Synthese (2021) 199:13689–13748
https://doi.org/10.1007/s11229-021-03395-y

ORIGINAL RESEARCH

A Battle in the Statistics Wars: a simulation-based comparison of Bayesian, Frequentist and Williamsonian methodologies

Mantas Radzvilas[1] · **William Peden**[2,3] · **Francesco De Pretis**[1,4]

Received: 31 October 2020 / Accepted: 31 August 2021 / Published online: 1 November 2021

Abstract

The debates between Bayesian, frequentist, and other methodologies of statistics have tended to focus on conceptual justifications, sociological arguments, or mathematical proofs of their long run properties. Both Bayesian statistics and frequentist ("classical") statistics have strong cases on these grounds. In this article, we instead approach the debates in the "Statistics Wars" from a largely unexplored angle: simulations of different methodologies' performance in the short to medium run. We used Big Data methods to conduct a large number of simulations using a straightforward decision problem based around tossing a coin with unknown bias and then placing bets. In this simulation, we programmed four players, inspired by Bayesian statistics, frequentist statistics, Jon Williamson's version of Objective Bayesianism, and a player who simply extrapolates from observed frequencies to general frequencies. The last player served a benchmark function: any worthwhile statistical methodology should at least match the performance of simplistic induction. We focused on the performance of these

This article belongs to the topical collection "Recent Issues in Philosophy of Statistics: Evidence, Testing, and Applications", edited by Sorin Bangu, Emiliano Ippoliti, and Marianna Antonutti.

✉ William Peden
peden@esphil.eur.nl; w.j.peden@durham.ac.uk

Mantas Radzvilas
mantas.radzvilas@mail.huji.ac.il

Francesco De Pretis
francesco.depretis@unimore.it

1 Department of Biomedical Sciences and Public Health, School of Medicine and Surgery, Marche Polytechnic University, 60126 Ancona, Italy

2 Erasmus Institute for Philosophy and Economics, Erasmus University Rotterdam, 3000 DR, Rotterdam, The Netherlands

3 Centre for Humanities Engaging Science and Society, Durham University, DH1 3HN, Durham, UK

4 Department of Communication and Economics, University of Modena and Reggio Emilia, 42121 Reggio Emilia, Italy

methodologies in guiding the players towards good decisions. Unlike an earlier simulation study of this type, we found no systematic difference in performance between the Bayesian and frequentist players, provided the Bayesian used a flat prior and the frequentist used a low confidence level. The Williamsonian player was also able to perform well given a low confidence level. However, the frequentist and Williamsonian players performed poorly with high confidence levels, while the Bayesian was surprisingly harmed by biased priors. Our study indicates that all three methodologies should be taken seriously by philosophers and practitioners of statistics.

Keywords Bayesianism · Decision theory · Formal epistemology · Frequentism · Philosophy of statistics · Probability

1 Introduction

If there is any suspicion that philosophy of science is an ivory tower subject, it should be extinguished by what Deborah Mayo has called the "Statistics Wars" between classical statisticians, Bayesians, and a prismatic assortment of variations of these views (Ioannidis, 2005; Howson and Urbach, 2006; Wasserstein and Lazar, 2016; Mayo, 2018; van Dongen et al., 2019; Sprenger and Hartmann, 2019; Romero and Sprenger, 2020). Even apparently recondite questions about concepts like evidence, probability, and rational belief are connected with questions of statistical practice. These questions have been given particular salience by the "replication crisis", in which the rates of replication in published statistical research across a range of scientific fields are apparently below what would be expected from random variation alone (Gelman, 2015; Open Science Collaboration, 2015; Smaldino and McElreath, 2016; Fidler and Wilcox, 2018; Trafimow, 2018). All factions within the Statistics Wars make plausible (but often incompatible!) cases that their particular methodology, properly applied, can mitigate some of the malpractices behind the replication crisis. Furthermore, far from being dry debates, the Statistics Wars are frequently characterised by the sort of aggressive rhetoric, bombastic manifestos, and political maneuvering that their name would suggest. And this war-like atmosphere is understandable, because the Wars affect statistical practice, and thereby the health, wealth, and happiness of nations.

Statisticians are often very practical people, so when they are faced with these debates, many of them naturally think that either classical or Bayesian methods can be appropriate in various contexts. Statisticians often say very reasonable things such as "What matters is not who is right about issues in philosophical analysis, but what *works best*". Some philosophers of science might be attracted to such attempts to pacify the conflicts. Unfortunately, by itself, this pacification strategy fails, because assessing whether a methodology will "work" in a context depends on standards for evaluating what constitutes "working", and as soon as we start determining those standards, we start making choices about the proper analysis of terms like "evidence", "test", "severe test", and for that matter "work".

Yet there is wisdom in the practicing statistician's pragmatism. The philosophical debates within the Statistics Wars are rumbling on with no end in sight, as philosophical debates tend to do. One problems is that the rival statistical methodologies often have

different goals, e.g. Bayesians often focus more on the decisions of an ideally rational agent, whereas frequentists are often more interested in the reliability of a particular type of test. Statisticians could reasonably use different tools for different contexts. This opens up the opportunity to examine particular types of problem and see which method performs better at achieving some goal that all statisticians share.

One approach to such a question would be historical case studies of real-life cases where one or other method (or both) were used, and then comparing the two. However, it is rarely, if ever, possible to make such comparisons fairly and rigorously. Instead, we employ and expand the use of simulation studies for comparing statistical methodologies. Our simulations use a simple decision problem to pit four players against each other: (1) a confidence interval approach, which classical ("frequentist") statisticians might use for such a problem; (2) conditionalization upon a beta distribution, which many Bayesians would regard as an appropriate family of priors in this context; (3) a hybrid approach based on Jon Williamson's "Objective Bayesian" methodology, which involves forming beliefs about the relative frequencies using confidence intervals and then combines them with an updated version of the Principle of Indifference in order to generate precise probabilities; and (4) a "Sample" player who simply uses the relative frequencies in their observed samples to make point estimates of probabilities, akin to maximum likelihood estimation.

All the methodologies we discuss have some intuitively attractive performance properties as our sample sizes tend towards infinity. Long run performance properties are often given as the *raison d'être* of frequentist methods, while—in the right sorts of problems and with the right sorts of priors—Bayesian methods will also lead to credences (also called "degrees of belief") that, in the long run, approximate or even match the true relative frequencies (De Finetti, 1980). That is to these theories' credit: we never reach the long run, but information about it arguably provides defeasible information about the short run. If all the methodologies do well in the long run, then the short run is a more promising place to look for divergent performances. For this reason, we used simulations to compare the performance of players inspired by the methodologies when these players were given only small or moderately large samples.

Vituperative rhetoric and dismissive criticisms are common in the Statistics Wars. However, our results indicate that such disrespect is unfounded, at least for the decision problem we study. We found that Bayesianism, frequentism, and (what we call) Williamsonianism can all perform well with suitable player settings. As the number of games becomes large, this similarity can be explained by similar decisions. In the short run, we found differences in decisions. However, with the right player settings, all three of the statistical methodologies can exceed the benchmark that we set for them, and this similarity should increase the respect that the different factions of the Wars have for each other. Our results thus are contrary to what some dogmatic statistical warriors might expect. From each perspective, the other factions might seem absurd, but our study shows how all of them can sometimes lead to good decisions in two senses: (1) the players based on methodologies of statistics do not do significantly worse than the naïve sample-extrapolating player and (2) given the right settings, the players' performances with small samples are not bad in comparison to their performances with more information.

We briefly review the relevant literature in Sect. 2. We then explain our methods and the "players" in Sect. 3. We display the results and analyse them in Sect. 4, discuss them in Sect. 5, and conclude in Sect. 6.

2 The Statistics Wars: a multi-dimensional dispute

The Statistics Wars have a long history, dating back centuries, and they feature a "Who's Who" of statistics. The Wars are sometimes framed in terms of a simple conflict between Bayesian statisticians and classical (or "frequentist") statisticians, but this greatly oversimplifies the debates, as Mayo has detailed (Mayo, 2018). Indeed, no brief summary is possible, but for understanding our simulation study, it helps to note that the participants disagree across multiple dimensions, which we now broadly summarise.

2.1 The concept of probability

From a philosopher's perspective, perhaps the most fundamental dimension is which concepts of probability that a statistician regards as appropriate within statistical reasoning. As a preliminary, we must distinguish two potentially divergent domains in which we use probabilistic language: (i) everyday language like "The traffic will probably not be bad today" and (ii) the use of mathematical probability in science, such as a statistic for the probable error of a test. According to some philosophers of probability (including many labelled "frequentist") there is no need for a formal analysis of probabilistic language of type (i). In everyday life, we use "probability" and cognate terms in all sorts of ways, but it is debatable whether there is much of a connection between this ordinary usage and the proper place of probability in scientific reasoning. Insofar as the Statistics Wars are disputes over the proper concept of probability, they are disputes about which concept is appropriate *in scientific reasoning*.

1a Some theorists adopt *physical* interpretations of probability, in which probability is variously defined as finite relative frequencies (Venn, 1876), long run hypothetical relative frequencies (von Mises, 1957), propensities (Popper, 1959) or some other type of objective physical magnitude. According to this family of theories, probabilistic statements are generally logically contingent and non-psychological.[1] Epistemologically, there is no fundamental difference between knowledge of probabilities and other types of knowledge. While probabilities might appear in scientific hypotheses, no hypothesis itself has a probability: it is either true or false, but there is no sense in which statements like "This hypothesis has a probability of 0.5" can be true with a physical interpretation.

1b *Epistemic* interpretations analyse probabilistic concepts in terms of an epistemological framework—a system of concepts to do with evidence, belief etc. Within this family, we can distinguish two subgroups. The first subgroup interpret probability psychologically, as either unique ("Objective Bayesian") (Jaynes, 1957)

[1] With two obvious exceptions: assertions that are true by definition and assertions within psychology.

or non-unique ("Subjective Bayesian") (De Finetti, 1980) rational credences. In both cases, probability is analysed as the degree of confidence of either idealized or actual reasoners. The second subgroup of epistemic theorists regard probabilities as "partial entailment" relations between evidential statements and hypotheses (Keynes, 1921; Benenson, 1984; Kyburg, 2001). These "logical" probabilities are arguably guides to rational credences,[2] but they are not strictly speaking the same thing, just as deductive logical relations (on any non-psychologist philosophy of logic) can often guide rational belief, but cannot be reduced to psychological relations. On either a psychological or partial entailment interpretation, scientific hypotheses can have probabilities.

1c Pluralist combinations of 1a and 1b. For example, Carnap (1945b) combined a partial entailment interpretation of the probabilities of hypotheses ("H's probability relative to our total evidence is 0.5" etc.) with a long run relative frequency interpretation of the probabilities *mentioned in* hypotheses, such as assertions about radioactive half-lives in physics. Pluralistic interpretations are very common among philosophers of science, e.g. Popper (2002), Howson and Urbach (2006), Williamson (2010).

To appreciate these differences in the conceptual analysis of probability, consider how these approaches might interpret some sentences that might occur within scientific reasoning:

> H_1 : *"There is a high probability that height will be approximately normally distributed in a large subset of humans."*

– *Physical*: There is a tendency (interpreted in terms of relative frequencies or propensities) of large subsets of humans to be approximately normally distributed with respect to their height.
– *Epistemic*: On a *pure* psychological interpretation (such as an uncompromising Subjective Bayesianism) this is an assertion of high confidence that height will be approximately distributed in a large subset of humans, or that a high degree of confidence is rationally appropriate. However, note that most contemporary Subjective Bayesians are pluralists: see below. On a *pure* partial entailment interpretation, H_1 is best interpreted as a claim about a probability relation between statements, plus the assertion that one of those statements is true. For instance, we can interpret H_1 as the combination of three assertions: (i) a statistical hypothesis H_s, which asserts a physical probability statement such as "There is a high relative frequency of an approximately normal distribution of height among large subsets of human beings", (ii) the claim that H_s is our best evidence regarding the single-case hypothesis H_t "A particular randomly selected large subset of humans is approximately normally distributed with respect to height", and (iii) a claim that H_s has a high partial entailment relation in favour of H_t (Benenson, 1984; Kyburg, 1990).
– *Pluralist*: Most pluralists would adopt a physical interpretation of H_1, in most contexts.

[2] Popper (2002) makes use of logical probabilities, without regarding them as mapping onto rational credences; indeed, he entirely rejects the latter concept.

H_2 : *"It is 99% probable that there is no life currently on Mars."*

- *Physical*: No relative frequency interpretation of H_2 is possible, without being false. H_2 might have some function in unscientific language as a way of indicating confidence, so a supporter of a relative frequency interpretation might interpret H_2 as an informal, extra-scientific expression of confidence. On a propensity interpretation, H_2 could be an assertion about the physical chances of life on Mars at this time.
- *Epistemic*: On a subjectivist psychological interpretation (as in Subjective Bayesianism) H_2 is an assertion about the speaker's confidence. On an objectivist psychological interpretation (as in Objective Bayesianism) H_2 asserts that 99% is the degree of confidence, for the hypothesis that there is life on Mars, that is uniquely rational relative to some body of evidence. On a partial entailment interpretation, H_2 is an assertion about a partial entailment relation between the hypothesis that there is life on Mars and some body of evidence, such as the total body of evidence in astrobiology.
- *Pluralist*: Most pluralists would adopt some epistemic interpretation of H_2, in most contexts.

2.2 The evaluation of statistical methods

If we are testing a statistical hypothesis, then we need to select a method. Often, we might look at what has worked well in the past, but that just pushes the question back: "This method generally works well" is itself a (vague) statistical hypothesis. Additionally, if we are investigating some new or relatively new topic, we might not know what "works". There are three broad families of answers to this question in the Statistics Wars.

2a One criterion is the method's *long run performance*. For example, suppose that we kept on using a particular procedure for estimating some relative frequency as our sample sizes grew larger and larger. Suppose also that there *is* a limit that the relative frequency would tend in the long run.[3] Will the margin of error of our estimates tend towards zero if we stick to this method? If so, does the estimation method have other attractive properties, e.g. does it not depend on the language in which we formulate our hypothesis (Reichenbach, 1938; Feigl, 1954; Salmon, 1967)? Alternatively, we might ask whether the test would almost always provide statistically significant[4] results in the long run (Fisher, 1947, p. 14)? Alternatively, we may ask whether, in the long run, the test provides a satisfactory[5] combination of mistaken rejections (Type I errors) and mistaken non-rejections (Type II errors) (Neyman, 1949)?

2b In terms of the *doxastic (belief-modifying) properties of the test* (Howson and Urbach, 2006; Williamson, 2010). Is the testing procedure a rational way to update

[3] If there is no limit, then obviously no method will work for estimating it.

[4] For some contextually determined standard of statistical significance.

[5] As with statistical significance, a "satisfactory" balance must be determined contextually, but Neyman went deep into the decision theory behind such a choice—what he called "inductive behavior" (Neyman, 1957).

our beliefs? Put differently, if we adopt a doxastic approach to evaluation, we assess the test based on how its use would affect the structure of our beliefs. For example, diachronic Dutch Book Arguments are intended to show that, if we adopt some systematic procedure for revising our beliefs that is not an application of Bayesian conditionalization, then we shall be vulnerable to accepting a series of bets in which we will inevitably lose money. Others advocate conditionalization based on its benefits for "epistemic" utilities (Greaves and Wallace, 2006). However, not all philosophers who adopt a doxastic approach to evaluating tests have endorsed conditionalization as a general norm (Bacchus et al., 1990).

2c In terms of whether the testing procedure provides a *severe test* in a particular context (Mayo, 1996; Spanos, 2010; Mayo, 2018). A "severe" test adequately probes the possible sources of error in our inference, like extrapolations from spurious correlations in the data. In particular, there should be a good chance that our test would uncover a flaw in the hypothesis that we are investigating, if such a flaw exists. A hypothesis might have an excellent fit for our data, but it has not been severely tested unless it was at risk of being rejected by our test, if it were false. Put another way, there should be a low probability that using the testing procedure will lead us to a mistaken inference from the data, i.e. a low "error probability". The role of probability in this methodology is to assess the extent to which tests have probed (or would probe) our hypotheses. Note that severity requires more than just the long term performance properties of 2a: a test might have good long run properties, but if the test is unlikely to detect a flaw in a single usage and we only evaluate a hypothesis once using that test, then that hypothesis has not been severely tested. In general, according to a severe testing ("error statistical", "probative") evaluation approach, there are problems (such as p-hacking) that do not involve the lack of good long run performance properties, but nonetheless can make our tests inadequate (Mayo, 2018, pp. 13–14).

Naturally, there are interrelations between the criteria used by these three approaches to evaluation. For example, adherents of 2a and 2c care about revising our beliefs in rational ways, but evaluate what is a "rational" method for belief revision in terms of long run performance or probative capacity. They believe that the formal models of belief used by adherents of 2b are not the right place to start.[6] Similarly, one intuitive feature of proper doxastic states is that there should be proper coherence between (a) beliefs about long run performance of tests and (b) beliefs in how to update one's beliefs given individual tests. Meanwhile, error statisticians (2c) think that the long run performance qualities of a test are one aspect of its probative quality (Mayo, 2018).

2.3 Testing procedures and statistics

Finally, there are testing procedures and test statistics that different methodologists judge to be appropriate in general or in some context. Even given a particular interpretation of probability and shared standards to evaluate our methods, selecting among types of tests can be controversial.

[6] And, for those methodologists who reject such formal models entirely, not the right place to finish either.

3a Classical tests and test statistics. The former include confidence interval estimation, null-hypothesis significance testing, maximum likelihood estimation, and so on. For our test statistics, we might use p-values, confidence levels, and other descriptive statistics that indicate the long run properties of the testing procedure that we shall implement.

3b Bayesian tests and test statistics, involving the use of conditional probabilities and Bayes' Theorem to update a prior using conditionalization. There are associated test statistics like Bayes factors and posterior probabilities to indicate, respectively, (i) the performance of the tested hypothesis in comparison to another hypothesis and (ii) the warranted degree of belief in the tested hypothesis given the initial prior and the acceptance of the new data.

3c Pluralist combinations of 3a and 3b. For example, according to Kyburg and Teng (2001), we should use Bayesian methods when we have prior probabilities based on previously inferred statistical information about relative frequencies. However, when such background knowledge is lacking, we should use classical statistics.

Often, in methodological debates, the immediate battleground is one of these dimensions, but disagreements in the others intrude and further complicate the discussions. A simple type of example occurs when, during debates about the adequacy of particular testing procedures, the term "probability" is insufficiently clarified. Researcher A can then say things to Researcher B that are obviously absurd given the other's interpretation of "probability", even though they might be able to develop a practical consensus on the immediate point of disagreement if they made some disambiguations. In such cases, a fundamental disagreement across dimension 1 (Sect. 2.1) unnecessarily frustrates the possibility of consensus on dimension 3 (*supra*, Sect. 2.3). Moreover, while some of these viewpoints are closely correlated in practice, a great variety of permutations are logically consistent.

Despite the complexities of this debate, we were able to identify three ideas to inspire the three non-benchmark players in our simulations. Thus, in addition to our benchmark player (who naïvely uses their observed sample frequencies) we based on our players on the following methodologies:

A. *Frequentism*, in the sense of adopting 3a, at least for the reasoning problem that we use in this paper. Thus, we are specifically referring to the *methods* of classicial statistics and their direct epistemic consequences of providing estimates of relative frequencies. For classical methods, we only need interpretation 1a, although these methods are compatible with other interpretations of probability. In particular, we shall focus on confidence interval estimation. Note that this choice of statistical method for our problem might be justified by any of the evaluative methodologies falling under the categories 2a, 2b, or 2c. For the sake of convenience, we shall assume that a frequentist will also interpret probabilities as relative frequencies, even though this is not necessitated by confidence interval estimation methods.

B. *Bayesianism*, in the sense of using 3b for the problem in our study. Bayesians can adopt either a strictly epistemic interpretation, 1b, or a pluralist interpretation, 1c. Bayesians more or less invariably arrive at their view via a doxastic approach to evaluating statistical methods, 2b.

C. *Williamsonianism*, by which we mean Jon Williamson's combination of "calibrating" via reasoning that is fundamentally based on classical methods, 3a, but using these relative frequency estimates to guide credences, in accordance with 1b. Given the constraints from beliefs about relative frequencies, the credences are subsequently determined by the Maximum Entropy Principle, described below in Sect. 3.3.3. The overall package is justified by Williamson on doxastic grounds (Williamson, 2010).[7] In other words, Williamson has an epistemic theory of probability, but there is a large role for classical statistical methods in his theory of inference, leading to divergences from conditionalization (Williamson, 2010, pp. 167–169). Williamsonianism hence offers an interesting combination of ideas that are typically attributed to either frequentism or Bayesianism alone.

