Stockholms universitet

Hur utvecklas jämna bordtennismatcher?

Konstantin Kalinichenko

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2009:12
Matematisk statistik
December 2009

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2009:12**,
http://www.math.su.se/matstat

Hur utvecklas jämna bordtennismatcher?

Konstantin Kalinichenko*

December 2009

Sammanfattning

En bordtennismatch leder till en sekvens av vunna och förlorade set. Ifall en match slutar exempelvis med 3-0, är det troligt att vinnaren var bättre än förloraren. Om däremot en bordtennismatch är jämn, dvs. sådan att antal set som är vunna av matchens vinnare och dess förlorare skiljer sig med ett mål (exempelvis 4-3), finns det en stor chans att seger av just denna spelare var rent slumpmässig. I denna uppsats undersöker vi ett stort antal jämna matcher (men även andra typer av matcher) för att ta reda på om det finns någon systematisk tendens inom dessa, dvs. matcher med en viss följd av segrar/förluster förekommer oftare än de borde göra, eller att vinnarna av jämna matcher vinner oftast även de första seten, eller tvärtom. Undersökningen är baserad på två probabilistiska modeller - Bernoulli-modellen och Markov-modellen. Analys av data med hjälp av dessa två modeller resulterade i att vi inte kunde konstatera att det finns någon/några systematiska tendenser i data, med ett undantag. I ett datamaterial, som avsåg barntävlingen Folksam-Talangen, kunde en signifikant inlärningseffekt påvisas, i att slutliga vinnaren oftare tagit över matcher efter att ha förlorat inledande set, än tvärtom.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: kk1986@gmail.com. Handledare: Rolf Sundberg.

Abstract

A table tennis match results in a sequence of won and lost sets. In case a match ends as 3-0, it is likely that the winner was better than the loser. If a table tennis match is "even", i.e. that the difference between the number of sets that are won by the winner of the match and the number of sets that are lost by that individual is one (e.g. 3:2 or 4:3), there is a greater chance that the victory by this particular player was due to pure chance. In this essay we are looking at a large number of even matches (but also less even matches) to investigate whether there is any systematic tendency in table tennis matches, for instance, that matches with a specific chain of victories / losses occur more often than expected, or that the winners of the even games usually win even the first sets, or vice versa. The study is based on two probabilistic models - Bernoulli-model and Markov-model. Analysis of the data using these two models resulted in that we could not conclude that there is / are any systematic trends in the data, with one exception. While analysing the data material for children's contest "Folksam-talangen", a significant learning effect has been observed, which means that the final winner won the games more often after losing the opening set, than vice versa.

Förord

Denna uppsats utgör ett examensarbete om 15 högskolepoäng och leder till en kandidatexamen i matematisk statistik vid Matematiska istitutionen, Stockholms Universitet. Arbetet med uppsatsen har pågått under perioden mars – november 2009.

Jag vill tacka min handledare, Rolf Sundberg vid Matematiska institutionen, Stockholms Universitet, för ovärderlig rådgivning och hjälp.

Innehåll

1 Introduktion **5**

1.1 Inledning . . . 5

1.2 Syfte och metod . . . 5

2 Beskrivning av data **7**

3 Statistisk undersökning av data **8**

3.1 Bernoulli-modellen . . . 8

3.1.1 Vinstsannolikheten . . . 8

3.1.2 Statistisk test . . . 11

3.1.3 Test av fördelning av vunna set . . . 12

3.1.4 Inlärningseffekterna inom jämna bordtennismatcher . . . 13

3.2 Markov-modellen . . . 18

3.2.1 Beskrivning av modellen . . . 18

3.2.2 Resultatena och undersökning av eventuell avvikelse från modellen . . . 21

4 Diskussion och slutsatser **23**

4.1 Slutsatser . . . 23

4.2 Diskussion . . . 24

A Härlendning av ML-skattningar **27**

B Sekvenser av jämna matcher **31**

1 Introduktion

1.1 Inledning

Bordtennismatcher spelas i bäst av 5 eller 7 set, alltså först till 3 respektive 4 vunna set. Matcher som slutar med 3-0 eller 4-0 säger i de flesta fall att vinnaren var bättre än förloraren. Matcher där förloraren tar set är mer intressanta, särskilt om de inte avgörs förrän i sista set, dvs. slutar med 3-2 eller 4-3. Låt oss kalla dem för jämna matcher. Faktumet att just den ena spelaren blir vinnare i en sådan match kan bero på psykologiska skäl, inlärningseffekter och möjligtvis något mer. Ett seger i en jämn bordtennismatch kan vara en ren tillfällighet också. Sådana effekter kan vi inte studera på något direkt sätt ur matchstatistik.

Vi kan dock undersöka förekomsten av eventuella systematiska tendenser på ett indirekt sätt. När brukar förlorarna ta sina set i sådana matcher? Vilka sekvenser är vanligast, eller är alla sekvenser i någon mening lika vanliga? Vad kan man tänka sig av enkla probabilistiska modeller, och passar de med data?

Exempelvis, om seten betraktas som Bernoulli-försök med vinstsannolikhet 0.5 för vardera spelaren, är då alla tänkbara sekvenser som leder till en 3-2 eller 4-3-vinst lika sannolika? Ifall man går igenom matchstatistik, kommer det då att visa sig att data är förenliga med den modellen? Eller finns det mönstren som signifikant avviker i frekvens? Till exempel, låt oss anta att vinnarna av jämna matcher spelar oftast bättre än deras motståndare i de sista seten, men ej i de första. I fall det verkligen ägde rum, skulle sekvenser där förlorarnas set kommer tidigt förekomma oftare än man skulle förvänta sig. Det kan alternativt vara så att spelarna slappnar av efter ett par inledningsvis vunna set och släpper något eller några set innan de koncentrerar sig och avgör matchen.

1.2 Syfte och metod

I detta arbete kommer vi att undersöka faktiska data med avseende på eventuell förekomst av systematiska tendenser, och dra slutsatser ur de resultat som kommer att uppnås. Detta kommer att göras med hjälp av två statistiska modeller: Bernoulli-modellen och Markov-modellen.

Bernoulli-modellen. Denna modell utgår från ett antagande att en sekvens av set i en bordtennismatch ses som oberoende Bernoulli-försök med en viss parameter p. Med detta antagande som bas kommer vi att konstruera en statistisk modell för hur en typisk bordtennismatch skall se ut. Vidare kommer vi att jämföra vår modell med faktiska data, undersöka om avvikelser i datamaterialet är statistiskt signifikanta, och dra slutsatser.

En del frågor som kommer upp då man ska använda Bernoulli-modellen avser parametern p. Vi kommer att klargöra dess betydelse i modellen, och undersöka ett eventuellt samband mellan den och sannolikheter för de olika sekvensernas förekomst.

Den allmänna Bernoulli-modellen kommer vi att tillämpa på alla typer av matcher vi har, det enda kravet är naturligtvis att förlorare ska ta minst ett set i dessa matcher. De jämna matcherna, som är av störst intresse för oss, kommer vi även att undersöka med en modifikation av allmänna undersökningsmetoden inom Bernoulli-modellen. Den går ut på att testa om det finns ett beroende inom olika set i jämna matcher.

Markov-modellen. Här kommer vi att se på seten i en bordtennismatch av typ 4-3 eller 3-2 som på element i en Markovkedjemodell. Kärnan av denna modell blir övergångssannolikheten, dvs. sannolikheten att ett utfall i nästa set inte blir samma som i det föregående. Vi kommer att skapa en teori-baserad modell för denna, givet att några systematiska tendenser i data inte existerar. Vidare kommer vi att få fram en empirisk skattning av denna parametern, och slutligen undersöka om den empiriska skattningen och modellen passar ihop.

Vi kommer att tillämpa Markov-modellen enbart på data avseende jämna matcher.

2 Beskrivning av data

Detta arbete är baserat på två omfattande datamaterial. Det första datamaterialet avser bordtennismatcher på elitnivå (matcherna spelas till 4 vunna set). Detta inkluderar material avseende Sveriges Mästerskap i Bordtennis år 2004 – 2009, Junior- och Ungdoms-SM år 2002-2009, samt Swedish Junior Open 2009. Detta datamaterial innehåller totalt 411 st. observationer, varav :

- 75 st. jämna matcher, dvs. sådana som slutade med 4-3
- 92 st. matcher med slutresultatet 4-2
- 115 st. matcher av typ 4-1
- 129 st. matcher av typ 4-0

Källa för dessa matchdata är Svenska Bordtennisförbundets matcharkiv tillgänglig via deras hemsida - www.svenskbordtennis.se.

