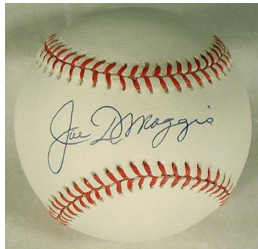


Chasing DiMaggio: Streaks in Simulated Seasons Using Non-Constant At-Bats



Bachelor-Seminar: Statistik im Sport

Betreuer: Sebastian Kaiser
Dozent: Prof. Dr. Friedrich Leisch

1. Die Sportart Baseball

Geschichte:

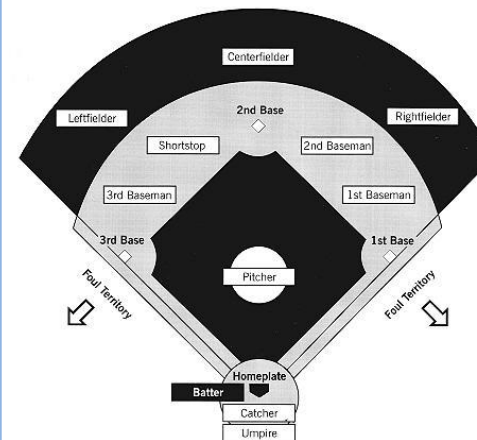
- 1744: Erstes Spiel mit Name „base ball“
- 1845: Gründung des ersten Baseballvereins
- traditionsreichste Sportart in den USA
- Baseball auch in Ostasien und der Karibik
- Baseball-Ligen in Deutschland, Österreich und Schweiz

Inhalt

1. Die Sportart Baseball
2. Wichtige Statistiken im Baseball
3. Statistische Methodik
4. Der Artikel von Rockoff und Yates
5. Fazit
6. Weitere mögliche Statistik
7. Abschließende Bemerkungen

1. Die Sportart Baseball

Das Spielfeld:



- Spielfeldform: Viertelkreis
- Zwei Mannschaften treten gegeneinander an
- Ein Team besteht aus neun Spielern
- Infielder: 4 Spieler
- Outfielder: 3 Spieler
- Homeplate: Mittelpunkt des Spielgeschehens

1. Die Sportart Baseball

Das New York Yankee Stadion:



Stephanie Bellinghausen

5

1. Die Sportart Baseball

Strike-Zone aus Sicht des Pitchers



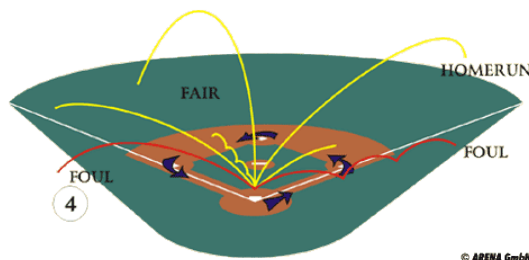
- Das Spiel beginnt beim Pitcher
- Ball ist durch unterschiedliche Wurfgeschwindigkeiten und Tricks unberechenbar
- Wurf zählt nur, wenn der Ball durch die Strike-Zone fliegt
- Schiedsrichter entscheidet über Gültigkeit des Wurfs

Stephanie Bellinghausen

6

1. Die Sportart Baseball

Mögliche Schlagrichtungen des Hitters



Foul – Ball landet außerhalb des Spielfelds

Fair – Ball landet im Spielfeld

Homerun – Ball fliegt über die hintere Spielfeldbegrenzung

Stephanie Bellinghausen

7

1. Die Sportart Baseball

Duell zwischen Hitter und Pitcher kann auf drei Arten enden:

1. Der Hitter ist „Out“, weil er den Ball dreimal nicht getroffen hat
2. Der Hitter rückt automatisch auf die erste Base vor, weil der Pitcher vier Balls geworfen hat
3. Der Hitter hat den Ball getroffen → wird zum Baserunner

Ziel der Verteidigung: Möglichst schnell drei „Outs“ erzielen

Stephanie Bellinghausen

8

1. Die Sportart Baseball

Möglichkeiten einen Angreifer „Out“ zu machen:

1. Hitter trifft den Ball dreimal nicht → „Out“
2. Der geschlagene Ball wird direkt aus der Luft gefangen
3. „Force out“: Verteidiger bekommen den geschlagenen Ball unter Kontrolle und dieser kommt noch vor dem Läufer an der ersten Base an → Läufer ist „Out“
4. Ein Verteidiger berührt einen Baserunner mit dem Ball, wenn dieser gerade nicht auf einer Base steht

