Data Science Lessons From a Predictive Chess Model

Smith College Computer Science

Kenneth W. Regan[1]
University at Buffalo (SUNY)

24 Mar. 2023

[1]With grateful acknowledgment to co-authors and UB's Center for Computational Research (CCR)

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with “95%” confidence that it’s between 5% and 15%.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

A Predictive Analytic Model

Means that the model:

- Addresses a series of events or decisions,each with possible outcomes $m_1, m_2, \ldots, m_j, \ldots$
- Assigns to each m_j a probability p_j.
- Projects risk/reward quantities associated to the outcomes.
- Also assigns *confidence intervals* for p_j and those quantities.

Example: An insurance company may estimate that:

- The probability of a given house having flood damage in a 5-year period is 10% with "95%" confidence that it's between 5% and 15%.
- This means is that out of 100 homes in similar and independent locations, they expect **10** to be flooded, with 95% confidence of no better than **5** but no worse than **15**.
- Homes being close together does not affect the expectation but does widen the confidence interval.

In my model, the m_j are possible moves in chess positions.

Inputs

Inputs

- The model is based on a **utility function / loss function** in a standard way—except for being log-log linear, not log-linear (why).

Inputs

- The model is based on a **utility function / loss function** in a standard way—except for being log-log linear, not log-linear (why).
- The (dis-)utility comes from (my heavily scaled version of) **average centipawn loss** of the played move compared to (what a powerful chess-playing program thinks is) the best move.

Inputs

- The model is based on a **utility function / loss function** in a standard way—except for being log-log linear, not log-linear (why).
- The (dis-)utility comes from (my heavily scaled version of) **average centipawn loss** of the played move compared to (what a powerful chess-playing program thinks is) the best move.
- **No chess knowledge other than the move values is input.**

The (only!) parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment.
- c for "consistency" in surviving tactical minefields.

Inputs

- The model is based on a **utility function / loss function** in a standard way—except for being log-log linear, not log-linear (why).
- The (dis-)utility comes from (my heavily scaled version of) **average centipawn loss** of the played move compared to (what a powerful chess-playing program thinks is) the best move.
- **No chess knowledge other than the move values is input.**

The (only!) parameters trained against chess Elo Ratings are:

- s for "sensitivity"—strategic judgment.
- c for "consistency" in surviving tactical minefields.
- h for "heave" or "Nudge"—obverse to depth of thinking.

Trained on all available in-person classical games in 2010–2019 between players within 10 Elo of a marker 1025, 1050, . . . , 275, 2800, 2825. Wider selection below 1500 and above 2500.

How it Works

How it Works

- Take s, c, h from a player's rating (or "profile").

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.
- (Correct after-the-fact for chess decisions not being independent.)

The statistical application then follows by math known since the 1700s.

How it Works

- Take s, c, h from a player's rating (or "profile").
- Generate probability p_i for each legal move m_i.
- Paint m_i on a 1,000-sided die, 1,000p_i times.
- **Roll the die**.
- (Correct after-the-fact for chess decisions not being independent.)

The statistical application then follows by math known since the 1700s. (Example of "Explainable AI" at small cost in power.)

Validate the model on millions of randomized trials involving "Frankenstein Players" to ensure conformance to the standard bell curve at all rating levels.

See: Published papers and articles on Richard J. Lipton's blog **Gödel's Lost Letter and P=NP** which I partner.

Text and Subtext

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.
- Subtext: Many deployed models **satisfice**—

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.
- Subtext: Many deployed models **satisfice**—
- —designed toward one prime objective but don't build in cross-checks or invest in the space of neighboring objectives.

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.
- Subtext: Many deployed models **satisfice**—
- —designed toward one prime objective but don't build in cross-checks or invest in the space of neighboring objectives.
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.
- Subtext: Many deployed models **satisfice**—
- —designed toward one prime objective but don't build in cross-checks or invest in the space of neighboring objectives.
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.
- *Going back to my model*, since it is fundamentally incorrect regarding independence, the cross-checks are a vital basis.

Text and Subtext

- Text: Despite being *severely underfitted*, the model works checkably well.
- Subtext: Many deployed models **satisfice**—
- —designed toward one prime objective but don't build in cross-checks or invest in the space of neighboring objectives.
- **Nonreproducibility**, **Mission Creep**, and **Shifting Sands**. E.g., I do not reproduce the longer conclusions of this study.
- *Going back to my model*, since it is fundamentally incorrect regarding independence, the cross-checks are a vital basis.
- Build not a Model but a Root System.

