Kramnik vs Nakamura or Bayes vs p-value.

Shiva Maharaj
Chess Ed

Nick Polson*
Booth School of Business
University of Chicago

Vadim Sokolov†
Department of Systems Engineering
and Operations Research
George Mason University

November 29, 2023

Abstract

We provide a statistical analysis of the recent controversy between Vladimir Kramnik (ex-world champion) and Hikaru Nakamura. Kramnik called into question Nakamura's 45.5 out of 46 win streak in a 3+0 online blitz contest at chess.com. We assess the weight of evidence using an a priori probabilistic assessment of Viswanathan Anand and the streak evidence of Kramnik. Our analysis shows that Nakamura has 99.6 percent chance of not cheating given Anand's prior assumptions. We also study the statistical fallacies prevalent in their analysis. Kramnik on the one hand, bases his argument on the fact that the probability of such a streak is very small. This falls into precisely the Prosecutor's Fallacy. Nakamura on the other hand, attempts to refute the argument, using a cherry-picking argument. This violates the likelihood principle. We conclude with the discussion of the relevant statistical literature on the topic of fraud detection and the analysis of streaks in sports data.

Key Words: Kramnik, Nakamura, island problem, Bayes factor, weight of evidence

1 Introduction

We provide a discussion of the statistical fallacies in probabilistic thinking of both Kramnik and Nakamura. Kramnik makes the blunder of the Prosecutor's Fallacy, well-known in expert testimony of confusing the p-value with the Bayes posterior probability. Whereas Nakamura attempts to refute Kramnik's argument using a cherry-picking argument. This violates the likelihood principle. We also provide a list of relevant research on the topic of streaks in sports.

We start by addressing the argument of Kramnik which is based on the fact that the probability of such a streak is very small. This falls into precisely the Prosecutor's Fallacy. The Prosecutor's Fallacy is a statistical reasoning error that occurs when the probability of one event is confused with the probability of another related event. It's often associated with the misinterpretation of evidence and conditional probabilities. It assumes that the probability of innocence given the evidence is the same as the probability of the evidence given the innocence. This can lead to wrongly convicting innocent defendants.

*Nick Polson is Professor of Econometrics and Statistics at Chicago Booth: ngp@chicagobooth.edu

†Vadim Sokolov is an Associate Professor at Operations Reaearch at George Mason University: vsokolov@gmu.edu

News / Sports / Others / 'Online, it must be 1 in 10,000. But there are millions of g...

'Online, it must be 1 in 10,000. But there are millions of games': Viswanathan Anand addresses 'cheating' in chess

Figure 1: Anand's prior. Source: Hindustan Times.

Let introduce the notations. We denote by G the event of being guilty and I the event of innocence. We use E to denote evidence. In our case the evidence is the streak of wins by Nakamura. The Kramnik's argument is that probability of observing the streak is very low, thus we might have a case of cheating. This is the prosecutor's fallacy

$$P(I \mid E) \neq P(E \mid I).$$

Kramnik's calculations neglects other relevant factors, such as the prior probability of the cheating. The prosecutor's fallacy can lead to an overestimation of the strength of the evidence and may result in an unjust conviction. In the cheating problem, at the top level of chess prior probability of $P(G)$ is small! According to a recent statement by Viswanathan Anand, the probability of cheating is 1/10000.

Section 2 provides calculations that use the prior odds of cheating and apply the Bayes rule to calculate the odds of cheating given the observed 45 win streak. We provide a sensitivity analysis to the inputs inherent in the analysis of the Nakamura-Kramnik controversy. Nakamura is one of the world's best 3+0 online blitz players, second only to Magnus Carlsen, the current world champion. We show that this winning streak, given he played against lower rated players, does not provide strong evidence of cheating. We also provide a list of relevant research on the topic of streaks in sports.

Furthermore, we discuss the likelihood principle. The likelihood principle is addressed in the seminal paper on testing by Edwards et al. [1963].

