Coin-Flip Experiment: A Regression-Based Approach

MATH-408 - Final Project

École Polytechnique Fédérale de Lausanne (EPFL)

Kalil Bouhadra
kalil.bouhadra@epfl.ch

Gabriel Marival
gabriel.marival@epfl.ch

January 11, 2025

Contents

1 Introduction **1**

2 Analysis **4**
2.1 Intercept-Only Binomial Regression 4
2.2 Normal Approximation Approach 5
2.3 Fixed, Random and Nested Effects 5

3 Discussion **7**
3.1 Outliers Treatment 7
3.2 Overdispersion 8
3.3 Learning Effect 10
3.3.1 Modeling Time Dynamics 10
3.3.2 Influence of Previous Flips 10

4 Conclusion **12**

References **13**

1 Introduction

At first glance, a coin appears to be the perfect symbol of chance - a 50-50 bet. However, the detailed work of Diaconis, Holmes, and Montgomery 2007 and Bartoš et al. 2023 reveals a surprising aspect: coins tend to "remember" their starting side. Through rigorous data collection of over 350'000 flips and Bayesian analysis, they have shown that even this most iconic of random processes may not be as impartial as we might think. The present analysis takes a different approach to studying the data collected in the study. Instead of assuming a Bayesian model and conducting statistical tests, we apply regression methods to extract insights from the coin-flip data. In the context of regression methods, by 'extracting insights', we refer to constructing statistical models, fitting them to the data, and interpreting the results. Both the construction and interpretation phases are as crucial as the fitting process, as two models with equally good fits may lead to different interpretations of the data (Davison 2024, p. 43, 60, Chapter 1.4, 1.5).

Before discussing suitable regression models, let's examine the content of our data (it consists of two datasets both retrieved from the work of Bartoš et al. 2023). The primary dataset used in our analysis is `data-agg.csv`. Minor formatting adjustments were made to obtain the following structure: each row in the dataset corresponds to a run of coin flips performed by one person with one coin and a specific starting face. The columns provide information on the number of flips in the run, the number of successes (number of times the coin lands on the same side as the initial throw in a given run), the starting face, the person who flipped the coin, and the type of coin. The dataset includes 48 participants, 44 coins, and 422 runs (i.e., 422 rows), evenly split between starting heads and starting tails (211 rows for each starting face). The other dataset we used is `df-time-agg.csv`. Each row represents a block of 100 coin flips performed by one person with one coin and a specific starting face. The dataset aggregates outcome counts, success numbers, and time-related information for each block, providing a structured basis for analyzing trends and learning effects over successive blocks of flips.

The empirical distribution of the success proportion in Figure 1 provides an initial insight into the same-side bias. We observe distributions slightly shifted to the right of $1/2$, with a larger proportion of values exceeding one-half. The mean success proportion is $\hat{\mu} = 0.508$ for both starting faces. It is also worth noting the presence of a few outliers, particularly on the right tail of the distribution. At this stage, one might question the extent of their impact on the results and whether it is appropriate to retain them in the sample. Lastly, because both empirical distribution are similar, the starting face may not have an impact on the success proportion, but this is something we will confirm later when evaluating models.

To gain a clearer understanding of the distribution of success proportions, particularly regarding outlier values, we grouped the data in three different ways: the first grouping is by person, where each value represents the average success proportion across all runs for a specific person. The second grouping follows the same process but it is for each coin. The final grouping is by person-coin combinations. This approach aims to identify whether certain persons, coins, or person-coin combinations exhibit extreme values, and to determine if one grouping shows a significantly higher variability. Figure 2 presents the kernel density estimation of the distributions, along with the corresponding boxplots, displayed as three violin plots. The left plot, representing the success proportion by coin, appears approximately normal, with a symmetric boxplot and distribution, and only one outlier. The middle plot, which shows the success proportion by person, exhibits significantly more variability and a less symmetric distribution, with three outliers positioned far from the boxplot's whiskers. Lastly, the person-coin combination plot presents a relatively symmetric distribution but

