Statistical Methods & Applications (2024) 33:439–469
https://doi.org/10.1007/s10260-023-00732-y

ORIGINAL PAPER

Multinomial Thompson sampling for rating scales and prior considerations for calibrating uncertainty

Nina Deliu[1,2]

Accepted: 27 October 2023 / Published online: 6 December 2023

Abstract
Bandit algorithms such as Thompson sampling (TS) have been put forth for decades as useful tools for conducting adaptively-randomised experiments. By skewing the allocation toward superior arms, they can substantially improve particular outcomes of interest for both participants and investigators. For example, they may use participants' ratings for continuously optimising their experience with a program. However, most of the bandit and TS variants are based on either binary or continuous outcome models, leading to suboptimal performances in rating scale data. Guided by behavioural experiments we conducted online, we address this problem by introducing *Multinomial-TS* for rating scales. After assessing its improved empirical performance in unique optimal arm scenarios, we explore potential considerations (including prior's role) for calibrating uncertainty and balancing arm allocation in scenarios with no unique optimal arms.

Keywords Adaptive experiments · Thompson sampling · Multi-armed bandits · Rating scales · Multinomial model · Dirichlet distribution · Incomplete learning

1 Introduction

Well-designed randomised experiments such as randomised controlled trials (RCTs) are considered the "gold standard" for evaluating and comparing interventions at the end of the study (Akobeng 2005; Rosenberger et al. 2019; Kim et al. 2021). The standard protocol for conducting an RCT is based on following a predefined design that does not allow modifications of the trial without approved amendments (Pallmann et al. 2018). For example, interventions are assigned with prespecified (typically equal) randomisation probabilities that cannot be changed *during the course of*

✉ Nina Deliu
nina.deliu@uniroma1.it

[1] Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF), Sapienza Università di Roma, Rome, Italy

[2] MRC - Biostatistics Unit, University of Cambridge, Cambridge, UK

the study to direct participants to the most promising options. This approach ensures strong statistical guarantees in terms of bias and type-I error control, among other properties (see e.g., Villar et al. 2015; Williams et al. 2021; Robertson et al. 2023), but allows no flexibility for data-driven *online* alterations. Adaptively-randomised experiments have the potential to utilise accumulating information within the trial to modify its course (randomisation probabilities, sample size, etc.) in accordance with predetermined rules (Pallmann et al. 2018). To illustrate, response-adaptive randomisation schemes may use participants' responses to skew the allocation toward more efficient or informative interventions with the aim of assigning superior ones to as many participants as possible (Robertson et al. 2023). While collecting high-quality data, these types of experiments can result in a more flexible, efficient, and ethical alternative compared to traditional randomised studies (Bothwell et al. 2018; Pallmann et al. 2018).

As a concrete example, consider a platform such as Netflix or Spotify: instead of just showing users their own TV show/music preferences (or display random content), the system's goal is to continuously interact with users to learn and offer them special recommendations that may enhance their overall experience and engagement with the platform. Clearly, to be able to efficiently provide the most ideal content, the outcomes of interest (e.g., some measure of appreciation or engagement) must be promptly observed so as to adjust to users' preferences on a continuous basis. It is thus becoming increasingly common to incorporate *preference information* that may provide insights on a longer-term outcome of interest such as engagement. Netflix has actually discovered significant business value in collecting users' feedback to personalise their movie experience (Amatriain and Basilico 2015). Similarly, university instructors saw great value in using student ratings to give them better or preferred explanations of a concept to enhance their understanding (Williams et al. 2018). In the healthcare setting, relying on feedback collected through mobile-health technologies is increasingly practised for healthcare delivery (Figueroa et al. 2022; Liu et al. 2023) and to increase medication adherence (Gandapur et al. 2016).

Starting from the pioneer work of Thompson (1933), *multi-armed bandit* (MAB) algorithms (Lattimore and Szepesvári 2020) have been argued for decades as useful tools to adaptively randomise experiments. This framework provides a succinct abstraction of the trade-off between *exploration* (learning enough information about the different options or *arms*) and *exploitation* (selecting the most promising arm(s) so far), inherent in many *online* sequential decision-making problems with incomplete knowledge. Specifically, in MAB problems, a *learner* or *decision-maker* must repeatedly select an arm from a given set of alternatives, say $A_t \in \mathscr{A}$, in an *online* manner for each round $t = 1, \ldots, T$, with T finite or infinite. After selecting an arm A_t, a numerical *reward* $Y_t(A_t) \in \mathbb{R}$ associated with that arm is observed, ideally before the next round. A typical goal in MAB problems is to learn how to efficiently use observations from previous rounds to improve decision making so as to maximise–under uncertainty on the best arm–the *expected total reward* $\mathbb{E}\left[\sum_{t=1}^{T} Y_t(A_t)\right]$.

Several algorithms have been proposed for the MAB problem; for a survey, we refer to Lattimore and Szepesvári (2020). In this work, we focus on a highly interpretable, computationally efficient, and asymptotically optimal (Agrawal and Goyal

2017) strategy originally introduced in the clinical trial arena: *Thompson sampling* (TS; Thompson 1933; Russo et al. 2018). Due to its competitive empirical and theoretical properties (see e.g., Chapelle and Li 2011; Agrawal and Goyal 2017, for details), TS is nowadays receiving renewed attention in several domains, including online recommendation (Chapelle and Li 2011), economics/finance (Charpentier et al. 2023), education (Williams et al. 2018), and healthcare (Deliu et al. 2023; Figueroa et al. 2021).

There exists a wide array of TS variants (see e.g., Russo 2016; Kasy and Sautmann 2021; Li et al. 2022); however, with a few exceptions, most of them are based on binary or continuous outcomes, with a persistent shortage of appropriate solutions for rating scale data. In such cases, typical practices consist in either dichotomising the ordinal reward variable according to a (often arbitrary) cutoff and then using a binary model (Williams et al. 2018), or using a Normal model directly on the rating outcome (Deliu et al. 2021; Parapar and Radlinski 2021). Such practices have long been recognised as suboptimal in terms of both reward efficiency (Williamson and Villar 2020) and statistical inference (Altman and Royston 2006).

Motivating example This work is directly motivated by a series of exploratory studies aimed at evaluating the feasibility of MAB algorithms to develop intelligent adaptive systems that continuously improve user experiences through their ratings. Without loss of generality and only for illustrative purposes, here we report on a two-armed behavioural experiment (*MTurk I*) we conducted on Amazon Mechanical Turk, an online platform widely used by academics as a quick and inexpensive means of collecting experimental data (Mason and Suri 2012). Participants were recruited online, upon invitation to complete the survey with a compensation of USD $10 per user. Among other personal information (such as age and gender), participants were asked to provide a *rating* defined on a 7-point Likert scale for two types of messages related to mood and mental health. In the *MTurk I* experiment, a traditional (balanced) randomised design was implemented with the aim of learning about arm effectiveness. Overall, $T = 110$ users have participated in the study, with $T_1 = 58$ and $T_2 = 52$ users receiving arm 1 ("Today is a new day to start fresh") and arm 2 ("Let the past make you better, not bitter"), respectively. A summary of the ratings provided to each arm is reported in Fig. 1, which shows a small superiority of arm 1 (sample mean of $\hat{\mu}_1 = 5.81$ vs $\hat{\mu}_2 = 5.08$). Subsequent experiments have been conducted adaptively using TS, but, despite the rating outcome, they were *all* based on a conventional Normal model, therefore motivating this work.

Our contribution and related work The contribution of this paper is threefold. First, we introduce *Multinomial Thompson sampling* (*Multinomial-TS*), a TS version specifically designed for rating scale data. The existence of a conjugate prior–namely the *Dirichlet* distribution (Kotz et al. 2000)–for this exponential family, allows an easy implementation of the proposed algorithm through its closed-form posteriors. Second, guided by the empirical behaviour of *Multinomial-TS* in particular skewed distributions, we also introduce *Multinomial-TS with augmented support*, a variation of *Multinomial-TS* that, based on a simple trick on priors' support, improves the ability of the algorithm to balance arms allocation when an optimal arm does not exist. Further investigations on how to better calibrate uncertainty in such a scenario are conducted by studying the role of

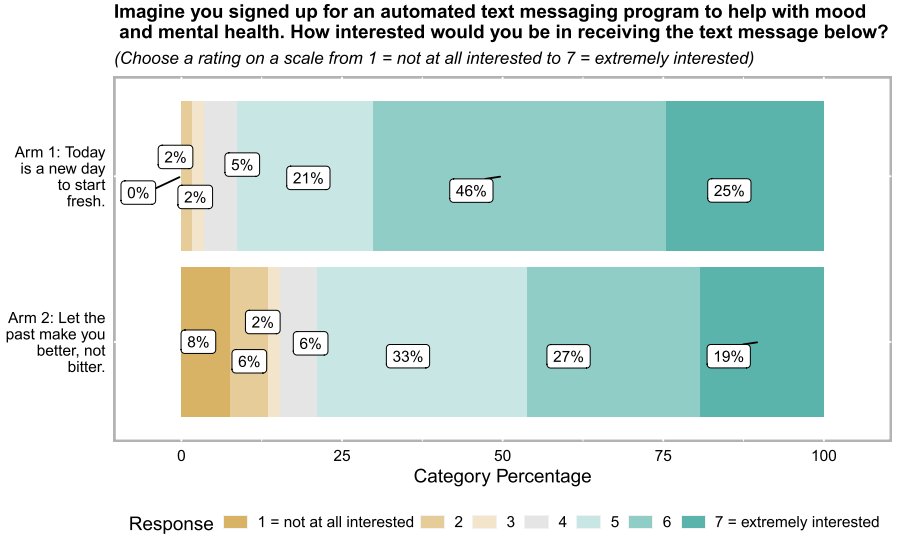

Fig. 1 Arm ratings distribution in the two-armed *MTurk I* experiment, defined on a 7-point Likert scale. A small superiority of arm 1 vs arm 2 is shown: sample means $\hat{\mu}_1 = 5.81$ vs $\hat{\mu}_2 = 5.08$, sample variances $\hat{\sigma}_1 = 1.04$ vs $\hat{\sigma}_2 = 1.72$, for an overall sample size $N = 110$

the prior distribution. As we discuss in Sects. 3 and 4, these considerations offer a potential solution for mitigating the so-called *incomplete learning* phenomenon (see e.g., Kalvit and Zeevi 2021), which refers to the inefficiency that many MAB algorithms, including TS, may have when exploring the arms. Finally, we evaluate the possibility of redesigning the motivating *MTurk I* experiment using the different *Multinomial-TS* variants and discuss their benefits and drawbacks. This paper is an extended version of the short paper presented at the 51st Scientific Meeting of the Italian Statistical Society on June, 2022 (Deliu 2022).

It should be noted that a few other works have studied TS under a multinomial model (Agrawal et al. 2022; Riou and Honda 2020; Zhang et al. 2021); however, they all differ from our work in several aspects, including the problem of interest. Specifically, Zhang et al. (2021) and Agrawal et al. (2022) both study a *selection problem*, where a decision system is faced with one choice among a set of K arms (online products and radio codebooks, respectively). Although these K arms may be considered analogous to the number of points of a Likert scale, we emphasise that in our (rating scale) problem each arm is itself defined on a scale, and the scale has an ordered nature. Riou and Honda (2020) cover a more general framework that can also suit our problem, but their work assumes bounded rewards in [0, 1]. Furthermore, it is worth noting that all of these works are primarily interested in regret analysis, performed under the assumption that there exists a unique optimal arm. No evaluations or considerations are made in scenarios with equal optimal arms, which represents a common scenario in many realistic applications, from image classification to personalised medicine (see e.g., Berry et al. 1997, for examples).

Structure of the work The paper is organised as follows: in Sect. 2, we introduce the general adaptive K-armed experimental setup along with the TS algorithm. The proposed *Multinomial-TS* strategy is detailed in Sect. 2.1 and its empirical performance is evaluated in Sect. 3. Section 4 is devoted to examining two potential strategies for calibrating the incomplete learning phenomenon and mitigating uncertainty within *Multinomial-TS*: one based on considerations about the skewness of the distribution (Sect. 4.1) and another one based on the information carried out by the prior distribution (Sect. 4.2). The potential of applying the proposed *Multinomial-TS* variants for the *MTurk I* experiment is discussed in Sect. 5. Finally, Sect. 6 presents concluding remarks and some research directions for future work.