1.1 Existing Literature

The dichotomy between the traditional p-value based hypothesis testing and Bayesian hypothesis testing has been widely discussed in the literature. The Bayesian analysis uses Bayes factors that can then be calculated once the researcher is willing to assess a prior predictive interval for the test statistic. In most experimental situations, this appears to be the most realistic way of assessing apriori information. For related discussion, see Berger [2003] and Lopes and Polson [2019].

A related concept to hypothesis testing is a concept of black swan, an unexpected or extreme outcome [Taleb, 2007]. As Lindley [2008] discusses in his review, the Bayesian framework provides a proper framework for calculating probabilities of those rare events.

Balding and Donnelly [1995] discusses the Bayes in forensic science. He provides a detailed discussion of the Bayes factor and the likelihood ratio. He also discusses the problem of the prior probability of guilt. Berry and Chastain [2004] discuss the use of Bayes factors (ratio of evidence under innocent and guilty hypotheses). They discuss the case of Mary Decker Slaney. Good [1996] in the O.J. Simpson case uses Bayes Factors, and Lindley [2008] shows how to assess probabilities in the Black Swan Case.

2 Bayesian Island Problem

Given there prior ratio of cheaters to not cheaters is $1/N$, meaning out of $N+1$ players, there is one cheater. The Bayes calculations requires two main terms. The first one is the prior odds of guilt:

$$O(G) = P(I)/P(G).$$

Here $P(I)$ and $P(G)$ are the prior probabilities of innocence and guilt respectively.

The second term is the Bayes factor, which is the ratio of the probability of the evidence under the guilt hypothesis to the probability of the evidence under the innocence hypothesis. The Bayes factor is given by

$$L(E \mid G) = \frac{P(E \mid I)}{P(E \mid G)}.$$

Product of the Bayes factor and the prior odds is the posterior odds of guilt, given the evidence. The posterior odds of guilt is given by

$$O(G \mid E) = O(G) \times L(E \mid G).$$

The odds of guilty is

$$O(G) = \frac{N/(N+1)}{1/(N+1)} = N.$$

The Bayes factor is given by

$$\frac{P(E \mid I)}{P(E \mid G)} = \frac{p}{1} = p.$$

Thus, the posterior odds of guilt are

$$O(G \mid E) = Np.$$

There are two numbers we need to estimate to calculate the odds of cheating given the evidence, namely the prior probability of cheating given via N and the probability of a streak $p = P(E \mid I)$.

There are multiple ways to calculate the probability of a streak. We can use the binomial distribution, the negative binomial distribution, or the Poisson distribution. The binomial distribution is the most natural choice. The probability of a streak of k wins in a row is given by

$$P(E \mid I) = \binom{N}{k} q^k (1-q)^{N-k}.$$

Here q is is the probability of winning a single game. Thus for a streak of 45 wins in a row, we have $k = 45$ and $N = 46$. We encode the outcome of a game as 1 for a win and 0 for a loss or a draw. The probability of a win is $q = 0.8916$ (Nakamura's Estimate, he reported on his YouTube channel). The probability of a streak is then 0.029. The individual game win probability is calculated from the ELO rating difference between the players.

Then we use the Anand's prior of $N = 10000$ to get the posterior odds of cheating given the evidence of a streak of 45 wins in a row. The posterior odds of being innocent are 285. The probability of cheating is then

$$P(G \mid E) = 1/(1 + O(G \mid E)) = 0.003491.$$

Therefore the probability of innocent

$$P(I \mid E) = \frac{Np}{Np+1} = 0.9965.$$

For completeness, we perform sensitivity analysis and also get the odds of not cheating for $N = 500$, which should be high prior probability given the status of the player and the importance of the event. We get

$$P(I \mid E) = \frac{Np}{Np+1} = 0.9445.$$

There are several assumptions we made in this analysis.

