Valid and powerful second-level group statistics for decoding accuracy: Information prevalence inference using the i-th order statistic (i-test)

Satoshi Hirose\textsuperscript{a,b}

\textsuperscript{a} Faculty of Psychology, Otemon Gakuin University, 2-1-15 Nishiai, Ibaraki-cho, Osaka 567-8502, Japan
\textsuperscript{b} Center for Information and Neural Networks (CiNet), Advanced ICT Research Institute, National Institute of Information and Communications Technology, 1-4, Yamadakka, Suita City, Osaka, Japan

\textbf{Abstract}

In functional magnetic resonance imaging (fMRI) decoding studies using pattern classification, a second-level group statistical test is typically performed after first-level decoding analyses for individual participants. In the second-level test, the mean decoding accuracy across participants is often tested against the chance-level accuracy (for example, one-sample Student t-test) to check whether information about the label, such as, experimental condition or cognitive content, is included in brain activation. Meanwhile, Allefeld et al., (2016) highlighted that significant results for such tests only indicate that “there are some people in the population whose fMRI data carry information about the experimental condition.” Therefore, such tests failed to conclude whether the effect is typical in the population. Based on this argument, they proposed an alternative method implementing the prevalence inference. In the present study, that method is extended to propose a novel statistical test called as the “information prevalence inference using the i-th order statistic” (i-test). The i-test has a high statistical power compared with the method proposed in Allefeld et al., (2016) and provides an inference regarding the typical effect in the population. In the i-test, the i-th lowest sample decoding accuracy (the i-th order statistic) is compared to the null distribution to verify whether the proportion of higher-than-chance decoding accuracy in the population (information prevalence) is higher than the threshold. Hence, a significant result in the i-test is interpreted as a majority of the population has information about the label in the brain. Theoretical details of the i-test are provided, its high statistical power is identified by numerical calculation, and the application of this method in an fMRI decoding is demonstrated.

\section{Introduction}

An ever-increasing number of functional magnetic resonance imaging (fMRI) studies use multi-voxel pattern classification analysis where the cognitive information (= label), such as visual input categories (e.g. Spiridon and Kanwisher 2002; Haxby et al., 2014; Nishida and Nishimoto 2018), types of movement (e.g. Gallivan et al., 2011; Nambu et al., 2015), and the effect of movement (e.g. Hirose et al., 2015; Hirose et al., 2018) are predicted (decoded) from fMRI signal patterns. The prediction accuracy (decoding accuracy; D-Acc) is considered as an index of the label information in the brain. The D-Acc is often compared to the chance-level accuracy, which is the theoretical expectation of the D-Acc without information about the label (e.g., 50% for binary classifications). When the D-Acc exceeds the chance level, the result is interpreted as information about the label is represented in the brain.

In real experiments, the D-Acc for multiple participants is combined to make conclusions regarding the D-Acc’s population distribution. This study deals with a statistical test for inferring the population proportion of the above-chance D-Acc based on the experimental results from multiple participants. By the analogy with the conventional univariate analysis for voxel activity (Friston et al., 1995), the mean D-Acc across participants has often been compared to the chance level by using Student’s t-test (e.g., Gilbert and Fung, 2018; Gallivan et al., 2011), or permutation test (Gallivan et al., 2013; Good, 2000; Nambu et al., 2015; Stelzer et al., 2013). These analyses are based on the belief that the sample mean is a good estimator of the population mean. Hence if the population mean is higher than the chance level, it is expected that the brain activation typically contains the information regarding the cognitive content. However, this interpretation is misleading. Indeed, in theory, the expectation of each participant’s D-Acc has to be equal or exceed that of the random classifier (= chance level) and can-

\textbf{Abbreviations:} D-Acc, decoding accuracy; fMRI, functional magnetic resonance imaging; i-test, information prevalence inference using the i-th order statistic; i-test-one, i-test with i=1; i-test-unif-bino, i-test with the assumption of a uniform distribution of true parameters and a binomial distribution of D-Acc.

\textbf{E-mail address:} satoshi.hirose@gmail.com

https://doi.org/10.1016/j.neuroimage.2021.118456.
Received 10 February 2020; Received in revised form 29 July 2021; Accepted 4 August 2021
Available online 5 August 2021
1053-8119/© 2021 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
not be lower than the chance level (Allefeld et al., 2016). Thus, a very small part of the population (e.g., one in a million) with higher-than-chance D-Acc results in the population mean higher than the chance level. Therefore, significant results of these tests can only indicate that “there are some people in the population whose fMRI data carry information about the experimental condition” (Allefeld et al., 2016), but fail to conclude whether the effect is typical in the population.

The above-mentioned problem is inevitable for statistical tests of the population mean. A possible solution is to assess the population information prevalence (i.e., the proportion of the population having label information in their brain activity), like other methods for evaluation of the population proportion for second-level group statistics, such as the dynamic causal model (Stephan et al., 2009), the second-level random effect of univariate analysis (Rosenblatt et al., 2014), and the conjunction analysis (Friston et al., 1999). Accordingly, a second-level group statistical test on D-Acc was proposed by Allefeld et al. (2016) where the population information prevalence is targeted instead of the population mean. In this method, the null hypothesis is that the proportion of the population having label information is not larger than the predetermined prevalence threshold (e.g., 0.5). Thus, the null hypothesis rejection means that proportion larger than the prevalence threshold (e.g., more than half) of the population has label information.

The method proposed by Allefeld et al. (2016) theoretically solved the above issue by implementing the population information prevalence and provided a meaningful inference regarding the typical effect. However, the method may have the disadvantage of a low statistical power (Section 3). Because the lowest D-Acc (minimum order statistic) among the participants is used as the test statistic, the method may fail to capture the population characteristics at the presence of one low D-Acc, leading to low statistical power.

