USE AND ABUSE OF INFORMATION IN SPORTS
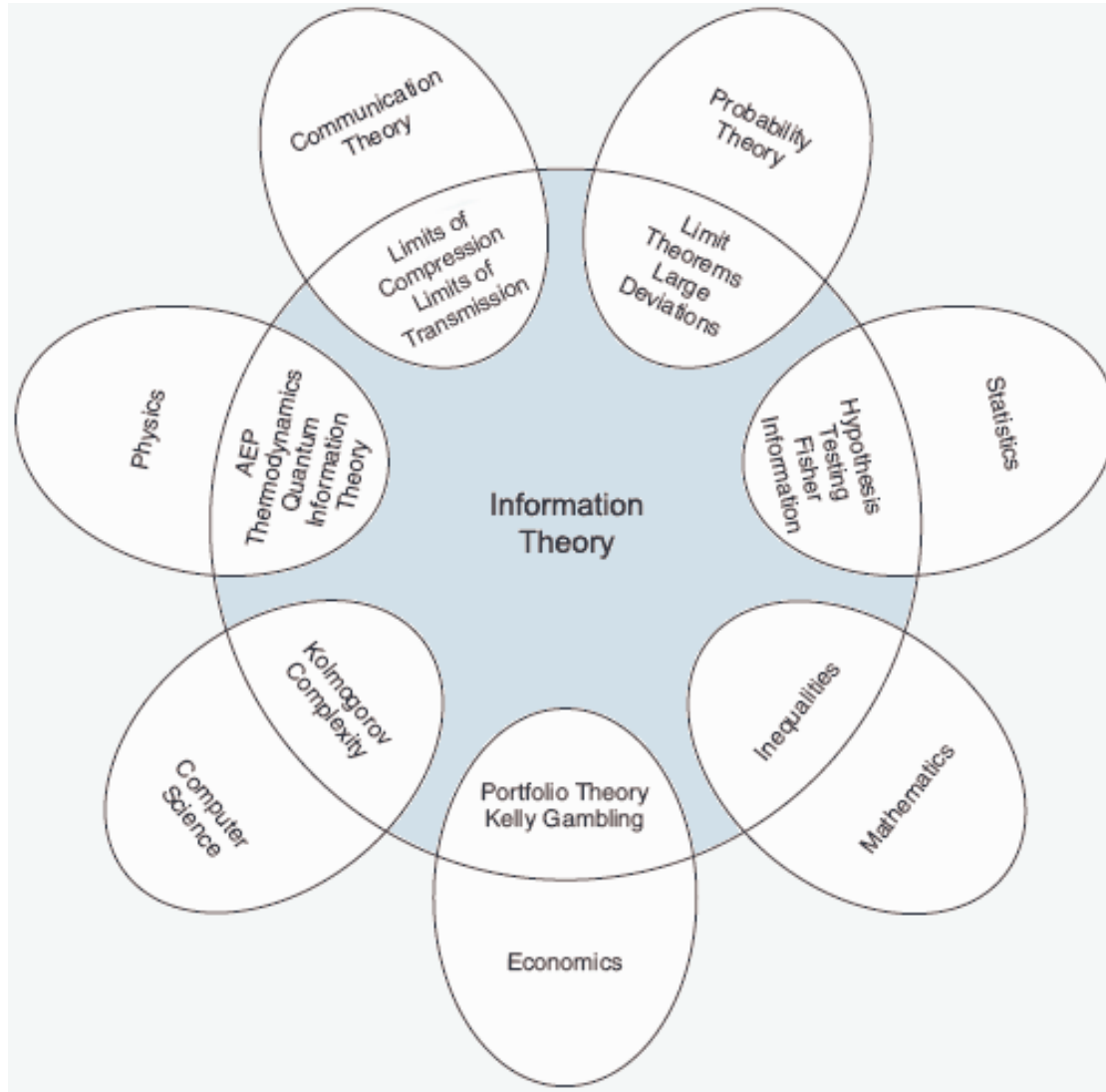
Hal Stern

Department of Statistics

University of California, Irvine

email: sternh@uci.edu

Communication Theory

Probability Theory

Limits of Compression Limits of Transmission

Limit Theorems Large Deviations

Physics

AEP Thermodynamics Quantum Information Theory

Hypothesis Testing Fisher Information

Statistics

Information Theory

Kolmogorov Complexity

Inequalities

Computer Science

Portfolio Theory Kelly Gambling

Mathematics

Economics

Stat 50 / MCS 100: Mathematics of Sports

- Instructor: Tom Cover

- Structure: The running theme will be to find new strategies, techniques and statistics for sports

- Topics

  - Statistics in sports (Anomalies, Developing new statistics)

  - Strategy in sports

  - Physics in sports

Today - Probability and statistics in sports

- Goal: Show how **information** (broadly defined) is used and abused by sports professionals/fans

- Topics:
  - Basic baseball statistics (the ugly, the bad, the good)
  - Cover's OERA and Markov Chain baseball
  - Batter-pitcher matchups: a beta-binomial approach
  - Ranking, prediction and information
  - Great comebacks and the probability of wining

Baseball statistics

- Statistics are a big part of the fabric of baseball

- Consider season-long 2001 statistics 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 scored 6.6 runs/game for Clemens vs 4.5 runs/game for Mussina)