- Instead of calculating game-by-game probability of winning, we used the average probability of winning of 0.8916, provided by Nakamura himself. This is a reasonable assumption given the fact that Nakamura is a much stronger player than his opponents. This assumption slightly shifts posterior odds in favor of not cheating. Due to Jensen inequality, we have $E(q^{50}) > E(q)^{50}$. Expected value of the probability of winning a single game is $E(q) = 0.8916$ and the expected value of the probability of a streak of 50 wins is $E(q^{50})$. We consider the difference between the two to be small. Further, there is some correlation between the games, which also shifts the posterior odds in favor of not cheating. For example, some players are on tilt. Given they lost first game, they are more likely to lose the second game.

- There are many ways to win 3+0 unlike in classical chess. For example, one can win on time. We argue that probability of winning calculated from the ELO rating difference is underestimated.

Next, we can use the Bayes analysis to solve an inverse problem and to find what prior you need to assume and how long of a sequence you need to observe to get 0.99 posterior? Small sample size, we have p close to 1. Figure 2 shows the combination of prior (N) and the probability of a streak (p) that gives posterior odds of 0.99.

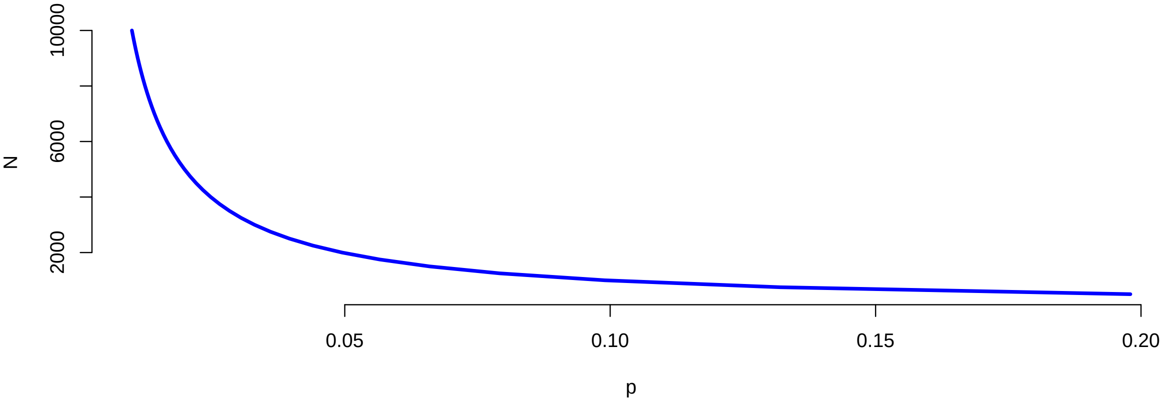


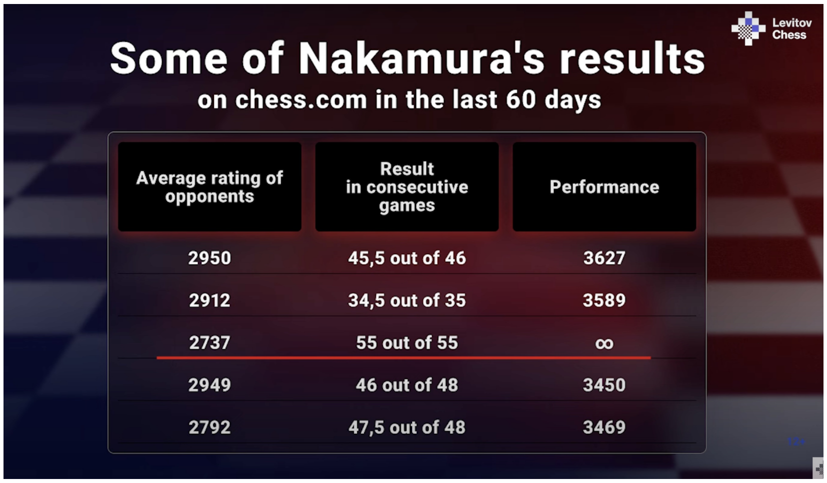
Figure 2: Combination of prior (N) and the probability of a streak (p) that gives posterior odds of 0.99.