Det andra datamaterialet avser matcher från lagtävlingen "Rookie-Cupen" (2009) och den individuella tävlingen "Folksam-Talangen" säsongen 2008 – 2009. Matcher i tävlingar av denna typ spelas till 3 vunna set, och deltagare av dessa är relativt oerfarna pojkar upp till 15 år gamla (i regel yngre) respektive barn upp till 10 år. Detta material består av 266 st. observationer av typ 3-2 samt 449 st. observationer av typ 3-1. Dessa data kommer ifrån matchprotokoll som Stockholms Bordtennisförbund har ställt till förfogande.

De två turneringarna som inkluderas i det andra materialet kommer även att undersökas separat, eftersom dessa är två stora datamaterial var för sig, och dessutom är åldern för deltagare av dessa två turneringar inte samma.

Materialet avseende turneringen "Folksam-Talangen" innehåller 201 observationer av typ 3-1 och 119 observationer av typ 3-2. Data gällande "Rookie-Cupen" består av 248 respektive 147 observationer av typ 3-1 respektive 3-2.

3 Statistisk undersökning av data

Undersökningarna genomfördes med hjälp av datorprogrammet MatLab. Uträkningarna av nödvändiga teststatistikor och p-värden genomfördes både i MatLab och i externa fördelningsminiräknare (distribution calculator), som finns tillgängliga på internet.

3.1 Bernoulli-modellen

Huvudantagande i denna modell är att seten i en bordtennismatch ses som Bernoulli-försök med en viss parameter p. Till exempel: antag att spelare A och B deltar i en bordtennismatch och spelare A vinner den. Om matchen spelades först till 4 vunna set, och förlorare (spelare B) har vunnit 2 set i den, kan den t.ex. se ut såhär: AABABA, BAAABA eller ABAABA.

För att kunna genomföra en statistisk undersökning av data med denna metod skall vi göra ett antagande att olika set i en match är oberoende, och även att resultaten i olika matcher är oberoende, men med parametern p som varierar mellan matcher.

3.1.1 Vinstsannolikheten

Vi börjar uppbyggnad av modellen med undersökningen av dess viktigaste komponent, nämligen sannolikhetsparametern p. Först måste vi ge en exakt tolkning till denna, vilket kommer att ge oss möjlighet att ta reda på dess egenskaper och dess inflytande på vår modell. Den tolkningen som man möjligtvis tror är den bästa, är att se p som en individuell parameter som uttrycker individens förmåga att vinna. En sådan definition av p medför dock en del komplikationer vid konstruering av en lämplig modell, eftersom det är två spelare inblandade, vilket innebär att p inte kan vara individuellt. Därför definierar vi p på ett mer allmänt sätt:

Definition. Antag att A och B spelar en bordtennismatch. I Bernoulli-modellen betraktar vi varje set som ett Bernoulliförsök, där A har (den okända) vinstsannolikheten p. Sannolikheten p får bero på det konkreta paret av individerna, och andra omständigheter, såsom ställe där matchen äger rum, deltagarnas psykologiska tillstånd osv.

Konstrueringen av den fullständiga modellen kräver dock att vi besvarar ytterligare frågor avseende p. Hur är vinstsannolikheten p kopplad till sannolikheterna för uppkomst av de olika sekvenserna i en bordtennismatch? Kommer en eller flera av sekvenserna att dyka upp oftare (eller mer sällant) än förväntat då man förändrar p, givet att matcher är av en viss typ, exempelvis 4-3? Vi besvarar dessa frågor med en sats som kommer nedan. Innan det ska vi dock införa följande två beteckningar:

Låt s vara antal set i en bordtennismatch som man behöver vinna för att vinna hela matchen. Låt vidare t stå för antal set som matchens förlorare har tagit / skall ta i denna match. I en (avslutad) bordtennismatch av typ *s:t* har alltså vinnaren vunnit s set, och förloraren t set.

Exempelvis om $s = 3$, och $t = 2$, är en bordtennismatch med dessa parametrar av typ *3:2*. Vi fortsätter med att illustrera satsen:

Sats 1. Givet en viss sluträkning, saknas det något samband mellan vinstsannolikheten p och sannolikheten att en av sekvenserna kommer upp.. Dessutom så är samtliga sekvenser lika sannolika, och sannolikheten att en viss sekvens förekommer är enbart beroende av s och t. Detta gäller för alla matcher av typ *s:t*, där s,t - valfria positiva heltal med $s > t$.

Bevis. Antag att en bordtennismatch är av typ *s:t* enligt ovan. Det totala antalet set i denna match blir $s+t$. Låt vidare $X_1 ... X_{s+t}$ vara stokastiska variabler sådana att $X_1 ... X_{s+t-1} \sim Be(p)$. Dessa variabler står alltsa för samtliga seten i matchen, exklusive det sista. Låt vidare $P(X_{s+t} = 1) = 1$. Vi gör alltså en betingning som går ut på att "titeln" *1* blir tilldelat till hela matchens vinnare, och *0* - till dess förlorare. Vi fortsätter med att konstatera att eftersom matchen spelas till s vunna set, och $P(X_{s+t} = 1) = 1$, måste

$$\sum_{i=1}^{s+t-1} X_i = s - 1$$

Sannolikheten att en viss sekvens dyker upp kan följaktligen skrivas på nedanstående sätt:

$$P(X_1 = x_1, ..., X_{s+t-1} = x_{s+t-1} \mid \sum_{i=1}^{s+t-1} X_i = s - 1) =$$

$$= \frac{P(X_1 = x_1, ..., X_{s+t-1} = x_{s+t-1} \cap \sum_{i=1}^{s+t-1} X_i = s - 1)}{P(\sum_{i=1}^{s+t} X_i = s - 1)} \qquad (1)$$

Övre delen av (1) kan skrivas som $p^{s-t}(1 - p)^t$. Eftersom $X_1 ... X_{s+t-1}$ är i.i.d., är:

$$\sum_{i=1}^{s+t-1} X_i \sim Bin(s + t - 1, p)$$

Detta medför att man kan skriva om nämnaren i uttrycket ovan som:

$$P(\sum_{i=1}^{s+t-1} X_i = s - 1) = \binom{s + t - 1}{s - 1} p^{s-t}(1 - p)^t$$

Vi får då att uttrycket för (1) kan skrivas om som:

$$P(X_1 = x_1, ..., X_{s+t-1} = x_{s+t-1} \mid \sum_{i=1}^{s+t-1} X_i = s - 1) =$$

$$\frac{p^{s-t}(1-p)^t}{\binom{s+t-1}{s-1}p^{s-t}(1-p)^t} = \frac{1}{\binom{s+t-1}{s-1}} = ... = \frac{(s-1)!t!}{(s+t-1)!}\square$$

Satsen är härmed bevisad, eftersom uttrycket ovan inte innehåller p.

Det kan dock vara intressant att veta att det finns ett starkt samband mellan p och sannolikheten att en slumpmässigt vald match är jämn. Vi börjar med följande definition:

Definition. Låt

$$C(s:t) = \frac{(s+t-1)!}{(s-1)!t!}$$

Stå för antalet möjliga sekvenser i en match av typ *s:t* .

Vi illustrerar vidare följande sats:

Sats 2. Sannolikheten att en slumpmässigt vald match är jämn går mot 0 då $p \to 0$ eller $p \to 1$ och har sitt maximum då $p = 0.5$.

Bevis. . Låt *P(s:t)* beteckna att givet att matcher spelas till s vunna set, är en slumpmässigt vald match just en *s:t* match. Med hjälp av formeln ovan får vi att:

$$P(s:s-1) = P(t = s-1| \sum_{i=1}^{s+t-1} X_i = s-1) =$$

$$P(sekvens\, av\, typ\; s:s-1| \sum_{i=1}^{s+t-1} X_i = s-1) =$$

$$p^{s-1}(1-p)^{s-1}C(s:s-1).$$

Vi deriverar detta med avseende på p. Eftersom *C(s:s-1)* inte innehåller p, får vi:

$$\frac{d}{dp}(p^{s-1}(1-p)^{s-1}C(s:s-1)) = ... = C(s:s-1)(s-1)(p-p^2)^{s-2}(1-2p)$$

Sätter man denna derivata lika med noll får man tre stationära punkter: $p =$ *0, 0.5, 1.* Det är lätt att se att *P(s:s-1)* som en funktion av p antar sitt minsta värde (= *0*) då p=*0* eller *1*, respektive sitt största värde då p=*0.5*. Uttrycket för uträkning av största värdet är lika med $0.5^{2s-2}C(s:s-1)$. Detta är ekvivalent med *0.3125* för matcher av typ 4-3, respektive *0.375* för matcher av typ 3-2 $\square$

3.1.2 Statistisk test

Vi börjar med att illustrera antal möjliga sekvenser för samtliga typer av matcher som våra datamaterial består av. Enligt resultatena som fåtts i den föregående avsnittet fås dessa genom att räkna ut $\frac{1}{C(s:t)}$.

