


CENTRAL LIMIT THEOREM IN COMPLETE FEEDBACK GAMES

ANDREA OTTOLINI ^{*, **} AND
RAGHAVENDRA TRIPATHI,^{*, ***} *University of Washington*

Abstract

Consider a well-shuffled deck of cards of n different types where each type occurs m times. In a complete feedback game, a player is asked to guess the top card from the deck. After each guess, the top card is revealed to the player and is removed from the deck. The total number of correct guesses in a complete feedback game has attracted significant interest in the past few decades. Under different regimes of m, n , the expected number of correct guesses, under the greedy (optimal) strategy, has been obtained by various authors, while there are not many results available about the fluctuations. In this paper we establish a central limit theorem with Berry–Esseen bounds when m is fixed and n is large. Our results extend to the case of decks where different types may have different multiplicity, under suitable assumptions.

Keywords: Central limit theorem; complete feedback games; card guessing

2020 Mathematics Subject Classification: Primary 60F05

Secondary 60G70; 91A60

1. Introduction

1.1. Lady tasting tea, revisited

Muriel Bristol, a biologist at Rothamsted Research at the dawn of the twentieth century, once claimed that she could taste whether a cup of tea was prepared by pouring the milk first. Ronald Fisher, in an attempt to disprove her claim, arranged the following simple experiment: Bristol was presented with eight cups of tea, half of which were prepared by pouring the milk first, and she was asked to taste them one by one and identify which were which. This episode became widely known as the ‘lady tasting tea’ experiment, the very first example appearing in Fisher’s seminal book [9] on the design of statistical experiments. Its analysis of the experiment is now a cornerstone of scientific thinking, being also the first appearance of the expression ‘null hypothesis’ in Fisher’s work. In this case, it assumes that Bristol’s guesses are random, so that the distribution of the score follows a hypergeometric distribution. On average, one expects four correct guesses, the underlying probability distribution being well understood.

In the original experiment, Bristol did not receive any kind of feedback during the experiment: What if she had been told the correct answer after each attempt? Clearly, since she knows the exact number of cups of each type, she can always guess the one that appeared the

Received 2 April 2023; accepted 9 August 2023.

* Postal address: Department of Mathematics, University of Washington, Seattle, WA 98195, USA.

** Email address: ottolini@uw.edu

*** Email address: raghavt@uw.edu

© The Author(s), 2023. Published by Cambridge University Press on behalf of Applied Probability Trust.

smallest number of times so far, and therefore increase her likelihood of a correct guess at each step. This does not require any special ability on her part, other than a clever exploitation of the information she is provided with. There has been a substantial flurry of interest in variations of this kind, owing to the connection with randomized clinical trials [1] and the testing of claims of extra-sensory perception [3], which we will review later. While most of the focus has been on the asymptotic expected score for large experiments, it is clear that a rigorous analysis of the experiments requires the understanding of the fluctuations of the score. This is the focus of our paper.

1.2. Model and main result

Consider a well-shuffled deck of cards consisting of n different types of cards where each card appears m times. Thus, there are mn cards in total. Consider the following *complete feedback game*: a player is asked to guess the type of card appearing on top of the deck. After each guess the top card is revealed to the player. The game continues until the deck is exhausted. Let $S_{m,n}$ denote the total number of correct guesses (also referred to as the score) at the end of the game. Obviously the score depends on the strategy. For instance, if the player keeps guessing, say, card of type 1, then $S_{m,n} = m$. Diaconis and Graham [4] have shown that the greedy algorithm maximizes the expected number of correct guesses, that is to say, a player should guess a card that has the maximum multiplicity in the remaining deck. We will refer to the greedy algorithm as the ‘optimal strategy’ throughout this paper, and we will tacitly assume that the player is performing this strategy. Our main result is a central limit theorem (CLT) for the optimal score $S_{m,n}$ with a Berry–Esseen bound that can be stated as follows.

Theorem 1.1. *Consider a deck of cards with n distinct types of cards where each card appears with a fixed multiplicity m . Let S_n be the total number of correct guesses under the greedy/optimal strategy. Then we have the following.*

- The mean $\mu_n := \mathbb{E}[S_n]$ and the variance $\sigma_n^2 := \text{Var}[S_n]$ satisfy

$$\mu_n \sim \sigma_n^2 \sim \left(1 + \frac{1}{2} + \dots + \frac{1}{m}\right) \ln n$$

as $n \rightarrow +\infty$.

- There exists a constant $C = C(m)$ depending only on m such that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_n - \mu_n}{\sigma_n} \leq x\right) - \Phi(x) \right| \leq C \frac{\ln \ln n}{\sqrt{\ln n}},$$

where Φ is the cumulative distribution function of a standard normal random variable.

More generally, we prove a CLT result analogous to Theorem 1.1 for a deck of cards where the cards of each type occur with (possibly) different multiplicity, that is, for a deck with n different types of cards where the cards of type i for $i \in [n]$ appear m_i times. We always assume that, for all n , the deck is shuffled so that all arrangements are equally likely.

Notation. To state our theorem clearly, we need to fix some notation.