In the present study, I propose an extension of the previous method. The idea is to generalize the method in Allefeld et al. (2016) for second or higher order statistics (i-th lowest D-Acc for \(i = 2, 3, \ldots \)). See Section 2.3.4 regarding the relationship between the method in Allefeld et al. (2016) and that proposed in this paper. The proposed method can be called as the “information prevalence inference using the i-th order statistic”, or i-test.

The paper is organized as follows. The proposed statistical test is explained in Section 2. The advantage of the method in terms of a high statistical power is demonstrated by providing numerical results for the artificial data in Section 3. An application of the method using a real fMRI dataset is demonstrated in Section 4. Finally, the results are summarized, and future directions of the study are discussed in Section 5.

2. Information prevalence inference using the i-th order statistic (i-test)

In this section, Section 2.1 clarifies the statistical model, defines the problem, and outlines the i-test calculation. Section 2.2 explains the i-test theoretical details. Section 2.3 introduces essential theoretical characteristics of the i-test, including its relationship with the previous method. Finally, a practical procedure for performing the i-test is explained in Section 2.4. Section 2.4 is self-contained and involves a minimum number of equations, so that readers may skip sections 2.2 and 2.3 to familiarize yourself with the practical procedure before switching to the theoretical details.

2.1. Notations and problem definition

Figure 1A displays the statistical model for the population and experimental result (sample D-Acc in an experiment). The population \(\Omega\) is composed of two subgroups, \(\Omega_+\) and \(\Omega_-\). People in \(\Omega_+\) have label information in their brain activity and, thus, their D-Acc expectations are higher than the chance level, while people in \(\Omega_-\) do not have label information and the expectations are at a chance level. A randomly chosen person with an index \(n\) from the population belongs to \(\Omega_+\) (\(n \in \Omega_+\)), with a probability \(\gamma\), or otherwise belongs to \(\Omega_-\) (\(n \in \Omega_-\)), with a significance threshold \(\alpha\). An experiment with \(N\) participants constitutes a two-step random sampling. First, \(N\) participants are independently and randomly sampled from the population (Random sampling of participants). Each sampled participant is associated with the D-Acc probability distribution, (probability mass function; \(p(a_n)\)). Second, the experimental results (sample D-Acc; \(\hat{a}_n\)) are randomly sampled from the distribution for each participant (Random sampling of D-Acc). Note that in this paper, a lowercase \(p\) depicts a probability mass function for discrete variables or probability density function for continuous variables, while an uppercase \(P\) stands for probabilities.

The i-test objective is to verify whether \(\gamma\) (\(P(n \in \Omega_+)\)) is larger than the predetermined threshold \(\gamma_0\) from \(N\) participants’ experimental results (\(\hat{a}_n, n = 1 \ldots N\)), with a significance threshold \(\alpha\).

To perform the i-test, we first identify the i-th order statistic \(\hat{a}_{i\gamma}\), which is the i-th lowest D-Acc observed in the experiment (Figure 1B). Then, \(P(a_n < \hat{a}_{i\gamma}|n \in \Omega_+)\), which is the probability that a participant without label information has D-Acc lower than \(\hat{a}_{i\gamma}\) (green area in Figure 1B) is estimated. The estimation can be performed by assuming a parametric distribution for D-Acc, such as the binomial distribution (Section 2.4), or by an empirical procedure, such as a permutation test (Supplementary Materials S3.1). Note that the i-test is performed without estimation of D-Acc distribution for \(\Omega_+\). After estimating \(P(a_n < \hat{a}_{i\gamma}|n \in \Omega_+)\), we check whether \(P(a_n \geq \hat{a}_{i\gamma}|H_0: \gamma \leq \gamma_0)\) is smaller than \(\alpha\), where \(P(a_n \geq \hat{a}_{i\gamma}|H_0: \gamma \leq \gamma_0)\) is the probability that the i-th order statistic \(a_n\) is not lower than the observed value \(\hat{a}_{i\gamma}\) under the null hypothesis (the green
area in Figure 1C). If it is smaller than $\alpha$, we conclude that the observed $\hat{a}_{ij}$ is so high that it has unlikely (with a probability smaller than the significance threshold $\alpha$) occurred under the null hypothesis. Thus, the null hypothesis is rejected and one may conclude that $\gamma$ is higher than the predetermined threshold $\gamma_0$ with the significance threshold $\alpha$.

The notations are tabulated in Table 1.

### Table 1
Notations used in the present study

| Notation | Description |
|----------|-------------|
| $N$ | number of participants |
| $\gamma_0$ | prevalence threshold |
| $\alpha$ | statistical threshold |
| $c$ | rank of order statistic |
| $H_0$ | null hypothesis ($\gamma \leq \gamma_0$) |
| $H^*$ | alternative hypothesis ($\gamma > \gamma_0$) |

Population
- $\Omega$: whole population
- $\Omega_S$: population subgroup with label information
- $\Omega_L$: population subgroup without label information

Population parameters
- $\gamma$: population information prevalence, i.e., $P(n \in \Omega)$
- $\theta$: parameters other than $\gamma$ that determine the true distribution of D-Acc
- $k$: the parameter space of $\theta$

Parameters of binomially distributed D-Acc
- $N_{null}$: number of trials
- $P_{\text{true}}$: probability of the correct decoding in a trial for the participant with label information
- $P_{\text{false}}$: probability of the correct decoding in a trial for the participant without label information ($=\text{chance level}$)

