Is It Luck or Skill: Establishing Role of Skill in Mutual Fund Management and Fantasy Sports

Vishal Misra[†], **Devavrat Shah**[‡], **and Sudarsan V. S. Ranganathan**[§]

Abstract: The emergence of Online Fantasy Sports Platforms (OFSP) has presented a challenge for the regulatory bodies across the globe: do they represent game of skill or game of chance? A game of skill is a game where the outcome is determined mainly by the predominance of mental or physical skill, rather than chance. Gambling or a game of chance is where outcomes are entirely driven by luck and skill has no role to play. In this work, we present a novel data-driven test that helps address this question. In particular, the failure of the test leads to the conclusion that the outcomes are based on the predominance of skill, and not based on luck.

The proposed test is based on a sound statistical hypothesis of chance that we introduce. It is robust against all sorts of biases that might be present in the data. We apply the test to data obtained from two fantasy sports platforms: Dream11 for Cricket and FanDuel for Basketball. In both cases, we conclude that there is overwhelming evidence that the outcomes are driven by the predominance of skill. Indeed, evidence of "survivor bias" can be found in such dataset suggesting the importance of a robust statistical approach presented in this work. We report similar observations in the context of mutual fund performances suggesting that they are being managed by skilful fund managers.

1. Introduction

A central question surrounding any *game* is whether a *player* in the game can influence the outcomes they experience in the game. In general, the ability of a player to influence the outcomes they experience in comparison to the performance of other players in the game or in comparison to a desired performance level indicates the skill level of a player. For instance, a soccer team that wins more often than other teams is perceived as more skilled than the rest of the teams. In roulette, on the other hand, the player cannot influence the outcome and the outcomes are based on predetermined odds of the game or simply *luck*.

The ability of a player to influence the outcomes in the game is used as the basis by which regulatory bodies across the globe determine whether a game is *gambling* or not. In particular, if the game is simply based on *luck*, then the participation in the game for loss or gain of money (or more generally any item of value) is considered as placing a *bet* or *wager*. Naturally, gambling regulatory considerations are very different from that of skill-based gaming. In many settings, determination of whether an activity is based on skill or simply luck is not so challenging, e.g. Roulette. However, it is not easy in the context of emerging fantasy sports. In particular, the question of luck versus skill has become of the central importance as regulatory bodies across the globe grapple with the explosive growth of *online fantasy sports platforms*.

According to data from the Fantasy Sports Trade Association [12], in 2017 there were an estimated 59.3 million people playing fantasy sports in the USA and Canada and the average player (age 18+) spent 653 USD annually on fantasy sports and related material. As of the time of writing this paper, the Unlawful Internet Gambling Enforcement Act (UIGEA), which regulates online financial transactions associated with betting or wagering in the USA [13], does not apply to fantasy sports. Questions around this exemption

*Author names appear in the alphabetical order of their last names. Their email addresses are: vishal.misra@columbia.edu, devavrat@mit.edu, sudarvsr@mit.edu.

[†]Department of Computer Science, Columbia University

[‡]Department of EECS, Statistics and Data Science, IDSS, MIT

[§]MIT Institute for Data, Systems, and Society

linger and there are legal cases pending decisions in federal and state courts in the USA that pertain to the issues of regulating fantasy sports games.

Likewise, such considerations are becoming of central importance in other countries including India where fantasy cricket has become widely popular [14, 15]. Sports in India is undergoing rapid change and given India's large fan base and economic power the future holds enormous potential for sports-related activities. A report published by KPMG and the Federation of Indian Fantasy Sports (FIFS)[1] estimates that the user base (18+ year olds) of fantasy sports platforms crossed 70 million Indians in 2018. Participants in the nascent sport spent around $ 1.73 billion (₹ 11,880 crore) in 2018.

Multiple High Courts in India and the Supreme Court of India has held Online Fantasy Sports (OFS) as 'Game of Skill' and clarified multiple times that OFS do not fall under the ambit of the Indian Public Gambling Act of 1867. The courts have specifically accorded the protection to the right to free trade and commerce guaranteed under Article 19(1) (g) of the constitution of India. There has been a recent landmark decision by the Supreme Court dismissing multiple SLPs filed against Dream11 on online fantasy sports format.

In summary, given the rapidly growing market of online fantasy sports and regulatory questions surrounding it, it is of urgent importance to develop a scientific and data-driven method that can help answer the question of luck versus skill in a comprehensive manner. As the main contribution of this work, we precisely address this challenge. Before we describe our contributions and their consequences, we review why this question is challenging and has remained unanswered thus far.

The Challenge. The fundamental challenge is the need to define what game of chance or luck means unambiguously. Given the involvement of luck or chance, it needs to be statistical in nature. Now if such a definition were available, the next challenge is to develop an appropriate statistical hypothesis test that verifies whether the outcomes are driven by luck. Such a hypothesis test needs to be based on available data. The available data might involve variety of biases and hence the test needs to be robust against such biases. Despite the importance and the urgent need for such a test, the above challenges have not been addressed thus far.

There have been various attempts made in the literature to address this challenge. We take note of work by Levitt et al. [7] and by Getty et al. [4] in the context of online fantasy sports. In particular, [7] proposes a formal definition and an associated test for what it means for the outcomes to be driven by luck. In particular, they suggest that if answer to all the following four questions is 'no' then the game is a game of chance: (1) Do players have different expected payoffs when playing the game? (2) Do there exist predetermined observable characteristics about a player that help one to predict payoffs across players? (3) Do actions that a player takes in the game have statistically significant impacts on the payoffs that are achieved? and (4) Are player returns correlated over time, implying persistence in skill? The work in [4] builds upon this framework, and in particular utilizes (4) as a guiding principle to propose a 'skill score' for the entire game based on the 'correlations' amongst players' performance over time. An attractive feature of 'skill score' proposed in [4] is the ability to simultaneously compare the role of skill in different types of games, e.g. bicycling seems to be more skill driven than football.

At its core, the approach put forth in [7] and [4] relies on evaluating the player-centric distributional properties of the outcomes. In practice, the data available is likely to be biased. For example, as reported in [4], in fantasy sports the individuals who end up having a losing streak tend to drop out while individuals who tend to win more continue to play. Furthermore, if a player starts the game in a losing streak, then they may not play a sufficient number of times that we need in order to compute the required metrics associated with such players with statistical significance. Losing streaks therefore inherently bias the datasets towards relatively skilled (or lucky) players. Subsequently, due to the biased sampling of data, the player-centric tests of [7] and [4] might suggest the role of skill even if the underlying game might be purely game of chance or luck. In short, we need statistical tests that are robust against the biases in the data generation.

Summary of Contributions. As the main contribution of this work, we address the above stated challenge: we propose a novel, unambiguous definition of luck and provide an associated data-driven hypothesis test that is robust to all sorts of biases in the data. Our definition of luck translates naturally to a formulation of a null hypothesis for our hypothesis test. Let each player of the fantasy sports game have distinct

[1] The Federation of Indian Fantasy Sports (FIFS) is India's first and only Fantasy Sports self-regulatory industry body which presently constitutes 95% of the Online Fantasy Sports market in India.

identification in terms of numbers assigned. This can be simply *hash* of their unique names or pre-assigned by the platform. Now, when two players participate in a competition together, it results in their head-on or pair-wise comparison. If the player with smaller identity wins, then we count it as a *head* else we count it as a *tail*. This mechanism provides an ability to utilize *all* pair-wise comparisons resulting from all competitions by reducing the outcome of each of them to either *head* or *tail*. If our proposed hypothesis of *luck* holds, then *head* and *tail* are equally likely. For *distinct enough* player comparisons, if the hypothesis of luck holds then we effectively obtain a sequence of independent fair coin tosses. This distributional property is not impacted by any bias, variations in the number of contests that a player participates in, etc. That is, this translation of outcome of competitions data into heads/tails through pair-wise comparisons provides a robust statistic under the hypothesis of outcomes being driven by luck. We utilize the pair-wise comparisons obtained from the overall population to compute a test statistic that allows us to reject or fail to reject the hypothesis of luck.