- (1) For each $n \in \mathbb{N}$, a vector $\mathbf{m} := \mathbf{m}^n = (m_1, \dots, m_n)$ denotes a deck of cards with n different types of cards and where a card of type $i \in [n]$ appears m_i times in the deck.
- (2) Let $|\mathbf{m}| = |\mathbf{m}^n|$ denote the total number of cards in the deck, that is, $|\mathbf{m}| = \sum_{i=1}^n m_i$.

- (3) Let \mathbf{m}_{\max}^n denote the highest multiplicity of a card in the deck, that is, $\mathbf{m}_{\max}^n = \max_{i \in [n]} m_i$.
- (4) Let ϵ_n denote the fraction of type(s) i such that the cards of type i occur with highest multiplicity \mathbf{m}_{\max}^n .
- (5) Let $S_{\mathbf{m}^n}$ be the total number of correct guesses (also referred to as the score) at the end of the game under the optimal strategy.

Theorem 1.2. *Let \mathbf{m}^n be a sequence of decks, indexed by n , with n distinct types of cards. Suppose that $\mathbf{m}_{\max}^n \leq m$ for some m that is independent of n , and that $\epsilon_n \geq \epsilon$ for some positive ϵ independent of n . Let $S_{\mathbf{m}^n}$ be the total number of correct guesses under the greedy/optimal strategy. Then we have the following.*

- The mean $\mu_n := \mathbb{E}[S_{\mathbf{m}^n}]$ and the variance $\sigma_n^2 := \text{Var}[S_{\mathbf{m}^n}]$ satisfy

$$\mu_n \sim \sigma_n^2 \sim \left(1 + \frac{1}{2} + \dots + \frac{1}{\mathbf{m}_{\max}^n} \right) \ln n$$

as $n \rightarrow +\infty$.

- There exists a constant $C = C(\epsilon, m)$ such that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left(\frac{S_{\mathbf{m}^n} - \mu_n}{\sigma_n} \leq x \right) - \Phi(x) \right| \leq C \frac{\ln \ln n}{\sqrt{\ln n}},$$

where Φ is the cumulative distribution function of a standard normal random variable.

Remark 1.1. It is clear that all the assumptions in the Theorem 1.1 are satisfied if $m_i = m$ for all $i \in [n]$. In particular, Theorem 1.1 follows from Theorem 1.2. The result about the mean was already shown in [4] and [10].

1.3. Related literature

The complete feedback game was originally motivated by clinical trials. For an in-depth discussion of the problem, a good reference is [8], though the first appearance is in a work by Blackwell and Hodges [1]. They considered the case where two types of medical treatments have to be assigned to a fixed number of people, say $2m$, who arrive one by one at the clinic. They were interested in the case where both treatments are provided in the same quantity and in a random order. However, they assume that the hospital may decide, at their discretion, whether to rule out some of the subjects because of their medical conditions. Since they have information on the treatments provided up to that point, they may decide to bias the result of the experiment towards a specific treatment. This can be done by ruling out a particularly sick subject if they know that it is more likely that their favorable treatment has to appear next.

In our language, this is precisely the complete feedback case with $n = 2$, with $\mathbf{m} = (m, m)$. Blackwell and Hodges [1] gave an asymptotic formula for the optimal expected score (in their language, the selection bias), which was then extended by Diaconis and Graham [4] to the generic case $\mathbf{m} = (m_1, \dots, m_n)$ with n fixed. As for the fluctuations, the latter reference shows that, for $n = 2$ and the $\mathbf{m} = (m, m)$ with m large, the limiting optimal score satisfies a central limit theorem. On the other hand, in the unbalanced case where $\mathbf{m} = (m_1, m_2)$, where m_1, m_2 grow with $m_1/m_2 \rightarrow p \neq 1/2$, they show that the fluctuations of the optimal score are *not Gaussian*. Related results in the case $m = 2$ also appeared in [11].

Another occurrence of the complete feedback game is related to the rigorous analysis for extra-sensory perception claims. In fact, one of the most celebrated experiments in this direction corresponds precisely to the complete feedback game with a deck of twenty-five cards (Zener cards), with five symbols each appearing five times. For a historical account, the interested reader is referred to [3]. Motivated by this, Diaconis and Graham [4] suggest studying the asymptotic optimal expected score for the complete feedback game with decks $\mathbf{m} = (m_1, \dots, m_n)$ where n grows. In the case $m_i \equiv 1$, the analysis becomes much simpler since the sequence of guesses becomes independent. In particular, it is easy to deduce that one obtains about $\ln n$ correct guesses in expectation, with a variance of the same order and normal fluctuations.

The case where some of the m_i are greater than one is more subtle, since the chance of a correct guess will depend on the history of the draws up to that moment. Diaconis *et al.* [5] analyzed the case where $m_i \equiv m$ is fixed and n grows to infinity, showing that asymptotically the expected optimal score is $(1 + \dots + 1/m) \ln n$. The result was substantially refined by He and Ottolini in [10], where the expected score is determined for decks $\mathbf{m} = (m_1, \dots, m_n)$ under the same assumptions as in Theorem 1.2. Moreover, their asymptotic result matches the optimal expected score up to an explicit error that goes to zero. Their main tool is the analysis of a certain variation of the birthday problem via Stein's methods, which will be our main tool here as well. In the case $m_i \equiv m$ where both m and n are growing, the asymptotic for the expected optimal score was obtained by Ottolini and Steinerberger [15], covering a variety of regimes that include the case $n = m$ (Zener's original setting).

