parametern som används i figurerna är

oddskvot θ =0.886 (i fig 4.2),

oddskvot θ =3.13 (i fig 4.3),

oddskvot θ =1 (i fig 4.4) och

oddskvot θ =2.2 (i fig 4.5).

4.13. Figurer

Fig 4.1 Utfallsmängd. Antal möjliga utfall och utfallen i en ordnad utfallsmängd

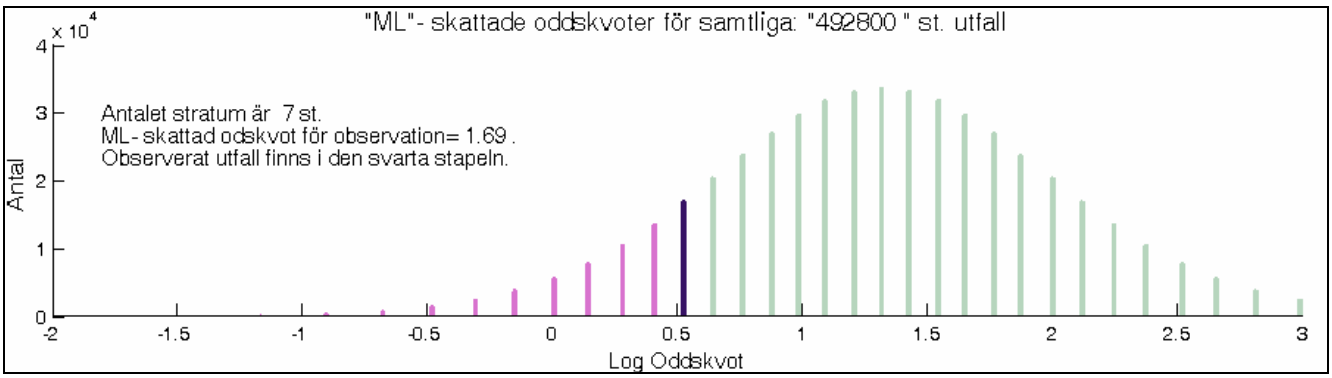


Fig 4.1
Kommentar 1. Datainput som används i fig 4.1 (och även i fig 4.2-4.5) hittas i tabell 5.2.2 och utgörs av 886 spelare i redovisningsgruppen "Förenade, samtliga spelare".
Kommentar 2. Fig 4.1 visar en ordnad mängd av alla möjliga utfall. Data är samtliga 886 spelare i 7 strata. Om antalet europeiska och amerikanska spelare skulle ha varit lika stora skulle centrum för stapelsamlingen förskjutas ifrån ~1.4 till noll.
Kommentar 3. Svart stapel innehåller c:a 20 000 olika möjliga utfall varav ett av dessa utfall är det observerade utfallet.

