How Fair is a Coin Toss?

Amir Sahrani

University Of Amsterdam

Abstract

Diaconis et al. (2007) hypothesized coin tosses to be slightly biased towards landing in the same position the coin started. We investigated this same-side bias by performing 134234 coin flips. These coin flips were performed by multiple tossers over the span of three monthṣ. Approximately 6000 of these coin tosses where evaluated to estimate the number of mistakes made when inputting the data. We found coins to have a 50.86% chance of landing on the same side they started, this did differ between the tossers. Using the 6000 evaluated tosses, an estimated 1% of the tosses was wrongly put in, further analysis provided weak evidence in favor of errors being biased. These results strongly support the hypothesis by Diaconis et al.

How Fair is a Coin Toss?

Introduction

Coin flipping has been a staple solution for any binary decision, whether it is to decide which sports team gets to choose the initial conditions of the match, or to solve a dispute between individuals. Coin flips have been an easy solution to small conflicts.

This is not without consequences. Jaffé et al. (2022) found the outcome of a coin flip to have an effect on the feelings of the involved parties . When the coin flip was used to make decisions, if the result of the coin flip was in-congruent with the desired outcome, feelings of forfeiture would increase. While feelings of validation increase when the outcome is congruent with the desired outcome.

In addition to these feelings, it seems people in general are bad at judging randomness, Williams and Griffiths (2013) explain this phenomena using a Bayesian framework, they found people generally do not have sufficient data to accurately assess whether any random outcome is actually random. This lack of data could for example be too few coin tosses to judge whether this outcome is systematic or random. This distinction between systematic processes and random processes is important when it comes to coin flips. According to Keller (1986) coin flips are deterministic, if one was to know the exact starting speed and rotations per second of the coin, one could predict the resulting outcome with 100% accuracy.

However in regular coin flips there are multiple factors that make coins flips unpredictable, firstly usually a human is flipping the coins, therefore the force that is used to hit the coin, and the exact position the coin is hit vary. Secondly, during a coin flip the hand of the thrower moves and tilts, considering this the position of the hand and coin can vary a lot from toss to toss, which introduces another source or randomness. Apart from the human related "error", there is also some error regarding the physics which Keller did not account for such as air resistance and coin thickness, since Keller assumed the coin to be infinitely thin.

Diaconis et al. (2007) investigated the trajectory of coin using a high speed camera to measured the speed and rotations of the coins, which they used to model the coin as rigid bodies. They found coins to have a slight bias toward landing with the same side up as the starting position. 50 coin flips where recorded, but only 27 provided useful information. Considering this Diaconis et al. concluded coins to have a 51% chance of landing face side up, this is the rounded up mean from 27 coin flips, the actual mean found was 50.8%, and when the point Diaconis et al. considered an outlier is removed this drops to 50.69%.

The goal of this thesis is the evaluate whether human performed coin flips are "fair", meaning they have the same chance of landing on either side, or if there is a same side bias. In order to investigate this 132234 coin flips have been collected.

Method

Data collection

The researchers will be flipping a coin, catching in their hand, and saving the outcome for each throw, every toss start in the position the last toss ended. A trial is a sequence of 100 coin tosses. All throws will be recorded and made public alongside the raw data. The researchers must show the outcome clearly to the camera after each flip, and invalid flips will be noted as an "x". The researchers are free to do any number of flips per day. Some researchers will be making use of a script to help the data entry process. All scripts, models and data can be found at the GitHub repository (Sahrani, 2023), all the records can be found in an OSF repository (Bartoš, 2023).