Typ av match (s : t)	4:1	4:2	4:3	3:1	3:2
Antal möjliga sekvenser ($\frac{1}{C(s:t)}$)	4	10	20	3	6

Nu, när vi har räknat ut antal möjliga sekvenser för samtliga typer av matcher, ska vi börja analysera våra faktiska data med avseende på om någon (några) av sekvenserna förekommer oftare eller mer sällant än de resterande.

Detta kommer att implementeras enligt följande: Vi tar fram data för hur ofta olika sekvenser förekommer i en viss typ av match. Sedan räknar man ut de förväntade frekvenserna av en sekvens förekomst, vilka är alltså samma för samtliga sekvenser enligt Sats 1. Slutligen jämför man de förväntade frekvenser med de faktiska, och undersöker om avvikelsen mellan dessa är statistiskt signifikant. Detta kommer att implementeras med Pearsons Chi-två test (Se [Lindgren,1993] för härledning av fördelning för teststatistika i denna test).

Definition. Låt L vara det totala antalet spelade matcher av en viss typ, och l – antal möjliga sekvenser. Låt vidare $i = 1 \ldots l$, O_i- antal sekvenser av typ i som förekom, och E_i- det förväntade antalet sekvenser. Vår noll-hypotes blir

$$H_0 : E_i = \frac{L}{l}$$

Teststatistikan

$$X^2 = \sum_{i=1}^{l} \frac{(O_i - E_i)^2}{E_i}$$

är asymptotiskt χ^2-fördelad med *l-1* frihetsgrader.

Dessa teststatistikor illustrerar alltså hur mycket faktiska data avviker från vår noll-hypotes. Slutligen räknar man ut p-värden för samtliga teststatistikor. Se tabell 1 på nästa sida för resultatena och p-värden.

Det måste noteras att ett av villkoren för tillämpning av Pearsons Chi-två test är att antalet observationer i en cell (E_i) behöver vara tillräckligt stort, vanligtvis större än *5*. Om antal celler är betydligt, anses det dock acceptabelt att tillämpa det ovannämnda testet även om *20* eller mindre procent celler innehåller *5* observationer eller mindre (dock inga celler med noll observationer). Dessutom

Typ av match (nivå, turnering)	Teststatistika	Antal frihetsgrader	P-värde
3-1 (Barn-samtliga)	3.1133	2	0.21
3-1 (Folksam-Talangen)	2.8955	2	0.24
3-1 (Rookie-Cupen)	4.8313	2	0.09
3-2 (Barn - samtliga)	6.1364	5	0.29
3-2 (Folksam-talangen)	7.6500	5	0.18
3-2 (Rookie-Cupen)	2.9200	5	0.71
4-1 (Elit)	1.1724	3	0.76
4-2 (Elit)	9.5556	9	0.39
4-3 (Elit)	14.2500	19	0.77

Tabell 1: Resultatena av dataundersökningen: Det allmänna testet

bör det totala antalet observationer inte vara för liten, *50* eller mer är tillräckligt. (Se t.ex. [Snedecor and Cochran,1989] eller [Lindgren,1993]).

Data som avser matcher av alla typer förutom 4-3 är tillräckligt omfattande både när det gäller det totala antalet observationer i varje typ av match, och antalet observationer inom celler (se Beskrivning av data). I matcher av typ 4-3 är det andra kravet inte uppfyllt, eftersom i en del av cellerna är antalet observationer lägre än *5* (inga av de är dock toma celler). Det resultatet som vi har fått för 4-3 matcher illustrerar dock ganska tydligt (P-värde = *77%*) att avvikelse från vår modell är irrelevant. Det är därför lämpligt att anta att det faktumet att antal observationer i en del av celler är för lågt kommer inte att förändra vår slutsats angående 4-3 matcher.

3.1.3 Test av fördelning av vunna set

Jämna matcher av typ 4-3 och 3-2 har vi undersökt med en lite annorlunda metod. Den går ut på att analysera beroendet mellan olika set genom att titta på procentuellt antal segrar i varje set, och titta sedan på avvikelserna mellan faktiska data och det som hade förväntats. Enligt vår modell, skall procentuellt antal segrar (och procentuellt antal förluster) i varje set utom det sista ligga runt 50%. Som förut, kommer vi att undersöka om avvikelserna är statistiskt signifikanta med hjälp av ett Chi-två test. Den empiriska datan som kommer att undersökas tabuleras på nästa sida (se tabell 2). Vi betecknar det observerade antalet segrar i set i med O_i , och låter E stå för medelvärden av summan av O_i. Data för förekomsten av varje sekvens för sig hittar man i Appendix B.

Teststatistikor för denna test har samma struktur som i det föregående avsnittet. Vi måste dock ta hänsyn till att resultat i en match är negativt korrelerade,och därför multiplicerar vi våra teststatistikor med

$$\frac{s+t-2}{s+t-1-a} = [t = s-1] = \frac{2s-3}{2s-2-a}$$

Typ av match	O_1	O_2	O_3	O_4	O_5	O_6	$E = \frac{\sum O_i}{n}$
3-2 (Barn-samtliga)	127	118	141	146			133
3-2 (Folksam-Talangen)	52	52	69	65			60
3-2 (Rookie-Cupen)	75	66	72	81			74
4-3 (Elit)	32	37	42	45	35	34	38

Tabell 2: Fördelningen av segrar bland olika set

Typ av match (nivå)	Teststatistika	Antal frihetsgrader	P-värde
3-2 (Barn-samtliga)	5.5714	3	0.13
3-2 (Folksam-Talangen)	5.8500	3	0.12
3-2 (Rookie-Cupen)	2.3919	3	0.5
4-3 (Elit)	5.5702	5	0.35

Tabell 3: Resultatena av dataundersökning: Test av fördelning av vunna set

, där $a =$ antal vunna set (exklusive det sista som alltid vinns) $= 2$ eller 3 (Personlig information: Rolf Sundberg). Resultatena av undersökningen illustreras i tabell 3.

3.1.4 Inlärningseffekterna inom jämna bordtennismatcher

Trots att dataundersökningen i det föregående avsnittet inte visade på några större avvikelser, kan man se att fördelningen av antal vunna set inte är helt jämn. När det gäller turneringen Folksam-Talangen, är det väldigt tydligt att matchernas vinnare är betydligt mer aktiva i de sista seten än i de första. En sådan ojämn fördelning av data kan peka på att det förekommer inlärningseffekter hos bordtennisspelare eller med andra ord att spelarna lär sig sin motstandares teknik och tar över matchen. I detta avsnitt kommer vi att testa om sådana effekter förekommer.

Låt oss anta att inlärningseffekterna verkligen äger rum. I sådant fall skulle avgörande majoritet av matcher av typ AABB .(s = 5) respektive AAABBB . (s = 7) vinnas av just spelare B.

Exempel. Antag att en bordtennismatch spelas till 3 vunna set. Antag vidare att spelare A vinner två set i rad, men B lyckas utjämna till 2-2. Om B har lärt sig sin motståndare under matchens gång och därför lyckats vända matchen så långt, så bör B ha större chans att vinna ett avgörande set. Dock i fall inlärningseffekterna inte förekommer, kommer både spelare A och B ha samma chans att vinna det sista setet, och därmed, hela matchen.

Alternativt, anta att spelare A skärper sig efter två eller tre förlorade set, och tar hem sista set. I fall en sådan företeelse verkligen äger rum, skulle majoriteten av matcher av typ AABB. (respektive AAABBB.) vinnas just av spelare A. (Punkten i slutet av sekvenserna svarar alltså mot det setet vars utfall vi är intresserade av). Motsvarande effekter inom matcher med inledningen AB... eller

A.... (5-setare), eller A...... (7-setare) borde vara svagare än i ovannämnda exemplet. Detta beror först och främst på att en kortare bas (antal förbestämda set i början av en match) innebär ett högre antal sekvenser som ligger inom ramer for vår modell, vilket leder till att samband mellan en viss tendens (t.ex. att A vinner matcher av en viss typ oftare) och vår eventuella definition av detta (A vinner oftare ⇒Spelare skärper sig vanligtvis och därför vinner de) inte blir lika konkret och tydligt.

I vår undersökning kommer vi att titta på matcher som börjar med AABB. (5-setare), samt AAABBB., och AAB... (7 setare). I de första två fallen, är det enbart ett par sekvenser som vi kommer att titta på. I det sista fallet, blir det dock tre olika par av sekvenserna, nämligen AABABB. , AABBAB. , och AABBBA. . Det som är gemensamt för dessa är att de samtliga börjar med AA, så man kan fortfarande tala om eventuella inlärnings (skärpnings) effekter med avseende på spelare A(B). Dock kommer resultat av undersökning av dessa sekvenser inte att ha samma styrka, som resultatet av det ”rena” fallet (AAABBB).

