Foundational questions about sports rating models

David Aldous

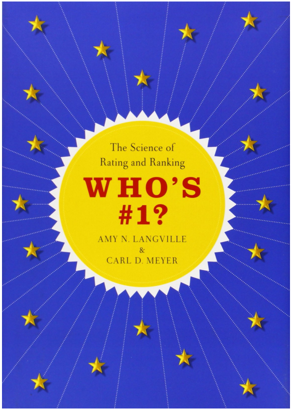
24 March 2015

This is a lightweight talk – no hard theorems – my purpose is to formulate questions.

Centered on two very simple ideas, which have been developed in certain directions, but with other aspects that no-one seems to have thought about . . . . . .

Inspired by a lecture in my "Probability in the Real World" course, in which I give 20 lectures on maximally different topics (mostly) not taught in other courses.

- Everyday perception of chance
- Ranking and rating
- Risk to individuals: perception and reality
- Luck
- A glimpse at probability research: spatial networks on random points
- Prediction markets, fair games and martingales
- Science fiction meets science
- Coincidences, near misses and one-in-a-million chances.
- Psychology of probability: predictable irrationality
- Mixing: physical randomness, the local uniformity principle and card shuffling
- Game theory
- The Kelly criterion for favorable games: stock market investing for individuals
- Toy models in population genetics: some mathematical aspects of evolution
- Size-biasing, regression effect and dust-to-dust phenomena
- Toy models of human interaction: use and abuse
- Short/Medium term predictions in politics and economics
- Tipping points and phase transitions
- Coding and entropy



This recent book "The Science of Ranking and Rating" treats methods using undergraduate linear algebra – not my cup of tea but a starting place.

Ratings are widely used in games where there is no organized scheduling, in particular online games.


Safari File Edit View History Bookmarks Develop Window Help
Thu 7:37 PM David Aldous
aldousdj in Jungle Gin > Social > Dr. Livingston, I Presume?
monthly.tex
Jungle Gin | Pogo.com Free Online Card Games
www.pogo.com/games/gin#game_gin2-plgin2sf027:table_23
Reader
Club Pogo Apple Yahoo! Google Maps YouTube Wikipedia News Popular
STAND UP
How to Play
tokens
1,043,250
KNOCK
aldousdj gets 28 points
yeloroses1
aldousdj
44
90
Who's Here (2)
Rating
yeloroses1 1446
aldousdj 1519
yeloroses1:
aldousdj:
Your rating has been updated. Your new rating is 1505. Your old rating was 1525.
Your rating has been updated. Your new rating is 1519. Your old rating was 1505.
100 POINT GAME
500 POINT MATCH
yeloroses1 44
aldousdj 90
SCORE summary
My Fav
See all f

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 (minimal natural?) conditions

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

Implicit in this setup:

- 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;
- strengths do not change with time.

Here are 3 comments.

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

Comment 1. There are many reasons one might want to estimate winning probabilities in professional sports:

- gambling
- fantasy sports
- actual sports – Moneyball etc

Our "foundational" model is not seeking to be realistic about estimating probabilities in professional sports – which obviously involves detailed modeling of specific sports – so I view the first three points as "engineering details of a particular sport" which could be incorporated later. But the fourth point – that strengths do in fact change with time – is conceptually important – what makes spectator sports interesting is hope your team does better next year ...

Comment 2. In teaching applied probability or applied stochastic processes we use several classical "toy models" – $M/M/1$ queue, Wright-Fisher, Galton-Watson, ... Why don't we teach this model?

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

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

Note that we can reinterpret the model as follows. Suppose the winner is determined in the usual way 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 the win probability function W can be interpreted as the distribution function of D, because

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

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, 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*.

Google Scholar records the Bradley-Terry paper as *cited by 1544*, though on a small minority are in the sports context.