1. Die Sportart Baseball

Die Spielzeit

- Ein Inning: Jede Mannschaft spielt einmal in der Verteidigung und einmal als Angreifer
 - Insgesamt: Neun Innings
 - Falls es nach neun Innings unentschieden steht: Verlängerung um jeweils ein Inning, bis ein Sieger feststeht
- Keine feste Spielzeit, da es immer darauf ankommt, wie schnell drei Schlagmänner „Out“ sind

1. Die Sportart Baseball

2. Wichtige Statistiken im Baseball

Baseball in Amerika:



Play-offs



World-Series
über 7 Spiele

National League:
-1876 gegründet
-15 Teams

American League:
-1901 gegründet
-15 Teams

Eine Saison dauert von April bis Oktober → 162 Spiele = 5 Spiele pro Woche

MLB.com
Team Sites | Scoreboard | Standings | Stats | Schedule | Players | News | Audio & Video | Fantasy | Tickets | Shop

Stats
Thursday, December 17, 2009
print this page

MLBPLAYERS.COM
Sortable Player Stats
Major League Baseball Hitting Stats, World Series 2009
Next Stats >>

PLAYERSEARCH
Active Players:
Enter Last Name [Go]
Historical Players:
Enter Last Name [Go]

COMPARE	Player	TEAM	POS	G	AB	R	H	2B	3B	HR	RBI	TB	BB	SO	SB	CS	OBP	SLG	AVG
<input type="checkbox"/>	1. D Jeter	NYN	SS	6	27	5	11	3	0	0	1	14	1	6	0	0	.429	.519	.407
<input type="checkbox"/>	2. P Feliz	PHI	3B	6	23	2	4	1	0	1	2	8	0	4	0	0	.174	.348	.174
<input type="checkbox"/>	3. R Howard	PHI	1B	6	23	3	4	2	0	1	3	9	2	13	1	0	.240	.391	.174
<input type="checkbox"/>	4. R Ibanez	PHI	OF	6	23	2	7	4	0	1	4	14	1	9	0	0	.333	.609	.304
<input type="checkbox"/>	5. J Rollins	PHI	SS	6	23	3	5	0	0	0	2	5	5	2	3	0	.345	.217	.217

Player	TEAM	POS	SF	SH	HBP	IBB	GDP	TPA	NP	XB%	SB%	GO	AO	GO/AO	OPS	
<input type="checkbox"/>	1. D Jeter	NYN	SS	0	0	0	0	1	28	95	3	0	7	4	2.00	.947
<input type="checkbox"/>	2. P Feliz	PHI	3B	0	0	0	0	1	23	76	2	0	9	6	1.67	.522
<input type="checkbox"/>	3. R Howard	PHI	1B	0	0	0	0	0	25	101	3	100	1	5	0.20	.631
<input type="checkbox"/>	4. R Ibanez	PHI	OF	0	0	0	0	0	24	90	5	0	3	4	0.75	.942
<input type="checkbox"/>	5. J Rollins	PHI	SS	1	0	0	0	1	29	112	0	100	7	11	0.73	.562

2. Wichtige Statistiken im Baseball

At-Bat (AB):

Wird zur Berechnung der Offensivleistungen eines Spielers benötigt

$$AB = PA - BB - HBP - SH - SF - IO$$

PA = Plate Appearance – Anzahl der Schlagdurchgänge

BB = Base on Balls

HBP = Hit by Pitch

SH = Sacrifice Hit

SF = Sacrifice Fly

IO = Interference/Obstruction

19.12.2009

Stephanie Bellinghausen

13

2. Wichtige Statistiken im Baseball

Slugging Percentage (SLG):

- Anzahl Bases, die ein Hitter pro At-Bat erreicht
- Kennzahl zur Berechnung der Schlagkraft eines Hitters

$$SLG = \frac{TB}{AB}$$

Die Total Bases (TB) berechnen sich wie folgt:

$$TB = (1B) + (2*2B) + (3*3B) + (4*HR) = H + 2B + (2*3B) + (3*HR)$$

Total Bases = Gesamtzahl der Bases, die ein Spieler durch seine Hits erlaufen konnte

19.12.2009

Stephanie Bellinghausen

15

2. Wichtige Statistiken im Baseball

Batting Average (AVG):

misst Fähigkeit eines Spielers durch Hits auf Base zu kommen

$$AVG = \frac{H}{AB} = \frac{1B + 2B + 3B + HR}{AB}$$

H = Anzahl der Hits

1B = Single

2B = Double

3B = Triple

HR = Homerun

- Wert zwischen null und eins
- Fähigkeit des Spielers gute Würfe von schlechten zu unterscheiden wird nicht beachtet
- Schlagkraft nicht berücksichtigt