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

Baseball statistics

- Traditional baseball milestones have limitations as statistical measures

- We tend to underestimate the effects of variability in small samples
  - 1991 Bill Gullickson of Tigers had 20 wins - 7 losses
    * never had success like that before or after
    * career winning pct $\approx$ .50
    * was this just a random event? $(p \approx .01)$
  - 1992 Mike Bordick of A's had .300 batting average
    * career .260 hitter
    * would a "true" .260 hitter ever hit .300?
    * yes, probability is .025

Baseball statistics

- Some good stuff

  - Moneyball by Michael Lewis

  - Writings of Bill James

  - baseballprospectus.com

  - Markov chain model of baseball
    * 25 "states" = (outs=0,1,2) x (bases=0,1,2,3,12,13,23,123) + end inning
    * use data, theory, or some combination to estimate the probability of moving from one state to another
    * use data or theory to estimate distribution of outcomes from any state
    * one use of model is Cover and Keilers OERA (1977)

George Lindsey's 1959-1960 data

Bases occup	Outs	# of obs	Distn of runs scored			Mean runs	StdErr mean
			Pr(0)	Pr(1)	Pr(> 1)		
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

Baseball statistics - Markov chain model
Probability-based results - Expected runs

- Can use probabilities of basic events (out, walk, single, double, triple, home run) to populate transition matrix and then solve for key quantities

- 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

Markov chain model – evaluating players

- Can use the Markov chain model to help evaluate players

- Example I - Cover/Keilers 1977 OERA:
  expected number of runs for a lineup of 9 copies of an individual player

- Example II - measure player performance relative to expected runs
  - relief pitcher enters game with 0 out and bases 123
  - leaves game after allowing 1 run to score with 2 out and bases 123
  - team expected to yield 2.31 runs when relief pitcher entered
  - team gave up 1 run and now expects to yield 0.82 runs more
  - this relief pitcher "saved" .49 runs
  - can accumulate or average such performance scores

Markov chain model – baseball strategy

- <u>Runner at 1st base, 0 out – Sacrifice bunt</u>
  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 $\Pr(\text{scoring}) = 0.39$

  - Sacrifice:
    expected runs $= 0.69$ and $\Pr(\text{scoring}) = 0.42$

  - 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 BUT ....

Markov chain model - difficulties

- Data is averaged over all players
  - different probabilities should be used in considering the usefulness of a strategy depending on the players involved (e.g., intentionally walking Barry Bonds is different than intentionally walking David Eckstein)
  - more complete analysis can use individual player's statistics (i.e., 9 different transition matrices)

Baseball statistics: batter-pitcher matchups

- Consider the following baseball anecdote

  - 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 "rests" are common ... but does it make sense?

  - Lofton's substitute has .273 avg

- Some major league managers believe strongly in the importance of such matchup data (e.g., Tony LaRussa in *Three Days in August*)

- But is putting a weaker player in the lineup really a better bet?

- What should we make of batter-pitcher matchups based on small samples?

Batter-pitcher matchups: the binomial view

- If we think of Lofton's batting as a Bernoulli process with constant probability of success .3, then we can use probability to assess how unusual is Lofton's poor performance against Milton

- $\Pr(\leq 1$ hit in 19 attempts$) = .010$ so 1-for-19 is an unusual outcome

- Not so unusual when we consider the large number of matchups we have not accounted for the multiplicity of such matchups

- Dan Fox on-line report of 30,000 pitcher-batter matchups finds that there is nothing in the data that one would not expect to occur by chance

- Thus traditional analyses suggest batter-pitcher matchup data are consistent with the null hypothesis of constant hitting ability

- BUT THIS CAN'T BE RIGHT !!!!!

- We KNOW probability of success varies from day-to-day due to:
  - skill of the opposing pitcher
 - left-right considerations
  - site of game

A Derek Jeter study

- Data thru July 23, 2006 from www.sports.yahoo.com

- Jeter had faced 382 pitchers at least five times

Pitcher	At-bats	Hits	ObsAvg
R. Mendoza	6	5	.833
H. Nomo	20	12	.600
A. J. Burnett	5	3	.600
E. Milton	28	14	.500
D. Cone	8	4	.500
R. Lopez	45	21	.467
K. Escobar	39	16	.410
J. Wetteland	5	2	.400
T. Wakefield	81	26	.321
P. Martinez	83	21	.253
K. Benson	8	2	.250
T. Hudson	24	6	.250
J. Smoltz	5	1	.200
F. Garcia	25	5	.200
B. Radke	41	8	.195
D. Kolb	5	0	.000
J. Julio	13	0	.000
TOTAL	6530	2061	.316

Batter-pitcher matchups: beta-binomial view

- Consider a fairly simple statistical model that allows for variation in ability
  - 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, $\text{Beta}(\alpha, \beta)$
    * mean of this distribution is $\mu = \alpha/(\alpha + \beta)$
    * variance is $\mu(1 - \mu)/(\alpha + \beta + 1)$
    * $\phi = 1/(\alpha + \beta + 1)$ measures degree to which beta is concentrated around the mean ($\phi = 0$ means highly concentrated and $\phi = 1$ means widely dispersed)
  - inference for α, β from the data (Bayesian analysis as in *Bayesian Data Analysis* by Gelman et al. 2003)