In the setting when data is limited, e.g. a nascent fantasy sports platform, we propose a refined test that attempts to get most information out of the data. This is achieved by developing a randomness extractor from a random permutation in form of pair-wise comparisons, which is a variation of the classical question posed by Von Neumann [11].

We utilize the proposed approach to evaluate the role of luck in two fantasy sports platforms: Dream11 [16] for Cricket, and FanDuel [17] for Basketball. In both cases, we find overwhelming evidence that the outcomes are not driven by pure luck – skill has a role to play. For Mutual Fund performance data [18], using our test we verify that indeed there is role of skill in the returns experienced in the mutual funds.

Organization. Section 2 provides a formal description of the problem setting and our null hypothesis of luck. Section 3 describes a simple test for the null hypothesis of luck and establishes its correctness. Section 4 provides further refinement of the test that extracts more information from given data to obtain a more powerful test. This section also describes the connection to randomness extraction from a random permutation. In Section 5, we put our framework to practice. In particular, we discuss the use of our method to evaluate the role of luck in fantasy cricket using data obtained from Dream11 and in fantasy basketball using data obtained from FanDuel. We also apply our test to the setting of mutual fund performance to verify whether the performance is driven by luck or skill. Section 6 discusses survivor bias in fantasy sports games using a player-centric test statistic. Section 7 presents our conclusions, discussion, and directions for future work. All proofs are relegated to the Appendix (supplementary material).

2. A Statistical Formulation of Luck

Formalism. We formally introduce the statistical definition of luck. We consider a setting where there is a universe of N participants or players with distinct identities. Without loss of generality, we shall enumerate them from 1 to N, i.e. identities of players are denoted as $[N] = \{1, \ldots, N\}$. As discussed earlier, in a game purely based on luck, a participant or player cannot have any influence on the outcomes of the game. In particular, when $m \geq 2$ players make the same "bet", "wager", or participate in the same "competition", "tournament" or "contest", the relative ranking of their outcomes or performance in the contest is completely random. We formalize this notion as follows.

Definition 2.1 (Luck)**.** *Consider a competition induced by a contest between any $m \geq 2$ players with identities $a_1, \ldots, a_m \in [N]$. Let $\sigma : [m] \to [m] \in \mathbb{S}_m$, where $\mathbb{S}_m$ is the set of all possible $m!$ permutations, denote the permutation or ranking of these m players based on the performance in the contest. That is, $\sigma(i) \in [m]$ is the ranking of player a_i amongst m players participating in the competition. Then, in a game of pure luck σ is a random permutation with uniform distribution over any of the $m!$ possible permutations.*

Data. We observe data in the form of ranking of players across competitions. Specifically, we observe outcomes from M competitions. Competition $i \in [M]$ has m_i players participating in it denoted by player identities $a_1^{(i)}, \ldots, a_{m_i}^{(i)}$ where $a_j^{(i)} \in [N]$ for $j \in [m_i]$. We observe the ranking $\sigma^{(i)} : [m_i] \to [m_i] \in \mathbb{S}_{m_i}$ based on player performance for each competition $i \in [M]$. That is, player $a_j^{(i)}$ in competition $i \in [M]$ has rank of $\sigma^{(i)}(j) \in [m_i]$.

The Goal. Given the data as described above, the goal is to verify whether the hypothesis of luck holds or not. Specifically, we wish to develop a statistical test that can be evaluated using the available data. The test should help decide whether the hypothesis of luck be rejected and if so, with what confidence. The outcome of the test ought to be robust with respect to all sorts of biases within the data including those mentioned as a challenge in prior work, cf. [7, 4].

Useful Bits. We propose to utilize the following binary variables from pair-wise comparisons: given a competition $i \in [M]$ with m_i players within it assume without loss of generality that $a_1^{(i)} < a_2^{(i)} < \cdots < a_{m_i}^{(i)}$. For any $\ell, \ell' \in [m_i]$ such that $\ell < \ell'$ define the binary variable $Z_{\ell,\ell'}^{(i)}$

$$Z_{\ell,\ell'}^{(i)} = \begin{cases} 1 & \text{if } \sigma_\ell^{(i)} < \sigma_{\ell'}^{(i)}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that $\ell < \ell'$ means that $a_\ell^{(i)} < a_{\ell'}^{(i)}$ since we assume $a_1^{(i)} < a_2^{(i)} < \cdots < a_{m_i}^{(i)}$. Hence, the binary variable is equal to 1 when the player with the smaller identification number wins and is equal to 0 otherwise.

The following proposition provides the distributional characterization of the binary variables $Z_{\ell,\ell'}^{(i)}$ under the hypothesis of luck.

Proposition 2.1. *Let the hypothesis of luck as defined in Definition 2.1 be satisfied. Then, for any $i \in [M]$ and $\ell, \ell' \in [m_i]$ with $\ell < \ell'$, $Z_{\ell,\ell'}^{(i)}$ is a binary random variable with $\Pr\left(Z_{\ell,\ell'}^{(i)} = 1\right) = \frac{1}{2}$, and $Z_{\ell,\ell'}^{(i)}$ is independent of all $Z_{j,j'}^{(i')}$ for $i' \neq i \in [M]$ with $j < j' \in [m_{i'}]$. Further, for a given $i \in [M]$, $Z_{k,k+1}^{(i)}, 1 \leq k \leq m_i - 1$ are mutually independent.*

3. A Hypothesis Test For Luck

We present a simple statistical hypothesis test with a test statistic that can be directly evaluated from the data. In particular, Proposition 2.1 suggests a natural test statistic. Given ranking of players over M competitions, let $Z_{1,2}^{(i)}, \ldots, Z_{m_i-1,m_i}^{(i)}$ be $m_i - 1$ pair-wise comparisons obtained as per (1) for competition $i \in [M]$. Define

$$S_M \triangleq \sum_{i=1}^{M} \sum_{k=1}^{m_i-1} Z_{k,k+1}^{(i)}.$$

Then, from Proposition 2.1, under the hypothesis of luck it follows that S_M is a Binomial random variable with parameters $n = (\sum_{i=1}^{M} m_i) - M$ and $p = \frac{1}{2}$. We propose test statistic T_M^{naive}, defined as

$$T_M^{\text{naive}}(n) \triangleq \frac{2(S_M - \frac{n}{2})}{\sqrt{n}}. \tag{2}$$

This leads to the following rejection criterion for the hypothesis of luck.

Rejection Criterion for Naive Test Statistic. The hypothesis of luck, as defined in Definition 2.1, is rejected with p-value

$$\Pr\left(\left|B\big(n, 0.5\big) - \frac{n}{2}\right| > \frac{1}{2}\sqrt{n}|T_M^{\text{naive}}(n)|\right),$$

where $B(n, 0.5)$ is a Binomial random variable with parameters $n \geq 1$ and 0.5. The Appendix (supplementary material) presents approaches to evaluate $\Pr\left(\left|B\big(n, \frac{1}{2}\big) - \frac{n}{2}\right| > t\right)$ for Binomial distribution with parameters $n, \frac{1}{2}$ with $t \geq 0$. Namely, we make use of the standard central limit theorem for Bernoulli random variables, Hoeffding's inequality, and Monte-Carlo probability estimation.

Robustness Against Biases in the Data. As noted in Proposition 2.1, the pair-wise comparisons obtained are independent of the players participating in the contest and are independent of the ranking of

players in other contests under the null hypothesis of luck. Hence, the test statistic (4) is not affected by any biases in the underlying dataset. If the null hypothesis was true, the test statistic will capture this (given the dataset is large enough) since it is computed at the *population level* and is different from *player-centric metrics*, which can be affected by biases in the data, as considered in the prior literature.