[Sensory evaluation practices]
H Stone, R Bleibaum, HA Thomas - 2012 - books.google.com
Understanding what the consumer wants and will accept are two of the most significant hurdles faced by anyone in new product development. Whether the concern is the proper mouth-feel of a potato chip, the sense of freshness" evoked by a chewing gum, or the ...
Cited by 2250 Related articles All 10 versions Cite Save More

[Multiway contingency tables analysis for the social sciences]
TD Wickens - 2014 - books.google.com
This book describes the principles and techniques needed to analyze data that form a multiway contingency table. Wickens discusses the description of association in such data using log-linear and log-multiplicative models and defines how the presence of ...
Cited by 481 Related articles All 5 versions Cite Save More

[Mendel 14.0 Documentation]
K Lange, R Cantor, S Horvath, JC Papp, C Sabatti… - 2014 - genetics.ucla.edu
Mendel is a comprehensive package for the statistical analysis of qualitative and quantitative genetic traits. On pedigree data, it internally incorporates both the Elston-Stewart [27, 60] and the Lander-Green-Kruglyak [53, 54] algorithms. In some applications, it will choose ...
Related articles All 7 versions Cite Save More

[Exploratory font selection using crowdsourced attributes]
P O'Donovan, J Lībeks, A Agarwala… - ACM Transactions on …, 2014 - dl.acm.org
Abstract This paper presents interfaces for exploring large collections of fonts for design tasks. Existing interfaces typically list fonts in a long, alphabetically-sorted menu that can be challenging and frustrating to explore. We instead propose three interfaces for font ...
Cited by 1 Cite Save

[Molecular population dynamics of DNA structures in a bcl-2 promoter sequence is regulated by small molecules and the transcription factor hnRNP LL]
Y Cui, D Koirala, HJ Kang, S Dhakal… - Nucleic acids …, 2014 - Oxford Univ Press
Abstract Minute difference in free energy change of unfolding among structures in an oligonucleotide sequence can lead to a complex population equilibrium, which is rather challenging for ensemble techniques to decipher. Herein, we introduce a new method, ...
Cited by 5 Related articles All 6 versions Cite Save

[A Revealed Preference Ranking of US Colleges and Universities*]
CN Avery, ME Glickman, CM Hoxby… - The Quarterly Journal …, 2012 - qje.oxfordjournals.org
Abstract We present a method of ranking US undergraduate programs based on students' revealed preferences. When a student chooses a college among those that have admitted him, that college "wins" his "tournament." Our method efficiently integrates the information ...
Cited by 14 Related articles All 7 versions Cite Save

[The meaning of "near" and "far": the impact of structuring design databases and the effect of distance of analogy on design output]
K Fu, J Chan, J Cagan… - Journal of …, 2013 - … .asmedigitalcollection.asme.org
Design-by-analogy is a practice in which designers use solutions from other domains in order to gain inspiration or insight for the design problem at hand, and has been shown to

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 is ridiculous;

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

Seem to be about 2-3 academic papers per year that (typically) introduce some extended model (typically allowing changes in strengths) and analyze some specific sports data. Useful as source of projects for my students. But rather than doing explicit data analysis, I want to think about how one might use the model in other ways, which apparently have not been studied systematically.

The model clearly can be considered for individual players in sports such as chess, tennis, or boxing, and we write *team* or *player* according to which seems more common in a particular context.

- There is a standard way to design a single-elimination tournament in terms of the seeding (ranking) of players. Within the basic probability model, does this particular design maximize the chance of the highest-ranked player winning the tournament? Does it optimize anything?
- Suppose each of n teams plays each other team once during a season. Each team i wins some number $q^*(i)$ of its matches. Under a given probability distribution over the entire season match results there is some expected number $q(i)$ of matches won by team i. It is not hard to characterize the possible values of the sequence $(q(i), 1 \leq i \leq n)$ for a completely arbitrary probability distribution. But under our basic probability model (with arbitrary W and strengths) what are the possible values?
- At the end of a league season or tournament we can ask *what is the chance the winner was actually the best team?*.

