Cheating Detection and Player Estimation

St. Louis Chess Conference 2024

Kenneth W. Regan[1]
University at Buffalo (SUNY)

25 October, 2024

[1]With grateful acknowledgment to co-authors Guy Haworth and Tamal Biswas, students in my graduate seminars, and UB's Center for Computational Research (CCR)

A Predictive Analytic Model

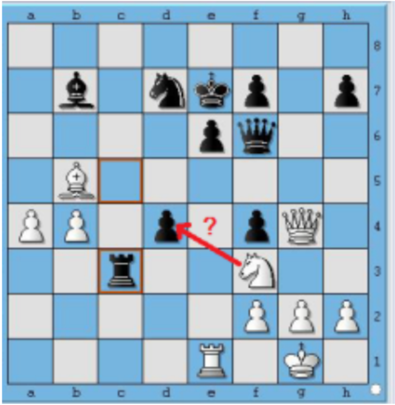
Means that the model:

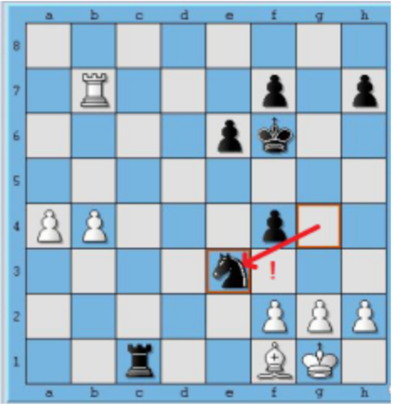
- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Example:

- In one application, the m_i were ways to get to downtown San Francisco, with utilities u_i based on time and cost. [McFadden et al.]
- Consumer profiles + $u_i \rightarrow$ projected probabilities p_i.
- **In my model, the m_j are possible moves in chess positions.**
- The utilities u_i are move values judged by strong chess **engines**.
- Player skill profiles (mainly Elo ratings) + $u_i \rightarrow$ move probabilities p_i.

Move Utilities Example (Kramnik-Anand, 2008)





Depths... Values by Stockfish 6

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

Inputs

- Main difference from McFadden is the **utility function / loss function** being **log-log linear**, not log-linear (why).
- So each p_i is a **power** not multiple of the best-move prob. p_1.
- Second important "differentiator": my heavily scaled version (**ASD**) of "*average centipawn loss.*"
- Other than move values, **my model knows nothing about chess.**

The (only!) player parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment.
- c for "consistency" in surviving tactical minefields.
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 between players within 10 Elo of a marker 1025, 1050, . . . , 2775, 2800, 2825. Wider selection below 1500 and above 2500.

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.
- (Correct after-the-fact for chess decisions not being independent.)

The statistical application then follows by math known since the 1700s. (Example of "Explainable AI" at small cost in power.)

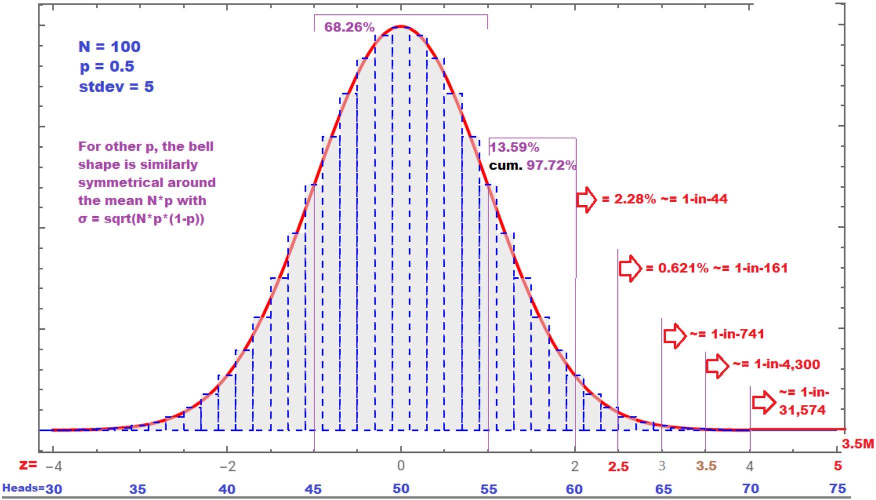
Validate the model on millions of randomized trials involving "Frankenstein Players" to ensure conformance to the standard bell curve at all rating levels.

