BASEBALL STATISTICS
MEETS
MATHEMATICAL STATISTICS

Hal Stern
Department of Statistics
University of California, Irvine
email: sternh@uci.edu

Symposium on Statistics and Operations Research in Baseball
CSU - East Bay
July 11, 2007

Baseball statistics

Introduction

- “Statistics” vs “statistics”
  - Statistics is the science concerned with the collection, analysis, interpretation, and presentation of numerical data
  - statistics refers to a collection of quantitative data (e.g., batting averages, points scored in basketball)
- Sports examples can be extremely useful in the teaching of probability and statistics
- Probability and statistics can also contribute to the study of sports

Baseball and statistics

- Statistics are a big part of the fabric of baseball
- Consider season-long statistics (2001) for two pitchers (on the same team!)
  - Pitcher A: 220.1 inns, 213 K, 205 H, 3.51 ERA, .309 OBP, .375 SLG
  - Pitcher B: 229.2 inns, 214 K, 202 H, 3.15 ERA, .274 OBP, .358 SLG
  - Fairly even performance ... slight edge to Pitcher B
- Pitcher A is Roger Clemens. His record was 20 wins and 3 losses.
- Pitcher B is Mike Mussina. His records was 17 wins and 11 losses.
- Roger Clemens won Cy Young award (best pitcher) but does he deserve it?
- Why does Clemens have more wins .... run support (team scores 6.6 runs/game for Clemens vs 4.5 runs/game for Mussina)

  (Rob Neyer of ESPN.com first pointed this out)

Understanding effects of uncertainty and variability

- Baseball has a number of sacred marks (20 wins, .300 average) but ....
- Sample sizes are small and effects of randomness/variability can be large
  - Bill Gullickson 20 wins - 7 losses in 1991
    - never had success like that before or after (career winning pct $\approx$ .50)
  - Luis Aparicio batted .313 (552 ABs) in 1970
    - career (18 yrs) .262 avg (next highest season is .280)
    - there is some probability (approx .03) that a "true" .262 hitter would hit .300+ by chance

Understanding effects of uncertainty and variability

- Non-random sampling is also an issue
- Gary Sheffield enters all-star break "hot" (11 for his last 24, .458 avg)
  - this reflects some selection
    (e.g., he is also 12/27 (.444) if we include one more game
    and 8/19 (.421) if we include one less game)
  - can account for this by simulation
  - assume we report best performance found by searching over intervals ranging from last 3 days to last 14 days
  - for career .300 batter the median "best" performance we would find is .389 (2.5%ile is .219 and 97.5%ile is .636)

Baseball statistics and mathematical statistics

- Can talk about impact of mathematical statistics in a few different areas (briefly)
- Baseball strategies and Markov models
  - George Lindsey's data
  - assessing the value of the sacrific bunt or intentional walk
- Player valuation
  - SABR-metrics
  - player win percentage
  - comparing player's across eras
- Modelling multiplicities
  - situational batting
  - batter-pitcher matchups

George Lindsey's 1959-1960 data

Bases occup	Outs	# of obs	Distn of runs scored Pr(0)	Pr(1)	Pr(> 1)	Mean runs	StdErr mean
0	0	6561	0.747	0.136	0.117	0.461	0.012
0	1	4664	0.855	0.085	0.060	0.243	0.011
0	2	3710	0.933	0.042	0.025	0.102	0.008
1	0	1728	0.604	0.166	0.230	0.813	0.031
1	1	2063	0.734	0.124	0.142	0.498	0.022
1	2	2119	0.886	0.045	0.069	0.219	0.016
2	0	294	0.381	0.344	0.275	1.194	0.083
2	1	657	0.610	0.224	0.166	0.671	0.043
2	2	779	0.788	0.158	0.054	0.297	0.024
3	0	67	0.12	0.64	0.24	1.39	0.09
3	1	202	0.307	0.529	0.164	0.980	0.072
3	2	327	0.738	0.208	0.054	0.355	0.040
12	0	367	0.395	0.220	0.385	1.471	0.087
12	1	700	0.571	0.163	0.266	0.939	0.051
12	2	896	0.791	0.100	0.109	0.043	0.032
13	0	119	0.13	0.41	0.46	1.94	0.15
13	1	305	0.367	0.400	0.233	1.115	0.077
13	2	419	0.717	0.167	0.116	0.532	0.054
23	0	73	0.18	0.25	0.57	1.96	0.18
23	1	176	0.27	0.24	0.49	1.56	0.10
23	2	211	0.668	0.095	0.237	0.687	0.080
123	0	92	0.18	0.26	0.56	2.22	0.20
123	1	215	0.303	0.242	0.455	1.642	0.105
123	2	283	0.671	0.092	0.237	0.823	0.085