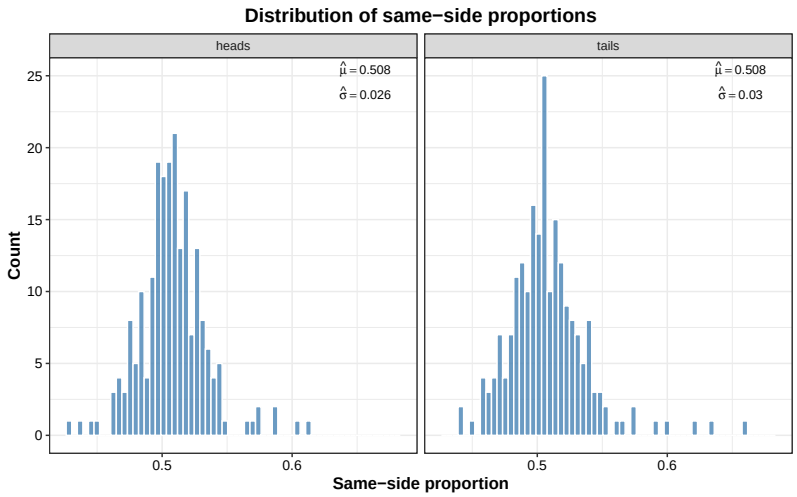


Figure 1: Graph illustrating the empirical distribution of success proportions for both starting faces: heads on the left and tails on the right. The empirical mean and standard deviation are displayed in the top-right corner of each plot. The distributions are nearly identical, with a slight shift in the mean away from 0.5, suggesting a potential same-side bias in the coin flips.

with several extreme outliers. As the next step involves constructing regression models and fitting them, we must carefully handle these outliers and check the impact they have on the fit, this will be described in the Discussion section.

Another aspect we would like to examine as part of the exploratory analysis is whether there is a temporal dependence in the success proportion. This could indicate, for example, the presence of a learning effect, as suggested by Bartoš et al. 2023. To investigate this, we computed the autocorrelation function (ACF) of the success proportions across aggregated runs using `df-time-agg.csv` (Davison 2024, p. 47, Chapter 1.4). Specifically, we grouped the coin flip runs into consecutive blocks of one hundred flips and calculated the success proportion for each block. The ACF plot, as shown in Figure 3, was then constructed by computing the correlations between the success proportions at different lags. As seen in Figure 3, the autocorrelation values fluctuate around zero or switch between positive and negative values, indicating an absence of a clear temporal trend in the success proportions. While these preliminary results do not show strong evidence of temporal dependence, further exploration of potential learning effects and temporal dynamics will be detailed in the Discussion section.

As previously discussed, while the analysis of Bartoš et al. 2023 utilizes a Bayesian framework for analyzing coin flips, focusing on prior and posterior distributions for the probability of landing on the same side, we will, in the subsequent sections, attempt to fit regression models to the data and extract meaningful information from these models. Our exploratory analysis has provided indications of influential factors in our data and guidance on which model to use. First, the empirical distribution in Figure 1 suggests a slight same-side bias and no significant impact of the starting face. Second, the violin plots in Figure 2 show few outliers but some variability, particularly when considering certain individuals and specific person-coin combinations. Finally, the autocorrelation plot in Figure 3 exhibits behavior that is difficult to interpret, it may be due

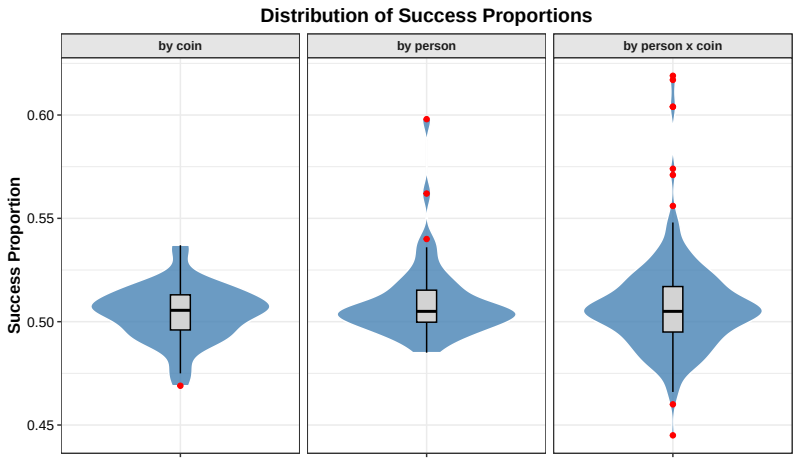


Figure 2: Violin plots showing the distribution of success proportions for three groupings: by person, by coin, and by person-coin combination. Each plot includes a kernel density estimate of the distribution and a boxplot. The central and the right plots exhibit a higher number of outliers compared to the left plot, indicating greater variability in these groupings.