Fig 4.2 Konfidensintervall (90 %) nedre gräns

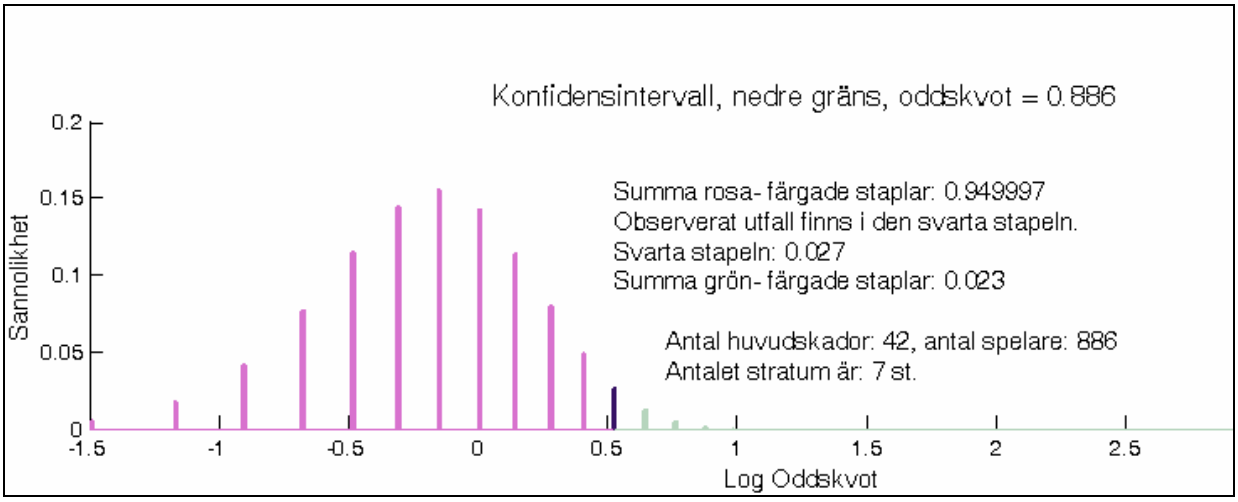


Fig 4.2
Kommentar 1. Fig 4.2 innehåller samma datainput som fig 4.1. Svarta stapeln innehåller c:a 20 000 stycken möjliga utfall. När sannolikheter ska beräknas, exempelvis för att vi ska erhålla ett utfall placerad i hopen av möjliga utfall i svarta stapeln så måste samtliga 20 000 separata sannolikheter för varje möjligt utfall i svarta stapeln beräknas och sedan adderas samman.
Kommentar 2. Sannolikhetsfördelning för oddskvotskattningar beror av värdet på oddskvot-parametern. Olika värden på oddskvot-parametern prövas iterativt ända tills att summan av rosa staplar blir 0.9500. En undre gräns för ett 90 % konfidensintervall för oddskvot har därmed beräknats.
Kommentar 3. Här ges en förklaring till denna tolkning av 90 % konfidensintervall. Skulle metoden upprepas med en bakomliggande oddskvot>0.886 och vi noterar ett nytt hypotetiskt observerat utfall, sannolikheten för att det nya observerade utfallet skulle vara lika stort eller större än det faktiska observerade värdet 1.69 (se fig 4.1) blir då >0.05. Genom att upprepa förfarandet ovan och även bilda en övre gräns så skapas ett konfidensintervall för oddskvot. Innanför intervallets gränser och för varje värde på den bakomliggande oddskvoten kommer ett bildat 90% sannolikhetsintervall (dubbelsi-

digt kring oddskvoten) för en ny hypotetisk observation och ny skattning av oddskvoten alltid inrymma det observerade värdet 1.69.

Fig 4.3 *Konfidensintervall (90 %) övre gräns*

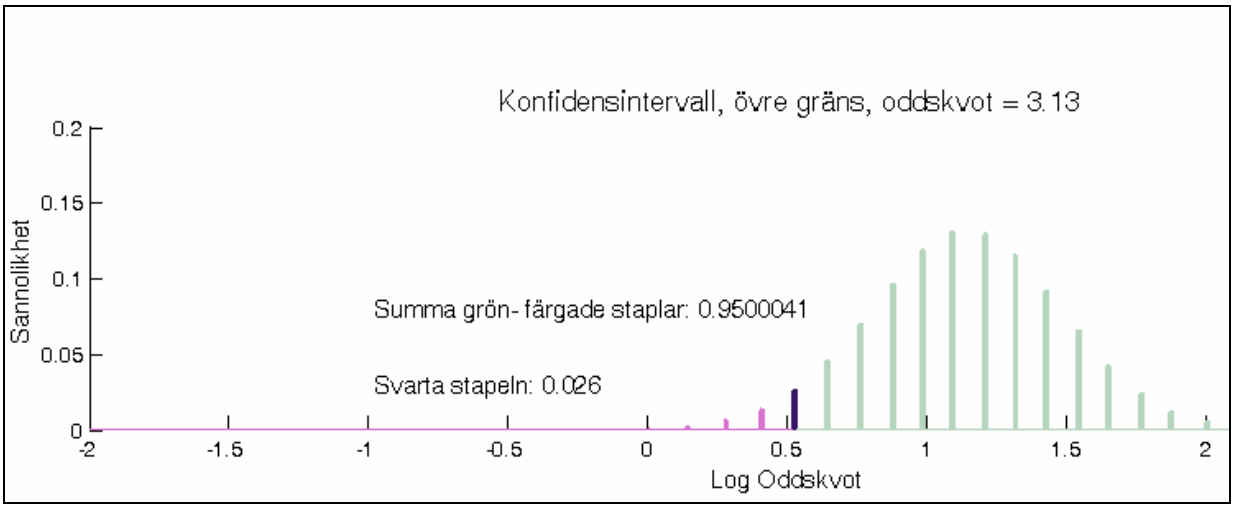


Fig 4.4 *P- värde, sannolikhet för observerat utfall (med test specificerat ovan)*