Determining baseball strategy

- Markov model treats baseball game as moves among series of states (bases occupied / no. of outs)
- Rules of baseball limit the number of ways to move from one state to another
- Example: transition from (0 out, bases=12) to (0 out, bases=123) is a walk or single
- Probabilities of the various transitions can be found from play-by-play data or via probabilities of common events

Event	1989 American	1989 National
Out	0.672	0.685
Walk	0.092	0.091
Single	0.171	0.159
Double	0.040	0.040
Triple	0.005	0.006
Home run	0.020	0.019
Errors	2% of outs	
Sac fly	12.5% of outs (if applicable)	
Double play	17.5% of outs (if applicable)	

- Then can use probability theory (Markov chains) to determine quantities of interest

Determining baseball strategy

Probability-based results - Expected runs

Results using 1989 American League statistics

Bases occup.	Probability of scoring			Expected runs		
	0 out	1 out	2 out	0 out	1 out	2 out
0	0.26	0.16	0.07	0.49	0.27	0.10
1	0.39	0.26	0.13	0.85	0.52	0.23
2	0.57	0.42	0.24	1.06	0.69	0.34
3	0.72	0.55	0.28	1.21	0.82	0.38
12	0.59	0.45	0.24	1.46	1.00	0.48
13	0.76	0.61	0.37	1.65	1.10	0.51
23	0.83	0.74	0.37	1.94	1.50	0.62
123	0.81	0.67	0.43	2.31	1.62	0.82

Determining baseball strategy

Applying the results

- Runner at 1st base, 0 out – Sacrifice bunt
  Offensive team can have the next hitter sacrifice himself to place a runner on 2nd base with 1 out. For now assume the sacrifice is always successful.
  - No sacrifice:
    expected runs = 0.85 and Pr(scoring) = 0.39
  - Sacrifice:
    expected runs = 0.69 and Pr(scoring) = 0.42
  - Sacrifice decreases chance of a big inning, but may help score a run

Determining baseball strategy

Applying our results

- Runner at 2nd base, 0 out – Intentional walk

  Defensive team can intentionally walk the next hitter to place runners on 1st base and 2nd base with 0 out.

  - No intentional walk:

    expected runs = 1.06 and Pr(scoring) = 0.57
  - Intentional walk:

    expected runs = 1.46 and Pr(scoring) = 0.59
  - Should not use the intentional walk

Determining baseball strategy

Discussion

- BIG PROBLEM – Data is averaged over all players
  - different probabilities should be used in considering the usefulness of a strategy depending on the player involved (e.g., intentionally walking Barry Bonds is different than intentionally walking David Eckstein)
  - more complete analysis can use individual player's statistics so that each individual has their own probability of transition matrix (e.g., Bellman 1977, Bukiet 1997)

Evaluating players

- A major use of statistics in baseball has always been the evaluation of players
- Recent developments, as chronicled in *Moneyball* and involving among others,
  - SABR
  - Bill James
  - Baseball Prospectus

  has advocated replacing traditional statistics (batting average, RBIs) with alternatives
- Some of the appproaches are VERY BRIEFLY described next

Evaluation of players - linear models

- One common approach is to regress team output (runs) on team events (1b, 2b, 3b, hr, bb) to identify appropriate weights
- This (or similar idea) has been often applied
- Leads to OPS (on-base pct + slugging) as an improvement over either alone
- Result "regress" below is for a regression approach applied by Albert and Bennett in their book Curve Ball

	weight assigned					
measure	sb	bb	1b	2b	3b	hr
on-base	0	1	1	1	1	1
slugging	0	0	1	2	3	4
total avg	1	1	1	2	3	4
regress	.31	.69	1	1.29	2.27	2.88

Evaluation of players - runs produced/saved

- Most performance measures are essentially attempts to measure run productivity of players
- We can use the Markov chain model to do this directly
  - Cover's 1977 OERA (plug in probabilities of different outcomes for an individual player and compute expected runs scored)
  - Or can apply Markov model expected runs matrix on a play-by-play basis
    - Example: relief pitcher enters game with 0 out and bases 123
    - relief pitcher leaves game with 2 outs after allowing 1 run to score and bases are still loaded
    - team expected to yield 2.31 runs when relieve pitcher entered
    - team gave up 1 run and now expects to yield 0.82 runs more
    - this relief pitcher "saved" .49 runs
    - can accumulute or average such performance scores
    - could also partition credit for each event to batter and pitcher

Evaluation of players - win probabilities

- Can also take things to the next level
- Goal is to win games rather than score runs
- It is possible to compute the probability of winning from any situation (Lindsey 1961, PROTRADE 2007, Bennett and Flueck 1984, etc.)
  - Can use earlier results to compute probability distribution on runs scored in remainder of half inning
  - Can use distribution of runs per half inning to find probability of a win after that point
- Can use idea above with win probabilities (the player game percentage of Bennett and Flueck)
- P(Team A wins) = .55 before a play and P(Team A wins) = .70 after, then the offensive/defensive players split 0.15 change
- Problem is that this puts very large weight on late game heroics