to randomness or may reflect a real trend. These observations inform our approach to constructing regression models. Formally, regression can be defined as a measure of the relationship between the mean value of one variable (the response variable) and corresponding values of other variables (explanatory variables) (Davison 2003, Chapter 8). In our case, the response variable is the success, or same-side proportion, which represents the number of times a coin lands on the same side that it started on a run. In other words, for run i, if we start with heads, let m_i be the total number of flips in run i and y_i the number of times the coin lands on heads. We then define our response variable as $p_i = y_i/m_i$. Explanatory variables may or may not be present depending on the model we fit. One possible assumption is that we observe a same-side bias independently of any other information. In that case, no explanatory variables are needed. This baseline model is the simplest form, with a straightforward formulation and only one parameter to estimate. However, one might argue that the person who flips the coin, the coin itself, or the starting face could impact the success proportion. In such cases, we would include categorical variables to properly model these factors. Determining whether to include these variables, which ones to include, and how to model them appropriately will be the focus of the subsequent Analysis section. The Discussion section will further delve into comparing these models, analyzing the variance and deviance exhibited by the models, assessing the impact of outliers, and discuss some models that highlight the temporal effect in our data. Finally, the Conclusion provides the take-away message from the full study. It discusses the strengths and limitations of our approach and possible extensions for future work.

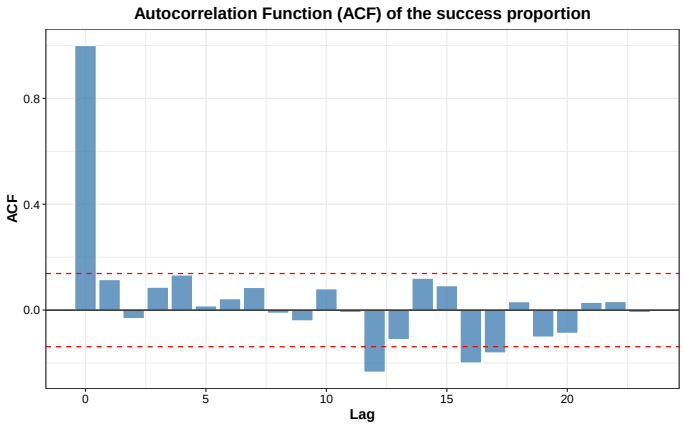


Figure 3: Autocorrelation Function (ACF) of the success proportion over time. The bars represent the autocorrelation values at different lags, where the lag indicates the time difference between observations. The value of the ACF at lag j indicates the correlation between success proportion at time t and time t+j.

2 Analysis

When performing regression, a common mistake is to focus too much on finding the model that best fits the data, while overlooking how the model interprets the relationship between predictors and the response. Two models may fit the data equally well yet offer different explanations of this relationship (Davison 2024, p. 60, Chapitre 1.5). In analyzing our coin flip data, our approach is guided by two primary aims: understanding the underlying phenomena and achieving accurate predictions. While our immediate focus is on selecting a well-fitting model, we recognize that no single model may perfectly balance understanding with prediction. Therefore, in our subsequent analysis, we will evaluate various regression models, compare them using criteria like Akaike information criterion (AIC), and weight both their fit and interpretability.

2.1 Intercept-Only Binomial Regression

The simplest model to fit includes only the intercept, without any covariates, indicating that the success proportion is independent of external factors. Since our data represents proportion data, taking values of either 0 or 1 with a certain probability (the probability of success we aim to estimate), we will use a logistic distribution with a logit link function (common choice) (Davison 2024, p. 135, Chapitre 2.4). This leads to the following expression for the success proportion p:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 \tag{2.1}$$

This implies that for every observation, we estimate the success proportion as defined by Equation 2.1, attributing any variation in the probability of success solely to random chance. When fitting the model, the estimate of β_0 is slightly greater than 0, resulting in a success proportion of $p \approx 0.508$. The Model 2.1 provides strong evidence for $p \neq 0$, as the 95% confidence interval using

normal approximation is $[0.5060, 0.5094]$. β_0 is the maximum likelihood estimator (MLE), found via the Iterative weighted least squares (IWLS) algorithm implemented in R's glm() function. Thanks to the asymptotic properties of MLE, the Wald confidence interval for p remains valid (provided we have sufficient data and the model assumptions hold).

2.2 Normal Approximation Approach

An alternative approach to the previous one is to treat each observation as normal random variables: $p_i = R_i/m_i \sim \mathcal{N}(p, 1/(4m_i))$ where R_i are binomial random variables and m_i to be estimated, this lead to the following linear Model 2.2:

$$p_i = p + \varepsilon_i, \quad \forall\, i = 1, \ldots, n \tag{2.2}$$

with weights m_i to reflect that a proportion based on m_i flips has variance $1/(4m_i)$ ($p(1-p) \approx 1/4$ when $p \approx 0.5$). As in the previous model, the coefficient p represents the grand mean of the observed proportions. Unlike in the logistic model, there is no link function here. We directly get an estimate of the overall same-side probability, analogous to what we obtained in the binomial GLM, but on a different modeling scale. For large m_i (which is assumed to be the case here as m_i usually higher than 400), we know from the central limit theorem that the normal approximation is reliable. So the Model 2.2 we define should provide similar estimate of the same-side probability. However this approach could be suboptimal in cases of overdispersion due to external factors and doesn't account for the binomial nature of the data. Indeed, the results show similar results with strong evidence for same-side bias but with a larger confidence interval for the same-side proportion. Also, Figure 4 displays the Q-Q plot of the pearson residuals from the previous Model 2.1. We observe some deviations from the diagonal in the tails, suggesting that the residuals are not perfectly normally distributed (Davison 2024, p. 167, Chapitre 2.9). This is not entirely surprising, given that our response variables are originally proportions or counts data, so we may prefer the previous binomial Model 2.1. This may also suggests additional heterogeneity (e.g., between persons or coins).

2.3 Fixed, Random and Nested Effects

Pursuing in the pointed direction, we would like to adapt our model to take into account the possible variability due to participant, coin, or initial face. In order to do that, we add fixed effect terms as categorical explanatory variables in our base GLM model, giving the following formulation:

$$\operatorname{logit}\big(p_{ijk}\big) = \beta_0 + \beta_i\, \mathbf{1}\{\text{initial throw} = i\} + \beta_j\, \mathbf{1}\{\text{person} = j\} + \beta_k\, \mathbf{1}\{\text{coin} = k\}. \tag{2.3}$$

In this Model 2.3, p_{ijk} denotes the probability of obtaining a same-side outcome when the initial throw is of type i (e.g., heads or tails), for the jth person and the kth coin. The parameter β_0 is the global intercept on the log-odds scale, representing the baseline same-side logit. The coefficients β_i, β_j, and β_k capture fixed effects for, respectively, the initial face of the coin, the individual flipping the coin, and the coin itself. In order to know if every variables is sufficiently explanatory and do not complexify the model too much, we'll perform variable selection using AIC to compare models. In general, testing every model could be computationally too expensive and we use stepwise methods (Davison 2024, p. 74, Chapitre 1.6), but because we have only three variables, we fit every 2^3 choices of covariates. By doing that we find that the model minimising the AIC is the one with the

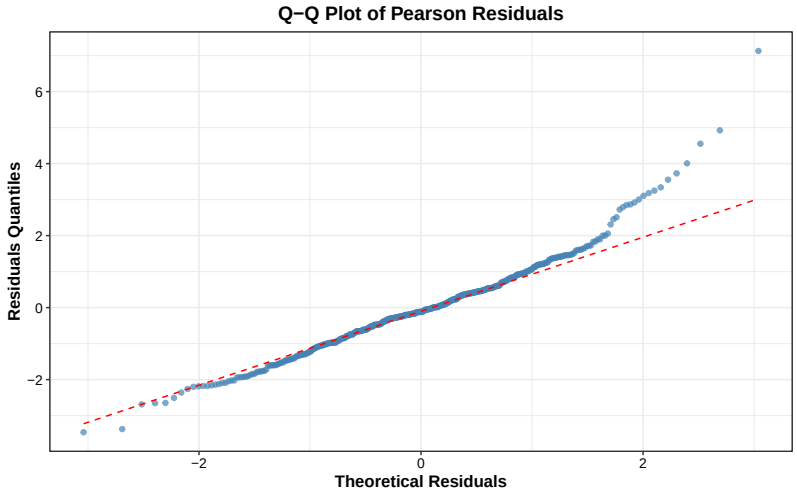


Figure 4: Q-Q Plot of the Pearson residuals from the simple binomial GLM. The plot assesses the normality of residuals, with deviations from the diagonal, particularly in the right tail.

intercept and the person variables, as detailed in Equation 2.4. The results for models evaluation are summarized in Table 1:

$$\text{logit}\big(p_j\big) = \beta_0 + \beta_j \, \mathbf{1}\{\text{person} = j\} \tag{2.4}$$

Model	**Variable Description**	**AIC**
$\text{Logit}\big(p_{ijk}\big)$	Initial face, coin and person	3443
$\text{Logit}\big(p_{ij}\big)$	Initial face and person	3426
$\text{Logit}\big(p_j\big)$	Person	3424
$\text{Logit}\big(p\big)$	Intercept only	3596

Table 1: Comparison of models based on AIC values. The model including the person contribution shows the lowest AIC value, indicating that it provides the best fit to the data.