2 Problem setting and methods

Experimental setup Consider a general K-armed experiment defined over a finite horizon T, in which N participants are accrued in a fully-sequential way, with an experimental size $N = T$. At accrual, each participant $t = 1, \dots, T$ is assigned to one of the K available arms $A_t \in \{1, \dots, K\}$ and subsequently an outcome $Y_t(A_t)$ associated with the assigned arm A_t is observed before the next round $t + 1$. We assume that $Y_t(A_t), t = 1, \dots, T$, does not depend on individual characteristics but only on arms, although our results may be generalised to a more general contextual MAB setting (Lattimore and Szepesvári 2020). The arms are drawn according to a policy $\boldsymbol{\rho_t} \doteq \{\rho_{t,k}, k = 1, \dots, K\}$, where $\rho_{t,k}$ is the allocation probability of arm k in the round t. Given the history of selected arms and associated rewards, say $\mathscr{H}_t \doteq \{A_\tau, Y_\tau(A_\tau), \tau = 1, \dots, t\}$, the goal is to find an (optimal) allocation policy so as to maximise the expected cumulative reward over the horizon T. Resembling this experimental setup, the MAB paradigm is framed on the efficient use of observations from previous rounds for estimating arm reward distributions and choosing which arm to select in the future. The most common measure of performance in MABs is the ability to maximise, under uncertainty on the best arm, the *expected total reward* $\mathbb{E}\left[\sum_{t=1}^{T} Y_t(A_t)\right]$, e.g., cumulative ratings provided by users over time. Equivalently, the goal is to minimise the so called *expected total regret* (more simply, total regret), i.e., how much we expect to regret in not knowing/selecting the optimal arm, when one exists. Formally, considering a stationary setting in which the mean reward associated with each arm does not change over different rounds, i.e., $\mathbb{E}[Y_t(A_t)] = \mathbb{E}[Y(A_t)] = \mu_{A_t}$, and the optimal arm $A_t^* = A^* \doteq \arg\max_{k \in \mathscr{A}} \mathbb{E}(Y(A_t = k)) = \arg\max_{k \in \mathscr{A}} \mu_k, \forall t$, we have

$$\text{Total Regret} = \mathbb{E}\left(\sum_{t=1}^{T} Y_t(A^*)\right) - \mathbb{E}\left(\sum_{t=1}^{T} Y_t(A_t)\right) = T\mu^* - \sum_{t=1}^{T} \mu_{A_t}, \tag{1}$$

with $\mu^* = \mu_{A^*} = \max_{k \in \mathscr{A}} \mu_k$. Basically, regret is defined as the difference between the maximum possible reward attainable and the reward resulting from the arms selected over the horizon T.

In such a setting, outcomes can be considered independent draws from a fixed but unknown distribution with the following conditional mean:

$$\mathbb{E}\big(Y_t(A_t) \mid \mathscr{H}_{t-1}\big) = \mathbb{E}(Y \mid A_t) = \sum_{k=1}^{K} \mu_k \mathbb{I}(A_t = k),$$

where $\{\mu_k, k = 1, \dots, K\}$ are the unknown arm parameters.

The focus of this work is on rating scale outcomes. Without loss of generality, we consider a J-point Likert scale with $Y_t \in \mathscr{Y} \doteq \{1, 2, \dots, J\}$ for $t = 1, \dots, T$, where higher values translate into better outcomes. Given the specific rating scale setting, we consider a data generation process that occurs according to a multinomial distribution such that $\mu_k = \sum_{j=1}^{J} j p_{j,k}$ for $k = 1, \dots, K$, where $p_{j,k} = \mathbb{P}(Y(A_t = k) = j), j = 1, \dots, J, k = 1, \dots, K$ are the unknown parameters defining the multinomial family. This will be discussed in more detail in Sect. 2.1.

Thompson sampling Rooted in a Bayesian framework, TS defines arm allocations in terms of their posterior probability of being associated with the maximum expected reward at each round $t = 1, \dots, T$. In a K-armed setting, denoted by $\rho_{t,k}^{\text{TS}}$ the TS probability of allocating arm k at round t, this is given by:

$$\begin{aligned}
\rho_{t,k}^{\text{TS}} &= \mathbb{P}\Big(\mathbb{E}\big(Y_t(A_t = k)\big) \geq \mathbb{E}\big(Y_t(A_t = k')\big), \forall k' \mid \mathscr{H}_{t-1}\Big) \\
&= \mathbb{P}\Big(\mu_k \geq \mu_{k'}, \forall k' \mid \mathscr{H}_{t-1}\Big) \\
&= \int_{\Omega_1 \times \cdots \times \Omega_K} \mathbb{I}[\mu_k \geq \mu_{k'}, \forall k'] \prod_{k=1}^{K} \pi_t(\mu_k \mid \mathscr{H}_{t-1}) d\mu_1 \times \cdots \times d\mu_K,
\end{aligned} \tag{2}$$

with π_t the posterior distribution of arm means, and Ω_k the parameter space of $\mu_k, \forall k$.

With a few exceptions (e.g., the Normal model), the exact computation of the quantity in Eq. (2) is not feasible. Thus, the typical way to implement TS (Russo et al. 2018) involves drawing at each round t a sample from the posterior distribution of the parameters of each of the unknown arms, and then selecting the arm associated with the highest posterior estimated mean reward, say $\tilde{\mu}_{tk} = \mathbb{E}(\tilde{Y}_t \mid A_t = k)$, that is,

$$\tilde{a}_t \doteq \text{argmax}_{k=1,\dots,K} \mathbb{E}(\tilde{Y}_t \mid A_t = k) = \text{argmax}_{k=1,\dots,K} \tilde{\mu}_{tk}. \tag{3}$$

To guarantee computationally efficient posterior sampling, given the repeated implementation of the strategy for each round t, a conjugate family is generally considered, with the most common distributional assumptions for the reward variable being the Bernoulli and the Normal model (Russo et al. 2018).

The TS algorithm is recognised as an asymptotically optimal strategy (Agrawal and Goyal 2017), meaning that it matches the asymptotic lower bound of the regret metric in Eq. (1) for adaptive allocation schemes. We refer to Lai and Robbins (1985) for the expression of the lower bound, which is beyond the scope of this paper.

2.1 Proposal: multinomial Thompson sampling

The multinomial distribution is the extension of the binomial distribution for categorical outcomes with more than two response categories (Agresti 2019). Consider a fixed number of J mutually exclusive categories of an outcome $Y_t, t = 1, \dots, T$, among which, one and only one category is observed at each round t, i.e., $\sum_{j=1}^{J} \mathbb{I}(Y_t = j) = 1$, with $\sum_{t=1}^{T} \sum_{j=1}^{J} \mathbb{I}(Y_t = j) = T$. Similarly to the binomial, the multinomial distribution models the probability of counts $X_j = \sum_{t=1}^{T} \mathbb{I}(Y_t = j), \forall j \in \{1, \dots, J\}$, over T trials. Denoted by Multinom(T;$\boldsymbol{p}$), with $T > 0$ the number of trials and $\boldsymbol{p} = (p_1, \dots, p_J) = \Big(\mathbb{P}(Y = 1), \dots, \mathbb{P}(Y = J)\Big)$ the unknown model parameters belonging to the standard simplex $\mathscr{P}^J = \{p \in [0, 1]^J : \sum_{i=1}^{J} p_j = 1\}$, its probability mass function is given by

$$f(x_1, \dots, x_J; T; p_1, \dots, p_J) = \left(\frac{T!}{\prod_{j=1}^{J} x_j!}\right) \prod_{j=1}^{J} p_j^{x_j}, \tag{4}$$

where $x_j = \{0, \dots, T\}$, for all $j = 1, \dots, J$, and $\sum_{j=1}^{J} x_j = T$. The expected number of times category j is observed over T trials is $\mathbb{E}(X_j) = Tp_j$, with variance $\mathbb{V}(X_j) = Tp_j(1 - p_j)$ and covariance $\mathbb{C}ov(X_j, X_i) = -Tp_jp_j$, for $i \neq j$. Theorem 1 bridges these properties of the multinomial family to the distribution of interest.

Theorem 1 *Assuming a fixed number J of mutually exclusive ordinal and real valued categories, and denoting with $Y_t, t = 1, \dots, T$, the category variable, this can be expressed as*

$$Y_t = \sum_{j=1}^{J} jX_j, \quad \mathbf{X} \sim \text{Multinom}(1; \boldsymbol{p}), \quad t = 1, \dots, T,$$

with

$$\mu_t = \mathbb{E}[Y_t] = \sum_{j=1}^{J} jp_j, \tag{5}$$

$$\mathbb{V}(Y_t) = \sum_{j=1}^{J} j^2 p_j(1 - p_j) - \sum_{i \neq j} ijp_ip_j. \tag{6}$$

The proof of Theorem 1 is based on noticing that in each single trial (consider, for simplicity, $T = 1$), $X_j = \{0, 1\}$, for all $j = 1, \dots, J$, with $\sum_{j=1}^{J} X_j = 1$. The expectation in Eq. (5) as well as the variance in Eq. (6) follow straightforwardly from operator properties and known results on the multinomial distribution (Agresti 2019).

We emphasise the key role of this result in light of the algorithm under study. *Multinomial-TS* is indeed implemented by using the result in Eq. (5) to directly compute the posterior mean outcome, as required by the definition in Eq. (2).

Posterior updates are based on the Dirichlet connection to multinomial-distribution counts. In fact, Eq. (4) can also be expressed using the gamma function Γ, directly showing its similarity to the Dirichlet distribution, its conjugate prior. Given a parameter vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)$, with $\alpha_j > 0, \forall j = 1, \dots, J$, the Dirichlet distribution, denoted by $\text{Dir}(\boldsymbol{\alpha})$, models our knowledge on the unknown parameters $(p_1, \dots, p_J)$ of the multinomial as:

$$f(p_1, \dots, p_J; \alpha_1, \dots, \alpha_J) = \left(\frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \right) \prod_{j=1}^{J} p_j^{\alpha_j - 1}, \tag{7}$$

where again $\boldsymbol{p}$ belongs to the standard $J-1$ simplex $\mathscr{P}^J = \{p \in [0,1]^J : \sum_{i=1}^{J} p_j = 1\}$.

We then iteratively update our beliefs about unknown parameters $\boldsymbol{\alpha}$ based on the observed outcome Y_t at each round t. Specifically, for $t = 1, \dots, T$, the posterior distribution of arms is defined by the vector $\boldsymbol{\alpha}$ with elements $\alpha_j \leftarrow \alpha_j + c_j$, $\forall j = 1, \dots, J$, where $c_j = \mathbb{I}(y_t = j)$ indicates whether the category j was observed at time t. The update is made independently for each arm and changes occur only for the selected one. The pseudocode of *Multinomial-TS* is given in Algorithm 1.

3 Simulation studies

For our empirical evaluation, we start with a set of simulation studies based on the setup introduced in Sect. 2, with $T = 1000$, $J = 7$, and $K = 2$. We consider four scenarios defined by the following data-generation processes.

Require: Time horizon T, number of categories of the rating scale outcome J, number of study arms K, prior parameters $\boldsymbol{\alpha_k} = (\alpha_{1k}, \dots, \alpha_{Jk})$ for each arm $k = 1, \dots, K$.

for $t = 1, 2, \dots, T$ **do**
  for $k = 1, \dots, K$ **do**
    Sample $\tilde{\boldsymbol{p}}_k = (\tilde{p}_{1k}, \dots, \tilde{p}_{Jk}) \sim \text{Dir}(\alpha_{1k}, \dots, \alpha_{Jk})$
    Compute $\tilde{\mu}_{tk} = \sum_{j=1}^{J} j \tilde{p}_{jk}$
  end for
  Select arm $\tilde{a}_t = \text{argmax}_{k=1,\dots,K} \tilde{\mu}_{tk}$ and observe the reward y_t
  for $j = 1, \dots, J$ **do**
    Compute $c_{jk} = \mathbb{I}(y_t = j, \tilde{a}_t = k)$
  end for
  for $k = 1, \dots, K$ **do**
    Update posteriors: $(\alpha_{1k}, \dots, \alpha_{Jk}) \leftarrow (\alpha_{1k}, \dots, \alpha_{Jk}) + (c_{1k}, \dots, c_{Jk})$
  end for
end for

Algorithm 1 *Multinomial-TS*

Scenario 1: Existence of a unique optimal arm ($H_1 : \mu_1 > \mu_2$). In line with the *MTurk I* example in Fig. 1 ($\mu_1 = 5.8$; $\mu_2 = 5.08$), we generate $Y_t, t = 1, \ldots, T$ as

$$Y_t \mid (A_t = k) = [1, \ldots, 7] \times \mathbf{X}_k, \quad \mathbf{X}_k \sim \text{Multinom}(1; \mathbf{p}_k),$$
$$\mathbf{p}_k = \begin{cases} (0.00, 0.02, 0.02, 0.05, 0.21, 0.45, 0.24) & k = 1, \\ (0.08, 0.06, 0.02, 0.06, 0.33, 0.27, 0.19) & k = 2. \end{cases}$$