4. A Refined Hypothesis Test

Under the hypothesis of luck, the method discussed in Section 3 extracts independent and identically distributed pair-wise comparisons from a given dataset. We evaluate a pre-determined $m_i - 1$ number of comparisons, $Z^{(i)}_{k,k+1}$ for $1 \le k \le m_i - 1$, for every competition $i \in [M]$ in order to compute the test statistic.

Under the hypothesis of luck, the ranking $\sigma^{(i)}$ of players in a competition i is any of $m_i!$ possibilities equally likely. Information theoretically speaking, cf. [1], uniformly random ranking or permutation over m_i elements contains $\log m_i! = \Theta(m_i \log m_i)$ bits of randomness under the null hypothesis. The mutually independent pair-wise comparisons extracted out of a random permutation effectively contain 1 bit of information each. Therefore, from an information theoretic perspective, given a random permutation or ranking, one ought to be able to extract more than $m_i - 1$ independent comparisons and, ideally, up to $\Theta(m_i \log m_i)$ of them. Formally, we are asking the following question:

Question. How many bits of randomness in the form of pair-wise comparisons can be extracted from a random permutation?

This question falls within the realm of *randomness extraction*, whose origins go back to Von Neumann [11] and with a rich literature in the sub communities of cryptography cf. see [10], information theory cf. see [20], and probability cf. see [8]. Despite this rich literature, the question of interest is not considered in the literature. Turns out, it has a simple solution which we present next. It will enable a stronger hypothesis test.

A Variation Of Quicksort For Randomness Extraction. We now present a randomness extractor from a random permutation. To do so, we utilize a variation of the classical quicksort, which is a pair-wise comparison sorting algorithm [9]. We are given a permutation $\sigma : [m] \to [m] \in \mathbb{S}_m$ that is drawn from the uniform distribution over $\mathbb{S}_m$. We wish to extract as many mutually independent pair-wise comparisons from σ as possible. We present the following adaptive algorithm.

Recursive Algorithm for Randomness Extraction from a Random Permutation. First, choose a *pivot* element $s \in [m]$ uniformly at random. Evaluate $m-1$ comparisons between $\sigma(s)$ and $\sigma(k), k \neq s$. Second, create two subsets of $[m]$ excluding s as $C^{<}(s) = \{k \in [m] : k \neq s, \sigma(k) < \sigma(s)\}$, $C^{>}(s) = \{k \in [m] : k \neq s, \sigma(k) > \sigma(s)\}$. Recursively apply the above two steps to sets $C^{<}(s)$ and $C^{>}(s)$ with restriction of σ to them as long as they are of size at least 2 and stop otherwise. In this process, the comparisons Z_{ℓ_k,ℓ'_k} produced according to (1) using the sequentially chosen index pairs $\ell_k, \ell'_k \in [m]$ such that $\ell_k < \ell'_k$ and their corresponding $\sigma(\ell_k)$ and $\sigma(\ell'_k)$ are the output of the algorithm.

For such an adaptive method, we characterize the average number of comparisons extracted.

Proposition 4.1. *Given a random permutation σ drawn as per uniform distribution over $\mathbb{S}_m$, let there be $m-1 \le C(\sigma) \le \binom{m}{2}$ comparisons produced by the algorithm sequentially denoted as $c_1, \ldots, c_{C(\sigma)}$. Then, they are mutually independent and equally likely to take either of the two possible values. Further,*

$$\mathbb{E}[C(\sigma)] = 2(m+1)\sum_{k=1}^{m}\frac{1}{k} - 4m. \tag{3}$$

Revised Test Statistic. The comparisons produced across competitions using the above adaptive algorithm are independent under the hypothesis of luck. However, the comparisons produced from the same competition are generated adaptively and the number of these comparisons is random. Therefore, the test statistic needs to be carefully designed.

Given ranking $\sigma^{(i)}, i \in [M]$ of players over M competitions, let $C(i) \triangleq C(\sigma^{(i)})$ be the random number of

pair-wise comparisons $Z^{(i)}_{\ell_k,\ell'_k}$ with $\ell_k, \ell'_k \in [m_i]$ and $\ell_k < \ell'_k$ for $k = 1, 2, \ldots, C(i)$. Define, $\widetilde{S}_M$ as

$$\widetilde{S}_M = \sum_{i=1}^{M} \sum_{k=1}^{C(i)} \left(Z^{(i)}_{\ell_k,\ell'_k} - \frac{1}{2} \right).$$

Let

$$V_M = \sum_{i=1}^{M} \mathcal{QS}(m_i)$$

where $\mathcal{QS}(m) = 2(m+1)\sum_{k=1}^{m} \frac{1}{k} - 4m$ is the average number of comparisons extracted by our adaptive scheme as established in Proposition 4.1 for a competition with m players. We define a refined test statistic T^{refine}_M as

$$T^{\text{refine}}_M \triangleq \frac{2 \cdot \widetilde{S}_M}{\sqrt{V_M}}. \tag{4}$$

We state the following result.

Theorem 4.1. *Let the hypothesis of luck, as defined in Definition 2.1, be satisfied. Given the universe with a fixed number of players N, as the number of competitions $M \to \infty$ the refined test statistic T^{refine}_M converges in distribution to $\mathcal{N}(0,1)$.*

Given Theorem 4.1, we propose to utilize the implied Gaussian approximation to compute the confidence as follows.

Refined Rejection Criteria. The hypothesis of luck, as defined in Definition 2.1 is rejected with p-value $\alpha \in [0,1]$ if $\Phi(|T^{\text{refine}}_M|) = 1 - \frac{\alpha}{2}$. Here $\Phi(t) = \Pr(\mathcal{N}(0,1) \leq t)$.

5. Data Analysis and Results

In this section, we put our statistical hypothesis test to practice. We use our framework to verify whether the outcomes of the fantasy sports, Cricket and Basketball, as well as stock market through mutual funds are driven by luck or there is a role of skill. We note that all the experiments reported here are using the naive statistic that is described in Section 3 as we have sufficient data and do not need the further refined test. We refer the reader to Appendix (supplementary material) for the methods used to compute the p-values in this section.

5.1. Fantasy Cricket

Data. *Dream11* [16] is a fantasy sports platform with Cricket being an extremely popular fantasy sport amongst the players on the platform. A player of the fantasy sport participates in a competition or a contest by entering one or more fantasy teams in a round. Each round is associated with an actual real-life cricket game and players on the platform are allowed to submit multiple teams (up to 6 different) to any number of contests that are open within the round. A contest could be as small as a head-to-head (2-player) or as large as thousands of players. The team that a player enters in a contest receives points or a score based on the outcome of the associated physical game. This scoring leads to a ranking of players in each contest.

We obtained a dataset[2] containing all such rounds, the players (in an anonymized manner) who participated, and their performance in each contest that they participated across all the rounds. The dataset contains the data of all rounds within the four calendar years from 2013 to 2016.

[2]The dataset was obtained from Dream11 for research purposes. Please contact Dream11 at policy@dream11.com to request data for research purposes.

TABLE 1
Pair-wise comparison statistics.

Year	Comparison = 1	Total Comparisons
2013	21,989	43,388
2014	126,003	247,536
2015	838,554	1,607,414
2016	7,881,168	14,836,570
Full dataset	8,864,463	16,731,783

TABLE 2
p-values obtained for the Dream11 Fantasy Cricket dataset. As can be seen, the outcomes do not resemble pure luck and clearly suggest that there is role of skill. Note: A p-value of 0 reflects that within the precision of floating point on computer, it is less than smallest possible positive value.

Year	p-value for Hypothesis of Luck
2013	0.0046
2014	0
2015	0
2016	0
Full dataset	0

Since players are allowed to submit the same team to different contests in a round there is an inherent dependency among contests in a round. Therefore, we ranked the players at the round level by considering the best-scoring team that a player submitted across all contests in a round. In this fashion, the full dataset contains a total of 3700 rounds with an average of 4400 players in each round. Approximately half of these rounds were in the calendar year 2016, a third of the rounds were in the year 2015 and the 2013 and 2014 datasets are a small fraction of the overall dataset. On obtaining the round-level ranking of players, each round is now represented by an array of ranks and an array of the associated player IDs that we then use directly to obtain pair-wise comparison random variables. In the following, we may refer to a round of a Dream11 as a contest or a competition for simplicity.