Variations of the game also include different types of feedback, the most relevant case being that of yes/no feedback (i.e. the card is shown only when a guess is correct). This becomes much harder to analyze even for balanced decks $\mathbf{m} = (m_1, \dots, m_n)$ with $m_i \equiv m$. It is known [4] that the optimal strategy is *not the greedy one* as soon as $m > 1$ and $n > 2$. Some limiting results were recently obtained in [5] and [13], for instance that the expected optimal score for $n \gg m \gg 1$ is of the form $m + \Theta(\sqrt{m})$ uniformly in n . Results on the fluctuations are currently unknown – except in the case $m = 1$ – where the limiting distribution has a non-normal behavior as shown in [4]. Since the optimal strategy is rather hard to implement, a fact ultimately due to its connection with permanents [2, 6], there has also been some interest in near-optimal strategies that are easier to implement [7].

Finally, we mention some other variations of the game in a similar flavor. The problem of minimizing the expected number of correct guesses was also addressed in [4], [5], and [10], for both complete feedback and yes/no feedback. Another natural set of questions comes from considering decks that have not been properly shuffled, such as the case of a deck which has been riffle shuffled [12].

2. Proofs

In the remainder of the paper we will sometimes drop the dependence on the deck \mathbf{m} and on n . The implicit constants in the notation O , Ω , Θ , \lesssim will depend on m and ϵ only, unless we specify otherwise. We will often identify \mathbf{m}_{\max}^n and ϵ_n with their upper/lower bound m or ϵ , unless there is ambiguity.

2.1. Main idea

It will be convenient to define the following random variables and set up some notation.

- $T_j = \max\{t \in \{0, 1, \dots, |\mathbf{m}|\} : \text{no card among the last } t \text{ appears more than } j \text{ times}\}$.
Here $0 \leq j \leq m$, with the convention $T_0 = 0$ and $T_m = |\mathbf{m}|$.

- $W_{j,t} = \sum_{\mathbf{t} \leq t} Y_{\mathbf{t}}$. Here $\mathbf{t} = (\mathbf{t}_1, \dots, \mathbf{t}_{j+1})$ is a $(j + 1)$ -tuple of strictly increasing positive integers, and the notation $\mathbf{t} \leq t$ means $\mathbf{t}_{j+1} \leq t$ (implying that $t_\ell \leq t$ for each $\ell = 1, 2, \dots, j + 1$). The binary random variable $Y_{\mathbf{t}}$ is one if and only if the cards at positions \mathbf{t} are equal (here, positions are considered from the *bottom* of the deck). Notice that $T_j > t$ if and only if $W_{j,t} = 0$. Here, $1 \leq t \leq |\mathbf{m}|$ and $0 \leq j \leq m$.
- $\tilde{W}_j = W_{j-1, T_j}$ denotes the number of cards that appear j times before some card appears $j + 1$ times. Again, this is done from the *bottom* of the deck. Here, $1 \leq j \leq m$.

Example 2.1. Consider a deck $\mathbf{m} = (3, 3, 2)$, and assume that the sequence of cards extracted, listed from the last to the first, is

$$(1, 2, 2, 1, 3, 1, 2, 3).$$

In this case, $T_0 = 0, T_1 = 2, T_2 = 5, T_3 = 8$. Correspondingly, we have $\tilde{W}_1 = W_{0,2} = 2, \tilde{W}_2 = W_{1,5} = 2, \tilde{W}_3 = W_{2,8} = 2$. For instance, $\tilde{W}_2 = 2$ reflects the fact that there are two pairs of identical cards among the last five and no triple of identical cards (those labeled one and two), while among the last six cards there is a triple of identical cards (those labeled one).

Remark 2.1. Notice that \tilde{W}_m is equal to the number of types that appear with multiplicity m , which is at least ϵn under the assumption of Theorem 1.2.

Remark 2.2. Since $W_{j,t}$ is a sum of indicators, $W_{j,t} \leq W_{j,|\mathbf{m}|} \leq \binom{nm}{j}$ is at most polynomial in n under the assumption of Theorem 1.2.

The random variables $W_{j,t}$ are important tools for understanding the asymptotic behavior of the total score. In fact, He and Ottolini’s main ingredient in [10] is an asymptotic result for the $W_{t,j}$, which behave like Poisson random variables with suitable parameters. This should come as no surprise, since the $Y_{\mathbf{t}}$ are indicators of rare events, most of which are weakly dependent. They obtained the following result.

Theorem 2.1. (Theorem 1.8 of [10].) *Let $1 \leq j < m$. Then there exists*

$$\lambda = \Theta\left(\frac{t^{j+1}}{n^j}\right)$$

such that

$$d_{\text{tv}}(W_{j,t}, \text{Poi}(\lambda)) \lesssim \frac{t}{n} \lesssim \left(\frac{\lambda}{n}\right)^{1/(j+1)}.$$

Here, $\text{Poi}(\lambda)$ represents a Poisson random variable with mean λ , and d_{tv} represents the total variation distance between probability measures (where we identify a random variable with its law). Moreover, the implicit constants in the error term and the definition of λ can be chosen to depend only on j and ϵ , the fraction of types that appear with multiplicity m .