Scenario 2: Identical arms ($H_0 : \mathbf{p}_1 = \mathbf{p}_2$)*—symmetric distribution.* Following the normal approximation of the binomial distribution, we define $\mathbf{p}_k$ so as to resemble a (symmetric) Normal distribution with parameters $\mu_k = 4$ (to get symmetry over the 7 categories) and $\sigma_k^2 = 1.7$ (aligned with the *MTurk I* study), for $k = 1, 2$. Specifically, we generate $Y_t, t = 1, \ldots, T$ as

$$Y_t \mid (A_t = k) = [1, \ldots, 7] \times \mathbf{X}_k, \quad \mathbf{X}_k \sim \text{Multinom}(1; \mathbf{p}_k),$$
$$\mathbf{p}_k = (0.02, 0.09, 0.23, 0.31, 0.23, 0.09, 0.02) \quad k = 1, 2.$$

Scenario 3: Identical arms ($H_0 : \mathbf{p}_1 = \mathbf{p}_2$)*—right-skewed distribution.* To obtain right-skewed data we consider $\mu_1 = \mu_2 = 3$ and generate $Y_t, t = 1, \ldots, T$ as

$$Y_t \mid (A_t = k) = [1, \ldots, 7] \times \mathbf{X}_k, \quad \mathbf{X}_k \sim \text{Multinom}(1; \mathbf{p}_k),$$
$$\mathbf{p}_k = (0.2, 0.3, 0.15, 0.15, 0.1, 0.05, 0.05) \quad k = 1, 2.$$

Scenario 4: Identical arms ($H_0 : \mathbf{p}_1 = \mathbf{p}_2$)*—left-skewed distribution.* To obtain left-skewed data we consider $\mu_1 = \mu_2 = 5$ and generate $Y_t, t = 1, \ldots, T$ as

$$Y_t \mid (A_t = k) = [1, \ldots, 7] \times \mathbf{X}_k, \quad \mathbf{X}_k \sim \text{Multinom}(1; \mathbf{p}_k),$$
$$\mathbf{p}_k = (0.05, 0.05, 0.1, 0.15, 0.15, 0.3, 0.2) \quad k = 1, 2.$$

To disentangle the individual role of mean, variance, and skewness, we relax the null cases of identical arms $H_0 : \mathbf{p}_1 = \mathbf{p}_2$. Specifically, an extended number of scenarios with different probability mass functions $\mathbf{p}_1 \neq \mathbf{p}_2$, but identical means $\mu_1 = \mu_2$ are sampled from the polytope of discrete distributions with mean $m \in \{1, 2, \ldots, 7\}$. These results are reported in Appendix B.

We evaluate the proposed *Multinomial-TS* using a uniform prior over simplexes for each of the two arms, i.e., $Dir(\boldsymbol{\alpha}_k = \mathbf{1})$ for $k = 1, 2$. Its performance is then compared with alternative modelling options that are part of the current common practice (see Sect. 1). More specifically, we consider:

- *Normal-TS*, assuming weakly-informative and identical priors $N(\mu_k = 4, \sigma_k^2 = 100)$, $k = 1, 2$, in all scenarios. Posterior arm means are updated following the conjugacy of the Normal reward model, with unknown means and known variances. The variances are set according to the formulation reported in Eq. (6) for each scenario.
- *Binary-TS* with a cutoff $y_c = 6$, meaning that a success occurs when $Y_t \geq 6$. Uniform and identical Beta($\alpha_k = 1, \beta_k = 1$), $k = 1, 2$, priors are assumed in all scenarios. Posterior arm means are updated following the conjugacy of the Beta-Bernoulli model, with the outcome variable following a Bernoulli distribution.

- *Binary-TS* with a cutoff $y_c = 7$, such that only the highest category $j = 7$ is considered a success. Uniform and identical Beta($\alpha_k = 1, \beta_k = 1$), $k = 1, 2$, priors are assumed in all scenarios. Posterior arm means are updated following the conjugacy of the Beta-Bernoulli model, with the outcome variable following a Bernoulli distribution.
- An *Oracle* that is assumed to know the underlying truth, which always assigns the arm with the highest mean reward (Besbes et al. 2014) when this exists. Under identical arm scenarios, the Oracle allocates arms with equal probability.

Regret and Optimal Arm Allocation—Scenario 1 ($H_1 : \mu_1 > \mu_2$) We evaluate standard bandit performances under a setting in which an optimal arm, defined as the one yielding the maximum outcome, exists and it is unique. Note that in a setting with equal arm distributions, thus equal arm means, regret becomes meaningless by definition as $T\mu^* - \sum_{t=1}^{T} \mu_{A_t} = 0$ when $\mu_k = \mu^*$, for all k. Empirical results for this setting are shown in Fig. 2, highlighting the increased performance of the proposed *Multinomial-TS* over both *Normal-TS* and *Binary-TS*.

Notably, *Binary-TS* results to be highly sensitive to changes in the cutoff value y_c, with greater values remarkably impacting regret and best-arm allocation. As a result, when the extreme category $j = 7$ is chosen as the cutoff value, *Binary-TS* focusses on discriminating between upper extreme values vs all other values (that is, $Y_t = 7$ vs $Y_t < 7$). The outcomes $Y_t < 7$ are treated equally, although they contain important information on arms distributions.

Arms Allocation—Scenario 2 ($H_0 : \mathbf{p}_1 = \mathbf{p}_2$; ***symmetric distribution)*** While in an identical arm setting with $\mathbf{p}_1 = \mathbf{p}_2$ regret is not of interest, it is instead

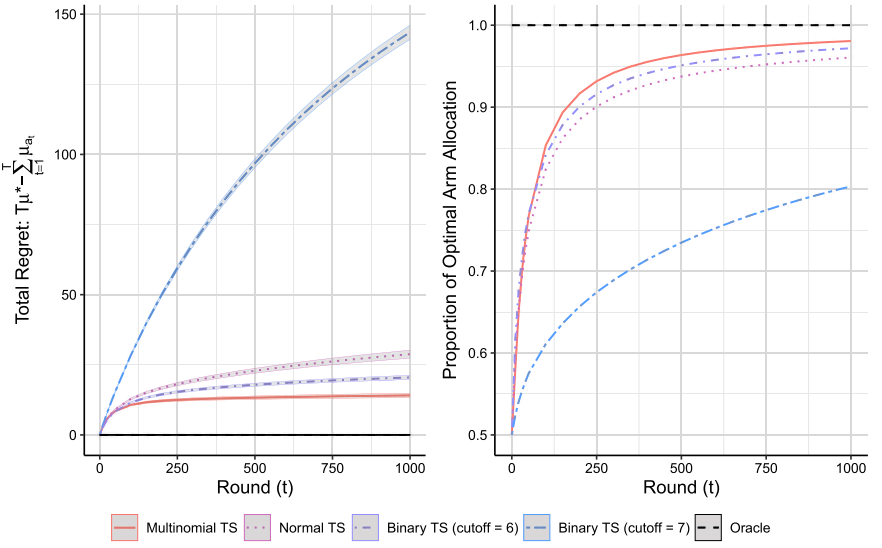

Fig. 2 Regret and proportion of optimal arm allocation in the proposed *Multinomial-TS* vs *Normal-TS and Binary-TS*. Values are obtained by averaging across 10^4 independent TS trajectories

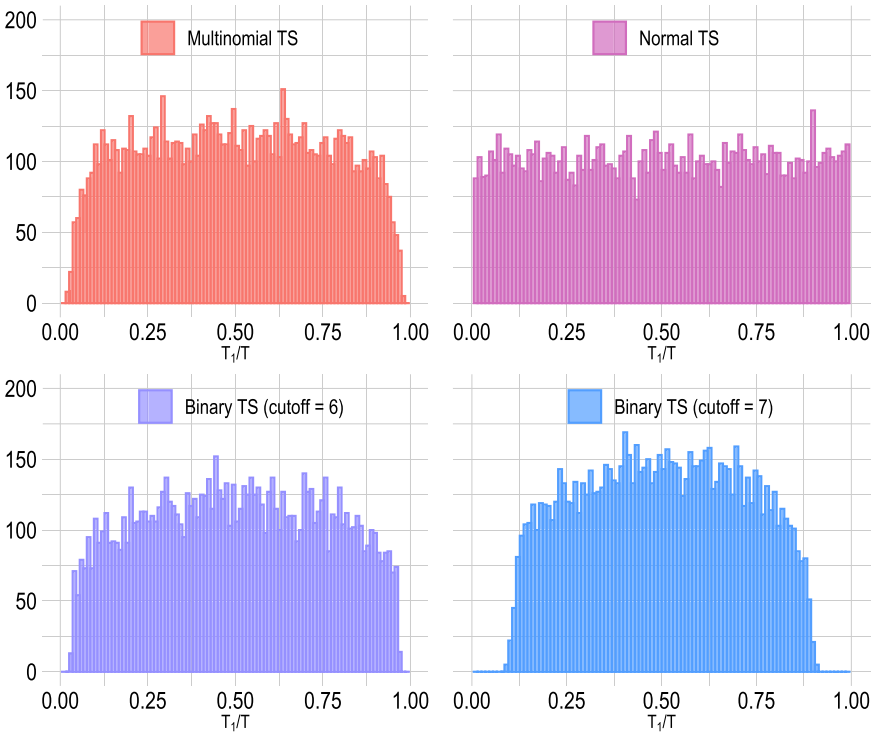

Fig. 3 Empirical allocation (T_1/T, with T_1 being the number of times arm 1 is allocated) under the identical arm case (Scenario 2). Values are obtained by averaging across 10^4 independent TS trajectories

helpful to understand how an algorithm balances the allocation of one arm over the other one when no one should be exclusively preferred. Results in Fig. 3 show that, differently from the Oracle that in such a case would assign arms with equal probability (by definition), TS would continuously search for an optimal arm, with the aim, and often the result, of assigning it more often only by chance. This intrinsic characteristic is reflected in all modelling strategies and is exacerbated in the Normal case (see Fig. 3; top-right plot).

In fact, in the Normal case, the empirical allocation of arms does not only not concentrate (does not converge to a unique value), but its distribution closely resembles a uniform distribution in the probability interval [0, 1]. This behaviour results in heavily unbalanced allocations in favour of one of the two arms–eventually fixing on selecting only a single arm–even when no underlying differences between arms exist. We emphasise that this is a well-studied phenomenon in MAB problems, referred to as *incomplete learning* (see e.g., Keskin and Zeevi 2018) and occurring when parameter estimates fail to converge to the true value. The main reason is the insufficient exploration of the arms, although recent work has pointed to some consequences of the sequential nature of data collection (see

e.g., Villar et al. 2015; Shin et al. 2019; Deshpande et al. 2018). As a result, when using standard statistical estimators, such as the sample mean, in adaptively collected data, unbiasness, consistency, and asymptotic normality are no longer guaranteed (Hadad et al. 2021), with negative impacts on hypothesis testing, that is, inflated type-I error and low power (Deliu et al. 2021).

Arms Allocation—Scenario 3 and Scenario 4 ($H_0 : \mathbf{p}_1 = \mathbf{p}_2$; ***skewed distribution)*** The skewness of the distribution and its direction can play a relevant role in the algorithm's behaviour and its resulting performances. This is particularly true for *Multinomial-TS*, which shows a significant sensitivity to the shape of the distribution in terms of its ability to balance arms allocation under a null (H_0) scenario. Specifically, it can be observed that, in the case of positive skewness (Scenario 3; Fig. 4), the arm allocation seems to concentrate, although with high variability, around the ideal $T_1/T = 1/2$ value, with T_1 being the number of times the arm 1 is allocated. We recall that this is the result that we would observe under the *Oracle* in a two-arm setting. However, for negatively skewed distributions (Scenario 4; Fig. 4), an opposite trend occurs, highlighting the incomplete learning phenomenon more intensely.

In fact, when looking at the empirical probabilities of extreme arm allocations in Scenario 4, these have values $\mathbb{P}\left(\frac{T_1}{T} > 0.95\right) = 8.7\%$ and $\mathbb{P}\left(\frac{T_1}{T} < 0.05\right) = 8.9\%$, which are substantial compared to the values $< 0.5\%$ characterising Scenario 3 (see Table 1).