Pre-Check: The "Screening" Stage

Pre-Check: The “Screening” Stage

- Makes a simple “box score” of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- **Completely data driven.**

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- **Completely data driven.** Rapid and Blitz trained on **in-person** events in 2019.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- **Completely data driven.** Rapid and Blitz trained on **in-person** events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- **Completely data driven.** Rapid and Blitz trained on **in-person** events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018.
- Example: The just-finished European Individual CC.

Pre-Check: The "Screening" Stage

- Makes a simple "box score" of agreements to the chess engine being tested and the **scaled** average centipawn loss from disagreements.
- Creates a **Raw Outlier Index (ROI)** on the same 0-100 scale as flipping a fair coin 100 times.
- Here 50 is the expectation *given one's rating* and 5 is the standard deviation, so the "two-sigma normal range" is 40-to-60.
- Like medical stats except **indexed** to common **normal** scale.
- 65 = amber alert, 70 = code orange, 75 = red. Example.
- **Completely data driven.** Rapid and Blitz trained on **in-person** events in 2019. Slow chess trained on in-person FIDE Olympiads from 2010 to 2018.
- Example: The just-finished European Individual CC.
- Does not account for the *difficulty* of games. That is the job of the full model.

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.
- In our homes and flooding example :
  - $z = 2$ indexes the probability that **15** *or more* homes get flooded. About **1-in-44**, which is somewhat under 2.5% probability.

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.
- In our homes and flooding example :
  - $z = 2$ indexes the probability that **15** *or more* homes get flooded. About **1-in-44**, which is somewhat under 2.5% probability.
  - $z = 3$ means at least "**17.5**" homes being flooded, 1-in-741 frequency.

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.
- In our homes and flooding example :
  - $z = 2$ indexes the probability that **15** *or more* homes get flooded. About **1-in-44**, which is somewhat under 2.5% probability.
  - $z = 3$ means at least "**17.5**" homes being flooded, 1-in-741 frequency.
  - $z = 4$ means **20** or more flooded, for **1-in-31,575** frequency. (Ignoring that "half a home" matters here too.)

Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.
- In our homes and flooding example :
  - $z = 2$ indexes the probability that **15** *or more* homes get flooded. About **1-in-44**, which is somewhat under 2.5% probability.
  - $z = 3$ means at least "**17.5**" homes being flooded, 1-in-741 frequency.
  - $z = 4$ means **20** or more flooded, for **1-in-31,575** frequency. (Ignoring that "half a home" matters here too.)
  - $z = 6$ means **25** or more. A "Six-Sigma Deviation": 1-in-a-billion.

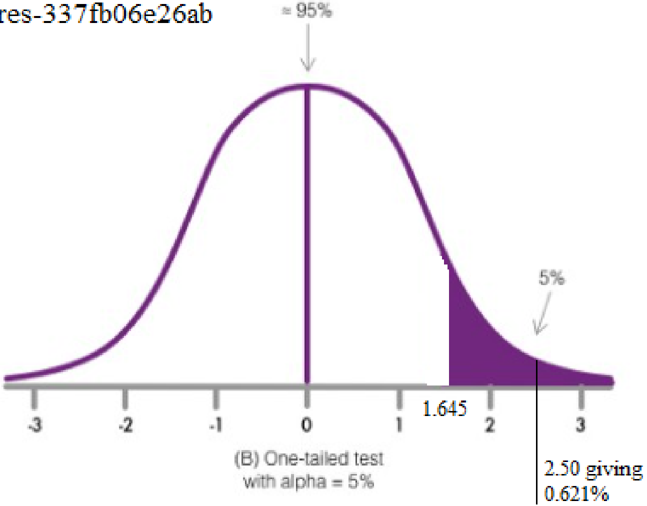
Z-scores

For **independent** situations whose results add up, one can replace probabilities by **Z-scores**, which quantify deviations of averages from expected means.

- Like how raw numbers are indexed by their logarithms on a slide rule.
- A z-value denotes the natural frequency of *at least* yea-much deviation.
- In our homes and flooding example :
  - $z = 2$ indexes the probability that **15** *or more* homes get flooded. About **1-in-44**, which is somewhat under 2.5% probability.
  - $z = 3$ means at least "**17.5**" homes being flooded, 1-in-741 frequency.
  - $z = 4$ means **20** or more flooded, for **1-in-31,575** frequency. (Ignoring that "half a home" matters here too.)
  - $z = 6$ means **25** or more. A "Six-Sigma Deviation": 1-in-a-billion.
- Like with a **Richter Scale**, +1 matters a lot.