Results. Table 1 provides the pair-wise comparison statistics obtained from the data following the setup described in Section 2. We describe how many total comparisons were extracted from the data using method of Section 3 and how many of it were equal to 1. Using the data presented in Table 1, we can obtain the naive test statistic as described in Section 3 and evaluate the p-value for rejecting the luck hypothesis as described using methods in Appendix (supplementary material); we present these p-values in Table 2. Since the dataset for the year 2013 is small, we utilize Monte-Carlo estimation to obtain an accurate p-value. For years 2014-2016, the dataset is larger and hence we utilize Hoeffding's inequality to compute the upper bound as reported. As can be seen, across years and for the entire dataset, the hypothesis of luck is overwhelmingly rejected. In particular, the data presents strong evidence that the outcomes of the fantasy cricket game administered by Dream11 are not driven by pure luck and skill has role to play.

Further Verification via Bootstrap and Comparison To Pure Luck. To evaluate the distribution of induced p-values for the test statistic, we utilize the standard approach of bootstrap, cf. [3]. The natural comparison is with the p-value distribution under the hypothesis of luck which is, by definition, uniformly distributed over $[0, 1]$ (the logarithm of such p-values is distributed as the negative of an exponential random variable with parameter 1).

Given the large size of dataset, we simulate bootstrap by selecting a random sample of size 80%. In particular, we sample each competition with probability 0.8 independently. From this sub-sampled dataset, we extract pair-wise comparisons, produce test statistics and compute the associated p-values using the central limit theorem. We plot the computed p-values across a number of such trials. As a representative example, we present the resulting histogram for data from year 2014 in Figure 1. Notice that the plot is in log-scale along the x-axis to allow for a finer view of the p-value distribution.

As seen from the histogram, the distribution of the logarithm of the p-values is far from the distribution of the negative of an exponential random variable. This further confirms our summary conclusion: the dataset

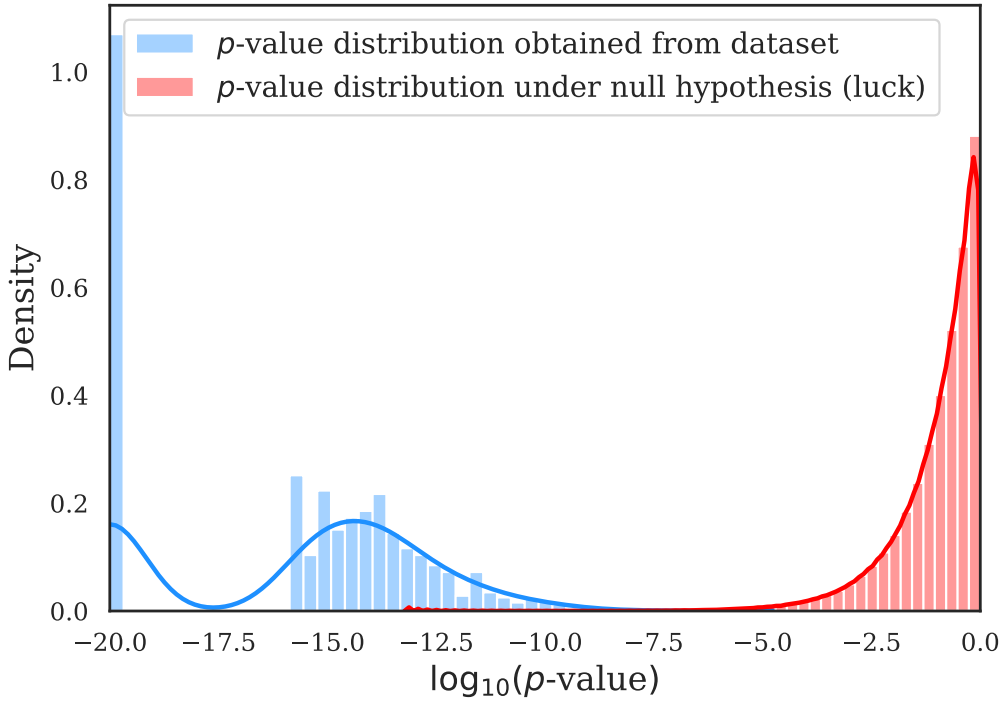

Fig 1: p-value histogram using bootstrap method over Dream11 dataset for year 2014.

of Dream11 players *rejects the hypothesis of pure luck* and *skill ought be involved in the outcomes.*

5.2. Fantasy Basketball

Data. *FanDuel* [17] is a fantasy sports platform with an active fantasy Basketball player base. The platform administers, among other major sports fantasy games, a fantasy game based on the actual National Basketball Association (NBA) games. We obtained a subset of the dataset utilized in the earlier research study [4]. The dataset contains fantasy basketball competitions held during the 2014-2015 NBA Basketball season. These competitions are called head-to-head contests. Specifically, in each contest a fantasy sports player enters a team and the team receives scores or points based on the performance of the actual game, just like in the setting of Dream11 [16] dataset. Each contest corresponds to a game-week and the teams submitted by players are scored based on the performance of the athletes in that game-week. The data is anonymized but each fantasy player has a consistent player identifier. For each competition, we obtain the score of a given player's entry within the head-to-head competition. This allowed us to create a ranking of players within a competition. Using this ranking we extract pair-wise comparisons, produce test statistics and evaluate the p-value for the null hypothesis of luck. The contest sizes are moderate in that each contest contained anywhere between 112 and around 7200 players.

Results. The dataset contained 397 competitions. Players that participated in a contest to match-up against players in the same contest in a head-to-head fashion were allowed to participate in more than one head-to-head match-up in a given contest. The dataset that we obtained does not contain the information of the list of opponents that each player participated against in head-to-head match-ups in a given contest. Therefore, we used the best possible score of all the teams submitted by each player in the contest as a proxy for the rank ordering of the players in the contest. Using this ranking we extract pair-wise comparisons, produce test statistics, and evaluate the p-value for the null hypothesis of luck. The resulting test statistic is

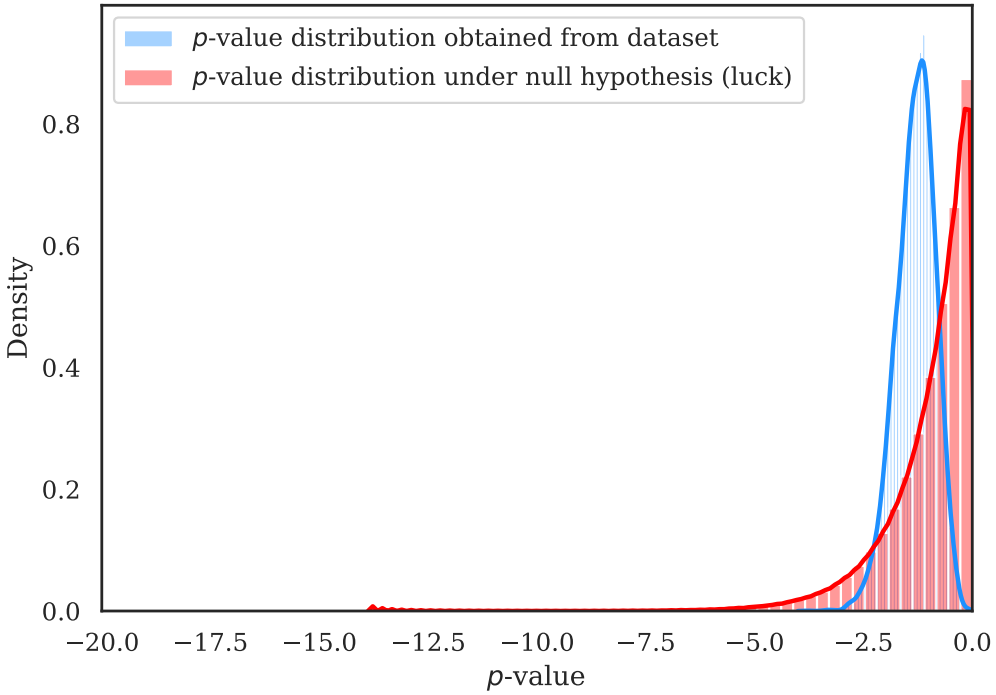

Fig 2: p-value histogram for the Fanduel basketball fantasy dataset.

based on a total of 810, 776 comparisons of which 404, 389 take the value 1. We find using the test statistic from Section 3 that the hypothesis of luck is rejected with a p-value of 0.0273. We computed this p-value using Monte-Carlo simulation because of the smaller data size. We note that the central limit theorem approximation resulted in a p-value of 0.0264, which is very close to the numerically simulated p-value.

