Agreements with and Counterarguments to "Response of Critique of Dream Investigation Results"

Photoexcitation

January 8, 2021

***Executive Summary*: The Minecraft Speedrunning Team wrote a Report that concluded that a subset of six livestreams had low probability events so extreme to establish that Dream's games were inappropriately modified. Photoexcitation provided a Dream-commissioned Critique of this Report that disagreed on a few key points and concluded that the numbers were not as extreme as proposed, though still with very low odds. The Minecraft Speedrunning Team has written a Response or rebuttal to the Photoexcitation Critique claiming multiple issues with the calculations. This current report, partially commissioned by Dream, provides counterarguments to the Response that support many of the conclusions of the original Critique. It also identifies many areas of agreement including that the low probabilities are very strong arguments for the hypothesis that Dream cheated. Incorporating the arguments from the Minecraft Speedrunning Team Response, a new estimate is provided of about 1 in 100 million odds that any livestreaming speedrunner would experience the low probability effects seen on Dream's six livestreams.**

1 Introduction

After my[1] "Critique of Dream Investigation Results", associated with a Response video from Dream[2], the Minecraft Speedrunning Team have responded (the "Response"`https://mcspeedrun.com/dream/rebuttal.pdf`). They viewed several aspects of my report unfavorably. I do not argue or believe that their original Report or the Response is unfair or biased, but hope that discussion and dialogue will help identify the most robust conclusions.

This response was partially commissioned by Dream, but mostly represents my desire to better explain my original Critique, to defend myself, and to help maintain the credibility of Photoexcitation.

2 Areas of Agreement

As I wrote in my original report, "Probability calculations are hard. There may not be one 'right' way to do something. It is easy to violate some hidden or unknown assumption. There is room for healthy debate about different methods and results." There will likely be areas of this debate where we don't fully reach mutually agreeable conclusions. However, there are many areas where we can agree. As everyone attempts to draw some reasonable conclusions and to speed the discussion along, I've decided to start with areas of agreement.

I won't speak for the Minecraft Speedrunning Team, but here are some conclusions that I believe that they and I would agree on.

[1]The discussion boards have been abuzz with questions about my identity, which is protected by Photoexcitation's anonymity policy. I am the same author as the original Critique. Dream knew more about my identity that I revealed in the original report or that can be found on Photoexcitation's website (`https://www.photoexcitation.com`). I do hold a degree from Harvard and use advanced statistics as an astrophysicist. I wish to maintain my anonymity, though I understand that for some this introduces the possibility that my credentials cannot be vetted.

[2]`https://www.youtube.com/watch?v=1iqpSrNVjYQ`

1. Even in the best-possible interpretation, the probability that any speedrunner had any streak of luck as strong as Dream's is extremely low, providing significant evidence that the ender pearl and/or blaze rod drop probabilities were modified, resulting in a very strong argument for the hypothesis that Dream cheated.

2. Anyone who claimed that I said "Dream did not cheat" was severely misrepresenting what I wrote. My report was focused on providing an additional probability calculation and the difference in our final numbers was large, but the difference between luck of 99.9999% vs. 99.9999999999% is only meaningful for people who are already quite inclined to believe Dream's claims that he did not cheat. My Critique also raised the possibility that the probabilities were modified, but without malicious intent.

3. Drawing conclusions about whether Dream cheated are unlikely to be substantively different even if there are discrepancies of a factor of 10 or 100 in the probability calculations. Anything smaller than a factor of about 10 may not be worth detailed investigation given the range of numbers under consideration.

4. Mixing in other runs from Dreams significantly increases the probability of an unmodified game, but is not directly relevant to the main hypothesis that Dream cheated during this particular sequence of runs. Deciding which subset of runs to investigate after noticing that a particular subset were lucky requires some kind of correction. The original Report clearly provides such a correction [it was never my intention to imply otherwise] by imagining that all possible substreaks were investigated, which increases the probability by a factor of 10-100.

5. Including other streams isn't entirely meaningless, as it does establish that Dream's probabilities were not modified for all of his livestreams. This conclusion is separate from any conclusion about the six streams in question.