Bell Curve and Tails

From https://towardsdatascience.com/hypothesis-testing-z-scores-337fb06e26ab



Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve "kicks in" at $N = 30$.

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve "kicks in" at $N = 30$.
- Shadable either way.

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve "kicks in" at $N = 30$.
- Shadable either way. Razieh Fathi used used 3 sets of $N = 15$ in her PhD work (paper at EDSIGCON 2022).

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve "kicks in" at $N = 30$.
- Shadable either way. Razieh Fathi used used 3 sets of $N = 15$ in her PhD work (paper at EDSIGCON 2022).
- In chess, the distribution D isn't the same for different chess positions.

Central Limit Theorem and “Rule of 30”

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve “kicks in” at $N = 30$.
- Shadable either way. Razieh Fathi used used 3 sets of $N = 15$ in her PhD work (paper at EDSIGCON 2022).
- In chess, the distribution D isn’t the same for different chess positions.
- But it stays “chessy.” I’m fully comfortable with $N = 50$.

Central Limit Theorem and "Rule of 30"

Theorem (CLT)

For **any** *probability distribution D, the mean of N* **independent** *samples from D is distributed more like the bell curve as $N \to \infty$.*

- Origin in the accuracy of N trials of any scientific measurement.
- **Convention**: closeness to bell curve "kicks in" at $N = 30$.
- Shadable either way. Razieh Fathi used used 3 sets of $N = 15$ in her PhD work (paper at EDSIGCON 2022).
- In chess, the distribution D isn't the same for different chess positions.
- But it stays "chessy." I'm fully comfortable with $N = 50$.
- The severe underfitting causes other problems for $N \gg 500$.

Demonstration

- At this point I showed data from the full model results, including from the recent European Individual Championships.
- The model is trained to make **MM%** (engine move-match) and **ASD** (scaled average centipawn loss) into **unbiased estimators**.
- Although the projections on the engine's second and third moves are moderately out of true, the 4th moves onward agree closely, while projections of various levels of mistakes are in fair agreement.
- In 10–15% of positions, the model projects an inferior move to be more likely than the engine's favored move. This yields 2–3 percentage points gain in predicting the played moves, compared to "betting the favorite" move. See this GLL blog article.
- *Advancing moves*, *capture moves*, and *moves with the knights* are played far more often than the model projects.
- Is it better to leave these human tendencies as "theorems" of the model in its minimalist form, or alter projections after-the-fact to match them?

How Well Does It Work?

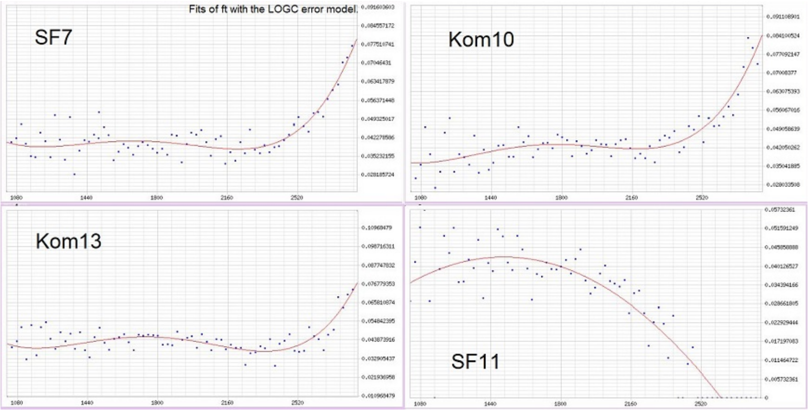
Internal evidence that it gives $(1 + \epsilon)$ relative error with $\epsilon \approx 0.04$ for most rating levels.

How Well Does It Work?

Internal evidence that it gives $(1 + \epsilon)$ relative error with $\epsilon \approx 0.04$ for most rating levels. Means it supports betting on chess moves with only 5% "vig" needed to avoid *arbitrage*.

How Well Does It Work?

Internal evidence that it gives $(1+\epsilon)$ relative error with $\epsilon \approx 0.04$ for most rating levels. Means it supports betting on chess moves with only 5% "vig" needed to avoid *arbitrage*. (SF11 issue corrected "by hand.")