Idea 2: Elo-type rating systems

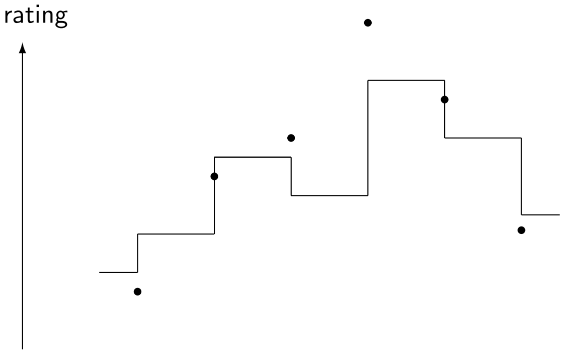
(not ELO). The particular type of rating systems we study are known loosely as Elo-type systems and were first used systematically in chess. The Wikipedia page *Elo rating system* is quite informative about the history and practical implementation. What we describe here is an abstracted "mathematically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

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



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, we want the functions

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

the rating updates when a player with (variable) strength x plays a player of fixed strength y, to be an *increasing* function 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.

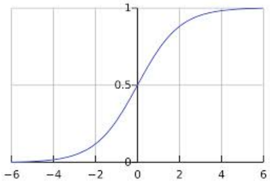
The logistic distribution function

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

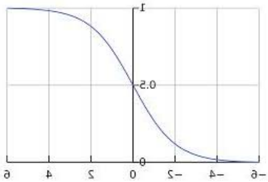
is a common choice for the "win probability" function $W(x)$ in the basic probability model; and its complement

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

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 $W(x)$

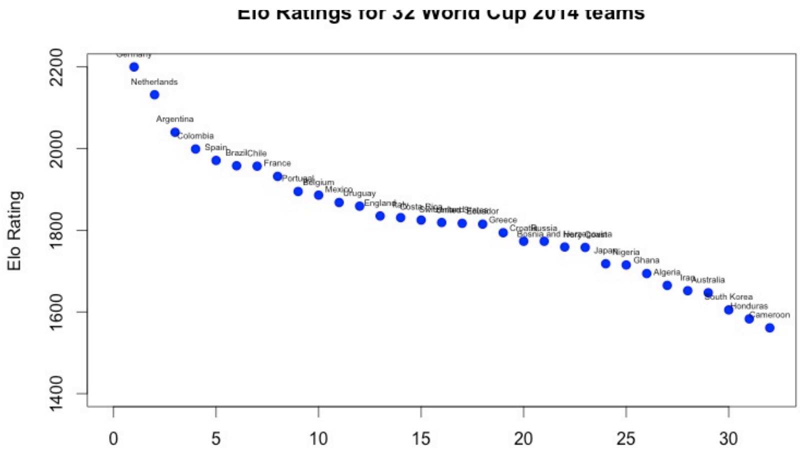


possible $\Upsilon(x)$

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

From e.g. Wikipedia *World Football Elo Rankings*

GROUP STAGE



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.

Elo-type algorithms have nothing to do with probability, *a priori*. But there is an obvious heuristic 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 player 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?"

Is there any relevant non-elementary math probability?

Assume the basic probability model and use Elo-type ratings – what happens? We need to specify how the matches are scheduled, use the mathematically simplest "random matching" scheme in which there are n players and for each match a pair of players is chosen uniformly at random. This gives a continuous-state Markov chain

$$\mathbf{Y}(t) = (Y_i(t), 1 \le i \le n), t = 0, 1, 2, \ldots$$

where $Y_i(t)$ is the rating of player i after a total of t matches have been played. We call this the *update process*. Note that this process is parametrized by the functions W and Υ, and by the vector $\mathbf{x} = (x_i, 1 \le i \le n)$ of player strengths. We center player strengths and rankings: $\sum_i x_i = 0$ and $\sum_i Y_i(0) = 0$.

The following convergence theorem is intuitively obvious; the technical point is that no further technical assumptions are needed for W, Υ.

Theorem

Under our standing assumptions (1, 3, 4) on W and Υ, for each $\mathbf{x}$ the update process has a unique stationary distribution $\mathbf{Y}(\infty)$, and for any initial ratings $\mathbf{y}(0)$ we have $\mathbf{Y}(t) \to_d \mathbf{Y}(\infty)$ as $t \to \infty$.

