Fraught Issues in Statistical Chess Cheating Detection

Physics Colloquium, Vanderbilt University

Kenneth W. Regan
University at Buffalo (SUNY)

16 November, 2023

\[1\text{With grateful acknowledgment to co-authors—including Tamal Biswas now of RKMVERI—and UB’s Center for Computational Research (CCR)}\]
Two Framing Issues

1. What does it mean to have statistical confidence in non-repeatable events?
   - whether $X$ exists in our accessible universe
   - whether $X$ cheated at chess.

2. Can regularities of human behavior reach the status of physical law?

Re. 2: We are physical systems, after all. “P.O.B.I.T.E. Lite.”
Re. 1: I hope to shed light on some current miseries not mysteries of physics—physics praxis, that is.
What Is a Physical Law?

Nervy Answer:

A severely underfitted model that works.

For example, consider three (or five) natural quantities:

\( m_1 \) : tendency of \( X \) to resist force.

\( m_2 \) : capacity of \( X \) to exert force.

\( m_3 \) : count of basic particles in \( X \).

Isaac N: “Let’s model all three by one variable \( m \) called mass.”
Elo Chess Ratings—and Why Cheat?

- Named for **Arpad Elo**, number $R_P$ rates skill of player $P$.
- Defined by Logistic Curve: expected win % $p$ given by
  \[
  p = \frac{1}{1 + \exp(c\Delta)}
  \]
  where $\Delta = R_P - R_O$ is the difference to your opponent’s rating.
- Taking $c = (\ln 10)/400$ makes $\Delta = 200$ give about 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = “Super GM,”; Carlsen only player over 2800. Adult beginner $\approx 600$, kids $\rightarrow 100$.
- **Stockfish 16 3544, Torch 1.0 3531, Komodo Dragon 3.3 3529**.
- So computers are at “Class 15.” $\Rightarrow$ a “Moore’s Law of Games.”
- Other Q: How do computer evaluations—in units of hundredths of a pawn (centipawns)—translate to chances of winning?
Model: Inputs and Parameters

- Based on a utility function / loss function $\delta$ in a standard way—except for being log-log linear, not log-linear.
- The (dis-)utility comes from (my heavily scaled version of) average centipawn loss of the played move compared to (what a powerful chess-playing program thinks is) the best move.
- No chess knowledge other than the move values is input.

The (only!) 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, …, 275, 2800, 2825. Wider selection below 1500 and above 2500.
Model: Lone Equation (*)

\[
\frac{\log(p_i)}{\log(p_1)} = r_i = \exp \left( \frac{\delta(\vec{v}_1, \vec{v}_i; e_v)}{s} \right)^c,
\]

where

- \( p_1 \) = projected probability of playing the move ranked first by the chess program.
- \( p_i \) = projected probability of the \( i \)-th ranked move.
- \( v_1 \) = value vector of first-ranked move across depths of search.
- \( v_i \) = value vector of \( i \)th-ranked move.
- \( e_v \) = “eagerness” of the player. Essentially a restriction of the idea to cases of deciding between equal-valued moves.

(*) Except for the separate training of a gaggle of hyper-parameters...
Why Not a Simpler Log-Linear Model?

\[
\log(p_i) = \alpha + \beta \left( \frac{\delta(\vec{v}_1, \vec{v}_i; e_v)}{s} \right)^c
\]

- Normalizing \(\sum_i p_i = 1\) drops out \(\alpha\).
- Fit \(\beta\), then compute \(p_i\) via \text{softmax}.
- Analogous to Gibbs Equations (well, if \(c = 1\)).
- Works in much of Machine Learning, but not in chess.
- Double-log model has perilous dynamics, needs careful hyperparameter settings. (Predictivity-robustness tradeoff.)
The lone equation fits $p_i$ as a **power** not a **multiple** of $p_1$.

\[ p_i = p_1^{r_i}; \quad \sum_i p_i = 1. \]

Yields **aggregate projections** over sets $T$ of game turns $t$ of:

\[
\frac{1}{T} \sum_{t=1}^{T} p_{1,t} = \text{“T1 match” to computer}
\]

\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{\ell} p_{i,t} \delta(-i-) = \text{“average centipawn loss”}
\]
Internal and External Confidence