We also obtain the p-value histogram by performing bootstrap sampling using 80% sampling rate. Figure 2 shows the p-value histogram thus obtained by using p-values obtained using the central limit theorem. The p-value histogram, coupled with the p-value of the overall dataset, clearly leads to the conclusion that *the hypothesis of pure luck should be rejected* and *some level of skill ought be involved in the outcomes* of FanDuel's fantasy basketball.

5.3. Mutual Funds

Data. The Center for Research in Security Prices (CRSP) Survivor-Bias Free US Mutual Fund Database [19] contains mutual fund performance data over the past few decades. We obtained the monthly performance data from this database for years 2005 through 2018 using Wharton Research Data Services (WRDS) [18]. We considered monthly returns of the mutual funds for this period and therefore we analyzed 168 virtual contests where a contest consists of the monthly return of the mutual funds over a single calendar month. The dataset contains performance of 55,789 mutual funds.

Results. We now turn our attention to the question of luck vs. skill in managing a mutual fund. It is worth noting that the size of each contest in this dataset is very large since all active mutual funds in a given month are participants in a contest. Specifically, the smallest contest size is 19, 345 and the largest contest size is 32, 671. As before, following our framework of ranking, we extract pair-wise comparison statistics and evaluate p-values. We obtain a total of 4, 574, 144 comparisons of which 2, 282, 693 have the value 1. Using the approach suggested in Section 3, we find that the hypothesis of pure luck driving performance of mutual

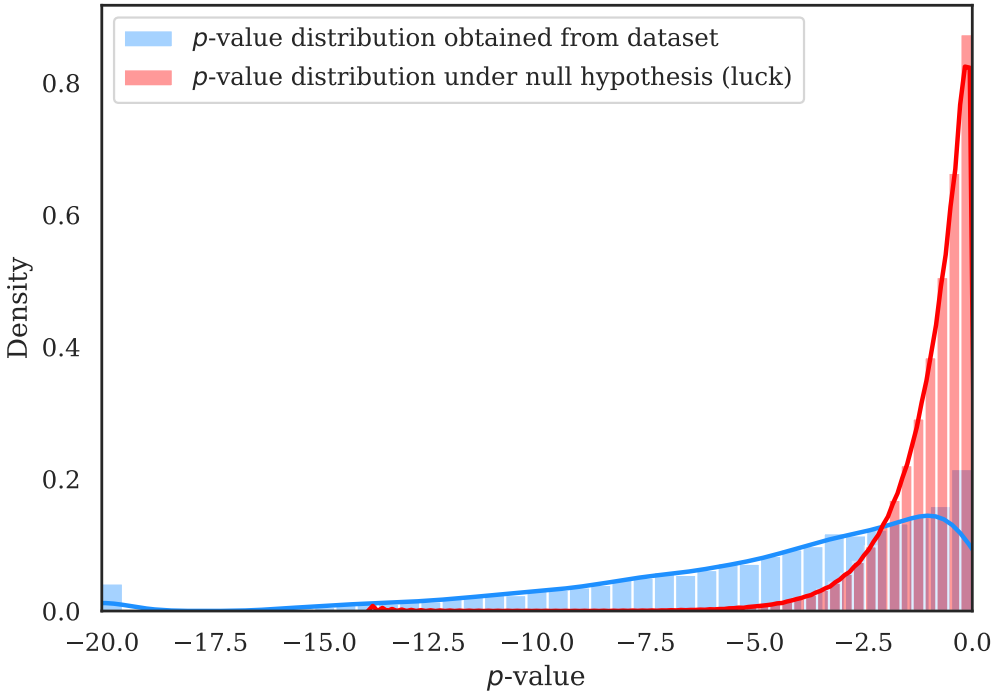

Fig 3: p-value histogram for the mutual fund dataset.

funds can be rejected with a p-value of 4×10^{-5}. This p-value was obtained using the central limit theorem; Hoeffding's inequality resulted in an upper bound of 4×10^{-4}.

Using bootstrap sampling with sampling rate of 80%, we find the p-value histogram (obtained using p-values computed via the central limit theorem) that is shown in Figure 3. From the overall p-value and p-value histogram, we can conclude that for the overall mutual fund performance *the hypothesis of pure luck should be rejected* and *some level of skill ought be involved in their performances.*

6. Survivor Bias

The discussions thus far utilize our approach for aggregating pair-wise comparison statistics across all players and competitions. As discussed, our hypothesis test is not impacted by behavioral biases such as the *survivor bias* where winners end up playing more compared to losers. We utilize the Dream11 dataset to inspect empirically whether such a phenomenon exists in reality or not.

To that end, we shall utilize a player-centric approach. In particular, for any given player $i \in [N]$, we count the number of head-to-head or pair-wise comparisons with other players across all contests in which player i has participated. Specifically, consider the subset of all contests $M_i \subset [M]$ in which player i has participated. For any given contest $j \in M_i$, and for any player $i' \neq i \in [N]$ that has participated in the contest j, we assign score $1, \frac{1}{2}$, or 0 to player i if they outperform player i', player i ties in performance with i' or player i' outperforms player i, respectively.

For player i, we compute the normalized score in $[0, 1]$ for each contest $j \in M_i$ by adding such pair-wise scores in the contest and dividing by the number of other players in that contest. Let k_j be the total number of players in contest $j \in M_i$ (including player i). Then, the normalized score of player i in contest j (assuming ties are resolved at random), denoted W_j, is uniformly distributed with possible values $0, \frac{1}{k_j-1}, \ldots, \frac{k_j-2}{k_j-1}, 1$

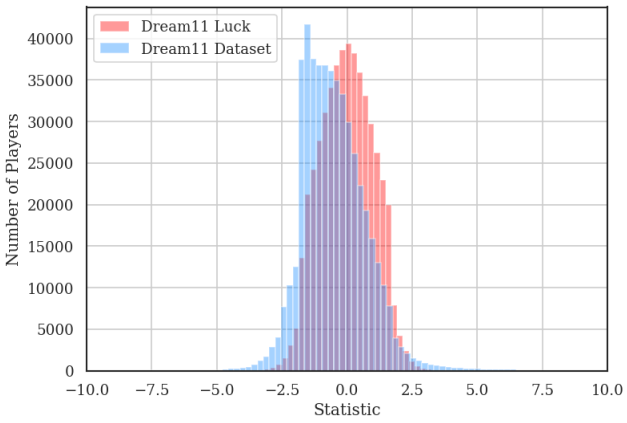

Fig 4: Distribution of statistic T_i of 500,000 random players in the Dream11 dataset and the emulated luck dataset.

under the null hypothesis of luck. The expected value of W_j under the null hypothesis of luck is

$$\begin{aligned}\mathbb{E}[W_j] &= \sum_{i=0}^{k_j-1} \frac{i}{k_j-1} \cdot \frac{1}{k_j} \\ &= \frac{1}{k_j(k_j-1)} \sum_{i=0}^{k_j-1} i = \frac{1}{2}.\end{aligned}$$

It can also be computed that the variance of W_j is

$$\mathrm{Var}(W_j) = \frac{1}{12} \cdot \frac{k_j+1}{k_j-1}.$$

Note that $\mathrm{Var}(W_j) \approxeq \frac{1}{12}$ when k_j is very large, which is the case in the Dream11 dataset. Hence W_j behaves like a continuous uniform random variable when the contest sizes are large. By the central limit theorem, under the null hypothesis the random variable

$$T_i \triangleq \frac{\sum_{j=1}^{|M_i|} W_j - \frac{|M_i|}{2}}{\sqrt{\sum_{j=1}^{|M_i|} \mathrm{Var}(W_j)}}$$

tends to a Gaussian with mean 0 and variance 1.