This is proved by standard methods – coupling and and Lyapounov functions. Note here we are not assuming the specific relation (6) between W and Υ. Note also that given non-random initial rankings $\mathbf{y}(0)$ the distribution of $\mathbf{Y}(t)$ has finite support for each t, so we cannot have convergence in variation distance, which is the familiar setting for Markov chains on $\mathbb{R}^d$ (Meyn-Tweedie text).

Alas these techniques do not give useful quantitative information about the stationary distribution. The theorem suggests a wide range of quantitative questions that we can't answer.

Let's fix W and Υ satisfying

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty \quad (6)$$

and update using $c\Upsilon$. Intuitively there is a trade-off between wanting the "rating errors" $\sum_i \mathbb{E}(Y_i(\infty) - x_i)^2$ to be small and wanting the Markov chain mixing time to be small; would like to formalize this as a single "cost" that we could optimize over c; this would provide insight into which solution of (6) to use.

Because "small mixing time" corresponds roughly to "better at tracking changes in strengths", a roughly equivalent and more "realistic" question is as follows.

Take some model of time-varying strengths $\mathbf{X}(t)$; we expect some asymptotically stationary joint process $(\mathbf{X}(t), \mathbf{Y}(t))$ and we would like to estimate asymptotic errors $\sum_i \mathbb{E}(Y_i(t) - X_i(t))^2$.

All these questions seem too hard as theory; we are playing around with numerical examples (projects for students).

Intuitively, instead of a "univariate" rating as in Elo, the following "bivariate" scheme should improve on any given Elo scheme with update function $c\Upsilon$. Use as one component the Elo rating $Y_i(s)$ with some $c' > c$; take the other component $Z_i(s)$ as the discounted past average of the $Y_i(\cdot)$, so we can update as

$$Z_i(s) = \lambda Z_i(s-1) + (1-\lambda) Y_i(s).$$

Announcing Z_i as the "rating" should reduce the effect of recent randomness of match results.


$59 99
Estimated Retail Price
XBOX
HALO 4
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.*

A different "bivariate" algorithm is used in the TrueSkill ranking system on Xbox Live. 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". In fact it implicitly assumes the ‘‘total order" model, that there is some (unknown or random, to frequentists or Bayesians) ranking such that a higher-ranked team always beats a lower-ranked team.`

Many mathematicians have thought about design of tournaments, as I discovered after first giving this talk. For instance, Lewis Carroll and me (prior to this project) . . . . . .

LAWN TENNIS TOURNAMENTS

The True Method of Assigning Prizes with a Proof of the Fallacy of the Present Method

I. INTRODUCTORY

AT a Lawn Tennis Tournament, where I chanced, some while ago, to be a spectator, the present method of assigning prizes was brought to my notice by the lamentations of one of the Players, who had been beaten (and had thus lost all chance of a prize) early in the contest, and who had had the mortification of seeing the 2nd prize carried off by a Player whom he knew to be quite inferior to himself. The results of the investigations, which I was led to make, I propose to lay before the reader under the following four headings—

(*a*) A proof that the present method of assigning prizes is, except in the case of the first prize, entirely unmeaning.

(*b*) A proof that the present method of scoring in matches is constantly liable to lead to unjust results.

(*c*) A system of rules for conducting Tournaments, which, while requiring even less time than the present system, shall secure equitable results.

(*d*) An equitable system for scoring in matches.

I have used the following for many years as a take-home exam problem, for a graduate course using *Durrett*. The model here is more general than the "basic probability model" in this talk.