Beta distributions
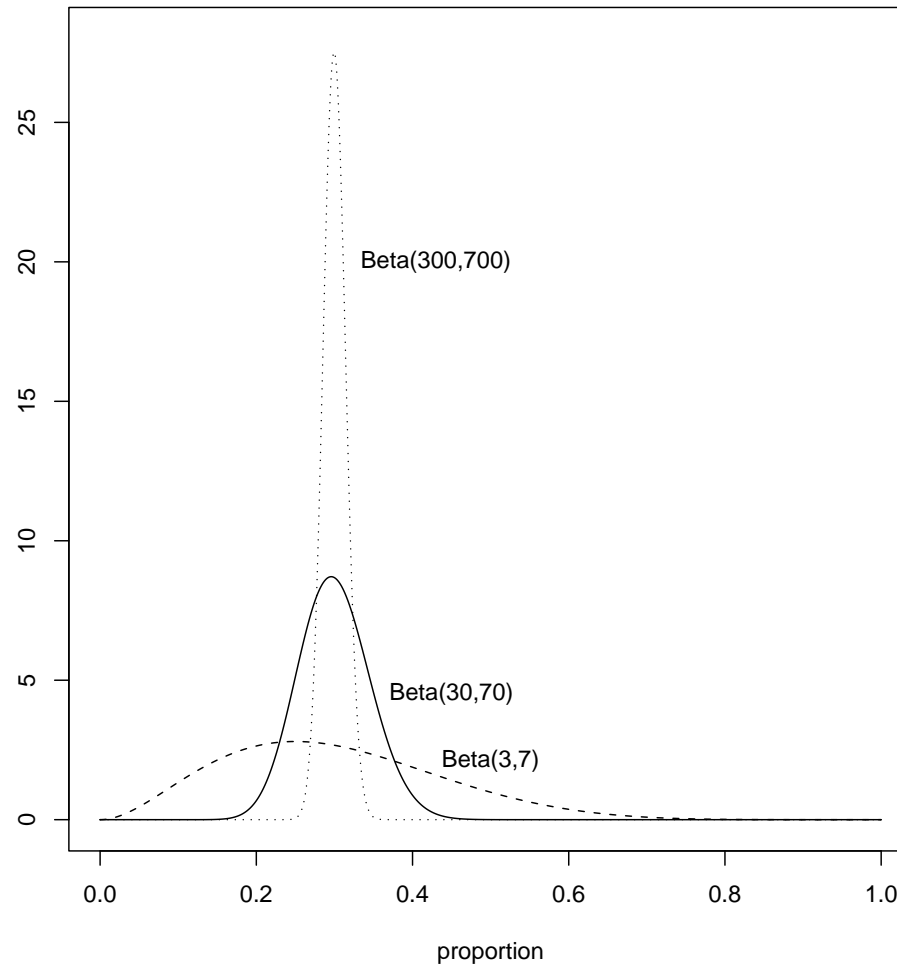
Derek Jeter study - results

- Posterior median for $\mu = .318$ (95% interval .310 to .327)

- Posterior median for $\phi = .002$ (95% interval .000 to .010)

  (note Beta(3,7) has $\phi = .09$;
  Beta(30,70) has $\phi = .01$;
  Beta(300,700) has $\phi = .001$)

- Jeter's performance is extremely consistent across pitchers

Derek Jeter study - results

Pitcher	At-bats	Hits	ObsAvg	EstAvg	95% int
R. Mendoza	6	5	.833	.322	(.282, .394)
H. Nomo	20	12	.600	.326	(.289, .407)
A. J. Burnett	5	3	.600	.320	(.275, .381)
E. Milton	28	14	.500	.324	(.292, .397)
D. Cone	8	4	.500	.320	(.278, .381)
R. Lopez	45	21	.467	.326	(.292, .401)
K. Escobar	39	16	.410	.322	(.281, .386)
J. Wetteland	5	2	.400	.318	(.275, .375)
T. Wakefield	81	26	.321	.318	(.279, .364)
P. Martinez	83	21	.253	.312	(.254, .347)
K. Benson	8	2	.250	.317	(.264, .368)
T. Hudson	24	6	.250	.315	(.260, .362)
J. Smoltz	5	1	.200	.317	(.263, .366)
F. Garcia	25	5	.200	.314	(.253, .355)
B. Radke	41	8	.195	.311	(.247, .347)
D. Kolb	5	0	.000	.316	(.258, .363)
J. Julio	13	0	.000	.312	(.243, .350)
TOTAL	6530	2061	.316		

Batter-pitcher matchups: other players

- If all players are like Jeter, then it might be reasonable to act as if the null hypothesis (constant ability) was true

- What happens with other players?