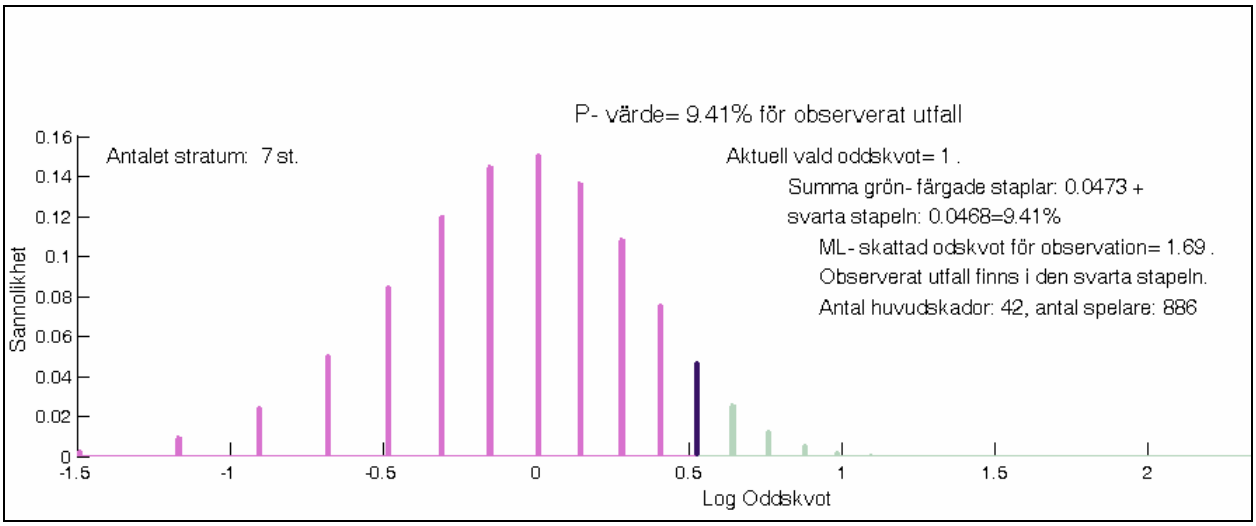
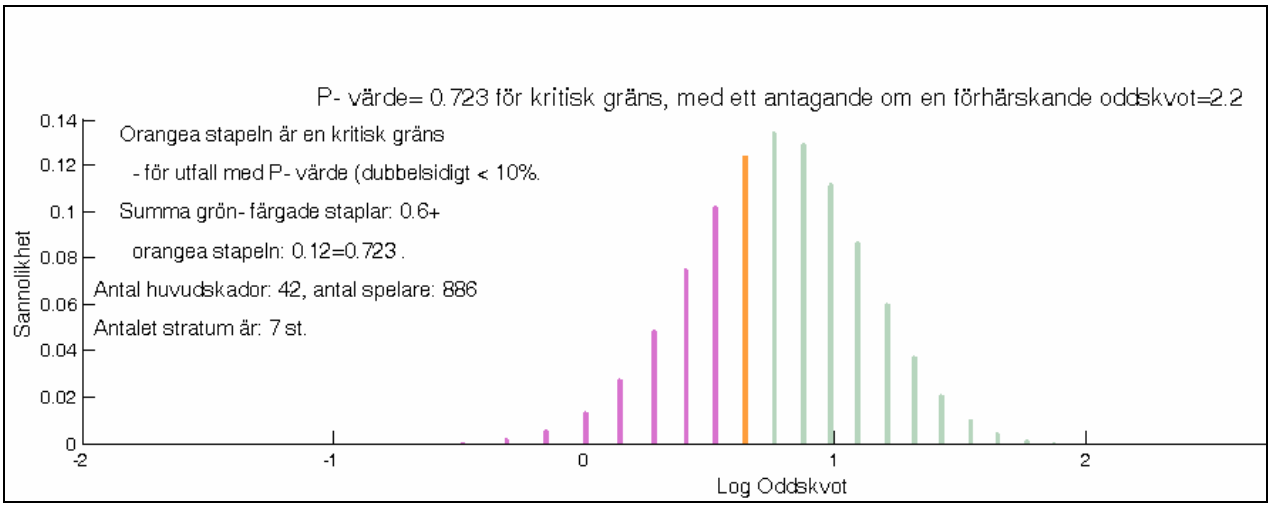
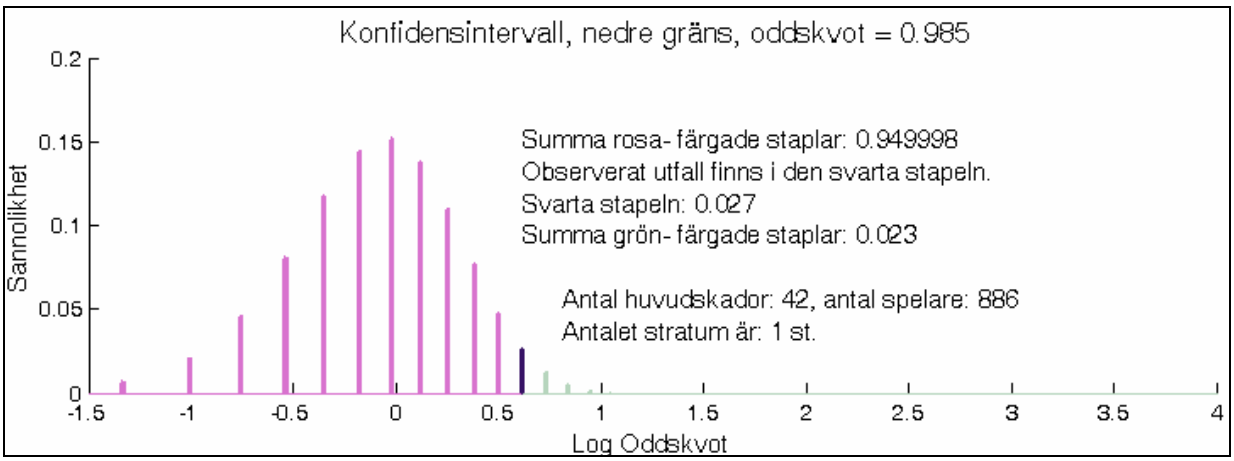