Rating Lag—Natural Versus Systematic

Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**

Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**
- But this has perforce been **post-normal science**.

Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**
- But this has perforce been **post-normal science**.
- My "back of the envelope" formula held up over two years with only one small revision for preteens.

Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**
- But this has perforce been **post-normal science**.
- My "back of the envelope" formula held up over two years with only one small revision for preteens.
- Larger revision in Oct. 2022 to curtail projections past Elo 2000 level.

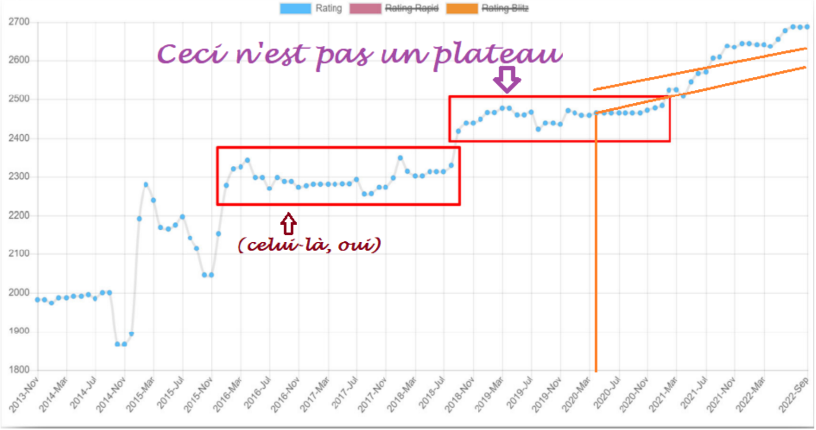
Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**
- But this has perforce been **post-normal science**.
- My "back of the envelope" formula held up over two years with only one small revision for preteens.
- Larger revision in Oct. 2022 to curtail projections past Elo 2000 level.
- Would have been more "normal" if comprehensive studies of the career arcs (measured by Elo rating) of young players were to hand.

Rating Lag—Natural Versus Systematic

- **The #1 scientific role I've played during the pandemic has been estimating the true skill growth of young players while their official ratings have been frozen.**
- But this has perforce been **post-normal science**.
- My "back of the envelope" formula held up over two years with only one small revision for preteens.
- Larger revision in Oct. 2022 to curtail projections past Elo 2000 level.
- Would have been more "normal" if comprehensive studies of the career arcs (measured by Elo rating) of young players were to hand.
- Lack of such studies exposed by the controversy over Hans Niemann's rise from 2465 Elo to 2700.

Hans Niemann: Platform or Plateau?



The Gender Gap in Chess

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
- How should one begin to address this question?

The Gender Gap in Chess

- Is clear: with Judit Polgar retired, there are no women in the top 100 by rating.
- Where/when does it begin?
- How should one begin to address this question?
- What data could corroborate a result—or a proposed explanation?

Q & A

The talk stopped here for Q & A. The remaining slides were in case of available time or separate interest for more on how the z-scores are interpreted in chess cheating cases.

Q & A

The talk stopped here for Q & A. The remaining slides were in case of available time or separate interest for more on how the z-scores are interpreted in chess cheating cases.

Two more concluding points in the meantime:

1. I have accepted lower sensitivity and predictivity in order to preserve *explainability* and gain *robustness*. Neural methods have been brittle in ways discussed here and here. I present a recent instance linked in an Update at the bottom of this GLL blog post.
2. Models should promote multiple paths of engagement with reality.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common “sigma” units allow combining z-scores of disparate events.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.
- **z = 3.00: 1-in-741** odds, 0.135% natural frequency.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.
- **z = 3.00: 1-in-741** odds, 0.135% natural frequency.
- **z = 4.00: 1-in-31,574** odds, 3.167/100,000 natural frequency.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.
- **z = 3.00: 1-in-741** odds, 0.135% natural frequency.
- **z = 4.00: 1-in-31,574** odds, 3.167/100,000 natural frequency.
- **z = 5.00: 1-in-3,486,914** odds, 2.87/10,000,000 natural freq.