Since these methodologies tend to differ across the evaluative dimension, it is extremely difficult to find ways of comparing them that do not beg crucial questions. However, a shared goal is that statistics should help scientists achieve their aims. Scientists have many aims where statistics could reasonably be expected to help: prediction, explanation, control etc. However, one aim of science that seems amenable to formal study is making good decisions under conditions of uncertainty. We mean "decisions" in a broad sense of choices among actions. As Jerzy Neyman pointed out, this is a broad enough definition to include most or all of scientific practice. Decision-making includes choosing to publish a paper, to make an expert pronouncement, to invest in developing an experiment, as well as non-scientific decisions like financial investing or trying to design a better mousetrap (Neyman, 1941).

In our simulation, we programmed four different "players" using belief models and choice models. These players were inspired by the three positions in the Statistics Wars that we identified above, plus a benchmark player. The latter enables to see if these methodological positions fulfill a reasonable basic criterion of adequacy: they should do at least as well as someone who just expects future frequencies to match the sample frequency that they have observed thus far. We focused on non-asymptotic decision making, i.e. performance in a finite number of decisions. This might seem to prejudge the case against frequentists, because many of them emphasise long run performance as the proper criterion for testing. However, frequentists are just as interested in making good decisions in the relatively short run as any other other sensible people; that a test procedure has good long run properties is arguably indicative that it will help us to make good decisions in the short run, and frequentists have acknowledged this goal as one objective of statistical inference.

We have no pretense that our study (or simulation studies in general) will or should end the Statistics Wars. Given the persistence of these conflicts throughout the history of statistics, it is probable (in more than one sense) that they are not going to be resolved any time soon. Nonetheless, simulations provide a useful way to assess which methods help guide us towards better decisions and under what conditions. In particular, they can help us evaluate whether frequentism, Bayesianism, or Williamsonianism will enable us to perform systematically better in a particular type of decision problem.

[7] Williamson uses "Objective Bayesianism" for his view, but we use his name for this approach to distinguish it from (B) and to avoid confusion with the panoply of very different "Objective Bayesianisms".

3 Methods

3.1 Simulations

Simulations are a relatively new tool in the philosophy of science. For a review up to 2009, see Winsberg (2009). There is an even sparser literature that applies simulations to the Statistics Wars. However, within statistics, there are many studies using simulations to assess Bayesian tools versus frequentist tools for particular problems. To give just a few examples, Avinash Singh *et al.* compare Bayesian and frequentist methods for estimating parameters of small sub-populations using simulations (Singh et al., 1998). Gilles Celeux *et al.* use simulations to compare Bayesian regularization methods against frequentist methods (Celeux et al., 2012). Daniel McNeish applies simulations within an investigation of the strengths and weaknesses of Bayesian estimation with small samples (McNeish, 2016).[8] There is also a literature on Bayesian estimation versus frequentist estimation in structural equation modelling (Smid et al., 2019).

Within philosophy, we found only two earlier studies comparing Bayesian and frequentist methodologies using simulations. Felipe Romero and Jan Sprenger employ simulations to argue in favour of a Bayesian approach to the replication crisis (Romero and Sprenger, 2020). Their study is interesting, but its topic is inferential success under different sorts of scientific institutions, and therefore very different from our own. Instead, our study builds on a philosophically-motivated use of simulations by Kyburg and Teng (1999) to compare frequentist statistics versus Bayesian statistics in a simple decision problem. They use simulations to investigate a game (against Nature) in which a frequentist player and a Bayesian player separately make bets based on a random binomial event—a toss of a coin with unknown bias or fairness. The frequentist player used confidence intervals and the Bayesian player used conditionalization. To serve as a benchmark, Kyburg and Teng also use a player, *Sample*, who simply estimates that the probabilities will be equal to the frequency in the sum of their samples. After observing some randomly generated coin tosses, players had the option to buy a ticket for heads at a price t or a reversed ticket (effectively betting that the toss will land tails) at a price $(1 - t)$. The variable t had a randomly generated price in the interval $[0, 1]$. Each player was given a decision algorithm for deciding when and how to bet.

The study then compares the average profits across a range of conditions. The Bayesian player performed at about the benchmark level. Across the full range of possible coin biases that Kyburg and Teng investigated, the Bayesian performed best when their prior corresponded to that required by the Principle of Indifference.[9] The frequentist tended to make better profits than the Bayesian and the benchmark. Kyburg and Teng do not explain their result, and it has been noted as a puzzle (Schoenfield, 2020, p. 2).

[8] While we are also interested in small sample inference in this article, we wanted to make comparisons in terms of decision making and also with individual predictions, so McNeish's study was not suitable as a basis for our own research.

[9] The Principle of Indifference requires that, if your evidence is equivocal with respect to $m \geq 2$ mutually exclusive and exhaustive states of the world, then your degree of belief in each of them should also be symmetrical: $1/m$ for each state of the world.

However, comparisons were difficult because the frequentist (unlike the Bayesian) would sometimes refuse to bet. In particular, the frequentist player's algorithm meant that it would not bet if t was within their confidence interval $[x, y]$ where $0 \leq x \leq 1$, $0 \leq y \leq 1$, and $x \leq y$. Kyburg and Teng believed that this feature of the frequentist algorithm meant that this player did much better than the Bayesian in absolute profits. While Kyburg and Teng attempted to address this issue by using average profits rather than total profits, the difficulty remains that the players are being compared over an unequal number of bets. Consequently, it is possible that their results are partly explained by this asymmetry.

Additionally, Kyburg and Teng's study was published in conference proceedings, so quite a few details are obscure, such as whether the players confronted the same sequences of coin tosses and ticket prices. Additionally, they only report the results of 100 simulations for each parameter setting, which means that many of the variations in performance that they find could be explained by random error.

Despite its limitations, the Kyburg and Teng study offers a novel and relatively simple way to compare the decision-theoretic performance of Bayesianism and frequentism in a short run reasoning problem. On the other hand, there was a lot of scope for the expansion, modification, and clarification of their study. Our study design enables comparisons with Kyburg and Teng (1999), but also explores several novel directions; it also attains a level of rigour and detail beyond their conference paper-style investigation. We also sought to reduce the problem of random error by conducting our simulations many times, as we detail in Sect. 3.5. Overall, we conducted two sets of 1000 simulations for each player setting. We also varied the random parameters—the ticket prices and the coin toss results—in the decision problem. Each simulation involved 1000 games, with 1000 simulations for each of five coin biases, and therefore there was a total of 5 million games per player setting in each set of simulations. In this way, by using modern computational power and software, we were able to use a "Big Data" approach that was missing in the study by Kyburg and Teng (1999).

3.2 Decision problem

The game that the players faced in our simulations is based around a finite sequence of Bernoulli trials with an initially unknown physical probability. For convenience, we use the terminology of coin tosses, but we stress that, whereas a real-world player would know that a coin is likely to land heads/tails with roughly equal long run frequency (they know that there is a very low proportion very biased coins in the world) in our game they have no such background knowledge about the coin being tossed. However, the players do all know that the tosses are random, in the sense that each toss has an equal (but unknown) chance of landing heads. They also know that the order of coin tosses is irrelevant, i.e. patterns in the sequences of tosses provide no information. Thus, they will only regard sample statistics as relevant evidence for the coin biases.

The basic unit of the decision problem is a decision to bet on *heads* (action b_h), bet on *tails* (action b_t), or to hold (action $\bar{b}$). Players have known and fixed payoffs for the decisions that we depict in Fig. 1.

Fig. 1 Player payoff matrix

	$heads$	$tails$
b_h	t	$-t$
b_t	$-(1-t)$	$(1-t)$
$\bar{b}$	$-\varepsilon$	$-\varepsilon$

The variable t is a US dollar (USD) value. It was randomly generated for each game; since it takes values in the [0, 1] interval, its average price over the full series of simulations was very probably around 0.5. Action b_h gives a return of t if the result is *heads*, but incurs a loss $-t$ if the results is *tails*, because the player bought the ticket at a price t and did not win any money. Action b_t gives a return of $(1-t)$ if the result is *tails* and incurs a loss $-(1-t)$ if the result is *heads*, again reflecting the price of purchasing the ticket and the absence of a return. Action $\bar{b}$ gives a guaranteed result of $-\varepsilon$. We set ε to be in the unit interval [0, 1].

In Kyburg and Teng (1999), a decision to hold has a guaranteed result of 0. However, as they note (pp. 364–365) it is interesting to see what happens if players are forced to bet. "Force" implies a sanction, and by setting ε high enough, we can make it rational for a payoff-maximising player to bet. In particular, if $\varepsilon > t$ or $\varepsilon > (1-t)$, then there will always be at least one bet that is rational for the player, because the bet will have a non-negative expected payoff, whereas $-\varepsilon$ is greater than the loss $-t$ or $-(1-t)$. By setting ε to 1, we can ensure that all players will bet in each game.

Our simulation presents players with sequences of coin tosses. After a certain number of trials, the players are given a choice whether and how to bet. The players' decisions are completely independent and non-interdependent, so they will not convert to another player's approach if they see that player outperforming them.

3.3 The players

Our simulations feature four players, *Bayes*, *Frequentist*, *Williamsonian*, and *Sample*. The player *Sample* does not correspond to a standard approach to our decision problem, but it is useful, because a statistical methodology should at least match *Sample*'s performance. *Sample* can also be interpreted as one way that a frequentist might act if they were forced to give precise relative frequency estimates and use these estimates to guide their decisions. The players have significant differences in their learning rules and decision rules, so we shall discuss them each individually.

3.3.1 The Bayesian

Bayes is inspired by Subjective Bayesianism, so there are infinitely many prior probability distributions that they might adopt. However, in our simulations, *Bayes* will use a

type of prior that many Subjective Bayesians would use for such a reasoning problem. *Bayes* knows that the tosses are random. Therefore, they can estimate the probability that a particular toss will land *heads* simply by deciding one degree of belief in *heads* for each and every toss. Since they know that each toss will land *heads* or *tails*, the probability of *tails* will be one minus the probability of *heads*. The beta distribution is a popular option in Bayesian statistics for this type of estimation problem (Mun, 2008, p. 906). The beta distribution enabled us to generate a wide variety of priors for *Bayes* using just two parameters. For these reasons, we used this procedure for characterising initial priors for *Bayes*.

Bayes's belief revision procedure in this particular game can be represented with an epistemic model $\mathrm{K} := \{\Omega, \Theta, H, c, p\}$, where $\Omega := \{heads, tails\}$ is the set of possible outcomes of a single coin toss (the decision-relevant states), $\Theta := \left\{\theta \in \mathbb{R}_{\geq 0} : 0 \leq \theta \leq 1\right\}$ is the set of values representing all the possible biases of the coin towards *heads* (from this point onward, we will simply call θ "coin bias"), $H := \{\mathrm{H} \cup \{\emptyset\}\}$ is the set of possible histories[10] with a typical element h, where $\mathrm{H} := \{heads, tails\}^n$ is the set of possible histories that can be generated by $n > 0$ coin tosses, such that each $\mathrm{h} \in \mathrm{H}$ is a sequence $\{e_1, \ldots, e_n\}$ where each $e_i \in \{heads, tails\}$, $c : H \to \mathbb{Z}_{\geq 0}$ is a *heads* event counting function and for every history $h \in H$, $c(h) = |\{e_i \in h : e_i = heads\}|$, and $p : H \to \Delta(\Theta)$ is a probability measure which assigns, to every history $h \in H$, a probability distribution on Θ. After observing some history $h \in H$, *Bayes* revises their beliefs about each possible coin bias $\theta \in \Theta$ via the standard Bayes rule

$$p(\theta|h) = \frac{p(\theta)\, p(h|\theta)}{p(h)}, \text{ where } p(\theta) = p(\theta|h = \emptyset) > 0 \text{ denotes a prior.} \quad (1)$$

Since each coin toss is a Bernoulli trial and the game generates a binomial distribution, we can use a counting function c to reformulate the Bayes rule for each history $h \in H$ as

$$p(\theta|c(h), n) = \frac{p(\theta)\, p(c(h), n|\theta)}{p(c(h), n)}, \quad (2)$$

where $p(c(h), n|\theta) = \binom{n}{c(h)} \theta^{c(h)} (1-\theta)^{n-c(h)}$.

Since our setup ensures that, for every history $h \in H$, the posterior probability distribution $p(\theta|c(h), n)$ will be in the same family of probability distributions as prior $p(\theta)$, we can use a conjugate prior and represent the prior by a standard beta distribution with parameters a and b. By denoting beta distribution as $\mathrm{B}(a, b)$, we can express the prior as

$$p(\theta|a, b) = \frac{\theta^{a-1} (1-\theta)^{b-1}}{\mathrm{B}(a, b)}. \quad (3)$$

[10] We use $\emptyset$ for "no observation", not for the empty set.

The Bayes rule for a setup with a beta distribution and some history $h \in H$ can be defined as

$$p\left(\theta|c\left(h\right),n\right)=\frac{\binom{n}{c\left(h\right)}\theta^{c(h)+a-1}\left(1-\theta\right)^{n-c(h)+b-1}/\mathrm{B}\left(a,b\right)}{\int_0^1\left(\binom{n}{c\left(h\right)}\bar{\theta}^{c(h)+a-1}\left(1-\bar{\theta}\right)^{n-c(h)+b-1}/\mathrm{B}\left(a,b\right)\right)\mathrm{d}\bar{\theta}}$$
$$=\frac{\theta^{c(h)+a-1}\left(1-\theta\right)^{n-c(h)+b-1}}{\mathrm{B}\left(a+c\left(h\right),b+n-c\left(h\right)\right)}. \tag{4}$$

Notice that this Bayes rule is just another beta distribution with parameters $a+c\left(h\right)$ and $b+n-c\left(h\right)$. Armed with the prior and the Bayes Rule, *Bayes* now has all they need both to assign initial priors in the coin toss simulations and to update them in accordance with conditionalization.

Bayes's choices can be represented with a choice model $\mathrm{C_b} := \{\Omega, \mathrm{D}, H, \mu, \pi\}$, where $\mathrm{D} := \left\{b_h, b_t, \bar{b}\right\}$ is the set of possible actions, while b_h represents the decision to bet on *heads*, b_t represents the decision to bet on *tails*, and $\bar{b}$ represents the decision to hold, Ω represents the outcomes of a single coin toss, $\mu : \Omega \times H \rightarrow X$, where $X := \left\{x \in \mathbb{R}_{\geq 0} : 0 \leq x \leq 1\right\}$ is the conditional probability function which assigns, to every history $h \in H$, some probability $q \in [0, 1]$ on *heads* and $1-q$ on *tails*, and $\pi : \mathrm{D} \times \Omega \rightarrow \mathbb{R}$ is the payoff function which assigns, to every possible action-state combination, a real number representing player's payoff. The payoff function assigns payoffs to the action-state combinations in the same way as they are represented in Fig. 1, so that

$$\begin{array}{ll}\pi\left(b_h, heads\right)=t; & \pi\left(b_h, tails\right)=-t;\\ \pi\left(b_t, heads\right)=-\left(1-t\right); & \pi\left(b_t, tails\right)=\left(1-t\right);\\ \pi\left(\bar{b}, heads\right)=-\varepsilon; & \pi\left(\bar{b}, tails\right)=-\varepsilon.\end{array}$$

The conditional probability function μ is such that, for history $h = \emptyset$, the probability of *heads* is

$$\mu\left(heads|h=\emptyset\right)=\int_0^1\left(p\left(\theta\right)\theta\right)\mathrm{d}\theta, \tag{5}$$

while $\mu\left(tails|h=\emptyset\right)=1-\mu\left(heads|h=\emptyset\right)$.

For every history $h \in \mathrm{H}$ the probability of *heads* is

$$\mu\left(heads|h\right)=\int_0^1\left(p\left(\theta|c\left(h\right),n\right)\theta\right)\mathrm{d}\theta, \tag{6}$$

while $\mu\left(tails|h\right)=1-\mu\left(heads|h\right)$.

The expected payoff from some action $\mathrm{d} \in \mathrm{D}$ given some history $h \in H$ is

$$\mathbb{E}_\pi\left[\mathrm{d}|h\right]=\pi\left(\mathrm{d}, heads\right)\mu\left(heads|h\right)+\pi\left(\mathrm{d}, tails\right)\mu\left(tails|h\right). \tag{7}$$

We assume that *Bayes* is a rational player who seeks to maximise the expected payoff. Thus, for any history $h \in H$, *Bayes* always chooses an action $\mathrm{d} \in \mathrm{D}$, such that

$$\mathbb{E}_\pi\left[\mathrm{d}|h\right] \geq \mathbb{E}_\pi\left[\bar{\mathrm{d}}|h\right], \text{ for all } \bar{\mathrm{d}} \in \mathrm{D}. \tag{8}$$

When more than one action satisfies this requirement, *Bayes* will make a random choice among these actions. Since hold is never a strictly dominant option, we simplified *Bayes* by supposing that they never choose this option.

3.3.2 The Frequentist

Just as Subjective Bayesians might adopt many different priors for a coin tossing problem, so there are multiple ways that a frequentist might estimate the relative frequencies involved, i.e. the frequencies of heads and tails among the tosses. However, many frequentists would regard estimation using confidence intervals as a reasonable method for our decision problem, in which *Frequentist* starts out with no initial estimate of what the relative frequency might be and where they want to estimate the general relative frequency for the purpose of guiding their betting behaviour on particular tosses. Additionally, it is this approach that Kyburg and Teng, who are frequentists (in the broad sense that we are using in this article) adopt for their frequentist player.

We explain the technical details of how *Frequentist* works below, but we shall also provide an informal summary beforehand. *Frequentist* will estimate that, in the population of coin tosses, the relative frequency of *heads* is within a confidence interval. Since they know that all the tosses will land *heads* or *tails*, that confidence interval also provides them with an estimate of the relative frequency of *tails*. A "confidence interval" is an estimate that a population frequency is within some range. Let us focus on the long run relative frequency of *heads* in the coin tosses. This is the relative frequency that a hypothetical infinite series of tosses would approach in the limit as the number of tosses increased. A confidence interval of [0.4, 0.6] for the long run frequency of *heads* in the population of coin tosses is an estimate that the relative frequency of the coin landing *heads* is at least 40% of the coin tosses and no more than 60% of them. *Frequentist* will tentatively believe that the relative frequency for the coin toss population is somewhere in the estimated interval. They will tentatively dismiss values outside this range, such as 0.2 or 0.9, as possible values for the relative frequency. *Frequentist*'s beliefs are "tentative" in the sense that, with more sample data, they can change the interval to very different values; a confidence interval using one sample for estimation could be [0, 0.5], but the interval using an enlarged sample could be [0.9, 0.99].

The confidence *level* is not a probability, but instead a value describing the procedure used to estimate the intervals. We stipulate that *Frequentist* knows that the sampling assumptions for confidence interval testing are satisfied. The confidence level is set via a parameter α, between 0 and 0.5, such that the level is $(1 - \alpha)$. The value of α is exogenous to the confidence interval methodology, so we simulated how *Frequentist* will perform with a range of values of α. The confidence level tells us, for a particular sample size, the *minimum* possible long run relative frequency of correct estimates of

a confidence interval of that width using that sample size. For example, if $\alpha = 0.05$, then the confidence level for the intervals estimated using this method is 0.95, or 95%. The level is only the minimum, because if the population is homogeneous—all tosses land *heads* or all tosses land *tails*—then the success rate will be 100%. Therefore, the success rate in the example could be anywhere from 95 to 100%. Once *Frequentist* has estimated a confidence interval given their existing observations, they will use it to guide their betting behaviour for the individual tosses on which they can bet.[11]

We now formulate this belief revision procedure in detail. *Frequentist* revises their estimate of the coin bias in a way that can be represented with a model $\mathrm{F} := \{\Omega, \mathrm{H}, \mathrm{k}_\alpha, \mathrm{c}\}$, where Ω is the set of possible outcomes of a single coin toss, $\mathrm{H} = \{heads, tails\}^n$ is the set of possible histories which can be generated with $n > 0$ coin tosses,[12] $\mathrm{c} : \mathrm{H} \to \mathbb{Z}_{\geq 0}$ is the *heads* event counting function and for every history $\mathrm{h} \in \mathrm{H}$, $\mathrm{c}(\mathrm{h}) = |\{e_i \in \mathrm{h} : e_i = heads\}|$, and $\mathrm{k}_\alpha : \mathrm{H} \to \mathrm{P}(X)$, where $X := \left\{x \in \mathbb{R}_{\geq 0} : 0 \leq x \leq 1\right\}$, is a function which assigns, to every history $\mathrm{h} \in \mathrm{H}$, an interval $\mathrm{k}_\alpha(\mathrm{h}) = (k_l, k_u)$ with lower bound k_l and upper bound $k_u \geq k_l$. Since the game is a sequence of Bernoulli trials that generates a binomial distribution, we can determine *Frequentist*'s confidence interval by calculating the Clopper-Pearson interval. For any history $\mathrm{h} \in \mathrm{H}$, the lower and upper boundaries of the Clopper-Pearson interval can be represented as the following beta distribution quantiles

$$k_l = \mathrm{B}\left(\frac{\alpha}{2}; \mathrm{c}(\mathrm{h}), n - \mathrm{c}(\mathrm{h}) + 1\right); \tag{9}$$

$$k_u = \mathrm{B}\left(1 - \frac{\alpha}{2}; \mathrm{c}(\mathrm{h}) + 1, n - \mathrm{c}(\mathrm{h})\right). \tag{10}$$

Thus, after observing a history of n coin tosses $\mathrm{h} \in \mathrm{H}$ with $\mathrm{c}(\mathrm{h}) \geq 0$ *heads*, if *Frequentist* adopts a an level $1 - \alpha$, they will estimate the actual coin bias to be within an interval $\mathrm{k}_\alpha(\mathrm{h}) = \left(\mathrm{B}\left(\frac{\alpha}{2}; \mathrm{c}(\mathrm{h}), n - \mathrm{c}(\mathrm{h}) + 1\right), \mathrm{B}\left(1 - \frac{\alpha}{2}; \mathrm{c}(\mathrm{h}) + 1, n - \mathrm{c}(\mathrm{h})\right)\right)$, and they will not take into account any value $k \notin \mathrm{k}_\alpha(\mathrm{h})$.

