Lecture 3: Sports rating models

David Aldous

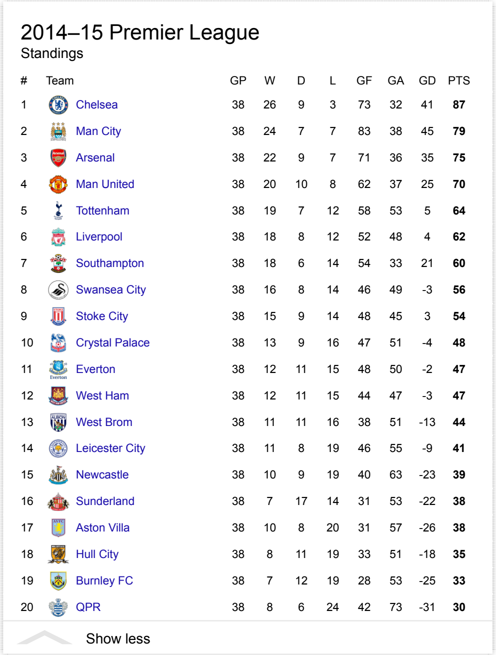
January 27, 2016

- **Sports** are a popular topic for course projects – usually involving details of some specific sport and statistical analysis of data.
- In this lecture we imagine some non-specific sport; either a team sport – *(U.S.) football, baseball, basketball, hockey; soccer cricket* – or an individual sport or game – *tennis, chess, boxing*.
- We consider only sports with matches between two teams/individuals. But similar ideas work where there are many contestants – *athletics, horse racing, automobile racing, online video games*.

Let me remind you of three things you already know about sports.

Reminder 1. Two standard "centralized" ways to schedule matches: league or tournament.

[redacted]



2014–15 Premier League

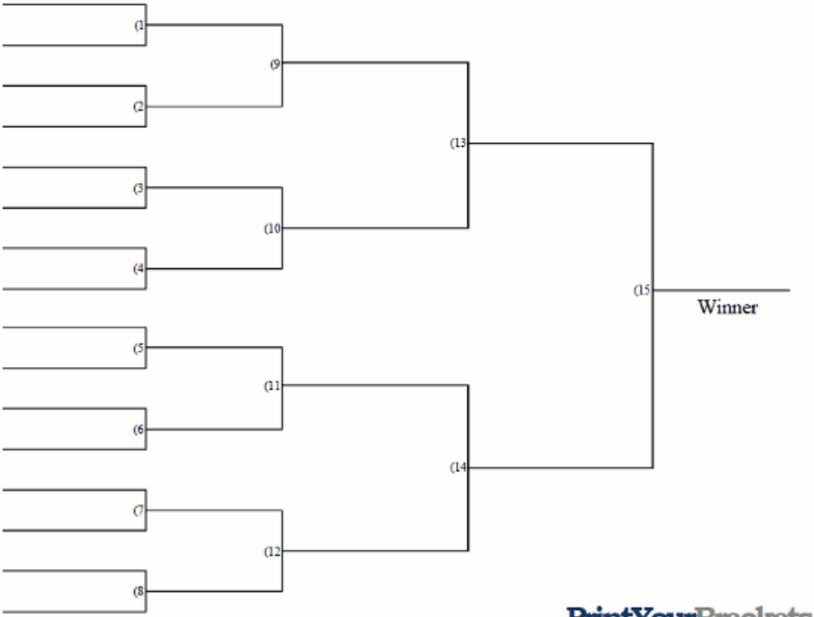
Standings

#	Team	GP	W	D	L	GF	GA	GD	PTS
1	Chelsea	38	26	9	3	73	32	41	**87**
2	Man City	38	24	7	7	83	38	45	**79**
3	Arsenal	38	22	9	7	71	36	35	**75**
4	Man United	38	20	10	8	62	37	25	**70**
5	Tottenham	38	19	7	12	58	53	5	**64**
6	Liverpool	38	18	8	12	52	48	4	**62**
7	Southampton	38	18	6	14	54	33	21	**60**
8	Swansea City	38	16	8	14	46	49	-3	**56**
9	Stoke City	38	15	9	14	48	45	3	**54**
10	Crystal Palace	38	13	9	16	47	51	-4	**48**
11	Everton	38	12	11	15	48	50	-2	**47**
12	West Ham	38	12	11	15	44	47	-3	**47**
13	West Brom	38	11	11	16	38	51	-13	**44**
14	Leicester City	38	11	8	19	46	55	-9	**41**
15	Newcastle	38	10	9	19	40	63	-23	**39**
16	Sunderland	38	7	17	14	31	53	-22	**38**
17	Aston Villa	38	10	8	20	31	57	-26	**38**
18	Hull City	38	8	11	19	33	51	-18	**35**
19	Burnley FC	38	7	12	19	28	53	-25	**33**
20	QPR	38	8	6	24	42	73	-31	**30**

