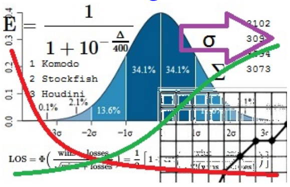
Home * Engine Testing * Match Statistics



Match Statistics [1]

Match Statistics,
the statistics of chess tournaments and matches, that is a collection of chess games and the presentation, analysis, and interpretation of game related data, most common game results to determine the relative playing strength of chess playing entities, here with focus on chess engines. To apply match statistics, beside considering statistical population, it is conventional to hypothesize a statistical model describing a set of probability distributions.

Table of Contents

Ratios / Operating Figures

Common tools, ratios and figures to illustrate a tournament outcome and provide a base for its interpretation.

Number of games

The total number of games played by an engine in a tournament.

-

Score

The score is a representation of the tournament-outcome from the viewpoint of a certain engine.

-
-

Win & Draw Ratio

-
-

These two ratios depend on the strength difference between the competitors, the average strength level, the color and the drawishness of the opening book-line. Due to the second reason given, these ratios are very much influenced by the timecontrol, what is also confirmed by the published statistics of the testing orgnisations CCRL and CEGT, showing an increase of the draw rate at longer time controls. This correlation was also shown by Kirill Kryukov, who was analyzing statistics of his test-games [2] . The program playing white seems to be more supported by the additional level of strength. So, although one would expect with increasing draw rates the win ratio to approach 50%, in fact it is remaining about equal.

Timecontrol	Draw Ratio	Win Ratio (white)	Source
40/4	30.9%	55.0%	CEGT
40/20	35.6%	54.6%	CEGT
40/120	41.3%	55.4%	CEGT
40/120 (4cpu)	45.2%	55.9%	CEGT

Timecontrol	Draw Ratio	Win Ratio (white)	Source
40/4	31.0%	54.1%	CCRL
40/40	37.2%	54.6%	CCRL

Doubling Time Control

As posted in October 2016 [3] , Andreas Strangmüller conducted an experiment with Komodo 9.3, time control doubling matches under Cutechess-cli, playing 3000 games with 1500 opening positions each, without pondering, learning, and tablebases, Intel i5-750 @ 3.5 GHz, 1 Core, 128 MB Hash [4] , see also Kai Laskos' 2013 results with Houdini 3 [5] and Diminishing Returns:

Time Control 2 vs 1	20+0.2 10+0.1	40+0.4 20+0.2	80+0.8 40+0.4	160+1.6 80+0.8	320+3.2 160+1.6	640+6.4 320+3.2	1280+12.8 640+6.4	2560+25.6 1280+12.8
Elo	144	133	112	101	93	73	59	51
Win	44.97 %	41.27 %	36.67 %	32.67 %	30.47 %	25.17 %	21.77 %	18.97 %
Draw	49.20 %	54.00 %	57.93 %	63.03 %	65.33 %	70.47 %	73.17 %	76.63 %
Loss	5.83 %	4.73 %	5.40 %	4.30 %	4.20 %	4.37 %	5.07 %	4.40 %

Elo-Rating & Win-Probability

see Pawn Advantage, Win Percentage, and Elo

-
-
-
-

Generalization of the Elo-Formula:
win_probability of player i in a tournament with n players

-

Likelihood of Superiority

See LOS Table

The likelihood of superiority (LOS) denotes how likely it would be for two players of the same strength to reach a certain result - in other fields called a p-value, a measure of statistical significance of a departure from the null hypothesis [6]. Doing this analysis after the tournament one has to differentiate between the case where one knows that a certain engine is either stronger or equally strong (directional or one-tailed test) or the case where one has no information of whether the other engine is stronger or weaker (non-directional or two-tailed test). The latter due to the reduced information results in larger confidence intervals.

Two-tailed Test
Null- and alternative hypothesis:

-
-
-

The probability of the null hypothesis being true can be calculated given the tournament outcome. In other words, how likely would it be for two players of the same strength to reach a certain result. The LOS would then be the inverse, 1 - the resulting probability.

For this type of analysis the trinomial distribution, a generalization of the binomial distribution, is needed. Whilest the binomial distribution can only calculate the probability to reach a certain outcome with two possible events, the trinominal distribution can account for all three possible events (win, draw, loss).

The following functions gives the probability of a certain game outcome assuming both players were of equal strength:

-
-

This calculation becomes very inefficient for larger number of games. In this case the standard normal distribution can give a good approximation:

-

where N(1 - draw_ratio) is the sum of wins and losses:

-

To calculate the LOS one needs the cumulative distribution function of the given normal distribution. However, as pointed out by Rémi Coulom, calculation can be done cleverly, and the normal approximation is not really required [7] . As further emphasized by Kai Laskos [8] and Rémi Coulom [9] [10] , draws do not count in LOS calculation and don't make a difference whether the game results were obtained when playing Black or White. It is a good approximation when the two players played the same number of games with each color:

-

[11] [12] [13]

One-tailed Test
Null- and alternative hypothesis:

-
-

Sample Program
A tiny C++11 program to compute Elo difference and LOS from W/L/D counts was given by Álvaro Begué [14] :

```
#include <cstdio>
#include <cstdlib>
#include <cmath>

int main(int argc, char **argv) {
  if (argc != 4) {
    std::printf("Wrong number of arguments.\n\n
```

```
Usage:%s <wins> <losses> <draws>\n", argv[0]);
    return 1;
  }
  int wins = std::atoi(argv[1]);
  int losses = std::atoi(argv[2]);
  int draws = std::atoi(argv[3]);

  double games = wins + losses + draws;
  std::printf("Number of games: %g\n", games);
  double winning_fraction = (wins + 0.5*draws) / games;
  std::printf("Winning fraction: %g\n", winning_fraction);
  double elo_difference = -std::log(1.0/
winning_fraction-1.0)*400.0/std::log(10.0);
  std::printf("Elo difference: %+g\n", elo_difference);
  double los = .5 + .5 * std::erf((wins-losses)/std::sqrt(2.0*(
wins+losses)));
  std::printf("LOS: %g\n", los);
}
```

Statistical Analysis

The trinomial versus the 5-nomial model

As indicated above a match between two engines is usually modeled as a sequence of independent trials taken from a trinomial distribution with probabilities (win_ratio,draw_ratio,loss_ratio). This model is appropriate for a match with randomly selected opening positions and randomly assigned colors (to maintain fairness). However one may show that under reasonable elo models the trinomial model is not correct in case games are played in pairs with reversed colors (as is commonly the case) and unbalanced opening positions are used.