Using Z-Scores

- Golf-shot analogy for why one uses the whole tail.
- The common "sigma" units allow combining z-scores of disparate events.
- The z-value gives "Face-Value odds" against the *null hypothesis* of the deviation occurring by natural chance.
- **z = 2.00: 1-in-44** odds, 2.275% natural frequency.
- **z = 3.00: 1-in-741** odds, 0.135% natural frequency.
- **z = 4.00: 1-in-31,574** odds, 3.167/100,000 natural frequency.
- **z = 5.00: 1-in-3,486,914** odds, 2.87/10,000,000 natural freq.
- But face-value odds need to be tempered against Bayesian priors, the look-elsewhere effect, and possible selection bias.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.
- "Sparse dependence" with exponential decay within a game.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.
- "Sparse dependence" with exponential decay within a game.
- Book between games is removed already.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.
- "Sparse dependence" with exponential decay within a game.
- Book between games is removed already.
- Can approximate effect of *covariance* by adjusting z 10–15% downward.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.
- "Sparse dependence" with exponential decay within a game.
- Book between games is removed already.
- Can approximate effect of *covariance* by adjusting z 10–15% downward.
- These are my **adjusted z-scores**.

Extremes, Dependence, and Adjustments

Going back to our homes-and-flooding example:

- All 100 homes being flooded gives $z = 18$. Beyond astronomical.
- But what if all 100 homes are together and a big storm comes?
- Problem is the home risks not being independent.
- Chess "homes" are like spaced 10km apart in a straight line from Kyushu to Hokkaido.
- "Sparse dependence" with exponential decay within a game.
- Book between games is removed already.
- Can approximate effect of *covariance* by adjusting z 10–15% downward.
- These are my **adjusted z-scores**.
- Both determined and vetted by millions of *resampling* trials—emphasizing 4-game, 9-game, and 16-game sets.

Sensitivity, Soundness, and Safety

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.
- Model is *safe* if in the absence of systematic deviations, the z-scores it gives follow a normal distribution

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.
- Model is *safe* if in the absence of systematic deviations, the z-scores it gives follow a normal distribution—or at least are *conservatively* within the $z \geq 2$ high end of the standard bell curve.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.
- Model is *safe* if in the absence of systematic deviations, the z-scores it gives follow a normal distribution—or at least are *conservatively* within the $z \geq 2$ high end of the standard bell curve.
- It is possible for models to be safe without being sensitive.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.
- Model is *safe* if in the absence of systematic deviations, the z-scores it gives follow a normal distribution—or at least are *conservatively* within the $z \geq 2$ high end of the standard bell curve.
- It is possible for models to be safe without being sensitive.
- My model has preserved safety while improving sensitivity.

Sensitivity, Soundness, and Safety

- Model is *sensitive* if whenever there is a high deviation in fact, the model registers a high z-score.
- Also termed: the model avoids *false negatives* / avoids *type-2 errors*.
- Model is *sound* if whenever it measures a high z-score there is a factual high deviation.
- Aka.: avoids *false positives* / avoids *type-1 errors*.
- Model is *safe* if in the absence of systematic deviations, the z-scores it gives follow a normal distribution—or at least are *conservatively* within the $z \geq 2$ high end of the standard bell curve.
- It is possible for models to be safe without being sensitive.
- My model has preserved safety while improving sensitivity.
- Safe models can still give false positives in (*normally rare*) cases.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
    - The **prior likelihood** of cheating.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000?

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(
  - The **look-elsewhere effect**: How many others could you have tested? How many in the tournament? How many others playing comparable-level chess that weekend? week? month? year?

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(
  - The **look-elsewhere effect**: How many others could you have tested? How many in the tournament? How many others playing comparable-level chess that weekend? week? month? year?
- Presence of **other, non-quality evidence** offsets these matters.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(
  - The **look-elsewhere effect**: How many others could you have tested? How many in the tournament? How many others playing comparable-level chess that weekend? week? month? year?
- Presence of **other, non-quality evidence** offsets these matters.
- OTB, divide 30,000 by 10,000 leaves just a "balance of probability." Insufficient. Need $z \geq 5$ for comfort.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
    - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(
    - The **look-elsewhere effect**: How many others could you have tested? How many in the tournament? How many others playing comparable-level chess that weekend? week? month? year?
- Presence of **other, non-quality evidence** offsets these matters.
- OTB, divide 30,000 by 10,000 leaves just a "balance of probability." Insufficient. Need $z \geq 5$ for comfort.
- Online, dividing by 100 leaves 300-to-1 "reckoned odds" against the *null hypothesis* of fair play.

Example Application and Reasoning

Suppose one gets a z-score of **4.00**.