Index
- $\kappa$: index for participants

Experimental result (constant)
- $\hat{a}_{ij}$: observed D-Acc of a participants with index $i$ ($i = 1 \ldots N$)
- $a_{ij}$: $i$-th order statistic of the D-Acc observed in the experiment
- $a_{ij}$: D-Acc of a participant with index $n$
- $a_{ij}$: $i$-th order statistic of the D-Acc in the experiment

Variables for derivation
- $Q$: lower bound of $P(a \leq \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0)$ see Eqs. (2.5) and (2.6)
- $Q$: the estimation of $Q$
- $L$: upper bound of $P(a \geq \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0)$ see Eqs. (2.8) and (2.9)
- $L$: the estimation of $L$
- $i_{max}$: maximum available $i$ determined by Eq. (2.10), $P(\text{Significant})$: Probability that the i-test reports a significant result (statistical power at $\gamma > \gamma_0$, false alarm rate at $\gamma \leq \gamma_0$)

Functions
- $\text{BPDF}_k(\theta, K, \rho) = \sum_{|\theta| = K} C_k^{|\theta|}(1 - \rho)^{|\theta|/2}$: binomial probability mass function
- $\text{BCDF}_k(\theta, K, \rho) = \sum_{|\theta| = K} \text{BPDF}_h(\theta, K, \rho)$: binomial cumulative probability mass function

#### 2.2. Theoretical details

As noted above, the objective of the i-test is to verify whether $\gamma$ is larger than the predetermined threshold $\gamma_0$. Thus, the alternative hypothesis of $H_1 : \gamma > \gamma_0$ is set, and the null hypothesis is its negation $H_0 : \gamma \leq \gamma_0$. Under this null hypothesis, we evaluate whether $P(a_{ij} \geq \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) < \alpha$ holds.

The i-test formulation is conducted in four steps (Sections 2.2.1-2.2.4). First, $P(a_{ij} < \hat{a}_{ij}|\gamma)$, which is the probability that a participant has D-Acc lower than $\hat{a}_{ij}$ for a given $\gamma$, is formulated by Eq. (2.1). Secondly, $P(a_{ij} \geq \hat{a}_{ij}|\gamma)$, which is the probability that the i-th lowest D-Acc ($\hat{a}_{ij}$) is not lower than the observed one ($\hat{a}_{ij}$) for a given $\gamma$ is formulated by Eq. (2.2). Third, the lower bound $Q$ of $P(a_{ij} < \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0)$, which is the probability that a participant has lower D-Acc than $\hat{a}_{ij}$ under the null hypothesis is formulated by Eq. (2.5). Finally, the upper bound $L$ of $P(a_{ij} \geq \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0)$, which is the probability that the i-th lowest D-Acc ($\hat{a}_{ij}$) is not lower than $\hat{a}_{ij}$ under the null hypothesis, is formulated by Eq. (2.8).

#### 2.2.1. Derivation of $P(a_{ij} < \hat{a}_{ij}|\gamma)$

Considering that a participant belongs to $\Omega_+\gamma$ with the probability $\gamma$ and otherwise belongs to $\Omega_-$, the probability that a participant has D-Acc lower than $\hat{a}_{ij}$ is formulated as:

$$P(a_{ij} < \hat{a}_{ij}|\gamma) = (1 - \gamma)P(a_{ij} < \hat{a}_{ij}|n \in \Omega_+) + \gamma \cdot P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-)$$

(2.1)

Note that it is assumed that the D-Acc are identically distributed among participants, i.e., $P(a_{ij} < \hat{a}_{ij}|n \in \Omega_+)$ and $P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-)$ are independent of $n$, and, therefore, $P(a_{ij} < \hat{a}_{ij}|\gamma)$ is independent of $n$.

#### 2.2.2. Derivation of $P(a_{ij} \geq \hat{a}_{ij}|\gamma)$

Then, the probability that the i-th lowest D-Acc is not lower than $\hat{a}_{ij}$ for a given $\gamma$ is derived as follows:

$$P(a_{ij} \geq \hat{a}_{ij}|\gamma) = \text{BCDF}(i - 1, N, P(a_{ij} < \hat{a}_{ij}|\gamma))$$

(2.2)

This is because the condition the i-th lowest D-Acc not being lower than $\hat{a}_{ij}$ is identical to the condition fewer than $i$ of the $N$ participants having D-Acc lower than $\hat{a}_{ij}$, whose probability is formulated with the binomial cumulative distribution function.

#### 2.2.3. Lower bound of $P(a_{ij} < \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0)$

Following the basic nature of probability, i.e., $P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-) \geq 0$, the lower bound of $P(a_{ij} < \hat{a}_{ij}|\gamma)$ can be easily derived from Eq. (2.1) as

$$P(a_{ij} < \hat{a}_{ij}|\gamma) \geq (1 - \gamma)P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-)$$

(2.3)

Consequently, a lower bound of the probability under the null hypothesis $H_0 : \gamma \leq \gamma_0$ is

$$P(a_{ij} < \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \geq (1 - \gamma_0)P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-)$$

(2.4)

I define $Q$ to equal the right side of Eq. (2.4), i.e.,

$$Q = (1 - \gamma_0)P(a_{ij} < \hat{a}_{ij}|n \in \Omega_-)$$

(2.5)
and Eq. (2.4) may be expressed as

\[ Q \leq P(a_\delta < \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \]  

(2.6)

2.2.4. Upper bound of \( P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \)

As \( BCDF(i-1, N, P(a_\delta < \hat{a}_{ij}|\gamma)) \) in Eq. (2.2) monotonically decreases with \( P(a_\delta < \hat{a}_{ij}|\gamma) \) and \( P(a_\delta < \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \) is bounded below by \( Q \) (Eq. (2.6)), we obtain the following upper bound for \( P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \):

\[ BCDF(i-1, N, Q) \geq P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \]  

(2.7)