Remark 2.3. In the case $m_i \equiv m$, we have

$$\lambda = \frac{t^{j+1}}{n^j} \frac{\binom{m}{j+1}}{m^{j+1}}.$$

Notice that for fixed j , the second term converges to $1/(j + 1)!$ as $m \rightarrow +\infty$.

Remark 2.4. Theorem 2.1 can be thought of as a variant of the classical birthday problem. In fact, for $j = 1$ and $m_j \equiv \mathbf{m}_{\max}^n \rightarrow +\infty$, we obtain the best possible approximation for the classical birthday problem.

In this section we state and prove some lemmas that will be useful in the proof of Theorem 1.2. We begin with a discussion of the idea of the proof. The total score $S_{\mathbf{m}}$ can be written as the sum of $|\mathbf{m}|$ (the total size of the deck) binary random variables, namely, the indicators that at a given time the player obtains a correct guess. If these random variables were independent, the CLT for the total score would follow at once.

However, even in the case $m = 2, n = 2$, it is clear that if the second guess is correct, then the remaining two cards are distinct (hence the third guess will be correct with probability $1/2$), while if the second guess is wrong, then the remaining two cards are the same. Hence the third guess will be correct with probability 1. Intuitively, the dependence becomes weak for large n , and it should be related to the concentration properties of the random variables introduced above. Indeed, the strategy will change depending on how many cards appear with a given multiplicity at a given time. The first crucial step towards the proof of the CLT will be the observation that, conditioned on \tilde{W}_j , the score can be written as a sum of independent random variables (see Lemma 2.1). This allows us to prove a CLT for the conditional score with a uniform Berry–Essen bound (Lemma 2.2).

The final issue is to understand the behavior of the \tilde{W}_j and show that they enjoy suitable concentration. The main difficulty is that Theorem 2.1 requires a *fixed* time, rather than the random time T_j appearing in the definition of the \tilde{W}_j . Moreover, the random variables T_j and $W_{j-1,t}$ are dependent, meaning that we cannot expect a straightforward limiting result expressed in terms of compound Poisson random variables. It is worth noticing that while in [10] there is a multivariate version of Theorem 2.1, it does not suffice for our purposes. To circumvent the problem, we will use a suitable monotonicity and concentration argument that allows us to prove the main Lemma 2.3.

2.2. A CLT for the conditional score

We start by showing a useful representation of the optimal score.

Lemma 2.1. *For any deck \mathbf{m} , the optimal score $S_{\mathbf{m}}$ can be written as*

$$S_{\mathbf{m}} = \sum_{j=1}^m \sum_{s=1}^{\tilde{W}_j} X_{j,s},$$

where the $X_{j,s}$ are conditionally independent – given the \tilde{W}_j – Bernoulli random variables with $\mathbb{P}(X_{j,s} = 1) = 1/s$.

Proof. Let $\tau_{j,s}$ be the time at which the maximum multiplicity of a symbol left in the deck is equal to j ; there are exactly s symbols with this multiplicity, and the symbol of the card on top of the deck is precisely one of those s symbols. Notice that a correct guess can only occur at the times $\tau_{j,s}$. If $X_{j,s}$ denotes the indicator that the guess at time $\tau_{j,s}$ is correct, then

$$\mathbb{P}(X_{j,s} = 1) = \frac{1}{s}.$$

To see this, notice that at time $\tau_{j,s}$, the symbol appearing on the card is one of the s symbols that appear with the maximum multiplicity, and each of them appears with the same likelihood

since the deck is uniformly shuffled. It is worth remarking that while the optimal strategy is not unique (the player is free to choose any of the s symbols that appear with the maximum multiplicity: he/she may always guess, for example, the symbol they like most among those s , or a uniformly random choice), the distribution of the $X_{j,s}$ remains uniform. Moreover, these random variables are conditionally independent, given the times $\tau_{s,j}$. Finally, observe that for each j , the number of relevant $\tau_{j,s}$ is $1 \leq j \leq m$ and $1 \leq s \leq \tilde{W}_j$, so that we obtain conditional independence given the \tilde{W}_j . \square

Remark 2.5. Following Remark 2.1, our assumption on ϵ guarantees that the expected number of correct guesses is lower-bounded by $\ln n + O(1)$. This can be seen by looking at \tilde{W}_m (i.e. guesses early on in the game).

Remark 2.6. Since $\tilde{W}_j \geq 1$ for all j , we have the deterministic bound $S_m \geq m$. These correspond to the correct guesses when there is only one card appearing with the maximum multiplicity, which will eventually result in a correct guess with certainty.

For convenience, we let S'_m denote the total score conditioned on the random variables \tilde{W}_j for $1 \leq j \leq m$. Lemma 2.1 says that S'_m is a sum of independent Bernoulli random variables. We can leverage this to obtain the following result.