Barclays Premier League table, current & previous standings

www.**premierleague**.com/en-gb/.../**league-table**.html ▾ Premier League ▾

16 Team Single Elimination



PrintYourBrackets.com

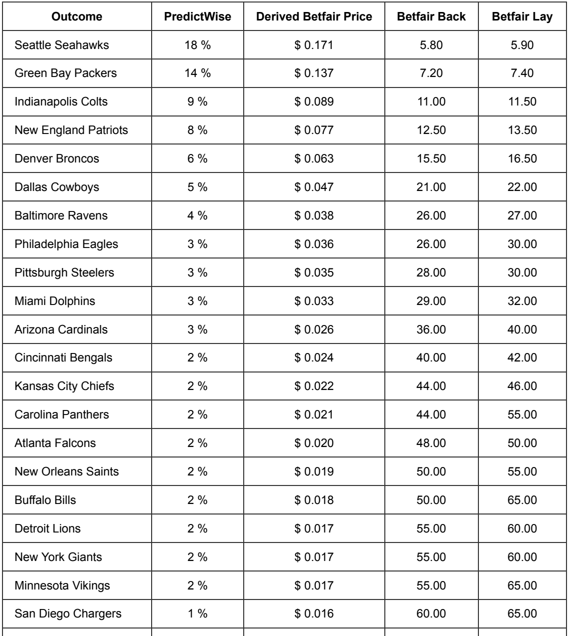
These schemes are clearly “fair” and produce a “winner”, though have two limitations

- Limited number of teams.
- Start anew each year/tournament.
- Require central organization – impractical for for games (*chess, tennis*) with many individual contestants.

Reminder 2. In most sports the winner is decided by *point difference*. One could model point difference but we won't. For simplicity we will assume matches always end in win/lose, no ties.

Reminder 3. A main reason why sports are interesting is that the outcome is uncertain. It makes sense to consider the **probability** of team A winning over team B. In practice one can do this by looking at gambling odds [next slide]. Another lecture will discuss data and theory concerning how probabilities derived from gambling odds change over time.

Winner of 2015-16 season Superbowl – gambling odds at start of season.



Outcome	PredictWise	Derived Betfair Price	Betfair Back	Betfair Lay
Seattle Seahawks	18 %	$ 0.171	5.80	5.90
Green Bay Packers	14 %	$ 0.137	7.20	7.40
Indianapolis Colts	9 %	$ 0.089	11.00	11.50
New England Patriots	8 %	$ 0.077	12.50	13.50
Denver Broncos	6 %	$ 0.063	15.50	16.50
Dallas Cowboys	5 %	$ 0.047	21.00	22.00
Baltimore Ravens	4 %	$ 0.038	26.00	27.00
Philadelphia Eagles	3 %	$ 0.036	26.00	30.00
Pittsburgh Steelers	3 %	$ 0.035	28.00	30.00
Miami Dolphins	3 %	$ 0.033	29.00	32.00
Arizona Cardinals	3 %	$ 0.026	36.00	40.00
Cincinnati Bengals	2 %	$ 0.024	40.00	42.00
Kansas City Chiefs	2 %	$ 0.022	44.00	46.00
Carolina Panthers	2 %	$ 0.021	44.00	55.00
Atlanta Falcons	2 %	$ 0.020	48.00	50.00
New Orleans Saints	2 %	$ 0.019	50.00	55.00
Buffalo Bills	2 %	$ 0.018	50.00	65.00
Detroit Lions	2 %	$ 0.017	55.00	60.00
New York Giants	2 %	$ 0.017	55.00	60.00
Minnesota Vikings	2 %	$ 0.017	55.00	65.00
San Diego Chargers	1 %	$ 0.016	60.00	65.00

Obviously the probability A beats B depends on the **strengths** of the teams – a better team is likely to beat a worse team. So the problems

- estimate the strengths of A and B
- estimate the probability that A will beat B

must be closely related. This lecture talks about two ideas for making such estimates which have been well studied. However the connection between them has not been so well studied, and is suitable for simulation-style projects.

Terminology. I write **strength** for some hypothetical objective numerical measure of how good a team is – which we can't observe – and **rating** for some number we can calculate by some formula based on past match results. Ratings are intended as estimates of strengths.

Idea 1: The basic probability model.