There is no consensus on how to model decisions with interval-valued beliefs. There are many options in the literature (Resnik, 1987). However, we programmed *Frequentist* to use what seems to be the decision procedure in Kyburg and Teng (Kyburg and Teng, 1999), which we also interpreted using other publications by Kyburg, in particular (Kyburg, 1990, 2003).

Frequentist's choices can be represented with a choice model $\mathrm{C_f} := \{\Omega, \mathrm{D}, \mathrm{H}, \mathrm{k}_\alpha, \pi, \phi\}$, where Ω is the set of possible outcomes of a coin toss, D is the set of possible actions, H is the set of possible histories, k_α is the function which assigns a Clopper-Pearson interval to every history $\mathrm{h} \in \mathrm{H}$, π is a payoff function identical to the one defined for *Bayes* player, and $\phi : \mathrm{D} \times \mathrm{H} \to \mathrm{P}(\mathbb{R})$ is a function

[11] In our decision problem, where the players all know that the coin tosses are random, we can safely assume that *Frequentist* should use the estimates of the general relative frequency to form their beliefs about particular tosses. In problems where *Frequentist* is confronted with e.g. different statistics for different overlapping reference classes, the applicability of this assumption becomes much more complicated, as frequentists have known since at least (Venn, 1876). For influential discussions, see also Salmon (1967) and Reichenbach (1971).

[12] The method used to compute the confidence interval cannot be employed when $h = \emptyset$. However, the first bet in the game occurs only after each player observes a number of coin tosses.

which assigns, to every action-history pair $(\mathrm{d}, \mathrm{h}) \in \mathrm{D} \times \mathrm{H}$, an expected payoff vector $\phi(\mathrm{d}, \mathrm{h}) = \left(\mathbb{E}_\pi\left[\mathrm{d}, \underline{k}\right], \ldots, \mathbb{E}_\pi\left[\mathrm{d}, \overline{k}\right]\right)$, where $\underline{k} = \min(\mathrm{k}_\alpha(\mathrm{h}))$, $\overline{k} = \max(\mathrm{k}_\alpha(\mathrm{h}))$, and $\mathbb{E}_\pi[\mathrm{d}, k] = k\pi(\mathrm{d}, heads) + (1-k)\pi(\mathrm{d}, tails)$ for every $k \in \mathrm{k}_\alpha(\mathrm{h})$.

The choice behaviour of *Frequentist* can be defined with the *interval-dominance* principle. Action $\mathrm{d} \in \mathrm{D}$ interval-dominates action $\bar{\mathrm{d}} \in \mathrm{D}$ if and only if, given history $\mathrm{h} \in \mathrm{H}$, $\min(\phi(\mathrm{d}, \mathrm{h})) > \max\left(\phi\left(\bar{\mathrm{d}}, \mathrm{h}\right)\right)$. For any pair of actions $\mathrm{d} \in \mathrm{D}$ and $\bar{\mathrm{d}} \in \mathrm{D}$, *Frequentist* always chooses action d if d interval-dominates action $\bar{\mathrm{d}}$, and is indifferent between d and $\bar{\mathrm{d}}$ when neither action interval-dominates the other. Reasoning this way almost corresponds extensionally to the frequentist player's behaviour in Kyburg and Teng (1999), where no general decision rule is provided.[13]

We now apply this approach to the coin tossing problem. We define $\triangleright$ as denoting the interval-dominance of one action over another and $\circ$ for when neither action interval-dominates the other. For a particular confidence interval and the fixed utilities given in Fig. 1, the actions b_h, b_t, and $\bar{b}$ can stand in the following relations:

(i) $b_h \triangleright b_t$ and $b_h \triangleright \bar{b}$;
(ii) $b_t \triangleright b_h$ and $b_t \triangleright \bar{b}$;
(iii) $\bar{b} \triangleright b_h$ and $\bar{b} \triangleright b_t$;
(iv) $b_h \triangleright b_t$ and $b_h \circ \bar{b}$;
(v) $b_h \circ b_t$ and $b_h \triangleright \bar{b}$;
(vi) $b_t \triangleright b_h$ and $b_t \circ \bar{b}$;
(vii) $b_h \circ b_t$ and $b_h \circ \bar{b}$.

Frequentist will choose b_h in case (i), b_t in case (ii), and $\bar{b}$ in case (iii). They will make random choices between b_h and $\bar{b}$ in case (iv), b_h and b_t in case (v), b_t and $\bar{b}$ in case (vi), and between all three actions in case (vii). Our decision algorithm for *Frequentist* in the code embodies this behaviour.[14]

3.3.3 The Williamsonian

Williamsonian is somewhat of a hybrid player. Consequently, after our discussion of the previous two players, there are few new notions involved in their reasoning. Williamson (2010) argues for three fundamental principles about probabilistic reasoning:

[13] However, in Kyburg and Teng (1999), if no action interval-dominates holding, their frequentist player will always choose to hold. Kyburg and Teng provide no justification for this feature of their frequentist player's decision algorithm, so we have interpreted it as a programming simplification, and one that was not helpful for our simulations.

[14] One might object that it is unfair to enable *Frequentist* to hold, while *Bayes* never makes this decision. This modelling choice has the advantage that our simulations are more comparable to those of Kyburg and Teng (1999). It is also acceptable from *Bayes*'s perspective, since hold is never strictly dominant—at best, *Bayes* is indifferent between this option and some other option. However, an anonymous referee points out that less conventional Bayesian reasoners will not necessarily share this perspective. In particular, "imprecise" Bayesians, who model agents using sets of probability functions, do not identify the expected value of not taking a bet to be the negative of the bet's expected value. In their model, the minimum price at which an agent is willing to sell a ticket and the minimum price at which they are willing to buy that ticket can differ, in a similar way to *Frequentist*. For some influential discussions and excellent expositions, see Levi (1974), Gilboa and Schmeidler (1989), Seidenfeld (2004), Troffaes (2007), Bradley (2017). We look forward to investigating the performance of such players in future research. In the present study, we address this issue using $\epsilon = 1$, which enables us to simulate what happens when we remove hold as an acceptable option for *Frequentist*.

1. **Probabilism**: Beliefs should be representable as credences satisfying the additive probability calculus.
2. **Calibration**: These credences should reflect any relevant knowledge about relative frequencies.
3. **Equivocation**: Subject to the constraints from Calibration, the credences should be maximally equivocal among the different possible states of the world.

The principle of Probabilism is a familiar Bayesian idea. For Calibration, Williamson endorses Kyburg's system of "Evidential Probability" (Wheeler and Williamson, 2011) and this system requires using confidence intervals where (as in our simulations) there is no background knowledge of the relevant conditional probabilities for events (Kyburg and Teng, 2001, p. 264). Consequently, *Williamsonian* will estimate confidence intervals in the same way as *Frequentist*.

However, Equivocation and Probabilism entail that *Williamsonian* will differ from *Frequentist* in the other parts of their learning procedure. To determine equivocal credences in a systematic manner, *Williamsonian* uses the "maximum entropy principle". Williamson gives the full formal connections between entropy maximisation and formal beliefs in Williamson (2010), building on research such as Jaynes (1957). The salient point is that, given a confidence inferred at some exogenously determined confidence level $(1-\alpha)$ and a set of $m \geq 2$ exhaustive and mutually exclusive states of the world that are consistent with the relative frequencies estimated using that confidence level, a Williamsonian will try to minimise the distance between their probabilities for each state of the world and the value $1/m$ implied by the Principle of Indifference.

Williamsonian's belief revision can be represented with a model $\mathrm{W} := \{\Omega, \Theta, \mathrm{H}, \mathrm{p}, \mathrm{k}_\alpha, \mathrm{c}, p_w\}$, where Ω is the set of possible outcomes of a single coin toss, Θ is the set of values representing all the possible coin biases, H is the set of possible histories that can be generated by $n > 0$ coin tosses, c is the *heads* event counting function, $\mathrm{p} : \Theta \rightarrow \Delta(\Theta)$ is a function which assigns a uniform probability distribution on Θ, k_α is a function which assigns a Clopper-Pearson interval to every history $\mathrm{h} \in \mathrm{H}$, and $p_w : \mathrm{H} \rightarrow X$, where $X := \left\{x \in \mathbb{R}_{\geq 0} : 0 \leq x \leq 1\right\}$, is a Williamsonian belief function which assigns, to every history $\mathrm{h} \in \mathrm{H}$, a belief $p_w(\mathrm{h})$, such that

$$p_w(\mathrm{h}) \in \mathrm{argmin}_{k \in \mathrm{k}_\alpha(\mathrm{h})} \left| k - \int_0^1 (\mathrm{p}(\theta)\,\theta)\,\mathrm{d}\theta \right|. \tag{11}$$

Since function p assigns a uniform distribution on Θ, we have a case where $\int_0^1 (\mathrm{p}(\theta)\,\theta)\,\mathrm{d}\theta = 1/2$. Thus, we can rewrite condition 11 as

$$p_w(\mathrm{h}) \in \mathrm{argmin}_{k \in \mathrm{k}_\alpha(\mathrm{h})} \left| k - \frac{1}{2} \right|. \tag{12}$$

Therefore, in our coin tossing problem, *Williamsonian* first estimates confidence intervals for *heads* and *tails* in general at a confidence level $(1-\alpha)$. Secondly, subject to this constraint, they set a Bayesian probability. For example, suppose that the Williamsonian has estimated a confidence interval [0.4, 0.6] for *heads*. This is consistent with assigning a probability of 0.5 for each particular toss landing *heads*, so

Williamsonian makes 0.5 their degree of belief. However, if the interval is [0.1, 0.4], then 0.5 is not consistent with their beliefs about the relative frequencies. Instead, 0.4 is the available value that maximises entropy, and hence *Williamsonian* makes 0.4 their degree of belief in each particular toss landing *heads*.

Williamsonian's choices can be characterised with choice model $\mathrm{C_w} := \{\Omega, \mathrm{H}, \mathrm{D}, \pi, p_w\}$, where Ω is the set of coin toss outcomes, H is the set of histories, D is the set of actions, π is the payoff function identical to that which we defined for *Bayes* and *Frequentist* players, and p_w is a Williamsonian belief function. *Williamsonian*'s expected payoff associated with some action $\mathrm{d} \in \mathrm{D}$ given some history $\mathrm{h} \in \mathrm{H}$ is

$$\mathbb{E}_\pi\left[\mathrm{d}|\mathrm{h}\right] = p_w\left(\mathrm{h}\right)\pi\left(\mathrm{d}, heads\right) + \left(1 - p_w\left(\mathrm{h}\right)\right)\pi\left(\mathrm{d}, tails\right). \tag{13}$$

We assume that *Williamsonian* is an expected payoff maximizer and thus, for any history $\mathrm{h} \in \mathrm{H}$, always chooses action $\mathrm{d} \in \mathrm{D}$, such that $\mathbb{E}_\pi\left[\mathrm{d}|\mathrm{h}\right] \geq \mathbb{E}_\pi\left[\bar{\mathrm{d}}|\mathrm{h}\right]$, for all $\bar{\mathrm{d}} \in \mathrm{D}$.

Despite the similarities in the learning procedures for *Williamsonian* and *Frequentist*, the principle of **Equivocation** leaves *Williamsonian* in a very different place, because the latter has a precise probability when they face each next coin toss. Therefore, *Williamsonian* makes decisions, for a particular set of credences, in an identical way to *Bayes*.

3.3.4 Sample

Our final player, *Sample*, serves two primary functions. Firstly, they can be interpreted as *Frequentist* in a situation where they are forced to make *precise* estimates of the relative frequencies of heads. In that case, *Frequentist* might estimate the general relative frequencies of heads to be the limit as $\alpha \to 0$, which is just the sample frequency. Thus, *Sample* may be interpreted as estimating that the long run frequency of heads is just the relative frequency in the total sample that they have observed so far. These estimates then guide *Sample*'s bets on each given toss. *Sample* also thereby has estimates for tails, given their knowledge of the coin tossing set-up. *Sample*'s second function is as a benchmark: if a statistical methodology's performance is inferior to just estimating a precise probability using the sample frequency, then this is a bad mark against that methodology.

Conceptually, there is nothing novel in *Sample*'s belief revision procedure. They can be interpreted in multiple ways: as a version of *Frequentist* with precise estimates of relative frequencies, as a quasi-Bayesian reasoner (similar to the "$\lambda = 0$" agent in Carnap, 1952) and so on. The divergence with *Frequentist* is mainly in how different sample sizes are treated: given a 0.5 sample frequency of heads and a sample of 4, *Sample* is certain that the limiting relative frequency of heads in the tosses is 0.5, whereas *Frequentist* has a very wide confidence interval around 0.5; given a 0.5 sample frequency of heads and a sample size of 1000, *Sample*'s beliefs are the same as when they had just 4 observations, whereas *Frequentist* has a very narrow confidence interval. Even this divergence can be reduced if we interpret *Sample* as a frequentist reasoner who uses the sample frequency as a decision-making tool, without interpreting the concomitant probability distributions as degrees of belief. This type of reasoner is

briefly discussed by Williamson (2007). The divergence with Bayesianism is that *Sample* has no initial priors, and thus does not update by conditionalization. While *Sample* is not a true Bayesian reasoner, in our simulations they will always have some sample data prior to making decisions, and therefore there is no need to specify the beliefs of *Sample* when they lack sample data.

For the choice behaviour of *Sample*, we model them as possessing a precise probability (interpreted either as a relative frequency estimate or as a non-Bayesian credence) for each possible outcome of each toss. Given this probability, it is natural for *Sample* to use Bayesian decision theory, because even many frequentists would say that such choices would be appropriate *if* one legitimately possessed the relevant relative frequency estimates, and *Sample* views themselves as having this information. Thus, for given credences, *Sample* makes decisions in the same way as *Bayes*. This common decision algorithm has the added advantage that *Sample* only differs from *Bayes* in their learning methods, which helps comparisons between the two players.

3.4 Coding and simulation architecture

Starting from the common decision problem and particular settings specified for each player in Sects. 3.2 and 3.3, we coded six different Python 3 scripts (version 3.8.1), based on the `statsmodel` econometric and statistical library (Seabold and Perktold, 2010) to perform the simulations. The first two scripts generated coin tosses and ticket prices. This data was subsequently employed as an input for the other four scripts. The output of the latter scripts gave us the cumulative monetary mean profits for each player, like those shown below in Tables 1, 2, 3, 4, 5 and 6 and, focusing on the Frequentist, the statistics of the conditions met by this player (see end of Sect. 3.3.2) in the simulations, as outlined in Tables 7 and 8. All simulations were performed on an Ubuntu Linux server powered by an 8 cores (16 threads) Intel Xeon (Skylake type) processor @ 2.2 GHz. High performance data multiprocessing meant that the overall computational time was approximately 1.5 hours. The simulation results were stored in 310,000 different `txt` files for about 3.5 gigabytes. The code is available from the authors upon request.

3.5 Variations

The games each consisted of a number of observations of coin tosses, which were used by the players to update their initial belief state. Once these beliefs were updated, players used their decision algorithms to choose an action. Players retained information from game-to-game within a particular simulation, but they did not retain any information from one simulation to another.

Table 1 Comparison of results for all player types

Coin bias	#Games	Sample	Bayesian		
		$	B (100, 1)	B (1, 1)	B (1, 100)
0.1	10	0.4068 ± 0.0082	0.0236 ± 0.0111	0.4060 ± 0.0081	0.4052 ± 0.0081
	25	0.4089 ± 0.0051	0.1898 ± 0.0066	0.4084 ± 0.0051	0.4078 ± 0.0050
	50	0.4097 ± 0.0035	0.2829 ± 0.0041	0.4094 ± 0.0034	0.4095 ± 0.0034
	100	0.4101 ± 0.0024	0.3405 ± 0.0027	0.4100 ± 0.0024	0.4097 ± 0.0024
	250	0.4103 ± 0.0016	0.3810 ± 0.0017	0.4103 ± 0.0016	0.4102 ± 0.0016
	500	0.4101 ± 0.0011	0.3952 ± 0.0012	0.4100 ± 0.0011	0.4100 ± 0.0011
	1000	0.4098 ± 0.0008	0.4024 ± 0.0008	0.4098 ± 0.0008	0.4097 ± 0.0008
0.3	10	0.2879 ± 0.0095	0.0573 ± 0.0112	0.2909 ± 0.0094	0.2518 ± 0.0101
	25	0.2883 ± 0.0060	0.1550 ± 0.0069	0.2896 ± 0.0060	0.2642 ± 0.0063
	50	0.2897 ± 0.0044	0.2120 ± 0.0047	0.2903 ± 0.0044	0.2758 ± 0.0044
	100	0.2898 ± 0.0031	0.2474 ± 0.0032	0.2902 ± 0.0031	0.2821 ± 0.0031
	250	0.2895 ± 0.0019	0.2715 ± 0.0019	0.2897 ± 0.0019	0.2865 ± 0.0019
	500	0.2900 ± 0.0014	0.2808 ± 0.0014	0.2901 ± 0.0014	0.2886 ± 0.0014
	1000	0.2895 ± 0.0010	0.2849 ± 0.0010	0.2896 ± 0.0010	0.2889 ± 0.0010
0.5	10	0.2508 ± 0.0102	0.1342 ± 0.0110	0.2525 ± 0.0102	0.1390 ± 0.0113
	25	0.2453 ± 0.0065	0.1844 ± 0.0067	0.2460 ± 0.0064	0.1831 ± 0.0065
	50	0.2466 ± 0.0048	0.2122 ± 0.0049	0.2470 ± 0.0048	0.2088 ± 0.0047
	100	0.2488 ± 0.0033	0.2293 ± 0.0035	0.2490 ± 0.0033	0.2281 ± 0.0033
	250	0.2495 ± 0.0021	0.2409 ± 0.0022	0.2495 ± 0.0021	0.2406 ± 0.0021
	500	0.2499 ± 0.0015	0.2457 ± 0.0015	0.2499 ± 0.0015	0.2454 ± 0.0015
	1000	0.2496 ± 0.0010	0.2476 ± 0.0010	0.2496 ± 0.0010	0.2473 ± 0.0010

Table 1 continued

Coin bias	#Games	Sample	Bayesian		
		$	B (100, 1)	B (1, 1)	B (1, 100)
0.7	10	0.2838 ± 0.0102	0.2538 ± 0.0105	0.2831 ± 0.0102	0.0610 ± 0.0111
	25	0.2888 ± 0.0063	0.2692 ± 0.0064	0.2889 ± 0.0063	0.1626 ± 0.0068
	50	0.2895 ± 0.0043	0.2779 ± 0.0044	0.2896 ± 0.0043	0.2154 ± 0.0047
	100	0.2898 ± 0.0031	0.2840 ± 0.0031	0.2898 ± 0.0031	0.2487 ± 0.0032
	250	0.2905 ± 0.0020	0.2882 ± 0.0020	0.2905 ± 0.0020	0.2728 ± 0.0020
	500	0.2903 ± 0.0014	0.2891 ± 0.0014	0.2903 ± 0.0014	0.2814 ± 0.0014
	1000	0.2898 ± 0.0010	0.2892 ± 0.0010	0.2898 ± 0.0010	0.2852 ± 0.0010
0.9	10	0.4119 ± 0.0082	0.4095 ± 0.0083	0.4102 ± 0.0082	0.0398 ± 0.0105
	25	0.4102 ± 0.0051	0.4089 ± 0.0051	0.4098 ± 0.0051	0.1967 ± 0.0064
	50	0.4091 ± 0.0036	0.4083 ± 0.0036	0.4089 ± 0.0036	0.2836 ± 0.0041
	100	0.4089 ± 0.0025	0.4086 ± 0.0025	0.4089 ± 0.0025	0.3395 ± 0.0027
	250	0.4092 ± 0.0016	0.4091 ± 0.0016	0.4092 ± 0.0016	0.3800 ± 0.0016
	500	0.4096 ± 0.0012	0.4095 ± 0.0012	0.4096 ± 0.0012	0.3945 ± 0.0012
	1000	0.4095 ± 0.0008	0.4094 ± 0.0008	0.4095 ± 0.0008	0.4018 ± 0.0008