With this player-centric approach, we sample 500,000 players at random from the Dream11 dataset and compute their statistics T_i as defined. The empirical distribution of the statistics of this random set of 500,000 players for the actual data and the corresponding luck dataset (obtained by sampling a random ranking array for each round) are plotted in Figure 4 – the two distributions show a significant difference. The empirical dataset for Dream11 has a significantly higher number of players with negative scores when compared to pure luck. We can also see that the right tail for the Dream11 dataset is a lot more spread out as compared to the luck dataset. This is significant as the right tail, i.e. players that perform better than predicted under the luck hypothesis, is evidence that skill plays a role in the performance of players. As discussed earlier, this could simply be due to the bias in data such as the survivor bias.

To that end, Figure 5 explores survivor bias in the Dream11 dataset. It is conceivable that in fantasy sports players might have "equal skill", and a few players simply get unlucky, lose a few contests and then leave, leaving the system biased in favor of players who got "lucky" and won. However, if the hypothesis of pure luck is true then after a sufficient number of contests that "luck" should revert. In Figure 5 we present a scatter plot of the statistic T_i plotted against the number of contests played for a random sample of 10,000

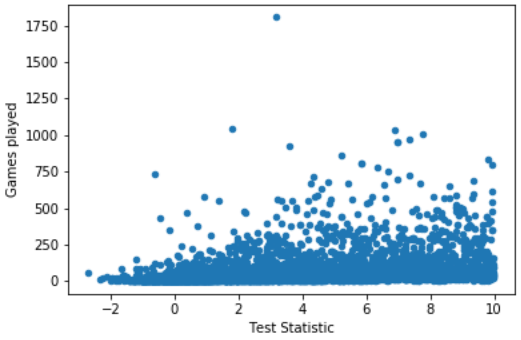

Fig 5: Scatter plot of statistic T_i vs number of rounds played.

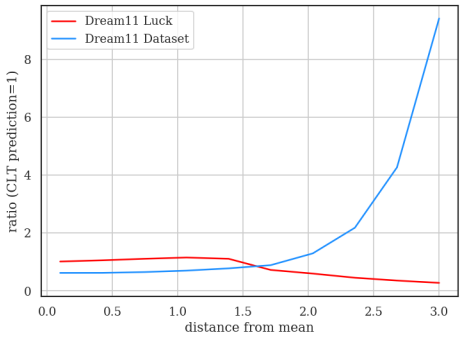

Fig 6: Predominance of Skill.

players. We can see clearly an increasing trend, i.e. players who win a lot play a lot and *keep* winning, indicating that skill plays a role in the outcomes. Therefore, while the dataset does have a survivor bias, it reinforces the presence of skill as skillful players keep playing and winning and do not revert to the scenario predicted by a game of "luck".

6.1. Predominance of Skill

The player-centric view helps empirically confirm survivor bias as discussed. However, it also appears to suggest existence of players with superior performance or "predominant skill". To that end, we define the "predominant skill" of the games as follows.

Definition 6.1 (Predominant Skill)**.** *A game is said to have pre-dominant skill at least* s *at level* x *if the fraction of empirical statistic* $T_i > x$ *is at least* $s \times (1 - \phi(x))$ *where* $\phi(x)$ *is the CDF of the standard normal distribution.*

In other words, the predominance of skill represents how "heavy" the tail of the empirical distribution of player scores is when compared to that predicted by the standard normal distribution. A predominant skill score of 1 at a level $x > 0$ represents the pure luck case, and a score higher than 1 indicates a game of skill.

As we can see in Figure 6, the Dream11 data *strongly* indicates a predominance of skill. While the luck data indicates a score of 1 around the mean, it starts going down as we go further into the tail because of a limited number of samples (for any finite dataset the support of the distribution of the statistic will be bounded). However, for the Dream11 data we see the predominant skill go *up* as we go further out in the tail. For instance, at 3 standard deviations away from the mean we find 10 times as many players with the

statistic $T_i > 3$ as compared to that predicted by CLT, and that is without even accounting for the finite sample effect.

7. Conclusion

In this work, we developed a robust statistical framework for evaluating hypothesis of luck in a data-driven manner. Our work takes inspiration from prior works [7] and [4], but overcomes their limitations by presenting a statistical test that is robust to variety of biases induced in data as well as provides test statistics that can be evaluated from data with provable performance guarantees.

We applied our test to the setting of fantasy sports and stock market. In particular, using data from Dream11, a fantasy cricket platform, we find that the test fails (p-value of 0 upto floating point precision) across years, suggesting that skill has a role to play in the fantasy cricket sports on Dream11 platform. A similar conclusion is found (p-value of 0.0264) using data from FanDuel, a fantasy Basketball platform. Using data about various mutual funds (2005 to 2018), we find that skill has role to play (p-value of 4×10^{-5}) in managing mutual funds. A further analysis of the Dream11 data where we test individual player performance reveals a predominance of skill in that dataset, where there is a direct and positive correlation between player performance and the number of contests played.

We strongly believe that our work should provide a robust, data-driven tool for regulatory bodies to decide that fantasy sports is based on predominance of skill and not based on pure luck. In particular, our refined test statistics are aimed at getting as much information as possible from given observations, and therefore, of particular use when the regulatory body needs to evaluate an emerging format to quickly decide whether to regulate it as gambling or game of skill.

Furthermore,in our analysis of the data, it can be concluded that a users making teams on a fantasy sports platform like Dream11, demonstrate a *higher* range of skill than what is required by a mutual fund manager to manage a mutual fund portfolio.

Acknowledgements

Devavrat Shah would like to thank Peko Hosoi for inspiring this work, numerous fruitful conversations as well as help in obtaining the FanDuel Fantasy dataset. Vishal Misra and Devavrat Shah would like to thank Dream11, Inc. for making the Fantasy dataset available for this work as well as to the broader research community. We would like to acknowledge help of Michael Fleder in obtaining Mutual Fund dataset. We thank Anuran Mankur for carefully reading the earlier draft and providing feedback to improve readability. This work was supported in parts by NSF TRIPODS Phase I project, NSF CNS project, NSF CMMI-1462158 project, NSF CMMI-1634259 project and IDSS MicroMasters Post-doctoral Fellowship. Vishal Misra and Devavrat Shah are technical advisors to Dream11, Inc. since January 2019.

References

[1] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[2] Durrett, R. (2019). *Probability: theory and examples* **49**. Cambridge university press.

[3] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[4] Getty, D., Li, H., Yano, M., Gao, C. and Hosoi, A. (2018). Luck and the Law: Quantifying Chance in Fantasy Sports and Other Contests. *SIAM Review* **60** 869-887.

[5] Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

[6] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13-30.

[7] Levitt, S. D., Miles, T. J. and Rosenfield, A. M. (2012). Is Texas Hold'Em a Game of Chance – A Legal and Economic Analysis. *Geo. LJ* **101** 581.

[8] Peres, Y. (1992). Iterating von Neumann's procedure for extracting random bits. *The Annals of Statistics* 590–597.

[9] Sedgewick, R. and Wayne, K. (2011). *Algorithms, Fourth Edition*. Addison-Wesley.