Lemma 2.2. Consider a deck as in the assumption of Theorem 1.2. Let S'_m denote the total score conditioned on the \tilde{W}_j , and let μ'_n and σ'_n denote the conditional mean and standard deviation. Then

$$\mu'_n = \sum_{j=1}^m \ln \tilde{W}_j + O(1), \quad \sigma_n'^2 = \mu'_n + O(1).$$

Moreover, uniformly over the \tilde{W}_j , we have

$$\left| \mathbb{P}\left(\frac{S'_m - \mu'_n}{\sigma'_n} \leq x\right) - \Phi(x) \right| \lesssim \frac{1}{(\ln n)^{3/2}},$$

where $\Phi(x)$ denotes the cumulative distribution function of a standard normal random variable.

Proof. By means of Lemma 2.1 and linearity of expectation, we can write

$$\begin{aligned} \mu'_n &= \sum_{j=1}^m \left(1 + \frac{1}{2} + \dots + \frac{1}{\tilde{W}_j}\right), \\ \sigma_n'^2 &= \sum_{j=1}^m \left[\left(1 + \frac{1}{2} + \dots + \frac{1}{\tilde{W}_j}\right) - \left(1 + \frac{1}{2^2} + \dots + \frac{1}{\tilde{W}_j^2}\right) \right]. \end{aligned}$$

Using the well-known facts

$$1 + \frac{1}{2} + \dots + \frac{1}{k} = \ln k + O(1), \quad \sum_{n=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6} < \infty,$$

we deduce that

$$\mu'_n = \sum_{j=1}^m \ln \tilde{W}_j + O(1), \quad \sigma_n'^2 = \mu'_n + O(1). \tag{2.1}$$

Moreover, using the fact that $\tilde{W}_m \geq \epsilon n$ (see Remark 2.1), we deduce that

$$\sigma_n'^2 \geq \ln n + O(1) \tag{2.2}$$

uniformly over all realizations of the \tilde{W}_j . Since the second and third moments of a Bernoulli random variable are the same, a standard Berry–Esseen bound for non-identically distributed random variables (see e.g. [16]) gives

$$\left| \mathbb{P}\left(\frac{S'_m - \mu'_n}{\sigma'_n} \leq x\right) - \Phi(x) \right| \lesssim \frac{1}{\sigma_n'^3} \lesssim \frac{1}{(\ln n)^{3/2}}. \quad \square$$

Remark 2.7. Notice that the proof already shows that the limiting fluctuations for the score S_m^n are distributed as mixtures of normal random variables. In order to prove Theorem 1.2, it will suffice to show suitable concentration for the conditional mean and variance.

2.3. Removing the conditioning

As pointed out in Remark 2.7, the conditional CLT for S'_m shown in Lemma 2.2 will suffice for our purposes if we can show a suitable concentration for the conditional means and variance μ'_n and σ'_n . Towards this, the main ingredient will be the following.

Lemma 2.3. Consider a deck \mathbf{m} satisfying the assumptions of Theorem 1.2. Let μ'_n be the conditional mean of the total score S_m given the random variables \tilde{W}_j . Then

$$\text{Var}[\mu'_n] = O((\ln \ln n)^2).$$

Proof. First, we claim that it suffices to show

$$\mathbb{E} \left[(\ln \tilde{W}_j - c_j)^2 \right] = O((\ln \ln n)^2), \tag{2.3}$$

where

$$c_j = c_j(n) = \frac{\ln n}{j + 1},$$

and $1 \leq j \leq m - 1$ (the case $j = m$ is obvious since \tilde{W}_m is deterministic). Here, the implicit constant in $O((\ln \ln n)^2)$ depends on the maximum multiplicity of the deck \mathbf{m}_{\max} , which is bounded by some constant m . Therefore, assuming (2.3), the desired conclusion follows, using the first bound in (2.1) together with the triangle inequality and the well-known fact that the variance minimizes the square discrepancy from any constant.

The goal now is to reduce concentration properties of \tilde{W}_j to those of $W_{j-1,t}$ and T_j , for which we can exploit Theorem 2.1. To this end, consider two sequences $f_n = \ln n$ and $g_n = 1/\ln n$. Recall that $T_j > t$ if and only if $W_{j,t} = 0$. In particular, Theorem 2.1 allows us to approximate probabilities of the form $T_j \in [a, b]$, with an explicit error term. More precisely, we have

$$\left| \mathbb{P}\left(\frac{T_j}{Cn^{j/(j+1)}} \notin [g_n, f_n]\right) - (e^{-f_n} + 1 - e^{-g_n}) \right| \lesssim \frac{f_n}{n^{1/(j+1)}} \lesssim \frac{\ln n}{n^{1/(j+1)}}. \tag{2.4}$$

Let $t_- := Cn^{j/(j+1)}g_n$ and let $t_+ := Cn^{j/(j+1)}f_n$. Using the fact that $1 \leq \tilde{W}_j \leq n$, as well as the definition of c_j , we have $\ln \tilde{W}_j, c_j \leq \ln n$. Therefore, on the event $T_j \notin [t_-, t_+]$, we obtain

$$\mathbb{E} \left[(\ln \tilde{W}_j - c_j)^2 1_{\{T_j \notin [t_-, t_+]\}} \right] = O((\ln \ln n)^2).$$