In the fixed-effects Model 2.4, we introduce a separate coefficient β_j for each person j, thereby interpreting each individual's flipping bias as distinct. Even if it may exhibit a better fit, this approach presents two drawbacks depending on the context: the number of parameters can rapidly become huge if the number of persons grows large, and such a model treats each person as a fixed quantity of interest, rather than assuming that these persons represent a sample from a broader population (Davison 2024, p. 266, Chapitre 3.6). To adress these limitations, a natural alternative is to consider random effects for persons, i.e., to assume that each person's deviation b_j from the global bias β_0 is drawn from a common distribution $b_j \sim \mathcal{N}(0, \sigma_b^2)$. Under this assumption, the model focuses on estimating a single variance term σ_b^2 (the population-level variance among

individuals) rather than all individual coefficients β_j. Mathematically, the fixed-effect Model 2.4 is replaced by

$$\text{logit}\big(p_j\big) = \beta_0 + b_j \quad \text{with} \quad b_j \sim \mathcal{N}(0, \sigma_b^2). \tag{2.5}$$

When computing the AIC of this new Model 2.5, we obtain 3477, which is lower than that of the baseline Model 2.1 but higher than that of the fixed-effect Model 2.3. However, due to its advantages, namely the reduced number of parameters and its ability to generalize to a broader population, one might consider retaining that model.

Carrying on the analysis, we observe that in our dataset, different coins are used by different participants, but they are not truly shared across individuals (eliminating the need to test crossed effects). Therefore, the coin factor can be nested within the person factor (Davison 2024, p. 267, Chapter 3.6): each participant has their own set of coins, which do not overlap with those of other participants. In this context, we extend the previous Model 2.5 to include a random effect where b_j represents the random intercept for person j, and $c_{j(k)} \sim \mathcal{N}(0, \sigma_c^2)$ represents the random intercept for coin k, specific to person j. This nested random-effects Model 2.6 acknowledges that each participant has their own coin-flipping bias, with additional small variations among coins within that participant, beyond the inter-person variability. This leads to the following formulation:

$$\text{logit}\big(p_{jk}\big) = \beta_0 + b_j + c_{j(k)} \quad \text{where} \quad b_j \sim \mathcal{N}(0, \sigma_b^2), \quad c_{j(k)} \sim \mathcal{N}(0, \sigma_c^2). \tag{2.6}$$

This Model 2.6 has a lower AIC value compared to the previous Model 2.5. Since it is also nested within that model, we perform a likelihood ratio test to evaluate the significance of the additional parameter $c_{j(k)}$ (Davison 2024, p. 106, Chapter 2.1). The test results show that including $c_{j(k)}$ significantly improves the model fit, indicating that the coin factor has a measurable impact on the response.

3 Discussion

3.1 Outliers Treatment

As discussed in the introduction, we observed some variability in the data, especially when looking at person individually, or for some combinations person - coin and this variability exhibit some outliers. Since the available information on the sampling process does not indicate any procedural issues, it would be inappropriate to dismiss certain person or coin contributions outright, as their high impact might reflect an important pattern we seek to capture (Davison 2024, p. 58, Chapter 1.4). However, we also need to ensure that our regression model's results remain robust even without these outliers. Therefore, for every models we discussed in the Analysis section, our approach is as follows: we first fit the model with all data points included. Then, we identify influential outliers by calculating Cook's distance and refit the model excluding points with Cook's distance greater than $8/(n-2p)$, where n is the number of observations and p is the number of parameters of the fit(Davison 2024, p. 52, Chapter 1.4).

Concerning the baseline Model 2.1, we observe in Table 2 that removing the outliers slightly reduces the effect but doesn't change the conclusion: There is indeed a same-side bias independently of any external factors with strong statistical evidence.

Removing the outliers does not drastically alter most of our conclusions regarding the choice between fixed and random effects. The best fixed-effect Model 2.4 (in terms of AIC) still includes

Model	95% Confidence Interval for p
Baseline GLM with full data	[0.506 , 0.5094]
Baseline GLM after outlier removal	[0.5042 , 0.5078]

Table 2: Comparison of 95% confidence intervals for p between the baseline GLM model and the refit GLM model after removing outliers. Removing outliers reduces the effect size but does not change the conclusion, there is still strong statistical evidence of a same-side bias.

the individual factor, although some of the estimated effects are less pronounced. Furthermore, our discussion of switching from a fixed-effect to a random-effect model remains largely relevant here: after outlier removal, the random-effect Model 2.5 becomes almost as competitive as the fixed-effect model.

One key difference is that once the outliers are excluded, the nested effect of "coin within person" is no longer statistically significant. In other words, any improvement that came from modeling coin-level variability within each person (as done in Model 2.6) disappears after removing these outliers. Table 3 presents a comparison of the likelihood-ratio test results before and after removing outliers: when the outliers are included, the nested effect is clearly significant ($p = 0.0025$), whereas after their exclusion, it becomes clearly insignificant ($p = 0.9246$).