Indeed, the results of the Bayesian analysis contradict the results of a traditional p-value based approach. A p-value is a measure used in frequentist statistical hypothesis testing. It represents the probability of obtaining the observed results, or results more extreme, assuming that the null hypothesis is true. The null hypothesis is a default position that Nakamura is not cheating and we compare the ELO-based expected win probability of $q = 0.8916$ to the observed on of $s = 45/46 = 0.978$. Under the null hypothesis, Nakamura should perform at the level predicted by q.

It is assumed that the ratio $(s-q)^2/q$ follows a Chi-squared distribution with one degree of freedom. The p-value is then calculated as the probability of s being smaller than or equal to q. A smaller p-value indicates a stronger likelihood that you should reject the null hypothesis. In other words, a small p-value suggests that the observed data is unlikely under the null hypothesis, providing evidence against it. In our case, the p-value is 0.009. Thus, under a traditional frequentist approach, we would reject the null hypothesis that Nakamura is not cheating.

However, it's important to note that a p-value does not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. It also does not measure the size of an effect or the importance of a result.

The screenshot below shows the table that Kramnik presented at in interview on *Levitov Chess World* YouTube channel he gave on November 27, 2023 [Levitov Chess World, 2023].


Levitov Chess

Some of Nakamura's results

on chess.com in the last 60 days

Average rating of opponents	Result in consecutive games	Performance
2950	45,5 out of 46	3627
2912	34,5 out of 35	3589
2737	55 out of 55	∞
2949	46 out of 48	3450
2792	47,5 out of 48	3469

Figure 3: Table Published by Kramnik

Although, it is not clear how the Performance columns was constructed, the analysis of Kramnik is more aligned with the p-value based approach or calculating $P(E \mid I)$ (prosecutor's fallacy). However, even those two approaches have non-zero probability assigned to observing a streak of 55 consecutive wins. This case $P(E \mid I) = 0.001023$ and the p-value is 0.0062. However, the posterior probability of cheating is still relatively low and is equal to 0.089. He probably can beat his opponents 55 out of 55, but he cannot beat the laws of probability. It is not infinity.

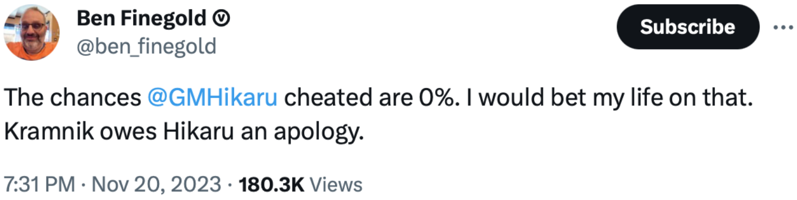
Nakamura himself, on the other hand claims that cherry-picking a sequence of 46 games out of more than 3500 he played on chess.com is not a fair approach to collect the data. We argue, Nakamura's statement is a violation of the likelihood principle.

The likelihood principle is addressed in the seminal paper on testing by Edwards et al. [1963]. The likelihood principle states that the evidence from an experiment is contained in the likelihood function. As Edwards-Lindman-Savage say it: "Often evidence which, for a Bayesian statistician, strikingly supports the null hypothesis leads to rejection of that hypothesis by standard classical procedures. The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience."

3 Discussion

Probability of a streak is widely discussed, but it has nothing to do with the probability of cheating. Kramnik is correct in ignoring the analysis of streaks in Nakamuras games —- although this is an interesting problem in and on itself. In section 1.1, we provide a relevant list of research and clearly it's an interesting study. Hot hands and streaks have been the topic of much research in many sports.

Finegold is a victim of another fallacy, called the Cromwell's rule. Cromwell's rule, states that predictions should be wary of assigning a prior probability of 0 (impossible) or 1 (certain) to anything except to statements that must be logically true.



Ben Finegold
@ben_finegold

The chances @GMHikaru cheated are 0%. I would bet my life on that. Kramnik owes Hikaru an apology.

7:31 PM · Nov 20, 2023 · **180.3K** Views

Figure 4: Screenshot of Ben Finegold's Twitt from November 20, 2023.