Additional results reported in Appendix B support these findings and allow for a better understanding of the individual role of the standard deviation vs the skewness of the distribution. Specifically, the higher the variability of one arm compared to the other, the lower its empirical probability of being assigned an extreme number of times (this can be inferred from Table 2). Analogously, the

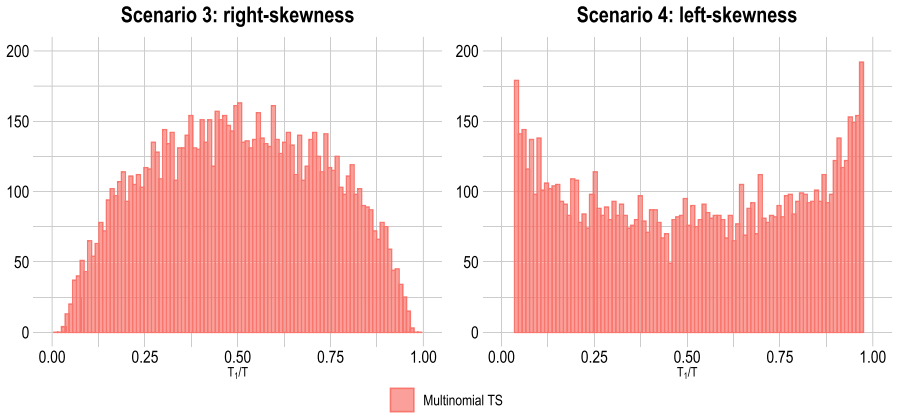

Fig. 4 Empirical allocation (T_1/T, with T_1 being the number of times arm 1 is allocated) under the identical arm case but with a non-symmetric distribution (Scenario 3: right-skewed distribution; Scenario 4: left-skewed distribution). Values are obtained by averaging across 10^4 independent TS trajectories

Table 1 Empirical probabilities (in percentage) of observing an arm allocation within a given range. Values are referred to *Multinomial-TS* and are obtained by averaging across 10^4 independent TS trials

	Scenario 2 (H_0; symmetric) (%)	Scenario 3 (H_0; right-skewed) (%)	Scenario 4 (H_0; left-skewed) (%)
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	1.1	0.2	8.9
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	5.2	2.3	15.4
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	11.6	14.6	8.0
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	23.6	28.6	15.8
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	5.3	2.5	15.3
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	1.2	0.3	8.7

higher the *left* skewness, the higher the chances of observing a large amount of extreme allocations.

Normal-TS and *Binary-TS* show similar patterns in these scenarios; however, the overall leaning is detrimental rather than ameliorable, particularly for *Binary-TS* (we refer to Figs. 16 and 17 in Appendix C).

Motivated by this distinctive nature of *Multinomial-TS*, in Sect. 4.1 we discuss a simple trick that may be useful in enhancing the safety of *online*-data collection when using this algorithm.

4 Mitigating the incomplete learning problem via prior considerations

As discussed in Sect. 3, the incomplete learning phenomenon characterising TS reveals that, in a context with unknown parameters, full knowledge of the parameter values may be precluded in the long run, with only one parameter (i.e., the one of the "optimal" arm) being consistently estimated. Although this behaviour may be considered acceptable when *one* optimal arm truly exists, it is detrimental and misleading in opposite cases, that is, when all arms are identical. Therefore, a good strategy should consider certain adjustments so that some *active experimentation* (Antos et al. 2008) is used to generate additional information about the undersampled parameters. We now explore two directions to address this challenge and better calibrate the uncertainty with the proposed *Multinomial-TS*.

4.1 Multinomial-TS with augmented prior support

As illustrated in Fig. 4 and Fig. 3, the more the reward distribution is positively skewed, the more *Multinomial-TS* is able to balance arm allocation under the null. This is depicted in: i) an allocation distribution with a peak at $1/K = 1/2$ and; ii) a very low probability of allocating arms with a proportion T_1/T equal to or close

to the extremes 0 or 1. This is shown both in Fig. 4 (left plot) and Table 1, where $\mathbb{P}(T_1/T < 0.05) = 0.02$ (in Scenario 3).

On the other side, when the reward distribution is negatively skewed (Scenario 4), arms allocation follow a U-shaped distribution (see Fig. 4-right plot), with peaks on the extremes, highlighting a relevant incomplete learning problem. In fact, in Scenario 4, we have that $\mathbb{P}(T_1/T < 0.05) = 0.08$, which is greater that the 0.05 probability occurring for a uniformly-distributed case. Such behaviour suggests a high instability and sensitivity of the algorithm to *relatively* high values of the outcome variable *with respect to its support*.

Motivated by this particular behaviour of *Multinomial-TS*, we introduce a simple algorithm modification that alters the skewness of the prior and, in turn, the posterior distribution. Note that within TS, the underlying skewness of the reward distribution is reflected in the arm posterior distributions. For example, a $Dir(\boldsymbol{\alpha})$ with a parameter vector with identical elements (e.g., $\boldsymbol{\alpha} = \mathbf{1}$) indicates uniformity, thus symmetry, over the support. We recall that the support of a Dirichlet $\{p_1, \dots, p_J\}$, with $p_j \in [0, 1]$ for each j and $\sum_{j=1}^{J} p_j = 1$, is the standard $(J-1)$-simplex specifying both the number of categories J and the set of their probability distributions. To induce some skewness in the distribution, it therefore suffices to increase the number of categories J, say from J to $J + s$, with s a nonnegative integer and $\alpha_i \neq \alpha_j$, for $i = J + 1, \dots, J + s$, $j = 1, \dots, J$. We term s the support-augmentation hyperparameter.

In the particular case of TS, where posteriors are iteratively updated as $\alpha_j \leftarrow \alpha_j + c_j$, with $c_j = \mathbb{I}(Y_t = j)$, the idea is to specify an augmented support of order $(J-1)+s$ for the $Dir(\boldsymbol{\alpha})$ prior. Given that $Y_t \in \{1, \dots, J\}$ for each $t = 1, \dots, T$, it is guaranteed by its possible realisations that the Dirichlet posteriors will be iteratively skewed to the right. In fact, ignoring for now the arm index, we have that:

$$\begin{aligned} \alpha_j + c_j &\geq \alpha_j, \quad j = 1, \dots, J, \\ \alpha_j + c_j &= \alpha_j, \quad j = J + 1, \dots, J + s, \end{aligned}$$

as $\mathbb{P}(Y_t > J) = 0, \forall t$, thus $c_j = 0, \forall j \in [J + 1, J + s]$.

We emphasise that, although such augmentation trick requires a change in the support of the prior distribution, involving thus a change in theoretical support of the reward model, the latter remains practically unaltered by virtue of $\mathbb{P}(Y_t > J) = 0, \forall t$. Also note that *Multinomial-TS* is a special case of this modified version-which we name *Multinomial-TS with augmented support*-when $s = 0$. The pseudocode of the latter is provided in Algorithm 2.

The plausibility of our solution is supported by an existing theoretical result (see Theorem 2) for *Binary-TS*, which nonetheless represents a particular case of *Multinomial-TS*, when $J = 2$ and $Y_t \in \{0, 1\}$, for all t.

Theorem 2 (Kalvit and Zeevi (2021)) *In a two-armed model where both arms yield rewards distributed as Bernoulli*(p), *the following holds under TS as* $n \to \infty$:

(I) *If* $p = 0$, *then* $\frac{T_1}{T} \to \frac{1}{2}$;
(II) *If* $p = 1$, *then* $\frac{T_1}{T} \to Unif[0, 1]$.

We conjecture that a similar result to Theorem 2–Case (I) applies to *Multinomial-TS* when the parameters are in the form $\mathbf{p_k} = (1, 0 \ldots, 0), \forall k$. Notice that since our aim is to balance allocation under the null scenarios, the asymptotic limit in Case (I) represents our main interest. Specifically, in a K-armed setting with identical arms yielding rewards that follow a *Multinom*($\mathbf{p}$) distribution, with $\mathbf{p} = (1, 0 \ldots, 0)$, we expect that, under *Multinomial-TS*, the allocation proportions will be such that $T_k/T \to 1/K$ as $n \to \infty$, for $k = 1, \ldots, K$. We support our conjecture with empirical evaluations for $K = 2, 3, 4$, where T_k/T is expected to converge to 1/2, 1/3 and 1/4, respectively. We refer to Fig. 5 and point to additional simulation studies explored in Appendix B (in particular, Fig. 15 and Table 7).

In light of this result, the augmentation trick detailed in this section is explored in Fig. 6, where we show the distribution of arm allocations of the modified *Multinomial-TS* for different values of the augmentation size s, including $s = 0$. Empirical results are based on Scenario 4, representing the worst-case scenario in terms of arm allocation under the null. As expected, the higher the hyperparameter s, the higher the overall balance in the allocation of the two arms. For $s = 20$, we already achieve more than satisfactory results, with $\mathbb{P}(T_1/T \in [0.45, 0.55]) = 78\%$ and $\mathbb{P}(T_1/T < 0.1) = \mathbb{P}(T_1/T > 0.9) = 0\%$, highlighting the value of the augmented-support solution to the incomplete learning phenomenon.

The computation cost in terms of running time of both *Multinomial-TS* and *Multinomial-TS with augmented support* is detailed in Appendix D. As depicted in Fig. 19 and Table 8, when compared to a standard *Multinomial-TS* ($s = 0$), the increasing (median) time complexity of *Multinomial-TS with augmented support* is negligible up to $s = 5$ (25.04 vs 26.79 milliseconds) and it is more than doubled for

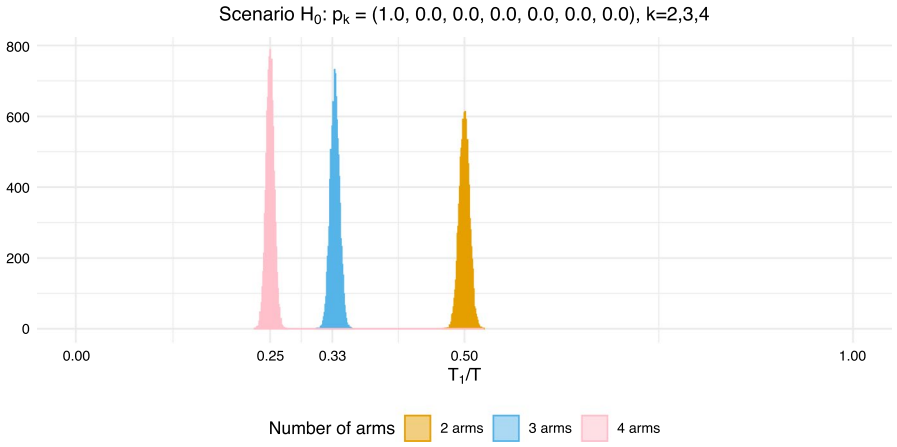

Fig. 5 Empirical allocation (T_1/T, with T_1 being the number of times arm 1 is allocated) under $H_0 : \mathbf{p_1} = \cdots = \mathbf{p_K}$, with $\mathbf{p_k} = (1, 0 \ldots, 0)$ and $k = 2, 3, 4$. Values refer to *Multinomial-TS* and are obtained by averaging across 10^4 independent TS trajectories

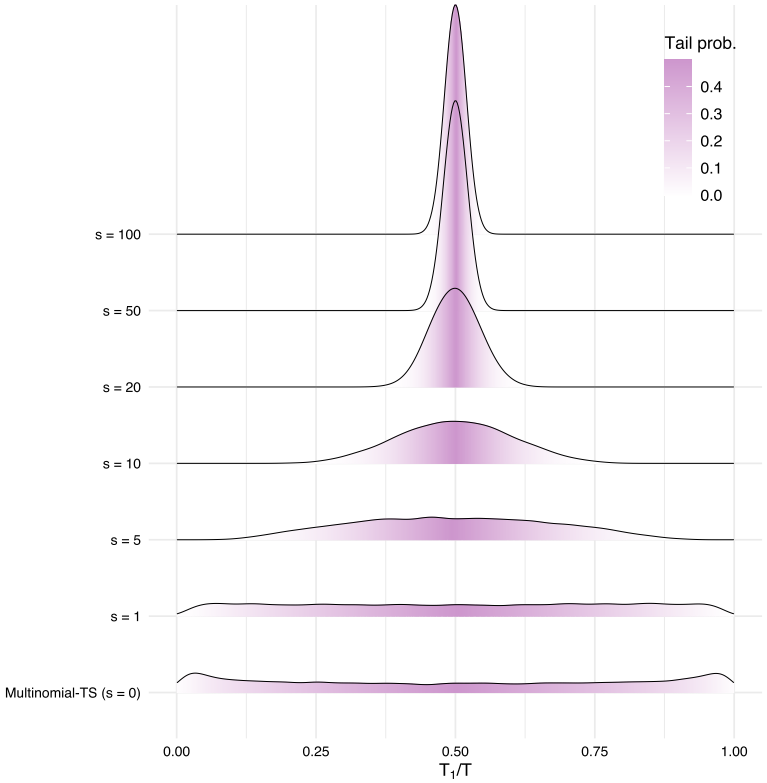

Fig. 6 Empirical allocation (T_1/T, with T_1 being the number of times arm 1 is allocated) under Scenario 4: identical arm means with a left-skewed distribution. Values refer to *Multinomial-TS with augmented support* (see Algorithm 2) and are obtained by averaging across 10^4 independent TS trajectories

$s = 100$ (25.04 vs 69.85 milliseconds). A similar trend is shown when J increases; notice that this is expected as J and s play the same role: they define the number of scale categories. Finally, the time complexity also increases with the number of arms K: in line with existing literature on TS (see e.g., Min et al. 2019), the run time is linear of order $\mathcal{O}(K)$; see also Fig. 20 in Appendix D.