6. When accounting for other possible random numbers that might have been investigated, going from 10 to 37 only provides an increase of about[3] a factor of 15. Not all of these are obvious methods for cheating or methods that could have been as easily investigated, so using a correction of $37 \times 36 = 1332$ is an overestimate. My Critique used 1000, while the original MST Report used 90.

7. One major difference in probability is due to the comparison of the number of speedrunners/streams/runs.

8. The Sampling Bias Corrections that I originally argued aren't appropriate to lucky streaks were actually fine. The arguments in this section of my Critique were weak, distracting, and had the effect of overcriticizing the original Report. Even so, since any disagreement on these corrections was not used in my probability calculations, the conclusions of my Critique are unaffected.

9. The Bonferroni correction is a reasonable way to correct for p-hacking and other after-the-fact corrections.

10. The Bonferroni and (Dunn-)Šidák corrections are equivalent when the p-values are very small.

11. We currently disagree on whether the "Barter Stopping" correction should be used and whether it was already accounted for in their original Stopping model.

12. The sum of negative binomial distributions is distributed as a negative binomial distribution. The Barter Stopping model I proposed in my Critique is effectively a combination of (sums of) negative binomial distributions.

13. I mixed Bayesian and frequentist concepts in my Critique.

I again reiterate that I have not discussed these points with the Minecraft Speedrunning Team, but that I believe that they would agree with me on the above points, based on their reports. I provide these

[3] Instead of 10×9, the factor would be 37×36, leading to an increase of $\frac{37 \times 36}{10 \times 9} = 14.8$

"agreed-upon" conclusions in the interest of expediency and with the goal of more quickly concluding our back-and-forth discussion.

I think it is cool that this discussion rose to the level of being posted on Andrew Gelman's famous blog. :) One positive outcome was influencing large numbers of people to think more about math and statistics. Hopefully, highlighting areas of agreement will help people realize that there is much objectivity in these calculations.

I now turn to four major areas of discussion: barter stopping, mixing Bayesian and frequentist methods, lucky streak probabilities, and corrections for number of streamers.

3 Is Barter Stopping accounted for in the original MST Report?

My Critique raised the point that Barter Stopping is arguably a higher fidelity representation[4] of what really happened in most of the speed runs than the Binomial Model used in the original MST Report. My Barter Stopping simulation is effectively a combination (depending on whether it takes 2 or 3 barters to reach 10 ender pearls) of negative binomial distributions. The Response claims that the original Report's binomial model with arbitrary stopping condition (from their Appendix B) already accounts for this. However, the Response focuses on showing that the sum of negative binomials is distributed as a negative binomial (which I agree with). This shows that the per-barter stopping criterion basically washes out to just being a single stopping criterion at the end, and I've done simulations which agree with this result.

I argue that there is an additional key point here: the original report uses the binomial distribution where arguably a negative binomial distribution should be used. Despite the similar names and related formulas, these are *not* identical distributions. Their Response asserts but does not prove that their original Stopping Criterion is enough to overcome the issue of using a different distribution. Further, the Response does not show that my model is an inappropriate choice.

In my Bayesian analysis, I perform both the barter-stopping criterion (equivalent to the negative binomial distribution) and the binomial distribution for ender pearls[5] and find that the former leads to a posterior probability higher by a factor of 60 even without applying a stopping criterion on the last run (which increases the probability further). In the original Report, using their Stopping Rule increases the probability by only a factor of 2. While these are clearly apples and oranges, I do not feel that the Response has demonstrated that their original Stopping Criterion with a Binomial Model is equivalent or comparable or more favorable than a Negative Binomial Model. My results are that the Barter Stopping Model (effectively a combination of Negative Binomial Models) is substantively different from the Binomial Model with the implemented Stopping Rule.