We define \( L \) to equal the left side of Eq. (2.7), i.e.,

\[ L \equiv BCDF(i-1, N, Q) \]  

(2.8)

and Eq. (2.7) may be expressed as

\[ L \geq P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \]  

(2.9)

2.2.5. i-test calculation

The i-test calculation is conducted as follows. First, \( P(a_\delta > \hat{a}_{ij}|n \in \Omega_1) \) is estimated by assuming a parametric distribution for D-Acc or by an empirical procedure. Then, \( Q \) is calculated with Eq. (2.5), and, consequently, \( L \) is calculated with Eq. (2.8). Assuming that the estimation is correct, the expression \( a > L \geq P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \) follows from Eq. (2.9) when \( L \) is smaller than the statistical threshold \( a \). Therefore, it is confirmed that \( a > P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \) and we may conclude that the observed i-th order statistic \( \hat{a}_{ij} \) is so high that it has unlikely (with probability smaller than \( a \)) occurred under the null hypothesis. Therefore, we reject the null hypothesis and accept \( H_i : \gamma > \gamma_0 \), i.e., a proportion larger than \( \gamma_0 \) of the population has label information in their brain activity. 

2.3. Theoretical characteristics

2.3.1. Parameter constraint

The i-test requires three predetermined parameters: \( a \) (threshold for statistical significance), \( \gamma_0 \) (threshold for the information prevalence), and \( i \) (rank of the order statistic). Together with the number of participants \( N \), they should satisfy the following inequality:

\[ a > BCDF(i-1, N, 1-\gamma_0) \]  

(2.10)

This is because \( Q \) is bounded above by \( 1-\gamma_0 \) from Eq. (2.5) and \( P(a_\delta < \hat{a}_{ij}|n \in \Omega_1) \leq 1 \). Therefore, given that \( L \) decreases with \( Q \) (Eq. (2.8)), one may find that \( L \) is bounded below as follows:

\[ L \geq BCDF(i-1, N, 1-\gamma_0) \]  

(2.11)

Herein, Eq. (2.10) must be satisfied; otherwise, \( L \) is never lower than \( a \) and the i-test never reports a significant result.

2.3.2. Expectation of the statistical power

The statistical power is expressed as \( P(\text{Significant}|H_i : \gamma > \gamma_0) \), which is the probability that the i-test reports a significant result under the alternative hypothesis. This can be formulated as:

\[ P(\text{Significant}|H_i : \gamma > \gamma_0) = \int_{\gamma_0}^{1} \int_{\theta} P(\text{Significant}|\gamma, \theta) p(\gamma, \theta) d\theta d\gamma \]  

(2.12)

where \( \theta \) is the parameter other than \( \gamma \) that determines the true distribution of D-Acc, \( \Theta \) is the parameter space, \( p(\gamma, \theta) \) is the joint prior distribution, and \( P(\text{Significant}|\gamma, \theta) \) is the probability that the i-test reports a significant result for given \( \gamma \) and \( \theta \). This is used for selection of parameter \( i \) (see Section 2.4.2). Refer to Appendix B for an example of the statistical power calculation and formulation of the selection of \( i \).

2.3.3. Control of the false alarm rate

It is analytically guaranteed that the false alarm rate \( P(\text{Significant}|H_0 : \gamma \leq \gamma_0) \), which is the probability that the i-test reports a significant result under the null hypothesis, does not exceed the statistical threshold \( a \), when the estimation of \( P(a_\delta|n \in \Omega_1) \) is correct.

When the estimation is correct, the estimation of \( P(a_\delta < \hat{a}_{ij}|n \in \Omega_1) \) is correct for any \( \hat{a}_{ij} \) and therefore, \( Q = Q \) following Eq. (2.5) and, consequently, \( L = L \) following Eq. (2.8). Keeping in mind that \( P(\text{Significant}|H_0 : \gamma \leq \gamma_0) = P(a > L|H_0 : \gamma \leq \gamma_0) \) by definition and \( L \) is the upper bound of \( P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \), one may determine that

\[ \begin{align*}
    P(\text{Significant}|H_0 : \gamma \leq \gamma_0) &= P(a > L|H_0 : \gamma \leq \gamma_0) \quad \text{[By definition]} \\
    &= P(a > L|H_0 : \gamma \leq \gamma_0) \\
    &\leq P(a > a|H_0 : \gamma \leq \gamma_0) \quad \text{[definition]} \\
    &\leq P(a > a|H_0 : \gamma \leq \gamma_0) \quad \text{[definition]} \\
    &\leq a
\end{align*} \]  

(2.13)

\( P(a > a|H_0 : \gamma \leq \gamma_0) \) means the probability that the sample i-th order statistic \( \hat{a}_{ij} \) is higher than the right \( a \) point of its distribution \( (a_\delta|n) \), and therefore, it is at or smaller than \( a \).

2.3.4. Relation with the previous method in Allefeld et al. (2016)

By fixing \( i = 1 \), Eq. (2.8) transforms simply to the N-th power of \( 1-Q \):

\[ L = (1-Q)^N \]  

(2.14)

and the i-test reduces to the method referred as the “Permutation-based information prevalence inference using the minimum statistic” proposed by Allefeld et al. (2016). Hereafter, this is called as i-test-one. 

2.4. Practical Procedure and implementation

The i-test procedure is performed by 1) setting two threshold parameters: \( a \) and \( \gamma_0 \) (Section 2.4.1), 2) determining \( i \) (Section 2.4.2), 3) identifying \( \hat{a}_{ij} \) and estimating \( P(a_\delta < \hat{a}_{ij}|n \in \Omega_1) \) (Section 2.4.3), and 4) calculating the upper bound \( (L) \) of \( P(a_\delta > \hat{a}_{ij}|H_0 : \gamma \leq \gamma_0) \) and comparing \( L \) with \( a \) (Section 2.4.4).