- The **primary meaning** is that the performance has a natural frequency of about **1-in-31,574**, for that quality or higher.
- Let's round that to what I call "Face-Value Odds" of 30,000-to-1.
- This needs to be rectified according to various factors:
  - The **prior likelihood** of cheating. In-person: 1-in-5,000 to 1-in-10,000? Online: 1-in-50 to 1-in-100. :-(
  - The **look-elsewhere effect**: How many others could you have tested? How many in the tournament? How many others playing comparable-level chess that weekend? week? month? year?
- Presence of **other, non-quality evidence** offsets these matters.
- OTB, divide 30,000 by 10,000 leaves just a "balance of probability." Insufficient. Need $z \geq 5$ for comfort.
- Online, dividing by 100 leaves 300-to-1 "reckoned odds" against the *null hypothesis* of fair play.
- Interpret 100-1 to 1,000-1 as range of **comfortable satisfaction** per CAS Lausanne.

Cancer and Covid (= in-person and online chess)

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
    - Among them, **1** has the cancer; expect that result to be positive.
    - But we can also expect about **100** false positives.
    - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a "Second Opinion."

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a "Second Opinion."
- IMPHO, 1-in-5,000 $\approx$ frequency of cheating in-person.

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a "Second Opinion."
- IMPHO, 1-in-5,000 $\approx$ frequency of cheating in-person.
- A positive from a "98%" test is like getting $z = 2.05$. *Not enough.*

Cancer and Covid (= in-person and online chess)

- Say you take a test that is **98%** accurate for a cancer that affects 1-in-5,000 people...
- ...and get a positive. *What are the odds that you have the cancer?*
- Not the same as the odds that any one test result is wrong.
- Consider giving the test to 5,000 people, including yourself.
  - Among them, **1** has the cancer; expect that result to be positive.
  - But we can also expect about **100** false positives.
  - All you know at this point is: you are **one** of **101** positives.
- So the odds are still 100-1 against your having the cancer.
- The test result knocked down your prior 5,000-to-1 odds-against by a factor of 50, but not all the way. Need a "Second Opinion."
- IMPHO, 1-in-5,000 ≈ frequency of cheating in-person.
- A positive from a "98%" test is like getting $z = 2.05$. *Not enough.*
- In a 500-player Open, **you should see ten such scores**.

The 99.993% Test

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
- Sweet spot IMHO is **99.5%**, meaning **1-in-200** ultimate chance of wrong decision.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
- Sweet spot IMHO is **99.5%**, meaning **1-in-200** ultimate chance of wrong decision. Same criterion used by **Decision Desk HQ** to "call" US elections.

The 99.993% Test

- Suppose our cancer test were 600 times more accurate: **1-in-30,000** error.
- That's the face-value error rate claimed by a $z = 4$ result.
- Still 1-in-6 chance of false positive among 5,000 people.
- (This is really how a "second opinion" operates in practice.)
- If the entire world were a 500-player Open, then 1-in-60 chance of the result being natural.
- Still not **comfortable satisfaction** of the result being unnatural.
- IMPHO, the interpretation of CAS comfortable-satisfaction range of **final odds** determination is **99%–99.9%** confidence.
- Target confidence should depend on gravity of consequences. (CAS)
- Sweet spot IMHO is **99.5%**, meaning **1-in-200** ultimate chance of wrong decision. Same criterion used by **Decision Desk HQ** to "call" US elections.
- Higher stringency cuts against timely public service.

Covid in Non-Surge and Surge Times

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
    - Only 2 false negatives will expect to come from the **100** dangerous people.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
  - Only 2 false negatives will expect to come from the **100** dangerous people.
  - From the **4,900** safe people, about **4,800** true negatives.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
  - Only 2 false negatives will expect to come from the **100** dangerous people.
  - From the **4,900** safe people, about **4,800** true negatives.
  - Odds that your negative is false are **2,400-to-1** against.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
  - Only 2 false negatives will expect to come from the **100** dangerous people.
  - From the **4,900** safe people, about **4,800** true negatives.
  - Odds that your negative is false are **2,400-to-1** against.
- *Fine to be on a plane.*

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
    - Only 2 false negatives will expect to come from the **100** dangerous people.
    - From the **4,900** safe people, about **4,800** true negatives.
    - Odds that your negative is false are **2,400-to-1** against.
- *Fine to be on a plane.* What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50.