See: Published papers and articles on Richard J. Lipton's blog **Gödel's Lost Letter and P=NP**.

Z-Scores

- A **z-score** measuresf performance relative to natural expectation.
- Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
- Expressed in units of standard deviations, called "sigmas" (σ).
- Correspond to statements of odds-against (**but see next slides**):
- "Six Sigma" (6σ) means about 1,000,000,000–1 odds;
- 5σ = about 3,500,000–1;
- 4.75σ = about 1,000,000–1;
- 4.5σ = about 300,000–1;
- 4σ = about 32,000–1;
- 3σ = about 750–1 (closest is 740–1);
- $2\sigma \doteq$ 43–1 (civil minimum standard, polling "margin of error").

Bell Curve and Tails (also Screening Stage)



Suppose We Get $z = 3.54$

- Natural frequency $\approx$ 1-in-5,000. *Is this Evidence?*
- Transposing it gives “raw face-value odds” of “5,000-to-1 against the null hypothesis of fair play. **But:**
- **Prior likelihood** of cheating is estimated at
  - 1-in-5,000 to 1-in-10,000 for in-person chess.
  - 1-in-50 (greater for kids) to 1-in-200 for online chess.
- **Look-Elsewhere Effect**: How many were playing chess that day? weekend? week? month? year?

Are these considerations orthogonal, or do they align?

If you’re “marked” by a previous incident, these recede.

If there is on-site evidence, z = 2.50 is enough (FIDE).

Evaluation Criteria and Demonstrations

1. Is it **safe**? That is, do its outputs conform to an expected (normal) distribution over populations that obey the null hypothesis? (Yes).
2. Is it **sensitive**? And are its positive results clearly pertinent to the desired inferences? (Can improve?)
3. How is it calibrated? Are the calibration—as well as positive results—**explainable**?
4. Can it be **cross-validated**? What sanity checks does it provide?
5. Does it model more than what its proximate application demands, so as to be robust against "mission creep"?

Show demos as time allows:

- **US Championships.**
- **David Smerdon's experiment.**
- **Budapest Olympiad.**

Player Estimation

- Model → **Intrinsic Performance Rating** (**IPR**) for any games.
- IPR still may overdo *accuracy*, undercut *challenge created.*
- The s, c, h... tradeoff that produces a given Elo IPR value judges positional versus tactical abilities.

Questions that IPR can answer:

1. Natural growth curves for young players? & arcs for older players?
2. Are there substantial geographical variations in ratings?
3. How does skill at fast chess correlate with ratings at slow chess?
4. Has there been rating **inflation**? Is there current **deflation**?

Rating estimation bias skews linearly, but my model has ample cross-checks by which to detect and correct it. The pandemic brought a truly monstruous situation where official ratings were frozen for years...

Rating Lag—Natural Versus Pandemic-Caused

- **The #1 scientific role I've played since the pandemic has been estimating the true skill growth of young players.**
- My "back of the envelope" formula held up over two years with only one small revision for preteens.
- Revision in Oct. 2022 to curtail projections past Elo 2000 level.
- Would have been more "normal" if comprehensive studies of the career arcs (measured by Elo rating) of young players were to hand.
- Lack of such studies exposed by the controversy over Hans Niemann's rise from 2465 Elo to 2700.
- Show this GLL article including example of Ms. Sarayu Velpula.
- Near-term to-do: **Improve gauging of difficulty**.
- To-do: **Use move-time information**. But absent in many cases. Updating Ludwig Wittgenstein's maxim: *On what we cannot model, we must remain silent.*

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating (to 2637).
- Hou Yifan is 2633 but semi-inactive; next is Ju Wenjun at 2563.
- (But are current top female players more distinctly underrated?)
- Where and when does the gap begin?
- "Nature versus Nurture"—or rather **Duration of Engagement**?
- I have not found differences between these improvement factors:
  - Playing in-person chess events—versus binging online blitz.
  - Study alone—versus with a regular chess coach (online).
- What data could test a simple "10,000 hours" hypothesis?
- Perhaps: time spent on major platforms, crosstabbed by age, rating, and gender. Alas not maintained as such?
- **Q&A**, and **Thanks**.