Table 1 continued

Coin bias	#Games	Frequentist			Williamsonian		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.4040 ± 0.0082	0.3873 ± 0.0082	0.3724 ± 0.0085	0.4010 ± 0.0082	0.3742 ± 0.0087	0.3585 ± 0.0089
	25	0.4080 ± 0.0050	0.3997 ± 0.0051	0.3919 ± 0.0051	0.4058 ± 0.0051	0.3933 ± 0.0053	0.3855 ± 0.0054
	50	0.4093 ± 0.0034	0.4051 ± 0.0034	0.4013 ± 0.0035	0.4083 ± 0.0035	0.4019 ± 0.0036	0.3978 ± 0.0036
	100	0.4098 ± 0.0024	0.4074 ± 0.0025	0.4054 ± 0.0024	0.4093 ± 0.0024	0.4056 ± 0.0025	0.4034 ± 0.0025
	250	0.4101 ± 0.0016	0.4091 ± 0.0016	0.4082 ± 0.0015	0.4099 ± 0.0016	0.4084 ± 0.0016	0.4074 ± 0.0016
	500	0.4100 ± 0.0011	0.4094 ± 0.0011	0.4090 ± 0.0011	0.4099 ± 0.0011	0.4092 ± 0.0011	0.4086 ± 0.0011
	1000	0.4097 ± 0.0008	0.4095 ± 0.0008	0.4092 ± 0.0008	0.4097 ± 0.0008	0.4094 ± 0.0008	0.4091 ± 0.0008
0.3	10	0.2865 ± 0.0095	0.2596 ± 0.0097	0.2446 ± 0.0096	0.2857 ± 0.0096	0.2701 ± 0.0098	0.2652 ± 0.0099
	25	0.2873 ± 0.0060	0.2744 ± 0.0060	0.2648 ± 0.0060	0.2883 ± 0.0061	0.2792 ± 0.0061	0.2749 ± 0.0062
	50	0.2893 ± 0.0044	0.2819 ± 0.0043	0.2750 ± 0.0043	0.2900 ± 0.0045	0.2843 ± 0.0045	0.2809 ± 0.0044
	100	0.2897 ± 0.0031	0.2851 ± 0.0031	0.2814 ± 0.0030	0.2901 ± 0.0031	0.2869 ± 0.0031	0.2846 ± 0.0031
	250	0.2893 ± 0.0019	0.2873 ± 0.0019	0.2856 ± 0.0019	0.2894 ± 0.0020	0.2878 ± 0.0019	0.2868 ± 0.0019
	500	0.2898 ± 0.0014	0.2887 ± 0.0014	0.2879 ± 0.0014	0.2899 ± 0.0014	0.2889 ± 0.0014	0.2884 ± 0.0014
	1000	0.2895 ± 0.0010	0.2889 ± 0.0010	0.2884 ± 0.0010	0.2894 ± 0.0010	0.2889 ± 0.0010	0.2887 ± 0.0010

Table 1 continued

Coin bias	#Games	Frequentist			Williamsonian		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2440 ± 0.0099	0.2219 ± 0.0101	0.2093 ± 0.0098	0.2548 ± 0.0103	0.2569 ± 0.0103	0.2576 ± 0.0103
	25	0.2434 ± 0.0063	0.2321 ± 0.0063	0.2251 ± 0.0063	0.2470 ± 0.0065	0.2473 ± 0.0065	0.2482 ± 0.0065
	50	0.2455 ± 0.0047	0.2388 ± 0.0047	0.2350 ± 0.0047	0.2476 ± 0.0048	0.2481 ± 0.0048	0.2487 ± 0.0048
	100	0.2477 ± 0.0033	0.2441 ± 0.0033	0.2414 ± 0.0033	0.2491 ± 0.0033	0.2496 ± 0.0034	0.2499 ± 0.0034
	250	0.2489 ± 0.0021	0.2470 ± 0.0021	0.2455 ± 0.0021	0.2495 ± 0.0021	0.2500 ± 0.0021	0.2501 ± 0.0021
	500	0.2496 ± 0.0014	0.2486 ± 0.0015	0.2476 ± 0.0015	0.2500 ± 0.0015	0.2501 ± 0.0015	0.2502 ± 0.0015
	1000	0.2495 ± 0.0010	0.2489 ± 0.0010	0.2483 ± 0.0010	0.2497 ± 0.0010	0.2498 ± 0.0010	0.2498 ± 0.0010
0.7	10	0.2798 ± 0.0100	0.2609 ± 0.0098	0.2519 ± 0.0098	0.2801 ± 0.0102	0.2658 ± 0.0105	0.2595 ± 0.0103
	25	0.2879 ± 0.0063	0.2784 ± 0.0062	0.2736 ± 0.0062	0.2865 ± 0.0063	0.2782 ± 0.0064	0.2737 ± 0.0064
	50	0.2891 ± 0.0043	0.2837 ± 0.0043	0.2803 ± 0.0043	0.2882 ± 0.0043	0.2832 ± 0.0044	0.2802 ± 0.0044
	100	0.2896 ± 0.0031	0.2869 ± 0.0031	0.2844 ± 0.0031	0.2891 ± 0.0031	0.2865 ± 0.0031	0.2848 ± 0.0031
	250	0.2903 ± 0.0020	0.2890 ± 0.0020	0.2876 ± 0.0020	0.2903 ± 0.0020	0.2891 ± 0.0020	0.2880 ± 0.0019
	500	0.2903 ± 0.0014	0.2896 ± 0.0014	0.2888 ± 0.0014	0.2903 ± 0.0014	0.2895 ± 0.0014	0.2889 ± 0.0014
	1000	0.2898 ± 0.0010	0.2893 ± 0.0010	0.2888 ± 0.0010	0.2898 ± 0.0010	0.2893 ± 0.0010	0.2890 ± 0.0010
0.9	10	0.4086 ± 0.0082	0.3976 ± 0.0083	0.3842 ± 0.0084	0.4105 ± 0.0083	0.4046 ± 0.0083	0.3974 ± 0.0085
	25	0.4093 ± 0.0051	0.4030 ± 0.0051	0.3977 ± 0.0052	0.4099 ± 0.0051	0.4066 ± 0.0051	0.4027 ± 0.0052
	50	0.4085 ± 0.0036	0.4049 ± 0.0036	0.4024 ± 0.0036	0.4090 ± 0.0036	0.4068 ± 0.0036	0.4049 ± 0.0037
	100	0.4086 ± 0.0025	0.4067 ± 0.0025	0.4053 ± 0.0025	0.4089 ± 0.0025	0.4078 ± 0.0025	0.4068 ± 0.0025
	250	0.4090 ± 0.0016	0.4082 ± 0.0016	0.4076 ± 0.0016	0.4092 ± 0.0016	0.4087 ± 0.0016	0.4081 ± 0.0016
	500	0.4094 ± 0.0012	0.4090 ± 0.0012	0.4087 ± 0.0012	0.4096 ± 0.0012	0.4093 ± 0.0012	0.4090 ± 0.0012
	1000	0.4094 ± 0.0008	0.4091 ± 0.0008	0.4089 ± 0.0008	0.4095 ± 0.0008	0.4093 ± 0.0008	0.4091 ± 0.0008

Bayesian with different initial priors and Frequentist and Williamsonian with different levels of α. 1000 simulations (each one including 1000 games, each one including 10 $(9+1)$ draws) have been run for all considered coin biases. Assuming normal distribution for the outcomes of our experiment, all results report the upper and lower confidence level for the mean value of the net profit in USD per game at 5% significance level computed over 1000 simulations, i.e. the mean value plus/minus the related standard error multiplied by $z = 1.96$

Table 2 Interval analysis of Table 1

Coin bias	#Games	Sample	Bayesian			Frequentist			Williamsonian		
		$	B (100, 1)	B (1, 1)	B (1, 100)	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.3986	0.0347*	0.4141	0.4133	0.4122	0.3955*	0.3809*	0.4092	0.3829*	0.3674*
	25	0.4038	0.1964*	0.4135	0.4128	0.4130	0.4048	0.3970*	0.4109	0.3986*	0.3909*
	50	0.4062	0.2870*	0.4128	0.4129	0.4127	0.4085	0.4048*	0.4118	0.4055*	0.4014*
	100	0.4077	0.3432*	0.4124	0.4121	0.4122	0.4099	0.4078	0.4117	0.4081	0.4059*
	250	0.4087	0.3827*	0.4119	0.4118	0.4117	0.4107	0.4097	0.4115	0.4100	0.4090
	500	0.4090	0.3964*	0.4111	0.4111	0.4111	0.4105	0.4101	0.4110	0.4103	0.4097
	1000	0.4090	0.4032*	0.4106	0.4105	0.4105	0.4103	0.4100	0.4105	0.4102	0.4099
0.3	10	0.2784	0.0685*	0.3003	0.2619*	0.2960	0.2693*	0.2542*	0.2953	0.2799	0.2751*
	25	0.2823	0.1619*	0.2956	0.2705*	0.2933	0.2804*	0.2708*	0.2944	0.2853	0.2811*
	50	0.2853	0.2167*	0.2947	0.2802*	0.2937	0.2862	0.2793*	0.2945	0.2888	0.2853
	100	0.2867	0.2505*	0.2933	0.2852*	0.2928	0.2882	0.2844*	0.2932	0.2900	0.2877
	250	0.2876	0.2734*	0.2916	0.2884	0.2912	0.2892	0.2875*	0.2914	0.2897	0.2887
	500	0.2886	0.2822*	0.2915	0.2900	0.2912	0.2901	0.2893	0.2913	0.2903	0.2898
	1000	0.2885	0.2859*	0.2906	0.2899	0.2905	0.2899	0.2894	0.2904	0.2899	0.2897

Table 2 continued

Coin bias	#Games	Sample	Bayesian			Frequentist			Williamsonian		
		$	B (100, 1)	B (1, 1)	B (1, 100)	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2406	0.1452*	0.2627	0.1503*	0.2539	0.2320*	0.2191*	0.2651	0.2672	0.2679
	25	0.2388	0.1911*	0.2524	0.1896*	0.2494	0.2384*	0.2314*	0.2535	0.2538	0.2548
	50	0.2418	0.2147*	0.2518	0.2135*	0.2502	0.2435	0.2397*	0.2524	0.2529	0.2535
	100	0.2455	0.2328*	0.2523	0.2314*	0.2510	0.2474	0.2447*	0.2524	0.2530	0.2533
	250	0.2474	0.2431*	0.2516	0.2427*	0.2510	0.2491	0.2476	0.2516	0.2521	0.2522
	500	0.2484	0.2474*	0.2514	0.2469*	0.2510	0.2501	0.2491	0.2515	0.2516	0.2517
	1000	0.2486	0.2486	0.2506	0.2483*	0.2505	0.2499	0.2493	0.2507	0.2508	0.2508
0.7	10	0.2736	0.2643*	0.2933	0.0721*	0.2898	0.2707*	0.2617*	0.2913	0.2763	0.2698*
	25	0.2825	0.2756*	0.2952	0.1694*	0.2942	0.2846	0.2798*	0.2989	0.2846	0.2801*
	50	0.2852	0.2823*	0.2939	0.2203*	0.2934	0.2880*	0.2846*	0.2925	0.2876	0.2846*
	100	0.2867	0.2871	0.2929	0.2519*	0.2927	0.2900	0.2875	0.2922	0.2896	0.2879
	250	0.2885	0.2902	0.2925	0.2748*	0.2923	0.2910	0.2896	0.2923	0.2911	0.2899
	500	0.2889	0.2905	0.2917	0.2828*	0.2917	0.2910	0.2902	0.2917	0.2909	0.2903
	1000	0.2888	0.2902	0.2908	0.2862*	0.2908	0.2903	0.2898	0.2908	0.2903	0.2900
0.9	10	0.4037	0.4178	0.4184	0.0503*	0.4168	0.4059	0.3926*	0.4188	0.4129	0.4059
	25	0.4051	0.4140	0.4149	0.2031*	0.4144	0.4081	0.4029*	0.4150	0.4117	0.4079
	50	0.4055	0.4119	0.4125	0.2877*	0.4121	0.4085	0.4060	0.4126	0.4104	0.4086
	100	0.4064	0.4111	0.4114	0.3422*	0.4111	0.4092	0.4078	0.4114	0.4103	0.4093
	250	0.4076	0.4107	0.4108	0.3816*	0.4106	0.4098	0.4092	0.4108	0.4103	0.4097
	500	0.4084	0.4107	0.4108	0.3957*	0.4106	0.4102	0.4099	0.4108	0.4105	0.4102
	1000	0.4087	0.4102	0.4103	0.4026*	0.4102	0.4099	0.4026*	0.4103	0.4101	0.4099

The **Sample** column represents the minimum mean value of the net profit in USD of the *Sample* player within the confidence interval. The **Bayesian**, **Frequentist** and **Williamsonian** columns represent the maximum mean values of the net profit in USD of the *Bayesian*, *Frequentist* and *Williamsonian* players within the confidence intervals. Stars indicate cases where **Sample** values exceed the **Bayesian**, **Frequentist** or **Williamsonian** values

Table 3 This table presents the same kind of information defined in Table 1 but with the same draws halved

Coin bias	#Games	Sample	Bayesian		
		$	B (100, 1)	B (1, 1)	B (1, 100)
0.1	10	0.4030 ± 0.0083	-0.1157 ± 0.0110	0.4010 ± 0.0083	0.4036 ± 0.0081
	25	0.4071 ± 0.0051	0.0590 ± 0.0071	0.4063 ± 0.0051	0.4065 ± 0.0050
	50	0.4093 ± 0.0035	0.1860 ± 0.0047	0.4088 ± 0.0035	0.4088 ± 0.0034
	100	0.4098 ± 0.0024	0.2795 ± 0.0030	0.4095 ± 0.0024	0.4094 ± 0.0024
	250	0.4101 ± 0.0016	0.3522 ± 0.0017	0.4100 ± 0.0016	0.4100 ± 0.0016
	500	0.4099 ± 0.0011	0.3801 ± 0.0012	0.4099 ± 0.0011	0.4099 ± 0.0011
	1000	0.4098 ± 0.0008	0.3946 ± 0.0008	0.4097 ± 0.0008	0.4097 ± 0.0008
0.3	10	0.2784 ± 0.0097	-0.0297 ± 0.0106	0.2816 ± 0.0097	0.2358 ± 0.0101
	25	0.2830 ± 0.0061	0.0785 ± 0.0069	0.2843 ± 0.0061	0.2508 ± 0.0064
	50	0.2861 ± 0.0044	0.1544 ± 0.0048	0.2868 ± 0.0044	0.2663 ± 0.0044
	100	0.2882 ± 0.0031	0.2103 ± 0.0032	0.2885 ± 0.0031	0.2760 ± 0.0031
	250	0.2888 ± 0.0019	0.2536 ± 0.0020	0.2889 ± 0.0019	0.2836 ± 0.0019
	500	0.2897 ± 0.0014	0.2711 ± 0.0014	0.2898 ± 0.0014	0.2870 ± 0.0014
	1000	0.2894 ± 0.0010	0.2798 ± 0.0010	0.2894 ± 0.0010	0.2880 ± 0.0010

Table 3 continued

Coin bias	#Games	Sample	Bayesian		
		$	B (100, 1)	B (1, 1)	B (1, 100)
0.5	10	0.2379 ± 0.0104	0.0927 ± 0.0113	0.2451 ± 0.0105	0.0928 ± 0.0115
	25	0.2389 ± 0.0065	0.1483 ± 0.0069	0.2421 ± 0.0065	0.1432 ± 0.0067
	50	0.2433 ± 0.0048	0.1849 ± 0.0051	0.2449 ± 0.0048	0.1788 ± 0.0047
	100	0.2467 ± 0.0034	0.2125 ± 0.0035	0.2475 ± 0.0034	0.2086 ± 0.0033
	250	0.2485 ± 0.0021	0.2325 ± 0.0022	0.2488 ± 0.0021	0.2310 ± 0.0021
	500	0.2493 ± 0.0015	0.2411 ± 0.0015	0.2495 ± 0.0015	0.2402 ± 0.0014
	1000	0.2493 ± 0.0010	0.2452 ± 0.0010	0.2494 ± 0.0010	0.2446 ± 0.0010
0.7	10	0.2761 ± 0.0104	0.2396 ± 0.0103	0.2801 ± 0.0104	-0.0246 ± 0.0112
	25	0.2853 ± 0.0064	0.2555 ± 0.0064	0.2873 ± 0.0064	0.0819 ± 0.0071
	50	0.2877 ± 0.0044	0.2684 ± 0.0044	0.2888 ± 0.0044	0.1548 ± 0.0049
	100	0.2890 ± 0.0031	0.2783 ± 0.0031	0.2896 ± 0.0031	0.2113 ± 0.0033
	250	0.2901 ± 0.0020	0.2852 ± 0.0019	0.2904 ± 0.0020	0.2555 ± 0.0020
	500	0.2901 ± 0.0014	0.2875 ± 0.0014	0.2903 ± 0.0014	0.2723 ± 0.0014
	1000	0.2897 ± 0.0010	0.2884 ± 0.0010	0.2898 ± 0.0010	0.2805 ± 0.0010

Table 3 continued

Coin bias	#Games	Sample	Bayesian		
		$	B (100, 1)	B (1, 1)	B (1, 100)
0.9	10	0.4074 ± 0.0084	0.4085 ± 0.0081	0.4063 ± 0.0085	-0.1085 ± 0.0107
	25	0.4079 ± 0.0051	0.4076 ± 0.0051	0.4074 ± 0.0052	0.0647 ± 0.0068
	50	0.4079 ± 0.0036	0.4075 ± 0.0036	0.4078 ± 0.0037	0.1864 ± 0.0046
	100	0.4083 ± 0.0025	0.4079 ± 0.0025	0.4082 ± 0.0025	0.2783 ± 0.0029
	250	0.4089 ± 0.0016	0.4087 ± 0.0016	0.4089 ± 0.0016	0.3512 ± 0.0017
	500	0.4094 ± 0.0012	0.4093 ± 0.0012	0.4094 ± 0.0012	0.3792 ± 0.0012
	1000	0.4094 ± 0.0008	0.4093 ± 0.0008	0.4094 ± 0.0008	0.3940 ± 0.0008

Coin bias	#Games	Frequentist			Williamsonian		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.3951 ± 0.0084	0.3638 ± 0.0087	0.3442 ± 0.0088	0.3882 ± 0.0085	0.3453 ± 0.0091	0.3249 ± 0.0093
	25	0.4037 ± 0.0051	0.3887 ± 0.0053	0.3797 ± 0.0053	0.4012 ± 0.0052	0.3793 ± 0.0054	0.3684 ± 0.0054
	50	0.4076 ± 0.0035	0.3991 ± 0.0035	0.3945 ± 0.0035	0.4063 ± 0.0035	0.3943 ± 0.0036	0.3879 ± 0.0036
	100	0.4088 ± 0.0025	0.4041 ± 0.0025	0.4016 ± 0.0024	0.4081 ± 0.0025	0.4014 ± 0.0025	0.3979 ± 0.0025
	250	0.4097 ± 0.0016	0.4075 ± 0.0016	0.4066 ± 0.0015	0.4093 ± 0.0016	0.4066 ± 0.0016	0.4049 ± 0.0016
	500	0.4097 ± 0.0011	0.4087 ± 0.0011	0.4080 ± 0.0011	0.4096 ± 0.0011	0.4082 ± 0.0011	0.4072 ± 0.0011
	1000	0.4097 ± 0.0008	0.4091 ± 0.0008	0.4087 ± 0.0008	0.4096 ± 0.0008	0.4088 ± 0.0008	0.4084 ± 0.0008
0.3	10	0.2746 ± 0.0094	0.2407 ± 0.0097	0.2060 ± 0.0098	0.2764 ± 0.0096	0.2606 ± 0.0098	0.2523 ± 0.0100
	25	0.2811 ± 0.0060	0.2620 ± 0.0061	0.2436 ± 0.0059	0.2836 ± 0.0061	0.2709 ± 0.0062	0.2630 ± 0.0062
	50	0.2848 ± 0.0044	0.2736 ± 0.0044	0.2631 ± 0.0042	0.2864 ± 0.0044	0.2781 ± 0.0044	0.2723 ± 0.0044
	100	0.2870 ± 0.0031	0.2805 ± 0.0031	0.2742 ± 0.0030	0.2880 ± 0.0031	0.2826 ± 0.0031	0.2789 ± 0.0031
	250	0.2882 ± 0.0019	0.2854 ± 0.0019	0.2820 ± 0.0019	0.2885 ± 0.0019	0.2859 ± 0.0019	0.2840 ± 0.0019
	500	0.2894 ± 0.0014	0.2877 ± 0.0014	0.2857 ± 0.0014	0.2894 ± 0.0014	0.2878 ± 0.0014	0.2866 ± 0.0014
	1000	0.2892 ± 0.0010	0.2883 ± 0.0010	0.2872 ± 0.0010	0.2892 ± 0.0010	0.2883 ± 0.0010	0.2876 ± 0.0010