Covid in Non-Surge and Surge Times

- Now suppose the factual positivity rate is **1-in-50**.
- We still have about **100** false positives, but now also **100** factual positives.
- A positive from a 98% test is here a 50-50 coinflip.
- But a negative is *good*:
  - Only 2 false negatives will expect to come from the **100** dangerous people.
  - From the **4,900** safe people, about **4,800** true negatives.
  - Odds that your negative is false are **2,400-to-1** against.
- *Fine to be on a plane.* What happened is that the 98%-test result multiplied your confidence in not having Covid by a factor of almost 50.
- **Now suppose the factual positivity rate is 20%.** Can we do this in our heads?

Back to Chess...

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also *generally* goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also *generally* goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.
- *Sensitivity and soundness generally remain separate criteria.*

Back to Chess...

- Suppose we get $z = 4$ in online chess with adult cheating rate 2%.
- Out of 30,000 people:
  - **1** false positive result.
  - **600** factual positives.
  - So **600-1** odds against the null hypothesis on the $z = 4$ person.
- A $z = 3.75$ threshold leaves about **200-1** odds. OK here, but not if factual rate is under 1%.
- This analysis does not depend on how many of the factual positives gave positive test results.
- If test is only 10% sensitive, then we will have only about 60 positive results. It sounds like the 1-in-60 case. But the chance of getting a $z = 4$ result on the 1 brilliant player also *generally* goes down to 1-in-10. The confidence ratio is $60/0.10 = 600$-to-1 even so.
- *Sensitivity and soundness generally remain separate criteria.*
- This is relevant insofar as I often get a lot of 3.00–4.00 range results.

Interpretations II: Multiple Factors

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.
- For repeated actions, CLT applies, so deviations can be expressed via z-scores.

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.
- For repeated actions, CLT applies, so deviations can be expressed via z-scores.
- If you get z_1 from quality metrics and z_2 from the interface ("telemetry"), weight these factors equally, and consider them independent, then the overall z-score is

$$z = \frac{z_1 + z_2}{\sqrt{2}}.$$

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.
- For repeated actions, CLT applies, so deviations can be expressed via z-scores.
- If you get z_1 from quality metrics and z_2 from the interface ("telemetry"), weight these factors equally, and consider them independent, then the overall z-score is

$$z = \frac{z_1 + z_2}{\sqrt{2}}.$$

- (If you give weights w_1, w_2 then the formula is $z = \frac{w_1 z_1 + w_2 z_2}{\sqrt{w_1^2 + w_2^2}}$.)

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.
- For repeated actions, CLT applies, so deviations can be expressed via z-scores.
- If you get z_1 from quality metrics and z_2 from the interface ("telemetry"), weight these factors equally, and consider them independent, then the overall z-score is

$$z = \frac{z_1 + z_2}{\sqrt{2}}.$$

- (If you give weights w_1, w_2 then the formula is $z = \frac{w_1 z_1 + w_2 z_2}{\sqrt{w_1^2 + w_2^2}}$.)
- E.g., if both z_1 and z_2 are 3.5 then $z = \frac{7.0}{1.414...} \simeq 4.95$.

Interpretations II: Multiple Factors

- Online platforms collect data on player behavior: clicks, changes in window focus, timing of moves.
- Independence is relative to profiled tendencies.
- For repeated actions, CLT applies, so deviations can be expressed via z-scores.
- If you get z_1 from quality metrics and z_2 from the interface ("telemetry"), weight these factors equally, and consider them independent, then the overall z-score is

$$z = \frac{z_1 + z_2}{\sqrt{2}}.$$

- (If you give weights w_1, w_2 then the formula is $z = \frac{w_1 z_1 + w_2 z_2}{\sqrt{w_1^2 + w_2^2}}$.)
- E.g., if both z_1 and z_2 are 3.5 then $z = \frac{7.0}{1.414...} \simeq 4.95$.
- Face-value odds about 1 in 2.7 million, enough for "any" prior.

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

- In (a), there do not exist 31,574 or even 500 players who do this normally (in any year).

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

- In (a), there do not exist 31,574 or even 500 players who do this normally (in any year).
- Can sanction for violation of rule in any event.

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

- In (a), there do not exist 31,574 or even 500 players who do this normally (in any year).
- Can sanction for violation of rule in any event.
- Far more likely that $z = 4$ means cheating.