Each team A has some "strength" x_A, a real number. When teams A and B play

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

for a specified "win probability function" W satisfying the conditions

$$\begin{gathered} W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing} \\ W(-x) + W(x) = 1; \quad \lim_{x \to \infty} W(x) = 1. \end{gathered} \tag{1}$$

Implicit in this setup, as mentioned before

- each game has a definite winner (no ties);
- no home field advantage, though this is easily incorporated by making the win probability be of the form $W(x_A - x_B \pm \Delta)$;
- not considering more elaborate modeling of point difference

and also

- strengths do not change with time.

Some comments on the math model.

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

$$W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing}$$

$$W(-x) + W(x) = 1; \quad \lim_{x\to\infty} W(x) = 1.$$

There is a reinterpretation of this model, as follows. Consider the alternate model in which the winner is determined by point difference, and suppose the random point difference D between two teams of equal strength has some (necessarily symmetric) continuous distribution not depending on their common strength, and then suppose that a difference in strength has the effect of increasing team A's points by $x_A - x_B$. Then in this alternate model

$$\mathbb{P}(\text{A beats B}) = \mathbb{P}(D+x_A-x_B \geq 0) = \mathbb{P}(-D \leq x_A-x_B) = \mathbb{P}(D \leq x_A-x_B).$$

So this is the same as our original model in which we take W as the distribution function of D.

This basic probability model has undoubtedly been re-invented many times; in the academic literature it seems to have developed "sideways" from the following type of statistical problem. Suppose we wish to rank a set of movies $A, B, C, \ldots$ by asking people to rank (in order of preference) the movies they have seen. Our data is of the form
(person 1): C, A, E
(person 2): D, B, A, C
(person 3): E, D
. . . .
One way to produce a consensus ranking is to consider each pair (A, B) of movies in turn. Amongst the people who ranked both movies, some number $i(A, B)$ preferred A and some number $i(B, A)$ preferred B. Now reinterpret the data in sports terms: team A beat B $i(A, B)$ times and lost to team B $i(B, A)$ times. Within the basic probability model (with some specified W) one can calculate MLEs of strengths $x_A, x_B, \ldots$ which imply a ranking order.

This method, with W the logistic function (discussed later), is called the *Bradley-Terry* model, from the 1952 paper *Rank analysis of incomplete block designs: I. The method of paired comparisons* by R.A. Bradley and M.E. Terry.

An account of the basic Statistics theory (MLEs, confidence intervals, hypothesis tests, goodness-of-fit tests) is treated in Chapter 4 of H.A. David's 1988 monograph *The Method of Paired Comparisons*.

So one can think of **Bradley-Terry as a sports model** as follows: take data from some past period, calculate MLEs of strengths, use to predict future win probabilities.

Considering Bradley-Terry as a sports model:

positives:

- allows unstructured schedule;
- use of logistic makes algorithmic computation straightforward.

negatives:

- use of logistic completely arbitrary: asserting

$$\text{if } \mathbb{P}(i \text{ beats } j) = 2/3,\ \mathbb{P}(j \text{ beats } k) = 2/3 \text{ then } \mathbb{P}(i \text{ beats } k) = 4/5$$

as a universal fact seems ridiculous;

- by assuming unchanging strengths, it gives equal weight to past as to recent results;
- need to recompute MLEs after each match.

The Bradley-Terry model could be used for interesting course projects – take the Premier league data and ask *what is the probability that Chelsea was actually the best team in 2014-15?*.

Reminder 4. Another aspect of what makes sports interesting to a spectator is that strengths of teams change over time – if your team did poorly last year, then you can hope it does better this year.

In the context of the Bradley-Terry model, one can extend the model to allow changes in strengths. Seem to be about 2-3 academic papers per year which introduce some such extended model and analyze some specific sports data. Possible source of course projects – apply to different sport or to more recent data.

Idea 2: Elo-type rating systems [show site]

(not ELO). The particular type of rating systems we study are known loosely as Elo-type systems and were first used systematically in chess. The Wikipedia page *Elo rating system* is quite informative about the history and practical implementation. What I describe here is an abstracted "mathematically basic" form of such systems.

