Ratings Variability Measure

Mark E. Glickman
Senior Lecturer on Statistics, Harvard University
Chair, US Chess Ratings Committee

Background:

The Ratings Committee (RC) was tasked by the US Chess Executive Board (EB) with proposing an algorithm to construct a measure of variability of a player's rating. Such a measure may be useful for tournament directors who may want to have a better idea of the range of abilities of players in the tournaments they organize. The goal of the proposed computation is to produce a numerical measure that characterizes the variability of a player's performance, with higher values indicating greater variability, as well as an interval measure that provides a range of rating values at which the player is capable of playing.

Variability algorithm:

The algorithm for the variability computation would be applied to US Chess players who compete in events rated in the over-the-board regular rating system. Below are the steps for computing the variability measure and the accompanying interval estimate:

- For each tournament $k = 1, \ldots, K$ played in the prior 3 years, compute a tournament performance rating (TPR_k). A TPR is the rating at which the sum of winning expectancies equals the player's attained score in the tournament.[1]
    - Use the opponents' post-tournament ratings to obtain the player's TPR for each tournament.
    - If the player has either all wins or all losses in a tournament, use the total score minus 0.25 in the case of all wins, and a total score of 0.25 in the case of all losses.

[1] The TPR calculation can be implemented efficiently as a slight modification of the Newton-Raphson algorithm, and takes only a few iterations to converge.

- Record the number of days, D_k, that have elapsed since tournament k's completion, and the number of completed games n_k in tournament k by the player.
- Determine weights connected to the TPR for each tournament. These are functions of both the number of games played in the tournament and the time elapsed since the completion of the tournament. Specifically, the weight for tournament k is computed as $w_k = n_k \exp(-\frac{0.36D_k}{365.25})$ for tournament k with n_k games and having been completed D_k days ago. The -0.36 factor in the exponential term is used in the staleness computation in the rating system. With this value, games played 1 year ago have about 70% weight compared to current games.
- Compute the weighted mean performance rating as

  $$R_w = \left(\sum_{k=1}^{K} w_k \mathrm{TPR}_k\right) / \sum_{k=1}^{K} w_k$$

  Compute the weighted variance of the performance rating as

  $$V_w = \frac{\sum_{k=1}^{K} w_k (\mathrm{TPR}_k - R_w)^2}{\left(1 - \frac{\sum_{k=1}^{K} w_k^2}{\left(\sum_{k=1}^{K} w_k\right)^2}\right) \sum_{k=1}^{K} w_k}$$

- The weighted standard deviation can be computed as $S_w = \sqrt{V_w}$, and this value can be reported as the variability measure.
- Two interval summaries can accompany the weighted standard deviation.
  - One summary is a 90% confidence interval, which would be formed as

    $$(R_w - 1.645S_w, R_w + 1.645S_w)$$

    This interval assumes that the variability in TPRs is symmetric around the mean performance rating.
  - A second approach is to calculate the 5th percentile and 95th percentile of the TPRs, accounting for the tournament weights. This is a standard calculation and results in asymmetric intervals of rating ranges.

Example calculation:

Suppose a player has competed in 8 tournaments in the prior 3 years, with a total of 35 rated games. Table 1 lists the results of the 35 games.

Tournament_ID	Days_Elapsed	Opponent_Rating	Score
1	30	1700	1.0
1	30	1800	0.5
1	30	1850	1.0
1	30	1900	0.0
1	30	1750	0.5
2	90	2100	0.0
2	90	1900	0.5
2	90	1950	0.5
2	90	2000	0.0
3	150	1800	0.5
3	150	1850	0.5
3	150	1825	1.0
4	210	1750	1.0
4	210	1800	1.0
4	210	1850	0.0
4	210	1725	0.5
5	280	1630	0.5
5	280	2070	0.0
5	280	1815	0.5
5	280	1840	1.0
6	370	1620	0.5
6	370	1960	1.0
6	370	1520	0.5
6	370	1750	0.0
7	450	1790	0.0
7	450	1560	1.0
7	450	1820	1.0
7	450	1540	1.0
7	450	2050	0.5
8	730	1520	1.0
8	730	1560	0.0
8	730	1515	0.5
8	730	2020	0.5
8	730	1550	0.0
8	730	2070	0.0

Table 1: Example game outcome data and opponents' ratings

The four columns are

- Tournament_ID: ID of tournament from 1 to 8.
- Days_Elapsed: Number of days ago the event was completed.
- Opponent_Rating: For each game, the post-tournament rating of the opponent.
- Score: Outcome of the game – 1 for a win, 0.5 for a draw, 0 for a loss.

Table 2 summarizes the details of the variability algorithm for each tournament.

Tournament	Number of Games	Days Elapsed	TPR	Weight
1	5	30	1873	4.85
2	4	90	1789	3.66
3	3	150	1946	2.59
4	4	210	1872	3.25
5	4	280	1837	3.04
6	4	370	1708	2.78
7	5	450	1935	3.21
8	6	730	1520	2.92

Table 2: Tournament summaries

The TPR column (tournament performance rating) is computed from the set of game outcomes in each tournament separately. The Weight column is computed as described in the algorithm outline, which depends on the number of games within the tournament and the days elapsed since its completion.

From the above table, the weighted mean TPR, R_w, is 1815, and weighted standard deviation of the TPR, S_w, is 133.5. A 90% confidence interval for the player's ability based on the weighted standard deviation is therefore (1595, 2035). That is, with 90% confidence, the player's ability ranges from 1595 to 2035. This interval is symmetric around the weighted mean TPR of 1815.

Using the approach of calculating the 5th and 95th percentiles for the distribution of TPRs with associated weights, the 90% confidence interval is (1736, 1942). This interval reflects the asymmetry in the distribution of TPRs.

Comments:

- The variability measure may be used by tournament directors to help decide whether a player should be placed in a higher-rated tournament. For example, a player with a rating of 1750 wants to enter an under-1800 section of a tournament. If the 90% confidence interval for the player is (1736, 1942), as in the example above, the tournament director may choose to place the player instead in the under-2000 section given that the expected variation in the player's performances can be at a 1900-rating level.
- The calculation of the standard deviation, S_w, and the associated confidence interval will remain unchanged as time passes, as long as the tournament game outcomes used in the calculation are still within a three-year window.
- The variability calculation is a measure of the expected variation in performances by a player. It is important to note, however, that this calculation does not account for trends in a player's performance, which is a much more difficult issue to address (e.g., the rating system does not account for trends either). Thus, for example, the variability calculation for a quickly improving young player will result in an interval that reflects how the player has been performing, and not how the player is projected to perform based on any inferred trends.

Next steps:

Assuming the EB approves in principle of this approach for determining estimated rating ranges, several tasks would need to follow:

1. A determination would need to be made which information to accompany published ratings. For example, should the weighted standard deviation be reported alone, should it be reported along with one of the confidence intervals, or should only one of the confidence intervals be reported?
2. Is it of interest to report a 90% confidence interval, of an interval at some other confidence level? Larger confidence levels (e.g., 95%, 99%) correspond to wider interval ranges, and smaller confidence levels correspond to narrower ranges. The value 90% is the lowest conventionally-used confidence level in practice.

3. The algorithm would need to be reviewed by US Chess staff to raise questions about implementation details, and the logistics of carrying out the computation.
4. Finally, the algorithm would need to be applied to the universe of regular rated games to examine the results for tournament players.