- Repeated above analysis for 230 players
  (those with at least 350 plate appearances in 2006)

- Used career data for each player

- Key parameter is ϕ and estimates range from $\phi = .0006$ to $\phi = .11$

- Small ϕ: Jeter (.002, around 10th percentile) has estimated ability that varies from .311 to .327

- Medium-to-high ϕ: Lofton (.008, around 65th percentile) has estimated ability that varies from .265 (vs Milton!!) to .340

Rating teams, prediction, and betting

- Long-time interest in rating teams and the effect of using different types of information

- Turns out to be relevant to:
  - college football and the BCS
  - predictability of sports
  - gambling / efficiency of betting market
  - probability of come-from-behind win

Least squares ratings
A simple idea for rating teams

Let R_i be the rating for team i

- Assume we have series of games with results Y

- Try to find/estimate R's using this data

- Suppose team i plays team j
  - prediction: $R_i - R_j \pm H$ (home field advantage)
  - actual outcome: $Y = i$'s score $- j$'s score

- Choose R's so that Y is "close" to $R_i - R_j \pm H$
  for most games

- One idea: minimize sum of squared prediction
  error, $\sum_{\text{games}} (Y - (R_i - R_j \pm H))^2$

- Turns out that R_i in this way is sum of
  avg performance (avg Y) + strength of schedule (avg R of opponents)

Least squares ratings

Bells and whistles

- Reduce impact of blowouts
  (e.g., truncate at M point margin)

$$Y^* = \begin{cases} M & Y \geq M \\ Y & -M < Y < M \\ -M & Y \leq -M \end{cases}$$

$M = 1$ ignores score as required by BCS

- Explicit reward for winning games
  (e.g., give winning team B extra pts)

$$Y^* = \begin{cases} B + Y & Y > 0 \\ 0 & Y = 0 \\ -B + Y & Y < 0 \end{cases}$$

- Weight recent games more than early games

$$\text{minimize} \quad \sum_g w_g (Y - (R_i - R_j \pm H))^2$$

where w_g is the weight for game g

- Combinations of the above

Rating college football teams

2004 regular season

Auburn		Oklahoma		USC		Utah	
opp	score	opp	score	opp	score	opp	score
La-M	31 - 0	BGSU	40 - 24	VaTech(N)	24 - 13	TexAM	41 - 21
@Miss St	43 - 14	Houston	63 - 13	Colo St	49 - 0	@Arizona	23 - 6
LSU	10 - 9	Oregon	31 - 7	@BYU	42 - 10	@Utah St	48 - 6
Citadel	33 - 3	TexTech	28 - 13	@Stanford	31 - 28	AF	49 - 35
@Tenn	34 - 10	Texas(N)	12 - 0	Cal	23 - 17	@NewMex	28 - 7
LaTech	52 - 7	@Kans St	31 - 21	Ariz St	45 - 7	UNC	46 - 16
Ark	38 - 20	Kansas	41 - 10	Wash	38 - 0	UNLV	63 - 28
Kent	42 - 10	@Okl St	38 - 35	@Wash St	42 - 12	@SD State	51 - 28
@Miss	35 - 14	@TexAM	42 - 35	@Oreg St	28 - 20	Colo St	63 - 31
Georgia	24 - 6	Nebraska	30 - 3	Arizona	49 - 9	@Wyom	45 - 28
@Ala	21 - 13	@Baylor	35 - 0	ND	41 - 10	BYU	52 - 21
Tenn(N)	38 - 28	Colo(N)	42 - 3	@UCLA	29 - 24		
12 - 0	401 - 134	12 - 0	433 - 164	12 - 0	441 - 150	11 - 0	509 - 227

The BCS

- BCS agreements among conferences/bowls:
  6 major conference champions and 4 at-large teams go to 5 predesignated "major" bowls

- BCS rating system designed to choose the top 2 teams (and "guide" selection of the at-large teams)

- Original BCS (1997) - 4 components
  1. human polls (AP writers, ESPN/USAT coachs)
  2. computer ratings (6 to 8 over the years)
  3. strength of schedule (based on NCAA's RPI)
  4. games lost

- Many changes to BCS rating system over time

BCS 2004 (Dec 2004)

BCS Rank	Associated Press			ESPN/USAToday			Computer Rankings							BCS
	Rk	Pts	Pct	Rk	Pts	Pct	AH	Bi	Co	Ma	Sa	Wo	Pct	Avg
1. USC	1	1599	.9840	1	1490	.9770	24	24	25	25	24	24	.97	.9770
2. OKL	2	1556	.9575	2	1459	.9567	25	25	24	24	25	25	.99	.9681
3. AUB	3	1525	.9385	3	1435	.9410	23	23	23	23	23	23	.92	.9331
4. TEX	6	1337	.8228	5	1281	.8400	21	22	22	22	22	22	.88	.8476
5. CAL	4	1399	.8609	4	1286	.8433	20	18	20	20	21	20	.80	.8347
6. UTA	5	1345	.8277	6	1215	.7967	22	20	21	21	20	21	.83	.8181
7. GA	8	1117	.6874	7	1117	.7325	17	19	18	17	15	15	.67	.6966
8. VPI	9	1111	.6837	9	1037	.6800	13	15	14	18	18	18	.65	.6712
9. BOI	10	960	.5908	10	943	.6184	19	21	19	19	19	19	.76	.6564
10. LOU	7	1183	.7280	8	1066	.6990	9	12	13	11	17	16	.52	.6490

2004 Least squares results

rnk	max = ∞		max = 40		max = 20		max = 1	
1	USC	53.8	USC	51.1	USC	32.8	USC	2.00
2	Cal	51.5	Cal	49.6	Okl	32.3	Okl	1.95
3	Okl	49.2	Okl	46.2	Cal	32.2	Tex	1.80
4	Lou	48.7	Uta	44.0	Uta	31.9	Cal	1.78
5	Uta	45.6	Lou	43.6	Tex	28.8	Uta	1.75
6	Tex	44.3	Tex	40.1	Aub	27.4	Aub	1.70
7	Mia	43.2	Mia	39.6	Lou	27.2	ASU	1.59
8	VPI	42.9	Aub	38.4	Mia	26.6	Boi	1.58
9	Aub	40.1	VPI	37.0	Virg	25.2	Iowa	1.48
10	Boi	39.5	ASU	35.6	FSU	25.0	TAM	1.46

Least squares ratings
A study

- Used least squares to predict professional
  football games (not enough college data)

- 9 seasons, 2nd half of season only,
  1027 games in all

- Algorithm: rate teams, predict a set of games,
  update ratings, predict next set, . . .