Fig 4.5 *Beräkning av styrka för metod*



4.14. Figurer, del 2, med ostratifierad datauppsättning

Fig 4.6 *Konfidensintervall (90 %) nedre gräns, ostratifierad analys*



5. Dataanalys och resultat

Varje redovisningsgrupp i datastudien finns både representerad i en ostratifierad representation av data och en stratifierad datauppsättning. Datatabeller med genomförda stratifieringar i två redovisningsgrupper redovisas nedan, redovisningsgrupperna är ”Forwards” och ”Förenade, samtliga spelare” (se tab 5.1.2 och 5.2.2). Data i tabellerna är uteslutande ifrån säsong 2013-14 i NHL för ishockeyspelare.

Utgående ifrån datauppsättningarna i datatabellerna nedan (i tab 5.1.2 - 5.2.2) ges underlag till att beskriva datautfallet med matematiska modeller i form av uppritade log likelihood-ratio kurvor för log oddskvot. Två hypergeomtriska likelihood-funktioner och två normala likelihood-funktioner används som matematiska modeller för aktuell data. Log likelihood-ratio kurvor för båda stratifierad data och ostratifierad data blir uppritade i figurer nedan, vilket blir fyra kurvor uppritade i varje graf och redovisningsgrupp (se fig 5.1 och fig 5.2). Uppritade log likelihood-ratio kurvor utgör ett komplement till dataanalysen för de fyra analysmetoderna av oddskvot som används i datastudien.

Resultat ifrån de fyra analysmetoderna för oddskvot redovisas i resultattabeller för de två redovisningsgrupperna ”Forwards” och ”Förenade, samtliga spelare” (se tab 5.3.1- 5.4.2). Slutligen kommer en sammanfattning av datastudien och några kommentarer kring vår exakta metod.

5.1. Datarepresentation och data för spelare 2013-14

Tab 5.1.1 Utan gruppindelning. Forwards. Primär redovisningsgrupp

Stratum	Européer (antal)		Amerikaner (antal)		Antal skadade		Observerat antal skadade européer	
	69	N_1	288	N_0	19	D	4	D_1

Tab 5.1.2 Gruppindelning av Forwards. Primär redovisningsgrupp

Stratum	Européer (antal)		Amerikaner (antal)		Antal skadade		Observerat antal skadade européer	
'Status'(+) 'Fysisk stil'(-) 1 (a)	18	N_1^1	42	N_0^1	5	D^1	2	D_1^1
'Status'(+) 'Fysisk stil'(+) 2 (a)	11	N_1^2	49	N_0^2	2	D^2	1	D_1^2
'Vikt'(+) 'Fysisk stil'(-) 3 (a)	16	N_1^3	51	N_0^3	7	D^3	1	D_1^3
'Fysisk stil'(-) 4 (a)	20	N_1^4	39	N_0^4	2	D^4	0	D_1^4
'Fysisk stil'(+) 5 (a)	4	N_1^5	54	N_0^5	2	D^5	0	D_1^5
Summa	69	N_1	235	N_0	18	D	4	D_1

Ett stratum föll bort som innehöll enbart 53 amerikanska spelare och med ett fall. Redovisningsgrupp: Forwards (a).

Tab 5.2.1 *Utan gruppindelning. Förenade, samtliga spelare*

Stratum	Européer (antal)		Amerikaner (antal)		Antal skadade		Observerat antal skadade européer	
	194	N_1	692	N_0	42	D	14	D_1

Tab 5.2.2 *Gruppindelning av Förenade, samtliga spelare*

Stratum	Européer (antal)		Amerikaner (antal)		Antal skadade		Observerat antal skadade européer	
'Status'(+) 'Fysisk stil'(-) 1 (u)	33	N_1^1	69	N_0^1	7	D^1	4	D_1^1
'Status'(+) 'Fysisk stil'(+) 2 (u)	20	N_1^2	81	N_0^2	9	D^2	3	D_1^2
'Vikt'(+) 'Fysisk stil'(-) 3 (u)	28	N_1^3	80	N_0^3	10	D^3	2	D_1^3
'Vikt'(+) 'Fysisk stil'(+) 4 (u)	6	N_1^4	87	N_0^4	3	D^4	0	D_1^4
'Fysisk stil'(-) 5 (u)	31	N_1^5	68	N_0^5	6	D^5	3	D_1^5
'Fysisk stil'(+) 6 (u)	13	N_1^6	85	N_0^6	3	D^6	0	D_1^6
Info saknas, Rookies 7 (u)	63	N_1^7	222	N_0^7	4	D^7	2	D_1^7
Summa	194	N_1	692	N_0	42	D	14	D_1

Redovisningsgrupp: Förenade, samtliga spelare (u).