This was also empirically observed by Kai Laskos [15] . He noted that the statistical predictions of the trinomial model do not match reality very well in the case of paired games. In particular he observed that for some data sets the variance of the match score as predicted by the trinomial model greatly exceeds the variance as calculated by the jackknife estimator. The jackknife estimator is a non-parametric estimator, so it does not depend on any particular statistical model. It appears the mismatch may even occur for balanced opening positions, an effect which can only be explained by the existence of correlations between paired games - something not considered by any elo model.

Over estimating the variance of the match score implies that derived quantities such as the number of games required to establish the superiority of one engine over another with a given level of significance are also over estimated. To obtain agreement between statistical predictions and actual measurements one may adopt the more general 5-nomial model. In the 5-nomial model the outcome of paired games is assumed to follow a 5-nomial distribution with probabilities

-

These unknown probabilities may be estimated from the outcome frequencies of the paired games and then subsequently be used to compute an estimate for the variance of the match score. Summarizing: in the case of paired games the 5-nomial model handles the following effects correctly which the trinomial model does not:

- Unbalanced openings
- Correlations between paired games

For further discussion on the potential use of unbalanced opening positions in engine testing see the posting by Kai Laskos [16] .

SPRT

The sequential probability ratio test (SPRT) is a specific sequential hypothesis test - a statistical analysis where the sample size is not fixed in advance - developed by Abraham Wald [17] . While originally developed for use in quality control studies in the realm of manufacturing, SPRT has been formulated for use in the computerized testing of human examinees as a termination criterion [18]. As mentioned by Arthur Guez in this 2015 Ph.D. thesis *Sample-based Search Methods for Bayes-Adaptive Planning* [19], Alan Turing assisted by Jack Good used a similar sequential testing technique to help decipher enigma codes at Bletchley Park [20]. SPRT is applied in Stockfish testing to terminate self-testing series early if the result is likely outside a given elo-window [21] . In August 2016, Michel Van den Bergh posted following Python code in CCC to implement the SPRT a la Cutechess-cli or Fishtest: [22] [23]

```
from __future__ import division

import math

def LL(x):
    return 1/(1+10**(-x/400))

def LLR(W,D,L,elo0,elo1):
    """
This function computes the log likelihood ratio of H0:elo_diff=elo0 ve
rsus
H1:elo_diff=elo1 under the logistic elo model

expected_score=1/(1+10**(-elo_diff/400)).

W/D/L are respectively the Win/Draw/Loss count. It is assumed that the
```

```
 outcomes of
the games follow a trinomial distribution with probabilities (w,d,l).
Technically
this is not quite an SPRT but a so-
called GSPRT as the full set of parameters (w,d,l)
cannot be derived from elo_diff, only w+(1/2)d. For a description and
properties of
the GSPRT (which are very similar to those of the SPRT) see

http://stat.columbia.edu/~jcliu/paper/GSPRT_SQA3.pdf

This function uses the convenient approximation for log likelihood
ratios derived here:

http://hardy.uhasselt.be/Toga/GSPRT_approximation.pdf

The previous link also discusses how to adapt the code to the 5-nomial
 model
discussed above.
"""
# avoid division by zero
    if W==0 or D==0 or  L==0:
        return 0.0
    N=W+D+L
    w,d,l=W/N,D/N,L/N
    s=w+d/2
    m2=w+d/4
    var=m2-s**2
    var_s=var/N
    s0=LL(elo0)
    s1=LL(elo1)
    return (s1-s0)*(2*s-s0-s1)/var_s/2.0

def SPRT(W,D,L,elo0,elo1,alpha,beta):
    """
This function sequentially tests the hypothesis H0:elo_diff=elo0 versu
s
the hypothesis H1:elo_diff=elo1 for elo0<elo1. It should be called aft
er
each game until it returns either 'H0' or 'H1' in which case the test
stops
and the returned hypothesis is accepted.

alpha is the probability that H1 is accepted while H0 is true
(a false positive) and beta is the probability that H0 is accepted
while H1 is true (a false negative). W/D/L are the current win/draw/lo
```

```
ss
counts, as before.
"""
    LLR_=LLR(W,D,L,elo0,elo1)
    LA=math.log(beta/(1-alpha))
    LB=math.log((1-beta)/alpha)
    if LLR_>LB:
        return 'H1'
    elif LLR_<LA:
        return 'H0'
    else:
        return ''
```

Tournament Manager

- Arena
- Amoeba Tournament Manager
- ChessGUI
- Cutechess-cli
- LittleBlitzer

See also

- Automated Tuning
- Bishop versus Knight - Winning Percentages
- Chess Server
- Depth | Diminishing Returns
- Draw
- Engine Rating Lists
- LOS Table
- Pawn Advantage, Win Percentage, and Elo
- Playing Strength
- Search Statistics
- Time Controls
- Who is the Master?

Publications

1920 ...

- Ernst Zermelo (**1929**). *Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der*

Wahrscheinlichkeitsrechnung. pdf (German)
- Abraham Wald (**1945**). *Sequential Tests of Statistical Hypotheses*. Annals of Mathematical Statistics, Vol. 16, No. 2, doi: 10.1214/aoms/1177731118
- Abraham Wald (**1947**). *Sequential Analysis*. John Wiley and Sons, AbeBooks
- Ralph A. Bradley, Milton E. Terry (**1952**). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika, Vol. 39, Nos. 3/4, doi: 10.2307/2334029, JSTOR 2334029

1960 ...