Evaluation of players - comparing across eras

- It has always been of great interest to compare players across seasons
- This is difficult because of the many differences in training, diet, ball parks, strategy over time
- Statistical modelling offers one possibility to build on the fact that careers overlap
- Berry et al. (JASA 1999) build such a model in which probability of a hit depends on
  - decade
  - ball park
  - age of player
- Sophisticated statistical methods are used to fit the model to the data (introduced in next few slides in a simpler context)
- Top 5 players by peak avg are Cobb (.368), Gwynn (.363), Williams (.353), Boggs (.353), Carew (.351)
- Also estimate variation in aging profiles (Gwynn stayed near peak longer)

Dealing with multiplicities (potential uses of modern methods)

- Recent contributions of newer statistical methods (like Berry et al.) seem to emphasize situations in which the goal is to estimate a large number of performance parameters simultaneously
- Examples – home/away splits for all players; batter-pitcher matchups
- Methods build models that treat individual parameters (home - away avg for a player) as coming from a population
- Such methods (often known as hierarchical models, random effects models and occasionally emphasizing Bayesian interpretations) allow for a compromise between two common extreme approaches
  - treat each player as an individual (but then subject to small sample vagaries)
  - treat all players as being identical (loses the main question of interest)
- Hierarchical models compromise allows us to distinguish among players but not over react to small samples

Batter-pitcher matchups

- Aug 29, 2006, LA Dodgers vs Cincinnati Reds
- Kenny Lofton (career .299 avg; season .308 avg) sits out for a “rest”
- Kenny Lofton career 1-for-19 vs Reds pitcher E. Milton
- Such “rest”’s are common ... but does it make sense?
- Lofton’s substitute has .273 avg
- Is putting in a weaker player really a better bet?

Batter-pitcher matchups

- Consider a detailed study of D. Jeter
- Data thru July 23, 2006 (Jeter had faced 382 pitchers at least five times)
- Table on next page lists Jeter's average in selected matchups
- Jeter's best avg is .833 against R. Mendoza (5 for 6)
- Can ask probability that .316 career hitter has 5 hits in 6 at-bats. Probability is .014 indicating this is highly unusual BUT ......
- No a priori reason to ask about Jeter vs Mendoza
- If we take into account the large number of pitchers Jeter has faced this ceases to be interesting

Pitcher	At-bats	Hits	ObsAvg	Est Avg	95% int
R. Mendoza	6	5	.833	.322	(.282, .394)
H. Nomo	20	12	.600	.326	(.289, .407)
R. Lopez	45	21	.467	.326	(.292, .401)
T. Wakefield	81	26	.321	.318	(.279, .364)
P. Martinez	83	21	.253	.312	(.254, .347)
J. Smoltz	5	1	.200	.317	(.263, .366)
B. Radke	41	8	.195	.311	(.247, .347)
D. Kolb	5	0	.000	.316	(.258, .363)
TOTAL	6530	2061	.316		

Batter-pitcher matchups

- Turns out that most traditional probability arguments indicate data are consistent with constant hitting ability across pitchers
- (See also, Dan Fox article at hardballtimes.com)
- BUT THIS CAN'T BE RIGHT !!!!!
- Statistical model to allow for variation
  - assume Jeter's ability against pitcher i is π_i
  - assume π_i's vary across the population of pitchers according to particular probability distribution (Beta)
  - estimate parameters of the Beta distribution and the individual π_i's
- Can go back to Table and look at estimates
- Some (but very little) variation is found for Jeter (avgs range from .311 to .326)

Batter-pitcher matchups

- What happens with other players?
- Key parameter is variance of underlying (Beta) distribution
- Have repeated above for 200+ players
- Jeter's avg is quite homogeneous (varies only from .311 to .327)
- Kenny Lofton is quite heterogeneous
- Lofton's estimated ability varies from .265 (for E. Milton) to .340
- Note: Can (and have) done the same analysis for pitchers

Batter-pitcher matchups

- Closely related to work of Jim Albert (JASA, 1994)
- As always one can do better, perhaps by incorporating
  - simultaneous inference for batters and pitchers
  - ballpark effects
  - left-right effects

References (generally all sports)

- Optimal Strategies in Sports - 1977 book edited by Ladany and Machol
- Management Science in Sports - 1976 book edited by Machol, Ladany, Morrison
- Statistics in Sport - 1998 book edited by Bennett
- Curve Ball - 2001 book by J. Albert and J. Bennett
- Anthology of Statistics in Sports - 2005 collection of articles edited by Albert, Bennett, Cochran
- Statistical Thinking in Sports - 2007 (?) book edited by Albert and Koning
- Chance magazine (column beginning in vol. 10)
- Articles in statistics and operations research journals (and other fields also)
- New electronic Journal of Quantitative Analysis in Sports (JQAS)