- Keep track of:
  - percent of outcomes predicted correctly
  - accuracy of score predictions

Least squares ratings
Pro football study

| Rating method	Percent correct	Typical error (RMS)
Least Squares (LS)	63.6	14.5
L S - reduce blowouts	63.5	14.4
L S - reduce blowouts 10 pt win bonus	62.9	—
L S - reduce blowouts weight recent gms	63.6	14.4

L S - reduce blowouts
     10 pt win bonus
     weight recent gms
	62.5	—

Las Vegas (LV) point spread	66.3	13.9

The value of information

- Compare four prediction strategies including least squares ratings
  - always predict a home win (not much to do with todays subject)
  - least squares with 1/-1 outcome
  - least squares with score
  - oddsmakers (use all info)

- Table below (number of games in parentheses)

Sport	Home	LS (1/-1)	LS (scores)	Oddsmaker
NFL 88-93	.58 (1342)	.63 (862)	.65 (862)	.67 (1303)
NBA 85-86	.66 (1886)	.69 (1255)	.70 (1255)	.71 (1827)
MLB 86	.53 (3884)	.56 (643)	.56 (643)	.55 (938)
NCAA F 92-96	.62 (3038)	.69 (2104)	.73 (2104)	.75 (1551)
NCAA B 95	.68 (2258)	.72 (1724)	.72 (1724)	.75 (2068)
NCAA H 91-92	.62 (1405)	.72 (998)	.73 (998)	—

The value of information

- Previous results displayed graphically
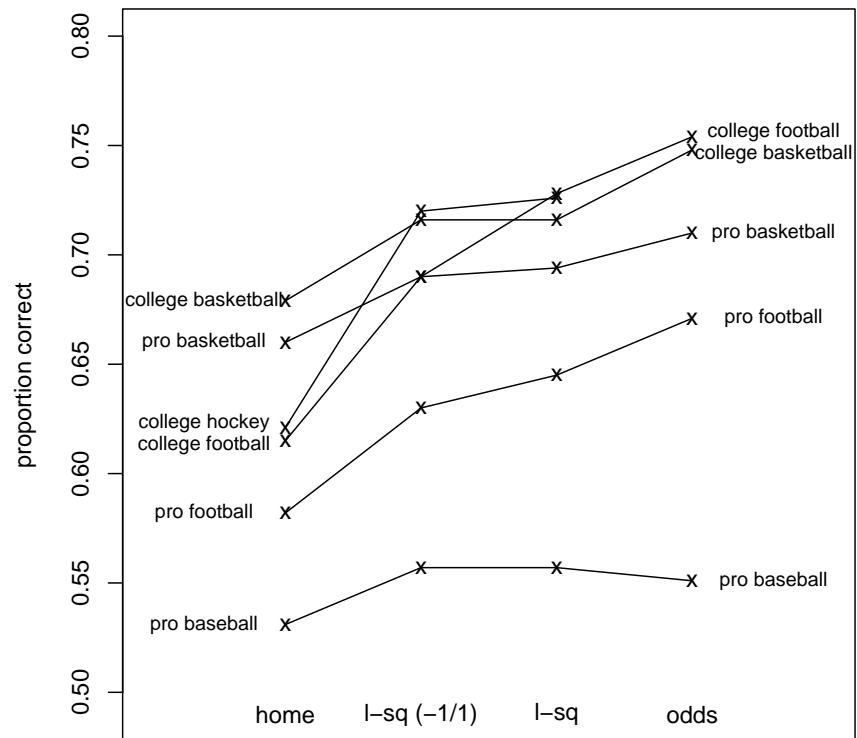
The value of even more information

- Note there is a great deal of information from previous seasons

- The same teams tend to be good each year

- Good for prediction but would this be legitimate in rating teams

- Some NFL results (σ^2 controls year-to-year stability with large σ^2 meaning less year-to-year stability in rating)

Rating method	Percent correct	Typical error (RMS)
Las Vegas (LV) point spread	66.3	13.9
Least Squares (LS) ($\sigma_s^2 = \infty$)	63.6	14.5
State-space ($\sigma_s^2 = 100$)	63.9	14.3
State-space ($\sigma_s^2 = 9$)	64.9	14.1
State-space ($\sigma_s^2 = 6.25$)	64.9	14.1
State-space ($\sigma_s^2 = 1$)	62.0	14.5

Efficiency of betting markets

- Previous results on rating teams all point to performance of Las Vegas odds as an upper limit

- Are sports betting markets efficient?