- Florence Nightingale David (**1962**). Games, Gods & Gambling: A History of Probability and Statistical Ideas. Dover Publications, ISBN-13: 978-0486400235
- Tony Marsland, Paul Rushton (**1973**). Mechanisms for Comparing Chess Programs. ACM Annual Conference, pdf
- James Gillogly (**1978**). *Performance Analysis of the Technology Chess Program*. Ph.D. Thesis. Tech. Report CMU-CS-78-189, Carnegie Mellon University, CMU-CS-77 pdf » Tech
- Arpad Elo (**1978**). *The Rating of Chessplayers, Past and Present*. Arco Publications [24]
- David Cahlander (**1979**). *Strength of a Chess Playing Computer*. ICCA Newsletter, Vol. 2, No. 1
- Jack Good (**1979**). *On the Grading of Chess Players*. Personal Computing, Vol. 3, No. 3, pp. 47

1980 ...

- John F. White (**1981**). Survey-Chess Games. Your Computer, August/September 1981 [25]
- Ken Thompson (**1982**). *Computer Chess Strength*. Advances in Computer Chess 3
- David Siegmund (**1985**). *Sequential Analysis. Tests and confidence intervals*. Springer
- Hans Berliner, Gordon Goetsch, Murray Campbell, Carl Ebeling (**1989**). *Measuring the Performance Potential of Chess Programs*, Advances in Computer Chess 5
- Eric Hallsworth (**1989**). *Playing Levels*. Computer Chess News Sheet 23, pp 2, pdf hosted by Mike Watters

1990 ...

- Hans Berliner, Gordon Goetsch, Murray Campbell, Carl Ebeling (**1990**). *Measuring the Performance Potential of Chess Programs.* Artificial Intelligence, Vol. 43, No. 1
- Hal Stern (**1990**). *Are all Linear Paired Comparison Models Equivalent*. pdf
- Eric Hallsworth (**1990**). *Speed, Processors and Ratings*. Computer Chess News Sheet 25, pp 6, pdf hosted by Mike Watters
- Hans Berliner, Danny Kopec, Ed Northam (**1991**). *A taxonomy of concepts for evaluating chess strength: examples from two difficult categories*. Advances in Computer Chess 6, pdf
- Steve Maughan (**1992**). *Are You Sure It's Better?* Selective Search 40, pp. 21, pdf hosted by Mike Watters
- Warren D. Smith (**1993**). *Rating Systems for Gameplayers, and Learning*. ps
- Robert Hyatt, Monroe Newborn (**1997**). *CRAFTY Goes Deep*. ICCA Journal, Vol. 20, No. 2 » Crafty

2000 ...

- Ernst A. Heinz (**2000**). New Self-Play Results in Computer Chess. CG 2000
- Ernst A. Heinz (**2001**). *Self-play Experiments in Computer Chess Revisited.* Advances in Computer Games 9
- Ernst A. Heinz (**2001**). *Modeling the "Go Deep" Behaviour of CRAFTY and DARK THOUGHT.* Advances in Computer Games 9 » Crafty, Dark Thought
- Ernst A. Heinz (**2001**). *Self-Play, Deep Search and Diminishing Returns.* ICGA Journal, Vol. 24, No. 2
- Guy Haworth (**2002**). Self-play: Statistical Significance. 7th Computer Olympiad Workshop
- Guy Haworth (**2003**). Self-Play: Statistical Significance. ICGA Journal, Vol. 26, No. 2
- Ernst A. Heinz (**2003**). *Follow-Up on Self-Play, Deep Search, and Diminishing Returns.* ICGA Journal, Vol. 26, No. 2
- David R. Hunter (**2004**). *MM Algorithms for Generalized Bradley-Terry Models*. The Annals of Statistics, Vol. 32, No. 1, 384–406, pdf [26] [27] [28] [29]

2005 ...

- Jan Renze Steenhuisen (**2005**). *New Results in Deep-Search Behaviour*. ICGA Journal, Vol. 28, No. 4, pdf
- Matej Guid, Ivan Bratko (**2007**). *Factors affecting diminishing returns for searching deeper*. CGW 2007 » Crafty, Rybka, Shredder, Diminishing Returns
- Jeff Rollason (**2007**). Statistical Minefields with Version Testing. AI Factory, Winter 2007 » Engine Testing
- Shogo Takeuchi, Tomoyuki Kaneko, Kazunori Yamaguchi, Satoru Kawai (**2007**). *Visualization and Adjustment of Evaluation Functions Based on Evaluation Values and Win Probability*. AAAI 2007, pdf
- Rémi Coulom (**2008**). Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. CG 2008, draft as pdf
- Giuseppe Di Fatta, Guy McCrossan Haworth, Kenneth Wingate Regan (**2009**). *Skill Rating by Bayesian Inference*. CIDM 2009, pdf [30]
- Guy McCrossan Haworth, Kenneth Wingate Regan, Giuseppe Di Fatta (**2009**). *Performance and Prediction: Bayesian Modelling of Fallible Choice in Chess*. Advances in Computer Games 12, pdf

2010 ...

- Diogo R. Ferreira (**2010**). Predicting the Outcome of Chess Games based on Historical Data. IST - Technical University of Lisbon [31]
- Kenneth W. Regan, Guy Haworth (**2011**). Intrinsic Chess Ratings. AAAI 2011, pdf, slides as pdf [32]
- Kenneth Wingate Regan, Bartłomiej Macieja, Guy McCrossan Haworth (**2011**). Understanding Distributions of Chess Performances. Advances in Computer Games 13, pdf
- Trevor Fenner, Mark Levene, George Loizou (**2011**). *A Discrete Evolutionary Model for Chess Players' Ratings*. Physics and Society, arXiv:1103.1530v2
- Rémi Coulom (**2012**). *Paired Comparisons with Ties: Modeling Game Outcomes in Chess*. pdf

preprint [33]
- Diogo R. Ferreira (**2012**). *Determining the Strength of Chess Players based on actual Play*. ICGA Journal, Vol. 35, No. 1
- Diogo R. Ferreira (**2013**). *The Impact of the Search Depth on Chess Playing Strength*. ICGA Journal, Vol. 36, No. 2
- Miguel A. Ballicora (**2014**). *ORDO v0.9.6 Ratings for chess and other games*. September 2014, pdf » Ordo [34]
- Don Dailey, Adam Hair, Mark Watkins (**2014**). Move Similarity Analysis in Chess Programs. Entertainment Computing, Vol. 5, No. 3, preprint as pdf [35]
- Kenneth W. Regan, Tamal T. Biswas, Jason Zhou (**2014**). *Human and Computer Preferences at Chess*. pdf
- Erik Varend (**2014**). *Quality of play in chess and methods for measuring*. pdf [36] [37]

2015 ...