Vi börjar undersökningen med att notera att enligt våra antaganden (se början av detta kapitel), anses seten i en bordtennismatch vara oberoende och Bernoulli-fördelade stokastiska variabler, eller tekniskt talat, $X_1...X_{s+t} \sim Be(p)$. Notera även att en av förutsättningarna är att det sista setet skall alltid vinnas av hela matchens vinnare, dvs. $P(X_{s+t} = 1) = 1$. Denna förutsättning innebär att samtliga sekvenser har samma utfall i set s+t, och därför bör vi använda något trick för att kunna undersöka våra datamaterial med avseende på förekomsten av inlärningseffekterna, eftersom denna undersökning går just ut på att analysera utfall i sista set.

Först noterar vi att i fall man inte gör någon betingning m.a.p. sista set, kan varje sekvens skrivas på två form. Dessa är sekvensen själv, samt dess spegelbild. Exempelvis, sekvenserna AABBA och BBAAB är egentligen en och samma sekvens. Detta innebär att vi kan genomföra undersökningen av inlärningseffekterna på följande sätt: Antag att vi vill undersöka förekomst av sekvenserna AABBA och AABBB. Den sista existerar inte i vår data, så vi ersätter den med dess spegelbild: BBAAA. Vi kommer alltså att titta på förekomsten av AABBA med avseende på BBAAA, och i fall den senare sekvensen förekommer mycket mer sällant än den första, är det ett tecken på att inlärningseffekterna äger rum i detta datamaterial.

Vi fortsätter med att konstatera att enligt Sats 1, är samtliga sekvenser lika sannolika i en bordtennismatch av typ s:t. Denna sannolikhet är enbart beroende av s och t, men ej av sannolikhetsparameter p. Exempelvis, sannolikheten att det första setet i en bordtennismatch vinns av spelare A är p. Men sannolikheten att det första setet i en bordtennismatch vinns av spelare A *givet att denna match kommer att avslutas med sluträkningen s:t* är helt oberoende av p. En naturlig fråga uppstår här: vad är sannolikheten ovan lika med? Detta besvaras med nedanstående sats, men först inför vi några viktiga begrepp. Vi börjar med att låta $\mathbf{X}_n = X_1, ..., X_n$, och $\mathbf{x}_n = x_1, ..., x_n$. Låt vidare J beteckna faktumet att en match är jämn och att det sista setet i den skall alltid vinnas av hela matchens vinnare. Beteckningen J är helt enkelt en förkortning för $X_{s+t} = 1 \cap t = s - 1$. Med hjälp av dessa beteckningar kan vi exempelvis illustrera Sats 1 som $P(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1}|J) = \frac{1}{C(s:t)}$. Vi förtsätter genom att definiera kärnparametern av vår underökning.

Definition. Låt

$$\mu = P(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1}|(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1} \cup \overline{\mathbf{X}_{s+t-1}} = \overline{\mathbf{x}_{s+t-1}}) \cap J)$$

Detta uttryck innebär sannolikheten att ur två sekvenser som är komplement till varandra exklusive sista set, kommer den ena att dyka upp givet att det sista setet slutar med A, och givet att matchen är jämn.

Som man ser, är μ precis den sannolikheten som vi är intresserade i. Vi fortsätter med att illustrera satsen:

Sats 3

$$\mu = \frac{1}{2}$$

Bevis. Parametern μ innebär sannolikheten att givet att någon av två sekvenserna skall förekomma med sannolikheten 1, skall just den ena (eller den andra) att dyka upp. Vi får:

$$\mu = P(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1}|(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1} \cup \overline{\mathbf{X}_{s+t-1}} = \overline{\mathbf{x}_{s+t-1}}) \cap J) =$$

$$\frac{P(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1}|J)}{P(\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1}|J) + P(\overline{\mathbf{X}_{s+t-1}} = \overline{\mathbf{x}_{s+t-1}}|J)} =$$

$$\frac{\frac{1}{C(s:t)}}{\frac{1}{C(s:t)} + \frac{1}{C(s:t)}} = \frac{1}{2} \square$$

Härmed har vi visat att parametern μ är ekvivalent med $\frac{1}{2}$ för alla sekvenser inom jämna matcher då samtliga set innan det sista är förbestämda. Vi kommer dock att behöva undersöka sekvenser som inte har samtliga set förbestämda. Dessa är sekvenserna av typ AAB...A/BBA...A, och avser matcher på elitnivå. Om man skall vara korrekt, skall man införa en egen parameter för sekvenser av denna typ. Det kommer dock att visa sig att denna parameter kommer att anta exakt likadant värde som den som definierades ovan, och dessutom så har de båda samma betydelse i vår modell.

Följdsats. Antag att μ avser nu både sekvenser där samtliga set är förbestämda, och sekvenser av typ AAB...A/BBA...A. Det gäller då fortfarande att

$$\mu = \frac{1}{2}$$

Bevis. När det gäller sekvenserna av typ AAB...A/BBA...A, kommer μ att innebära

$$\mu = P(\mathbf{X}_3 = \mathbf{x}_3, \bullet\bullet\bullet | (\mathbf{X}_3 = \mathbf{x}_3, \bullet\bullet\bullet \cup \overline{\mathbf{X}_3} = \overline{\mathbf{x}_3}, \bullet\bullet\bullet) \cap J)$$

Notera vidare att $P(\mathbf{X}_3 = \mathbf{x}_3, \bullet\bullet\bullet | J) = P(\overline{\mathbf{X}_3} = \overline{\mathbf{x}_3}, \bullet\bullet\bullet | J) = \frac{3}{C(4:3)}$. Genom att använda samma teknik som i bevis för Sats 3, får man att $\mu = \frac{1}{2}$ $\square$

Så länge har vi konstruerat vår undersökning på antagandet att några inlärningseffekter inte äger rum. Sannolikheten μ är i sådant fall ekvivalent med *0.5*, och denna skattning är baserad på antaganden och resultaten av Bernoulli-modellen. Antag nu att vi har observerat inlärningseffekter i något/några av våra datamaterial. Vad skulle detta innebära? En direkt följd av detta är att om man skattar μ ur dessa datamaterial, kommer skattningen att signifikant avvika från *0.5*. En sådan avvikelse skulle i sin tur tyda på att någon eller några av modellantaganden inte håller, och den måste följaktligen överges.

Vi kommer att konstruera vår dataundersökning på följande sätt: vi kommer att ha nollhypotesen $\mu = 0.5$, vilket i sin tur innebär att modellen håller och några inlärningseffekter inte äger rum. Alternativhypotesen blir att $\mu \neq 0.5$, vars direkta följd är att modellen inte håller och inlärningseffekter (eller skärpningseffekter) verkligen äger rum.

Vi fortsätter konstrueringen av modellen genom att definiera

$$Y = \sum_{i=1}^{n} (\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1} | (\mathbf{X}_{s+t-1} = \mathbf{x}_{s+t-1} \cup \overline{\mathbf{X}_{s+t-1}} = \overline{\mathbf{x}_{s+t-1}}) \cap J)_i$$

Här betecknar n det totala antalet observerade sekvenser av typ $\mathbf{x}_{s+t-1}$ och $\overline{\mathbf{x}_{s+t-1}}$. I fall vi undersöker förekomsten av sekvenserna AAB...A/BBA...A, kommer uttrycket för Y innehålla 3 förbestämda set, och 3 "valfria" set, som vi inte är intresserade i. Parametern n blir i detta fall lika med summan av samtliga observerade sekvenser som börjar med AAB eller BBA.

I båda dessa fall, blir $Y \sim Bin(n, \mu)$, eftersom matcherna anses vara oberoende. Låt vidare y stå för antal sekvenser av typ $\mathbf{x}_{s+t-1}$ (eller $\mathbf{x}_3$). I fall y avviker tillräckligt mycket från $E[Y] = \frac{n}{2}$, kommer det att innebära att inlärnings- eller skärpningseffekter förekommer. Vi testar detta genom att räkna ut p-värden för samtliga datamaterial, dvs. sannolikheter att observera ännu extremare data än det har observerats. Vi illustrerar detta med ett exempel.

