WHY IS THE MATCH SCORE A GOOD STATISTIC FOR COMPARING ENGINES?

MICHEL VAN DEN BERGH

1. Introduction

Assume we have a parametrized distribution $p(\theta, x)$ A statistical test for comparing hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (or similarly $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$) typically consists of two parts:

(1) the choice of a test statistic which can be computed from a sample;
(2) the choice of a decisision rule for choosing between H_0 and H_1.

In statistical theory the merits of test statistics (1) can be evaluated independently of their use (2). This is the subject of "efficiency". See e.g. [1]. In fact the relative efficiency of test statistics can even be evaluated empirically when the underlying distribution $p(\theta, x)$ is unknown. However empirical testing alone can not tell us what is an "optimal" test statistic in a particular situation. This requires some statistical theory.

It is known that the so-called "score statistic[1]" $S(x)$ [4]

$$S(x) := \left.\frac{\partial \log p(\theta, x)}{\partial \theta}\right|_{\theta=\theta_0}$$

is an optimal statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (or similar onesided comparisons). Thus for a sample $(x_i)_{i=1}^n$ the total score statistic is

$$S(\underline{x}) := \sum_{i=1}^{n} S(x_i)$$

2. Application to engine testing

In the case of engine testing the test statistic of choice is the match score [3]. This is obvious for the traditional fixed length tests but it is also true, albeit less obvious, for the SPRT test as implemented for example in `cutechess-cli`. We will show that under some common scenarios and assumptions the match score contains the same information as the score statistic introduced above. Hence it is an optimal test statistic in a suitable sense.

2.1. **The trinomial model.** Assume that the outcome of a game between two engines is modeled by a trinomial distribution with outcome probabilities $(w(\theta), d(\theta), l(\theta)$ where θ is the elo difference in some unspecified elo model. We assume that $\theta = 0$

[1] In the context of this note the standard terminology "score statistic" is somewhat unfortunate as we will also be talking about a "match score".

corresponds to engines of equal strength. We also assume that test conditions are fair, i.e.

$$\begin{aligned} w(\theta) &= l(-\theta) \\ d(\theta) &= d(-\theta) \\ l(\theta) &= w(-\theta) \end{aligned} \tag{2.1}$$

These assumptions are reasonable in the following contexts

(1) Randomly chosen opening positions and randomly assigned colors.
(2) (Nearly) balanced randomly chosen opening positions and paired games with reversed colors.[2]

Under the trinomial model we have

$$S(\text{win}) = \left.\frac{d\log w(\theta)}{d\theta}\right|_{\theta=0} \qquad S(\text{draw}) = \left.\frac{d\log d(\theta)}{d\theta}\right|_{\theta=0} \qquad S(\text{loss}) = \left.\frac{d\log l(\theta)}{d\theta}\right|_{\theta=0}$$

From (2.1) we easily deduce

$$\begin{aligned} \lambda := S(\text{win}) &= -S(\text{draw}) \\ S(\text{draw}) &= 0 \end{aligned}$$

so that the score statistic of a match with outcome frequencies (W, D, L) is given as $(W-L)\lambda$. Since the constant scale factor λ does not matter, the quantity $(W-L)\lambda$ contains the same information as the match score $\hat{w} + (1/2)\hat{d} = 1/2 + 1/2(\hat{w} - \hat{l})$.

2.2. **Paired games.** Now we assume that we are using paired games with (possibly) unbalanced positions. Let us write the outcome probabilities in a game pair as $(w_1(\theta), d_1(\theta), l_1(\theta), w_2(\theta), d_2(\theta), l_2(\theta))$. We assume that there are no correllations between paired games. The fairness assumption can be expressed as

$$\begin{aligned} w_1(\theta) &= l_2(-\theta) \\ d_1(\theta) &= d_2(-\theta) \\ l_1(\theta) &= w_2(-\theta) \end{aligned} \tag{2.2}$$

We now write

$$S_1(\text{win}) = \left.\frac{d\log w_1(\theta)}{d\theta}\right|_{\theta=0} \qquad S_1(\text{draw}) = \left.\frac{d\log d_1(\theta)}{d\theta}\right|_{\theta=0} \qquad \text{etc}\ldots$$

Then we get from (2.2)

$$\begin{aligned} a := S_1(\text{win}) &= -S_2(\text{loss}) \\ b := S_1(\text{draw}) &= -S_2(\text{draw}) \\ c := S_1(\text{loss}) &= -S_2(\text{win}) \end{aligned}$$

Hence a game pair should now be scored according to the following table

$$\begin{array}{cccc} & \text{win} & \text{draw} & \text{loss} \\ \text{win} & a-c & a-b & 0 \\ \text{draw} & b-c & 0 & b-a \\ \text{loss} & 0 & c-b & c-a \end{array} \tag{2.3}$$

[2] Even though we assume the positions are nearly balanced it would still be very hard to ensure that the average position bias in an opening book is perfectly zero (this would be a mathematical fiction anyway). This is why it is still necessary to replay games with reversed colors in order to achieve fairness.

At this point it seems we cannot continue to work in complete generality and we have to assume a specific underlying elo model. Under the Davidson elo model [2] we have

$$b = \frac{a+c}{2}$$

In that case if we put $\lambda := \frac{1}{2}(a-c)$, (2.3) can be written as

	win	draw	loss
win	2λ	λ	0
draw	λ	0	$-\lambda$
loss	0	$-\lambda$	-2λ

and we see that the score statistic contains again the same information as the match score.

There are indications that the Davidson elo model fits reality better than the more commonly used Rao-Kupper[3] elo model [2]. However as far as I know this has not been firmly established in the specific case of engine-engine matches with unbalanced opening positions. So for completeness let us describe what happens for the Rao-Kupper elo model. In that case we have

$$b = a + c$$

and now (2.3) becomes

	win	draw	loss
win	$a-c$	$-c$	0
draw	a	0	c
loss	0	$-a$	$c-a$

Now if $a + c \neq 0$ (as would be the case with unbalanced positions) the score statistic is different from the match score. Note however that in order to use the true score statistic we would have to know the ratio $-a/c$ which depends on the drawelo parameter and the average position bias, both of which would have to be estimated from the sample. This leads to a test statistic which is substantially more complicated to compute than the simple match score.

As indicated in the introduction the relative efficiency of the true score statistic (assuming the Rao-Kupper elo model) versus the naive match score statistic can be evaluated empirically. Such an analysis has not been carried out yet.

References

1. P. Bartlett, *Theoretical Statistics. Lecture 24.*, http://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/24notes.pdf, 2013.
2. D. Shawul1 and R. Coulom, *Paired Comparisons with Ties: Modeling Game Outcomes in Chess* , http://www.grappa.univ-lille3.fr/~coulom/ChessOutcomes.pdf.
3. Chess Programming WIKI, *Match statistics*, https://chessprogramming.wikispaces.com/Match+Statistics.
4. Wikipedia, *Score (statistics)*, https://en.wikipedia.org/wiki/Score_(statistics).

[3] The Rao-Kupper model is implemented in the software package “BayesElo”.