Each player i is given some initial rating, a real number y_i. When player i plays player j, the ratings of both players are updated using a function Υ (`Upsilon`)

$$\begin{aligned} &\text{if } i \text{ beats } j \text{ then } y_i \to y_i + \Upsilon(y_i - y_j) \text{ and } y_j \to y_j - \Upsilon(y_i - y_j) \\ &\text{if } i \text{ loses to } j \text{ then } y_i \to y_i - \Upsilon(y_j - y_i) \text{ and } y_j \to y_j + \Upsilon(y_j - y_i)\ . \end{aligned} \tag{2}$$

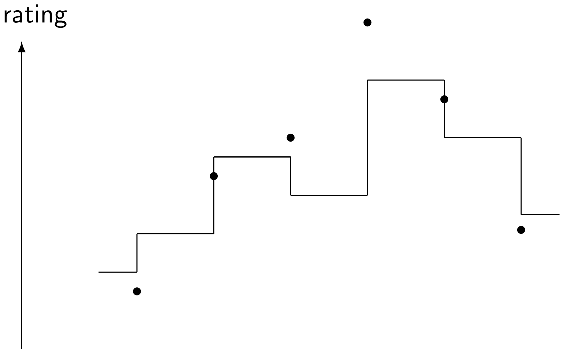
Note that the sum of all ratings remains constant; it is mathematically natural to center so that this sum equals zero.

The Elo ratings are based on the following formulas:

$R_n = R_o + K \times (W - W_e)$

- R_n is the new rating, R_o is the old (pre-match) rating.
- **K** is the weight constant for the tournament played:
- **60** for World Cup finals;
- **50** for continental championship finals and major intercontinental tournaments;
- **40** for World Cup and continental qualifiers and major tournaments;
- **30** for all other tournaments;
- **20** for friendly matches.
- **K** is then adjusted for the goal difference in the game. It is increased by **half** if a game is won by two goals, by **3/4** if a game is won by three goals, and by **3/4 + (N-3)/8**if the game is won by four or more goals, where **N** is the goal difference.
- **W** is the result of the game (**1** for a win, **0.5** for a draw, and **0** for a loss).
- W_e is the expected result (win expectancy), either from the chart or the following formula:
- $W_e = 1 / (10^{(-dr/400)} + 1)$
- **dr** equals the difference in ratings plus **100** points for a team playing at home.

Schematic of one player's ratings after successive matches. The ● indicate each opponent's rating.



Math comments on the Elo-type rating algorithm.
We require the function $\Upsilon(u), \ -\infty < u < \infty$ to satisfy the qualitative conditions

$$\Upsilon : \mathbb{R} \to (0, \infty) \text{ is continuous, strictly decreasing, and } \lim_{u \to \infty} \Upsilon(u) = 0. \tag{3}$$

We will also impose a quantitative condition

$$\kappa_\Upsilon := \sup_u |\Upsilon'(u)| < 1. \tag{4}$$

To motivate the latter condition, the rating updates when a player with (variable) strength x plays a player of fixed strength y are

$$x \to x + \Upsilon(x - y) \text{ and } x \to x - \Upsilon(y - x)$$

and we want these functions to be *increasing* functions of the starting strength x.

Note that if Υ satisfies (3) then so does $c\Upsilon$ for any scaling factor $c > 0$. So given any Υ satisfying (3) with $\kappa_\Upsilon < \infty$ we can scale to make a function where (4) is satisfied.

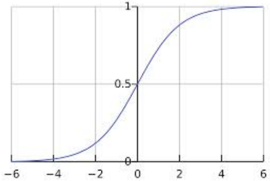
The logistic distribution function

$$F(x) := \frac{e^x}{1 + e^x}, -\infty < x < \infty$$

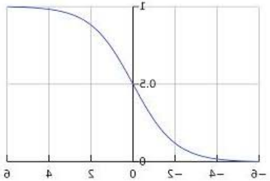
is a common choice for the “win probability” function $W(x)$ in the basic probability model; and its complement

$$1 - F(x) = F(-x) = \frac{1}{1 + e^x}, -\infty < x < \infty$$

is a common choice for the “update function shape” $\Upsilon(x)$ in Elo-type rating systems. That is, one commonly uses $\Upsilon(x) = cF(-x)$.



possible $W(x)$



possible $\Upsilon(x)$

Whether this is more than a convenient choice is a central issue in this topic.

Elo is an algorithm for producing ratings (and therefore rankings) which (unlike Bradley-Terry) does not assume any probability model. It implicitly attempts to track changes in strength and puts greater weight on more recent match results.

How good are Elo-type algorithms? This is a subtle question – we need to

- use Elo to make predictions
- choose how to measure their accuracy
- compare accuracy with predictions from some other ranking/rating scheme (such as Bradley-Terry or gambling odds).

The simplest way to compare schemes would be to look at matches where the different schemes ranked the teams in opposite ways, and see which team actually won. But this is not statistically efficient [board]. Better to compare schemes which predict **probabilities**.

Although the Elo algorithm does not say anything *explicitly* about probability, we can argue that it *implicitly* does predict winning probabilities.