Exempel. I turneringen Rookie-Cupen förekommer det 20 sekvenser av typ AABBA, och 26 sekvenser av typ AABBB. I detta fall blir $y = 20$, och $n = 20 + 26 = 46$. Sannolikheten att observera ännu extremare data blir i detta fall lika med $P(Y \leq 20)$. Denna sannolikhet är ett ensidigt p-värde, dvs. detta tar hänsyn enbart till eventuella avvikelser med avseende på små y. Vi kunde dock ha observerat stora y också, och därför bör p-värdet inkludera den spegelvända sannolikheten med avseende på $E[Y] = \frac{n}{2}$, eller, med andra ord, $P(Y \geq \frac{n}{2} +$

Typ av tävling	Sekvens	Antal sekvenser	$\widehat{\mu}$ (slh. för ...A)	Testets p-värde
Rookie-Cupen (3-2)	AABB.	46	0.43	0.46
Folksam-Talangen (3-2)	AABB.	39	0.31	0.024
Barn-Samtliga (3-2)	AABB.	85	0.37	0.03
Elitnivå (4-3)	AAABBB.	8	0.38	0.72
Elitnivå (4-3)	AAB....	22	0.41	0.52

Tabell 4: Resultatena av dataundersökning: Inlärningseffekterna inom jämna matcher

$|\frac{n}{2} - y|) = P(Y \geq 26)$. Eftersom binomialfördelning är symmetrisk då dess sannolikhetsparameter är ekvivalent med 0.5, kommer de båda p-värdena ovan att vara lika. Det tvåsidiga p-värdet, som vi är intresserade i, kommer följaktligen att fås som $2P(Y \leq 20) = 2P(Y \geq 26)$.

Det dubbelsidiga p-värdet räknar man alltså ut som $2P(Y \geq \frac{n}{2} + |\frac{n}{2} - y|)$ eller $2P(Y \leq \frac{n}{2} - |\frac{n}{2} - y|)$. Resultatena av undersökningen hittar man i tabell 4. Skattningarna av μ fick man helt enkelt genom att dividera y med n (detta är en ML-skattning av sannolikhetsparametern för Binomialfördelade data).

Det verkar vara uppenbart att det låga p-värdet för samtliga data avseende barnnivå beror på ett ännu lägre p-värde avseende datamaterialet ”Folksam-Talangen”. Den sista frågan som bör besvaras här är vad det låga p-värdet egentligen innebär. Enligt tabellen i Appendix B förekommer sekvensen AABBB 27 gånger, och sekvensen AABBA – enbart 12 gånger. En sådan fördelning av data innebär att de spelarna som förlorade två set först, men lyckades utjämna till 2-2 i turneringen Folksam-Talangen, vann sista setet och därmed hela matchen mer än dubbelt så ofta än deras motståndare. Alltså är det låga p-värdet ett tecken på att det förekommer inlärningseffekter i turneringen Folksam-Talangen.

3.2 Markov-modellen

I det föregående kapitlet har vi baserat den statistiska analysen på det antagandet att det inte finns något beroende mellan olika set. Nu ska dataundersökningen gå till en bit annorlunda. Vi kommer nämligen att konstruera en modell med en Markovkedja som bas. Som i Bernoulli-modellen, är det en sannolikhetsparameter som kommer att skattas. Denna parameter kommer dock inte att innebära en sannolikhet för att vinna ett set i en bordtennismatch, utan en sannolikhet att givet att ett set har vunnits av en av spelarna, kommer även nästa set att vinnas av denna individ. Analys av data kommer att bestå av två delar: först ska vi härleda en teoretisk skattning av denna parameter, och därefter undersöka om denna skattning är kompatibel med empiriska data.

Endast jämna matcher kommer att undersökas med Markov-modellen. En anledning till detta är att modellen blir betydligt krångligare om man undersöker samtliga matcher vi har, men huvudskälet är att det blir mycket svårare att tolka resultaten av undersökningen av icke-jämna matcher, än resultaten avseende jämna matcher.

Vi kommer inte att gå in i detaljer angående definition och egenskaper av Markovkedjor, men en intresserad läsare hänvisas till [Ross,2006].

3.2.1 Beskrivning av modellen

Vi börjar med att definiera några viktiga begrepp som Markov-modellen kommer att bygga på:

Definition. Antag att två individer spelar en bordtennismatch. Vi kallar en av individerna för spelare 0 och den andra för spelare 1. Låt X_i beteckna nummer på spelare som vinner set i i denna match. Slutligen låter vi

$$P(X_{i-1} = X_i) = p\,; \quad P(X_{i-1} \neq X_i) = 1 - p\,; \quad \pi = P(X_1 = 1)\,.$$

Parametern p står alltså för sannolikhet att givet att ett set har vunnits av någon av matchens deltagare (0 eller 1), kommer denna person även att vinna nästa set. Notera att matchen kan vinnas av både spelare 0 och 1, så det antagandet att matchens sista set måste vinnas av en viss spelare (som vi gjorde i det föregående kapitlet) är inte längre aktuellt. Detta leder även till att vi måste på något sätt binda ihop spelarnas nummer (0 och 1), och begreppen seger och nederlag i matchen. Om exempelvis spelare 1 vinner ett set, innebär det ju samtidigt att spelare 0 har förlorat det, och frågan vem som är egentligen setets vinnare inte blir besvarad. Parametern π är skapad för att lösa detta problem - den står för sannolikheten att det första setet kommer att vinnas av spelare 1. I fall π är ekvivalent med 0 eller 1, kan man se detta ungefär som setbetingningen i Bernoulli-modellen. Skillnaden blir att betingningen görs med avseende på det första setet i stället för det sista.

Notera att eftersom parametern π inte är beroende av p, är det tillåtet att välja den fritt. Konstruering av modellen förenklas betydligt då vi låter de båda spelarna vinna matcher med en och samma sannolikhet, och därför bestämmer vi oss att sätta $\pi = 1 - \pi = 0.5$. En följd av detta blir att vi inte behöver ta hänsyn till värden som X_i kommer att anta i formlerna ovan, eftersom $X_{i-1} = X_i$ kan innebära både två segrar eller två nederlag i rad med lika stor sannolikhet. Vi fortsätter med att definiera Markovkedjemodellen:

Definition. Antag att π är tilldelad värde $\frac{1}{2}$ enligt ovan. Markovkedjemodellen för jämna bordtennismatcher definieras som

$$(\frac{1}{2}\,\frac{1}{2})\begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}$$

, där vektorn till vänster motsvarar startsannolikheter i en match, och matrisen till höger innebär en Markovkedja där den lodräta riktningen motsvarar utfallet av set *i-1*, och den horisontella riktningen innebär den betingade sannolikheten för resultatet av set i.

Exempel. Antag att en bordtennismatch spelades enligt följande sekvens: 1101. Sannolikheten för ett sådant resultat blir enligt modellen $P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1) = \frac{1}{2}p(1-p)^2$. På motsvarande sätt blir sannolikheten för sekvens 111 $P(X_1 = 1, X_2 = 1, X_3 = 1) = \frac{1}{2}p^2$.

Denna modell är allmän och är tillämpbar för alla möjliga typer av matcher. Vi är dock endast intresserade i jämna matcher, och därför bör vi exkludera komplement till dessa (dvs. samtliga icke-jämna matcher) ur fortsatt analys. Vi börjar med att definiera några viktiga parametrar:

Definition. Antag att *s*, *t*, och *C(s:t)* har samma betydelse som i det föregående kapitlet. Låt vidare $D(s:t) = 2C(s:t)$, vilket står för antal möjliga sekvenser i en match i Markov-modellen. Detta följer av att vi inte gör en betingning att en match skall vinnas av en viss spelare, vilket fördubblar det totala antalet möjliga sekvenser. Slutligen definierar vi

$$g_i(p) = P(X_1 = x_1, X_2 = x_2, ..., X_s = x_s), \quad i = 1...D(s:s-1),$$

$$f_i(p) = P(X_1 = x_1, X_2 = x_2, ..., X_s = x_s | t = s-1), \quad i = 1...D(s:s-1).$$

Det översta utrycket står för den allmänna sannolikheten att en viss sekvens med $t = s - 1$ dyker upp. Det nedersta uttrycket innebär sannolikheten för en viss sekvens givet att en match är jämn. Vi fortsätter med att konstatera följande:

Lemma.

$$f_i(p) = \frac{g_i(p)}{\sum_{j=1}^{D(s:s-1)} g_j(p)}$$

Bevis. Det totala sannolikheten att en match blir jämn är $P(s = t-1) = \sum_{j=1}^{D(s:s-1)} g_j(p)$. Enligt formeln för betingad sannolikhet får man att

$$f_i(p) = \frac{P(X_1 = x_1, X_2 = x_2, ..., X_s = x_s \cap t = s-1)}{P(s = t-1)} = \frac{g_i(p)}{\sum_{j=1}^{D(s:s-1)} g_j(p)} \square$$

Vår nästa uppgift är att hitta ett uttryck som skulle skatta p med avseende på faktiska data. En sats som ska illustreras nedan kommer att vara till hjälp för detta. Innan det behöver vi dock att införa två ytterligare begrepp.

Definition. Låt o_i stå för antal förekomster av sekvens i, $i \in 1...D(s:s-1)$. Låt vidare n beteckna det totala antalet bordtennismatcher i det datamaterialet som skall undersökas.