19.12.2009

Stephanie Bellinghausen

14

2. Wichtige Statistiken im Baseball

On-Base-Percentage (OBP):

Fähigkeit des Hitters auf eine Base zu kommen

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

→ beachtet auch Walks

→ Wichtig bei Festlegung der Schlagreihenfolge

On-Base plus Slugging (OPS):

misst gesamte Offensivleistung eines Spielers

$$OPS = OBP + SLG$$

19.12.2009

Stephanie Bellinghausen

15

19.12.2009

Stephanie Bellinghausen

16

2. Wichtige Statistiken im Baseball

Hitting Streak:

- In aufeinander folgenden Spielen mindestens einen Base Hit erzielen
- Serie endet: Plate Appearance und kein Hit oder Sacrifice Fly
- Serie wird fortgesetzt: Hit und Plate Appearances, die in einem Walk, Hit by Pitch, Interference der Verteidiger oder Sacrifice Hit enden
- Longest Hitting Streak: längste bisher erreichte Hitting Streak → Joe DiMaggio (56 aufeinander folgende Spiele in der Saison von 1941)

19.12.2009

Stephanie Bellinghausen

17

3. Statistische Methodik

3. Statistische Methodik

Binomialverteilung

- beschreibt wahrscheinlichen Ausgang einer Folge von gleichartigen und unabhängigen Versuchen mit nur zwei möglichen Ergebnissen
- gewünschtes Ergebnis eines Versuchs: Wahrscheinlichkeit p
- Zahl der Versuche: n
- Wahrscheinlichkeit von insgesamt k Erfolgen:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Schreibweise der Binomialverteilung:

$$\text{Bin}(n, p)$$

19.12.2009

Stephanie Bellinghausen

18

3. Statistische Methodik

Maximum-Likelihood-Methode

- wichtige unbekannte Parameter einer Verteilung schätzen
 - Stichprobe mit bestimmter Anzahl von Objekten einer Population
 - Elemente einer Population sind Realisierung eines Zufallsexperiments
 - Interessante Kennwerte ausschließlich vom unbekanntem Parameter abhängig → als Funktion vom Parameter darstellbar
- Maximum-Likelihood-Schätzer maximiert die Wahrscheinlichkeit die Stichprobe zu erhalten

19.12.2009

Stephanie Bellinghausen

19

Monte-Carlo-Simulation

- komplexe Probleme mithilfe von Algorithmen lösen
- Verhalten des Modells durch mehrmaliges Aufrufen des Algorithmus testen
- Monte-Carlo-Simulation bei komplexen Modellen, nichtlinearen Modellen oder Modellen mit unsicheren Parametern
- Ebenfalls geeignet zur Bestimmung der Likelihood einer Binomialverteilung

19.12.2009

Stephanie Bellinghausen

20

4. Der Artikel von Rockoff und Yates

Entstehung

- Grundlage ist ein Artikel von Arbesman und Strogatz (März 2008)
 - Baseball-Geschichte 10.000mal simuliert, um Wahrscheinlichkeit einer langen Hitting Streak herauszufinden
 - Annahme: At-Bat eines Spielers pro Spiel in allen Spielen einer Saison ist konstant
 - Hitting Streak von mindestens 56 Spielen in insgesamt 42% der simulierten Baseball-Geschichte
- Wahrscheinlichkeit wird überschätzt

19.12.2009

Stephanie Bellinghausen

21

4. Der Artikel von Rockoff und Yates

Das Modell

- zu zeigen: konstante At-Bats überschätzen Likelihood von langen Hitting Streaks
- Schätzung der Wahrscheinlichkeit, dass ein Spieler mindestens einen Hit in Spiel a erzielt:

Methode 1:

p_i ist Wahrscheinlichkeit eines Schlagmanns mit Batting Average A während eines Spiels i-mal zum Schlag zu kommen

$$\hat{p} = \sum p_i (1 - (1 - A)^i)$$

→ Erwartungswert der Wahrscheinlichkeit eines Spielers, mindestens einen Hit zu erzielen

19.12.2009

Stephanie Bellinghausen

22

4. Der Artikel von Rockoff und Yates

Methode 2:

Sei B die durchschnittliche Anzahl der At-Bats pro Spiel

$$\hat{p} = 1 - (1 - A)^B$$

→ Wahrscheinlichkeit eines Spielers, mindestens einen Hit mehr als ihre durchschnittliche Anzahl an At-Bats pro Spiel zu erzielen