4.2 Multinomial-TS with more informative priors

Useful insights into the general sensitivity of the algorithm to the choice of priors has been evaluated in previous literature (see e.g., Liu and Li 2016; Russo et al. 2018). Russo et al. (2018), for example, points out that ignoring any useful knowledge from past experience may increase the time it takes for TS to identify the most effective arms. Therefore, a careful choice of the prior can improve learning performance under H_1, or arm allocations under H_0. While the effects of the prior distribution should wash out as $T \to \infty$, they may be decisive at the beginning of the experiment, where TS reacts very sensitively to small variations in arm outcomes,

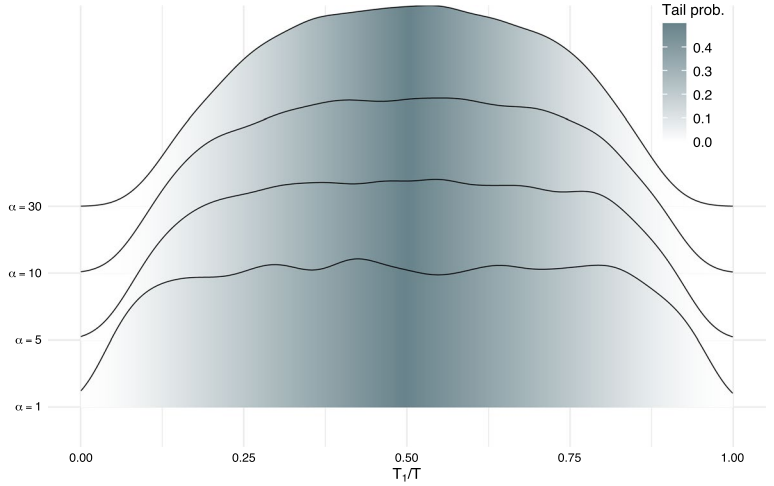

Fig. 7 Empirical allocation (T_1/T, with T_1 being the number of times arm 1 is allocated) under Scenario 2: identical arm means with symmetric distribution. Values refer to *Multinomial-TS* (see Algorithm 1) with different prior choices, and are obtained by averaging across 10^4 independent TS trajectories

even if these are simply dictated by noise. Thus, selecting a prior that still maintains a certain uniformity on the plausible response categories but is less variable may be helpful for the problem at hand. This is well illustrated in Fig. 7, where an increase in the value α_j, for $j = 1, \ldots, J$, leads to a decreased variability in the empirical allocation of the arms, with values concentrated mainly around $1/K = 1/2$. Note that, under uncertainty on the best arm, priors are chosen to be identical for both arms.

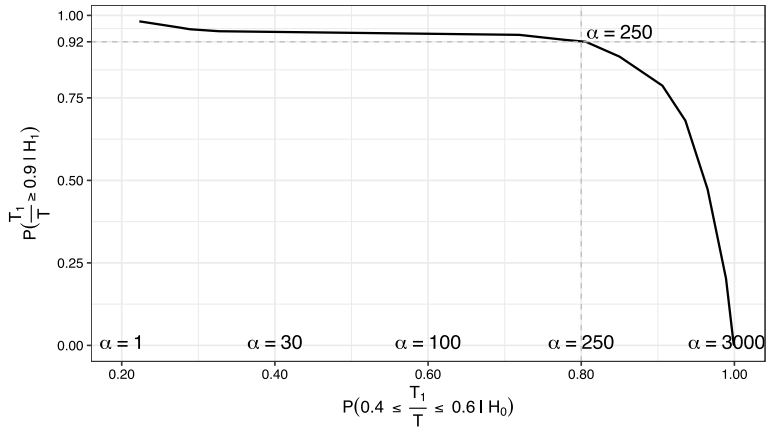

Fig. 8 Trade-off between between optimal arm allocation under H_1 (Scenario 1; y-axis) and arms balancing under H_0 (Scenario 2; x-axis). Comparisons are quantified with the empirical probability of observing an empirical allocation in a given range. Values refer to *Multinomial-TS* (see Algorithm 1) with different prior choices, and are obtained by averaging across 10^4 independent TS trajectories

Clearly, the higher the value of the hyperparameters α_j, the greater the extent of prior information carried over by the algorithm and the slower its ability to adapt to each new observation. Therefore, while beneficial in null scenarios, a natural consequence is a potential reduction in the overall reward efficiency under an alternative scenario, due to the higher weight assigned to the prior. To understand the overall benefit of using a more informative prior, we thus quantify the trade-off between optimal arm allocation under H_1 and arms balancing under H_0. Fig. 8 compares the empirical probability of observing an allocation proportion close to 1/2 when the arms are identical (H_0; Scenario 2) and of the optimal arm in a setting where an optimal arm exists (H_1; Scenario 1). Note the analogy between the curve in Fig. 8 and a ROC curve: here, the x- and y-axes may be interpreted in a similar fashion to sensitivity (true negative rate) and specificity (true positive rate), respectively. The ideal results are those lying in the right corner: both sensitivity and specificity close to 1.

As we can notice, the effect of the prior values is substantial under H_0, where the increasing values of α_j are increasingly translated into a higher probability of having a more balanced allocation. However, under H_1 the effect is less relevant: up to a value of $\alpha = 250$, the algorithm is still efficiently (with probability higher than 0.9) allocating the optimal arm in more than 90% of the times. Such a different behaviour of TS between H_0 and H_1 allows us to determine an optimal value of α that would achieve desirable guarantees in both scenarios. For example, with $\alpha = 250$, we achieve good regret performances, while also ensuring 80% confidence that a sufficiently balanced allocation in [0.4, 0.6] will be achieved under H_0.

5 Redesigning the *MTurk I* experiment

In this section, we discuss the potential of redesigning the motivating *MTurk I* experiment detailed in Sect. 1 (see Fig. 1) with an adaptive design guided by the proposed *Multinomial-TS* versions. We use a nonparametric simulation-based approach following a resampling strategy (*bootstrapping*; Efron and Tibshirani 1993) from the real data collected in the *MTurk I* experiment. More specifically, consider the set of $T = 110$ participants data split into two urns: Urn 1, with the $T_1 = 58$ data points associated with arm 1 and Urn 2, with the $T_2 = 52$ data points associated with arm 2. Anytime arm 1 is selected by *Multinomial-TS*, a data point (rating) from Urn 1 is sampled with replacement and vice versa. The resampling approach is replicated a number of 10, 000 independent times, each based on a horizon of $T = 110$ as in the original *MTurk I* experiment.

A comparison between the *actual* arm allocations observed in the *MTurk I* experiment and the *expected* allocation in a redesigned experiment with *Multinomial-TS* and *Multinomial-TS with augmented support* is presented in Fig. 9. We focus on the proportion of allocation of the optimal arm (arm 1: "Today is a new day to start fresh") over the horizon T. Compared to the balanced allocation of the original design, both *Multinomial-TS* variants show an increased allocation of arm 1, with 86% and 83% participants assigned to it at the end of the experiment. However, if we evaluate the uncertainty of the two *Multinomial-TS* variants (assessed in terms of the

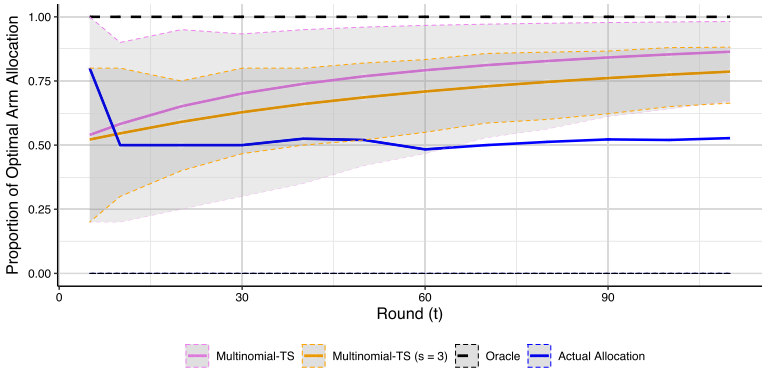

Fig. 9 Comparison between the actual allocation of optimal arm 1 (T_1/T) observed within the *MTurk I* experiment and the *expected* allocation, with their uncertainty ([0.1, 0.9] percentile confidence intervals; dotted lines), attainable with *Multinomial-TS* and *Multinomial-TS with augmented support* ($s = 3$). Results are obtained by averaging across 10^4 TS trajectories

[0.1, 0.9] percentile confidence intervals), we can notice a higher uncertainty in the standard version compared to the augmented-support one. This is particularly true at the beginning of the experiment (for $T < 60$), where the possibility of allocating the inferior arm 1 can be substantial. For example, at $T = 30$, in 10% of the cases, the optimal arm 1 is allocated only 30% and 45% of the times with *Multinomial-TS* and *Multinomial-TS with augmented support*, respectively. For $T > 60$, the benefits of the proposed strategies, particularly *Multinomial-TS with augmented support*, which shows reduced uncertainty, are relevant compared to the original study. Notice that, in light of the computational complexity assessments made in Appendix D, *Multinomial-TS with augmented support* comes at a negligible running time cost: for an horizon $T = 110$, we observe a median time (in milliseconds) of 9.91 ($s = 3$) vs 9.70 ($s = 0$) for a single run (see Table 8), translating into a median time (in minutes) of 1.65 ($s = 3$) vs 1.62 ($s = 0$) for 10, 000 runs.

For reproducibility of the results, the R codes and data are made available at the following repository: https://github.com/nina-DL/MultinomialTS.

6 Conclusion and future work

In this work, motivated by the *MTurk I* field experiment, we extended the applicability of TS to rating scale data, introducing *Multinomial-TS*. We demonstrated that, in scenarios with a unique optimal arm, it can outperform the widely used TS variants with a Normal or a Bernoulli model, which results in being highly sensitive to the dichotomisation threshold. In scenarios with identical arm means, *Multinomial-TS* can offer a more balanced solution in terms of arm allocation, but its performance depends on the shape of the underlying reward distribution, more specifically on its skewness. Motivated by this, we introduced an alternative variant on *Multinomial-TS*, called *Multinomial-TS with augmented support*,

which artificially injects skewness into the prior distribution, and thus the posterior, through a support extension. Additionally, the role of an increased informative content provided by the Dirichlet prior to calibrating the arms uncertainty is discussed. Both approaches demonstrated substantial benefits in an identical arm means scenario, highlighting a potential solution for the incomplete learning problem in this setting. The impact of the prior informative content is further evaluated to quantify the trade-offs under an optimal arm scenario, demonstrating satisfactory results. By illustrating the sensitivity of the algorithm to different prior choices, this work also fills an important knowledge gap, often neglected within the bandit literature.

Further work is required to understand how the proposed versions would behave in a contextual MAB setting, especially when different user characteristics are associated with different preferences. To this end, a future line of research may integrate tools from the literature on regression for categorical and ordinal data (Agresti 2019; Hedeker 2008; Tutz 2011) into MABs and adaptive experiments. Among the existing frameworks, it is worth mentioning the vast literature on mixture models that account for the uncertainty of the respondents. These include the CUB (combination of uniform and shifted binomial) class (see Piccolo and Simone 2019, and the related discussion and rejoinder, for a state-of-the-art survey), and other mixture model configurations that also factor in different response attitudes (see e.g., Colombi et al. 2019). Questions about whether the assumption of stationarity is plausible in these contexts and to what extent it influences the decision-making within the experiment represent an additional challenge. Along this line, a flexible approach for incorporating time dependencies is given by hidden Markov models, recently explored in longitudinal rating data with dynamic response styles (Colombi et al. 2023).