- Series of results suggest that it is "weakly" efficient (some data follows)

- This efficiency result leads to some interesting uses of point spread data

Baseball odds / outcomes

Favorite's Odds	Estimated probability favorite wins	Games played	Games won by favorite	Observed proportion won by favorite
-110	0.51	89	41	0.46
-115	0.52	98	56	0.57
-120	0.53	71	32	0.45
-125	0.54	82	47	0.57
-130	0.55	70	32	0.46
-135	0.56	80	41	0.51
-140	0.57	75	43	0.57
-145	0.58	79	41	0.52
-150	0.59	54	34	0.63
-155	0.60	58	29	0.50
-160 to -165	0.61	67	41	0.61
-170 to -185	0.63	51	36	0.71
-190 to -300	0.67	63	43	0.68

Data are 969 National League baseball games from 1986.

Football odds / outcomes

Point spread P	Games played	Proportion won by favorite	Proportion favorite beat spread
0.0	85	–	–
1.0	141	0.52	0.51
1.5	128	0.47	0.46
2.0	201	0.55	0.51
2.5	231	0.51	0.44
3.0	386	0.62	0.52
3.5	284	0.57	0.47
4.0	200	0.63	0.48
4.5	110	0.67	0.50
5.0	145	0.69	0.50
5.5	111	0.78	0.55
6.0	186	0.65	0.44
6.5	198	0.71	0.50
7.0	209	0.73	0.50
8.0	75	0.80	0.49
9.0	106	0.75	0.41
10.0	82	0.74	0.43
> 10	307	0.83	0.48

- Data from 3408 NFL games, 1981-1997 (except 1982, 1986)

- Similar results for college/pro basketball

An interesting distributional result

Consider difference between game outcome and pointspread

Year	Number of games	Mean point spread error	Std.dev. point spread error
1981	224	−0.7	14.1
1982	224	−0.4	13.8
1984	224	1.3	13.6
1985	224	1.1	13.5
1986	224	−0.5	13.8
1988	224	−0.3	14.3
1989	224	0.7	13.6
1990	224	0.9	13.6
1991	224	0.8	12.6
1992	224	−0.7	13.9
1993	224	−1.0	13.1
1994	224	−1.4	12.3
1995	240	−1.1	12.5
1996	240	−0.5	12.9
1997	240	−1.6	13.3
Total	3408	−0.24	13.42

Margin of victory over the pointspread
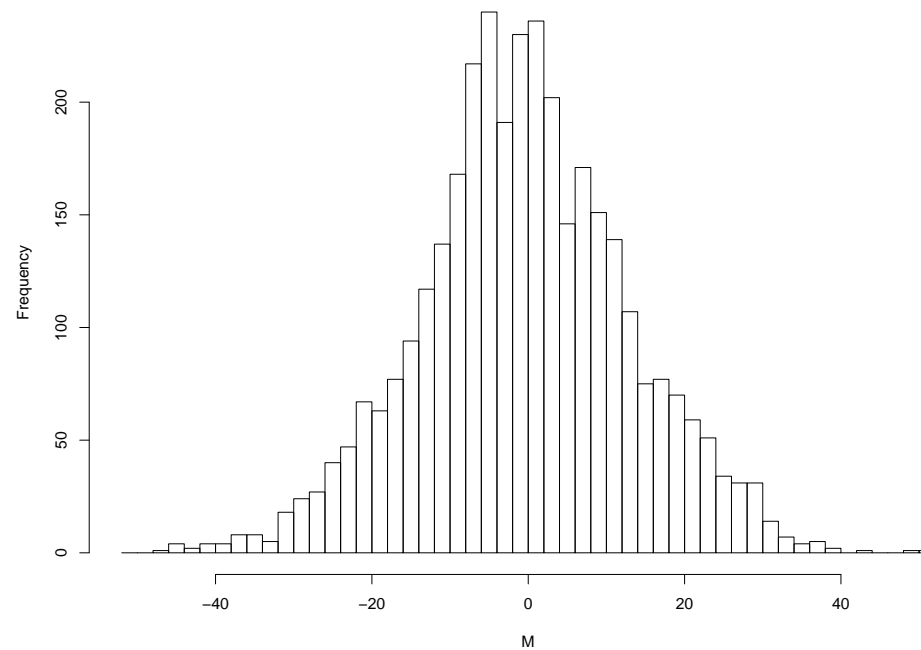
- Outcome minus pointspread is approx $N(0, 13.5^2)$ (pro football)

- Similar results in other sports
  - s.d. in pro basketball = 11.6 (2 years of data)
  - s.d. in college basketball = 11.0 (1 year of data)

Applications of the normal distribution

- Using pro football as an example, if we accept normal distribution then
  - can simulate game outcomes with P(p-point favorite wins) $= \Phi(p/13.5)$
  - home advantage $= 3$ points .... implies home team wins 59%
  - teaser bet gives you 6 points .... should win 67% of such bets
  - Super Bowl 2008 - NE Patriots 12 point favorite .... P(win) $= .81$
    Moneyline odds NE -450 imply P(win) $= .818$
    Moneyline odds NYG +350 imply P(win) $= .778$

- Normal distribution can also tell us about comebacks/reversals

On the probability of winning

- Occasionally see amazing comebacks or hear interesting statistics
  - NBA: NY trails Milwaukee by 18 pts with 6 minutes to play and scores the last 19 pts to win (Nov 18, 1972)
  - SF Chronicle (1985): Pr(NBA team ahead at half wins) = 0.80

- What can statistics tell us about such facts?