We can thus restrict our attention to the event $T_j \in [t_-, t_+]$. Since $W_{j-1,t}$ is weakly increasing in t , in the regime $T_j \in [t_-, t_+]$, we have

$$W_{j-1,t_-} \leq \tilde{W}_j \leq W_{j-1,t_+}.$$

The key gain is that we overcome the complicated dependence mechanism behind the definition of the \tilde{W}_j ; recall that T_j and $W_{j-1,t}$ are *not* independent. Using Theorem 2.1, we know

$$d_{\text{tv}}(W_{j-1,t_-}, \text{Poi}(\lambda_-)) \lesssim \frac{t_-}{n} \lesssim \frac{g_n}{n^{1/(j+1)}}, \quad d_{\text{tv}}(W_{j-1,t_+}, \text{Poi}(\lambda_+)) \lesssim \frac{t_+}{n} \lesssim \frac{f_n}{n^{1/(j+1)}},$$

where λ_+ and λ_- satisfy

$$\lambda_+ = \Theta(n^{1/(j+1)}f_n), \quad \lambda_- = \Theta(n^{1/(j+1)}g_n). \quad (2.5)$$

In particular, since $g_n = 1/\ln n$ we obtain that

$$\mathbb{P}(W_{j-1,t_-} = 0) \lesssim e^{-\lambda_-} + \frac{g_n}{n^{1/(j+1)}} \lesssim \frac{1}{n^{1/(j+1)}}.$$

Using the fact that $1 \leq \tilde{W}_j \leq n$, as well as the definition of c_j , we have $\ln \tilde{W}_j, c_j \leq \ln n$. This allows us to bound

$$\mathbb{E}\left[|\ln \tilde{W}_j - c_j|^2 \mathbf{1}_{\{W_{j-1,t_-}=0\}}\right] \leq (\ln n)^2 \mathbb{P}(W_{j-1,t_-} = 0) \lesssim \frac{(\ln n)^2}{n^{1/(j+1)}}.$$

Therefore we can restrict our attention to the event $\{T_j \in [t_-, t_+]\} \cap \{W_{j-1,t_-} > 0\}$. On this event we have

$$|\ln \tilde{W}_j - c_j|^2 \leq |\ln W_{j-1,t_-} - c_j|^2 \mathbf{1}_{\{W_{j-1,t_-} > 0\}} + |\ln W_{j-1,t_+} - c_j|^2 \mathbf{1}_{\{W_{j-1,t_+} > 0\}}. \quad (2.6)$$

Using the fact that

$$c_j = \frac{1}{j+1} \ln n$$

and Remark 2.2, we have $|\ln W_{j-1,t_{\pm}} - c_j| \lesssim \ln n$. In particular, we can replace $W_{j-1,t_{\pm}}$ with $\text{Poi}(\lambda_{\pm})$ in (2.6) up to an error of order

$$\ln n \left(e^{-f_n} + 1 - e^{-g_n} + \frac{f_n}{n^{1/(j+1)}} + \frac{g_n}{n^{1/(j+1)}} \right).$$

Recall that $f_n = \ln n$, $g_n = 1/\ln n$. Therefore, using $1 - e^{-x} \leq x$, the above expression is $O(1)$. We are thus left to show that

$$\mathbb{E}\left[(\ln(\text{Poi}(\lambda_+)) - c_j)^2 \mathbf{1}_{\{\text{Poi}(\lambda_+) > 0\}}\right] + \mathbb{E}\left[(\ln(\text{Poi}(\lambda_-)) - c_j)^2 \mathbf{1}_{\{\text{Poi}(\lambda_-) > 0\}}\right] = O((\ln \ln n)^2).$$

Thanks to (2.5), we can replace c_j with $\ln(\lambda_{\pm})$ up to an error. That is,

$$|c_j - \ln(\lambda_-)|^2 \lesssim (\ln g_n)^2 = (\ln \ln n)^2, \quad |c_j - \ln(\lambda_+)|^2 \lesssim (\ln f_n)^2 = (\ln \ln n)^2.$$

The result thus follows by showing

$$\mathbb{E}[(\ln(\text{Poi}(\lambda_{+})) - \ln \lambda_{+})^2 1_{\text{Poi}(\lambda_{+}) > 0}] + \mathbb{E}[(\ln(\text{Poi}(\lambda_{-})) - \ln \lambda_{-})^2 1_{\text{Poi}(\lambda_{-}) > 0}] \lesssim (\ln \ln n)^2.$$

In fact we can do better than this (notice that, because of (2.5) and our choices of f_n and g_n , we know that $\lambda_{\pm} > 1$ for all n sufficiently large). We claim the following.

Claim 2.2. *Let $X = \text{Poi}(\lambda)/\lambda$ for $\lambda > 1$. Then*

$$\mathbb{E}[(\ln X)^2 1_{X > 0}] \leq C$$

for some absolute constant C .