Table 3 continued

Coin bias	#Games	Frequentist			Williamsonian		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2277 ± 0.0101	0.1953 ± 0.0100	0.1692 ± 0.0096	0.2473 ± 0.0106	0.2574 ± 0.0103	0.2577 ± 0.0103
	25	0.2361 ± 0.0064	0.2161 ± 0.0064	0.2007 ± 0.0063	0.2430 ± 0.0066	0.2474 ± 0.0065	0.2482 ± 0.0065
	50	0.2414 ± 0.0048	0.2291 ± 0.0047	0.2194 ± 0.0047	0.2451 ± 0.0049	0.2482 ± 0.0048	0.2487 ± 0.0048
	100	0.2456 ± 0.0034	0.2380 ± 0.0033	0.2326 ± 0.0032	0.2475 ± 0.0034	0.2495 ± 0.0034	0.2499 ± 0.0034
	250	0.2478 ± 0.0021	0.2441 ± 0.0021	0.2411 ± 0.0021	0.2489 ± 0.0021	0.2498 ± 0.0021	0.2500 ± 0.0021
	500	0.2490 ± 0.0015	0.2469 ± 0.0015	0.2451 ± 0.0014	0.2496 ± 0.0015	0.2501 ± 0.0015	0.2502 ± 0.0015
	1000	0.2492 ± 0.0010	0.2480 ± 0.0010	0.2470 ± 0.0010	0.2494 ± 0.0010	0.2497 ± 0.0010	0.2498 ± 0.0010
0.7	10	0.2655 ± 0.0102	0.2303 ± 0.0099	0.2103 ± 0.0101	0.2705 ± 0.0103	0.2522 ± 0.0107	0.2453 ± 0.0107
	25	0.2800 ± 0.0064	0.2613 ± 0.0061	0.2481 ± 0.0062	0.2815 ± 0.0064	0.2677 ± 0.0065	0.2593 ± 0.0065
	50	0.2846 ± 0.0044	0.2728 ± 0.0042	0.2649 ± 0.0042	0.2853 ± 0.0043	0.2766 ± 0.0044	0.2710 ± 0.0044
	100	0.2874 ± 0.0031	0.2803 ± 0.0031	0.2759 ± 0.0030	0.2875 ± 0.0031	0.2827 ± 0.0031	0.2798 ± 0.0031
	250	0.2894 ± 0.0020	0.2860 ± 0.0020	0.2836 ± 0.0019	0.2897 ± 0.0020	0.2869 ± 0.0019	0.2853 ± 0.0019
	500	0.2898 ± 0.0014	0.2879 ± 0.0014	0.2864 ± 0.0014	0.2899 ± 0.0014	0.2882 ± 0.0014	0.2873 ± 0.0014
	1000	0.2895 ± 0.0010	0.2885 ± 0.0009	0.2877 ± 0.0010	0.2896 ± 0.0010	0.2887 ± 0.0010	0.2882 ± 0.0010
0.9	10	0.4004 ± 0.0084	0.3697 ± 0.0086	0.3447 ± 0.0088	0.4032 ± 0.0083	0.3872 ± 0.0086	0.3733 ± 0.0088
	25	0.4048 ± 0.0051	0.3902 ± 0.0053	0.3784 ± 0.0053	0.4062 ± 0.0051	0.3986 ± 0.0052	0.3924 ± 0.0053
	50	0.4066 ± 0.0037	0.3979 ± 0.0037	0.3921 ± 0.0037	0.4070 ± 0.0036	0.4029 ± 0.0037	0.3992 ± 0.0037
	100	0.4075 ± 0.0025	0.4028 ± 0.0025	0.3994 ± 0.0025	0.4078 ± 0.0025	0.4054 ± 0.0025	0.4033 ± 0.0025
	250	0.4086 ± 0.0016	0.4065 ± 0.0016	0.4050 ± 0.0016	0.4087 ± 0.0016	0.4076 ± 0.0016	0.4065 ± 0.0016
	500	0.4092 ± 0.0012	0.4082 ± 0.0011	0.4073 ± 0.0012	0.4093 ± 0.0012	0.4087 ± 0.0012	0.4081 ± 0.0012
	1000	0.4092 ± 0.0008	0.4087 ± 0.0008	0.4082 ± 0.0008	0.4093 ± 0.0008	0.4089 ± 0.0008	0.4086 ± 0.0008

It shows the outcomes produced by 1000 simulations, each one including 1000 games, each one including 5 (4 + 1) draws, all belonging to the same sequences analyzed in Table 1

Table 4 Interval analysis of Table 3

Coin bias	#Games	Sample	Bayesian			Frequentist			Williamsonian		
		$	B (100, 1)	B (1, 1)	B (1, 100)	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.3947	-0.1047*	0.4093	0.4117	0.4035	0.3725*	0.3530*	0.3967	0.3544*	0.3342*
	25	0.4020	0.0661*	0.4114	0.4115	0.4088	0.3940*	0.3850*	0.4064	0.3847*	0.3738*
	50	0.4058	0.1907*	0.4123	0.4122	0.4111	0.4026*	0.3980*	0.4098	0.3979*	0.3915*
	100	0.4074	0.2825*	0.4119	0.4118	0.4113	0.4066*	0.4040*	0.4106	0.4039*	0.4004*
	250	0.4085	0.3537*	0.4116	0.4116	0.4113	0.4091	0.4081*	0.4109	0.4082*	0.4065*
	500	0.4088	0.3813*	0.4110	0.4110	0.4108	0.4098	0.4091	0.4107	0.4093	0.4083
	1000	0.4090	0.3954*	0.4105	0.4105	0.4105	0.4099	0.4095	0.4104	0.4096	0.4092
0.3	10	0.2687	-0.0191*	0.2913	0.2459*	0.2840	0.2504*	0.2158*	0.2860	0.2704	0.2623*
	25	0.2769	0.0854*	0.2904	0.2572*	0.2871	0.2681*	0.2495*	0.2897	0.2771	0.2692*
	50	0.2817	0.1592*	0.2912	0.2707*	0.2892	0.2780*	0.2673*	0.2908	0.2825	0.2767*
	100	0.2851	0.2135*	0.2916	0.2791*	0.2901	0.2836*	0.2772*	0.2911	0.2857	0.2820*
	250	0.2869	0.2556*	0.2908	0.2855*	0.2901	0.2873	0.2839*	0.2904	0.2878	0.2859*
	500	0.2883	0.2725*	0.2912	0.2884	0.2911	0.2891	0.2871*	0.2908	0.2892	0.2880*
	1000	0.2884	0.2808*	0.2904	0.2890	0.2902	0.2893	0.2882*	0.2902	0.2893	0.2886

Table 4 continued

Coin bias	#Games	Sample	Bayesian			Frequentist			Williamsonian		
		$	B (100, 1)	B (1, 1)	B (1, 100)	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2275	0.1040*	0.2556	0.1033*	0.2378	0.2053*	0.1788*	0.2579	0.2677	0.2680
	25	0.2324	0.1552*	0.2486	0.1499*	0.2425	0.2225*	0.2070*	0.2496	0.2539	0.2547
	50	0.2385	0.1900*	0.2497	0.1835*	0.2462	0.2338*	0.2241*	0.2500	0.2530	0.2518
	100	0.2433	0.2160*	0.2509	0.2119*	0.2490	0.2413*	0.2358*	0.2509	0.2529	0.2533
	250	0.2464	0.2346*	0.2509	0.2331*	0.2499	0.2462*	0.2432*	0.2510	0.2519	0.2521
	500	0.2478	0.2426*	0.2510	0.2416*	0.2505	0.2484	0.2465*	0.2511	0.2516	0.2517
	1000	0.2483	0.2462*	0.2504	0.2456*	0.2502	0.2490	0.2480*	0.2504	0.2507	0.2508
0.7	10	0.2657	0.2499*	0.2905	-0.0134*	0.2757	0.2402*	0.2204*	0.2808	0.2629*	0.2560*
	25	0.2789	0.2619*	0.2937	0.0890*	0.2864	0.2674*	0.2543*	0.2879	0.2742*	0.2658*
	50	0.2833	0.2728*	0.2932	0.1597*	0.2890	0.2770*	0.2691*	0.2896	0.2810*	0.2754*
	100	0.2859	0.2814*	0.2927	0.2146*	0.2905	0.2834*	0.2789*	0.2906	0.2858*	0.2829*
	250	0.2881	0.2871*	0.2924	0.2575*	0.2914	0.2880*	0.2855*	0.2911	0.2888	0.2872*
	500	0.2887	0.2889	0.2917	0.2737*	0.2912	0.2893	0.2878*	0.2913	0.2896	0.2887
	1000	0.2887	0.2894	0.2908	0.2815*	0.2905	0.2894	0.2887	0.2906	0.2897	0.2892
0.9	10	0.3990	0.4166	0.4148	-0.0978*	0.4088	0.3782*	0.3535*	0.4115	0.3958*	0.3821*
	25	0.4028	0.4127	0.4126	0.0715*	0.4099	0.3955*	0.3837*	0.4113	0.4038	0.3977*
	50	0.4043	0.4111	0.4115	0.1910*	0.4103	0.4016*	0.3958*	0.4106	0.4066	0.4029*
	100	0.4058	0.4104	0.4107	0.2812*	0.4100	0.4053*	0.4019*	0.4103	0.4079	0.4058
	250	0.4073	0.4103	0.4105	0.3529*	0.4102	0.4081	0.4066*	0.4103	0.4092	0.4081
	500	0.4082	0.4105	0.4106	0.3804*	0.4104	0.4093	0.4085	0.4105	0.4099	0.4093
	1000	0.4086	0.4101	0.4102	0.3948*	0.4100	0.4095	0.4090	0.4101	0.4097	0.4094

The **Sample** column represents the minimum mean value of the net profit in USD of the *Sample* player within the confidence interval. The **Bayesian**, **Frequentist** and **Williamsonian** columns represent the maximum mean values of the net profit in USD of the *Bayesian*, *Frequentist* and *Williamsonian* players within the confidence intervals. Stars indicate cases where **Sample** values exceed the **Bayesian**, **Frequentist** or **Williamsonian** values

Table 5 This table presents the same kind of information defined in Table 1 but only for the Frequentist player and its penalty variations (parameter ϵ)

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.4040 ± 0.0082	0.3873 ± 0.0082	0.3724 ± 0.0085	0.4027 ± 0.0082	0.3839 ± 0.0084	0.3689 ± 0.0086
	25	0.4080 ± 0.0050	0.3997 ± 0.0051	0.3919 ± 0.0051	0.4072 ± 0.0050	0.3979 ± 0.0051	0.3904 ± 0.0052
	50	0.4093 ± 0.0034	0.4051 ± 0.0034	0.4013 ± 0.0035	0.4092 ± 0.0034	0.4039 ± 0.0035	0.4000 ± 0.0036
	100	0.4098 ± 0.0024	0.4074 ± 0.0025	0.4054 ± 0.0024	0.4098 ± 0.0024	0.4067 ± 0.0025	0.4047 ± 0.0025
	250	0.4101 ± 0.0016	0.4091 ± 0.0016	0.4082 ± 0.0015	0.4101 ± 0.0016	0.4088 ± 0.0016	0.4080 ± 0.0016
	500	0.4100 ± 0.0011	0.4094 ± 0.0011	0.4090 ± 0.0011	0.4099 ± 0.0011	0.4093 ± 0.0011	0.4089 ± 0.0011
	1000	0.4097 ± 0.0008	0.4095 ± 0.0008	0.4092 ± 0.0008	0.4097 ± 0.0008	0.4094 ± 0.0008	0.4091 ± 0.0008
0.3	10	0.2865 ± 0.0095	0.2596 ± 0.0097	0.2446 ± 0.0096	0.2836 ± 0.0096	0.2568 ± 0.0094	0.2396 ± 0.0096
	25	0.2873 ± 0.0060	0.2744 ± 0.0060	0.2648 ± 0.0060	0.2861 ± 0.0061	0.2719 ± 0.0060	0.2610 ± 0.0061
	50	0.2893 ± 0.0044	0.2819 ± 0.0043	0.2750 ± 0.0043	0.2888 ± 0.0044	0.2796 ± 0.0044	0.2734 ± 0.0043
	100	0.2897 ± 0.0031	0.2851 ± 0.0031	0.2814 ± 0.0030	0.2895 ± 0.0032	0.2839 ± 0.0031	0.2806 ± 0.0030
	250	0.2893 ± 0.0019	0.2873 ± 0.0019	0.2856 ± 0.0019	0.2894 ± 0.0019	0.2869 ± 0.0019	0.2852 ± 0.0019
	500	0.2898 ± 0.0014	0.2887 ± 0.0014	0.2879 ± 0.0014	0.2899 ± 0.0014	0.2885 ± 0.0014	0.2876 ± 0.0014
	1000	0.2895 ± 0.0010	0.2889 ± 0.0010	0.2884 ± 0.0010	0.2895 ± 0.0010	0.2887 ± 0.0010	0.2883 ± 0.0010
0.5	10	0.2440 ± 0.0099	0.2219 ± 0.0101	0.2093 ± 0.0098	0.2428 ± 0.0101	0.2193 ± 0.0101	0.1979 ± 0.0102
	25	0.2434 ± 0.0063	0.2321 ± 0.0063	0.2251 ± 0.0063	0.2432 ± 0.0063	0.2295 ± 0.0064	0.2177 ± 0.0063
	50	0.2455 ± 0.0047	0.2388 ± 0.0047	0.2350 ± 0.0047	0.2454 ± 0.0047	0.2370 ± 0.0048	0.2301 ± 0.0046
	100	0.2477 ± 0.0033	0.2441 ± 0.0033	0.2414 ± 0.0033	0.2481 ± 0.0033	0.2427 ± 0.0034	0.2387 ± 0.0033
	250	0.2489 ± 0.0021	0.2470 ± 0.0021	0.2455 ± 0.0021	0.2491 ± 0.0021	0.2463 ± 0.0021	0.2444 ± 0.0021
	500	0.2496 ± 0.0014	0.2486 ± 0.0015	0.2476 ± 0.0015	0.2497 ± 0.0015	0.2483 ± 0.0015	0.2471 ± 0.0015
	1000	0.2495 ± 0.0010	0.2489 ± 0.0010	0.2483 ± 0.0010	0.2495 ± 0.0010	0.2489 ± 0.0010	0.2480 ± 0.0010

Table 5 continued

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.7	10	0.2798 ± 0.0100	0.2609 ± 0.0098	0.2519 ± 0.0098	0.2775 ± 0.0101	0.2571 ± 0.0100	0.2414 ± 0.0102
	25	0.2879 ± 0.0063	0.2784 ± 0.0062	0.2736 ± 0.0062	0.2853 ± 0.0062	0.2739 ± 0.0062	0.2657 ± 0.0062
	50	0.2891 ± 0.0043	0.2837 ± 0.0043	0.2803 ± 0.0043	0.2880 ± 0.0043	0.2815 ± 0.0043	0.2750 ± 0.0043
	100	0.2896 ± 0.0031	0.2869 ± 0.0031	0.2844 ± 0.0031	0.2890 ± 0.0031	0.2845 ± 0.0031	0.2813 ± 0.0031
	250	0.2903 ± 0.0020	0.2890 ± 0.0020	0.2876 ± 0.0020	0.2903 ± 0.0020	0.2880 ± 0.0020	0.2862 ± 0.0020
	500	0.2903 ± 0.0014	0.2896 ± 0.0014	0.2888 ± 0.0014	0.2902 ± 0.0014	0.2889 ± 0.0014	0.2881 ± 0.0014
	1000	0.2898 ± 0.0010	0.2893 ± 0.0010	0.2888 ± 0.0010	0.2897 ± 0.0010	0.2890 ± 0.0010	0.2885 ± 0.0009

Table 5 continued

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.9	10	0.4086 ± 0.0082	0.3976 ± 0.0083	0.3842 ± 0.0084	0.4078 ± 0.0082	0.3915 ± 0.0085	0.3796 ± 0.0085
	25	0.4093 ± 0.0051	0.4030 ± 0.0051	0.3977 ± 0.0052	0.4088 ± 0.0051	0.4002 ± 0.0052	0.3947 ± 0.0052
	50	0.4085 ± 0.0036	0.4049 ± 0.0036	0.4024 ± 0.0036	0.4084 ± 0.0037	0.4034 ± 0.0036	0.4005 ± 0.0037
	100	0.4086 ± 0.0025	0.4067 ± 0.0025	0.4053 ± 0.0025	0.4084 ± 0.0025	0.4058 ± 0.0025	0.4042 ± 0.0025
	250	0.4090 ± 0.0016	0.4082 ± 0.0016	0.4076 ± 0.0016	0.4089 ± 0.0016	0.4080 ± 0.0016	0.4071 ± 0.0016
	500	0.4094 ± 0.0012	0.4090 ± 0.0012	0.4087 ± 0.0012	0.4094 ± 0.0012	0.4090 ± 0.0012	0.4084 ± 0.0012
	1000	0.4094 ± 0.0008	0.4091 ± 0.0008	0.4089 ± 0.0008	0.4094 ± 0.0008	0.4091 ± 0.0008	0.4088 ± 0.0008
Coin bias	#Games	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.4031 ± 0.0082	0.3876 ± 0.0085	0.3724 ± 0.0086	0.4049 ± 0.0082	0.3863 ± 0.0085	0.3766 ± 0.0086
	25	0.4075 ± 0.0051	0.3996 ± 0.0052	0.3925 ± 0.0053	0.4080 ± 0.0050	0.3991 ± 0.0051	0.3943 ± 0.0052
	50	0.4090 ± 0.0034	0.4055 ± 0.0035	0.4011 ± 0.0035	0.4092 ± 0.0034	0.4049 ± 0.0035	0.4026 ± 0.0035
	100	0.4097 ± 0.0024	0.4075 ± 0.0024	0.4052 ± 0.0024	0.4097 ± 0.0024	0.4073 ± 0.0025	0.4062 ± 0.0024
	250	0.4101 ± 0.0016	0.4092 ± 0.0016	0.4083 ± 0.0016	0.4101 ± 0.0016	0.4091 ± 0.0016	0.4086 ± 0.0016
	500	0.4100 ± 0.0011	0.4094 ± 0.0011	0.4091 ± 0.0011	0.4099 ± 0.0011	0.4095 ± 0.0011	0.4091 ± 0.0011
	1000	0.4098 ± 0.0008	0.4095 ± 0.0008	0.4092 ± 0.0008	0.4097 ± 0.0008	0.4095 ± 0.0008	0.4092 ± 0.0008
0.3	10	0.2837 ± 0.0095	0.2551 ± 0.0103	0.2380 ± 0.0103	0.2831 ± 0.0097	0.2630 ± 0.0101	0.2453 ± 0.0105
	25	0.2861 ± 0.0060	0.2711 ± 0.0063	0.2622 ± 0.0063	0.2861 ± 0.0061	0.2754 ± 0.0062	0.2655 ± 0.0063
	50	0.2887 ± 0.0045	0.2801 ± 0.0045	0.2745 ± 0.0045	0.2889 ± 0.0044	0.2826 ± 0.0044	0.2767 ± 0.0044
	100	0.2896 ± 0.0031	0.2840 ± 0.0031	0.2810 ± 0.0031	0.2895 ± 0.0031	0.2859 ± 0.0031	0.2824 ± 0.0031
	250	0.2894 ± 0.0019	0.2869 ± 0.0019	0.2856 ± 0.0019	0.2894 ± 0.0019	0.2874 ± 0.0019	0.2859 ± 0.0020
	500	0.2900 ± 0.0014	0.2885 ± 0.0014	0.2879 ± 0.0014	0.2899 ± 0.0014	0.2889 ± 0.0014	0.2880 ± 0.0014
	1000	0.2895 ± 0.0010	0.2887 ± 0.0010	0.2884 ± 0.0010	0.2895 ± 0.0010	0.2889 ± 0.0010	0.2884 ± 0.0010