- Some basic data: probability of winning for a team that is ahead after

	Football (154 gms) 1993	Basketball (493 gms) 1991	Baseball (962 gms) 1986
1 qtr (2 inn)	0.67	0.67	0.71
2 qtr (5 inn)	**0.76**	**0.75**	0.81
3 qtr (7 inn)	0.85	0.82	0.89

On the probability of winning

- Notice that for basketball and football the
  probability of winning given that a team
  is ahead at halftime is 0.75
  (SF Chronicle 1985 gave 0.80)

- Cover: a simple argument for the probability of winning if ahead at halftime
  - assume 2 independent identically distributed halves
  - A wins both halves - 1/4 of the time
  - B wins both halves - 1/4 of the time
  - A and B each win one half - 1/2 of the time
  - in the last case assume the team that wins
    the first half wins the game half the time
  - then Pr(team ahead at half wins) = 0.75

- But previous table and this argument ignores important information ...
  the score!

On the probability of winning

- Have seen a normal distribution model for game outcomes (with mean equal to pointspread)

- Does this apply during the game?

- Some NBA data:
  - 493 games of '91-'92 season (no pointspreads available)
  - Summaries for distribution of home - visitor margin

	mean	s.d.
1st qtr	1.4	7.6
2nd qtr	1.6	7.4
3rd qtr	1.5	7.3
4th qtr	0.2	7.0
Total	4.6	13.2

  - Data are nearly normal for each quarter
  - Quarters are slightly negatively correlated
  - Suggests a kind of random walk with drift in favor of home team

On the probability of winning
A normal random walk model model

- Model:
  - let $X(t) = $ lead of p-point favorite at time t
    (where assume t is scaled to run from 0 to 1)
  - assume $X(t)$ has a normal distribution with mean pt and std.dev. $\sigma\sqrt{t}$
  - assume results beyond time t don't depend on what happened earlier in the game
  - can calculate

$$
\begin{aligned}
P(l, t) &= \text{Probability } p\text{-point favorite wins} \\
& \quad \text{given a lead } l \text{ at time } t \\[2mm]
&= \Phi\left(\frac{l + (1-t)p}{\sqrt{(1-t)\sigma^2}}\right)
\end{aligned}
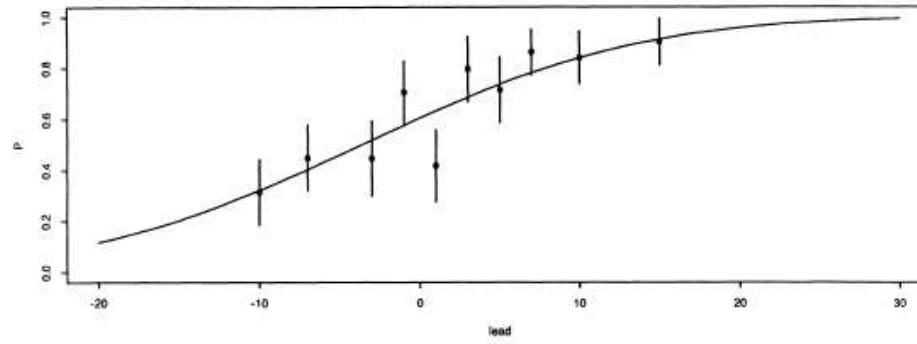$$

  where Φ is the standard normal cdf
  - potentially useful for in-game betting

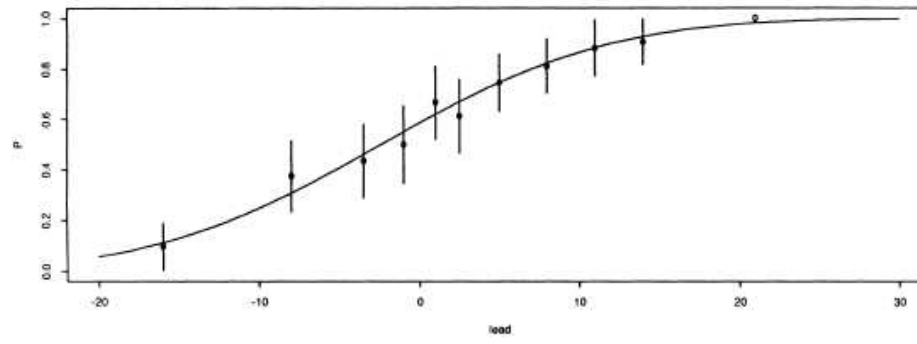## On the probability of winning
Using a normal distribution model