As an example of the i-test implementation, i-test-unif-bino (i-test with the assumption of uniform distribution of true parameters and the binomial distribution of D-Acc) is presented, which can be applied within realistic computational time (<5 seconds for \( N \leq 100 \), \( N_{\text{max}} \leq 1000 \) with MATLAB 2018 on iMacPro, 64GB Memory, 10-core 3GHz processor). Following the binomial distribution assumption, D-Acc distribution \( p(a_\delta) \) is determined by two unknown parameters: the true information prevalence, \( \gamma \), and \( F_{\text{correct}} \), which is the probability of the correct decoding in a trial for participants without label information. These parameters are assumed to follow the uniform prior distribution. Pseudocode of the procedures is presented in Appendix C. Refer to Supplementary Material S3 for other implementations.

2.4.1. Setting the \( a \) and \( \gamma_0 \) threshold parameters

As the first step of the i-test, the two threshold parameters, \( a \) (the threshold for the statistical significance) and \( \gamma_0 \) (the threshold for the information prevalence), should be set. These parameters explicitly appear in the concluding statement of the test. Namely, with the significant result of the i-test, one may conclude that a proportion larger than \( \gamma_0 \) of the population has label information in their brain activity with a false-alarm rate less than \( a \). Therefore, users may select these values based on their purpose.

As the primary choice, \( a = 0.05 \) and \( \gamma_0 = 0.5 \) are suggested, because \( a = 0.05 \) is the most commonly used value in neuroscience studies, while \( \gamma_0 = 0.5 \) may be intuitively acceptable because significant result leads us to conclude that a majority (more than half) of the population has label information, and at present this is a standard value in the existing statistical methods on the population prevalence (e.g. Friston et al., 1999; see also Discussion in Allefeld et al., 2016).
2.4.2. Determination of $i$ (the rank of the order statistic)

After the two threshold parameters are set, the upper limit of $i (i_{\text{max}})$ is determined from Eq. (2.10). The selection of $i$ among $i = 1, 2, \ldots i_{\text{max}}$ does not affect the concluding statement, but it affects the statistical power (see Section 3). Therefore, the optimal $i$ that provides a maximal expected statistical power (Eq. (2.12)) is suggested.

The above-mentioned assumptions enable numerical calculation of the optimal $i$ that maximizes the expected statistical power (refer to Appendix B.2 for full derivation).

2.4.3. Identification of $\tilde{a}_i$ and estimation of $P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i)$

Next, $\tilde{a}_i$ is identified from experimental results, and $P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i)$ is estimated. Following the binomial assumption, this estimation can be performed on the base of the known parameters: number of trials ($N_{\text{trial}}$), sample order statistic ($\tilde{a}_i$), and probability of the correct decoding in a trial for participants without label information ($P_{\text{correct}}$), which is the chance level (e.g., 0.5 for binary).

2.4.4. Calculation of $L$ and its comparison with $a$

Finally, by substituting $P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i)$ into Eqs. (2.5) and (2.8) the upper bound $L$ of $P(\tilde{a}_i \geq \tilde{a}_i | H_0 : \gamma \leq \gamma_0)$ is estimated. If the estimation $L$ satisfies the inequality $L < a$, the null hypothesis $H_0 : \gamma \leq \gamma_0$ is rejected certifying that $\gamma$ is higher than the predetermined threshold $\gamma_0$ with the significance threshold $a$.

3. Numerical calculation in artificial experiments

The $t$-test was designed to improve the statistical power compared to the $t$-test-one. The results of the statistical power calculation in the synthetic situations are presented below to demonstrate the advantage of $t$-test (see the full derivation in Appendix A). Note that the numerical calculation results reported in this section were replicated in simulations with the same parameters (Supplementary Material S1.1). Furthermore, the results were replicated with different prevalence threshold values ($\gamma_0 = 0.1, 0.3, \text{and } 0.7$; Supplementary Material S2).

3.1. Definition of artificial data

Let us consider artificial experiments where each participant performed $N_{\text{trial}}$ trials with a binary choice (chance level: 0.5). Each participant had label information ($n \in \Omega_i$) with probability $\gamma_i$ or did not have label information ($n \in \Omega_i$) otherwise. In a trial, the decoder for participants with label information ($n \in \Omega_i$) predicted the label correctly with probability $P_{\text{correct}}$ independently across trials, while for participants without label information ($n \in \Omega_i$), the label was predicted randomly and independently. Therefore, the D-Acc for $n \in \Omega_i$ and $n \in \Omega_i$ followed the binomial distributions, i.e., $p(a_i | n \in \Omega_i) = \text{BPDF}(N_{\text{trial}}|a_i, N_{\text{trial}}, P_{\text{correct}})$ and $p(a_i | n \in \Omega_i) = \text{BPDF}(N_{\text{trial}}|a_i, N_{\text{trial}}, P_{\text{correct}})$, where $P_{\text{correct}}$ is the chance level, i.e., 0.5. Note that the decoding accuracies do not always follow the binomial distribution in the real fMRI decoding, because the trials are not completely independent (refer to Discussion and Supplementary Material S1.2).