Recording. All coin flips must be recorded on camera, in order to allow for auditing after data collection. The recordings have to clearly show how to coin landed and preferably show which side of the coin was facing up when the coin flip was initiated, since this would give the auditors another way to verify the outcome of the coin. This could be done in anyway convenient to the researcher. E.g. Some used an overhead camera, where the camera was looking directly down on their hand. Others had a camera in front of them and showed the coin every toss. Apart from showing the start and final position of the coin, the researchers also have to show part of the trajectory of the coin, this would not be very clear on camera, but it would make sure all recordings actually contained coin tosses and not someone turning the coin off camera. Preferably all the trials would have their own video, but it was also allowed to have videos of multiple trials if the trials could easily be found during auditing, by using timestamps for example.

validity. Determining when a coin toss has failed is troublesome. Some circumstances are very clear, any coin that touches another surface before touching the hand counts as a failed toss. Other instances can be more unclear, a coin flipping over once it has hit your hand could be an issue according to the literature. Diaconis et al. noted that spinning a coin is often very biased towards one side and therefore one should not let the coin hit the ground after tossing it because the spinning bias would be introduced in the result of the coin toss. In a hand however, the coin rarely spins and more often just flips over once. Apart from this fact, In practice the procedure becomes close to impossible when trying to make sure no flipped coins are included in the data analysis. Therefore any result that ends with one of the to side facing up in the hand, without touching another surface, was counted as a valid coin toss. invalid coin tosses were noted, but their results where not, this would help with the auditing process later.

Script. I developed a python script which would ask the user for the coin they are currently using, and their preferred key-bindings for inputting the data. This simplified the data collection process, by automatically only counting successful tosses and giving an easy way to correct potential mistakes. This script automatically writes all the data into a csv file including: date, time, the coin and the starting position of that trial, and the sequence of head and tails of that trial. Researchers who opted not to use the script were allowed to use any other method. As long as they could provide the same information as the script output would, date and time could however be substituted with a sequence number.

Analysis

Our primary focus is the probability of a coin to land same side up, therefore the first analysis is a simple Bayesian model. Using the Bayes factor the probability of the coin landing same side up being is evaluated. Comparing a point prior of 0.5 for the null hypothesis against a beta distribution with parameters $\alpha = 5069$ and $\beta = 4031$. The possible bias in researchers, and coins will also be considered in later models. All of these models are written in the Stan language. The Rstan package (v2.26.1; Stan Development Team, 2020), and R are used to pass the data into Stan and to calculate the Bayes Factors. Python is used to transform the data into the right format for the model.

Priors. All three models use the same prior for the θ_1 parameter, with $\alpha = 5069$ and $\beta = 4931$ these were chosen based on the findings of Diaconis, preserving the ratio found in the paper, but scaling both parameters in order to make the distributions sufficiently narrow. All θ_1 priors are truncated at 0.5 as to represent the expectation of only finding a positive same-side bias. The priors for the hierarchical models will all be the same. All being normally distributed around 0, and a nested prior for the standard deviation with mean 0 and standard deviation 0.01.

Models. The null model is:

$$X_i \sim Ber(N_i, 0.5)$$

The Alternative model is:

$$X_i \sim Ber(N_i, \theta_1)$$

Where:

$$\theta \sim Beta_{[0.5,1]}(5070, 4930)$$

All models will be compared using Bayes factors.The first model containing just the same-side bias of a coin flip, using θ_1. All hierarchical models will extend the simple model by adding the bias terms to θ_1 in the following fashion:

$$X_i \sim Ber((N_i, \theta_1 + \eta_t + \eta_c)$$

Where X_i is success, θ is the global mean chance for landing same side up, η_t is the bias term for tosser t and η_c is the bias term for coin c. One model will only add η_t, another will only add η_c, and one model using both parameters Where:

$$\eta_t \sim Norm(0, \sigma_t)$$

$$\eta_c \sim Norm(0, \sigma_c)$$

Where:

$$\sigma_t \sim Norm(0, 0.01)$$

$$\sigma_c \sim Norm(0, 0.01)$$

Audit

Ideally all data would be reviewed to check for any fraudulent data entries, and for any possible errors the tossers made in the data entry. Due to the large amount of data this would be infeasible, therefore different methods will be used to ensure the validity and reliability of the data.