- Tamal T. Biswas, Kenneth W. Regan (**2015**). *Quantifying Depth and Complexity of Thinking and Knowledge*. ICAART 2015, pdf
- Tamal T. Biswas, Kenneth W. Regan (**2015**). *Measuring Level-K Reasoning, Satisficing, and Human Error in Game-Play Data*. IEEE ICMLA 2015, pdf preprint
- Guy Haworth, Tamal T. Biswas, Kenneth W. Regan (**2015**). A Comparative Review of Skill Assessment: Performance, Prediction and Profiling. Advances in Computer Games 14
- Shogo Takeuchi, Tomoyuki Kaneko (**2015**). Estimating Ratings of Computer Players by the Evaluation Scores and Principal Variations in Shogi. ACIT-CSI
- Jean-Marc Alliot (**2017**). *Who is the Master*? ICGA Journal, Vol. 39, No. 1, draft as pdf » Stockfish , Who is the Master?

Forum & Blog Postings

1996 ...

- Theoretical chess rating question... by Cyber Linguist, rgcc, April 17, 1996
- Statistical validity of medium-length match results by Bruce Moreland, rgcc, February 15, 1997
- ELO performance? by Stefan Meyer-Kahlen, CCC, May 22, 1999 » Pawn Advantage, Win Percentage, and Elo, Playing Strength

2000 ...

- Some thougths about statistics by Martin Schubert, CCC, February 24, 2000
- Who is better? Some statistics... by Gian-Carlo Pascutto, CCC, June 11, 2001
- Simulating the result of a single game by random numbers by Christoph Fieberg, CCC, July 03, 2001
- Simulating the result of a single game by random numbers - Update! by Christoph Fieberg, CCC, July 10, 2001
- Simulating the result of a single game by random numbers - Update! by Christoph Fieberg, CCC,

August 02, 2001
- ELO & statistics question by Gian-Carlo Pascutto, CCC, April 26, 2002
- Statistical significance of a match result by Rémi Coulom, CCC, November 23, 2002
- Value of playing different versions of a program against each other by Tom King, CCC, January 06, 2003
- A question about statistics... by Roger Brown, CCC, January 04, 2004
- New tool to estimate the statistical significance of match results by Rémi Coulom, CCC, July 17, 2004
- ELOStat algorithm ? by Rémi Coulom, Winboard Forum, December 10, 2004 » EloStat

2005 ...

- bayeselo: new Elo-rating tool, applied to CCT7 by Rémi Coulom, CCC, February 13, 2005 » CCT7
- table for detecting significant difference between two engines by Joseph Ciarrochi, CCC, February 03, 2006 [38]
- Observator bias or... by Alessandro Scotti, CCC, May 29, 2007
- a beat b,b beat c,c beat a question by Uri Blass, CCC, May 16, 2007
- how to do a proper statistical test by Rein Halbersma, CCC, September 19, 2007
- Elo Calcuation by Edmund Moshammer, CCC, April 19, 2009
- Likelihood of superiority by Marco Costalba, CCC, November 15, 2009

2010 ...

- Engine Testing - Statistics by Edmund Moshammer, CCC, January 14, 2010
  Re: Engine Testing - Statistics by John Major, CCC, January 14, 2010
- Chess Statistics by Edmund Moshammer, CCC, June 17, 2010
- Do You really need 1000s of games for testing? by Jouni Uski, CCC, November 04, 2010
- GUI idea: Testing until certainty by Albert Silver, CCC, December 07, 2010
- SPRT and Engine testing by Adam Hair, CCC, December 13, 2010 » SPRT

2011

- Ply vs ELO by Andriy Dzyben, CCC, June 28, 2011
- One billion random games by Steven Edwards, CCC, August 27, 2011
- Increase in Elo ..Question For The Experts by Steve B, CCC, December 05, 2011

2012

- Advantage for White; Bayeselo (to Rémi Coulom) by Edmund Moshammer, CCC, March 03, 2012
- Human Elo ratings: averages and standard deviations by Jesús Muñoz, CCC, March 18, 2012 [39]
- Elo uncertainties calculator by Jesús Muñoz, CCC, March 24, 2012
- Elo versus speed by Peter Österlund, CCC, April 02, 2012

- Rybka odds matches and the strength of engines by Kai Laskos, CCC, June 09, 2012 » Rybka
- A new way to compare chess programs by Larry Kaufman, CCC, June 21, 2012 » Komodo
- EloStat, Bayeselo and Ordo by Kai Laskos, CCC, June 24, 2012 » EloStat, Bayeselo, Ordo
- about error margins? by Fermin Serrano, CCC, August 01, 2012
- normal vs logistic curve for elo model by Daniel Shawul, CCC, August 02, 2012
- Derivation of bayeselo formula by Rémi Coulom, CCC, August 07, 2012 [40]
- Yet Another Testing Question by Brian Richardson, CCC, September 15, 2012
- margin of error by Larry Kaufman, CCC, September 16, 2012
- Average number of plies in {1-0, ½-½, 0-1} by Jesús Muñoz, CCC, September 21, 2012
- Another testing question by Larry Kaufman, CCC, September 23, 2012
- LOS calculation: Does the same result is always the same? by Marco Costalba, CCC, October 01, 2012
- LOS (again) by Ed Schröder, CCC, October 30, 2012
- Elo points gain from doubling time by Kai Laskos, CCC, December 10, 2012
- A word for casual testers by Don Dailey, CCC, December 25, 2012