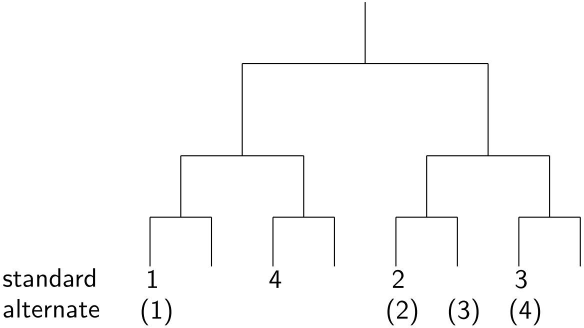
Proposition

Suppose that amongst n players the winning probabilities p_{ij} are unknown to us, and are arbitrary except that for some unknown "best" player i^* we have $p_{i^*j} \geq 0.5 + \delta \ \forall j \neq i$, for known $\delta > 0$. Then we can design a (broad sense) tournament which in $C(\delta)n$ matches will enable us to determine the best player with probability $\leq \delta$ of error, where $C(\delta)$ is a constant depending only on δ.

Returning to earlier questions about the basic probability model . . .

There is a standard way to design a single-elimination tournament in terms of the seeding (ranking) of players. Within the basic probability model, does this standard design optimize anything (compared to other designs) uniformly over the ratings?

I believe the answer is "no". For instance it does not maximize the chance that the top-seeded player wins the tournament.



In the alternate design in this figure, player 2 is less likely to get to the final and player 1 is more likely to win the final.

By a recursive argument, the standard design does not maximize the chance that the final is played between seeds 1 and 2.

The final topic in this talk.

Suppose each of n teams plays each other team once during a season. Each team i wins some number $q^*(i)$ of its matches. Under a given probability distribution over the entire season match results there is some expected number $q(i)$ of matches won by team i. Characterize the possible values of the sequence $(q(i), 1 \leq i \leq n)$
(a) for a completely arbitrary probability distribution
(b) under our basic probability model (with arbitrary W and strengths).

I will outline the answer to (a). It's an **open problem** whether (b) has the same answer.
Recall the notion of *convex order* on distributions: $Y \preceq U$, means

$$\mathbb{E}\phi(Y) \leq \mathbb{E}\phi(U) \ \text{ for all convex } \phi$$

or equivalently

$$Y =_d \mathbb{E}(U|Z) \text{ for some } Z. \tag{7}$$

Write U_n for the uniform distribution on $\{0, 1, 2, \ldots, n-1\}$. If there were a "total order" on results – the better team always wins, deterministically, then the numbers $\{q^*(i)\}$ of wins would be $\{0, 1, 2, \ldots, n-1\}$, so we write this as

$$q^*(I_n) \stackrel{d}{=} U_n$$

where I_n denotes a uniform random team.

Proposition

Consider an arbitrary probability distribution over season results, and write

$$q(i) = \mathbb{E}(\text{number of wins for team } i). \tag{8}$$

Then

$$q(I_n) \preceq U_n \tag{9}$$

in the sense of convex order. Conversely, given a function q from the set of teams to $[0, n-1]$ such that (9) holds, there exists a probability distribution over season results such that (8) holds.

Outline of proof.
(i) Given arbitrary deterministic results, inductively change the results as follows

make the team with worst record lose all its matches
then make the eam with second worst record lose all its matches except 1
. . . . . .

At each step $q^*(I_n)$ can only increase in convex order, and eventually it $=_d U_n$ because we get to a "total order" results.

Jensen's inequality establishes the "random results" case.

(ii) Write I_n and J_n for RVs with the uniform distribution on $\{1, 2, \ldots, n\}$. We are given a function $q : \{1, \ldots, n\} \to [0, n-1]$ such that $q(I_n) + 1 \preceq J_n$; using the martingale interpretation of convex order (7) we can construct a joint distribution for (I_n, J_n) such that

$$\mathbb{E}(J_n | I_n = i) = q(i) + 1, \ 1 \leq i \leq n.$$

Now the matrix with entries

$$p_{ij} := n\mathbb{P}(I_n = i, J_n = j)$$

is doubly stochastic, so by Birkhoff's theorem it is a mixture of permutation matrices; there is a *random* permutation $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ such that

$$p_{ij} = \mathbb{P}(\pi(i) = j).$$

Now use the realization of π to define a "total order" on season results – i beats i' iff $\pi(i) > \pi(i')$. So

$$\text{number of wins for team } i = \pi(i) - 1$$

and the identities above imply

$$\mathbb{E}(\text{number of wins for team } i) = q(i).$$