

Program Help

DecisionTimeStats.exe -h

DecisionTimeStats - statistics for cuteschess Decision Time PGN comments

Usage:

DecisionTimeStats <input.pgn> [--output report.txt] [--sort mean]

Options:

-o, --output <file> Also write the report to a UTF-8 text file
--sort <mode> mean, zero, score, name, missing
-h, --help Show this help

Examples:

DecisionTimeStats UHO_Ratinglist.pgn

DecisionTimeStats UHO_Ratinglist.pgn -o DT-Stats.txt --sort zero

Explanation of the Columns

Engine	Games	Score%	DT moves	Missing	Miss%	Mean	Median	P9
--------	-------	--------	----------	---------	-------	------	--------	----

Engine

Engine name and version as listed in the PGN header.

Games

Number of games evaluated for this engine, regardless of whether it played White or Black.

Score%

The engine's tournament score as a percentage:

- Win = 1 point
- Draw = 0.5 points
- Loss = 0 points

DT moves

Number of moves for which a valid Decision Time was available and could be included in the statistics.

Missing

Number of moves for which a move time was present, but no Decision Time was included in the

comment.

Miss%

Percentage of missing Decision Times among all timed non-book moves of that engine.

Mean

Arithmetic mean of all recorded Decision Times.

Median

The middle value of the DT distribution. At least 50 percent of all Decision Times are less than or equal to this value.

P90

The 90th percentile. For 90 percent of all evaluated moves, the Decision Time is less than or equal to this value. Only 10 percent of the moves are above it.

Max

Highest recorded Decision Time for the engine.

DT=0%

Percentage of Decision Times stored as `0s`. According to the source format, this means a DT below 0.01 seconds, not necessarily an exact mathematical zero.

>1s%, >3s%, >5s%, >15s%

Percentage of moves whose Decision Time was strictly greater than 1, 3, 5, or 15 seconds.

These ranges are cumulative. A Decision Time of 16 seconds is therefore counted in all four columns.

Plausibility Check of the Results

The output appears to be internally very consistent.

Out of 19,356,972 total moves, 1,440,000 were book moves. This leaves 17,916,972 non-book moves.

These are distributed exactly as follows:

- 17,773,142 comments with a valid Decision Time
- 135,271 comments with a move time but no Decision Time
- 8,559 unrecognized or statistically unusable comments

The sum is exactly equal to the total number of non-book moves. Therefore, no move appears to have been lost or counted twice.

The engine assignment is also plausible:

- 120,000 games result in 240,000 engine participations.
- 16 engines with 15,000 games each also result in exactly 240,000 participations.
- All missing Decision Times belong exclusively to Integral and Obsidian.
- Their combined number of missing values exactly matches the global Missing counter.

This agrees with the information that only the Integral versus Obsidian head-to-head set did not yet contain Decision Times.

Meaning of Mean, Median, and P90

The Decision Time distributions are strongly right-skewed.

For many engines, the median is $0.00s$, while the mean is still approximately between 0.6 and 0.9 seconds. This means:

For more than half of all moves, the engine had already selected the move it would eventually play almost immediately. However, a smaller number of much longer changes of opinion pull the average upward.

Cinder is a good example:

- Mean: 1.66 seconds
- Median: 0.13 seconds
- P90: 5.60 seconds
- Maximum: 53 seconds

The mean is therefore far above the median. A relatively small number of long Decision Times have a considerable effect on the average.

For this reason, Mean, Median, and P90 should always be considered together. The Mean alone does not describe the distribution sufficiently well.

DT=0% Is Particularly Informative

This column shows enormous differences:

- Viridithas: 2.26%
- Cinder: 37.40%
- Quanticade: 66.01%
- Stockfish 260720: 80.27%

Viridithas stands out strongly. Only 2.26 percent of its Decision Times were recorded as 0s , while its median is still only 0.07 seconds.

This does not necessarily indicate an error. Viridithas probably stabilizes very frequently after approximately 0.05 to 0.10 seconds, while other engines already report the move they will eventually play within the first 0.01 seconds.

Possible causes include:

- different frequencies of `info pv` output
- different iterative-deepening behavior
- later output of the first serious principal move
- a peculiarity of the Cutechess measurement in combination with the respective engine

The statistics therefore do not measure only abstract search quality. They can also reveal differences in engine search and output behavior.

Making exactly these kinds of differences visible is one of the main purposes of the analysis.

First Interesting Findings

The table suggests a possible relationship between playing strength and Decision Time:

- Stockfish 260720 has the highest score and the lowest mean DT.
- Stockfish 18 has the second-highest score and also a very low mean DT.
- Reckless and PlentyChess achieve strong scores with relatively low mean Decision Times.
- Cinder has a comparatively weak score and by far the highest mean DT.

This does not yet prove a general statistical relationship. However, Stefan's assumption appears plausible:

Stronger modern engines often seem to identify the move they will eventually play very early.

There are exceptions. Torch achieves a strong score, but its mean DT is still relatively high.

For a more reliable conclusion, a formal correlation analysis could later be added:

- Pearson correlation for a possible linear relationship
- Spearman rank correlation for a relationship between ranking position and DT behavior

Possible relationships to examine include:

- Score% versus Mean DT
- Score% versus Median DT
- Score% versus DT=0%
- Score% versus the percentage of DT values above 1 second
- Score% versus the percentage of DT values above 5 seconds

Assessment of the Selected Time Ranges

The thresholds $>1s$, $>3s$, $>5s$, and $>15s$ are well chosen for a first version.

They divide the behavior into understandable categories:

- more than 1 second: noticeable change of opinion
- more than 3 seconds: longer uncertainty
- more than 5 seconds: pronounced late change
- more than 15 seconds: exceptionally late stabilization

These thresholds should be retained.

Useful future additions would be:

- P95
- P99
- average DT considering only values greater than zero

P95 and P99 would be more robust than the maximum, because a single extreme value can strongly affect the maximum.

The average of positive Decision Times could also help distinguish between:

- engines with many small positive DT values
- engines with very many $0s$, but a few extremely long changes

Preliminary Overall Assessment

The program already appears to be correct and practically useful in its first version.

Particularly convincing are:

- the complete and exactly balanced accounting of all moves
- the correct separation of book moves, valid DT values, missing DT values, and unrecognized comments
- the plausible assignment of all missing values to Integral and Obsidian
- the consistent number of games per engine
- the clearly different and substantively interesting DT distributions

The selected statistical ranges are appropriate. The results already reveal differences that would not be visible from a normal playing-strength rating list alone.