5.2. Log likelihood-ratio grafer till de fyra metoderna

Uppritade log likelihood-ratio kurvor för oddskvot i log skala ger en kompletterande bild över hur de fyra metoderna fungerar för aktuell data. Log likelihood-ratio kurvorna kan ge ytterligare analyssvar om man skulle välja att använda kurvorna till det ändamålet. Några analyssvar har dock inte hämtats ur uppritade kurvor. Uppritade kurvor kan användas till en översiktlig grafisk jämförelse med de siffervärden som har levererats ifrån de fyra olika analysmetoderna för oddskvot. För samma redovisningsgrupp ritas fyra log likelihood-ratio plottar i samma graf. Plottarna är uppritade för den likelihood-funktion vilken respektive analysmetod utgår ifrån. Det är två normala och två hypergeometriska log likelihood-funktion kurvor uppritade i samma figur. Graferna ger bl.a. information om hur närliggande normal likelihood-funktionen är kurvan för hypergeometrisk likelihood-funktion för given datauppsättning. Graferna illustrerar även uppnådd effekt av stratifiering, förändring av varians samt värde på skattningen av oddskvot går att utläsa genom att jämföra plottar med eller utan stratifiering.

5.3. Uppritade log likelihood-ratio kurvor

Fig 5.1-5.2 Uppritade log likelihood kurvor som avser de fyra analysmetoderna:
1)Exakt metod ostratifierad data analys (hypergeometrisk log likelihood).
2)Normal approximation av Profile- likelihood och ostratifierad dataanalys.
3)Exakt metod stratifierad (hypergeometrisk log likelihood).
4)Normal approximation med Mantel-Haenszel metod i stratifierad dataanalys.
(Stratifierade datauppsättningen 5.1.2 är uppritad i fig 5.1. Ostratifierade datauppsättningen i tab 5.2.1 och stratifierade datauppsättningen i tab 5.2.2 är uppritad i fig 5.2.)

Fig 5.1 Fyra analysmetoders olika log likelihood-ratio kurvor, givna för data "Forwards"

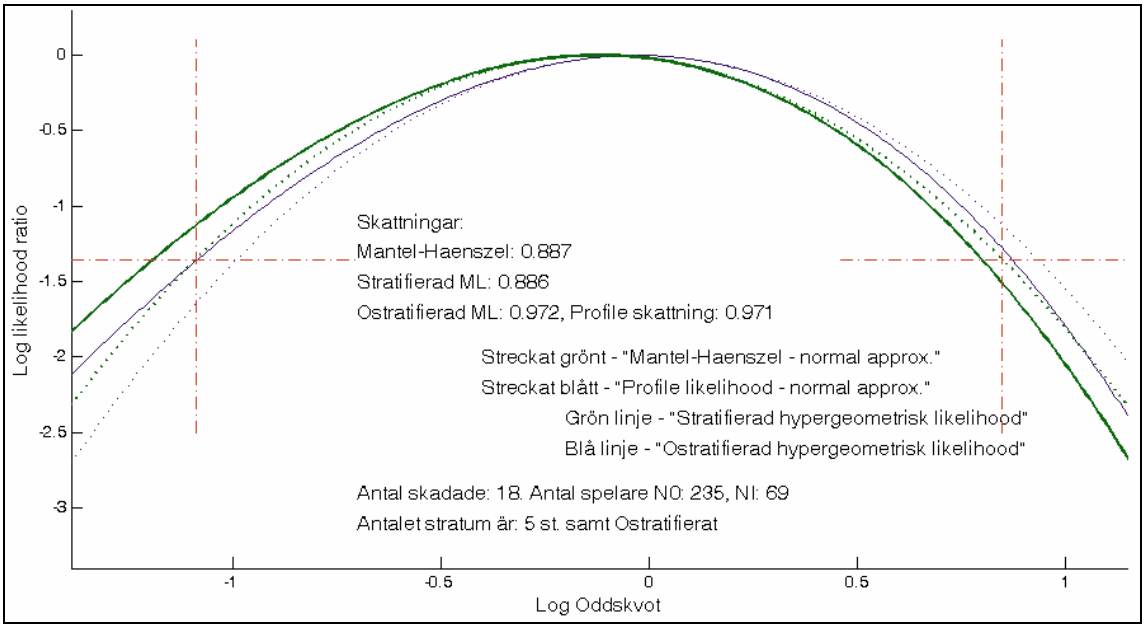


Fig 5.1
Anm 1. Skevhet beror av ojämn proportion av spelarkategorier 235/69.
Anm 2. Datainput för de fyra plottarna är enbart ifrån tabell 5.1.2. Forwards.