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

- In (a), there do not exist 31,574 or even 500 players who do this normally (in any year).
- Can sanction for violation of rule in any event.
- Far more likely that $z = 4$ means cheating. The false-positive guy under this combination won't arise in 60 years.

Interpretations III: Other Distinguishing Marks

Suppose we have one of thse two situations with player giving $z = 4$:

(a) Player found with cellphone on person.

(b) Player stowed cellphone in bag under chair, switched off [but it still rang].

- In (a), there do not exist 31,574 or even 500 players who do this normally (in any year).
- Can sanction for violation of rule in any event.
- Far more likely that $z = 4$ means cheating. The false-positive guy under this combination won't arise in 60 years.
- Logic goes for $z = 3$ and $z = 2.75$ and even $z = 2.5$ (1-in-161 frequency).

But in situation (b), it matters *how many* players do it, and whether it is *neutral* or *material*.

Distinguishing Marks, continued

Distinguishing Marks, continued

- If (b) is also material (or otherwise “covariant”) with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
  - the behavior is infrequent, *and*

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
    - the behavior is infrequent, *and*
    - we are not keeping a large catalogue of arbitrary/impertinent behaviors.

Distinguishing Marks, continued

- If (b) is also material (or otherwise “covariant”) with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
  - the behavior is infrequent, *and*
  - we are not keeping a large catalogue of arbitrary/impertinent behaviors.
- Suppose only 1,000 players do (b) in any year.

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
  - the behavior is infrequent, *and*
  - we are not keeping a large catalogue of arbitrary/impertinent behaviors.
- Suppose only 1,000 players do (b) in any year.
- Then the false-positive guy for $z = 4 \wedge (b)$ comes only once per 31.5 years.

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
  - the behavior is infrequent, *and*
  - we are not keeping a large catalogue of arbitrary/impertinent behaviors.
- Suppose only 1,000 players do (b) in any year.
- Then the false-positive guy for $z = 4 \wedge (b)$ comes only once per 31.5 years.
- So **30-to-1** odds against this year—especially if this is the first year of the policy.

Distinguishing Marks, continued

- If (b) is also material (or otherwise "covariant") with cheating, then I argue the face-value odds from the z-score become true odds, same as in situation (a).
- Even if (b) is *neutral*, still a problem if:
  - the behavior is infrequent, *and*
  - we are not keeping a large catalogue of arbitrary/impertinent behaviors.
- Suppose only 1,000 players do (b) in any year.
- Then the false-positive guy for $z = 4 \wedge (b)$ comes only once per 31.5 years.
- So **30-to-1** odds against this year—especially if this is the first year of the policy.
- Not enough for comfortable satisfaction, but $z = 4.265$ gives 1-in-100, $z = 4.42$ gives 1-in-200 (round number $z = 4.5$).

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?
- Together with $z = 3.29$, how the case alluded to in my "Doomsday Argument in Chess" article stood.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?
- Together with $z = 3.29$, how the case alluded to in my "Doomsday Argument in Chess" article stood.
- The low frequency—maybe at most 10 players per year do this?—does influence whether material.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?
- Together with $z = 3.29$, how the case alluded to in my "Doomsday Argument in Chess" article stood.
- The low frequency—maybe at most 10 players per year do this?—does influence whether material.
- But even if *neutral*, at 1-in-2,000 face-value odds, the false positive for this combination comes once every **200** years.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?
- Together with $z = 3.29$, how the case alluded to in my "Doomsday Argument in Chess" article stood.
- The low frequency—maybe at most 10 players per year do this?—does influence whether material.
- But even if *neutral*, at 1-in-2,000 face-value odds, the false positive for this combination comes once every **200** years.
- If we have a catalogue of **10** things like this, we err once in **20** years.

Distinguishing Marks, continued

- Suppose it's (b'): player wears green sneakers.
- Less frequent but completely neutral, arbitrary, impertinent.
- Judging based on that would be *selection bias*.
- How about (b''): player wears heavy sweater in hot June weather?
- Together with $z = 3.29$, how the case alluded to in my "Doomsday Argument in Chess" article stood.
- The low frequency—maybe at most 10 players per year do this?—does influence whether material.
- But even if *neutral*, at 1-in-2,000 face-value odds, the false positive for this combination comes once every **200** years.
- If we have a catalogue of **10** things like this, we err once in **20** years.
- (As it happens, my sharper August 2019 model gave some $z > 5$ readings, then more games were found which made $z > 6$ overall.)