Statistics and Analytics in Chess

Skill Rating and Cheating Detection

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

5 October, 2013

[1] Includes joint work with Guy Haworth and GM Bartlomiej Macieja. Sites: http://www.cse.buffalo.edu/~regan/chess/fidelity/ (my homepage links), http://www.cse.buffalo.edu/~regan/chess/ratings/ (not yet linked).

Outline

1. Cheating detection and much more.
2. Two aspects of cheating detection:
   - General: Idea and necessity of **z-score** concept.
   - Specific: Operation of my particular model.
3. Three tiers of application (partly depending on z-score):
   1. Hint to arbiters during competitions
   2. Support of observational evidence of cheating
   3. Standalone indication of cheating (needs $\mathbf{z > 5}$, maybe 4.75 or 4.5).
4. Analytics: specific moves; Intrinsic Performance Ratings (IPRs).

Why Z-Scores? I. Absolutes don't work

Actual Matching and Average Error in PEPs (Pawns in Equal Positions)

Elo	MM%	AE
2800	57.8	0.048
2700	56.3	0.055
2600	54.8	0.063
2500	53.3	0.070
2400	51.8	0.077
2300	50.3	0.084
2200	48.8	0.091
2100	47.3	0.098
2000	45.8	0.105

Hence a fixed rule like "70% matching = sanction" won't work.
But how about "70% for 2600+, 65% for rest" or "MM + 15%"?

II. Baseline Depends on Players' Games

1. Anatoly Karpov and Mikhail Tal co-won Montreal 1979 with 12-6 scores.
   - Tal matched 61.6%, which was best.
   - Karpov matched 49.9%, which was worst—by over 2%!
   - But my model **projects** only 51.0% for Karpov (56.8% for Tal).
2. Le Quang Liem matched **69.1%** at Aeroflot 2012.
   - My model gives a baseline of 61.7% for 2700 player;
   - His Multi-PV figure **regresses** to 64%;
   - He scored 3.5/9 in the tournament.
3. Top 3 Rybka-matchers in the *entire series of famous Lone Pine tournaments* are: Doug Root 62.8%, Ed Formanek 62.7%, and *moi*, 62.4% tied with Gennadi Sosonko. My 2700 baseline: 64.0%.
4. On positions faced by **Stockfish 4** in the current **nTCEC** tournament, a 2700 player would match under 47%.

Z-Scores

1. A **z-score** is a measure of performance relative to natural expectation.
2. Used extensively by business in Quality Assurance, Human Resources Management, and by many testing agencies.
3. Expressed in units of standard deviations, called "sigmas" (σ).
4. Every z-value includes a statement of odds against that-or-higher deviation. E.g.:
   - "Six Sigma" (6σ) means about 500,000,000–1 odds;
   - 5σ = 3,000,000–1;
   - 4.75σ = 1,000,000–1;
   - 4.5σ = 300,000–1;
   - 4σ = 32,000–1;
   - 3σ = 740–1;
   - 2σ = 43–1 (civil minimum-significance standard)
5. Example: Poll says Obama $52.0\% \pm 3.0\%$—if he had got 46% that would have been a 4σ deviation, probable sign of fraud. Ditto 58%.

Applying Z-Scores

1. **Statistical Test:** A quantity μ that follows a *distribution*.
2. If μ is an average of a **sample** taken from any distribution, then μ itself obeys **normal distribution**, and the general "p-test" theory becomes the well-traveled z-test theory.
3. You need a statistical model that upon analyzing a series of games gives **both μ and σ as internal projections**.
4. Then the projections must be tested against 10,000s of trials of games—presumably by non-cheating players—to verify conformance. OK to err conservatively.
5. Main statistical tests I use:
   - Move-matching (MM%).
   - Average error per move (AE), scaled in units of PEPs.
   - Equal-Top matching (TM%), usually 3–4% higher than MM%.
   - ~~Top-3 matching~~: AE test is more robust.
6. Online chess servers use specialized tests on greater information, such as exact time per move, "telltales," particular engine profiles...

Understanding and Applying Z-Scores

Main principle:

> *The odds that come with z-scores really represent frequencies of natural occurrence.*

1. Can measure freqency in units of "Weeks of TWIC."
2. One week = about 1,000 player-performances.
3. So 4σ = 32000–1 odds = 32 weeks of TWIC.
4. Thus we should see a 4σ-deviation *up* by a non-cheating player once every half-year or so, and also a 4σ deviation *down*.
5. But 5σ = 3,000,000–1 odds = 60 years of TWIC = more than the entire history of chess. (Actually closer to 3.5M–1, 70 years.)
6. While in an Open tournament 2σ is *nothing*: if 22 games are going on, you'll see a 2σ deviation.
7. Propose 3σ as the threshold for hints to TDs to watch a player more closely and meaningful support for observational evidence.

My actual presentation stopped here...

I had expected to give a general talk before the main meeting, updating my slides below, but in fact it was part of the main meeting, and the preliminary meetings in Paris also brought home to me the need to focus new slides on the topics above. That talk took about 30 minutes, then during 45 minutes of questions I was able to show other examples from my large data sets.