Fig 5.2 Fyra analysmetoders olika log likelihood-ratio kurvor, givna för data ”Förenade, samtliga spelare”

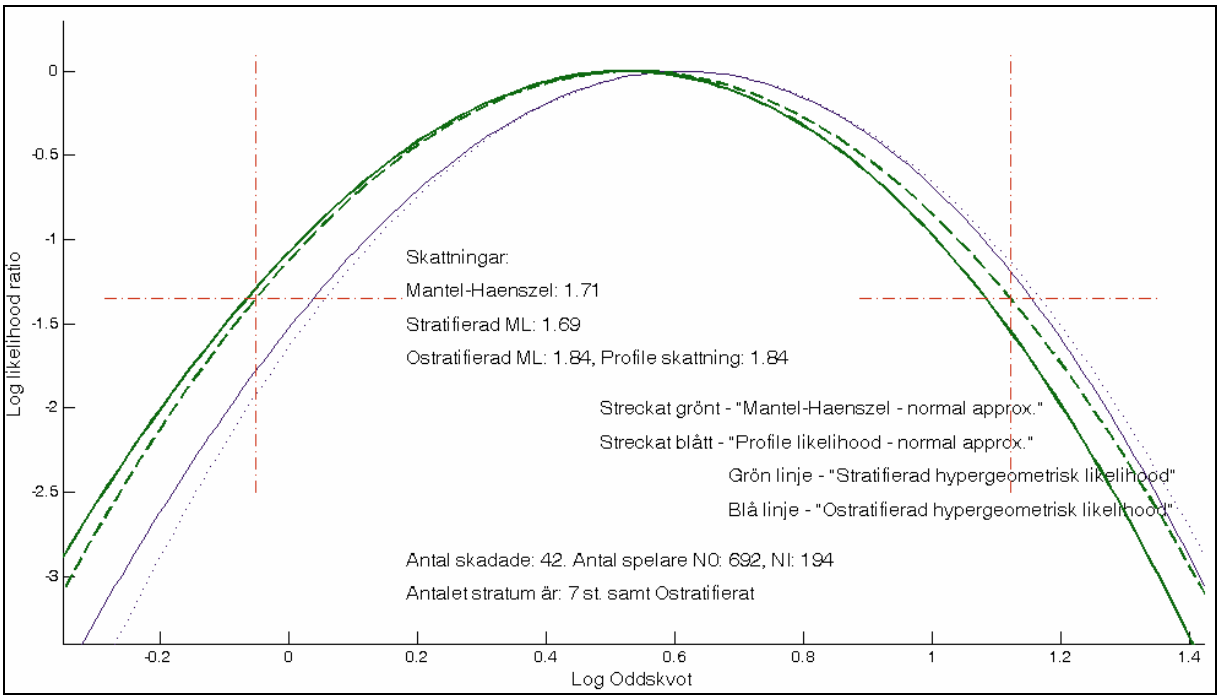


Fig 5.2
Anm. 1. Skevheten i blå plotten ostratifierad är nästan borta beroende av ökad datamängd.
Anm 2. Skevheten i (gröna) plotten för stratifierad hypergeometrisk likelihood består (men med minskad skevhet). Kvarvarande skevhet är inte enbart oönskad, utan tvärtom, skulle fördelningen i varje stratum vara identisk skulle vikigaste anledningen till stratifieringen inte vara verksam. Kvarvarande skevhet kan ses som ett tecken på en lyckad stratifiering.
Anm 3. Den horisontala streckade röda linjen är av ett intresse för att jämföra analysmetodernas 90% konfidensintervall. Där likelihood-ratio kurvan korsar den horisontala röda linjen motsvaras (motsvaras exakt eller approximativt får lämnas osagt) av gränserna i ett 90% konfidensintervall för en kontinuerlig fördelning som följer uppritad likelihood.
Anm 4. En konsekvens av att exakta metoden utgår ifrån en diskret utfallsmängd resulterar i att metoden också har en mer konservativ tolkning av konfidensintervall. Ostratifierad log likelihood-kurvor ger ett stöd för ett intervall med oddskvot>1. (Blåa plottarna korsar röda linjen när log odds >0.) Vår exakta metod ger inte ett stöd för oddskvot>1, i tabell 5.4.1. hittas istället värdet 0.985. Orsaken till att exakta metoden levererar ett värde mindre än =1 för nedre intervallgräns i ett 90 % konfidensintervall får en förklaring och kanske kan inses genom att granska den diskreta sannolikhetsfördelningen som bildas med en oddskvot =0.985 (i fig4.6) och konstatera att sannolikheten för det diskreta observerade utfallet är så stort som 2.7 % .