2013

- A poor man's testing environment by Ed Schröder, CCC, January 04, 2013 [41] » Engine Testing
- Noise in ELO estimators: a quantitative approach by Marco Costalba, CCC, January 06, 2013
- Updated Dendrogram by Kai Laskos, CCC, February 02, 2013
- Experiment: influence of colours at fixed depth by Jesús Muñoz, CCC, March 10, 2013
- LOS by BB+, OpenChess Forum, March 31, 2013
- Fishtest Distributed Testing Framework by Marco Costalba, CCC, May 01, 2013
- The influence of the length of openings by Kai Laskos, CCC, July 14, 2013
- Scaling at 2x nodes (or doubling time control) by Kai Laskos, CCC, July 23, 2013 » Doubling TC, Diminishing Returns, Playing Strength, Houdini
- Type I error in LOS based early stopping rule by Kai Laskos, CCC, August 06, 2013 [42]
- How much elo is pondering worth by Michel Van den Bergh, CCC, August 07, 2013 » Pondering
- Contempt and the ELO model by Michel Van den Bergh, CCC, September 05, 2013 » Contempt Factor
- 1 draw=1 win + 1 loss (always!) by Michel Van den Bergh, CCC, September 19, 2013
- SPRT and narrowing of (elo1 - elo0) difference by Jesús Muñoz, CCC, October 05, 2013 » SPRT
- sprt and margin of error by Larry Kaufman, CCC, October 15, 2013 » SPRT
- How (not) to use SPRT ? by BB+, OpenChess Forum, October 19, 2013
- Houdini, much weaker engines, and Arpad Elo by Kai Laskos, CCC, November 29, 2013 » Houdini, Pawn Advantage, Win Percentage, and Elo [43]
- Testing on time control versus nodes | ply by Ed Schröder, CCC, December 04, 2013

2014

- Calculating the LOS (likelihood of superiority) from results by Robert Tournevisse, CCC, January 22, 2014

- LOS --> Draws are irrelevant by User923005, OpenChess Forum, January 24, 2014
- Empirically 1 win + 1 loss ~ 2 draws by Kai Laskos, CCC, June 24, 2014
- Ordo 0.9.6 by Miguel A. Ballicora, CCC, September 10, 2014 » Ordo
- Using the Stockfish position evaluation score to predict victory probability by unavoidablegrain, Tumblr, September 28, 2014 » Pawn Advantage, Win Percentage, and Elo, Stockfish
- Elo estimation using quasi-Monte Carlo integration by Branko Radovanovic, CCC, September 30, 2014
- SPRT question by Robert Hyatt, CCC, November 13, 2014 » SPRT
- Usage sprt / cutechess-cli by Michael Hoffmann, CCC, November 16, 2014 » Cutechess-cli, SPRT

2015 ...

- 2-SPRT by Michel Van den Bergh, CCC, January 28, 2015 » SPRT
- Script for computing SPRT probabilities by Michel Van den Bergh, CCC, April 05, 2015
- Maximum ELO gain per test game played? by Forrest Hoch, CCC, April 20, 2015
- Getting SPRT right by Alexandru Mosoi, CCC, April 22, 2015 » SPRT
- SPRT questions by Uri Blass, CCC, May 15, 2015 » SPRT
- Adam Hair's article on Pairwise comparison of engines by Charles Roberson, CCC, May 19, 2015 [44]
- computing elo of multiple chess engines by Alexandru Mosoi, CCC, August 09, 2015
- Some musings about search by Ed Schröder, CCC, August 14, 2015 » Automated Tuning, Search
- Bullet vs regular time control, say 40/4m CCRL/CEGT by Ed Schröder, CCC, August 29, 2015
- The SPRT without draw model, elo model or whatever... by Michel Van den Bergh, CCC, September 01, 2015 » SPRT
  Re: The SPRT without draw model, elo model or whatever.. by Michel Van den Bergh, CCC, August 18, 2016
- Name for elo without draws? by Marcel van Kervinck, CCC, September 02, 2015
- The future of chess and elo ratings by Larry Kaufman, CCC, September 20, 2015 » Opening Book
- Depth of Satisficing by Ken Regan, Gödel's Lost Letter and P=NP, October 06, 2015 » Depth, Match Statistics, Pawn Advantage, Win Percentage, and Elo, Stockfish, Komodo [45]
- ELO error margin by Fabio Gobbato, CCC, October 17, 2015
- testing multiple versions & elo calculation by Folkert van Heusden, CCC, October 27, 2015
- A simple expression by Kai Laskos, CCC, December 09, 2015
- Counting 1 win + 1 loss as 2 draws by Kai Laskos, CCC, December 15, 2015

2016

- A Chess Firewall at Zero? by Ken Regan, Gödel's Lost Letter and P=NP, January 21, 2016
- Ordo 1.0.9 (new features for testers) by Miguel A. Ballicora, CCC, January 25, 2016
- Why the errorbar is wrong ... simple example! by Frank Quisinsky, CCC, February 23, 2016
- a direct comparison of FIDE and CCRL rating systems by Erik Varend, CCC, February 22, 2016 » FIDE, CCRL
- Some properties of the Type I error in p-value stopping rule by Kai Laskos, CCC, March 01, 2016
- A Visual Look at 2 Million Chess Games - Thinking Through the Party by Buğra Fırat, March 02, 2016