Table 5 continued

Coin bias	#Games	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2460 ± 0.0104	0.2168 ± 0.0104	0.1997 ± 0.0105	0.2440 ± 0.0102	0.2221 ± 0.0104	0.2119 ± 0.0107
	25	0.2435 ± 0.0064	0.2292 ± 0.0065	0.2212 ± 0.0067	0.2427 ± 0.0064	0.2314 ± 0.0065	0.2268 ± 0.0067
	50	0.2459 ± 0.0047	0.2367 ± 0.0048	0.2318 ± 0.0049	0.2454 ± 0.0047	0.2382 ± 0.0048	0.2345 ± 0.0048
	100	0.2482 ± 0.0033	0.2430 ± 0.0033	0.2401 ± 0.0034	0.2479 ± 0.0033	0.2432 ± 0.0034	0.2416 ± 0.0034
	250	0.2492 ± 0.0021	0.2468 ± 0.0021	0.2450 ± 0.0021	0.2490 ± 0.0021	0.2468 ± 0.0021	0.2458 ± 0.0021
	500	0.2498 ± 0.0015	0.2483 ± 0.0015	0.2476 ± 0.0015	0.2497 ± 0.0015	0.2483 ± 0.0015	0.2478 ± 0.0015
	1000	0.2495 ± 0.0010	0.2488 ± 0.0010	0.2482 ± 0.0010	0.2495 ± 0.0010	0.2488 ± 0.0010	0.2484 ± 0.0010
0.7	10	0.2801 ± 0.0102	0.2603 ± 0.0106	0.2374 ± 0.0106	0.2791 ± 0.0104	0.2588 ± 0.0105	0.2508 ± 0.0106
	25	0.2878 ± 0.0063	0.2761 ± 0.0064	0.2673 ± 0.0064	0.2870 ± 0.0062	0.2752 ± 0.0065	0.2705 ± 0.0066
	50	0.2889 ± 0.0043	0.2823 ± 0.0045	0.2758 ± 0.0045	0.2889 ± 0.0043	0.2810 ± 0.0044	0.2775 ± 0.0045
	100	0.2895 ± 0.0031	0.2858 ± 0.0032	0.2826 ± 0.0031	0.2898 ± 0.0031	0.2854 ± 0.0031	0.2830 ± 0.0032
	250	0.2904 ± 0.0020	0.2884 ± 0.0020	0.2871 ± 0.0020	0.2905 ± 0.0020	0.2886 ± 0.0020	0.2872 ± 0.0020
	500	0.2903 ± 0.0014	0.2892 ± 0.0014	0.2883 ± 0.0014	0.2903 ± 0.0014	0.2893 ± 0.0014	0.2885 ± 0.0014
	1000	0.2898 ± 0.0010	0.2891 ± 0.0010	0.2887 ± 0.0010	0.2897 ± 0.0010	0.2892 ± 0.0010	0.2888 ± 0.0010
0.9	10	0.4088 ± 0.0083	0.3945 ± 0.0085	0.3823 ± 0.0089	0.4080 ± 0.0082	0.3958 ± 0.0085	0.3844 ± 0.0089
	25	0.4086 ± 0.0051	0.4020 ± 0.0052	0.3964 ± 0.0053	0.4093 ± 0.0051	0.4027 ± 0.0052	0.3971 ± 0.0053
	50	0.4081 ± 0.0036	0.4045 ± 0.0037	0.4014 ± 0.0037	0.4086 ± 0.0037	0.4050 ± 0.0036	0.4021 ± 0.0037
	100	0.4083 ± 0.0025	0.4068 ± 0.0025	0.4049 ± 0.0025	0.4088 ± 0.0025	0.4066 ± 0.0025	0.4050 ± 0.0025
	250	0.4089 ± 0.0016	0.4083 ± 0.0016	0.4073 ± 0.0016	0.4091 ± 0.0016	0.4081 ± 0.0016	0.4073 ± 0.0016
	500	0.4094 ± 0.0012	0.4091 ± 0.0012	0.4085 ± 0.0012	0.4095 ± 0.0012	0.4090 ± 0.0012	0.4086 ± 0.0012
	1000	0.4094 ± 0.0008	0.4092 ± 0.0008	0.4088 ± 0.0008	0.4094 ± 0.0008	0.4091 ± 0.0008	0.4089 ± 0.0008

In particular, results have been reported for different values of penalties and α. The parameter $r = \mathrm{R}(0, 0.1)$ means that a random real number between 0 and 0.1 is taken as penalty for each game. Other penalties are fixed throughout each game

Table 6 This table presents the same kind of information defined in Table 5 but with the same draws halved

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.3951 ± 0.0084	0.3638 ± 0.0087	0.3442 ± 0.0088	0.3937 ± 0.0083	0.3577 ± 0.0086	0.3387 ± 0.0089
	25	0.4037 ± 0.0051	0.3887 ± 0.0053	0.3797 ± 0.0053	0.4034 ± 0.0051	0.3846 ± 0.0053	0.3756 ± 0.0053
	50	0.4076 ± 0.0035	0.3991 ± 0.0035	0.3945 ± 0.0035	0.4079 ± 0.0035	0.3971 ± 0.0035	0.3917 ± 0.0035
	100	0.4088 ± 0.0025	0.4041 ± 0.0025	0.4016 ± 0.0024	0.4089 ± 0.0024	0.4026 ± 0.0025	0.3995 ± 0.0025
	250	0.4097 ± 0.0016	0.4075 ± 0.0016	0.4066 ± 0.0015	0.4097 ± 0.0016	0.4069 ± 0.0016	0.4055 ± 0.0016
	500	0.4097 ± 0.0011	0.4087 ± 0.0011	0.4080 ± 0.0011	0.4097 ± 0.0011	0.4082 ± 0.0011	0.4075 ± 0.0012
	1000	0.4097 ± 0.0008	0.4091 ± 0.0008	0.4087 ± 0.0008	0.4096 ± 0.0008	0.4088 ± 0.0008	0.4084 ± 0.0008
0.3	10	0.2746 ± 0.0094	0.2407 ± 0.0097	0.2060 ± 0.0098	0.2638 ± 0.0093	0.2293 ± 0.0097	0.2036 ± 0.0097
	25	0.2811 ± 0.0060	0.2620 ± 0.0061	0.2436 ± 0.0059	0.2771 ± 0.0060	0.2561 ± 0.0060	0.2393 ± 0.0059
	50	0.2848 ± 0.0044	0.2736 ± 0.0044	0.2631 ± 0.0042	0.2831 ± 0.0044	0.2697 ± 0.0043	0.2603 ± 0.0043
	100	0.2870 ± 0.0031	0.2805 ± 0.0031	0.2742 ± 0.0030	0.2863 ± 0.0031	0.2781 ± 0.0031	0.2727 ± 0.0031
	250	0.2882 ± 0.0019	0.2854 ± 0.0019	0.2820 ± 0.0019	0.2880 ± 0.0019	0.2839 ± 0.0019	0.2811 ± 0.0019
	500	0.2894 ± 0.0014	0.2877 ± 0.0014	0.2857 ± 0.0014	0.2893 ± 0.0014	0.2867 ± 0.0014	0.2851 ± 0.0014
	1000	0.2892 ± 0.0010	0.2883 ± 0.0010	0.2872 ± 0.0010	0.2891 ± 0.0010	0.2878 ± 0.0010	0.2868 ± 0.0010

Table 6 continued

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2277 ± 0.0101	0.1953 ± 0.0100	0.1692 ± 0.0096	0.2299 ± 0.0103	0.1907 ± 0.0099	0.1617 ± 0.0098
	25	0.2361 ± 0.0064	0.2161 ± 0.0064	0.2007 ± 0.0063	0.2341 ± 0.0063	0.2138 ± 0.0064	0.1974 ± 0.0061
	50	0.2414 ± 0.0048	0.2291 ± 0.0047	0.2194 ± 0.0047	0.2399 ± 0.0047	0.2266 ± 0.0048	0.2158 ± 0.0046
	100	0.2456 ± 0.0034	0.2380 ± 0.0033	0.2326 ± 0.0032	0.2446 ± 0.0033	0.2363 ± 0.0034	0.2299 ± 0.0033
	250	0.2478 ± 0.0021	0.2441 ± 0.0021	0.2411 ± 0.0021	0.2475 ± 0.0021	0.2430 ± 0.0021	0.2396 ± 0.0021
	500	0.2490 ± 0.0015	0.2469 ± 0.0015	0.2451 ± 0.0014	0.2489 ± 0.0015	0.2466 ± 0.0015	0.2442 ± 0.0015
	1000	0.2492 ± 0.0010	0.2480 ± 0.0010	0.2470 ± 0.0010	0.2491 ± 0.0010	0.2478 ± 0.0010	0.2465 ± 0.0010
0.7	10	0.2655 ± 0.0102	0.2303 ± 0.0099	0.2103 ± 0.0101	0.2643 ± 0.0103	0.2305 ± 0.0101	0.1998 ± 0.0103
	25	0.2800 ± 0.0064	0.2613 ± 0.0061	0.2481 ± 0.0062	0.2794 ± 0.0064	0.2603 ± 0.0060	0.2412 ± 0.0063
	50	0.2846 ± 0.0044	0.2728 ± 0.0042	0.2649 ± 0.0042	0.2844 ± 0.0044	0.2713 ± 0.0043	0.2608 ± 0.0044
	100	0.2874 ± 0.0031	0.2803 ± 0.0031	0.2759 ± 0.0030	0.2872 ± 0.0031	0.2792 ± 0.0031	0.2726 ± 0.0031
	250	0.2894 ± 0.0020	0.2860 ± 0.0020	0.2836 ± 0.0019	0.2894 ± 0.0020	0.2852 ± 0.0020	0.2821 ± 0.0019
	500	0.2898 ± 0.0014	0.2879 ± 0.0014	0.2864 ± 0.0014	0.2897 ± 0.0014	0.2873 ± 0.0014	0.2856 ± 0.0014
	1000	0.2895 ± 0.0010	0.2885 ± 0.0009	0.2877 ± 0.0010	0.2895 ± 0.0010	0.2882 ± 0.0010	0.2872 ± 0.0009

Table 6 continued

Coin bias	#Games	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.9	10	0.4004 ± 0.0084	0.3697 ± 0.0086	0.3447 ± 0.0088	0.3985 ± 0.0084	0.3676 ± 0.0087	0.3463 ± 0.0089
	25	0.4048 ± 0.0051	0.3902 ± 0.0053	0.3784 ± 0.0053	0.4041 ± 0.0051	0.3880 ± 0.0053	0.3768 ± 0.0053
	50	0.4066 ± 0.0037	0.3979 ± 0.0037	0.3921 ± 0.0037	0.4058 ± 0.0037	0.3971 ± 0.0037	0.3901 ± 0.0037
	100	0.4075 ± 0.0025	0.4028 ± 0.0025	0.3994 ± 0.0025	0.4071 ± 0.0025	0.4018 ± 0.0025	0.3982 ± 0.0025
	250	0.4086 ± 0.0016	0.4065 ± 0.0016	0.4050 ± 0.0016	0.4085 ± 0.0016	0.4059 ± 0.0016	0.4042 ± 0.0016
	500	0.4092 ± 0.0012	0.4082 ± 0.0011	0.4073 ± 0.0012	0.4092 ± 0.0012	0.4078 ± 0.0012	0.4069 ± 0.0012
	1000	0.4092 ± 0.0008	0.4087 ± 0.0008	0.4082 ± 0.0008	0.4093 ± 0.0008	0.4084 ± 0.0008	0.4080 ± 0.0008
Coin bias	#Games	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	10	0.3960 ± 0.0084	0.3544 ± 0.0089	0.3276 ± 0.0094	0.3971 ± 0.0083	0.3636 ± 0.0090	0.3410 ± 0.0092
	25	0.4034 ± 0.0052	0.3837 ± 0.0054	0.3724 ± 0.0054	0.4044 ± 0.0051	0.3882 ± 0.0054	0.3779 ± 0.0054
	50	0.4077 ± 0.0035	0.3966 ± 0.0036	0.3907 ± 0.0036	0.4079 ± 0.0035	0.3989 ± 0.0036	0.3930 ± 0.0036
	100	0.4089 ± 0.0024	0.4030 ± 0.0025	0.3995 ± 0.0025	0.4088 ± 0.0025	0.4036 ± 0.0025	0.4008 ± 0.0025
	250	0.4097 ± 0.0016	0.4071 ± 0.0016	0.4055 ± 0.0016	0.4096 ± 0.0016	0.4076 ± 0.0016	0.4060 ± 0.0016
	500	0.4098 ± 0.0011	0.4084 ± 0.0011	0.4075 ± 0.0012	0.4097 ± 0.0011	0.4087 ± 0.0011	0.4078 ± 0.0012
	1000	0.4097 ± 0.0008	0.4089 ± 0.0008	0.4084 ± 0.0008	0.4096 ± 0.0008	0.4090 ± 0.0008	0.4086 ± 0.0008
0.3	10	0.2681 ± 0.0095	0.2259 ± 0.0102	0.1915 ± 0.0103	0.2697 ± 0.0098	0.2354 ± 0.0102	0.2142 ± 0.0103
	25	0.2798 ± 0.0061	0.2562 ± 0.0064	0.2391 ± 0.0062	0.2794 ± 0.0062	0.2615 ± 0.0062	0.2467 ± 0.0062
	50	0.2841 ± 0.0044	0.2715 ± 0.0046	0.2616 ± 0.0044	0.2839 ± 0.0045	0.2739 ± 0.0044	0.2658 ± 0.0045
	100	0.2866 ± 0.0031	0.2796 ± 0.0031	0.2736 ± 0.0030	0.2867 ± 0.0031	0.2803 ± 0.0031	0.2763 ± 0.0031
	250	0.2880 ± 0.0019	0.2847 ± 0.0019	0.2818 ± 0.0019	0.2883 ± 0.0020	0.2850 ± 0.0019	0.2833 ± 0.0019
	500	0.2892 ± 0.0014	0.2871 ± 0.0014	0.2855 ± 0.0014	0.2893 ± 0.0014	0.2873 ± 0.0014	0.2861 ± 0.0014
	1000	0.2891 ± 0.0010	0.2881 ± 0.0010	0.2871 ± 0.0010	0.2892 ± 0.0010	0.2881 ± 0.0010	0.2874 ± 0.0010

Table 6 continued

Coin bias	#Games	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	10	0.2277 ± 0.0103	0.1800 ± 0.0105	0.1338 ± 0.0107	0.2320 ± 0.0101	0.2017 ± 0.0107	0.1704 ± 0.0113
	25	0.2354 ± 0.0066	0.2102 ± 0.0066	0.1866 ± 0.0067	0.2349 ± 0.0064	0.2181 ± 0.0065	0.2028 ± 0.0068
	50	0.2412 ± 0.0049	0.2261 ± 0.0048	0.2132 ± 0.0049	0.2412 ± 0.0048	0.2306 ± 0.0048	0.2203 ± 0.0049
	100	0.2450 ± 0.0034	0.2370 ± 0.0034	0.2296 ± 0.0034	0.2451 ± 0.0034	0.2390 ± 0.0033	0.2321 ± 0.0034
	250	0.2476 ± 0.0021	0.2440 ± 0.0021	0.2400 ± 0.0022	0.2476 ± 0.0021	0.2440 ± 0.0021	0.2408 ± 0.0022
	500	0.2490 ± 0.0015	0.2467 ± 0.0015	0.2446 ± 0.0015	0.2487 ± 0.0015	0.2468 ± 0.0015	0.2450 ± 0.0015
	1000	0.2491 ± 0.0010	0.2478 ± 0.0010	0.2468 ± 0.0010	0.2491 ± 0.0010	0.2479 ± 0.0010	0.2470 ± 0.0010
0.7	10	0.2659 ± 0.0102	0.2153 ± 0.0105	0.1895 ± 0.0107	0.2668 ± 0.0106	0.2332 ± 0.0107	0.2082 ± 0.0107
	25	0.2797 ± 0.0063	0.2535 ± 0.0063	0.2398 ± 0.0065	0.2794 ± 0.0064	0.2624 ± 0.0064	0.2482 ± 0.0064
	50	0.2844 ± 0.0044	0.2690 ± 0.0044	0.2613 ± 0.0045	0.2842 ± 0.0043	0.2738 ± 0.0045	0.2660 ± 0.0044
	100	0.2873 ± 0.0031	0.2791 ± 0.0031	0.2740 ± 0.0032	0.2867 ± 0.0032	0.2810 ± 0.0031	0.2763 ± 0.0031
	250	0.2894 ± 0.0020	0.2858 ± 0.0020	0.2828 ± 0.0020	0.2892 ± 0.0020	0.2864 ± 0.0020	0.2840 ± 0.0019
	500	0.2898 ± 0.0014	0.2878 ± 0.0014	0.2858 ± 0.0014	0.2896 ± 0.0014	0.2879 ± 0.0014	0.2866 ± 0.0014
	1000	0.2895 ± 0.0010	0.2884 ± 0.0010	0.2873 ± 0.0010	0.2893 ± 0.0010	0.2884 ± 0.0010	0.2877 ± 0.0010
0.9	10	0.4013 ± 0.0085	0.3640 ± 0.0089	0.3341 ± 0.0094	0.3983 ± 0.0086	0.3705 ± 0.0089	0.3449 ± 0.0093
	25	0.4049 ± 0.0052	0.3875 ± 0.0054	0.3732 ± 0.0054	0.4037 ± 0.0051	0.3895 ± 0.0053	0.3779 ± 0.0055
	50	0.4061 ± 0.0036	0.3971 ± 0.0038	0.3895 ± 0.0037	0.4058 ± 0.0036	0.3978 ± 0.0038	0.3909 ± 0.0038
	100	0.4073 ± 0.0025	0.4024 ± 0.0025	0.3981 ± 0.0025	0.4073 ± 0.0025	0.4026 ± 0.0025	0.3987 ± 0.0026
	250	0.4085 ± 0.0016	0.4064 ± 0.0016	0.4044 ± 0.0016	0.4086 ± 0.0016	0.4066 ± 0.0016	0.4045 ± 0.0016
	500	0.4092 ± 0.0012	0.4081 ± 0.0012	0.4068 ± 0.0012	0.4093 ± 0.0012	0.4082 ± 0.0012	0.4071 ± 0.0012
	1000	0.4092 ± 0.0008	0.4086 ± 0.0008	0.4080 ± 0.0008	0.4093 ± 0.0008	0.4086 ± 0.0008	0.4081 ± 0.0008

In a similar manner to Table 3, it shows the outcomes of 1000 simulations, each one including 1000 games, each one including 5 (4 + 1) draws, all belonging to the same sequences analyzed in Table 1

Table 7 Statistics regarding the conditions met by the Frequentist player, as defined at the end of Sect. 3.3.2

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	i	97.320 ± 0.636	90.548 ± 0.612	87.366 ± 0.601	97.320 ± 0.636	90.548 ± 0.612	87.366 ± 0.601
	ii	893.870 ± 0.662	885.542 ± 0.679	881.361 ± 0.690	893.870 ± 0.662	885.542 ± 0.679	881.361 ± 0.690
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	7.500 ± 0.199	16.824 ± 0.480	19.557 ± 0.641
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	8.810 ± 0.181	23.910 ± 0.289	31.273 ± 0.335	1.310 ± 0.131	7.086 ± 0.446	11.716 ± 0.605
0.3	i	293.666 ± 0.939	282.852 ± 0.942	277.641 ± 0.937	293.666 ± 0.939	282.852 ± 0.942	277.641 ± 0.937
	ii	693.396 ± 0.953	681.540 ± 0.968	675.945 ± 0.974	693.396 ± 0.953	681.540 ± 0.968	675.945 ± 0.974
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	10.439 ± 0.266	21.271 ± 0.707	24.805 ± 0.914
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	12.938 ± 0.216	35.608 ± 0.358	46.414 ± 0.398	2.499 ± 0.199	14.337 ± 0.692	21.609 ± 0.911

Table 7 continued

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	i	493.090 ± 1.038	480.636 ± 1.040	474.681 ± 1.034	493.090 ± 1.038	480.636 ± 1.040	474.681 ± 1.034
	ii	492.778 ± 1.034	480.359 ± 1.044	474.572 ± 1.039	492.778 ± 1.034	480.359 ± 1.044	474.572 ± 1.039
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	11.233 ± 0.283	22.974 ± 0.780	25.866 ± 0.981
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	14.132 ± 0.232	39.005 ± 0.367	50.747 ± 0.409	2.899 ± 0.232	16.031 ± 0.738	24.881 ± 0.977
0.7	i	693.101 ± 0.991	681.504 ± 0.996	675.893 ± 1.003	693.101 ± 0.991	681.504 ± 0.996	675.893 ± 1.003
	ii	293.882 ± 0.978	282.762 ± 0.961	277.671 ± 0.953	293.882 ± 0.978	282.762 ± 0.961	277.671 ± 0.953
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	10.549 ± 0.279	21.621 ± 0.706	25.027 ± 0.904
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	13.017 ± 0.218	35.734 ± 0.349	46.436 ± 0.398	2.468 ± 0.200	14.113 ± 0.680	21.409 ± 0.882