Fraud. Random generation of tosses is verified by looking at the success rate of an individual tosser, and comparing it the expected distribution of tosses. This is done by comparing every tossers distribution of number of successes against the expected distribution of successes based on the mean of all their data. When the distribution of the sum of successes in each trial did not follow the same shape as the expected distribution, the tosser would be flagged and more trials would be audited as compared to the other tossers

Instead of fabricating all the data, one could also collect a part of the data and duplicate it. A python script will be used which will compare every trial to every other trial on the longest uninterrupted matching sequence, all invalid tosses were excluded since those do not effect the results of the analysis. Data that was completely duplicated, would very obviously be detected with this method, showing 100% matches for all trials. One could however cut up the videos and trials in such a way that the trials have less overlap, for a small number of cuts, this would still result in abnormally high overlap, but if a person would mix and match smaller and smaller sequences this method would become less and less effective. But this method of splicing video's increases the general chance of the splicing being picked up during the random auditing that will be performed later, thereby balancing it out.

Collection errors. Finally we also have to account of the inevitable human error in data collection. Considering each researcher will be manually flipping and recording the hundreds or even thousands of coin flips in a session, mistakes are bound to occur. In order to estimate the number of errors, random trials are re-encoded by watching the recording of that trial and compared to the trial submitted by the tossers. Using the reported trial every reported outcome can be compared to the trial from the audit.

One issue that arises when comparing the reported sequences to the sequences found during the audit, is that of using head and tails to denote the different sides of the coins. Most of the time the side of coin is marked, this combined with sub-part video quality makes it practically impossible to tell which side is head most of the time. In order to solve this the Levenshtein distance was used Levenshtein, 1965, this distance represents the minimum number of changes that have to be made to get from one string of characters to another. If a sequence from the audit would have a lower Levenshtein distance to the reported sequence after switching all heads for tails and vice versa, the switched sequence would be used.

Another issue with one to one comparing of the trials is inevitable miss alignment, if a tosser or auditor misses 1 toss, the number of mismatched between the original data and the audited data will increase. This alignment issue will be solved using the Needleman and Wunch Algorithm Needleman and Wunsch, 1970. This algorithm places spaces in the data set to optimize overlap between the two sequences. A penalty can be set for adding extra spaces and for mismatches in the data. For this analysis the penalty for spaces was set to 3 and the penalty for mismatches was set to 1, this reflects that adding spaces should be done more sparingly. The code to perform this algorithm was taken from "Sequence Alignment problem - GeeksforGeeks" (n.d.).

The final consideration is that a camera might not always perfectly capture the result, therefore errors have to be treated with relative caution. If no fraud is detected, any outcomes that were not captured by the camera are more likely to be accurately captured by the tossers. Therefore spaces placed by the algorithm will count as not count as a mistake when they are present in the auditor's sequence, but they will when they are present in the reported sequence.

Bias in Errors. Considering the design of the experiment errors will only be an issue when they are biased. Considering the expected effect size is less than 1%, if the error rate is of the same order of magnitude, bias in the errors could exaggerate or minimize the actual effect. In order to evaluate bias in the errors, we can use the model that is used for the bias of each researcher, where θ is the error rate, and η is the variance in error rate. Using a similar model as for the bias in tossers, bias in successes can be treated as a new term. Resulting in a model which doesn't account for bias in errors and one that does, which we can then compare. If there is no evidence to include bias we can conclude that there is no bias in the error rate.

The model to estimate the error rate is:

$$X_i \sim Ber(N_i, \theta)$$

Where:

$$\theta \sim Beta(1, 1)$$

The model to estimate the error rate and bias in these errors is:

$$X_i \sim Ber(N_i, \theta + \eta_s)$$

Where:

$$\theta \sim Beta(1, 1)$$

Where:

$$\eta_s \sim Norm(0, \sigma)$$

Where:

$$\sigma \sim Norm(0, 01)$$

Results

Descriptive statistics. We collected a total of 134,234 coin flips from 14 different tossers. The data was collected over the course of three months, with six individual tossers collecting 75036 flips, and eight different tossers collecting 59197 flips in one day. Out of all the flips, 50.86% landed on the same side they started. We also found that there were an approximately equal number of reported heads and tails, with 0.498% being heads (66885 heads, 67348 tails). 28 different coins were used. In total 60 sequences were audited resulting in approximately 6000 coins flips being checked.