Require: Time horizon T, number of categories of the rating scale outcome J, number of study arms K, augmentation size s, augmented prior parameters $\boldsymbol{\alpha_k} = (\alpha_{1k}, \ldots, \alpha_{Jk}, \ldots, \alpha_{(J+s)k})$, for $k = 1, \ldots, K$.

for $t = 1, 2, \ldots, T$ **do**
  for $k = 1, \ldots, K$ **do**
    Sample $\tilde{\boldsymbol{p}}_k = (\tilde{p}_{1k}, \ldots, \tilde{p}_{Jk}, \ldots, \tilde{p}_{(J+s)k}) \sim \text{Dir}(\alpha_{1k}, \ldots, \alpha_{Jk}, \ldots, \alpha_{(J+s)k})$
    Compute $\tilde{\mu}_{tk} = \sum_{j=1}^{J+s} j\tilde{p}_{jk}$
  end for
  Select arm $\tilde{a}_t = \text{argmax}_{k=1,\ldots,K} \tilde{\mu}_{tk}$ and observe the reward y_t
  for $j = 1, \ldots, J, \ldots, J+s$ **do**
    Compute $c_{jk} = \mathbb{I}(y_t = j, \tilde{a}_t = k)$
  end for
  for $k = 1, \ldots, K$ **do**
    Update posteriors $(\alpha_{1k}, \ldots, \alpha_{Jk}, \ldots, \alpha_{(J+s)k}) \leftarrow (\alpha_{1k}, \ldots, \alpha_{Jk}, \ldots, \alpha_{(J+s)k}) + (c_{1k}, \ldots, c_{Jk}, \ldots, c_{(J+s)k})$
  end for
end for

Algorithm 2 Multinomial-TS with Augmented Support

Appendix A Pseudocode Algorithm 2

Appendix B Supportive Simulation Studies

This Appendix expands the scenarios evaluated in Sect. 3 to account for a wider number of possible settings that may occur in real practice. In particular, we relax the null scenarios of identical arms distribution (either symmetric, left- or right-skewed cases) to account for cases where arms share the same mean but may have a different shape and a different variability, thus a different discrete distribution over the support categories $j = 1, \dots, J$. This represents, in fact, the typical null hypothesis in many statistical and MAB problems (see e.g., Deliu et al. 2021). These cases are of the form: $H_0 : \mu_1 = \mu_2 = m$, for $m \in \{1, 2, \dots, J\}$ and $J = 7$, but $\mathbf{p}_1 \neq \mathbf{p}_2$. In all cases, different degrees of variability and skewness are considered. All different scenarios are illustrated in Figs. 10, 11, 12, 13 and 14 and the results, in terms of arm allocation, are reported in Tables 2, 3, 4, 5, 6 and 7. The values obtained are in line with the results depicted in Table 1. In particular, the role of standard deviation σ and sample skewness g_1 (Agresti 2019) is now more evident. For a given mean $\mu_1 = \mu_2 = m \in \{2, 3, 4, 5, 6\}$, when $g_1 = 0$, the higher the variability of one arm compared to the other, the lower its empirical probability of being assigned an extreme number of times (see, e.g., the last two rows of S1–S6 in Table 2). In terms of skewness, the higher the *left* skewness, the higher the chances of observing a large amount of extreme allocation (compare Table 2 to Tables 3 and 4). On the contrary, an increased *right* skewness (see Tables 5 and 6), translates into a reduced chance of extreme allocation, and therefore an increased balance in arm allocation. For example, considering the case $\sigma = 1$, we find that $\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$ goes from 22.9% (scenario S6, with $g_1 = 0$ and $H_0 : \mu_1 = \mu_2 = 4$; Table 2) to 29.5% (scenario R6, with $g_1 = 0.35$ and $H_0 : \mu_1 = \mu_2 = 3$; Table 5) and 36.1% (scenario RR4, with $g_1 = 1.15$ and $H_0 : \mu_1 = \mu_2 = 2$; Table 6). In cases with left skewness, this probability decreases to 15.2% (scenario L6, with $g_1 = -0.38$ and $H_0 : \mu_1 = \mu_2 = 5$; Table 3) and ultimately to 5.3% (scenario LL4, with $g_1 = -1.16$ and $H_0 : \mu_1 = \mu_2 = 6$; Table 4).

Finally, when analysing the extreme cases, that is, $H_0 : \mu_1 = \mu_2 = 1$ and $H_0 : \mu_1 = \mu_2 = 7$, which represent the most skewed scenarios in our experimental setup defined on $J = 7$ categories, we can notice an empirical convergence aligned with the previous results, also supporting the conjecture we made in Sect. 4.1 (Fig. 15).

Table 2 Empirical probabilities (%) of observing an arm allocation within a given range in scenarios S1-S8 ($H_0 : \mu_1 = \mu_2 = 4$). Values are averaged across 10^4 independent *Multinomial-TS* trials

	S1	S2	S3	S4	S5	S6	S7	S8
σ	2.8	2.5	2.0	1.7	1.4	1.0	1.5	2.0
g_1	0.0	0.0	0.0	0.0	0.0	0.0	0.37	− 0.24
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	0.8%	0.8%	1.0%	0.7%	1.1%	0.8%	0.9%	0.8%
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	3.4%	4.1%	4.9%	4.1%	6.4%	4.4%	4.8%	3.8%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	11.5%	11.6%	11.4%	11.5%	11.5%	11.5%	11.4%	11.8%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	22.4%	22.9%	22.9%	22.7%	22.5%	22.9%	23.0%	23.1%
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	10.1%	9.3%	7.7%	8.2%	5.1%	4.7%	7.0%	7.8%
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	3.9%	3.5%	2.9%	2.5%	1.3%	0.7%	1.9%	2.6%

Table 3 Empirical probabilities (in percentage) of observing an arm allocation within a given range in scenarios L1–L6 ($H_0 : \mu_1 = \mu_2 = 5$). Values are averaged across 10^4 independent *Multinomial-TS* trials

	L1	L2	L3	L4	L5	L6
σ	2.8	2.5	2.0	1.7	1.4	1.0
g_1	− 0.75	− 0.70	− 0.67	− 0.59	− 0.50	− 0.38
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	5.7%	8.2%	8.7%	8.5%	9.5%	9.8%
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	10.3%	14.9%	14.8%	14.3%	16.5%	17.0%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	8.4%	8.1%	7.6%	8.1%	7.8%	7.8%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	16.9%	16.1%	15.7%	16.1%	15.4%	15.2%
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	18.1%	14.4%	15.9%	16.3%	14.1%	14.4%
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	10.5%	8.7%	9.6%	9.2%	8.3%	7.7%

Table 4 Empirical probabilities (in percentage) of observing an arm allocation within a given range in scenarios LL1–LL6 (mean = 6). Values are averaged across 10^4 independent *Multinomial-TS* trials

	LL1	LL2	LL3	LL4	LL5	LL6
σ	2.2	1.8	1.3	1.0	0.9	0.8
g_1	− 1.87	− 1.78	− 1.52	− 1.16	− 0.99	− 0.81
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	27.2%	28.0%	33.4 %	31.5 %	32.5 %	33.8%
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	32.7%	32.9 %	39.4 %	37.2 %	38.5 %	40.0%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	4.3 %	4.1 %	3.1 %	2.7 %	2.6 %	2.4%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	8.5 %	8.1 %	6.5 %	5.3 %	5.7 %	5.1 %
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	28.0 %	29.8 %	29.1 %	32.6 %	31.6 %	31.0 %
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	23.3 %	25.2 %	25.3 %	28.7 %	27.9 %	27.6%

Table 5 Empirical probabilities (in percentage) of observing an arm allocation within a given range in scenarios R1–R6 ($H_0 : \mu_1 = \mu_2 = 3$). Values are averaged across 10^4 independent *Multinomial-TS* trials

	R1	R2	R3	R4	R5	R6
σ	2.8	2.5	2.0	1.7	1.4	1.0
g_1	0.75	0.73	0.66	0.60	0.47	0.35
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	0.2%	0.3%	0.2%	0.1%	0.2%	0.3%
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	1.7%	2.1 %	1.4%	1.1%	2.0%	3.1%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	14.0%	13.4%	13.5%	12.2%	14.7%	15.3%
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	26.9%	26.9%	26.7%	25.0%	29.0%	29.5%
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	6.2%	5.0%	5.1%	4.6%	2.2%	0.5%
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	1.4 %	1.0 %	0.8 %	0.5 %	0.2%	0.0%

Table 6 Empirical probabilities (in percentage) of observing an arm allocation within a given range in scenarios RR1–RR6 ($H_0 : \mu_1 = \mu_2 = 2$). Values are averaged across 10^4 independent *Multinomial-TS* trials

	RR1	RR2	RR3	RR4	RR5	RR6
σ	2.2	1.8	1.3	1.0	0.9	0.8
g_1	1.79	1.76	1.47	1.15	1.00	0.80
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	0.9 %	0.8 %	0.5 %	0.6 %	1.0 %	0.7 %
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	15.4 %	16.6 %	17.2 %	18.6 %	18.5 %	18.7 %
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	30.4 %	32.9 %	34.1%	36.1 %	35.2 %	36.3%
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	0.9 %	0.6 %	0.2 %	0.1 %	0.0 %	0.0 %
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

Table 7 Empirical probabilities (in percentage) of observing an arm allocation within a given range in scenarios LE ($H_0 : \mu_1 = \mu_2 = 1$) and RE ($H_0 : \mu_1 = \mu_2 = 7$). Values are averaged across 10^4 independent *Multinomial-TS* trials

	LE (%)	RE (%)
$\mathbb{P}\left(\frac{T_1}{T} < 0.05\right)$	49.1	0.0
$\mathbb{P}\left(\frac{T_1}{T} < 0.1\right)$	49.9	0.0
$\mathbb{P}\left(\frac{T_1}{T} \in [0.45, 0.55]\right)$	0.0	100
$\mathbb{P}\left(\frac{T_1}{T} \in [0.4, 0.6]\right)$	0.1	100
$\mathbb{P}\left(\frac{T_1}{T} > 0.9\right)$	49.3	0.0
$\mathbb{P}\left(\frac{T_1}{T} > 0.95\right)$	48.3	0.0

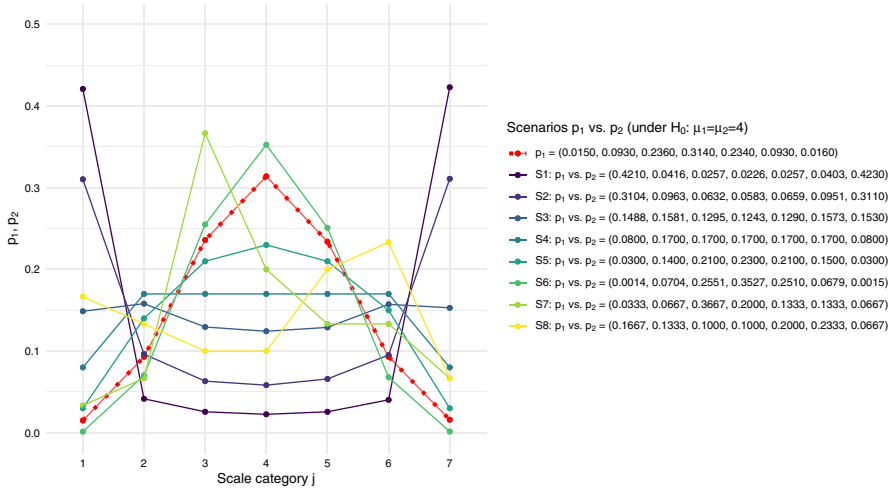

Fig. 10 Scenarios S1-S8 with $H_0: \mu_1 = \mu_2 = 4$ (equal arm means but different distribution)

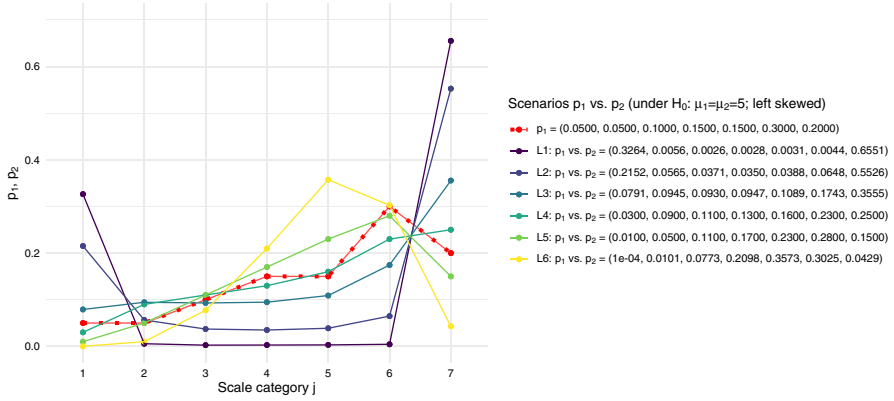

Fig. 11 Scenarios L1-L6 with $H_0: \mu_1 = \mu_2 = 5$ (equal arm means but different distribution)