Essentially, $P(G) = 0$ implies $P(G \mid E) = 0$ for any evidence E.

In conclusion, the Bayesian analysis of the Nakamura-Kramnik controversy shows that Nakamura has 99.9 percent chance of not cheating given Anand's prior assumptions. We also study the statistical fallacies prevalent in their analysis. Kramnik on the one hand, bases his argument on the fact that the probability of such a streak is very small. This falls into precisely the Prosecutor's Fallacy. Nakamura on the other hand, attempts to refute the argument, using a cherry-picking argument. This violates the likelihood principle.

Finally, in the light of new evidence we allow the possibility of updating our believes. As Kaynes said, "Sir, if the facts change, I change my opinion".

References

Vladimir Kramnik. *Wikipedia*, November 2023.

Jim Albert. A Statistical Analysis of Hitting Streaks in Baseball: Comment. *Journal of the American Statistical Association*, 88(424):1184–1188, 1993.

David Aldous. *Probability Approximations via the Poisson Clumping Heuristic*, volume 77 of *Applied Mathematical Sciences*. Springer, New York, NY, 1989. ISBN 978-1-4419-3088-0 978-1-4757-6283-9.

David Aldous. Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability? *Statistical Science*, 32(4):616–629, November 2017.

David J. Balding and Peter Donnelly. Inference in Forensic Identification. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(1):21–40, January 1995.

Ben Finegold (V) [@ben_finegold]. The chances @GMHikaru cheated are 0%. I would bet my life on that. Kramnik owes Hikaru an apology., November 2023.

James O. Berger. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, 18(1):1–32, February 2003.

Donald A. Berry and Leeann Chastain. Inferences about Testosterone Abuse among Athletes. *CHANCE*, 17(2):5–8, March 2004.

Colin F. Camerer. Does the Basketball Market Believe in the 'Hot Hand,'? *The American Economic Review*, 79(5):1257–1261, 1989.

Ward Edwards, Harold Lindman, and Leonard J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242, 1963.

Epic Chess. Kramnik Demands Hikaru Investigation By Chess.com, November 2023.

Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3):295–314, July 1985.

GMHikaru. Is Hikaru Cheating?? Kramnik "YES!!!", November 2023.

I. J. Good. When batterer becomes murderer. *Nature*, 381(6582):481–481, June 1996.

William H. Jefferys and James O. Berger. Ockham's Razor and Bayesian Analysis. *American Scientist*, 80(1):64–72, 1992.

Harold Jeffreys. *Theory of Probability*. Oxford Classic Texts in the Physical Sciences. Oxford University Press, Oxford, New York, third edition, third edition edition, November 1998. ISBN 978-0-19-850368-2.

John Maynard Keynes and Mathematics. *A Treatise on Probability*. Dover Publications, Mineola, NY, April 2004. ISBN 978-0-486-49580-4.

Vladimir Kramnik (VladimirKramnik). Sign the petition. https://www.chess.com/blog/VladimirKramnik/sign-the-petition, November 2023.

Levitov Chess World. Kramnik demands to examine Nakamura's games, November 2023.

Dennis Lindley. Books reviewed: The Black Swan the Impact of the Highly Improbable. *Significance*, 5(1):42–43, 2008.

Hedibert F. Lopes and Nicholas G. Polson. Bayesian hypothesis testing: Redux. *Brazilian Journal of Probability and Statistics*, 33(4):745–755, November 2019.

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK ; New York, illustrated edition edition, October 2003. ISBN 978-0-521-64298-9.

Edward Simpson. Edward Simpson: Bayes at Bletchley Park. *Significance*, 7(2):76–80, June 2010.

Hal S. Stern Sugano, Adam. Inference about batter-pitcher matchups in baseball from small samples. In *Statistical Thinking in Sports*. Chapman and Hall/CRC, 2007. ISBN 978-0-429-14790-6.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, New York. N.Y, annotated edition edition, April 2007. ISBN 978-1-4000-6351-2.