Proof of Claim 2.2. Let $A = \{|X - 1| \geq 1/2, X \neq 0\}$. Then we have

$$\mathbb{E}[(\ln X)^2 1_{X > 0}] \leq (\ln 2)^2 + \mathbb{E}[(\ln X)^2 1_A].$$

Notice that if $X > 1$, we can bound $(\ln X)^2 \leq X$. On the other hand, if $1/\lambda \leq X < 1$ then $(\ln X)^2 \leq (\ln \lambda)^2$. Thus, on the event A , we have $(\ln X)^2 \leq (\ln \lambda)^2 + X$. Therefore, combining with the Cauchy–Schwarz inequality,

$$\mathbb{E}[(\ln X)^2 1_A] \leq \mathbb{E}[X 1_A] + (\ln \lambda)^2 \mathbb{P}(A) \leq (\mathbb{E}[X^2 1_A])^{1/2} (\mathbb{P}(A))^{1/2} + (\ln \lambda)^2 \mathbb{P}(A).$$

For the first term, we note that

$$\mathbb{E}[X^2] = \frac{\lambda^2 + \lambda}{\lambda^2} \leq 2.$$

On the other hand, the Chernoff bound for Poisson random variables gives

$$\mathbb{P}(A) \leq \mathbb{P}\left(|\text{Poi}(\lambda) - \lambda| \geq \frac{\lambda}{2}\right) \leq 2e^{-\lambda/12}$$

from which the claim follows at once. □

This completes the proof. □

We are now ready to prove the main result, which will be an easy consequence of the lemmas we have proved so far.

Proof of Theorem 1.2. We start with the first part. The asymptotic result for the mean is the main result in [10]. As for the variance σ_n^2 , we can use the law of total variance to write in terms of the conditional mean and variance (see Lemma 2.2) as

$$\sigma_n^2 = \text{Var}[\mu'_n] + \mathbb{E}[\sigma_n'^2].$$

Therefore the asymptotic result follows by using (2.1), which shows that the second term is asymptotically the same as μ_n , together with Lemma 2.3, which shows that the first term is negligible.

We now move to the second part, namely the proof of the CLT. Let us define a random variable

$$y = y_n(x) = \frac{\sigma}{\sigma'} x + \frac{\mu - \mu'}{\sigma'}.$$

Now observe that

$$\begin{aligned} \left| \mathbb{P}\left(\frac{S_{\mathbf{m}^n} - \mu_n}{\sigma_n} \leq x\right) - \Phi(x) \right| &= \left| \mathbb{E} \left[\mathbb{P}\left(\frac{S'_{\mathbf{m}^n} - \mu_n}{\sigma_n} \leq x \mid \tilde{W}_1, \dots, \tilde{W}_m\right) - \Phi(x) \right] \right| \\ &= \left| \mathbb{E} \left[\mathbb{P}\left(\frac{S'_{\mathbf{m}^n} - \mu'_n}{\sigma'_n} \leq y \mid \tilde{W}_1, \dots, \tilde{W}_m\right) - \Phi(x) \right] \right| \\ &\leq \|\tilde{F} - \Phi\|_\infty + |\mathbb{E}[\Phi(y) - \Phi(x)]|, \end{aligned} \quad (2.7)$$

where

$$\tilde{F}(z) := \mathbb{P}\left(\frac{S'_{\mathbf{m}} - \mu'}{\sigma'} \leq z\right).$$

We know that

$$\|F - \Phi\|_\infty \lesssim \frac{1}{(\ln n)^{3/2}}$$

from Lemma 2.2. Therefore it suffices to bound $|\mathbb{E}[\Phi(y) - \Phi(x)]|$. Using (2.1) and (2.2), we observe that

$$\sigma_n - \sigma'_n = \frac{\sigma_n^2 - \sigma_n'^2}{\sigma_n + \sigma'_n} = o(|\mu_n - \mu'_n| + 1)$$

with probability one. Therefore

$$|y(x) - x| \leq \frac{|\sigma'_n - \sigma_n|}{\sigma'_n} |x| + \frac{|\mu_n - \mu'_n|}{\sigma'_n} \lesssim (1 + |x|) \frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}} \quad (2.8)$$

with probability one.

If $|x| \leq 1$, using $|\Phi(x) - \Phi(y)| \leq |x - y|$ and (2.8) we conclude that

$$|\Phi(y) - \Phi(x)| \lesssim \frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}},$$

with a constant that does not depend on x . Therefore Lemma 2.3 allows us to conclude

$$\mathbb{E}[|\Phi(y) - \Phi(x)|] \lesssim \frac{\ln \ln n}{\sqrt{\ln n}}.$$

If, instead, $|x| \geq 1$, define the event (recall that $y(x)$ is a random variable)

$$A_n = A_n(x) = \{|y(x) - x| \leq |x|/2\}.$$

On the event A_n , x and y have the same sign, and hence we have the bound

$$|\Phi(y) - \Phi(x)| \leq |x - y| \max_{z \in [\min(x,y), \max(x,y)]} \Phi'(z) \lesssim |x - y| e^{-\min(x^2, y^2)/2}.$$

Therefore, with probability one and for all n sufficiently large, we have

$$|\Phi(y) - \Phi(x)| \lesssim (1 + |x|) e^{-\min(x^2, y^2)} \frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}}$$

On this event, the pre-factor depending on x is uniformly bounded from above, and thus

$$|\Phi(y) - \Phi(x)|1_{A_n} \lesssim \frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}},$$

with the implicit constant in \lesssim independent of x . Using Lemma 2.3 together with Cauchy–Schwarz, we conclude that

$$\mathbb{E}[|\Phi(y) - \Phi(x)|1_{A_n}] \lesssim \frac{\ln \ln n}{\sqrt{\ln n}}.$$