Vi inför slutligen beteckning $L(p) = p(o_1...o_{D(s:s-1)}|p)$ för likelihood-funktionen av p.

Sats 4.

$$L(p) \propto \prod_{i=1}^{D(s:s-1)} f_i^{o_i}(p) \tag{2}$$

Bevis. Eftersom matcherna förutsätts vara sinsemellan oberoende, och

$$\sum_{i=1}^{D(s:s-1)} f_i(p) = 1$$

, har de observerade antalen av de möjliga sekvenserna en multinomialfördelning. Fördelningsfunktionen för detta är

$$L(p) = \frac{n!}{o_1!o_2!...o_{D(s:s-1)}!} \prod_{i=1}^{D(s:s-1)} f_i^{o_i}(p).$$

Den delen som inte innehåller p betraktas som konstant med avseende på p. Detta resulterar i (2)$\square$

Parametern av intresse skattar vi genom att tillämpa maximum likelihood principen på (2). Man skall alltså hitta ett sådant $\hat{p}$ som maximerar $L(p)$, eller, med annan notation, vi söker $\hat{p}$ =argmax$_p L(p)$. Algoritmen för detta är välkänd: man deriverar log-likelihooden med avseende på p, sätter detta lika med noll och löser slutligen den deriverade log-likelihood ekvationen:

$$\frac{d}{dp} \log L(p) = 0 \quad \Longleftrightarrow \quad \frac{d}{dp} \log \prod_{i=1}^{D(s:s-1)} f_i^{o_i}(p) = 0.$$

En detaljerad härledning av ML-skattningen samt uträkningarna för $\hat{p}$ avseende samtliga datamaterial illustreras i Appendix A. Här hoppar vi över detta och

Datamaterial	$\hat{p}$
Rookie-Cupen	0.5102
Folksam-Talangen	0.4966
Samtliga på barnnivå	0.5041
Elitnivå	0.4897

Tabell 5: Resultatena av dataundersökning: ML-skattningarna av p

går direkt över till illustrering av skattningarna av p och fortsätter med den statistiska undersökningen.

Slutligen noterar vi att i fall några avvikelser i data inte förekommer, kommer samtliga o_i att vara ungefär lika med varandra. Detta kommer i sin tur innebära att (sträck ovan står för medelvärden):

$$o_i \to \bar{o} \ \Rightarrow \ f_i(p) \to \bar{f}(p) \ \Rightarrow \ g_i(p) \to \bar{g}(p) \ \Rightarrow \ p \to 0.5$$

Med andra ord, kommer p att närma sig 0.5 då alla sekvenser förekommer lika ofta, vilket skulle alltså innebära att det inte finns någon avvikelse från vår modell.

3.2.2 Resultatena och undersökning av eventuell avvikelse från modellen

Lösningarna av ML-ekvationer för jämna matcher ur samtliga datamaterial illustreras i tabell 5.

Vi ser direkt att samtliga $\hat{p}$ ligger väldigt nära *0.5*, och har därför inga förväntningar om att avvikelsen från modellen kommer att vara betydlig. Att se det är dock inte tillräckligt, utan vi måste även undersöka om avvikelsena inte är signifikanta med hjälp av ett statistiskt test. En kärna för detta test är en sats som följer nedan:

Sats 5. Antag att vi vill testa ett hypotes att $p = p_0$, givet att den fullständiga modellen förutsätter $p = \hat{p}$. Deviansen

$$\Lambda = -2 \log \frac{L(p_0)}{L(\hat{p})}$$

är då approximativt χ^2 fördelad med k frihetsgrader, där k står för antal parametrar som tilldelas specifika värden under noll-hypotesen.

Bevis. Se [Lindgren,1993].

Vi noterar vidare att $L(p)$ är enbart baserad på en enda parameter, så vår hypotes blir simpel. Med andra ord, är det endast en (och den enda) parameter som blir tilldelad ett specifikt värde, så en följd av detta blir naturligtvis att

Datamaterial	95% konfidensintervall för p
Rookie-Cupen	[0.466 ; 0.560]
Folksam-Talangen	[0.444 ; 0.548]
Samtliga på barnnivå	[0.470 ; 0.540]
Elitnivå	[0.438 ; 0.539]

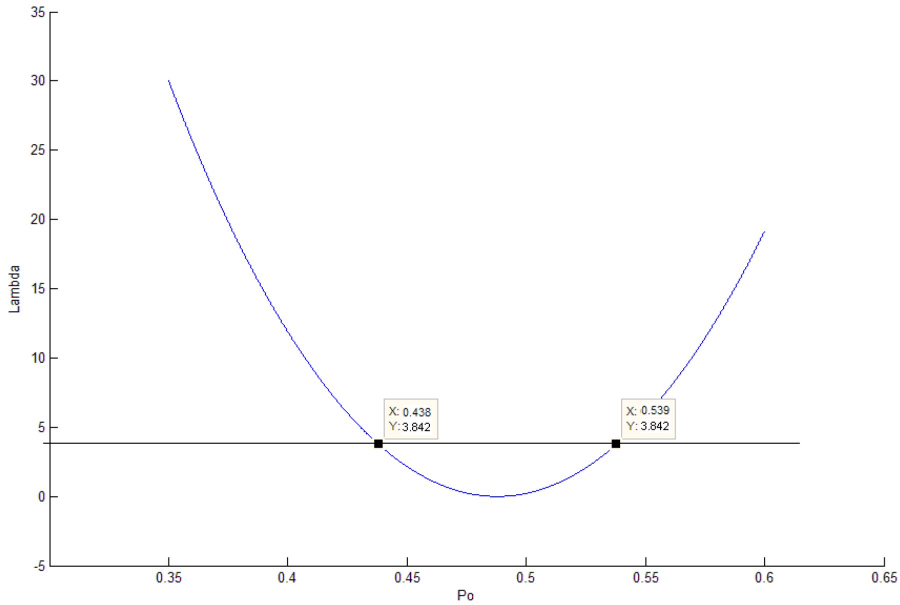
Tabell 6: Resultatena av dataundersökning: 95% konfidensintervall för p

$k = 1$. Givet Sats 5, kan den statistiska signifikansen av avvikelser i data testas med ett antal olika metoder. Här väljer vi att implementera detta genom att konstruera konfidensintervall för p med signifikansnivå $\alpha = 5\%$. Konfidensintervall kommer konstrueras på ett sådant sätt att dessa ska innehålla samtliga p_0 som uppfyller $\arg_{p_0} \Lambda \leq \chi^2_{1-\frac{\alpha}{2}}(1)$. Vi vet även att en χ^2 fördelad stokastisk variabel med en frihetsgrad är inget annat än en standard normalfördelad variabel i kvadrat (Se exempelvis [Gut,1995]). Konfidensintervallen kommer alltså bestå av p_0 sådana att

$$\Lambda \leq (inv\,\Phi(1 - \frac{\alpha}{2}))^2 \quad \Leftrightarrow \quad \Lambda \leq 3.8416 = 1.96^2, \quad \alpha = 0.05$$

Man ska alltså undersöka Λ som en funktion av p_0. Konfidensintervall för samtliga datamaterial visas i tabell 6.

Vi visar slutligen en figur som illustrerar sambandet mellan p_0 och Λ. Sannolikhetsparametern illustreras på den horisontella alxeln, och deviansen på den lodräta axeln. Denna figur avser matcher på elitnivå. (Se Figur 1)



Figur 1: Samband mellan p_0 och Λ

4 Diskussion och slutsatser

4.1 Slutsatser

Målet med detta arbete var att undersöka om det existerar några systematiska tendenser inom bordtennismatcher, och i fall det var så, skulle vi ta reda på karaktär av dessa. Data som arbetet byggde på avsåg ett antal matcher som spelades av individer ur olika åldersklasser och spelnivåer. Med hjälp av varierande statistiska verktyg kunde vi undersöka om den faktiska datamaterialens struktur inte följde vår hypotes om avsaknad av systematiska trend inom bordtennismatcher. Dessa verktyg grundade sig på olika antaganden och var väldigt skilda varann.

De jämna matcherna, som var av störst intresse för oss, har vi undersökt med hjälp av tre olika metoder. Dessa var:

- Undersökning av förekomsten av olika sekvenser inom bordtennismatcher (Bernoulli-modellen)
- Undersökning av fördelning av vunna set (Bernoulli-modellen)
- Dataundersökning där man inte har gjort ett antagande om oberoende mellan olika set (Markov-modellen)

De matcherna som inte var jämna undersökte vi med enbart med avseende på förekomsten av olika sekvenser.