Condition	Pr($\geq$ Chisq)
Without removing outliers	0.003
After removing outliers	0.925

Table 3: Likelihood ratio test results comparing the significance of the nested "coin within person" effect before and after outlier removal. After removing outliers, there is no significant effect of "coin within person".

In conclusion, removing these outliers reduces the significance of certain effects (such as the nested coin factor) but highlights others (such as the importance of the random-effect model). In practice, the decision to remove outliers should consider whether those observations represent true variability or experimental noise.

3.2 Overdispersion

In a Binomial GLM, we generally assume to have $\text{Var}(Y_i) = n_i\, p_i\, (1-p_i)$. If the empirical variance exceeds this quantity, we speak of overdispersion. In order to detect this we rely on the following theory: If the model is correct, we get the Pearson statistic from Equation 3.1:

$$P = \sum_{j=1}^{n} \frac{(y_j - \hat{\mu}_j)^2}{\hat{\mu}_j(1-\hat{\mu}_j)} \tag{3.1}$$

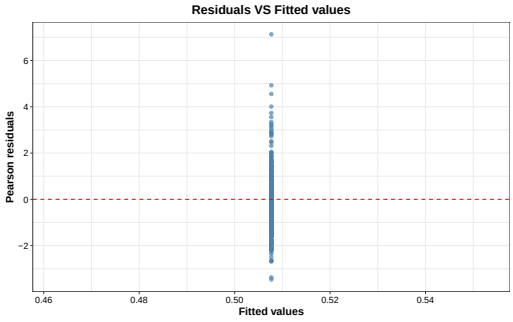
follows a χ^2_{n-p} distribution asymptotically, where n is the number of observations and p the number of parameters in the model. Hence, if the model is correct, there is no need for a dispersion

parameter for extra-variability, its value is 1. Applying the *method of moments* to estimate the dispersion parameter ϕ (better than maximum likelihood estimate here) amounts to solving $P = (n-p)$, since we know $\mathbb{E}\left[\chi^2_{n-p}\right] = n-p$ (Davison 2024, p. 121, Chapter 2.3). Thus, if the model is correctly specified, we obtain $\hat{\phi} = P/(n-p) \approx 1$. The dispersion parameters for Model 2.1 and Model 2.4 are presented in Table 4. We observe a value noticeably above 1 (1.63) for the simpler model, suggesting that the binomial assumption alone may be too simplistic and that overdispersion is present. When we add individual-level covariates as fixed effects, the dispersion ratio decreases to 1.12, indicating that much of the extra-variability is now captured by these additional parameters.

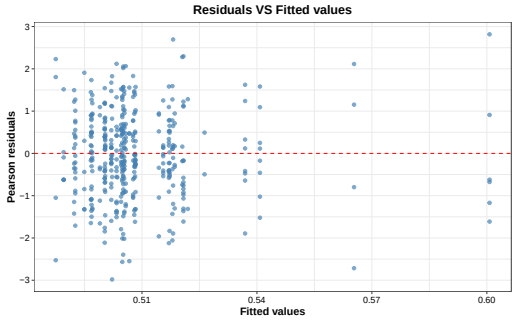
Model	Dispersion Parameter
Baseline Binomial GLM	1.63
Fixed-Effect Binomial GLM	1.12

Table 4: Comparison of dispersion parameters between the baseline binomial GLM and the fixed-effect binomial GLM. Introducing fixed effects reduces the dispersion.

Although this dispersion parameter provides a straightforward way to assess overdispersion in a GLM, a more classical approach involves plotting Pearson residuals against fitted values. Figure 5 shows this comparison for the same two models. In panel (a), the baseline Model 2.1 exhibits Pearson residuals up to 6 in magnitude, as well as a noticeable asymmetry (many large positive residuals compared to fewer large negative ones). This indicates that the model underestimates some success proportions for certain observations. In contrast, in panel (b) for Model 2.4, the residuals are smaller in absolute value and symmetrically distributed, suggesting that including individual-level effects mitigates much of the overdispersion and bias.



(a) Pearson residuals vs. fitted values for the baseline model.



(b) Pearson residuals vs. fitted values for the fixed-effect model.

Figure 5: Comparison of Pearson residuals vs. fitted values for the baseline (a) and fixed-effect (b) models. The baseline model shows higher residuals and a clear asymmetry. Including fixed effects for individuals reduces both the magnitude and asymmetry of the residuals.