For $|x| \geq 1$ and on the complement of event A_n , we use the bound $|\Phi(x) - \Phi(y)| \leq 1$ to deduce

$$\mathbb{E}[|\Phi(y) - \Phi(x)|1_{A_n^c}] \leq \mathbb{P}(A_n^c).$$

However, using (2.8), on the complement of A_n we have

$$\frac{|x|}{2} \leq |y(x) - x| \lesssim (1 + |x|) \frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}}.$$

In particular, this entails that for some constant $K > 0$ (independent of x), we have

$$\frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}} \geq K.$$

Using Lemma 2.3 once more, together with Chebyshev’s inequality, we conclude that

$$\mathbb{P}(A_n^c) \leq \mathbb{P}\left(\frac{|\mu_n - \mu'_n| + 1}{\sqrt{\ln n}} \geq K\right) \lesssim \frac{\ln \ln n}{\sqrt{\ln n}}.$$

This completes the proof. □

3. Discussion

It is natural to ask whether the assumptions of our main Theorem 1.2 are needed. The very first obstacle is given by the use of Theorem 2.1. One could keep track of the dependence of m (resp. ϵ) in their error bounds, and thus extend our result by allowing some moderate growth (resp. decay).

It is worth remarking that Lemma 2.1 holds for all decks, while the bound in Lemma 2.2 continues to hold as long as the conditional variance goes to infinity. In particular, under this condition we are guaranteed to have convergence to a mixture of normal random variables. Notice that the fact that m remains bounded plays no role in this part of the proof, while we needed our condition on ϵ . While this may not be sharp, some care has to be taken if one type of card has a much higher multiplicity than all the others: for instance, in the case of finite n , this may be an obstacle to the convergence to a mixture of normal random variables (see [4] for the case where $n = 2$).

As for the convergence to a normal random variable (i.e. a trivial mixture), our method relies on m being finite, and it is an interesting problem to determine if this is a true limitation. This is intimately related to the understanding of the concentration properties of \tilde{W}_j , i.e. the ‘number of ties at the top’, when j grows. At least for some regimes of j , a closely related

result is available in [14], where asymptotics for \tilde{W}_j are shown if the hypergeometric process is replaced by the multinomial process (i.e. if cards are reinserted into the deck).

It is worth pointing out that while there is hope for a CLT to hold even when m grows much faster than n , the asymptotic of the expected value (first part of Theorem 1.2) eventually breaks down, as shown in [15].

Acknowledgements

We thank Persi Diaconis for suggesting the problem and Jimmy He for the idea behind Lemma 2.1.

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

- [1] BLACKWELL, D. AND HODGES, J. L. (1957). Design for the control of selection bias. *Ann. Math. Statist.* **28**, 449–460.
- [2] CHUNG, F., DIACONIS, P., GRAHAM, R. AND MALLOWS, C. L. (1981). On the permanents of complements of the direct sum of identity matrices. *Adv. Appl. Math.* **2**, 121–137.
- [3] DIACONIS, P. (1978). Statistical problems in ESP research. *Science* **201**, 131–136.
- [4] DIACONIS, P. AND GRAHAM, R. (1981). The analysis of sequential experiments with feedback to subjects. *Ann. Statist.* **9**, 3–23.
- [5] DIACONIS, P., GRAHAM, R., HE, X. AND SPIRO, S. (2022). Card guessing with partial feedback. *Combinatorics Prob. Comput.* **31**, 1–20.
- [6] DIACONIS, P., GRAHAM, R. AND HOLMES, S. P. (2001). Statistical problems involving permutations with restricted positions. In *State of the Art in Probability and Statistics* (Lecture Notes: Monograph Series **36**), pp. 195–222. Institute of Mathematical Statistics.
- [7] DIACONIS, P., GRAHAM, R. AND SPIRO, S. (2022). Guessing about guessing: practical strategies for card guessing with feedback. *Amer. Math. Monthly* **129**, 607–622.
- [8] EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403–417.
- [9] FISHER, R. A. (1936). Design of experiments. *British Med. J.* **1**, 554.
- [10] HE, J. AND OTTOLINI, A. (2021). Card guessing and the birthday problem for sampling without replacement. Available at [arXiv:2108.07355](https://arxiv.org/abs/2108.07355).
- [11] KUBA, M. AND PANHOLZER, A. (2023). On card guessing with two types of cards. Available at [arXiv:2303.04609](https://arxiv.org/abs/2303.04609).
- [12] LIU, P. (2021). On card guessing game with one time riffle shuffle and complete feedback. *Discrete Appl. Math.* **288**, 270–278.
- [13] NIE, Z. (2022). The number of correct guesses with partial feedback. Available at [arXiv:2212.08113](https://arxiv.org/abs/2212.08113).
- [14] Ottolini, A. (2020). Oscillations for order statistics of some discrete processes. *J. Appl. Prob.* **57**, 703–719.
- [15] OTTOLINI, A. AND STEINERBERGER, S. (2023). Guessing cards with complete feedback. *Adv. Appl. Math.* **150**, 102569.
- [16] SHEVTSOVA, I. G. (2010). An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Math.* **82**, 862–864.