5.4. Resultat. Forwards spelare, 2013-14, primär redovisningsgrupp

Tab 5. 3. 1 **Forwards. Oddskvot. Utan gruppindelning. Sambandsförhållande mellan huvudskador och spelares ursprung, européer eller amerikaner** [1]

Olika beräkningsmetoder för konfidensintervall och sannolikhetsvärde samt olika metoder för skattning av oddskvot [2]

	Oddskvot (skattad) [c] $\hat{\theta}$	95 % Konf. Int. θ		90 % Konf. Int. θ		P- värde [d]
Numeriskt beräknade intervall, ML-skattning av oddskvot	1.12	0.262	3.67 [a]	0.332	3.15 [a]	52%
Profile likelihood, punkt skattning	1.12	0.36	3.49 [b]	0.432	2.91 [b]	~1/2

a. Exakta konfidensintervall.

b. Approximativa konfidensintervall.

c. Oddskvot (skattat värde). > 1 antyder en viss numerär övervikt 'concussions' bland europeiska spelare.

d. Sannolikhetsvärde för ett högre eller lika stort värde på skattning av oddskvot när ingen bakomliggande tendens om överrepresentation föreligger. (Enkelsidig mothypotes.)

1. Forwards. Spelande under säsong 2013-14 och med minst 10 matcher spelade under 2012-13. Ingen stratifiering. Antal strata =1. Se Tabell 5.1.1.

2. Dimensionener: Antal spelare =357, antal spelare med 'concussion' =19.

Tab 5. 3. 2 **Forwards. Oddskvot. Med gruppindelning. Sambandsförhållande mellan huvudskador och spelares ursprung, européer eller amerikaner** [1]

Olika beräkningsmetoder för konfidensintervall och sannolikhetsvärde samt olika metoder för skattning av oddskvot [2]

	Oddskvot (skattad) $\hat{\theta}$ [c]	95 % Konf. Int. θ		90 % Konf. Int. θ		P- värde [d]
Numeriskt beräknade intervall, ML-skattning av oddskvot	0.886	0.201	3.04 [a]	0.255	2.58 [a]	68%
Mantel-Haenszels skattning	0.887	0.28	2.81 [b]	0.337	2.33 [b]	~1/2

a. Exakta konfidensintervall.

b. Approximativa konfidensintervall.

c. Oddskvot (skattat värde). < 1 antyder en viss numerär övervikt 'concussions' bland amerikanska spelare.

d. Sannolikhetsvärde för ett högre eller lika stort värde på skattning av oddskvot när ingen bakomliggande tendens om överrepresentation föreligger. (Enkelsidig mothypotes.)

1. Forwards. Spelande under säsong 2013-14 och med minst 10 matcher spelade under 2012-13. Gruppindelning efter status($), vikt(kg) och 'aggressiv' spelstil'. Antal strata =5. Se Tabell 5.1.2.

2. Dimensionener: Antal spelare =304, antal spelare med 'concussion' =18. (Prospektiv data behandling, innebärande att inget urval av icke- skadade spelare är genomförd, detta är inte en fall- och kontrollstudie.)

5.5. Resultat. Förenade, samtliga spelare

Tab 5. 4. 1 **Förenade, samtliga spelare. Oddskvot. Utan gruppindelning. Sambandsförhållande mellan huvudskador och spelares ursprung, européer eller amerikaner** [1]

Olika beräkningsmetoder för konfidensintervall och sannolikhetsvärde samt olika metoder för skattning av oddskvot [2]

	Oddskvot (skattad) $\hat{\theta}$ [c]	95 % Konf. Int. θ		90 % Konf. Int. θ		P- värde [d]
Numeriskt beräknade intervall, ML-skattning av oddskvot	1.84	0.877	3.71 [a]	0.985	3.35 [a]	5.5%
Profile likelihood, punkt skattning	1.84	0.951	3.58 [b]	1.06	3.22 [b]	3.3%

a. Exakta konfidensintervall.

b. Approximativa konfidensintervall.

c. Oddskvot (skattat värde). > 1 antyder en viss numerär övervikt 'concussions' bland europeiska spelare.

d. Sannolikhetsvärde för ett högre eller lika stort värde på skattning av oddskvot när ingen bakomliggande tendens om överrepresentation föreligger. (Enkelsidig mothypotes.)