Mit Jensen'scher Ungleichung folgt:

$$\hat{p} \text{ der Methode 1} \leq \hat{p} \text{ der Methode 2}$$

→ Konstante At-Bats überschätzen die Likelihood von langen Hitting Streaks

19.12.2009

Stephanie Bellinghausen

23

4. Der Artikel von Rockoff und Yates

Beispiel:

- Ein Spieler hat in zwei Spielen insgesamt 8 At-Bats
 - Batting Average AVG= .300
- Wahrscheinlichkeit in beiden Spielen einen Hit zu erzielen:

$$\hat{p} = \{1 - (1 - .300)^i\} \{1 - (1 - .300)^j\}$$

Spiel 1 At-Bat	Spiel 2 At-Bat	\hat{p}
1	7	0.275
2	6	0.450
3	5	0.547
4	4	0.577
5	3	0.547
6	2	0.450
7	1	0.275

→ Höchste Wahrscheinlichkeit: Spiel 1 und Spiel 2 jeweils vier At-Bats

→ Überschätzung der Wahrscheinlichkeit im nächsten Spiel einen Hit zu erzielen

19.12.2009

Stephanie Bellinghausen

24

4. Der Artikel von Rockoff und Yates

Die Daten

- Daten stammen von Retrosheet
- Durch die Vielzahl an Informationen im Datenbestand: Ermittlung der At-Bats eines jeden Spielers für jedes Spiel in jeder Saison möglich
- Play-by-Play-Daten:
 - gesamte Major League von 1954 bis 2007
 - National League von 1911, 1921, 1922 und 1953

19.12.2009

Stephanie Bellinghausen

25

4. Der Artikel von Rockoff und Yates

Annahmen für die Simulation

- Batting Average p_{ij} für Schlagmann i in Saison j
- Anzahl der At-Bats (AB) jeden einzelnen Spiels für Spieler i mit k Spielen in j Saisonen:

$$AB_{ij} = (AB_{ij1}, AB_{ij2}, \dots, AB_{ijk})$$

- At-Bats im Verlauf eines Spiels sind unabhängig voneinander
 - H_{ijk} ist Anzahl der Hits (H) von Spieler i in Saison j in Spiel k
 - Hits sind binomialverteilt mit $n=AB_{ijk}$ und $p=p_{ij}$

$$H_{ijk} \sim Bin(AB_{ijk}, p_{ij})$$

19.12.2009

Stephanie Bellinghausen

26

4. Der Artikel von Rockoff und Yates

Annahmen für die Simulation

- Jeder Spieler besitzt für jede Saison eine Verteilung der At-Bats über die k gespielten Spiele in einer Saison
- Aus diesen At-Bats: Ziehen einer Stichprobe mit Zurücklegen → „simulierter“ Wert für die At-Bats einer Saison

$$AB_{ij}^* = (AB_{ij1}^*, AB_{ij2}^*, \dots, AB_{ijk}^*)$$

- Bei m simulierten Saisonen: simulierte At-Bats für Spieler i in Saison j

$$AB_{ij}^{*1}, AB_{ij}^{*2}, \dots, AB_{ij}^{*m}$$

19.12.2009

Stephanie Bellinghausen

27

4. Der Artikel von Rockoff und Yates

Annahmen für die Simulation

- Anzahl der Hits eines Spielers in jedem Spiel der m -ten simulierten Saison:

$$H_{ij}^{*m} \sim Bin(AB_{ij}^{*m}, p_{ij})$$

- Eine Hitting Streak ist also eine Reihe von Hits in H_{ij}^{*m} , die alle größer als null sind
- In jeder simulierten Saison wird für jeden Spieler jeweils die maximale Hitting Streak betrachtet

19.12.2009

Stephanie Bellinghausen

28

4. Der Artikel von Rockoff und Yates

Simulation und Ergebnisse

- nur wenige Daten von vor 1953 vorhanden
- Jede der 58 vorliegenden Saisonen wurde 1.000mal simuliert → 58.000 simulierte Saisonen
- Hitting Streak von mindestens 56 Spielen wurde in 30 der simulierten Saisonen erreicht $\approx 0,00517\%$

19.12.2009

Stephanie Bellinghausen

29

5. Fazit

4. Der Artikel von Rockoff und Yates

Simulation und Ergebnisse

- Vergleich des variablen At-Bat-Modells mit dem konstanten At-Bat-Modell: DiMaggios Saison von 1941 10.000mal simulieren

Methode	Max	40+	50+	56+
konstantes At-Bat-Modell	75	57	8	2
variables At-Bat-Modell	57	41	2	1