Same-side bias. The evidence for same-side bias in a coin toss is very strong when comparing the simple alternative model to the null model, with $\text{BF}_{01} = -1.05 \times 10^{-}8$. However, when comparing the simple alternative model to models that account for variability in tossers, coins or both, the model that accounts for tosser variability has strongest evidence with $\text{BF}_{01} = 1.75 \times 10 - 16$.The models containing variability in coins were both less likely compared to the model containing only tosser variability. In table 1 all models are compared to the null model and the model account for tosser variability.

Audit

Fraudulent data generation. Figure 2 compares the sum of success for every trial against the probability mass function based on the tossers success rate. This approach incorporates possible differences in tossers. Comparing the expected and the observed distributions of success rate for each person raised no suspicion for fraudulent data generation.

Data duplication. Comparing the sequences found in the data set amongst each other we found the longest duplicate sequence to be 23 coin flips long. This is comparing every sequence in the data set to every other sequence in the data set. This cross checking was done because the tossers had access to a part of the data set before the collection had concluded. Lastly the number of input errors made by the tossers was analysed, any misalignment issues either due to the tosser or the auditor missing inputs were solved using the Needleman and Wunch Algorithm. If after re-aligning with the algorithm more than 10% of the comparison between original and audited inputs were wrong, the audit was not used. This resulted in 5 audits with 134 mistakes being removed, resulting in 55 trials or 5531 coin tosses remaining. Considering the previous steps revealed no evidence of fraud during the data collection process, this removal of data ensures the auditing data which was used was of high quality. The estimated error rate of after this removal of low quality data was 1.05% (95%CI = 0.80%, 1.34%). Comparing this to the model that accounts for bias between successful and unsuccessful tosses, there is very little evidence in adding a bias term for the success($\text{BF}_{10} = 1.05$). Adding the tosser as a bias term instead of the success produces similar results ($\text{BF}_{10} = 1.09$), for both Bayes factors use the simple model was used as the null model, and the model with variability as the alternative model.

Conclusion

Our results strongly support the hypothesis from Diaconis et al. (2007). It seems the type of coin does not matter for this same side bias to show. There does seem to be variability between tossers on the chances of landing same side up. This is most likely due to a difference in techniques, whether that is in the form of differing angles of the trajectories, or due to some tossers hitting the coin with different levels of force. It could be the case that the tossers learned to control the coin tosses, in which case a future study could replicate the methods performed in the current study with different participants. These participants could be told the goal of the study is to study the fairness of landing head or tails, which might prevent the participants from trying to land same side up, instead focusing their attention on the face of the coin. Another possible explanation could be differences in technique are innate, although none of the tossers' technique was consistent enough to have a major influence on the success rate, it could be that some of the tossers were more likely to toss a coin with a certain speed, spin or angle to begin with. In order to investigate this, high quality and high speed camera recordings could be used to estimate the starting parameters of coin tosses, and how they differ between people. The audit does introduce an interesting problem, although there is strong reason to believe all the tossers were honest and made few mistakes, and the auditors did the same. This procedure hinges upon these assumptions. If there was a tosser with many mistakes, the current procedure might have over-correct and removed some mistakes that it should not have. The main issue with the auditing was that of video quality, bad lighting and sub-optimal angles resulted in the auditing being difficult at times. This would be solved by equipping the tossers with the tools necessary to improve the recording, this could include a better camera, or a tripod for a phone to be held above the hand, which would make sure the outcome of the coin toss is always visible. If an auditor was inaccurate or hasty, this procedure would over correct for their findings. if they miss a lot of tosses and mark them as unclear, this tosses will be considered correct, even when they are not. Being more strict about when an audit is considered correct, or cutting out all tosses that were deemed unclear, in the case of a bad auditor, would lead to punishing the good auditors in the cases where the video's really were hard to distinguish. If this procedure were to be replicated it would most likely be a good idea to take still frames from the video's, this would allow for more specific checking, and easier comparison between auditors. If the still frames are ordered, any part of a sequence that is especially noisy can quickly be re-coded and compared again.