Table 7 continued

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.9	i	895.470 ± 0.649	887.276 ± 0.669	883.225 ± 0.677	895.470 ± 0.649	887.276 ± 0.669	883.225 ± 0.677
	ii	95.886 ± 0.630	89.090 ± 0.608	86.066 ± 0.598	95.886 ± 0.630	89.090 ± 0.608	86.066 ± 0.598
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	7.355 ± 0.198	16.636 ± 0.475	19.355 ± 0.621
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	8.644 ± 0.179	23.634 ± 0.292	30.709 ± 0.338	1.289 ± 0.130	6.998 ± 0.433	11.354 ± 0.597
Coin bias	Condition	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	i	97.320 ± 0.636	90.548 ± 0.612	87.366 ± 0.601	97.320 ± 0.636	90.548 ± 0.612	87.366 ± 0.601
	ii	893.870 ± 0.662	885.542 ± 0.679	881.361 ± 0.690	893.870 ± 0.662	885.542 ± 0.679	881.361 ± 0.690
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	8.810 ± 0.181	23.869 ± 0.288	31.117 ± 0.335	8.810 ± 0.181	23.910 ± 0.289	31.273 ± 0.335
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.000 ± 0.000	0.041 ± 0.012	0.156 ± 0.023	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.3	i	293.666 ± 0.939	282.852 ± 0.942	277.641 ± 0.937	293.666 ± 0.939	282.852 ± 0.942	277.641 ± 0.937
	ii	693.396 ± 0.953	681.540 ± 0.968	675.945 ± 0.974	693.396 ± 0.953	681.540 ± 0.968	675.945 ± 0.974
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	12.938 ± 0.216	35.424 ± 0.359	45.940 ± 0.399	12.938 ± 0.216	35.608 ± 0.358	46.414 ± 0.398
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.000 ± 0.000	0.184 ± 0.024	0.474 ± 0.035	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 7 continued

Coin bias	Condition	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	i	493.090 ± 1.038	480.636 ± 1.040	474.681 ± 1.034	493.090 ± 1.038	480.636 ± 1.040	474.681 ± 1.034
	ii	492.778 ± 1.034	480.359 ± 1.044	474.572 ± 1.039	492.778 ± 1.034	480.359 ± 1.044	474.572 ± 1.039
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	14.132 ± 0.232	38.758 ± 0.366	50.134 ± 0.409	14.132 ± 0.232	39.005 ± 0.367	50.747 ± 0.409
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.000 ± 0.000	0.247 ± 0.027	0.613 ± 0.038	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.7	i	693.101 ± 0.991	681.504 ± 0.996	675.893 ± 1.003	693.101 ± 0.991	681.504 ± 0.996	675.893 ± 1.003
	ii	293.882 ± 0.978	282.762 ± 0.961	277.671 ± 0.953	293.882 ± 0.978	282.762 ± 0.961	277.671 ± 0.953
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	13.017 ± 0.218	35.559 ± 0.350	45.976 ± 0.399	13.017 ± 0.218	35.734 ± 0.349	46.436 ± 0.398
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.000 ± 0.000	0.175 ± 0.024	0.460 ± 0.035	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.9	i	895.470 ± 0.649	887.276 ± 0.669	883.225 ± 0.677	895.470 ± 0.649	887.276 ± 0.669	883.225 ± 0.677
	ii	95.886 ± 0.630	89.090 ± 0.608	86.066 ± 0.598	95.886 ± 0.630	89.090 ± 0.608	86.066 ± 0.598
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	8.644 ± 0.179	23.601 ± 0.293	30.560 ± 0.338	8.644 ± 0.179	23.634 ± 0.292	30.709 ± 0.338
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.000 ± 0.000	0.033 ± 0.011	0.149 ± 0.022	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

1000 simulations (each one including 1000 games, each one including 10 $(9 + 1)$ draws) have been run for all considered coin biases. Results have been reported for different values of penalties and α. $r = \mathrm{R}(0, 0.1)$ means that a random real number between 0 and 0.1 is taken as penalty for each game. Other penalties are fixed throughout each game. All results report the mean value plus/minus the related standard error multiplied by $z = 1.96$ of the number of times a given condition occurred

We made two sets of simulations. Each set consisted of 1000 simulations. In the first set of simulations, there were 1000 games. In each game, players observed 9 tosses and then bet on the 10th toss. In the initial game, players had no prior observations. They updated using their 9 observations and made a decision on how (and whether) to bet on the result of the 10th toss. After the 10th toss, players updated on the 10th toss. In subsequent games, retained knowledge of their past observations of tosses. Thus, for example, in a particular simulation, players had 500 observations after 50 games. In total, there were 1000 opportunities to bet per simulation, corresponding to the 1000 games.

In the second set of simulations, there were also 1000 games in each of the 1000 simulations. In these simulations, players had just 4 observations in each game, and they bet on the result of the 5th toss. The players' updating was otherwise identical to the first set of simulations. Therefore, after 50 games, players had 250 observations in a particular simulation. In total, there were 1000 opportunities to bet per simulation. We conducted this second set of simulations to investigate how players performed with very small samples. In each simulation, every player observed the same sequence of tosses.

For each of the 1000 games in a simulation we used 1000 randomly generated ticket prices. We used different ticket prices for each simulation. However, for a given simulation, every player faced the same ticket prices. We recorded the overall mean payoffs and standard deviations for each player in each simulation. We also recorded the overall mean payoffs and standard deviations at different points during the simulations, as we show in the tables below. Finally, we recorded how often *Frequentist* chose to bet and how often they chose to withhold from betting.

The sample sizes for players' observations (9 new observations per game for the first set of simulations, 4 for the second set) seemed to offer a reasonable balance between enabling us to look at short run performances and yet also giving players some data to use. No player will update their beliefs in a radical way in response to just several tosses. Therefore, with extremely small samples and just a few games, we would mainly be comparing the players' initial choice models—interesting in some respects, but not very informative about the differences in the players' statistical learning procedures. On the other hand, the players will almost always make identical decisions with very large samples, due to washing-out of priors and narrowing of confidence intervals. Our choice of sample sizes is intended to find a reasonable medium between comparisons with extremely large and extremely small samples.

In addition to varying the sample sizes, we also varied the coin bias across the values 0.1, 0.3, 0.5, 0.7, and 0.9. For each simulation and each particular coin bias, we randomly generated a single history of coin tosses and simulated each player's behaviour with that history. Our choice to keep the histories fixed ensured that any differences in performance between players were not due to random variation in the coin toss histories that they faced.

We also made variations in players' exogenous parameters—the parameters in their belief and decision algorithms that are not set by the methodologies. For *Bayes*, we varied the values for the beta distribution parameters of (a, b) to (100, 1), (1,1), and (100,1). For *Frequentist* and *Williamsonian*, we varied the values for α to 0.01, 0.05, and 0.5. Finally, following the suggestion of Kyburg and Teng (1999), we varied the

Table 8 This table presents the same kind of information defined in Table 7 but with the same draws halved

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	i	95.486 ± 0.671	86.229 ± 0.643	82.134 ± 0.628	95.486 ± 0.671	86.229 ± 0.643	82.134 ± 0.628
	ii	891.717 ± 0.715	879.641 ± 0.736	873.600 ± 0.753	891.717 ± 0.715	879.641 ± 0.736	873.600 ± 0.753
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	10.159 ± 0.265	21.009 ± 0.669	24.253 ± 0.851
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	12.797 ± 0.220	34.130 ± 0.346	44.266 ± 0.399	2.638 ± 0.200	13.121 ± 0.638	20.013 ± 0.846
0.3	i	291.105 ± 1.026	275.866 ± 1.006	268.884 ± 1.002	291.105 ± 1.026	275.866 ± 1.006	268.884 ± 1.002
	ii	690.411 ± 1.043	673.650 ± 1.061	665.431 ± 1.069	690.411 ± 1.043	673.650 ± 1.061	665.431 ± 1.069
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	13.666 ± 0.363	25.763 ± 0.972	27.859 ± 1.200
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	18.484 ± 0.257	50.484 ± 0.418	65.685 ± 0.471	4.818 ± 0.312	24.721 ± 0.969	37.826 ± 1.207

Table 8 continued

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	i	490.092 ± 1.115	472.629 ± 1.110	464.419 ± 1.104	490.092 ± 1.115	472.629 ± 1.110	464.419 ± 1.104
	ii	489.744 ± 1.114	472.386 ± 1.121	464.168 ± 1.120	489.744 ± 1.114	472.386 ± 1.121	464.168 ± 1.120
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	14.473 ± 0.405	26.598 ± 1.041	27.195 ± 1.311
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	20.164 ± 0.273	54.985 ± 0.434	71.413 ± 0.490	5.691 ± 0.357	28.387 ± 1.052	44.218 ± 1.312
0.7	i	690.395 ± 1.074	673.682 ± 1.091	665.746 ± 1.089	690.395 ± 1.074	673.682 ± 1.091	665.746 ± 1.089
	ii	291.096 ± 1.049	275.837 ± 1.029	268.732 ± 1.024	291.096 ± 1.049	275.837 ± 1.029	268.732 ± 1.024
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	13.576 ± 0.358	25.199 ± 0.967	27.095 ± 1.216
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	18.509 ± 0.255	50.481 ± 0.404	65.522 ± 0.470	4.933 ± 0.307	25.282 ± 0.975	38.427 ± 1.226

Table 8 continued

Coin bias	Condition	Penalty = 0			Penalty = r		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.9	i	893.233 ± 0.719	881.251 ± 0.753	875.230 ± 0.770	893.233 ± 0.719	881.251 ± 0.753	875.230 ± 0.770
	ii	94.190 ± 0.683	85.041 ± 0.652	80.963 ± 0.635	94.190 ± 0.683	85.041 ± 0.652	80.963 ± 0.635
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	9.890 ± 0.264	19.964 ± 0.688	22.876 ± 0.870
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	12.577 ± 0.215	33.708 ± 0.357	43.807 ± 0.415	2.687 ± 0.200	13.744 ± 0.665	20.931 ± 0.878
Coin bias	Condition	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.1	i	95.486 ± 0.671	86.229 ± 0.643	82.134 ± 0.628	95.486 ± 0.671	86.229 ± 0.643	82.134 ± 0.628
	ii	891.717 ± 0.715	879.641 ± 0.736	873.600 ± 0.753	891.717 ± 0.715	879.641 ± 0.736	873.600 ± 0.753
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	12.796 ± 0.219	33.750 ± 0.345	43.515 ± 0.401	12.797 ± 0.220	34.130 ± 0.346	44.266 ± 0.399
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.001 ± 0.002	0.380 ± 0.032	0.751 ± 0.042	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.3	i	291.105 ± 1.026	275.866 ± 1.006	268.884 ± 1.002	291.105 ± 1.026	275.866 ± 1.006	268.884 ± 1.002
	ii	690.411 ± 1.043	673.650 ± 1.061	665.431 ± 1.069	690.411 ± 1.043	673.650 ± 1.061	665.431 ± 1.069
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	18.472 ± 0.258	49.751 ± 0.419	64.292 ± 0.470	18.484 ± 0.257	50.484 ± 0.418	65.685 ± 0.471
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.012 ± 0.007	0.733 ± 0.042	1.393 ± 0.054	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 8 continued

Coin bias	Condition	Penalty = 0.5			Penalty = 1		
		$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.05$	$\alpha = 0.01$
0.5	i	490.092 ± 1.115	472.629 ± 1.110	464.419 ± 1.104	490.092 ± 1.115	472.629 ± 1.110	464.419 ± 1.104
	ii	489.744 ± 1.114	472.386 ± 1.121	464.168 ± 1.120	489.744 ± 1.114	472.386 ± 1.121	464.168 ± 1.120
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	20.149 ± 0.274	54.089 ± 0.435	69.709 ± 0.493	20.164 ± 0.273	54.985 ± 0.434	71.413 ± 0.490
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.015 ± 0.008	0.896 ± 0.043	1.704 ± 0.055	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.7	i	690.395 ± 1.074	673.682 ± 1.091	665.746 ± 1.089	690.395 ± 1.074	673.682 ± 1.091	665.746 ± 1.089
	ii	291.096 ± 1.049	275.837 ± 1.029	268.732 ± 1.024	291.096 ± 1.049	275.837 ± 1.029	268.732 ± 1.024
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	18.500 ± 0.255	49.724 ± 0.404	64.082 ± 0.471	18.509 ± 0.255	50.481 ± 0.404	65.522 ± 0.470
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.009 ± 0.006	0.757 ± 0.042	1.440 ± 0.054	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
0.9	i	893.233 ± 0.719	881.251 ± 0.753	875.230 ± 0.770	893.233 ± 0.719	881.251 ± 0.753	875.230 ± 0.770
	ii	94.190 ± 0.683	85.041 ± 0.652	80.963 ± 0.635	94.190 ± 0.683	85.041 ± 0.652	80.963 ± 0.635
	iii	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	iv	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	v	12.575 ± 0.215	33.335 ± 0.355	43.022 ± 0.414	12.577 ± 0.215	33.708 ± 0.357	43.807 ± 0.415
	vi	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	vii	0.002 ± 0.003	0.373 ± 0.033	0.785 ± 0.042	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

It shows the outcomes of 1000 simulations, each one including 1000 games, each one including 5 (4 + 1) draws, all belonging to the same sequences analyzed in Table 7

penalty parameter ϵ for *Frequentist*, with the intention of varying the willingness of *Frequentist* to place bets rather than hold. For *Sample*, there were no parameters to vary.

Each simulation's results were independent of every other simulation, and their basic stochastic parameters were the same for a given setting, e.g. a given coin bias. Consequently, our simulations might be regarded as independent and identically distributed draws from the overall population of possible simulations under these settings. In the next section, we use the standard error under that interpretation and a 95% confidence interval. However, we strongly stress our study is descriptive rather than inferential: we have provided some relevant descriptive statistics, but proper testing of hypotheses regarding our simulations is a project for further research. Like our simulation's code, our datasets are available upon request.

4 Results and analysis

We report our results in Tables 1, 2, 3, 4, 5, 6, 7 and 8. We begin with some comparative points, always using the standard errors reported in the tables. To evaluate the players in relation to the benchmark *Sample* player, we compared the *best* performance of a player in the confidence interval for that player's results with the *worst* performance of *Sample* in its confidence interval. Using this analysis, if a player fails to outperform *Sample* for a particular number of games (on average over the 1000 simulations) then we have starred the cell in Tables 2 and 4. In starless cells, the player outperforms *Sample* according to the interval analysis. This interval analysis acknowledges the random errors involved in such an assessment, while also indicating the descriptive basis for the evaluations that we make in this section.

When *Bayes* sets a B(1,1) prior[15] and *Frequentist* sets a high value of α, we found that they both reliably either matched or exceeded the performance of the *Sample* player: see Tables 2 and 4 for B(1,1). In this respect, our results differ from those of Kyburg and Teng (1999). The very similar performance is unsurprising in a large number of games, where these players' performances were more or less identical. As sample sizes become large, then *Bayes*'s posterior probabilities become very close the cumulative observed sample frequency, while *Frequentist*'s confidence interval estimates are very narrow around that frequency. However, it is surprising with 10 games, where players had less than 100 observations to make their decisions in each game. *Bayes* with a B(1,1) prior and *Frequentist* with $\alpha = 0.5$ are behaving differently, but neither performed detectably better across all the coin biases. The same was true for *Williamsonian* with $\alpha = 0.5$. Overall, for each of the three statistical methodologies, there were player settings under which they passed the benchmark that we set.

We did find inferior performance relative to *Sample* when *Bayes*'s beta prior was biased towards either heads, B(100,1), or tails, B(1,100). Indeed, in Tables 1 and 3, we can see that this player setting is the only one under which any player made a loss in the short-term. It is not surprising that, with a biased prior that differs from the coin bias, *Bayes* would perform badly. It is perhaps more surprising that these biases never

[15] A beta prior with parameters $a = 1$ and $b = 1$.

yielded detectable special advantages relative to the flat B(1,1) prior. Even in the short run, there were no simulation settings where *Bayes* performed less well (in either the short run or the long run) than *Sample* using a B(1,1) prior but not with a biased prior for the 1000 simulations as a whole. Consequently, a biased prior created the risk of some very bad performances for *Bayes*, without identifiable benefits relative to flat priors. Furthermore, these costs of biased priors were persistent: even with thousands of observations over many games, the *Bayes* with a badly inappropriate prior was still performing badly. While a washing-out of priors would occur in the long run, it can take a long time in this sort of decision problem.

The absence of special benefits from a biased prior can be explained by the "flatness" of the B(1,1) Bayesian prior. Although this prior is an equivocal and thus would generally have some advantages for a 0.5 coin bias, it is not this equivocation that is the key to its success in our decision problem. Instead, the advantages come from the speed with which *Bayes* revises their beliefs using this prior. Suppose that the coin is biased towards heads, such that its frequency of heads is 0.9. Such a bias will tend to produce samples that are themselves usually biased towards heads, and the Bayesian will quickly update on this sample data if they have a B(1,1) prior, thus quickly eliminating any special advantages that the B(100,1) prior will have from being biased towards heads. Some other equivocal priors—e.g. a B(100,100) prior—would not perform so well when observing the tosses of a biased coin.

In terms of common points, all players tended to do better when the coin bias was set further away from 0.5. This result is unsurprising, because the chance of a very unrepresentative sample is greatest when the coin is fair. In contrast, if the coin bias is 0.1 or 0.9, then the randomisation process in the Python code will generate very unrepresentative samples at a lower rate. Kyburg and Teng found the same result (Kyburg and Teng, 1999, pp. 362–363).

We now turn to particular players. As previously noted, *Bayes* did best with a B(1,1) beta prior. While this result might seem to be favourable to Objective Bayesian approaches, according to which such a prior would be mandatory, note that Subjective Bayesians regard flat priors as permissible, provided that they are coherent with the total credence distribution. They just deny that they are rationally required. Furthermore, as a matter of contingent fact, most real-world Subjective Bayesian statisticians would choose such a prior if faced with the decision we describe. Since Bayesian reasoning with such a "flat" prior is similar to maximum likelihood estimation, it is also unsurprising that *Bayes* with a B(1,1) beta prior would at least match the performance of *Sample*.

We now consider the *Williamsonian*. They differ from the players studied in Kyburg and Teng (1999). Although Williamson's theory of statistics is not as prominent as the Bayesian or frequentist methodologies, it provided us with the basis for an intriguing "hybrid" player.[16] Notably, *Williamsonian* matched the performance of *Bayes* and *Frequentist*. In some cases, with low values of α, the Williamsonian failed to match the performance of *Sample* in the very short run. However, this varied with different coin bias settings, so it was not a consistent pattern. Much more investigation is needed

[16] We stress that it was inspired by Williamson's views; we do not presume that Williamson would agree with every detail of it, anymore than all Bayesians or all frequentists will approve of *Bayes* and *Frequentist* respectively.

to infer anything definite about the best values of α for *Williamsonian*'s performance. The problem might be that low values of α increase the importance of the Equivocation norm as their confidence intervals will detect coin biases more slowly. If the coin bias is 0.5, then this behaviour is unproblematic: see Tables 2 and 4. However, when the coin is biased, the Equivocation norm can repeatedly drag *Williamsonian*'s credences towards 0.5. Our results are consistent with Williamson's principal justification for the Equivocation norm: its advantages in minimising loss for the worst-case scenarios (Williamson, 2007). For combinatorial reasons, 0.5 is the least favourable stochastic condition for binomial sampling, because it maximises the number of unrepresentative samples.

At least in our simulations, *Williamsonian* was able to "have their cake and eat it" by using a (maximally) low value of $\alpha = 0.5$. They were still able to match or exceed the performance of *Sample* when the coin bias was 0.5, in both the short and long run. Yet the greater receptiveness to sample data when $\alpha = 0.5$ meant that they were also able to pass our benchmark with other coin biases. Nonetheless, one cannot extrapolate to all decision problems and say that a low value of α will always be better for *Williamsonian*. In some decision problems, perhaps with rare extreme risks, being firmly equivocal could be beneficial.

Regarding biased priors, *Williamsonian* excludes these via the Equivocation norm, and thus they do not face the problem of having such a prior in this decision problem. Consequently, even the worst performances of *Williamsonian* were not as bad as the worst performances of *Bayes* with a prior that was biased in the wrong direction. For instance, see Table 4, for coin bias = 0.1 or 0.9, and runs of 10 games, where *Williamsonian* with $\alpha = 0.01$ does far better than *Bayes* with B(100,1).