The threshold for statistical significance $a$ was fixed at 0.05 and the prevalence threshold $\gamma_0$ was fixed at 0.5. The numerical calculation of $P(\text{Significant})$, which is the probability that $t$-test reports the significant result, was performed for all possible $i (i = \ldots i_{\text{max}})$ with all combinations of the following parameters: $N_{\text{trial}} = 10, 100, 1000, N = 5, 6, \ldots 100, \gamma$ from 0 to 1 with 0.01 increment and $P_{\text{correct}}$ from 0.51 to 1 with 0.01 increment. $N < 5$ was omitted because the $t$-test cannot be applied for any $i$ due to the constraint of Eq. (2.10). The $t$-test calculation was made with the correct estimation of $P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i)$, i.e., $P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i) = P(\tilde{a}_i < \tilde{a}_i | n \in \Omega_i)$. Note that $P(\text{Significant})$ indicates the statistical power at $\gamma > \gamma_0$ and the false alarm rate at $\gamma \leq \gamma_0$.

3.2. Extending $t$-test-one to $t$-test can improve the statistical power

First, the relationship between $i$ and the statistical power with $N_{\text{trial}} = 100$ and $N = 50$ was illustrated (Figure 2A–E for $i = 1, 5, 10, 15, 19$, SupplementaryGIF1 for $i = 1 \ldots 19$). As demonstrated in Figure 2A, the $t$-test-one provided a high statistical power in the restricted area near $\gamma = 1$. Particularly, the statistical power of the $t$-test-one exceeded 0.8 (Cohen, 2013) only at $\gamma \geq 0.97$, indicating that the $t$-test-one could report a significant result with sufficient probability only when almost all people in the population have label information although the prevalence threshold was $\gamma_0 = 0.5$. The area expanded downward (toward smaller $\gamma$) as $i$ is increased (Figure 2B–E), indicating that the $t$-test with larger $i$ can report a significant result with high probability at smaller $\gamma$.

In the implementation of the $t$-test-unif-bino, we used the optimal $i$, which maximizes the expectation of the statistical power under the assumption of the binomial distribution of D-Acc and uniform distribution of the parameters for true D-Acc distribution (Section 2.4.2). With this implementation, the optimal $i$ was 15 for this set of parameters (Figure 2D). Figure 2E illustrates the improvement of the $t$-test-unif-bino statistical power (Figure 2D) compared with the $t$-test-one (Figure 2A).

As observed, the statistical power improved in the large area (red area in Figure 2F) at $\gamma = 0.6-0.9$ and $P_{\text{correct}}>0.6$, while degraded in the smaller area (blue area) at $\gamma > 0.9$ and $P_{\text{correct}} < 0.6$.

The same comparison was performed for the combinations of $N_{\text{trial}} = 10, 100, 1000$ and $N = 5, 6, \ldots 100$. The statistical power of the $t$-test-one exceeded 0.8 in all cases only at $\gamma \geq 0.96$. In contrast, the statistical power of the $t$-test-unif-bino exceeded 0.8 with smaller $\gamma$ in many cases (see Figure 2G, H, and I for excerpted results, SupplementaryGIF2–4 for all results). Exceptions were observed when $i = 1$ was used for the $t$-test-unif-bino; therefore $t$-test-unif-bino was identical to $t$-test-one.

As expected from the analytical analysis (Section 2.3.3), the false alarm rate was below the statistical threshold $a$ for all the tested parameters with $i$ at $N_{\text{trial}} = 10, 100, 1000, N = 5, 6, \ldots 100, \gamma$ from 0 to 0.5 with 0.01 increment, and $P_{\text{correct}}$ from 0.51 to 1 with 0.01 increment (not shown).

In summary, extension from the $t$-test-one to the $t$-test (implemented as the $t$-test-unif-bino) improved the statistical power making it possible to report a significant result with a smaller $\gamma$, while the power was reduced in the smaller range around $\gamma = 1$ and the relatively low $P_{\text{correct}}$.

4. Application to empirical fMRI data

An example of the $t$-test application to a real fMRI dataset reported by Hirose et al., 2015 is presented below.

4.1. Experiment and analysis procedures

For the experimental details, please refer to the original paper (Hirose et al., 2015). Briefly, 8 participants ($N = 8$) performed a finger-tapping task using either his/her right index or middle finger for 1.5 s. Each participant underwent 10 sessions, and each session consisted of 20 trials comprising 10 index finger trials and 10 middle finger trials in a random order. Trials were separated by inter-trial intervals (ITI) of 6 s plus 0.5 s instruction periods just before the trials. The finger was predicted from the preprocessed fMRI volume measured in the time range of 2–6 s after the end of each trial. The classifier was trained by the sparse logistic regression algorithm (Yamashita et al., 2008) on voxel activities in the whole brain (33,270 ± 1,570 voxels). Classification accuracy was evaluated by the hold-out validation procedure, in which the classifiers were trained using 100 trials from the odd sessions (training dataset), and 100 trials from the even sessions (test dataset) were used to evaluate D-Acc. D-Acc was calculated as the number of test trials with correct prediction divided by the number of all test trials. The preprocessing and first-level decoding analyses for each participant...
were performed using the Multi-Voxel Pattern Classification Toolbox (https://github.com/satoshi-hirose/MVPCToolbox).

Then, the i-test-unif-bino and the i-test-one were applied to the experimental results. For both tests $P(\delta < \delta_{ij}|n \in \Omega)$ was estimated based on the binomial assumption, and the thresholds were set as $a = 0.05$ and $\gamma_0 = 0.5$.

4.2. Results

The D-Acc for the participants were 0.52, 0.65, 0.73, 0.74, 0.75, 0.7, 0.8, 0.9.

i-test-unif-bino: First, the optimal $i$ was chosen as follows (Section 2.4.2). The maximum available $i$ was 2 for $a = 0.05$, $\gamma_0 = 0.5$, and $N = 8$. The estimated statistical power was 0.27, 0.39 for $i = 1$ and 2, respectively. Thus, $i = 2$ was chosen.