Concerning the random-effect model, any previously unmodeled person-to-person (and coin-to-coin) heterogeneity may now be absorbed into the random effects through parameters $b_j \sim \mathcal{N}(0, \sigma_b^2)$

and $c_{j(k)} \sim \mathcal{N}(0, \sigma_c^2)$. An other alternative to decrease the overdispersion would be to use a quasi-binomial approach (Davison 2024, p. 172, Chapter 2.9). However, based on our current results, we assume this is not necessary.

3.3 Learning Effect

This section presents two distinct yet interconnected models that capture different features of how the same-side bias evolves over time. The first approach suggests that the bias weakens as more flips are performed, while the second explores how the same-side proportion of recent flips impacts the current same-side proportion.

3.3.1 Modeling Time Dynamics

A simple way to account for a learning effect is to incorporate a numerical predictor for time. To perform that, we use the dataset `df-time-agg` and consider the variable `mean_toss`, which represents the mean index of each batch of 100 flips. We then fit a binomial model of the form

$$\text{logit}\big(p_t\big) \;=\; \beta_0 + \beta_1 \,\text{mean\_toss}_t \tag{3.2}$$

where a negative β_1 indicates that the same-side probability decreases as the toss index grows. This linear Model 3.2 shows a modest but statistically significant negative slope, confirming a decline in same-side probability over time.

However, the linear form may underfit if the learning process is more complex. To test this, we replace the term $\beta_1 \,\text{mean\_toss}_t$ with a penalized spline, leading to the following model:

$$\text{logit}\big(p_t\big) \;=\; \beta_0 + \mu\big(\text{mean\_toss}_t\big) \tag{3.3}$$

Here, the smooth function $\mu(\text{mean\_toss})$ in our Model 3.3 is represented by a cubic regression splines. Specifically, we express the smooth effect as $\mu(\text{mean\_toss}) \;=\; B(\text{mean\_toss})\,\beta$, where $B(\text{mean\_toss})$ is a matrix of size $n \times p$ whose columns consist of evaluations of p chosen basis functions at each observed value of mean_toss and β is a $p \times 1$ vector of coefficients corresponding to these basis functions (Davison 2024, p. 226, Chapter 3.4). The number of internal knots and the degree of the spline (polynomial of degree 3 here) determine the flexibility of the basis $B(\cdot)$ and are chosen with the library `mgcv` in `R`. It also includes a smoothing parameter λ controlling overfitting by shrinking some coefficients in β (Davison 2024, p. 205, Chapter 3.2). In Figure 6, we show the fitted spline curve and its 95% confidence band; the curve confirms a gradual but clear downward trend in same-side probability over time. In particular, the plot indicates a strong initial learning. As more tosses accumulate, the curve suggests that the bias shrinks after a certain point. The 95% confidence intervals widen where the data are less informative, indicating increased uncertainty for higher mean toss.

3.3.2 Influence of Previous Flips

A second way to investigate learning or adaptation is to add information about recent flips. More precisely, we wanted to examine how the proportion of "same-side" outcomes in the previous block of 100 tosses ($\text{prop\_same\_side}_{t-1}$) affects the probability that the current block of 100 tosses also ends in a "same-side" outcome. In order to do that, we fit the following logistic regression Model 3.4:

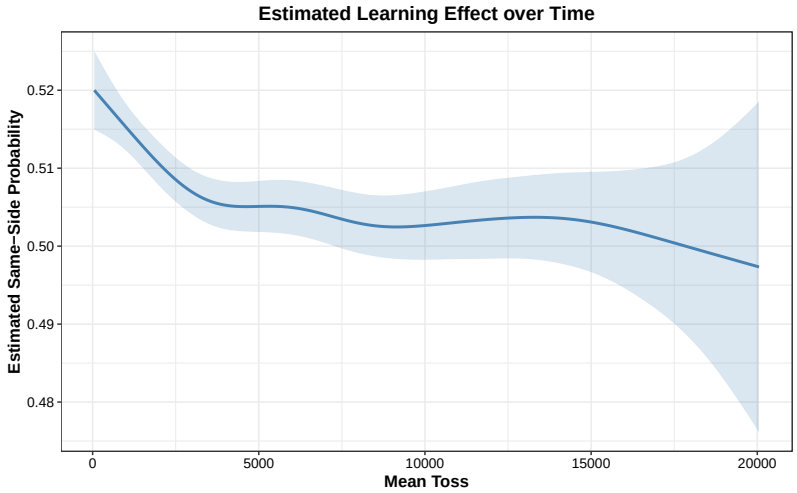


Figure 6: Estimated smooth effect of time on the same-side probability. The shaded region represents a pointwise 95% confidence interval. The plot shows a strong initial learning effect, with the bias decreasing as more tosses are made. The 95% confidence intervals widen where there is less data, reflecting greater uncertainty for higher toss counts.