[10] Shaltiel, R. (2011). An introduction to randomness extractors. In *International colloquium on automata, languages, and programming* 21–41. Springer.
[11] Von Neumann, J. (1951). Various techniques used in connection with random digits. *Appl. Math Ser* **12** 5.
[12] Fantasy Sports Trade Association. A Middleton, Wisconsin-based trade group representing the fantasy sports and gaming industries.
[13] Unlawful Internet Gambling Enforcement Act of 2006. The Unlawful Internet Gambling Enforcement Act of 2006 is United States legislation regulating online gambling.
[14] K.R. Lakshmanan v. State of Tamil Nadu AIR 1996 SC 1153. A game of skill, on the other hand although the element of chance necessarily cannot be entirely eliminated, is one in which success depends principally upon the superior knowledge, training, attention, experience and adroitness of the player.
[15] Fantasy Cricket Legality in India. Fantasy cricket is classified as a "game of skill", similar to fantasy sports in the United States. Fantasy Cricket for Cash is at the hub of three dynamic industry spokes ? Internet, gaming and cricket.
[16] Dream11, Inc. A fantasy sports platform that allowed millions of cricket fans in the country to finally have their own teams, albeit virtual!
[17] FanDuel, Inc. A daily fantasy sports provider from the United States and bookmaker based in New York City.
[18] Wharton Research Data Services. WRDS provides the leading business intelligence, data analytics, and research platform to global institutions.
[19] CRSP Survivor-Bias-Free US Mutual Funds. The CRSP Survivor-Bias-Free US Mutual Fund Database serves as a foundation for research and benchmarking for this asset class.
[20] Zhou, H. and Bruck, J. (2011). Efficient generation of random bits from finite state Markov chains. *IEEE Transactions on Information Theory* **58** 2490–2506.

Appendix

Appendix B: Proofs

Proof of Proposition 2.1. Under the hypothesis of luck as defined in Definition 2.1, for any $i \in [M]$, $\sigma^{(i)}$ is equally likely to be any of the $m_i!$ permutations. Therefore, when restricted to any $\ell \neq \ell' \in [m_i]$ with $\ell < \ell'$, $\sigma^{(i)}(\ell) < \sigma^{(i)}(\ell')$ with probability $\frac{1}{2}$. Therefore, it follows that $Z^{(i)}_{\ell,\ell'}$ is a binary random variable with $\Pr\left(Z^{(i)}_{\ell,\ell'} = 1\right) = \frac{1}{2}$.

Under the hypothesis of luck per Definition 2.1, the outcomes of two different competitions are independent of each other. Therefore, it follows that $Z^{(i)}_{\ell,\ell'}$ is independent of all $Z^{(i')}_{j,j'}$ for $i' \neq i \in [M]$.

Finally, when restricted to a specific competition $i \in [M]$, $Z^{(i)}_{1,2}$ is a Binary random variable with $\Pr\left(Z^{(i)}_{1,2} = 1\right) = \frac{1}{2}$. Now given $Z^{(i)}_{1,2}$, we know the relative order of $\sigma^{(i)}_1$ and $\sigma^{(i)}_2$. However, given this information, the relative order of $\sigma^{(i)}_3$ with respect to $\sigma^{(i)}_2$ is equally likely to be smaller or larger than it due to $\sigma^{(i)}$ being permutation that is distributed uniformly at random over all possible $m_i!$ possibilities. That is, given $Z^{(i)}_{1,2}$, $Z^{(i)}_{2,3}$ is still a binary random variable with $\Pr\left(Z^{(i)}_{2,3} = 1 \mid Z^{(i)}_{1,2} = 0\right) = \Pr\left(Z^{(i)}_{2,3} = 1 \mid Z^{(i)}_{1,2} = 1\right) = \frac{1}{2}$. In a similar manner, we can argue that $Z^{(i)}_{k,k+1}$ is uniform binary variable independent of $(Z^{(i)}_{1,2}, \ldots, Z^{(i)}_{k-1,k})$ for all $2 \leq k \leq m_i - 1$. This completes the proof of Proposition 2.1.

Proof of Proposition 4.1. To start with, we shall establish the distributional properties of the comparisons. Consider the comparisons produced in the first phase: we choose a "pivot" element $s \in [m]$ uniformly at random and then compare $\sigma(k)$ and $\sigma(s)$ for $k = 1, \ldots, m$ with $k \neq s$. Since σ is completely random, clearly $\sigma(1) < \sigma(s)$ is equally likely to be true or false. Given the outcome of comparison $\sigma(1) < \sigma(s)$, not the actual values $\sigma(1)$ and $\sigma(s)$, $\sigma(2)$ is equally likely to be smaller or large compared to $\sigma(s)$. And this continues for all $k = 2, \ldots, m$ with $k \neq s$. Next we produce comparisons between elements within $C^{<}(s)$ and between elements within $C^{>}(s)$, where both the subsets do not contain s. Further all elements in $C^{<}(s)$ (respectively

$C^{>}(s)$) have smaller (respectively larger) rank than $\sigma(s)$. But the ranking between the elements of $C^{<}(s)$ (respectively $C^{>}(s)$) is equally likely to be any of the $|C^{<}(s)|!$ (respectively $|C^{>}(s)|!$) feasible possibilities. Now, recursively applying the same argument, we conclude that the comparisons produced sequentially are independent of the information revealed earlier and they are likely to take either of the two values 0 or 1 with equal probability. This completes the proof of the sequential distributional independence property of the comparisons produced as desired.

The random number of comparisons, $C(\sigma)$ depends on the randomness in σ and choice of pivots in the algorithm. To analyze $\mathbb{E}[C(\sigma)]$, we shall utilize a recursive argument. Let $f(m) \triangleq \mathbb{E}[C(\sigma)]$ where σ is a random variable with uniform distribution over $\mathbb{S}_m$. As discussed, the first step of algorithm is to choose a pivot $s \in [m]$ uniformly at random. This means that the pivot s is such that $\sigma(s)$ is distributed uniformly over $[m]$. Therefore, after the decomposition, we have that $|C^{<}(s)|$ is uniformly distributed over $\{0, \ldots, m-1\}$. By symmetry, the same applies for $|C^{>}(s)|$. Therefore, using the sequential nature of the process and law of total expectation,

$$f(m) = m - 1 + \frac{2}{m} \sum_{k=0}^{m-1} f(k).$$

That is,

$$mf(m) = m(m-1) + 2 \sum_{k=0}^{m-1} f(k).$$

Equivalently,

$$(m-1)f(m-1) = (m-1)(m-2) + 2 \sum_{k=0}^{m-2} f(k).$$

Continuing,

$$mf(m) - (m-1)f(m-1) = 2(m-1) + 2f(m-1).$$

The above can be written equivalently as

$$mf(m) = (m+1)f(m-1) + 2m - 2.$$

Using $2m - 2 = 4m - 2(m+1)$, we have

$$\frac{f(m)}{m+1} = \frac{f(m-1)}{m} + \frac{4}{m+1} - \frac{2}{m}.$$

Define $g(m) \triangleq f(m)/(m+1)$. Then,

$$g(m) = g(m-1) + \frac{4}{m+1} - \frac{2}{m}.$$

Recursing this relation, we have

$$\begin{aligned} g(m) &= g(m-1) + \frac{4}{m+1} - \frac{2}{m} \\ &= g(m-2) + \frac{4}{m+1} + \frac{4}{m} - \frac{2}{m} - \frac{2}{m-1} \\ &= g(2) + 4 \sum_{k=4}^{m+1} \frac{1}{k} - 2 \sum_{k=3}^{m} \frac{1}{k} \\ &= 2 \sum_{k=4}^{m+1} \frac{1}{k} + \frac{2}{m+1} - \frac{1}{3} \end{aligned}$$

$$= 2\sum_{k=1}^{m} \frac{1}{k} + \frac{4}{m+1} - 4,$$

since $g(2) = \frac{1}{3}$. From this, $f(m) = g(m)(m+1) = 2(m+1)\sum_{k=1}^{m} \frac{1}{k} - 4m$. This completes the proof of Proposition 4.1.