- Type I error for p-value stopping: Balanced and Unbalanced by Kai Laskos, CCC, June 16, 2016
- Empirically Logistic ELO model better suited than Gaussian by Kai Laskos, CCC, July 12, 2016
- Testing resolution and combining results by Daniel José Queraltó, CCC, July 28, 2016
- Error margins via resampling (jackknifing) by Kai Laskos, CCC, August 12, 2016 [46] [47]
- Properties of unbalanced openings using Bayeselo model by Kai Laskos, CCC, August 27, 2016 » Opening Book
- ELO inflation ha ha ha by Henk van den Belt, CCC, September 16, 2016 » Delphil, Stockfish, Playing Strength, TCEC Season 9 [48]
  About expected scores and draw ratios by Jesús Muñoz, CCC, September 17, 2016
- The scaling with time of opening books by Kai Laskos, CCC, September 23, 2016 » Opening Book
- Perfect play by Patrik Karlsson, CCC, September 28, 2016
- Stockfish underpromotes much more often than Komodo by Kai Laskos, CCC, October 05, 2016 » Komodo, Promotions, Stockfish
- Differences between top engines related to "style" by Kai Laskos, October 07, 2016
- SPRT when not used for self testing by Andrew Grant, CCC, October 21, 2016
- Doubling of time control by Andreas Strangmüller, CCC, October 21, 2016 » Doubling TC, Diminishing Returns, Playing Strength, Komodo
- Stockfish 8 - Double time control vs. 2 threads by Andreas Strangmüller, CCC, November 15, 2016 » Doubling TC, Diminishing Returns, Playing Strength, Stockfish
- When Data Serves Turkey by Ken Regan, Gödel's Lost Letter and P=NP, November 30, 2016
- Magnus and the Turkey Grinder by Ken Regan, Gödel's Lost Letter and P=NP, December 08, 2016 » Pawn Advantage, Win Percentage, and Elo [49]
- Regan's conundrum by Carl Lumma, CCC, December 09, 2016
- Statistical Interpretation by Dennis Sceviour, CCC, December 10, 2016
- Absolute ELO scale by Nicu Ionita, CCC, December 17, 2016
- A question about SPRT by Andrew Grant, CCC, December 25, 2016 » SPRT
- Diminishing returns and hyperthreading by Kai Laskos, CCC, December 27, 2016 » Diminishing Returns, Playing Strength, Thread

2017

- Progress in 30 years by four intervals of 7-8 years by Kai Laskos, CCC, January 19, 2017 » Playing Strength
- sprt tourney manager by Richard Delorme, CCC, January 24, 2017 » Amoeba Tournament Manager, SPRT
- Binomial distribution for chess statistics by Lyudmil Antonov, CCC, March 03, 2017
- Higher than expected by me efficiency of Ponder ON by Kai Laskos, CCC, March 06, 2017 » Pondering
- What can be said about 1 - 0 score? by Kai Laskos, CCC, March 28, 2017
- 6-men Syzygy from HDD and USB 3.0 by Kai Laskos, CCC, April 04, 2017 » Komodo, Playing Strength, Syzygy Bases, USB 3.0
- Scaling of engines from FGRL rating list by Kai Laskos, CCC, April 07, 2017 » FGRL
- Low impact of opening phase in engine play? by Kai Laskos, CCC, April 18, 2017 » Opening
- How to simulate a game outcome given Elo difference? by Nicu Ionita, CCC, April 25, 2017

- Wilo rating properties from FGRL rating lists by Kai Laskos, CCC, May 01, 2017 » FGRL
- MATCH sanity by Ed Schroder, CCC, May 03, 2017 » Portable Game Notation
- Symmetric multiprocessing (SMP) scaling - SF8 and K10.4 by Andreas Strangmüller, CCC, May 05, 2017 » Lazy SMP, Komodo, Stockfish
- Symmetric multiprocessing (SMP) scaling - K10.4 Contempt=0 by Andreas Strangmüller, CCC, May 11, 2017 » SMP, Komodo, Contempt Factor
- Symmetric multiprocessing (SMP) scaling - SF8 Contempt=10 by Andreas Strangmüller, CCC, May 13, 2017 » SMP, Stockfish, Contempt Factor
- Likelihood Of Success (LOS) in the real world by Kai Laskos, CCC, May 26, 2017
- Opening testing suites efficiency by Kai Laskos, CCC, June 21, 2017 » Engine Testing, Opening
- Testing A against B by playing a pool of others by Andrew Grant, CCC, June 24, 2017
- Engine testing & error margin ? by Mahmoud Uthman, CCC, July 05, 2017
- Invariance with time control of rating schemes by Kai Laskos, CCC, July 22, 2017 [50]
- Ways to avoid "Draw Death" in Computer Chess by Kai Laskos, CCC, July 25, 2017
- SMP NPS measurements by Peter Österlund, CCC, August 06, 2017 » Lazy SMP, Parallel Search, Nodes per second
  ELO measurements by Peter Österlund, CCC, August 06, 2017 » Playing Strength
- What is a Match? by Henk van den Belt, CCC, September 01, 2017
- Scaling from FGRL results with top 3 engines by Kai Laskos, CCC, September 26, 2017 » FGRL, Houdini, Komodo, Stockfish
- Statistical interpretation of search and eval scores by J. Wesley Cleveland, CCC, November 18, 2017 » Pawn Advantage, Win Percentage, and Elo, Score
- "Intrinsic Chess Ratings" by Regan, Haworth -- seq by Kai Middleton, CCC, November 19, 2017
  Re: "Intrinsic Chess Ratings" by Regan, Haworth -- by Kenneth Regan, CCC, November 20, 2017 » Who is the Master?
- ELO progression measured by year by Ed Schroder, CCC, December 13, 2017

2018

- Wrong use of SPRT by Uri Blass, FishCooking, February 09, 2018 » Contempt Factor, SPRT
- Feed bayeselo with pure game results without PGN by Sergei S. Markoff, CCC, March 08, 2018
- Elo measurement of contempt in SF in self-play by Michel Van den Bergh, CCC, March 10, 2018 » Contempt, Playing Strength, Stockfish
- Time control envelope in top engines could be improved? by Kai Laskos, CCC, March 13, 2018 » Time Management
- LCZero: Progress and Scaling. Relation to CCRL Elo by Kai Laskos, CCC, March 28, 2018 » LCZero
- Elostat Question by Michael Sherwin, CCC, March 30, 2018

External Links

- Testing a chess engine from the ground up from Home of the Dutch Rebel by Ed Schröder [51] » Engine Testing
- MATCH - eng-eng utility by Ed Schröder

* Statistics of material imbalances in chess games by Alessandro Scotti » Material