Another quick test is to calculate the cumulative distribution function evaluated for the negative binomial distribution for ender pearls. For 42 ender pearls and $\leq$262 barters, the negative binomial CDF is 6.7×10^{-10}, an order of magnitude higher that their "best-case" stopping criterion and similar to my 3×10^{-10}. For 262 barters and 42 or more successes, the cumulative probability is 9.7×10^{-11}, again larger than their "best-case" values. In my opinion, this provides more evidence that the Response has not demonstrated that their Stopping Criterion appropriately accounts for barter stopping which is arguably more appropriate for the probabilities at hand.

All that said, different modeling methods only move the needle by a factor of 10-100 and don't change the overall conclusion that the odds that any speedrunner in any set of streams experienced such an unusual outcome is extremely low.

4 Is mixing Bayesian and frequentist methods allowed?

I admitted above to mixing Bayesian and frequentist methods, but I disagree that this makes the answers "uninterpretable" or that Bayesian analysis can not, does not, or need not account for bias corrections.

[4] Actually, I look at each run separately and assign it to Barter Stopping or Binomial, which is even higher fidelity than assuming one model for all the runs.

[5] One criticism of my model is that the way I generate values for ender pearls cannot produce 8 ender pearls, even though this is seen in actual play. As I stated, "variations in this model were not significant," which I validated by testing various distributions, including one that gave 4-8 ender pearls. Note that the choice of never giving 8 ender pearls causes the probabilities to be lower (less favorable to Dream), so I see little reason to criticize this choice.

While I agree that it would be better to perform the entire calculation within one probabilistic paradigm, that does not mean that the results of my analysis are invalid or uninterpretable. While I have not performed an end-to-end full Bayesian analysis, I have good reasons to suspect that these analyses are in the regime where the Bonferroni/Šidák corrections that we both use are appropriate in either the Bayesian or the frequentist paradigm. Bias corrections must be included and I propose that they can be applied to a posterior probability from a Bayesian analysis or a p-value from a frequentist analysis. Or, if you like, you can imagine my Bayesian analysis a different way of computing a probability that is then interpreted in the frequentist paradigm.

Bayesian methods are susceptible to p-hacking and including information on how and why the data were gathered is appropriate, e.g., `http://datacolada.org/13`. That article quotes from the authoritative Bayesian Data Analysis textbook by Gelman et al. that it is erroneous to claim that "because all inference in conditional on the observed data, it makes no difference how those data are collected, ..., the essential flaw in the argument is that a complete definition of 'the observed data' should include infomration on how the observed values arose ...." I believe that the Response's claims in Section 7.1 fall mostly into this category and argue that including the corrections in my Critique is allowed and interpretable.

5 Lucky Streak Probabilities

Keeping in mind that my Critique was written in a very short time, my initial investigations in early versions of the article were concerned with whether lucky streaks were appropriately accounted for. Most of the online arguments with my Critique are focused on this section, which I admit is weak and missing some details. As I continued to study, this point became less important and then disconnected with the rest of the analysis. But I thought I had found some results that were interesting to the discussion which I left in. Honestly, I did not spend as much time triple-checking this part of the document because it was irrelevant to the final numbers. However, this section still criticizes the original MST Report, which was unnecessary and for which I apologize. In retrospect, I should have either toned down this section or removed it entirely.

One major focus has been on the coin flips question. As it seems like a simple test, I can see why disagreement on this question would be of concern. First, in my code, I accidentally calculated the probability of a streak of 19 heads in a row.[6] Other possible issues are how to deal with streaks that are longer than the desired streak length. There are some good pedagogical statistics and probability discussions here, but in the interest of focusing on what is most important, I'm going to skip to the end and say *I made a mistake*. I will also apologize to the MST team for implying that the sampling bias correction methodology was inappropriate. Especially when I end up using the Bonferroni-style corrections myself in the final analysis. And especially when the differences are comparable to the factor of 2 that I propose be ignored.

Making this mistake has resulted one strongly-worded reddit post to declare me an "amateur" and "unreliable." While mistakes on basic calculations aren't confidence-boosting, I would propose that identifying a single weak point in a paper that is unconnected to the rest of the analysis and then concluding that the entire paper is untrustworthy (without identifying specific issues on other aspects) is itself unprofessional. Even peer-reviewed journal articles are not held to this standard. Especially when considering the nature of the mistake: a minor error on an unimportant point during an analysis completed in a very short amount of time.