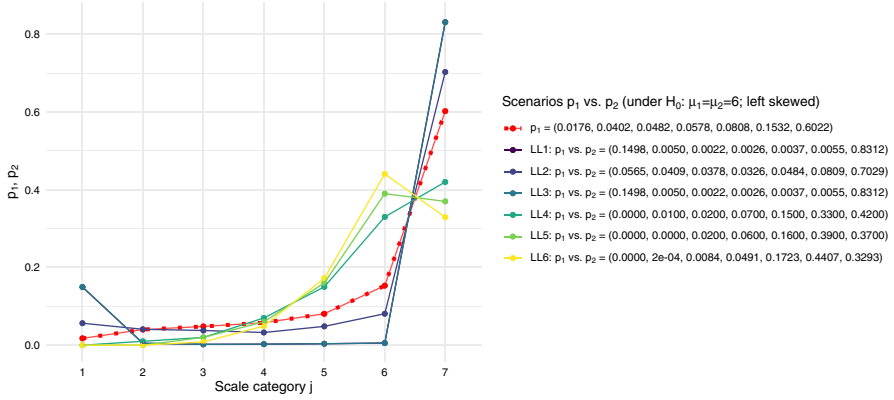

Fig. 12 Scenarios LL1-LL6 with $H_0: \mu_1 = \mu_2 = 6$ (equal arm means but different distribution)

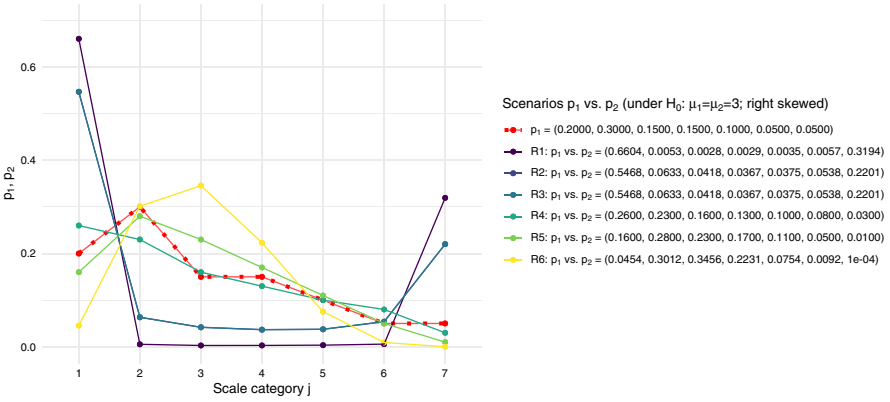

Fig. 13 Scenarios R1-R6 with $H_0 : \mu_1 = \mu_2 = 3$ (equal arm means but different distribution)

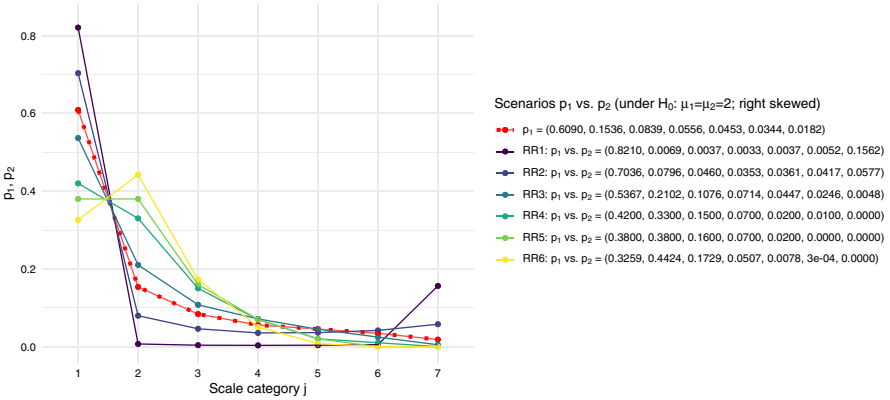

Fig. 14 Scenarios RR1-RR6 with $H_0 : \mu_1 = \mu_2 = 2$ (equal arm means but different distribution)

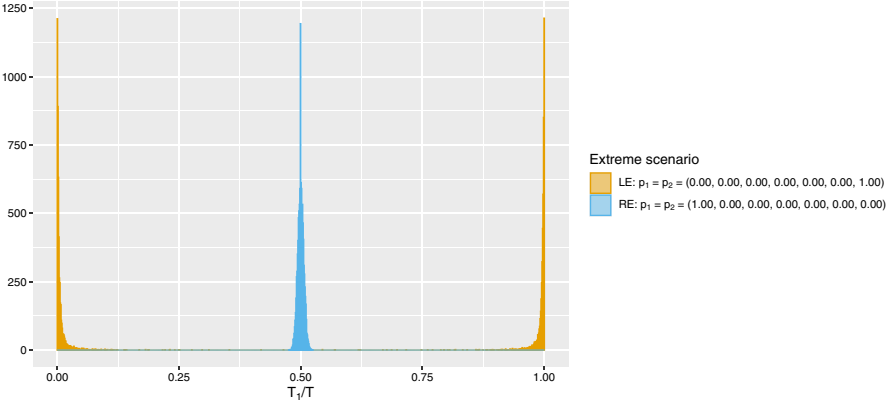

Fig. 15 Empirical allocation of arm 1 (T_1/T) under scenarios LE ($H_0 : \mu_1 = \mu_2 = 7$) and RE ($H_0 : \mu_1 = \mu_2 = 1$). Values are averaged across 10^4 independent *Multinomial-TS* trials

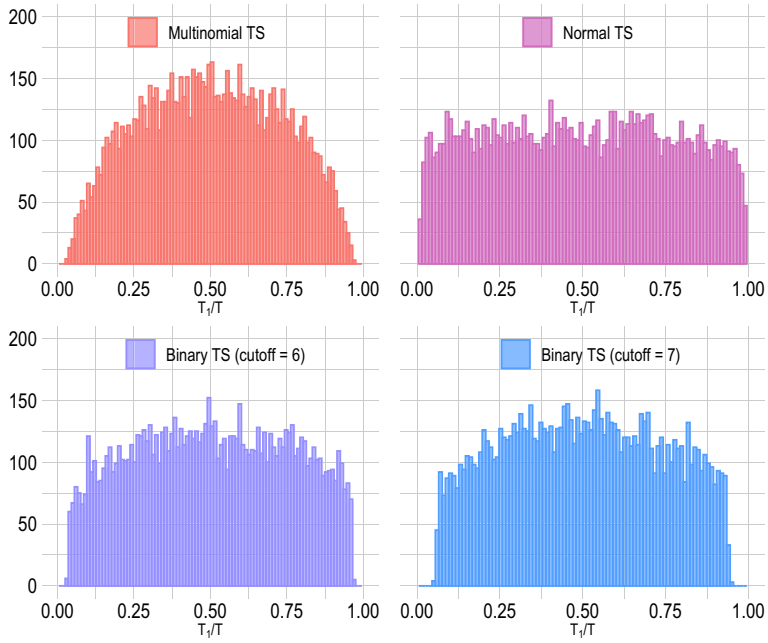

Fig. 16 Empirical allocation (T_1/T) under the identical arm case with a right skewed distribution (Scenario 3). Values are obtained by averaging across 10^4 independent TS trajectories

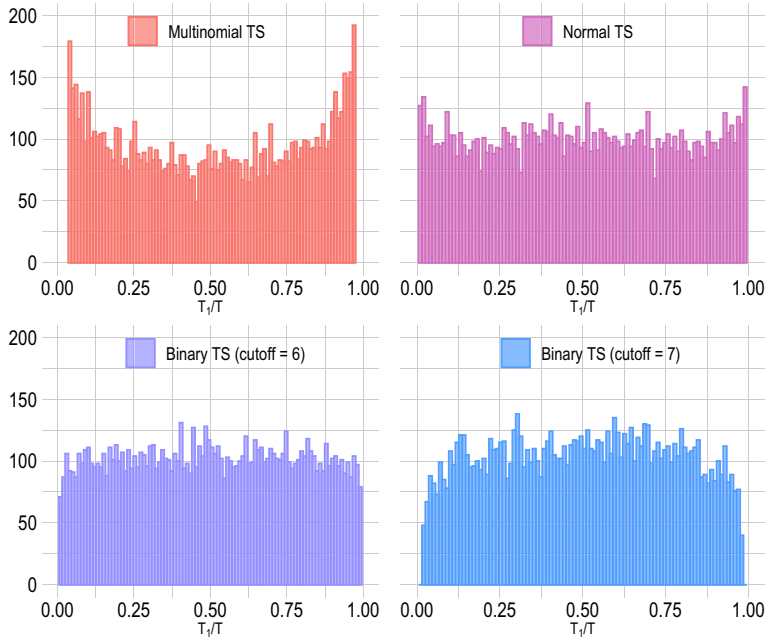

Fig. 17 Empirical allocation (T_1/T) under the identical arm case with a left skewed distribution (Scenario 4). Values are obtained by averaging across 10^4 independent TS trajectories

Table 8 (b) Summary of running times (based on 100 replicas) wrt the hyperparameter s and the horizon T. The setting is the same as in Fig. 19

Support augmentation	Min	First quartile	Mean	Median	Third quartile	Max
$T = 1000$						
s = 0	22.73	24.00	28.25	25.04	27.71	49.69
s = 1	23.30	24.53	29.03	25.14	27.65	71.59
s = 3	23.79	25.28	32.65	26.38	34.53	209.76
s = 5	23.93	25.42	30.43	26.79	29.79	76.59
s = 10	27.38	29.10	36.31	30.17	34.86	267.92
s = 20	30.09	32.08	41.36	33.78	42.93	340.70
s = 50	39.39	41.90	53.13	44.61	59.08	235.47
s = 100	54.15	59.16	72.10	69.85	80.41	208.43
$T = 110$						
s = 0	8.12	9.27	10.24	9.70	10.50	22.83
s = 3	8.73	9.41	10.98	9.91	10.92	24.44

Appendix C Supportive Plots

The following two plots support the statements in Sect. 3 comparing *Multinomial-TS* with *Binary-TS* and *Normal-TS* in Scenario 3 and Scenario 4 (skewed distributions) (see Figs. 16, 17).

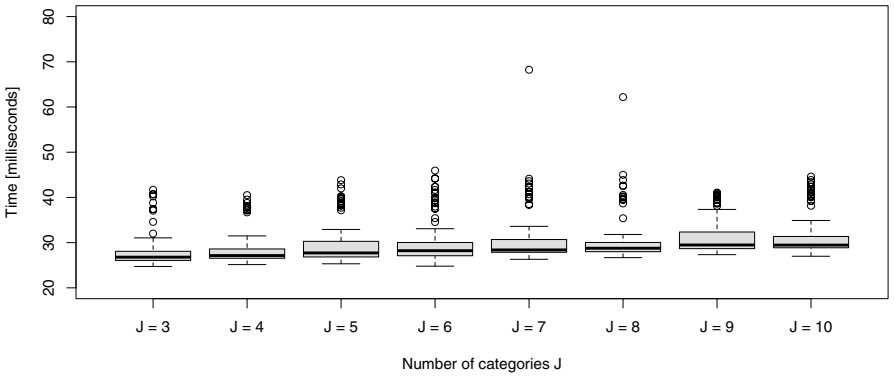

Fig. 18 (a) Boxplots of running times (based on 100 replicas) wrt the number of categories J. We focus on a two-armed setting under Scenario $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = (1, \underbrace{0, \dots, 0}_{J-1})$ and let J vary in the set $\{3, 4, \dots, 10\}$

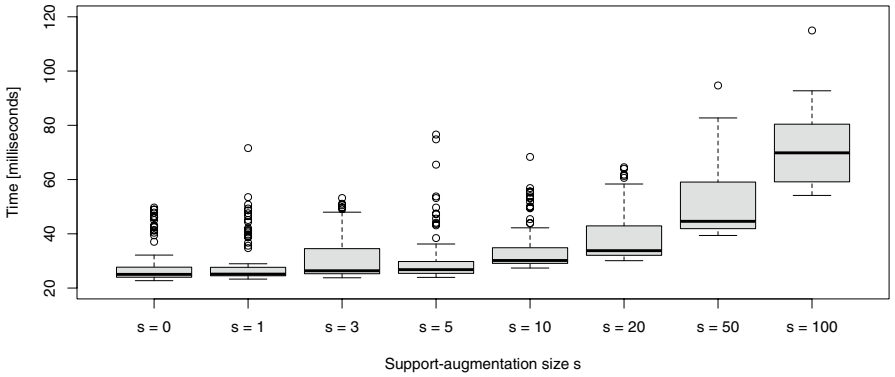

Fig. 19 (b) Boxplots of running times (based on 100 replicas) wrt the hyperparameter s. We focus on a two-armed setting with $J = 7$ categories under a Scenario $H_0 : \mathbf{p}_1 = \mathbf{p}_2 = (1, 0, 0, 0, 0, 0, 0, \underbrace{0, \dots, 0}_{s})$ and let s vary in the set $\{0, 1, 3, 5, 10, 20, 50, 100\}$

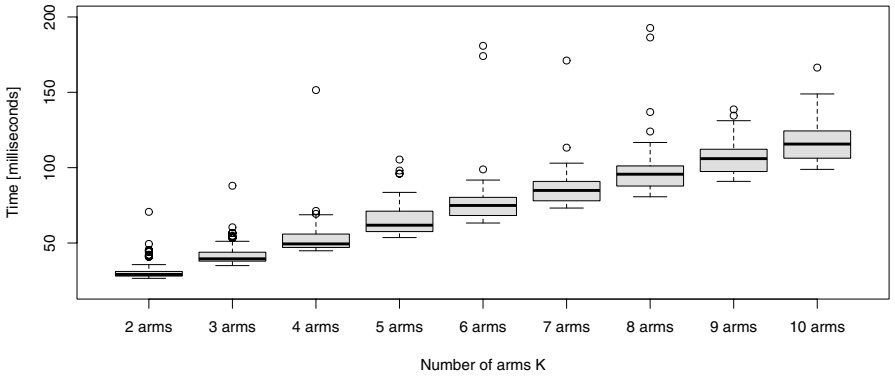

Fig. 20 (c) Boxplots of running times (based on 100 replicas) wrt the number of arms K. We focus on a setting with $J = 7$ categories under Scenario $H_0 : \mathbf{p}_k = (1, 0, 0, 0, 0, 0, 0)$, for $k = 1, 2, \dots, K$, and let K vary in the set $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Appendix D Computation Times

Here, we report the computation cost in terms of the running times (unit: milliseconds) of *Multinomial-TS*. Details vary according to:

(a) The number of categories J – see Fig. 18;
(b) The support augmentation hyperparameter s – see Fig. 19 and Table 8;
(c) The number of arms K – see Fig. 20.

