NORMALIZED ELO

MICHEL VAN DEN BERGH

1. Introduction

It is often discussed how one should compare the suitability for engine testing of opening books with different characteristics (balanced versus unbalanced, drawishness). Here we answer this question. A lot of inspiration came from Kai Laskos' empirical results posted on talkchess as well as from his comments. Credit goes also to him for first noting that the trinomial model is insufficient for working with unbalanced openings.

2. Normalized Elo

The primary purpose of an engine test is to show that one engine is stronger than another. The amount of evidence gathered from a test can be expressed by its "t-value".

$$t = \frac{\hat{s} - 1/2}{\hat{\sigma}(\hat{s})} \tag{2.1}$$

where $\hat{s} = \hat{w} + (1/2)\hat{d}$ is the score of a match (with $\hat{w}, \hat{d}, \hat{l}$ being the empirical WDL-ratios) and $\hat{\sigma}(\hat{s})$ is empirical standard deviation of this score. Note that the null-hypothesis is $s = 1/2$ and hence $t = 0$. The relation between the t-value and the more commonly used Likelihood Of Superiority is

$$\text{LOS} = \Phi(t)$$

where Φ is the cumulative distribution function of a normal distribution with unit variance and zero mean. The corresponding p-value is

$$p = 1 - \text{LOS}$$

The disadvantage of (2.1) is that $\hat{\sigma}(\hat{s})$ depends on the number of games N in the test via

$$\hat{\sigma}(\hat{s}) = \frac{\hat{\sigma}_0}{\sqrt{N}}$$

where σ_0 refers to the standard deviation of a single game. So

$$t = \sqrt{N}\frac{\hat{s} - 1/2}{\hat{\sigma}_0}$$

So to compare opening books we should really use the normalized quantity.

$$\boxed{t_0 = \frac{\hat{s} - 1/2}{\hat{\sigma}_0}}$$

In the case of balanced openings we find

$$\boxed{t_0 = \frac{\hat{s} - 1/2}{\sqrt{\hat{w} + \frac{1}{4}\hat{d} - \hat{s}^2}}}$$

Remark 2.1. In the case of unbalanced openings with paired games with reversed colors the formula $\hat{\sigma}_0 = \sqrt{\hat{w} + \frac{1}{4}\hat{d} - \hat{s}^2}$ over estimates $\hat{\sigma}_0$ and we should use the pentanomial model instead [1]. See §4.

Remark 2.2. Note

$$\sigma(t_0) = \frac{1}{\sqrt{N}}$$

so a 95%-confidence interval for t_0 is given by

$$\left[t_0 - \frac{1.96}{\sqrt{N}}, t_0 + \frac{1.96}{\sqrt{N}}\right]$$

It makes sense to call t_0 "normalized elo" since the result of an engine match is usually expressed as elo $\pm$ error where a 95%-confidence interval is assumed and then in good approximation

$$t_0 \cong \frac{1.96\,\text{elo}}{\text{error}\sqrt{N}}$$

Other reasonable names for t_0 are "sensitivity" (when measured under a standard set of conditions) or "signal-to-noise ratio" (see §3).

Remark 2.3. A typical setup to measure sensitivity is a self match with time odds using a well known engine. This reduces the number of parameters involved in setting up the testing environment. Time doubling seems reasonable although it is unknown if results for large elo differences truly extrapolate to small elo differences.

3. Signal-to-noise ratio

Here we present a different point of view on "normalized elo". In engineering a signal is decomposed[1] as

$$X = X_s + X_n$$

where X_s is the true signal and X_n is the noise added to it. The power ratio between the two components of X is the signal-to-noise ratio (SNR)

$$\text{SNR} = \frac{\int X_s(t)^2\,dt}{\int X_n(t)^2\,dt}$$

As noise will typically have zero mean we may also write

$$\text{SNR} = \left(\frac{\text{RMS}(X_s)}{\sigma(X_n)}\right)^2$$

where RMS stands for "root mean square":

$$\text{RMS}(X_s) = \sqrt{\frac{1}{T}\int_0^T X_s(t)^2\,dt}$$

[1]A signal will typically be sent over the wire by modulating the properties of a carrier signal, which has its own power content. Here we assume that the signal has already been demodulated.

We can view an engine match as a noisy DC-signal where X_s is the expected score s minus $1/2$ and X_n is the difference between the game outcome $\in \{0, 1/2, 1\}$ and s. We then find

$$\text{SNR} \cong t_0^2$$

4. The pentanomial model

Let $O := \{0, 1/2, 1, 3/2, 2\}$ be the set of possible outcomes of game pairs and let $(\hat{p}_i)_{i\in O}$ be the corresponding sample distribution. Let

$$\hat{s}_2 := \sum_{i\in O} \hat{p}_i i = 2\hat{s}$$

be the average outcome of a game pair. Then we get

$$\hat{\sigma}(\hat{s}_2) = \frac{1}{\sqrt{N/2}} \sqrt{\sum_{i\in O} \hat{p}_i i^2 - \hat{s}_2^2}$$

So for the t-value we obtain

$$t = \frac{\hat{s}_2 - 1}{\frac{1}{\sqrt{N/2}} \sqrt{\sum_{i\in O} \hat{p}_i i^2 - \hat{s}_2^2}}$$

Normalizing (dividing by $\sqrt{N}$) we get

$$\boxed{t_0 = \sqrt{2} \frac{\hat{s} - 1/2}{\sqrt{\sum_{i\in O} \hat{p}_i i^2 - 4\hat{s}^2}}}$$

References

[1] Chess Programming WIKI, *Match statistics*, https://chessprogramming.wikispaces.com/Match+Statistics.