Medan vi genomförde vår undersökning, kunde en intressant trend noteras, nämligen, att data avseende jämna matcher i turneringen Folksam-Talangen inte hade en jämn spridning, utan de sista seten vanns mycket oftare av hela matchens vinnare än de första. Ett test av detta datamaterial med avseende på fördelning av vunna set resulterade i ett p-värde = *0.07*. Ett sådant lågt p-värde betydde att det möjligtvis finns en tendens i data som vi inte har undersökt, och därför gjorde vi en ytterligare undersökning av jämna matcher. Denna undersökning grundade sig på att analysera förekomsten av inlärnings / skärpnings effekter inom bordtennis, och resultatet av detta blev att hypotesen om att sådana effekter inte förekommer förkastades på 5% signifikansnivå.

Detta resultat blir ännu tydligare om man tittar på den råa data. Nämligen, sekvenserna av typ AABBB förekom mer än dubbelt så ofta (27 vs. 12 gånger) än sekvenserna av typ AABBA avseende Folksam-Talangen - data (se Appendix B).

Det ska även noteras att inlärningseffekts-trenden kunde observeras i samtliga jämna matcher, både på barn- och elitnivå (se tabell 2). Dessa trend blev dock inte statistiskt signifikanta (exklusive Folksam-Talangen data) vilket kan möjligtvis bero på ett relativt lågt antal observationer, vars följd blev att sannolikheter att observera sådana fördelningar av data fortfarande var betydligt stora.

Resultat av samtliga statistiska test förutom denna speciella undersökning visar dock att det inte finns någon statistiskt signifikant avvikelse från våra modeller.

När det gäller Bernoulli-modellen, har det visat sig att samtliga datamaterial (förutom det ovannämnda) följer våra antaganden nästan perfekt, vilket stödjs först och främst av de höga p-värdena i resultattabeller.

Denna tendens (som innebär att det inte finns någon tendens, dvs. att data följer vår modell) blir ännu tydligare då man tittar på resultaterna av Markovundersökningen: hypotesvärdet *0.5* ligger bekvämt i samtliga konfidensintervall för sannolikheten p.

4.2 Diskussion

Slutsatserna som vi drog efter vi hade undersökt vår data känns naturliga och förväntade. I fall en av deltagarna i en bordtennismatch är betydligt duktigare än den andra, kommer det troligen att resultera i en 3-0 / 4-0 match, som vi inte kan och inte skulle undersöka här. Om en match är av typ 3-1; 4-1 eller 4-2, är det mycket mer troligt att spelarna åtminstone ligger på samma nivå. Slutligen, om en match är jämn, finns det stora anledningar att tro att spelarna har ungefär samma spelerfarenhet och är ungefär lika duktiga. Dessa påståenden avspeglas även av Sats 2.

Att spelarna är lika duktiga innebär i sin tur att seten i en match mellan dessa kommer snarast att vinnas av slump men ej på grund av någon systematisk tendens.

Med andra ord, även om någon sådan tendens verkligen äger rum, behöver den inte gälla för alla matcher eller deras avgörande majoritet, utan snarare för någon del av dessa. En sådan tendens kommer alltså att döljas på grund av ett stort antal helt slumpmässiga matcher.

Trots det så har vi kunnat visa att det föreligger en systematisk trend i datamaterialet Folksam-Talangen. Med hjälp av en undersökning som skulle testa just denna tendens har vi kunnat påvisa att givet att en match började med att en av spelarna (spelare A) vann 2 set i rad, men B lyckades utjämna till 2-2, är sannolikheten mycket stor att B kommer att vinna denna match. Detta beror möjligtvis på att B lärde sig sin motståndare under de första seten, och lyckades följaktligen vinna matchen trots underläge 0-2.

Det ska dock noteras att vi inte har observerat någon liknande tendens inom turneringen Rookie-Cupen, som också är en barntävling. Detta kan möjligtvis bero på att Rookie-Cupen - deltagare är i regel äldre än deras kollegor från Folksam -Talangen. Ungdomarna, som deltar i Rookie-Cupen har troligtvis mycket mer spelerfarenhet, och är möjligtvis bättre på att anpassa sig till varierande situationer inom bordtennismatcher.

Vi har inte heller observerat någon avvikelse från modeller avseende matcher på elitnivå. En av möjliga förklaringar till detta faktum är att i fall en individ är en bordtennisspelare på elitnivå, och är tillräckligt bra för att kunna vinna två eller tre set i en match mot en annan elitspelare, innebär det nog automatiskt att denna person har betydlig erfarenhet inom bordtennis och kan anpassa sig till sin motståndare. Så i fall två bordtennisspelare på elitnivå möts i en bordtennismatch, och en av de är bättre än den andra, kommer den som är bättre att vinna matchen med 4-0 / 4-1. I fall motståndarna är jämna, beror nog utfall av en match mellan dessa på en ren slump.

Slutligen ska det också konstateras att inlärningseffekts-trenden har vi observerat inte enbart för Folksam-Talangen data, utan för samtliga turneringar, både på barn- och elitnivå. Tabell 2 illustrerar detta tydligt, nämligen parametrarna $\hat{\mu}$ avseende samtliga matcher är mycket lägre än det förväntade värdet *0.5*. Vi har inte tagit upp detta som ett resultat eftersom dessa trend inte blev statistiskt signifikanta, vilket innebär att vi inte kan göra några konkreta påståenden angående dessa. Möjligtvis beror dessa höga p-värden på att det inte förekommer några inlärningseffekter i respektive dataset. Men det finns också en sannolikhet att avsaknad av statistisk signifikans av en sådan trend hos dessa datamaterial beror på att antalet observationer för dessa var relativt lågt. Det gäller speciellt data för elitnivå.

I fall en ny undersökning skulle göras avseende detta ämne, skulle första rekommendationen vara att använda ett större datamaterial, om det är möjligt. Man skulle även kunna tillämpa andra, mer avancerade undersökningsmetoder, till exempel skapa en generaliserad linjär modell för data, eller förstora det ursprungliga datamaterialet med hjälp av t.ex. bootstrap-simuleringarna. Tillämpning av dessa metoder kommer dock inte nödvändigtvis att ge ett resultat som skulle avvika från ett som kommer ur metoder presenterade i denna uppsats.

A Härlendning av ML-skattningar

Härledning och uträkning av ML-skattningar för p *i Markov-modellen.*

Vi söker lösning till ML-ekvationen $\frac{d}{dp}\log\prod_{i=1}^{D(s:s-1)} f_i^{o_i}(p) = 0$.

Vi omformar vänster ledet av denna ekvation genom att gå över till $g_i(p)$:

$$\frac{d}{dp}\log\prod_{i=1}^{D(s:s-1)} f_i^{o_i}(p) =$$

$$\frac{d}{dp}\sum_{i=1}^{D(s:s-1)} o_i \log f_i(p) = \frac{d}{dp}\sum_{i=1}^{D(s:s-1)} o_i \log(\frac{g_i(p)}{\sum_{j=1}^{D(s:s-1)} g_j(p)}) =$$

$$\frac{d}{dp}(\sum_{i=1}^{D(s:s-1)} o_i \log g_i(p) - \frac{d}{dp}\sum_{i=1}^{D(s:s-1)} o_i \log \sum_{j=1}^{D(s:s-1)} g_i(p)) =$$

$$\frac{d}{dp}(\sum_{i=1}^{D(s:s-1)} o_i \log g_i(p) - n \log \sum_{j=1}^{D(s:s-1)} g_i(p)).$$

Vi separerar uttrycket inom parentes i två delar med avseende på minustecknet, och deriverar sedan de nya uttrycken :

$$\frac{d}{dp}(\sum_{i=1}^{D(s:s-1)} o_i \log g_i(p) - n \log \sum_{j=1}^{D(s:s-1)} g_i(p)) =$$

$$\sum_{i=1}^{D(s:s-1)} \frac{o_i}{g_i(p)}(\frac{d}{dp}g_i(p)) - n\frac{1}{\sum_{j=1}^{D(s:s-1)} g_j(p)}(\frac{d}{dp}\sum_{j=1}^{D(s:s-1)} g_j(p)).$$

Uttrycket till höger av minustecknet behöver ej att förenklas ytterligare, eftersom i denna form är det enklast att arbeta med. Vi kommer dock att skriva om det andra uttrycket, vilket kommer att göra det ännu mer simplifierat.

Varje funktion $g_i(p)$ skrivs på form $g_i(p) = \frac{1}{2}p^{u_i}(1-p)^{v_i}$, där parametrarna u och v står för antal upprepningar av p respektive *(1-p)* inom en viss sekvens, eller, med andra ord, antal övergångar av typ $X_{i-1} = X_i$, respektive $X_{i-1} \neq X_i$.