- Konstantes At-Bat-Modell weist überall höhere Zahlen auf als das variable At-Bat-Modell
- Eine Hitting Streak ist im variablen At-Bat-Modell seltener

19.12.2009

Stephanie Bellinghausen

30

5. Fazit

- Nur Play-by-Play-Daten ab 1953 vorhanden
- Möglichkeit der Verwendung der Play-by-Play At-Bat-Verteilung ab 1953 zur Modellierung der At-Bat-Verteilung bis 1953
→ basierend auf den durchschnittlichen At-Bats pro Spiel
- Problem: die betrachteten Modelle haben eine schlechte Anpassung an die tatsächlichen Daten
→ Modellierung einer statistischen Verteilung nicht möglich
- Komplexere Modellierung der At-Bat-Verteilung
- Möglichkeit der Verbesserung der Simulation: Batting Average eines Spielers als Zufallsvariable, die sich von Spiel zu Spiel und At-Bat zu At-Bat verändert

19.12.2009

Stephanie Bellinghausen

31

- Keine Berücksichtigung von Entscheidungen im richtigen Baseballspiel, die die At-Bats eines Spielers beeinflussen z.B. wird ein Spieler nicht mitten unter einem Spiel herausgenommen, wenn er eine Hitting Streak am Laufen hat und bis dahin noch keine Hit erzielt hat
- Möglichkeit von mehreren Hitting Streaks während einer Saison werden nicht berücksichtigt
→ nur die maximale Hitting Streak wird betrachtet
- Bessere Anpassung an die realen Werte mit der Bootstrap-Methode: bereits existierende Daten verwenden und Zufallsstichprobe mit Zurücklegen ziehen und notieren
→ Bestimmung der jeweils längsten Hitting Streak

19.12.2009

Stephanie Bellinghausen

32

6. Weitere mögliche Statistik

Defensivstatistik: „Grounder Balls-in-Play“

- Berücksichtigung der letzten Ereignisse auf dem Spielfeld
- Positionen der Verteidiger müssen geschätzt werden
→ Beschreibung durch x- und y-Koordinate
- Berechnung des Aufschlagpunktes des Baseballs durch Erfassung des Winkels und der Geschwindigkeit
- Ein Winkel von 0° entspricht der ersten Base-Linie und ein Winkel von 90° entspricht der dritten Base-Linie
→ Es wird bestimmt, welcher Verteidiger auf welcher Position im Spielfeld, die höchste Wahrscheinlichkeit besitzt, den Ball zu bekommen

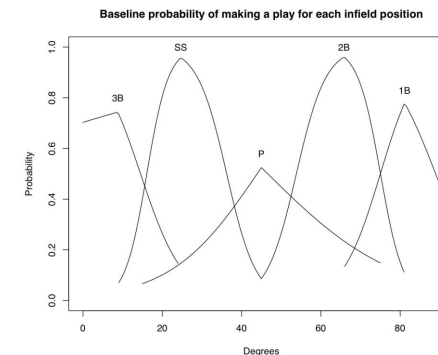
Stephanie Bellinghausen

33

6. Weitere mögliche Statistik

Defensivstatistik: „Grounder Balls-in-Play“

Wahrscheinlichkeit auf den Infielder-Positionen:



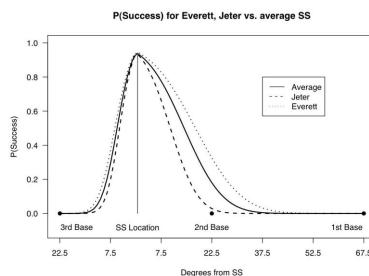
Stephanie Bellinghausen

34

6. Weitere mögliche Statistik

Defensivstatistik: „Grounder Balls-in-Play“

Die Position Shortstop für Derek Jeter und Adam Everett:



SAFE-Wert: gibt an, ob ein Spieler besser oder schlechter als der Durchschnitt auf dieser Position spielt

Stephanie Bellinghausen

35

7. Abschließende Bemerkungen

- Eine weitere mögliche Defensivstatistik: Ultimate Zone Rate (UZR) – gibt an, ob ein Spieler mehr Runs für seine Mannschaft erzielen konnte oder ob er der gegnerischen Mannschaft mehr Runs ermöglicht hat

Derek Jeter:

- 2007: UZR = -15,3 → gezieltes Defensivtraining nötig
- 2008: UZR = -0,5
- 2009 nach den ersten 90 Spielen: UZR = 1,8

→ Baseball-Spieler beachten ihre Statistiken und engagieren gegebenenfalls einen Trainer zur Verbesserung ihrer Leistungen

Stephanie Bellinghausen

36