Instead, I feel that the appropriate conclusions on this mistake are that:

1. That entire section of the Critique should be ignored. This gives strength to the original MST Report. My apologies again for including this section.

2. Including this section in the Critique gives concern that there are errors in the rest of the Critique. I don't think that's a particularly strong argument for saying "the original Report is completely correct and the Critique is completely wrong." I think it is more appropriate to then scrutinize each argument, especially those that are most important. In this regard, the Response identifies what I would consider to be the key points and hence those are the ones that I focus on in this document.

Of course, the reader is free to draw their own conclusions.

[6] For those unfamiliar with programming, I'll mention that such mistakes are relatively common, e.g., `https://en.wikipedia.org/wiki/Off-by-one_error`. This was a mistake and not intentional.

6 Corrections for Number of Streamers/Streams/Runs

Let's quickly review how the bias corrections for the number of streamers, streams, and runs was done. Since Dream's runs were investigated specifically because they appeared lucky, it is important to provide a bias correction. Using a Bonferroni/Šidák correction, the final probability for Dream's runs are multiplied by the number of comparable different possible investigations that could have been done. In the original Report, this was given as 1000 streamers (see Equation 13) and 66 possible consecutive subsets of Dream's 11 streams (see Equation 12) that could have been investigated (for ender pearls only, which I agree with). In my Critique, I proposed instead considering all possible (consecutive) streams, using an estimate of 300 livestreamed speedruns per day or, effectively 100000 total. I also mention a correction for choosing the length of the streams, to give another factor of 10.

One main difference is whether you count by first choosing a speedrunner and then studying a subset of their streams or whether you count the subsets of all streams. I'm not sure there's an obvious reason to prefer one over the other in terms of how you determine what all the possible subsets that could have been investigated. In fact, they probably would give similar numbers if we assumed similar numbers of speedrunners... even as it is, the total correction is 66,000 vs 1,000,000 which is only a factor of 15. I agree that this is one of the important numbers for my probability being higher than theirs.

The Response gives two very good points in estimating the number of livestreaming speedrunners. First, they point out that the number of livestreamers is not a "steady state" but is rapidly growing so that my assumption of typical leaderboard entry ages of one month probably overestimates the number of speedrunning livestreams. That's a great point. Even more meaningful is the idea that they present that when considering the sampling bias, each speedrunner need not be considered as a "binary" in or out, but that the correction should be weighted by the probability that the livestreaming speedrunner would be investigated. That is an excellent point. Most speedrunners, even those that livestream, would not have been investigated at this level of scrutiny. I think this deserves decreasing the odds by a factor of 10-100. On the other hand, the Response seems to be focus on 1.16 speedruns specifically, whereas the conclusions (in the original Report and in mine) are on all Minecraft speedrunning. So, leaving it at all Minecraft speedrunning and integrating over all past years, I'll lower my number of meaningful streams to 10,000 and draw a new conclusion that the odds of any small subset of any livestreamed speedrunner ever receiving as low a probability as Dream is **1 in 100 million**. Note that our reports now basically agree on how large of a correction to apply to go from Dream's six runs to all possible subsets of all speedruns.

I'll note here for the layperson that when we say "that any speedrunner was ever this lucky" it feels natural to conclude that Dream's odds of being this lucky were much worse than this. That is not a correct conclusion because it "undoes" the very real sampling bias that Dream specifically was investigated because his runs seemed very "lucky." You could also write this as 1 in 100 million odds of Dream ever receiving as low a probability accounting for the fact that he was investigated because he seemed too lucky.

7 Conclusion

While some of the issues of the Minecraft Speedrunning Team Response to my original Critique were valid, I disagree with their assessment on others. After including their considerations, especially of the number of speedrunners to compare to, I re-evaluate the odds from 1 in 10 million to about 1 in 100 million and still think that an upper board of 1 in 7.5 trillion is too strong. But either way, the probabilities were almost certainly modified and this provides very strong evidence that Dream cheated.