- Projections also automatically give additive variance, hence $\sigma$ and confidence intervals, if we assume turn decisions are independent.
- [Voiceover: They’re not.]
- But it’s a sparse dependence on neighboring moves. (Not across games—common “opening book” is removed from the sample.)
- $\Rightarrow$ covariance matrix is banded, hence approximable by scalars.
- Could treat as a “reduced-entropy” sample size $T' < T$.
- What I actually do is adjust $\sigma$ up to $\sigma'_E$ with dependence on Elo rating $E$ determined by millions of randomized resampling trials from the training sets.
- With this patched, justified in saying the model paints chess moves on a 1,000-sided die and simply rolls it. $\Rightarrow$ multinomial Bernoulli trials.
Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the \textit{scaled} average centipawn loss from disagreements.
- Creates a \textbf{Raw Outlier Index (ROI)} on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation \textit{given one’s rating} and 5 is the standard deviation, so the “two-sigma normal range” is 40-to-60.
- Like medical stats except \textbf{indexed} to common \textbf{normal} scale.
- 65 = amber alert, 70 = code orange, 75 = red. \textbf{Example}.
- \textbf{Completely data driven}—no theoretical equation.
- Does not account for the \textit{difficulty} of games. That is the job of the full model.
Recent Performance Examples
Z-Scores and Cheating Tests

For the aggregate quantities, the Central Limit Theorem in practice allows treating

\[ z' = \frac{\text{(actual)} - \text{(predicted)}}{\sigma'} \]

as a \textit{z-score} (after adjustment).

Evaluation Criteria:

- **Safety**: Over fair-playing populations, \( z' \sim \text{bell curve.} \)
- **Sensitivity**: Factual cheaters yield “high enough” \( z' \).

\textit{From this point on, let’s suppose my model has these properties.} What about interpreting the results?
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
  - 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?
Fraught Issue #1

What should be the target confidence?

1. Proof beyond reasonable doubt?
2. “Comfortable satisfaction”
3. “Balance of Probability”

CAS Lausanne recognizes all three, but inclines toward 2.
- Still doesn’t specify a corresponding confidence target.
- Science, of course, demands criterion 1.
Fraught Issue #2: Confidence For Chess

- I interpret the range of comfortable satisfaction as 99–99.9% final confidence.
- For calling elections, Decision Desk HQ uses 99.5% confidence.
- Not quite right to say 1-in-200 error, i.e. a “Florida” every 4 cycles, because returns often blast past that instantly.
- So maybe truer chess analogue is 1-in-500 error.
- Judge by “Countenanced Error Rate Per Year.”
- E.g. if 10 cases per year reach judgment stage, and you can tolerate 1 error per 20 years, then 99.5
- But online chess has 10,000+ cases per year...
Issue # 3: Accounting “Look Elsewhere”

- Approximately 100,000 players-in-event per year among “notable” events.
  - notable \(\equiv\) some or all gamescores preserved.
- A highly computerlike game is a “shiny marble”—players do notice.
- Accounted over a year, suggests to divide odds by 100,000.
  - 4.75 sigma \(\rightarrow\) only 90% confidence.
  - 5.00 sigma \(\rightarrow\) 1-in-35 error.
- Sounds like 1-in-35 error is still too high based on confidence target.
- But reckon against time-scale of actual cases and tolerated error rate.
Why stop at a year? Why not consider “look elsewhere” over an entire 50-year span?

- IMHO, the notorious Doomsday Argument kicks in for real to fend off this level of skepticism...at least for now.
- Key point: What are the odds of getting this once-in-50-years event this (early) year?
- (My formal IP agreement with FIDE is 20 months old.)
- (But I deployed my model in 2011.)
- Better argument?: Balance against the arrival rate of real cases.
- Aligns with Bayesian prior on average, but should allow for variance in the rate.
- Figure discount by 25,000 to 50,000. Then 5-sigma is OK.
Issue #4: Event Tiers

But what if we have a *top-tier* event?

- World Championships.
  - Many of these per year, down to Under-8 Cadets.
- Qualifying events for championships.
- Major international Opens.
- The Carlsen Online Chess Tour.
- Chess.com “Titled Tuesdays” ...

The combination of the online 100-1 prior and marquee online events amps up the calculus.
Issue #5: Distinguishing Marks

What if the $z = 3.54$ is on Hans Niemann? Is he a “marked man”? Even granting he’s never cheated at in-person chess?