Rating Systems

* Chessmetrics from Wikipedia
* Chess rating system from Wikipedia
* Elo rating system from Wikipedia
* Arpad Elo and the Elo Rating System by Dan Ross, ChessBase News, December 16, 2007
* Elo Win Probability Calculator by François Labelle
* LOS Table by Joseph Ciarrochi from CEGT [52]
* Kirr's Chess Engine Comparison KCEC - Draw rate » KCEC
* Chessanalysis homepage by Erik Varend [53]
* Who is the Master? by Jean-Marc Alliot » Who is the Master?
* How Should Chess Players Be Rated? by Martin Koppe, CNRS News, April 25, 2017
* Ranking chess players according to the quality of their moves by Frederic Friedel, ChessBase News , April 27, 2017
* Rating List Stats by Ed Schroder » CCRL

Tools

* BayesElo by Rémi Coulom builds tournament-statistics from a pgn-file
* Elostat by Frank Schubert builds tournament-statistics from a pgn-file
* Online Elo-Calculator by Pradu Kannan
* Ordo by Miguel A. Ballicora

Statistics

* Statistics from Wikipedia
* Statistical assumption from Wikipedia
* Statistical inference from Wikipedia
* Probability theory from Wikipedia
* Probability from Wikipedia
* Probability density function from Wikipedia
* Likelihood function from Wikipedia
* p-value from Wikipedia
* Misunderstandings of p-values from Wikipedia
* Probability distribution from Wikipedia
* Binomial distribution from Wikipedia
* Cumulative distribution function from Wikipedia
* Multinomial distribution from Wikipedia
* Multivariate normal distribution from Wikipedia
* Normal distribution from Wikipedia
* Standard deviation from Wikipedia
* Standard error from Wikipedia
* Error bar from Wikipedia

- Margin of error from Wikipedia
- Confidence interval from Wikipedia
- Statistical hypothesis testing from Wikipedia
- Sequential probability ratio test from Wikipedia
- Null hypothesis from Wikipedia
- Alternative hypothesis from Wikipedia
- Two-tailed test from Wikipedia
- Type I and type II errors from Wikipedia
- Type 1 Errors | Hypothesis testing with one sample | Khan Academy

Data Visualization

- A Pairwise Comparison of Chess Engine Move Selections by Adam Hair, hosted by Ed Schröder [54] [55] [56] [57]
- A Visual Look at 2 Million Chess Games - Thinking Through the Party by Buğra Fırat, March 02, 2016 [58]
- GitHub - ebemunk/chess-dataviz: chess visualization library written for d3.js by Buğra Fırat » JavaScript
- GitHub - ebemunk/pgnstats: parses PGN files and extracts statistics for chess games by Buğra Fırat » Go (Programming Language), Portable Game Notation

Misc

- Jeff Beck, Jan Hammer, Fernando Saunders, Simon Phillips - Definitely Maybe, YouTube Video ARMS Charity Concert, Madison Square Garden, December 08, 1983

References

1. ^ Image based on Standard deviation diagram by Mwtoews, April 7, 2007 with R code given, CC BY 2.5, Wikimedia Commons, Normal distribution from Wikipedia
2. ^ Kirr's Chess Engine Comparison KCEC - Draw rate » KCEC
3. ^ Doubling of time control by Andreas Strangmüller, CCC, October 21, 2016
4. ^ K93-Doubling-TC.pdf
5. ^ Scaling at 2x nodes (or doubling time control) by Kai Laskos, CCC, July 23, 2013
6. ^ Re: Likelihood Of Success (LOS) in the real world by Álvaro Begué, CCC, May 26, 2017
7. ^ Re: Calculating the LOS (likelihood of superiority) from results by Rémi Coulom, CCC, January 23, 2014
8. ^ Re: Calculating the LOS (likelihood of superiority) from results by Kai Laskos, CCC, January 22, 2014
9. ^ Re: Likelihood of superiority by Rémi Coulom, CCC, November 15, 2009
10. ^ Re: Likelihood of superiority by Rémi Coulom, CCC, November 15, 2009
11. ^ Error function from Wikipedia
12. ^ The Open Group Base Specifications Issue 6IEEE Std 1003.1, 2004 Edition: erf
13. ^ erf(x) and math.h by user76293, Stack Overflow, March 10, 2009

14. ^ Re: Calculating the LOS (likelihood of superiority) from results by Álvaro Begué, CCC, January 22, 2014
15. ^ Error margins via resampling (jackknifing) by Kai Laskos, CCC, August 12, 2016
16. ^ Properties of unbalanced openings using Bayeselo model by Kai Laskos, CCC, August 27, 2016
17. ^ Abraham Wald (**1945**). *Sequential Tests of Statistical Hypotheses*. Annals of Mathematical Statistics, Vol. 16, No. 2, doi: 10.1214/aoms/1177731118
18. ^ Sequential probability ratio test from Wikipedia
19. ^ Arthur Guez (**2015**). *Sample-based Search Methods for Bayes-Adaptive Planning*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, pdf
20. ^ Jack Good (**1979**). Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in World War II. Biometrika, Vol. 66, No. 2
21. ^ How (not) to use SPRT ? by BB+, OpenChess Forum, October 19, 2013
22. ^ Re: The SPRT without draw model, elo model or whatever.. by Michel Van den Bergh, CCC, August 18, 2016
23. ^ GSPRT approximation (pdf) by Michel Van den Bergh
24. ^ Elo's Book: The Rating of Chess Players by Sam Sloan
25. ^ The Master Game from Wikipedia
26. ^ Handwritten Notes on the 2004 David R. Hunter Paper 'MM Algorithms for Generalized Bradley-Terry Models' by Rémi Coulom
27. ^ Derivation of bayeselo formula by Rémi Coulom, CCC, August 07, 2012
28. ^ Mm algorithm from Wikipedia
29. ^ Pairwise comparison from Wikipedia
30. ^ Bayesian inference from Wikipedia
31. ^ How I did it: Diogo Ferreira on 4th place in Elo chess ratings competition | no free hunch
32. ^ "Intrinsic Chess Ratings" by Regan, Haworth -- seq by Kai Middleton, CCC, November 19, 2017
33. ^ Re: EloStat, Bayeselo and Ordo by Rémi Coulom, CCC, June 25, 2012
34. ^ Ordo by Miguel A. Ballicora
35. ^ A Pairwise Comparison of Chess Engine Move Selections by Adam Hair, hosted by Ed Schröder
36. ^ Questions regarding rating systems of humans and engines by Erik Varend, CCC, December 06, 2014
37. ^ chess statistics scientific article by Nuno Sousa, CCC, July 06, 2016
38. ^ LOS Table by Joseph Ciarrochi from CEGT
39. ^ Arpad Elo and the Elo Rating System by Dan Ross, ChessBase News, December 16, 2007
40. ^ David R. Hunter (**2004**). *MM Algorithms for Generalized Bradley-Terry Models*. The Annals of Statistics, Vol. 32, No. 1, 384–406, pdf
41. ^ Testing a chess engine from the ground up from Home of the Dutch Rebel by Ed Schröder
42. ^ Type I and type II errors from Wikipedia
43. ^ Arpad Elo - Wikipedia
44. ^ A Pairwise Comparison of Chess Engine Move Selections by Adam Hair, hosted by Ed Schröder
45. ^ Regan's latest: Depth of Satisficing by Carl Lumma, CCC, October 09, 2015
46. ^ Resampling (statistics) from Wikipedia
47. ^ Jackknife resampling from WIkipedia
48. ^ Delphil 3.3b2 (2334) - Stockfish 030916 (3228), TCEC Season 9 - Rapid, Round 11, September 16, 2016
49. ^ World Chess Championship 2016 from Wikipedia
50. ^ Normalized Elo (pdf) by Michel Van den Bergh