$$\text{logit}(p_t) \;=\; \beta_0 + \beta_1 \,\text{prop\_same\_side}_{t-1}, \tag{3.4}$$

where $\text{prop\_same\_side}_{t-1}$ represents the proportion of tosses that resulted in a same-side outcome in the previous block. The results of the Model 3.4 are summarized in Table 5, both coefficients are strongly statistically significant. The intercept $\beta_0 = -0.092$ represents the log-odds of obtaining a "same-side" outcome in the current block when the "same-side" proportion in the previous block is zero. This result is really interesting as according to the fit, having no same-side outcome in the 100 previous flips would tend to give a less than half same-side proportion in the current 100 flips, so overcoming, the same-side bias discussed extensively in the study. The coefficient $\beta_1 = 0.240$ measures the effect of an increase in the "same-side" proportion in the previous block on the log-odds of a "same-side" outcome in the current block. Specifically, an increase of 10% in the "same-side" proportion in the previous block increases the log-odds of a "same-side" outcome in the current block by 2.4%. These results suggest that a higher "same-side" proportion in the previous block is associated with an increased probability of observing "same-side" outcomes in the following block.

Parameter	**Estimate (Std. Error)**
Intercept β_0	-0.092 ± 0.033
Coefficient β_1	0.240 ± 0.065

Table 5: Results of the logistic regression model evaluating the effect of the "same-side" proportion in the previous block on the "same-side" probability in the current block. Both parameters have statistically significant effect.

4 Conclusion

This study highlights that even seemingly simple random experiments like coin flips can reveal intricate patterns upon close examination. By fitting various regression models to the data collected in Bartoš et al. 2023, we identified a consistent same-side bias across runs, confirming the findings of the original study through a different methodological lens. Our exploration underscores the importance of balancing model complexity with interpretability (Davison 2024, p. 60, Chapter 1.5). Selecting a suitable regression model heavily depends on the specific goals of the analysis, requiring careful evaluation of the relationship between the response and the explanatory variables. While the simplest binomial GLM captures the overall same-side bias, it does not account for individual-level variability, as evidenced by overdispersion and patterns in the residual plots. Adding fixed and random effects for participants notably improved the model fit and revealed substantial heterogeneity across individuals. The variability due to the coin factor nested within persons is more subtle, and its significance depends on whether we retain certain outliers. In examining time dynamics, we observed a non-linear decrease in bias as the number of flips increases, suggesting a potential learning effect. The coin-flip experiment also shows short-term memory effects when recent flips are taken into account.

Despite these insights, our analysis has its limitations. While we've discussed and generally trust the model's assumptions, some aspects require caution. For example, the assumption of a normal distribution for individual deviations in the random-effect model may not always hold. Aggregating data into blocks of 100 flips in our learning effect analysis might blur finer temporal details. Moreover, we did not account for external factors such as fatigue, environmental conditions, or the specific context of the experiment. Another limitation arises from the experimental design: with a relatively small number of participants and coins, and only a few runs per person or coin despite a high total number of flips, performing robust cross-validation and generalizing the model's findings becomes challenging. However, this scarcity of repeated measures also points to a strength of the fixed-effect model: its modest number of parameters allows it to effectively capture individual-level variation without overfitting.

While our analysis tested different models independently to clearly interpret individual factors and assess their singular significance, exploring a combined model that integrates time effects with person-coin contributions could reveal subtler interactions and suggest that some effects are more nuanced when considered together. Additionally, applying Ridge or Lasso penalties to the fixed-effect model might help shrink non-significant parameters, reduce overfitting, and improve interpretability (Davison 2024, Chapter 3). Regarding the learning effect, using a Markov chain framework could provide new insights into how past outcomes influence future ones; similarly, time series models like ARIMA might capture more complex dynamics. Furthermore, although our spline models for time dependencies include regularization to prevent overfitting, cross-validation could ensure that the fitted splines are not overly tailored to our data. These approaches, although not explored in our current study, represent promising directions for future research.

References

Bartoš, František et al. (2023). "Fair coins tend to land on the same side they started: Evidence from 350,757 flips". In: *arXiv*.

Davison, Anthony Christopher (2003). *Statistical models*. Vol. 11. Cambridge university press.

— (2024). *MATH-408 - Regression Methods*.

Diaconis, Persi, Susan Holmes, and Richard Montgomery (2007). "Dynamical bias in the coin toss". In: *SIAM review* 49.2, pp. 211–235.