Finally, we consider *Frequentist*. For ϵ, we firstly found that this parameter achieved its intended function of increasing *Frequentist*'s propensity to bet, as we report in Tables 7 and 8. On the other hand, ϵ does not seem to reliably affect *Frequentist*'s performance, as we see in Table 6. As for α, we found the same result as Kyburg and Teng (1999). In the very short run, there were some suggestive but indefinite signs that very low values of α—particularly $\alpha = 0.01$—could lead to a poor performance relative to *Sample*, but more research is needed; the differences were not as clear as with *Bayes* using a biased prior. For $\alpha = 0.05$ and $\alpha = 0.5$, there were at least some occasions where *Frequentist* matched *Sample*. When $\alpha = 0.5$, this above-benchmark performance was consistent.

Why might high values of α hurt the performance of *Frequentist*? This question was not a focus of our study and thus it would benefit from more comprehensive analysis. Nonetheless, one notable point from Tables 7 and 8 is that a lower value of α made an identifiable difference to the rate of using randomisation. Recall that, in *Frequentist*'s mixed strategy decision rule, they randomise between buying a ticket, buying a reversed ticket, or withholding from betting in situation (vii), and between buying a ticket or a buying a reversed ticket in situation (v), as we detail in Sect. 3.3.2.[17] In situation (v), neither betting heads nor betting on tails interval dominates the other given *Frequentist*'s estimated confidence interval. Thus, they randomise between these two

[17] Due to the incentive structure of the game, there was no occurrence of situations (iii), (iv) and (vi), which correspond to cases where withholding interval dominates at least one other action. This clearly appears also in Tables 7 and 8.

bets using an additional fair coin. In situation (vii), none of their three possible actions is interval dominated, and therefore *Frequentist* randomises between all three. Both these randomisation procedures mean that *Frequentist* is acting equivocally between heads and tails. Consequently, if α is low, then they are not making much use of sample data indicating bias, but if α is high, then *Frequentist* quickly detects bias and exceeds the minimum performance of *Sample* even in the very short run, as in Tables 2 and 4. It does not follow that a high value of α is better for decision problems in general. Instead, our results suggest high values are better for this kind of decision problem, because it makes prompter use of the sample information, and hence helps avoid acting equivocally when the coin is biased.

5 Discussion

Although our simulations did not detect any reason to use one statistical methodology rather than another, it does not follow that the choice is underdetermined. Firstly, Bayesians might note that, unlike Bayesian decision theory, *Frequentist*'s decision algorithm has no axiomatic derivation; Kyburg acknowledges that it is produced by unsystematic intuitive considerations (Kyburg, 2003, p. 148). We were able to use these intuitive considerations to infer how *Frequentist* should extend the algorithm from Kyburg and Teng (1999) to cases where there was a non-zero penalty for holding, but it would be an extreme exaggeration to call it a decision "theory". More generally, there is no agreement on how to make decisions with interval-valued beliefs.[18] Secondly, frequentists might argue that there are broader aspects like long run decisions, social decisions, or epistemological points that we do not address in our simulations, which only involve individual decision-making in the short run. Thirdly, Williamsonians also think that their approach has long run virtues that favour their view for decision-making (Williamson, 2007). Our results suggest that either (1) we need a different sort of challenge for comparing their short run performance or (2) the choice between them—even when we focus on the relatively narrow issue of making good decisions—must be approached via questions of systematic coherence, the social consequences of their implementation in science, asymptotic strengths or weaknesses, results of debates in formal learning theory (Schulte, 2018) and other angles.

Our results have some good news for all three methodologies. For Bayesians, our results are better than those of Kyburg and Teng (1999). Unlike their study, we did not find that the Bayesian player will generally be outperformed by the frequentist player. For frequentists, they can take comfort from the fact that *Frequentist* can match *Bayes* in performance, even when the Bayesian's prior is set to the true relative frequency, and even though guiding us to good decisions is often regarded as a strength of Bayesianism. Finally, for Williamsonians, we have shown how they can match the performance of the more well-known Bayesian and frequentist methodologies. This strong performance by *Williamsonian* might encourage a wider audience of philoso-

[18] For some discussions, see Kyburg (1990), Seidenfeld (2004), Troffaes (2007), Haenni et al. (2011), Bradley (2017).

phers, statisticians, and other interested groups to inquire further into the relatively marginal Williamsonian position.

The similarity in performance is particularly intriguing because it is not always caused by players making the same decisions. It is true that, in the long run, players have almost identical performances because they have more or less the same beliefs, and they are making the same decisions. That occurs because, with large samples, the initial priors lose their significance; the confidence intervals narrow; and the players concur on the best decisions for the greater proportion of a given coin toss history and sequence of ticket prices. However, with smaller samples, the players have different beliefs and make different decisions, which is why their results vary among particular coin toss histories. The lack of a systematic difference is thus occurring in spite of the players' contrasting decisions. Therefore, in the short run, the players have genuinely different statistical methodologies, but in our problem, their methodologies do not guide them to relatively better decisions.

The problem of priors is also a heated topic within the Statistics Wars. Overall, *Bayes* did badly with an extreme prior, and their performance was no better (relative to *Sample*) than the *Bayes* setting B(1,1), even when their biased prior was in the right direction relative to the coin bias. However, one point that our simulations illustrate is that the effects of "washing-out" of the priors can take a long time to manifest. In the long run, we would expect to see *Bayes* catch up even with a very "bad" prior. Nonetheless, in the real world, relative performance in the short run can make an important difference for aims like attracting investment in business, attracting funding in science, or winning an arms race.[19]

Moreover, the decision problem and priors we used is arguably favourable to *Bayes*'s catching up in such cases. *Bayes* views the coin tosses as an exchangeable. This means that the tosses are independently and identically distributed according to some distributional form, which implies that the order of the tosses does not matter. More fundamentally, their prior was always sensitive to evidence. Subjective Bayesianism as such does not require such priors. Most obviously, extreme priors of 0 or 1 cannot be changed by conditionalization. However, the further possibility of non-extreme priors that share this dogmatism have been known since the early days of Bayesian epistemology (Carnap, 1945a, pp. 80–81). The limits of what can be learned by conditionalization has been a topic of research in recent formal epistemology (Rédei and Gyenis, 2019). For instance, if a Bayesian player was estimating the general relative frequencies in a massively large number of tosses, then clearly there are prior distributions that will be totally insensitive to new evidence. Thus, we recognise that it would be possible to create tougher learning challenges for the Bayesian—just as it *may* be possible to construct choice problems where the frequentist decision theory we consider entails inconsistent or systematically exploitable choices. On the other hand, it is possible that Bayesians might do better in problems involving more background knowledge than the sparse information that they (and the other players) possess in this article's simulations.

[19] In general, relative short run performance is very significant for problems that are analogous to being at the front of a stampede of animals, where it is your *relative* speed that determines whether you are trampled.

The use of the penalty ϵ enabled us to remove an asymmetry between the Bayesian and frequentist players in Kyburg and Teng (1999). In their simulations, the frequentist player would refuse to bet until they either had relatively large samples or they were faced with very attractive (given their beliefs) ticket prices. In contrast, their Bayesian player would always bet, even with very small samples. It was possible that this asymmetry produced Kyburg and Teng's favourable results for the frequentist. Perhaps surprisingly, in our simulations, forcing *Frequentist* to bet using ϵ made no systematic difference to their performance relative to *Sample*. However, there is still a possibly important asymmetry when ϵ is high. *Frequentist*, when forced to bet, will randomise unless they have either relatively large samples or very attractive ticket prices. In contrast, *Bayes* almost never randomises, because they will only randomise if two bets have the same expected payoff. One might think that this reflects a methodological difference, in which case this asymmetry is not a problem for our comparisons. However, it indicates the dangers of generalising from our results to other decision problems, such as those in which there is a yet another penalty, this time directed at randomisation. Additionally, one reason for thinking that the asymmetry is not due to anything specific to the methodologies is that other types of prior for *Bayes* could reduce this asymmetry, even within our setup. For instance, *Bayes* could have an equivocal prior that was "sticky" in the sense of not changing (or not changing much) until they had relatively large samples. While a beta distribution was a reasonable way to model *Bayes*, there is nothing in Subjective Bayesianism that mandates this family of priors. *Frequentist* could also be modified to randomise less, by giving them a rule like maximin, which would involve less randomisation than the interval-dominance that we used in this study. We leave such alternative simulations for future research, but their possibility should warn against generalising from our simulations to other decision problems or different player settings.

Another contrast with Kyburg and Teng (1999) is that we have extended the simulations in some new ways. In addition to those already mentioned (such as the addition of *Williamsonian*) we have explored the players' performances in relatively large numbers of games—while Kyburg and Teng (1999, p. 362) stopped after 50 games, we have compared players' performances over 100, 250, 500, and 1000 games. We thus investigated both the relatively short and the medium run. Even though our primary interest was short run performances, the latter provides a useful check of both our model (if we expect convergence of performance in the long run) and whether clear differences emerged in the long run.

The principal divergence of our results and those of Kyburg and Teng (1999) is that we do not find, for any setting, that *Frequentist* detectably outperforms *Bayes*. We can explain the results in Kyburg and Teng (1999) in terms of randomness. They performed only 100 simulations, so the modest differences between players in their results can easily be explained by random error. By performing many simulations, we have greatly reduced the effects of randomness on our results. Their study retains some interest as a possibility result: our results do not contradict the possibility that *Frequentist* will sometimes perform better than *Bayes* in a relatively small number of simulations. What we have found is that such an outcome does not systematically occur for the best settings of these players out of the settings that we investigated.

Without an extensive theoretical analysis, we can only speculate briefly on why (with the appropriate settings) the players had similar results. Relative to *Sample*, the best performing players were *Bayes* with a B(1,1) prior, *Frequentist* with $\alpha = 0.5$, and *Williamsonian* with $\alpha = 0.5$. Their performances were very similar under every coin bias and with either games with 5 tosses or with 10 tosses. How can this similarity be explained?

For these three player settings, a common trait is their use of sophisticated inductive inference from their evidence. Like *Sample*, they are all willing to revise their beliefs quickly given the background knowledge and evidence that we gave them in our study. In the case of *Frequentist* and *Williamsonian*, this readiness was due to the low value of α, which respectively reduced the importance of randomisation and Equivocation. For *Bayes*, this readiness was due to a flat prior. Yet, unlike *Sample*, their early decisions were moderated by some factor that encouraged equivocal (and thus asymptotically cautious) reasoning in the early part of the game, thus downplaying the importance of early uniform samples. For *Bayes* and *Williamsonian*, this equivocation is achieved via the determination of their degrees of belief, using the beta prior and the Equivocation norm respectively. For *Frequentist*, it is achieved in their decision-making procedure: recall that their mixed strategy requires randomisation using a fair coin when no action interval-dominates any other; this underdetermination occurs most often early in the simulations, when sample sizes are small and confidence intervals can be wide even with $\alpha = 0.5$.

Hence, for philosophers, our results indicate the power of sophisticated statistical induction, in the sense of (1) revising one's initial beliefs about the world to be closer to observed frequencies, but also (2) not extrapolating too strongly from small samples. Sophisticated induction can be achieved by multiple methods, including as conditionalization and confidence interval estimation. It then only needs an adequate decision procedure, and we found that both maximising expected payoffs (the approach of *Bayes* and *Williamsonian*) and the interval dominance rule of *Frequentist* were adequate for the problem we studied, provided that the player's other settings were also adequate.

6 Conclusion

Our simulation study had the result of a Caucus Race (Carroll, 1920, p. 34), in which "*everybody* has won, and all must have prizes." This result obtained even in the very short run, before convergence theorems could manifest. In contrast to Kyburg and Teng (1999), we did not find that the frequentist player will generally outperform the other players. However, contrary to what some philosophers might expect, our frequentist player did no worse than the Bayesian. We also found that this result obtained after we introduced new features to the simulations, such as a penalty for not betting that encouraged *Frequentist* to place more bets. We also introduced a hitherto unstudied player, based on Williamson's version of Objective Bayesianism, and we found that they performed just as well as the more well-known methodologies.

There are many limitations of our simulations that should be noted in order to avoid overgeneralisations. We simulated just one type of decision problem. The data and the

problem itself were both very simple. The players could also be modified in ways that would be consistent with the statistical methodology that inspired them. We also only considered one algorithm for *Frequentist* and this algorithm lacked a systematic derivation from a generally accepted non-Bayesian theory of decision, because the latter does not exist. Another issue is that, like Kyburg and Teng (1999), our players differ in both their inference rules and their decision rules.[20] We analysed our results informally; we welcome the use of either our results or code for more systematic analyses.

The type of simulations we conducted are largely unstudied and there are many novel dimensions for further exploration. It would be interesting to investigate more alternative players. For example, there are many, many types of Bayesians: how would Imprecise Bayesians, whose beliefs are characterised by sets of probability functions, perform in comparison to *Bayes*? How would *Frequentist* perform with other decision algorithms? The game set-up could also be modified in various ways. Currently, it features a relatively simple problem space and simple data, but would the players still match each other's performance when faced with more complex challenges? What if we make the game interactive? These directions indicate the fertility of this field of research.

The strong performance of all four players in our study should remind philosophers and those working in applications of statistics to respect the capacity of Bayesianism, frequentism, and Williamsonianism to guide us towards good decisions. Our little skirmish does not win the Statistics Wars for any side, but it indicates that all sides have firepower that is worth taking seriously.

Acknowledgements We thank Daniele Tortoli (University of Modena and Reggio Emilia, Italy) for his very valuable support in accelerating computations in our research. We are grateful to two anonymous referees for their helpful comments. We also thank the Erasmus Institute for Philosophy and Economics for their advice on the development of this article. Likewise, we thank Durham University for providing open access for this article.

References

Bacchus, F., Kyburg, H. E., & Thalos, M. (1990). Against conditionalization. *Synthese, 85*(3), 475–506. https://doi.org/10.1007/bf00484837

Benenson, F. C. (1984). *Probability, objectivity and evidence*. Routledge and Kegan Paul.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press. https://doi.org/10.1017/9780511760105

[20] We are grateful to an anonymous referee, who points out that the aforementioned "imprecise" Bayesian approach offers a neater comparison with *Frequentist* in this respect, which provides a further reason to examine such a player in future research.

Carnap, R. (1945a). On inductive logic. *Philosophy of Science, 12*(2), 72–97. https://doi.org/10.1086/286851
Carnap, R. (1945b). The two concepts of probability: The problem of probability. *Philosophy and Phenomenological Research, 5*(4), 513. https://doi.org/10.2307/2102817
Carnap, R. (1952). *The continuum of inductive methods*. The University of Chicago Press.
Carroll, L. (1920). *Alice's adventures in wonderland*. The Macmillan Company.
Celeux, G., Anbari, M. E., Marin, J. M., & Robert, C. P. (2012). Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis, 7*(2), 477–502. https://doi.org/10.1214/12-ba716
De Finetti, B. (1980). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg Jr. & H. Smokler (Eds.), *Studies in subjective probability*. Krieger Publishing Company.
Feigl, H. (1954). Scientific method without metaphysical presuppositions. *Philosophical Studies, 5*(2), 17–29. https://doi.org/10.1007/bf02223254
Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, winter 2018 edn*, Metaphysics Research Lab. Stanford University. Retrieved May 13, 2021, from https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/
Fisher, R. (1947). *The design of experiments* (4th ed.). Oliver and Boyd.
Gelman, A. (2015). Working through some issues. *Significance, 12*(3), 33–35. https://doi.org/10.1111/j.1740-9713.2015.00828.x
Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics, 18*(2), 141–153. https://doi.org/10.1016/0304-4068(89)90018-9
Greaves, H., & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind, 115*(459), 607–632. https://doi.org/10.1093/mind/fzl607
Haenni, R., Romeijn, J. W., Wheeler, G., & Williamson, J. (2011). *Probabilistic logics and probabilistic networks*. Springer. https://doi.org/10.1007/978-94-007-0008-6
Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach*. Open Court Publishing.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124
Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*(4), 620–630. https://doi.org/10.1103/physrev.106.620
Keynes, J. M. (1921). *A treatise on probability*. Macmillan and Co.
Kyburg, H. E. (1990). *Science and reason*. Oxford University Press.
Kyburg, H. E. (2001). Probability as a guide in life. *The Monist, 84*(2), 135–152. https://doi.org/10.5840/monist200184210
Kyburg, H. E. (2003). Are there degrees of belief? *Journal of Applied Logic, 1*(3–4), 139–149. https://doi.org/10.1016/s1570-8683(03)00010-7
Kyburg, H. E., & Teng, C. M. (1999). Choosing among interpretations of probability. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Francisco CA, USA, UAI'99* (pp. 359–365). https://doi.org/10.5555/2073796.2073837
Kyburg, H. E., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press. https://doi.org/10.1017/cbo9780511612947
Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy, 71*(13), 391. https://doi.org/10.2307/2025161
Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. The University of Chicago Press.
Mayo, D. G. (2018). *Statistical inference as severe testing. How to get beyond the Statistics Wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184
McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549
Mun, J. (2008). *Advanced analytical models: Over 800 models and 300 applications from the Basel II Accord to Wall Street and Beyond, Wiley finance series* (vol. 419). Wiley. https://doi.org/10.1002/9781119197096
Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika, 32*(2), 128–150. https://doi.org/10.1093/biomet/32.2.128
Neyman, J. (1949). *First course in probability and statistics*. University of California Press.

Neyman, J. (1957). "Inductive Behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique/Review of the International Statistical Institute, 25*(1/3), 7. https://doi.org/10.2307/1401671

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251):aac4716. https://doi.org/10.1126/science.aac4716

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25–42. https://doi.org/10.1093/bjps/x.37.25

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge.

Rédei, M., & Gyenis, Z. (2019). Having a look at the Bayes blind spot. *Synthese*. https://doi.org/10.1007/s11229-019-02311-9

Reichenbach, H. (1938). *Experience and prediction*. The University of Chicago Press.

Reichenbach, H. (1971). *The theory of probability*. University of California Press.

Resnik, M. D. (1987). *Choices: An introduction to decision theory*. University of Minnesota Press.

Romero, F., & Sprenger, J. (2020). Scientific self-correction: The Bayesian way. *Synthese*. https://doi.org/10.1007/s11229-020-02697-x

Salmon, W. (1967). *The foundations of scientific inference*. University of Pittsburgh Press.

Schoenfield, M. (2020). Can imprecise probabilities be practically motivated? A challenge to the desirability of ambiguity aversion. *Philosophers Imprint, 20*(30), 1–21. http://hdl.handle.net/2027/spo.3521354.0020.030

Schulte, O. (2018). Formal learning theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy, spring 2018 edn*. Metaphysics Research Lab, Stanford University. Retrieved May 21, 2021, from https://web.archive.org/web/20210521152126/https://plato.stanford.edu/entries/learning-formal/

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In S. van der Walt, & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 92–96). https://doi.org/10.25080/majora-92bf1922-011

Seidenfeld, T. (2004). A Contrast Between two Decision Rules for use with (Convex) Sets of Probabilities: Γ-Maximin Versus E-Admissibilty. *Synthese, 140*(1/2), 69–88. https://doi.org/10.1023/b:synt.0000029942.11359.8d

Singh, A. C., Stukel, D. M., & Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60*(2), 377–396. https://doi.org/10.1111/1467-9868.00131

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*(9), 160384. https://doi.org/10.1098/rsos.160384

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Spanos, A. (2010). Is frequentist testing vulnerable to the base-rate fallacy? *Philosophy of Science, 77*(4), 565–583. https://doi.org/10.1086/656009

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology, 31*(8), 1188–1214. https://doi.org/10.1080/09515089.2018.1490707

Troffaes, M. C. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning, 45*(1), 17–29. https://doi.org/10.1016/j.ijar.2006.06.001

van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Haucke, M. N., Lakens, D., Hennig, C., Morey, R. D., Homer, S., Gelman, A., Sprenger, J., & Wagenmakers, E. J. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician, 73*(sup1), 328–339. https://doi.org/10.1080/00031305.2019.1565553

Venn, J. (1876). *The logic of chance*. Macmillan and Co.

von Mises, R. (1957). *Probability, statistics and truth*. Allen & Unwin.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wheeler, G., & Williamson, J. (2011). Evidential probability and objective Bayesian epistemology. In *Philosophy of statistics* (pp. 307–331). Elsevier. https://doi.org/10.1016/b978-0-444-51862-0.50009-5

Williamson, J. (2007). Motivating objective Bayesianism: From empirical constraints to objective probabilities. In W. Harper, & G. Wheeler (Eds.), *Probability and inference: Essays in Honour of Henry E. Kyburg, Jr., Texts in Philosophy* (vol. 2, pp. 155–183). College Publications.

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press.
Winsberg, E. (2009). Computer simulation and the philosophy of science. *Philosophy Compass, 4*(5), 835–845. https://doi.org/10.1111/j.1747-9991.2009.00236.x