1. Förenade, samtliga spelare. Spelande under säsong 2013-14. Ingen stratifiering. Antal strata =1. Se Tabell 5.2.1.

2. Dimensionener: Antal spelare =886, antal spelare med 'concussion' =42.

Tab 5. 4. 2 **Förenade, samtliga spelare. Oddskvot. Med gruppindelning. Sambandsförhållande mellan huvudskador och spelares ursprung, européer eller amerikaner** [1]

Olika beräkningsmetoder för konfidensintervall och sannolikhetsvärde samt olika metoder för skattning av oddskvot [2]

	Oddskvot [c] (skattad) $\hat{\theta}$	95 % Konf. Int. θ		90 % Konf. Int. θ		P- värde [d]
Numeriskt beräknade intervall, ML-skattning av oddskvot	1.69	0.788	3.47 [a]	0.886	3.13 [a]	9.4%
Mantel-Haenszels skattning	1.71	0.85	3.43 [b]	0.951	3.07 [b]	6.4%

a. Exakta konfidensintervall.

b. Approximativa konfidensintervall.

c. Oddskvot (skattat värde). > 1 antyder en viss numerär övervikt 'concussions' bland europeiska spelare.

d. Sannolikhetsvärde för ett högre eller lika stort värde på skattning av oddskvot när ingen bakomliggande tendens om överrepresentation föreligger. (Enkelsidig mothypotes.)

1. Förenade, samtliga spelare. Spelande under säsong 2013-14. Antal strata =7. Se Tabell 5.2.2.

2. Dimensionener: Antal spelare =886, antal spelare med 'concussion' =42. (Prospektiv data behandling, innebärande att inget urval av icke- skadade spelare är genomförd, detta är inte en fall- och kontrollstudie.)

5.6. Kommentarer. Datastudie

Stöd saknades för sökta sambandet mellan ishockeyspelare och huvudskador i tillgänglig data och i primär redovisningsgrupp bestående av forwards spelande i NHL säsong 2013-14. Men för två redovisningsgrupper som inte varit i fokus för datastudien gav dataanalysen två ”funna resultat” som antyder ett samband och en bakomliggande tendens om att europeiska spelare i en högre grad är drabbade av hjärnskakningar än de amerikanska spelarna. Ett av de ”funna resultaten” som avses hittas i tabell 5.4.1, det andra ”funna resultatet” fanns i en redovisningsgrupp som har utelämnats av utrymmesskäl. Den approximativa dataanalysen gav signifikanta värden men motsvarande dataanalys för data med vår exakta metod gav inte signifikanta värden. Bortsett ifrån att en slutsats om signifikans inte fick stöd av båda metoderna och att betydelsen reduceras av att redovisningsgruppen inte varit i fokus i datastudien så finns ytterligare ett skäl att fästa liten vikt vid de ”funna resultaten”. Data bakom dessa ”funna resultat” är ostratifierad och data har inte blivit korrigerad genom en stratifiering som rensar undan effekter ifrån ovidkommande faktorer som eroderar innebörden av ett eventuellt samband.

5.7. Slutkommentarer. Metod

En exakt metod har utformats och givits namnet ”Numerisk analysmetod för oddskvot i stratifierad modell”. Metoden analyserar samband eller en förhärskande tendens som genererar överrepresentation. Som i stycket ovan nämnda exemplet med värden hämtade i tabell 5.4.1 så kan exakt metod och approximativ metod lämna olika svar. Olikheterna i exemplet i tabell 5.4.1 kan bäst förklaras av att den exakta metoden använder en konservativ tolkning av p-värden och konfidensintervall vilket är en direkt konsekvens av att metodiken i den exakta metoden vid beräkningen av p-värden och konfidensintervall beaktar att maximum likelihood-skattningar av oddskvot i aktuell modell har en påfallande diskret (och inte kontinuerlig) utfallsmängd. En ytterligare visad olikhet mellan analysmetoderna för oddskvot illustreras av att studera uppritade log likelihood-ratio kurvor för log oddskvot. Uppritade kvadratiska (symmetriska) normala log likelihood-ratio kurvor avviker ifrån uppritade något skeva hypergeomtriska log likelihood-ratio kurvor vid aktuell data. Uppritade log likelihood-ratio kurvor för log oddskvot (figurer 5.1 och 5.2) visar att när datamängden var liten (~20 fall) så var området för oddskvot som observerad data ger störst stöd för att inrymma sanna värdet på oddskvoten skevt. Med en större datamängd (~40 fall) minskade skevheten men försvann inte. Den exakta metodens analys av oddskvot är anpassad efter skevhet.

I redovisad datastudie, och generellt när datamängden inte är allt för stor, levererar vår exakta metod maximum likelihood-skattning av gemensam oddskvot för gruppindelad data med exakta konfidensintervall för oddskvot samt exakta sannolikhetsvärden för observerad data.

6. Käll- och litteraturförteckning

6.1. Tryckt material

[1 Clay] Clayton, David. Statistical Models in Epidemiology. Oxford University, sidor 166-178, 1998.

6.2. Internet och i uppsatsförfattarens ägo

[DB01] Vollman, Robert. Stats compiled by Robert Vollman. Place, URL ://www.hockeyabstract.com. Databas, 887 poster med samtliga aktiva spelare säsongsdata 2013-14 och 2012-13, 2014.