Coins bouncing during the toss introduces another issue, as justified in the data collection section, in the current study the coin flip was valid as long as the coin did not leave the hand before you could see the outcome of the toss. However according to Diaconis et al. (2007)

coins show much stronger bias when bouncing and spinning. Although this bias relates to the side of the coin which is heavier, not to which side of the coin started up, it could have affected the results of the current study. A simple solution would be to only count coin tosses which do not bounce at all, but this is tricky in practice, a coin might move a little bit but not switch sides, or it might it part of the hand and slide along. Not only that but it might make data collection very inefficient by having to discard many coin tosses. This along with the fact that in everyday life people are rarely concerned with the bouncing, we considered the bouncing to be of little importance to natural human performed coin tosses

In conclusion, the findings of this study provide strong support for the current literature on coin tosses and the same side bias coin tosses have. The results suggest that the type of coin does not significantly affect the bias, but the person performing the toss may play a role. Future studies could explore this further by recruiting participants and controlling for factors such as attention and technique. One potential limitation of the study is the quality of the audit data, which could be improved by providing better recording equipment. Additionally, the issue of coin bouncing was not fully addressed in this study, and it may have had an impact on the results. Overall, the current study has added to the existing literature by further investigating the factors that contribute to the same side bias in coin tossing, and has provided insight into potential areas for further research.

References

Diaconis, P., Holmes, S., & Montgomery, R. (2007). Dynamical Bias in the Coin Toss. *SIAM Review*, *49*(2), 211–235. http://www.jstor.org.proxy.uba.uva.nl/stable/20453950

Jaffé, M. E., Douneva, M., & Greifeneder, R. (2022). When Choosing Implies Losing: Does Flipping a Coin Increase Forfeiture Thoughts? *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.33941

Keller, J. B. (1986). The Probability of Heads. *The American Mathematical Monthly*, *93*(3), 191. https://doi.org/10.2307/2323340

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Sequence Alignment problem - GeeksforGeeks. (n.d.). https://www.geeksforgeeks.org/sequence-alignment-problem/

Williams, J. J., & Griffiths, T. L. (2013). Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1473–1490. https://doi.org/10.1037/a0032397

Table 1

Bayes factors for different models

Model	**XXX**	**XXX**	**XXX**
Simple model	-93025.5	1.04152e-08	59411597
Variability in Tossers	-93007.64	1.753058e-16	-
Variability in Coins	-93013.55	6.463631e-14	368
Variability in Coins and Tossers	-93007.65	1.770676e-16	1.01005
Null model	-93043.92	-	5.704319e+15

Table 2

Mean, Standard Deviation, and Number of tosses

Participant	**Mean**	**Std**	**Number of Tosses**
IrmaT	.504	.500	701
JonasP	.510	.500	4996
JillR	.5045	.500	6463
MadlenH	.522	.500	7099
HenrikG	.511	.500	7382
FrantisekB	.499	.500	7900
AdamF	.520	.500	8328
AlexandraS	.534	.499	8434
IngeborgR	.505	.500	8596
KaleemU	.493	.500	14324
DavidV	.506	.500	14999
DavidKL	.526	.500	15000
PierreG	.500	.500	15000
AmirS	.497	.500	15012

Table 3

Mean, Standard Deviation, and Number of Tosses

Coin	**Mean**	**Std**	**Number of Tosses**
0.05EUR	.514	.500	4213
0.05NZD	.514	.500	2011
0.10EUR	.510	.500	5165
0.20EUR	.500	.500	11968
0.20GBP	.479	.500	2005
0.20GEL	.510	.500	5598
0.25USD	.491	.500	1000
0.50GBP	.491	.500	1504
0.5EUR	.506	.500	16298
100JPY	.505	.500	1500
10CZK	.495	.500	1000
1CHF	.488	.500	1500
1CNY	.498	.500	1500
1CZK	.510	.500	1000
1EUR	.498	.500	11728
1HRK	.505	.500	8596
1MAD	.491	.500	2000
1MXN	.534	.499	8434
1SGD	.526	.499	15000
20DEM-silver	.519	.500	1000
2CHF	.497	.5000	4503
2EUR	.508	.500	10384
2MAD	.506	.500	3000
5CZK	.496	.500	2500
5JPY	.506	.500	1500
5MAD	.508	.500	2001
5ZAR	.517	.500	7326