- Niemann plays $\approx 25$ events per year.
- Like giving drug test to same athlete 25x.
- But what about a player wearing a heavy winter overcoat in hot weather?
- Or a player wearing neon-green sneakers??
- Yet another separate matter from the Bayesian prior.
Super-Fraught Issue #6: Multi-Testing Samples

- Includes Cherry-Picking and other forms of \( p \)-hacking.
- What if a player seems to have cheated only in games 5–8 of a nine-game Open?
- Or maybe games 4–6 and 8–9?
- Proper domain of Bonferroni Correction if it doesn’t wipe out significance altogether.
- Well, \( z \)-hacking/\( p \)-hacking is a huge area...
What if you get $z = 3.54$ on three different players in a 500-player Open?

Not enough to convict any one player.

But odds against all being fair can be estimated by aggregating $z$-scores, presuming (under the null hypothesis of fair play) that the players’ actions are independent:

$$z = \frac{z_1 + z_2 + z_3}{\sqrt{3}} \approx 6.13 \text{ Billion-to-one}$$

Applying “Look-Elsewhere” still leaves astronomical confidence that some cheating occurred. Still leaves the question of who.
Issue #8: Scaling of Estimation Error

- My formulas—“screening” as well as the predictive analytic model—scale as $O(\sqrt{n})$ gracefully to any sample size $n$ of games/moves:
  - 5-game weekend tournaments;
  - 9-game international Opens;
  - 13-game invitational round-robbins;
  - 12–24 game championship matches.
- But how about 300+ games played in “Titled Tuesdays” over a half-year span?
- Skew from rating estimation error scales linearly as $\Omega(n)$.
- Overflows the $O(\sqrt{n})$ levees... Validation by myriad resampling trials done on $n = 4, 9, 16$. 
Issue #9: Biased Inputs

- Lag in ratings of rapidly improving young players.
- Was exponentiated by the pandemic. “Pandemic Lag” article on the GLL blog.
- Cause of many unwarranted suspicions, even recently.
- Also geographical variations in ratings.
- As in issue 8, rating estimation bias skews linearly.
- My model has enough cross-checks to detect and correct the bias—mainly need only assume not everyone is cheating. No “interstellar dust” issue.
Going Post-Normal

- Arguments over the Niemann-Carlsen fracas a year ago exposed the lack of any rigorous studies of the growth curves of young improving players.
- In Sept.-Nov. 2020, I fitted a simple formula from observations of players in multi-age youth events 5–7 months since their official ratings were frozen.
- I am still using fairly much the same formula, now 43 months in. Well, with some tweaks:
  - Reduced multiplier for players under age 12 from 30 Elo per month to 25; later filled in 20x for ages 12 and 13 as of April 2020.
  - Gains above Elo 2000 reduced by treating formula as a differential.
  - Formula for teenagers (with 15 multiplier) otherwise unchanged.
- Adjusted players are often over half the entrants in large Opens.
- Basically running a more accurate rating system from the back of an envelope.
The pandemic drove major tournaments online—where chess is played faster.

Not enough reliable training data for (in-person) fast chess across skill levels.

Panoply of different speeds anyway: $\tau =$ time you can use to play 60 moves.

FIDE standard slow chess gives $\tau = 150$ minutes.

Postulate: Elo reduction $R_E(\tau)$ if largely independent of the player’s Elo rating $E$.

Reasonable \textit{a-priori} since chess rating system is designed for additive invariance: only the difference in ratings to the opponent matters for predictions.
Reliable data for $\tau = 25$ and $\tau = 5$ (as well as $\tau \geq 150$) from the elite annual World Rapid and Blitz Championships.

Guess that $R(\tau)$ is logistic in $\log \tau$, so polynomial rational in $\tau$.

Gives four unknowns to fit, but only three equations. Try getting fourth from:

- Rating estimate of $\tau = 0$, i.e., of completely random chess. Implicitly done here.
- Aitken Extrapolation.

Lo and behold—the two methods agree!

Is the resulting “Rating Time Curve” thereby a natural law?

Does this make time fungible with difficulty, the latter as modeled by Item Response Theory?
Stance on Data Science

- Extreme Corner of Data Science—since I need ultra-high confidence on any claim. Well, so do you.
- Concern: Data modelers in less-extreme settings *satisfice*.
- That is, their models are designed up to one particular goal but don’t explore much of the harder adjacent metaspace. (Compare what Scott Aaronson calls the Meatspace.)
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.
- Here is a way of phrasing the question that comes from this stance:

  When is it important that our models include gravity?
Q & A

And Thanks.