51. ^ A poor man's testing environment by Ed Schröder, CCC, January 04, 2013
52. ^ table for detecting significant difference between two engines by Joseph Ciarrochi, CCC, February 03, 2006
53. ^ an interesting study from Erik Varend by scandien, Hiarcs Forum, August 13, 2017
54. ^ UPGMA from Wikipedia
55. ^ UPGMA Worked Example by Richard Edwards
56. ^ Adam Hair's article on Pairwise comparison of engines by Charles Roberson, CCC, May 19, 2015
57. ^ Don Dailey, Adam Hair, Mark Watkins (**2014**). Move Similarity Analysis in Chess Programs. Entertainment Computing, Vol. 5, No. 3, preprint as pdf
58. ^ A Visual Look at 2 Million Chess Games by Brahim Hamadicharef, CCC, November 02, 2017

What links here?

Page	Date Edited
Adam Hair	Jan 14, 2017
AI Factory	Dec 26, 2015
Alessandro Scotti	Feb 27, 2014
Alexandru Mosoi	Jan 8, 2018
Álvaro Begué	Feb 27, 2018
Amoeba	Mar 19, 2018
Andreas Strangmüller	May 15, 2017
Andrew Grant	Jan 13, 2018
Arthur Guez	Dec 6, 2017
Automated Tuning	Feb 27, 2018
Book Learning	Jan 1, 2017
Buğra Fırat	Nov 2, 2017
CCRL	Dec 13, 2017
Chess	Jan 21, 2018
Chess Game	Mar 28, 2017
Chess Server	Jun 10, 2016
CLOP	Nov 24, 2016
Contempt Factor	Mar 10, 2018
Cutechess-cli	Jan 6, 2018
Delphil	Sep 17, 2016
Dennis Sceviour	Feb 14, 2018
Depth	Feb 25, 2018
Dictionary	Aug 24, 2017
Diogo R. Ferreira	Sep 4, 2016
Don Dailey	Dec 7, 2016
Draw	Apr 14, 2018
Ed Schroder	Feb 26, 2018
Edmund Moshammer	Sep 23, 2017
Engine Rating Lists	Dec 13, 2017
Engine Testing	Feb 1, 2018

Page	Date Edited
FGRL	Oct 1, 2017
Fire	Oct 12, 2017
Go (Programming Language)	Nov 2, 2017
Houdini	Apr 14, 2018
JavaScript	Dec 11, 2017
Jeff Rollason	Dec 23, 2016
Kai Laskos	Jan 21, 2018
Kenneth Wingate Regan	Feb 19, 2018
Kirill Kryukov	Apr 16, 2014
Komodo	Mar 6, 2018
Larry Kaufman	Jan 28, 2018
Lazy SMP	Jan 4, 2018
LittleBlitzer	Jan 24, 2017
LOS Table	Jan 10, 2015
Lyudmil Antonov	Aug 31, 2017
Marcel van Kervinck	Dec 6, 2017
Marco Costalba	Feb 28, 2018
Mark Watkins	Dec 6, 2017
Match Statistics	Mar 31, 2018
Material	Mar 31, 2018
Mathematician	Apr 9, 2018
Michael Hoffmann	Aug 9, 2015
Michel Van den Bergh	Mar 10, 2018
Miguel A. Ballicora	Jun 19, 2017
Nicu Ionita	May 3, 2017
Opening	Jun 25, 2017
Opening Book	Feb 26, 2018
Parallel Search	Dec 30, 2017
Pawn Advantage, Win Percentage, and Elo	Dec 20, 2017
Playing Strength	Mar 31, 2018
Pondering	Dec 9, 2017
Portable Game Notation	Jan 11, 2018
ProbCut	Mar 25, 2016
Promotions	Jun 6, 2017
Rémi Coulom	Feb 5, 2018
Richard Delorme	Sep 30, 2017
Score	Apr 14, 2018
Search	Feb 1, 2018
Search Statistics	Jan 16, 2018
Search with Random Leaf Values	Apr 23, 2018
Software	Jul 28, 2017
Stefan Meyer-Kahlen	Jun 19, 2017
Steve Maughan	Jan 31, 2018
Stockfish	Apr 7, 2018
Tamal T. Biswas	Jun 19, 2017

Page	Date Edited
TCEC Season 9	Nov 6, 2017
Test-Positions	Feb 25, 2018
Thread	Apr 22, 2018
Time Management	Mar 20, 2018
Tournaments	Dec 21, 2017
UCT	Jan 22, 2018
Utilities	Jan 20, 2018
Web Broadcast	Jul 30, 2017
WEB-MATCH	Jul 28, 2017

Up one level