A math connection between the probability model and the rating algorithm.

Consider n teams with unchanging strengths $x_1, \ldots, x_n$, with match results according to the basic probability model with win probability function W, and ratings (y_i) given by the update rule with update function Υ. When team i plays team j, the expectation of the rating change for i equals

$$\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i). \tag{5}$$

So consider the case where the functions Υ and W are related by

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

In this case

> *(*) If it happens that the difference $y_i - y_j$ in ratings of two players playing a match equals the difference $x_i - x_j$ in strengths then the expectation of the change in rating difference equals zero*

whereas if unequal then (because Υ is decreasing) the expectation of $(y_i - y_j) - (x_i - x_j)$ is closer to zero after the match than before.

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty. \tag{6}$$

These observations suggest that, under relation (6), there will be a tendency for player i's rating y_i to move towards its strength x_i though there will always be random fluctuations from individual matches. So if we believe the basic probability model for some given W, then in a rating system we should use an Υ that satisfies (6).

Recall that in the probability model we can center the strengths so that $\sum_i x_i = 0$, *and similarly we will initialize ratings so that* $\sum_i y_i = 0$.

What is the solution of (6) for unknown Υ?

This can be viewed as the setup for a mathematician/physicist/statistician joke.

`Problem`

$$\text{solve } \Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

`Solution`

- physicist (Elo): $\Upsilon(u) = cW(-u)$
- mathematician: $\Upsilon(u) = W(-u)\phi(u)$ for arbitrary symmetric $\phi(\cdot)$.
- statistician: $\Upsilon(u) = c\sqrt{W(-u)/W(u)}$ (variance-stabilizing ϕ).

These answers are all "wrong" for different reasons. And so in fact it's hard to answer "what Υ to use?"

However we can work backwards; when people use the "complement of logistic" as the update function Υ, we can use (6) to argue that they are implicitly imagining the Bradley-Terry with W the logistic function. So this gives a way to use Elo ratings to predict win probabilities.

There is a link to my more mathematical write-up of the topic above.

There are several related possible **simulation projects**. For instance, consider a league on n teams, whose strengths has some SD σ, the strengths change in time via some rule with rate parameter λ. We take the probability model with logistic W and the update model with function $c\Upsilon$ for logistic Υ. Study how the optimal value of c depends on (n, σ, λ).

Some other aspects of rating models.

1. Recent book "The Science of Ranking and Rating" treats methods using undergraduate linear algebra. The lecture in this course in 2014 was based more on that book (link on web page).

2. People who attempt realistic models of particular sports, using e.g. statistics of individual player performance, believe their models are much better than general-sport methods based only on history of wins/losses or point differences. But a recent paper *Statistics-free sports prediction* claims that (using more complex prediction schemes) they can do almost as well using only match scores.

3. I have talked about comparing different schemes which predict **probabilities** – after we see the actual match results, how do we decide which scheme is better? I will discuss this in a different context, the lecture on Geopolitics forecasting.

4. Both schemes are poor at assessing new players. Xbox Live uses its "TrueSkill ranking system" [show page] which estimates both a rating and the uncertainty in the rating, as follows.

$\$59^{99}$
Estimated Retail Price
XBOX
HALO 4
RATING PENDING
RP
May contain content inappropriate for children. Visit www.esrb.org for rating information.
Online Interactions Not Rated by the ESRB
Halo® 4 sold separately.
*See back for details.
Microsoft Studios
343
12 MONTH
XBOX LIVE GOLD MEMBERSHIP
+1 BONUS MONTH
Play Halo 4 online with friends with 13 months of Xbox LIVE Gold.*

Here a rating for player i is a pair (μ_i, σ_i), and the essence of the scheme is as follows. When i beats j
(i) first compute the conditional distribution of X_i given $X_i > X_j$, where X_i has Normal(μ_i, σ_i^2) distribution
(ii) then update i's rating to the mean and s.d. of that conditional distribution.
Similarly if i loses to j then i's rating is updated to the mean and s.d. of the conditional distribution of X_i given $X_i < X_j$.

Discussion. The authors seem to view this as an approximation to some coherent Bayes scheme, but to me it fails to engage both ‘‘uncertainty about strength" and ‘‘uncertainty about match outcome".

So another simulation project is to compare this to other schemes. Note this implicitly predicts winning probabilities via $\mathbb{P}(X_i > X_j)$.

[show slides from previous year lecture]

People often think that bookmakers adjust their offered odds so that, whatever the outcome, they never lose money. This just isn't true. [show ESPN article]

Finally, there is an interesting paradox involved in designing matches to be exciting to spectators [board].