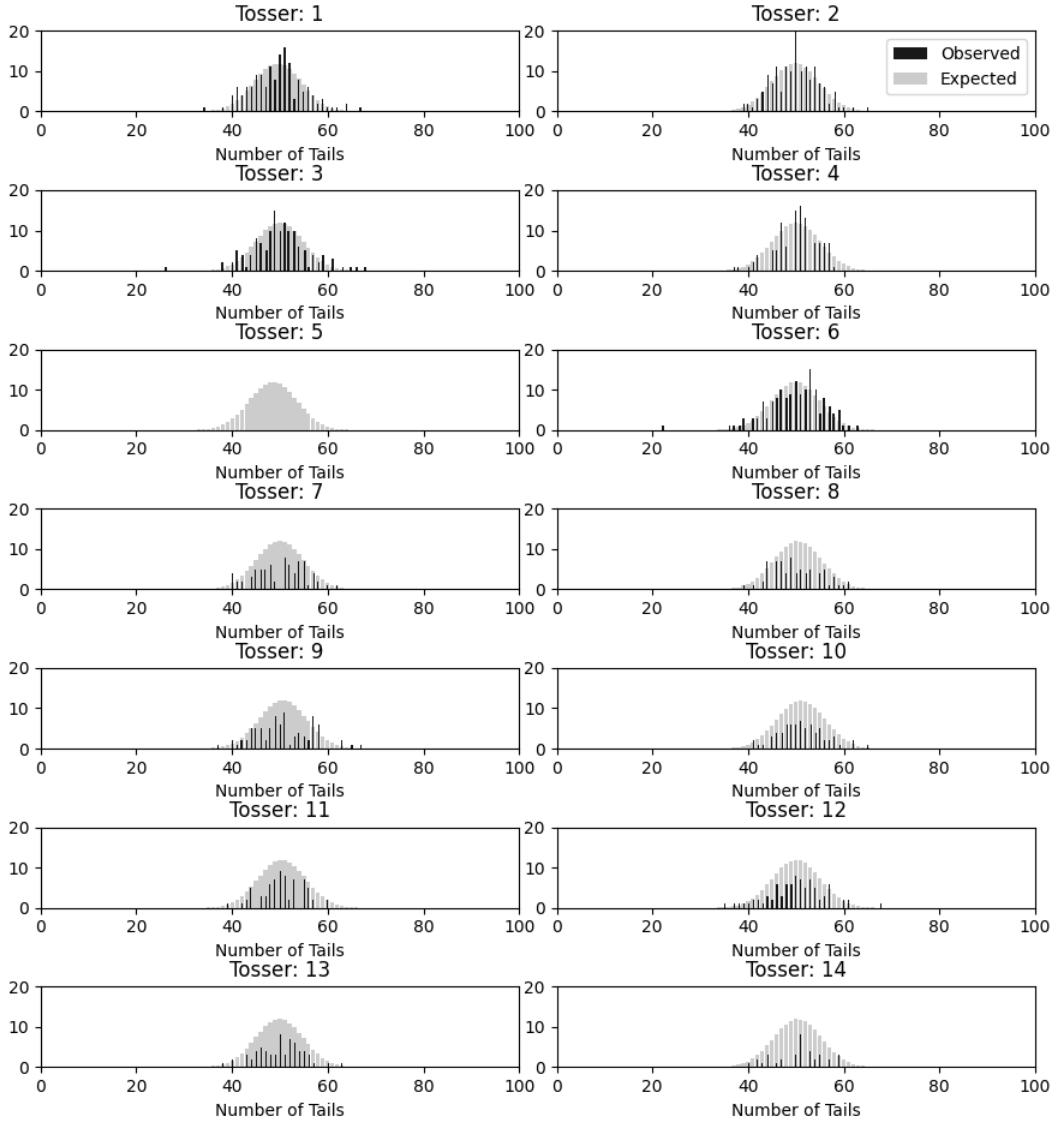

Figure 1. Distribution of various tossers