The results in (a) and (c) refer to *Multinomial-TS* and times are reported for a single run over an horizon $T = 1000$. The results in (b) relate to *Multinomial-TS*

with augmented support and times are computed for both $T = 1000$ as above and $T = 110$ (reflecting the *MTurk I* example in Sect. 1).

All analyses were performed on a Darwin (macOS) Kernel Version 22.5.0; root:xnu-8796.121.3-7/RELEASE_X86_64. We used the `microbenchmark()` package in `R` version 4.2.0 (2022-04-22).

Acknowledgements The author sincerely thanks Brunero Liseo for the constructive feedback on the manuscript, especially on the considerations on the Dirichlet prior parameters. The author also thanks Joseph Jay Williams and the IAI Lab for the motivating example and the *MTurk I* deployments data. Financial support by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) is acknowledged.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Declarations

Conflict of interest The author has no conflicts of interest to declare. I certify that the submission is original work and is not under review at any other publication.

References

Agrawal S, Goyal N (2017) Near-optimal regret bounds for Thompson sampling. J ACM (JACM) 64(5):30:1-30:24. https://doi.org/10.1145/3088510

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2022) The MNL-bandit problem. In: Chen X, Jasin S, Shi C (eds) The elements of joint learning and optimization in operations management. Springer Series in Supply Chain Management. Springer, Cham, pp 211–240. https://doi.org/10.1007/978-3-031-01926-5_9

Agresti A (2019) An introduction to categorical data analysis, 3rd edn. Wiley series in probability and statistics. John Wiley & Sons, Hoboken

Akobeng AK (2005) Understanding randomised controlled trials. Arch Dis Child 90(8):840–844. https://doi.org/10.1136/adc.2004.058222

Altman DG, Royston P (2006) The cost of dichotomising continuous variables. BMJ 332(7549):1080.1. https://doi.org/10.1136/bmj.332.7549.1080

Amatriain X, Basilico J (2015) Recommender systems in industry: a netflix case study. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer, Boston, pp 385–419. https://doi.org/10.1007/978-1-4899-7637-6_11

Antos A, Grover V, Szepesvári C (2008) Active learning in multi-armed bandits. In: Freund Y, Györfi L, Turán G, Zeugmann T (eds) Algorithmic learning theory. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp 287–302. https://doi.org/10.1007/978-3-540-87987-9_25

Berry DA, Chen RW, Zame A, Heath DC, Shepp LA (1997) Bandit problems with infinitely many arms. Ann Stat 25(5):2103–2116. https://doi.org/10.1214/aos/1069362389

Besbes O, Gur Y, Zeevi A (2014) Stochastic multi-armed-bandit problem with non-stationary rewards. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (eds) Advances in neural information processing systems, vol 27. Curran Associates Inc., Red Hook

Bothwell LE, Avorn J, Khan NF, Kesselheim AS (2018) Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. BMJ Open 8(2):e018320. https://doi.org/10.1136/bmjopen-2017-018320

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ (eds) Advances in neural information processing systems, vol 24. Curran Associates Inc., Red Hook

Charpentier A, Élie R, Remlinger C (2023) Reinforcement learning in economics and finance. Comput Econ 62(1):425–462. https://doi.org/10.1007/s10614-021-10119-4

Colombi R, Giordano S, Gottard A, Iannario M (2019) Hierarchical marginal models with latent uncertainty. Scand J Stat 46(2):595–620

Colombi R, Giordano S, Kateri M (2023) Hidden markov models for longitudinal rating data with dynamic response styles. Stat Methods Appl, 1–36

Deliu N (2022) Multinomial Thompson Sampling for adaptive experiments with rating scales. In: Book of short papers SIS 2022, pp 1065–1070. Pearson, London

Deliu N, Williams JJ, Villar SS (2021) Efficient inference without trading-off regret in bandits: an allocation probability test for Thompson Sampling. arXiv:2111.00137

Deliu N, Williams JJ, Chakraborty B (2023) Reinforcement learning in modern biostatistics: constructing optimal adaptive interventions. arXiv:2203.02605

Deshpande Y, Mackey L, Syrgkanis V, Taddy M (2018) Accurate inference for adaptive linear models. In: Proceedings of the 35th international conference on machine learning, pp 1194–1203. PMLR. https://proceedings.mlr.press/v80/deshpande18a.html

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Monogr Stat Appl Probab 57:158

Figueroa CA, Aguilera A, Chakraborty B, Modiri A, Aggarwal J, Deliu N, Sarkar U, Jay Williams J, Lyles CR (2021) Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. J Am Med Inf Assoc JAMIA 28(6):1225–1234. https://doi.org/10.1093/jamia/ocab001

Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal J, Jay Williams J, Lyles C, Aguilera A (2022) Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. Ann Behav Med 56(2):212–218. https://doi.org/10.1093/abm/kaab028

Gandapur Y, Kianoush S, Kelli HM, Misra S, Urrea B, Blaha MJ, Graham G, Marvel FA, Martin SS (2016) The role of mHealth for improving medication adherence in patients with cardiovascular disease: a systematic review. Eur Heart J Qual Care Clin Outcomes 2(4):237–244. https://doi.org/10.1093/ehjqcco/qcw018

Hadad V, Hirshberg DA, Zhan R, Wager S, Athey S (2021) Confidence intervals for policy evaluation in adaptive experiments. Proc Natl Acad Sci 118(15):e2014602118. https://doi.org/10.1073/pnas.2014602118

Hedeker D (2008) Multilevel models for ordinal and nominal variables. In: Leeuw JD, Meijer E (eds) Handbook of multilevel analysis. Springer, New York, pp 237–274. https://doi.org/10.1007/978-0-387-73186-5_6

Kalvit A, Zeevi A (2021) A closer look at the worst-case behavior of multi-armed bandit algorithms. Adv Neural Inf Process Syst 34:8807–8819

Kasy M, Sautmann A (2021) Adaptive treatment assignment in experiments for policy choice. Econometrica 89(1):113–132. https://doi.org/10.3982/ECTA17527

Keskin NB, Zeevi A (2018) On incomplete learning and certainty-equivalence control. Oper Res 66(4):1136–1167. https://doi.org/10.1287/opre.2017.1713

Kim K, Bretz F, Cheung YKK, Hampson LV (2021) Handbook of statistical methods for randomized controlled trials, 1st edn. CRC Press, Boca Raton

Kotz S, Balakrishnan N, Johnson Nl (2000) Continuous multivariate distributions, volume 1: models and applications. Wiley Series in Probability and Statistics, 1st edn. Wiley. https://doi.org/10.1002/0471722065

Lai Y, Robbins H (1985) Asymptotically efficient adaptive allocation rules. Adv Appl Math 6(1):4–22. https://doi.org/10.1016/0196-8858(85)90002-8

Lattimore T, Szepesvári C (2020) Bandit algorithms. Cambridge University Press, Cambridge

Li T, Nogas J, Song H, Kumar H, Durand A, Rafferty A, Deliu N, Villar SS, Williams JJ (2022) Algorithms for adaptive experiments that trade-off statistical analysis with reward: combining uniform random assignment and reward maximization. arXiv:2112.08507

Liu C-Y, Li L (2016) On the prior sensitivity of Thompson sampling. In: Ortner R, Simon HU, Zilles S (eds) Algorithmic learning theory. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 321–336. https://doi.org/10.1007/978-3-319-46379-7_22

Liu X, Deliu N, Chakraborty B (2023) Microrandomized trials: developing just-in-time adaptive interventions for better public health. Am J Public Health 113(1):60–69. https://doi.org/10.2105/AJPH.2022.307150

Mason W, Suri S (2012) Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 44(1):1–23. https://doi.org/10.3758/s13428-011-0124-6

Min S, Maglaras C, Moallemi CC (2019) Thompson sampling with information relaxation penalties. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates Inc., Red Hook

Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 16(1):29. https://doi.org/10.1186/s12916-018-1017-7

Parapar J, Radlinski F (2021) Diverse user preference elicitation with multi-armed bandits. In: Proceedings of the 14th ACM international conference on web search and data mining, WSDM '21, Association for Computing Machinery, New York, pp 130–138. https://doi.org/10.1145/3437963.3441786

Piccolo D, Simone R (2019) The class of cub models: statistical foundations, inferential issues and empirical evidence. Stat Methods Appl 28(3):389–435. https://doi.org/10.1007/s10260-019-00461-1

Riou C, Honda J (2020) Bandit algorithms based on Thompson sampling for bounded reward distributions. In: Proceedings of the 31st international conference on algorithmic learning theory, pp 777–826. PMLR. https://proceedings.mlr.press/v117/riou20a.html

Robertson DS, Lee KM, López-Kolkovska BC, Villar SS (2023) Response-adaptive randomization in clinical trials: from myths to practical considerations. Stat Sci 38(2):185–208. https://doi.org/10.1214/22-STS865

Rosenberger WF, Uschner D, Wang Y (2019) Randomization: the forgotten component of the randomized clinical trial. Stat Med 38(1):1–12. https://doi.org/10.1002/sim.7901

Russo D (2016) Simple Bayesian algorithms for best arm identification. In: Conference on learning theory, pp 1417–1418. PMLR. https://proceedings.mlr.press/v49/russo16.html

Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z (2018) A tutorial on Thompson sampling. Found Trends Mach Learn 11(1):1–96

Shin J, Ramdas A, Rinaldo A (2019) Are sample means in multi-armed bandits positively or negatively biased? In: Proceedings of the 33rd international conference on neural information processing systems, No. 638, Curran Associates Inc., Red Hook, pp 7102–7111

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25(3–4):285–294. https://doi.org/10.1093/biomet/25.3-4.285

Tutz G (2011) Regression for categorical data, 1st edn. Cambridge University Press, Cambridge

Villar SS, Bowden J, Wason J (2015) Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Stat Sci 30(2):199–215. https://doi.org/10.1214/14-STS504

Williams JJ, Rafferty AN, Tingley D, Ang A, Lasecki WS, Kim J (2018) Enhancing online problems through instructor-centered tools for randomized experiments. In: Proceedings of the 2018 CHI conference on human factors in computing systems, CHI '18, Association for Computing Machinery, New York, pp 1–12. https://doi.org/10.1145/3173574.3173781

Williams JJ, Nogas J, Deliu N, Shaikh H, Villar SS, Durand A, Rafferty A (2021) Challenges in statistical analysis of data collected by a bandit algorithm: an empirical exploration in applications to adaptively randomized experiments. arXiv:2103.12198

Williamson SF, Villar SS (2020) A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. Biometrics 76(1):197–209. https://doi.org/10.1111/biom.13119

Zhang Y, Basu S, Shakkottai S, Heath RW (2021) MmWave codebook selection in rapidly-varying channels via multinomial Thompson sampling. In: Proceedings of the twenty-second international symposium on theory, algorithmic foundations, and protocol design for mobile networks and mobile computing, MobiHoc '21, Association for Computing Machinery, New York, pp 151–160. https://doi.org/10.1145/3466772.3467044