Next, $\tilde{P}(a_k < \delta_{ij}|n \in \Omega)$ is estimated (Section 2.4.3). The 2nd order statistic was 0.65 ($\delta_{ij} = 0.65$). With the assumption of the binomial distribution of D-Acc, $\tilde{P}(a_k < \delta_{ij}|n \in \Omega)$ was estimated as 0.9999.

Finally, $\tilde{L}$ is calculated and compared with $a$ (Section 2.4.4). Substituting the estimation into Eq. (2.5) and Eq. (2.8), we found $L < 0.05$. Therefore, $\tilde{L}$ was lower than the statistical threshold $a = 0.05$ and the i-test reported a significant result.

i-test-one: The lowest observed D-Acc ($\tilde{a}_{(1)} = 0.52$) was used for the i-test-one. With the assumption of the D-Acc binomial distribution, $P(a_k < 0.52|n \in \Omega)$ was estimated as 0.62. Substituting this into Eq. (2.5) and Eq. (2.14), $L < 0.052$ was obtained, which is higher than $a$. Thus, the i-test-one did not report a significant result.

Although the true $\gamma$ was unknown, the results may suggest the statistical power improvement of the i-test compared with the i-test-one. In the experiment, 7 of the 8 participants had D-Acc above the upper 1% point (0.62) of the D-Acc distribution for $\Omega$, i.e., BPDF(100a, 0.05). Under this condition, i-test reported a significant result, while the i-test-one did not. This may be in agreement with the results of the previous section demonstrating the statistical power improvement by extending the i-test-one to the i-test.

5. Discussion

I proposed a novel second-level group statistical test for decoding accuracy, named “i-test.” The i-test targets the population information prevalence, i.e., the proportion of the population that has higher-than-chance D-Acc. Therefore, a significant result of the i-test can infer the typical effect that the brain activation contains label information in a proportion larger than $\gamma_0$ of the population. This was in contrast to the fact that a significant result of the statistical test on the population mean of the D-Acc, such as t-test, can only infer that “there are some people in the population whose fMRI data carry information about the experimental condition” (Allefeld et al., 2016).

i-test is an extension of the test for the population prevalence proposed by Allefeld et al., 2016 (i-test-one), in which the minimum statistic was necessarily used. By using a higher-order statistic, i-test can improve the statistical power (Section 3).

5.1. Limitation of the statistical power evaluation in the current study

My numerical evaluation of the statistical power (Section 3) was limited to the situation when the true D-Acc follows the binomial distribution and we can correctly estimate the distribution. In real situations, the D-Acc does not always follow the binomial distribution because of the lack of independence among trials and we cannot know the exact distribution. Particularly, when the D-Acc was evaluated with cross-validation, the distribution is known to differ from the binomial distribution (Combrinon and Jerbi, 2015, Supplementary Material S1.2.2.2.1). In Supplementary Material S1.2, I present a fragmental simulation evidence that the i-test provided a high statistical power for the non-binomial cross-validated D-Acc (CV-D-Acc). But fairly speaking, the D-Acc distribution of empirical neuroscience experiments can be more different from the binomial distribution than the simulated CV-D-Acc, because the dependence and difference between trials may arise not only due to cross-validation procedure, but also due to empirical issues in the experiment, such as drift of participants’ arousal level.

An important issue is that the i-test could not perfectly control the false alarm rate when the estimated D-Acc distribution did not match the true distribution (Supplementary Material S1.2.2.3). Further theoretical studies and accumulation of empirical evidence are needed for evaluation of the i-test’s robustness to violations of the distribution assumptions in real situations.

5.2. Implementation of i-test

In the present study, a mathematically simple implementation of the i-test with the assumption of the binomial distribution for D-Acc (i-test-unif-bino) was proposed. The computational merit was that such D-Acc distribution is fully determined by only two unknown parameters, $\gamma$ and...
$P_{correct}$. However, as mentioned above, D-Acc does not always follow the binomial distribution in empirical situations. Following this concern, another possible variation is to use a nonparametric empirical estimation of the distribution, e.g., estimation using permutation test for $\tilde{p}(a_n|n \in \Omega_\gamma)$ as proposed by Allefeld et al., 2016 for the $t$-test-one. As an example of such implementations, a permutation-based $t$-test ($t$-test-unif-perm) is introduced in Supplementary Material S3.1.

The other assumption used in the $t$-test-unif-bino is the noninformative (uniform) prior distributions of $\gamma$ and $P_{correct}$ that does not require any prior knowledge. Another idea is to use experimental results to estimate prior distribution of parameters $\gamma$ and $P_{correct}$, which could lead selection of better optimized $i$. From this viewpoint I propose another implementation of the $t$-test, in which a maximum likelihood point estimates of $\gamma$ and $P_{correct}$ are used ($t$-test-ml-bino; Supplementary Material S3.2).

Further studies are needed to evaluate the efficacy, including the robustness discussed in Section 5.1, of these alternative implementations, as well as other possibilities, e.g., $t$-test-ml-perm.

Conclusion

A novel second-level group statistical test for decoding accuracy named “$t$-test” was proposed. I provided the theoretical derivation, the empirical implementation, and evidence that this test can improve the statistical power compared with the $t$-test-one proposed by Allefeld et al., 2016. The advantages of the $t$-test comprise a mathematically guaranteed and meaningful population inference, and a high statistical power. Although the $t$-test was introduced as a statistical method for the D-Acc, it is also applicable to other “information-like” measures, including continuous variables, such as Mahalanobis distance (Kriegeskorte et al., 2006, Nili et al., 2014), linear discriminant $t$ (Nili et al., 2014), or pattern distinctness $D$ (Allefeld and Haynes, 2014). This study may provide a robust, second-level group statistical test for information-based neuroimaging studies.

Data and code availability

All analyses in this study can be replicated by MATLAB program codes available at https://github.com/satoshi-hirose/i-test/releases.