Notera även att $2s-2=u+v$, och $u_i, v_i \geq 0$. Vi använder denna beteckning i den vänstra delen av uttrycket ovan, och förenklar det :

$$\sum_{i=1}^{D(s:s-1)} \frac{o_i}{\frac{1}{2}p^{u_i}(1-p)^{v_i}}(\frac{d}{dp}\frac{1}{2}p^{u_i}(1-p)^{v_i}) =$$

$$\sum_{i=1}^{D(s:s-1)} \frac{\frac{1}{2}o_i p^{u_i-1}(1-p)^{v_i-1}(u_i(1-p)-v_i p)}{\frac{1}{2}p^{u_i-1}(1-p)^{v_i-1}} =$$

$$\sum_{i=1}^{D(s:s-1)} \frac{o_i(u_i(1-p)-v_i p)}{p(1-p)} =$$

$$\frac{1}{p(1-p)}\sum_{i=1}^{D(s:s-1)} o_i(u_i(1-p)-v_i p) = [v_i = 2(s-1)-u_i] =$$

$$\frac{1}{p(1-p)}\sum_{i=1}^{D(s:s-1)} o_i(u_i - 2sp + 2p) = ... = \frac{\sum_{i=1}^{D(s:s-1)} o_i u_i - 2np(s-1)}{p(1-p)}.$$

Skattningen av p fås således genom att lösa ekvationen

$$\frac{A-2np(s-1)}{p(1-p)} - n\frac{B'}{B} = 0; \quad A = \sum_{i=1}^{D(s:s-1)} o_i u_i; \quad B = \sum_{i=1}^{D(s:s-1)} g_i(p).$$

Notera att B är ej beroende av o_i, vilket innebär att det inte förändras givet att s är samma. Parametern A är dock direkt beroende på vår empiriska data – med andra ord, på o_i. Samtliga värden av o_i och u_i kan en intresserad läsare hitta i Appendix B.

Nästa steg är att räkna ut samtliga A och B. Observera att 3 av våra datamaterial är av samma typ, så vi kommer att endast ha 2 värden för parametern B. Vi börjar med att få fram B för båda typer av matcher. Först räknar vi ut detta för matcher på barnnivå: $\sum_{i=1}^{D(s:s-1)} g_i(p) = 2(1-p)^4 + 4p(1-p)^3 + 4p^2(1-p)^2 + 2p^3(1-p) = ... = 2(1-p)(p^2-p+1)$. Derivata av B med avseende på p blir då $-2(3p^2-4p+2)$, och kvoten av B' och B blir:

$$\frac{B'}{B} = (-1)\frac{3p^2-4p+2}{(1-p)(p^2-p+1)}.$$

Motsvarande kvot för 4:3 matcher räknas ut på ett liknande sätt: $\sum_{i=1}^{D(s:s-1)} g_i(p) = 2(1-p)^6 + 6p(1-p)^5 + 12p^2(1-p)^4 + 12p^3(1-p)^3 + 6p^4(1-p)^2 + 2p^5(1-p) =$ ...$= 2(1-p)(2p^4 - 4p^3 + 4p^2 - 2p + 1)$. Derivator av denna polynom blir $(-2)(10p^4 - 24p^3 + 24p^2 - 12p + 3)$, vilket innebär att kvoten av *B'* och B är ekvivalent med:

$$\frac{B'}{B} = \frac{10p^4 - 24p^3 + 24p^2 - 12p + 3}{(1-p)(2p^4 - 4p^3 + 4p^2 - 2p + 1)}.$$

Uträkning av skattningarna av p kräver att man löser ett antal av 3- respektive en 5-gradsekvation. Detta gjordes med hjälp av speciella ekvationslösare, som finns på internet (Se Referenser för detaljer). Vi fortsätter med att skatta p:

Rookie-Cupen

Parametern A för detta datamaterial beräknades vara lika med *226*. Vår ML-ekvation blir då följande ($n = 147$):

$$\frac{226 - 588p}{p(1-p)} + 147\frac{3p^2 - 4p + 2}{(1-p)(p^2 - p + 1)} = 0$$

Förenkling av detta gav oss följande 3-gradsekvation :

$$-147p^3 + 226p^2 - 520p + 226 = 0$$

Det visade sig att denna ekvation har endast en reell lösning. Denna är lika med $\hat{p} = 0.5102$

Folksam-Talangen

Här är parametern A lika med *177*, och det totala antalet observationer är lika med *119*. Vi fick att:

$$\frac{177 - 476p}{p(1-p)} + 119\frac{3p^2 - 4p + 2}{(1-p)(p^2 - p + 1)} = 0 \quad \Leftrightarrow$$

$$-119p^3 + 177p^2 - 415p + 177 = 0$$

Den enda reella roten av denna ekvation är $\hat{p} = 0.4966$

Samtliga matcher på barnnivå

Det totala antalet matcher är här *266*, och $A = 403$. ML-ekvation blev

$$\frac{403 - 1064p}{p(1-p)} + 266\frac{3p^2 - 4p + 2}{(1-p)(p^2 - p + 1)} = 0 \quad \Leftrightarrow$$

$$-266p^3 + 403p^2 - 935p + 403 = 0$$

Precis som vi har förväntat, har denna ekvation endast en reell lösning, och den är lika med $\hat{p} = 0.5041$

Matcher på elitnivå

Vår parameter A är här lika med *183*, och $n = 75$. Denna typ av tävling har ett annat s – värde ($s = 4$) än de föregående , vilket innebär att vi ska använda det andra uttrycket för $\frac{B'}{B}$. ML-ekvationen blir:

$$\frac{183 - 450p}{p(1-p)} + 75\frac{10p^4 - 24p^3 + 24p^2 - 12p + 3}{(1-p)(2p^4 - 4p^3 + 4p^2 - 2p + 1)} = 0$$

Efter ett antal omskrivningar och förenklingar resulterar denna i följande 5-gradsekvation : $-150p^5 + 366p^4 - 732p^3 + 732p^2 - 591p + 183 = 0$. Det visade sig att denna ekvation har endast en reell rot, som är lika med $\hat{p} = 0.4897$.

B Sekvenser av jämna matcher

Samtliga sekvenser avseende 4:3 samt 3:2 matcher, samt respektive u_i[1] *och antal förekomster* (o_i).

Tabell 6: 3-2 matcher

Typ av sekvens	u	Antal förekomster		
		Rookie-Cupen	Folksam-Talangen	Samtliga
1 1 0 0 1	2	20	12	32
1 0 1 0 1	0	24	24	48
1 0 0 1 1	2	31	16	47
0 1 1 0 1	1	22	18	40
0 1 0 1 1	1	24	22	46
0 0 1 1 1	3	26	27	53

Tabell 7: 4-3 matcher

Typ av sekvens	u	Antal f.	Typ av sekvens (forts.)	u	Antal f.
1 1 1 0 0 0 1	4	3	0 1 1 1 0 0 1	3	8
1 1 0 1 0 0 1	2	5	0 1 1 0 1 0 1	1	5
1 1 0 0 1 0 1	2	1	0 1 1 0 0 1 1	3	2
1 1 0 0 0 1 1	4	3	0 1 0 0 1 1 1	3	1
1 0 0 0 1 1 1	4	2	0 1 0 1 0 1 1	1	4
1 0 0 1 0 1 1	2	5	0 1 0 1 1 0 1	1	5
1 0 0 1 1 0 1	2	2	0 0 0 1 1 1 1	5	5
1 0 1 1 0 0 1	2	3	0 0 1 0 1 1 1	3	5
1 0 1 0 1 0 1	0	4	0 0 1 1 0 1 1	3	3
1 0 1 0 0 1 1	2	4	0 0 1 1 1 0 1	3	5

OBS!: Sekvenserna i tabeller ovan är betingade på att spelare 1 vinner sista set. Det påverkar dock inte resultaten av Markovmodellen, eftersom antalet övergångar av typ ($0 \to 1$ / $1 \to 0$)(respektive $0 \to 0$ / $1 \to 1$) i en match förblir samma oavsett om man gör denna betingning eller icke.

[1]Se Appendix A för definition av u_i

Referenser

[Gut,1995] GUT A. (1995), An Intermediate Course in Probability, Second Edition, Springer.

[Lindgren,1993] LINDGREN B., (1993), Statistical theory, Fourth Edition, Chapman & Hall.

[Ross,2006] ROSS S. (2006), Introduction to Probability Models, Ninth Edition, Academic Press.

[Snedecor and Cochran,1989] SNEDECOR G. and COCHRAN W., (1989), Statistical Methods, Eighth Edition, Iowa State University Press.

[1] Quintic Equation Solver, http://www.freewebs.com/brianjs/quinticequationcalculator.htm

[2] Cubic Equation Solver, http://www.freewebs.com/brianjs/ultimateequationsolver.htm

[3] Chi Square Calculator, http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

[4] Svenska Bordtennisförbundets hemsida, www.svenskbordtennis.se