Proof of Theorem 4.1. We establish the proof using Martingale Central Limit Theorem. To that end, the competitions are ordered in a sequence in an arbitrary manner (or the order in which they are processed). For each competition, we apply the adaptive comparison extraction algorithm as described. Let $Z_k \in \{0, 1\}, k \geq 1$ be the outcome of the k^{th} comparison and define the centered comparison variable $Q_k = Z_k - \frac{1}{2} \in \{-\frac{1}{2}, \frac{1}{2}\}$. Let $\mathcal{F}_k$ denote the smallest sigma algebra containing all the information pertinent till and including Q_k. Then, we get from Proposition 4.1 that

$$\mathbb{E}[Q_{k+1}|\mathcal{F}_k] = 0.$$

Define $\bar{Q}_k = \sum_{t=1}^{k} Q_t$. Then it can be checked that it is a Martingale with respect $\mathcal{F}_k,\ k \geq 0$. Let $R \geq 1$ to be the time when we stop extracting further comparisons using our algorithm from the data. It can be seen that $\{R = t\} \subset \mathcal{F}_t$ because once we have processed a particular pair-wise comparison we know we need to stop or continue extract more comparisons based on the outcomes till then. That is, R is a stopping time with respect to $\mathcal{F}_k,\ k \geq 1$. Further, for any given number of competitions M, the number of comparisons is no more than MN^2. Therefore, $R \leq MN^2$, which implies that $R < \infty$ with probability 1. Therefore, by Doob's optional stopping theorem [2], it follows that under the hypothesis of luck,

$$\mathbb{E}[\bar{Q}_R] = \mathbb{E}[\bar{Q}_1] = \mathbb{E}[Q_1] = 0.$$

Next, we compute the variance of $\bar{Q}_R$. In order for this, define $U_k \triangleq \bar{Q}_k^2 - \frac{k}{4}$. Then,

$$\begin{aligned}
\mathbb{E}[U_{k+1}|\mathcal{F}_k] &= \mathbb{E}\left[(\bar{Q}_k + Q_{k+1})^2 - \frac{k+1}{4}\Big|\mathcal{F}_k\right] \\
&= \bar{Q}_k^2 + \mathbb{E}[Q_{k+1}^2|\mathcal{F}_k] - \frac{k+1}{4} \\
&= \bar{Q}_k^2 - \frac{k}{4} \ = \ U_k,
\end{aligned}$$

where we used the fact that Q_{k+1} is uniform over $\{-\frac{1}{2}, \frac{1}{2}\}$ and independent of $\mathcal{F}_k$. This means that U_k is also a Martingale. Again, by optional stopping theorem, it follows that

$$\mathbb{E}[U_R] = \mathbb{E}[U_1] \ = \ 0 \ = \ \mathbb{E}[\bar{Q}_R^2] - \frac{\mathbb{E}[R]}{4}.$$

Using Proposition 4.1 and in particular (3), it follows that

$$\mathbb{E}[R] = \sum_{i=1}^{M} \mathcal{QS}(m_i),$$

where $\mathcal{QS}(m) = 2(m+1)\sum_{k=1}^{m} \frac{1}{k} - 4m$. Since by definition $\bar{Q}_R = \widetilde{S}_M$, we have that

$$\mathbb{E}[\widetilde{S}_M^2] = \frac{1}{4}\sum_{i=1}^{M} \mathcal{QS}(m_i).$$

We recall the following basic version of the Martingale Central Limit theorem, cf. see [5] for example.

Theorem B.1 (Martingale CLT)**.** *Let X_t be a discrete-time Martingale with respect to filtration $\mathcal{F}_t$ for $t \geq 0$ with $X_0 = 0$. For $t \geq 0$, let $|X_{t+1} - X_t| \leq c$ with probability 1 for some constant $c > 0$ and let $\sigma_t^2 = \mathbb{E}[(X_{t+1} - X_t)^2|\mathcal{F}_t]$. Define $\tau_\nu = \min\{t : \sum_{s \leq t} \sigma_s^2 \geq \nu\}$. Then, $\frac{X_{\tau_\nu}}{\sqrt{\nu}}$ converges in distribution to $\mathcal{N}(0, 1)$ as $\nu \to \infty$.*

As established, $\bar{Q}_k$ is a discrete-time Martingale sequence satisfying all the conditions of Theorem B.1. By definition, τ_ν for $\nu = \frac{1}{4}\sum_{i=1}^{M} \mathcal{QS}(m_i)$ is precisely equal to the stopping time R. Since each competition has at least 2 players, as $M \to \infty$, $R \to \infty$. Therefore, by an application of Martingale CLT Theorem B.1, it follows that T_M^{refine} converges to $\mathcal{N}(0,1)$ in distribution as $M \to \infty$. This completes the proof of Theorem 4.1.

Appendix C: Evaluating p-Value of Naive Test Statistic

We present different approaches used in this paper to evaluate the p-value of the naive test statistic.

Using Central Limit Theorem (CLT). Under the null hypothesis of luck, $T_M^{\text{naive}}(n)$ is a random variable such that $\mathbb{E}[T_M^{\text{naive}}(n)] = 0$, $\mathbb{E}[(T_M^{\text{naive}})^2(n)] = 1$. For n large enough, by Central Limit Theorem, we have under the null hypothesis of luck that $T_M^{\text{naive}}(n) \to \mathcal{N}(0,1)$, where $\to$ denotes convergence in distribution and $\mathcal{N}(0,1)$ denotes the standard Normal or Gaussian distribution with mean 0 and variance 1. Given this, we propose the following natural rejection criterion using the central limit theorem for the hypothesis of luck.

The luck hypothesis, as defined in 2.1, is rejected with p-value $\alpha \in [0,1]$ if $\Phi(|T_M^{\text{naive}}(n)|) = 1 - \frac{\alpha}{2}$.

Using Hoeffding's Inequality. We recall the following probabilistic bound for Binomial random variable.

Proposition C.1. *For any $n \geq 1$, let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in $\{0,1\}$ with $\Pr(X_1 = 1) = \frac{1}{2}$. Then, for $t \geq 0$,*

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2}\Big| > t\Big) \leq 2\exp\big(-2nt^2\big).$$

That is,

$$\begin{aligned}\Pr\Big(|T_M^{\text{naive}}(n)| > t\Big) &= \Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2}\Big| > \frac{t}{2\sqrt{n}}\Big) \\ &\leq 2\exp\Big(-\frac{t^2}{2}\Big).\end{aligned}$$

Proof. The proposition follows immediately from the classical Hoeffding's inequality [6].

The upper bound in Hoeffding's inequality provides an exact upper bound on the p-value and therefore the luck hypothesis is rejected with p-value at most equal to

$$\min\Big\{1, 2\cdot\exp\Big(-\frac{(T_M^{\text{naive}}(n))^2}{2}\Big)\Big\}.$$

An advantage of using Hoeffding's inequality is that if the upper bound is extremely small, then we have exact evidence that the null hypothesis is statistically rejected. This is in contrast to using the central limit theorem, which is not *exact* as it is only an asymptotic result. However, it is known that a binomial distribution with parameters (n,p) is well-approximated by a Gaussian distribution if n is relatively large and p is not too close to 0 or 1. In our case, under the null hypothesis $p = 0.5$.

Using Monte-Carlo Estimation. If the p-value computed according to the central limit theorem and Hoeffding's inequality is not small, then a simulation should provide an accurate estimation with reasonably small number of steps. For that reason, in such scenarios, the standard Monte-Carlo estimation is an excellent method to obtain accurate p-value.

To that end, to evaluate p-value corresponding to $T_M^{\text{naive}}(n)$, we obtain samples from Binomial distribution, $B(n, \frac{1}{2})$; evaluate $|B(n,\frac{1}{2}) - n/2|$; record 1 if $|B(n,\frac{1}{2}) - n/2| > \sqrt{n}|T_M^{\text{naive}}(n)|/2$ and else record 0. Over a number of such simulations, the fraction of times we record 1 is an unbiased estimate of the desired p-value.