- Probability of winning for a generic home team $(p = 4.5)$

Time elapsed, t	Lead for home team, l						
	-10	-5	-2	0	2	5	10
0.00				.62			
0.25	.32	.46	.55	.61	.66	.74	.84
0.50	.25	.41	.52	.59	.65	.75	.87
0.75	.13	.32	.46	.56	.65	.78	.92
0.90	.03	.18	.38	.54	.69	.86	.98
1.00	.00	.00	.00	.5?	1.00	1.00	1.00

time = 0.25

time = 0.50

time = 0.75
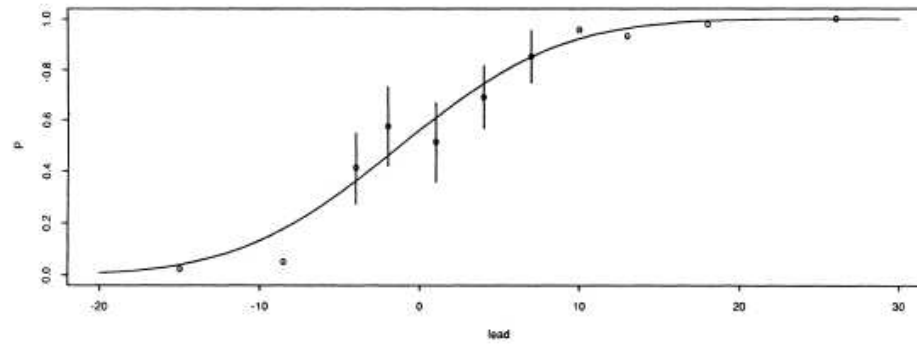
P(l,t) versus t for basketball data

lead = 12

lead = 10

lead = 7

lead = 5

lead = 2

lead = 0

lead = −2

lead = −5

lead = −10   lead = −7

lead = −12

P

time
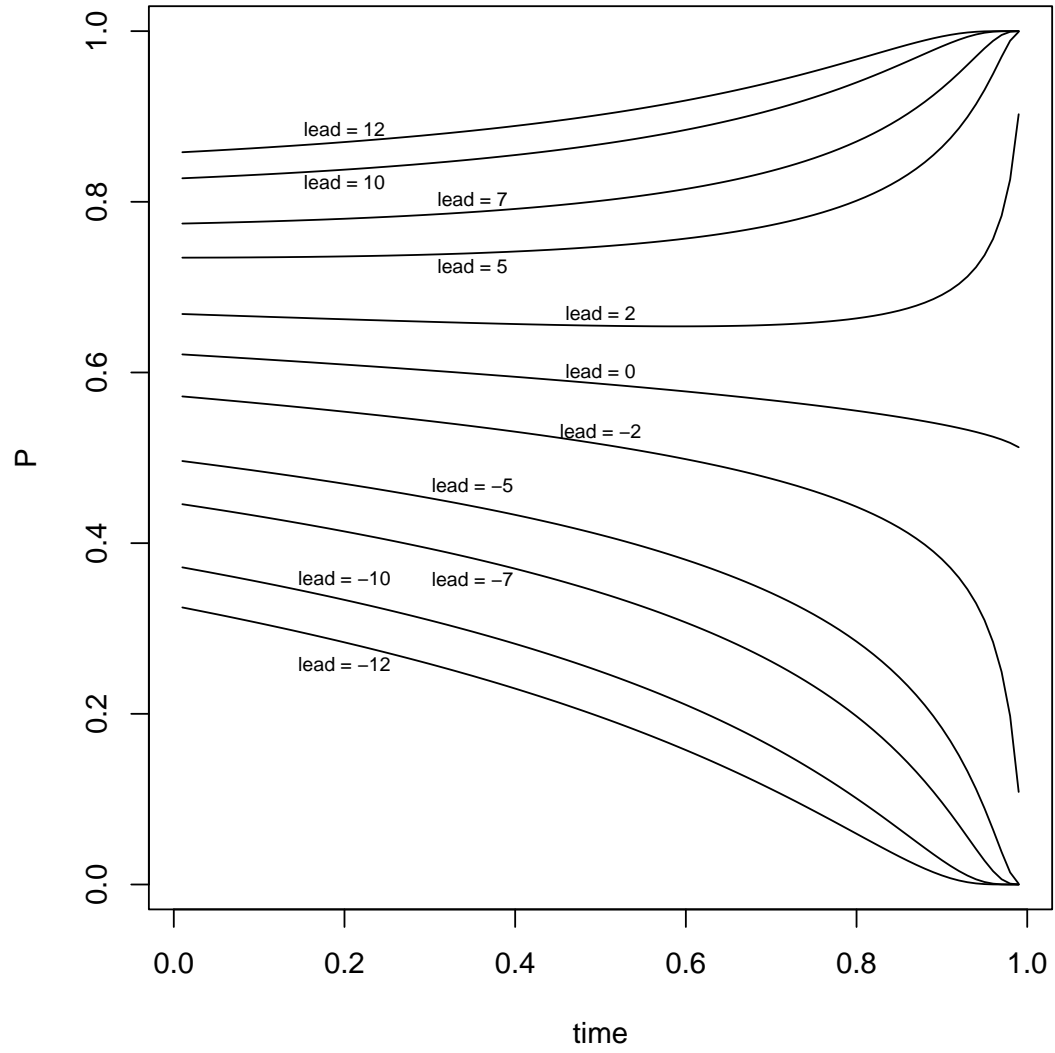
Summary / References

- Thinking carefully (and statistically) about sports information helps avoid some common misunderstandings that arise from randomness and/or small samples

- Some reference
  - 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
  - Statistical Theory in Sports - 2007 book
    edited by Albert and Koning
  - Chance magazine (column beginning in vol. 10)
  - Journal of Quantitative Analysis in Sports