Declaration of Competing Interest

None

Credit authorship contribution statement

Satoshi Hirose: Conceptualization, Methodology, Software, Resources, Writing – original draft, Visualization, Funding acquisition.

Acknowledgments

I would like to acknowledge Dr. Atsushi Yokoi, Center for Information and Neural Networks, Advanced ICT Research Institute, National Institute of Information and Communications Technology, Dr. Isao Nambu, Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Dr. Matthew de Brecth, Graduate School of Human and Environmental Studies, Kyoto University, and Dr. Carsten Allefeld, Department of Psychology, City, University of London, for helpful suggestions. This work was supported by a Grant-in-Aid for Young Scientists (B) (JSPS KAKENHI 16K16649) to the author.

Data and code availability statement

The participants of the experiment (Section 4) were explicitly informed that “the neural data measured in the experiment will not be released in any form” when they provided informed consent. Therefore, it is impossible to publish the neuronal data.

All analyses in this study can be replicated by MATLAB program codes available at https://github.com/satoshi-hirose/i-test/releases, which include the result of the decoding analysis (decoding accuracy) for each participant obtained from the experiment. I think this is sufficient because this study focuses on the second-level group statistical test using the decoding accuracies of the individual participants.

Appendix A: Numerical calculation of the probability that the $t$-test reports a significant result

I describe the calculation of $P$(Significant), which is the probability that the $t$-test reports a significant result, given the predetermined parameters ($\gamma_0$, $\alpha$, and $i$), number of participants ($N$), the estimated distribution of the D-Acc without label information ($\tilde{p}(a_n|n \in \Omega_\gamma)$), and the true distribution ($p(a_n|n \in \Omega_\gamma)$, $\tilde{p}(a_n|n \in \Omega_\gamma)$, and $\gamma$). The outline of the calculation is as follows. First, a threshold, $T$, of the $i$-th order statistic is introduced such that the $t$-test reports a significant result when and only when $a_{ij} > T$ (Section A.1). Then, $(P(a_{ij} > T)$ is derived (Section A.2).

A.1. The threshold of the $i$-th order statistic, $T$

From Eqs. (2.5) and (2.8), the $t$-test calculation is formulated as follows:

$$Q \equiv (1 - \gamma_0)P(a_n < a_{ij}|n \in \Omega_\gamma)$$ (A.1)

$$L \equiv BCDF(i - 1, N, Q)$$ (A.2)

$$P(a_n < a_{ij}|n \in \Omega_\gamma)$$ obviously increases with $a_{ij}$. Also, $Q$ increases with $P(a_n < a_{ij}|n \in \Omega_\gamma)$ (Eq. (A.1)) and $L$ decreases with $Q$ (Eq. (A.2)). Therefore, $L$ monotonically decreases with $a_{ij}$. Because of the monotonic dependence of $L$ on $a_{ij}$, we can define the threshold $T$, such that $a > L$ when and only when $a_{ij} > T$. Because $a_{ij}$ is a discrete variable, one may easily calculate $L$ numerically, by calculating $L$ for every possible $a_{ij}$ using Eqs. (A.1) and (A.2) and defining $T$ as the largest value of $a_{ij}$ that satisfies $a \leq L$ (Figure A.1).

A.2. Calculation of the probability that $i$-th order statistic is higher than $T$

The probability of a significant result is defined as

$$P$$(Significant) $\equiv P(a > L)$ (A.3)

From the equivalence of $a > L$ and $a_{ij} > T$, it follows that

$$P$$(Significant) $= P(a_{ij} > T)$ (A.4)

To calculate $P(a_{ij} > T)$, first, the probability that the D-Acc of a participant does not exceed $T$ is formulated as follows:

$$P(a_{ij} \leq T) = (1 - \gamma)P(a_n \leq T|n \in \Omega_\gamma) + \gamma P(a_n \leq T|n \in \Omega_\gamma)$$ (A.5)

Then, by accounting that the condition $i$-th order statistic being higher than $T$ is identical to the condition fewer than $i$ of the $N$ participants having D-Acc not exceeding $T$, we can formulate $P(a_{ij} > T)$ as follows:

$$P(a_{ij} > T) = BCDF(i - 1, N, P(a_n \leq T))$$ (A.6)

Therefore, $P$(Significant) is derived as follows:

$$P$$(Significant) $= BCDF(i - 1, N, P(a_n \leq T))$ (A.7)

In summary, $P$$(Significant)$ is numerically calculated with the following steps: $T$ is calculated from $\tilde{p}(a_n|n \in \Omega_\gamma)$, $\gamma_0$, $\alpha$, $i$, and $N$. Then, $P(a_n \leq T)$ is calculated with Eq. (A.5), the value of $T$, and the true distribution ($p(a_n|n \in \Omega_\gamma)$, $\tilde{p}(a_n|n \in \Omega_\gamma)$, and $\gamma$). Finally, $P$(Significant) is calculated with Eq. (A.7).
A.3. Numerical calculation example

I present the numerical calculation of the example with the following situation. The parameters were set as $N = 50$, $\alpha = 0.05$, $\gamma_0 = 0.5$ and $i = 1$. $p(a_i|n \in \Omega_1)$ and $p(a_i|n \in \Omega_2)$ were the binomial distribution, i.e., $p(a_i|n \in \Omega_1) = BPDF(a_i, N_{\text{trial}}, P_{\text{correct}})$ and $p(a_i|n \in \Omega_2) = BPDF(a_i, N_{\text{trial}}, N_{\text{trial}}, P_{\text{correct}})$, where $N_{\text{trial}} = 1000$, $P_{\text{correct}} = 0.5$ and $P_{\text{correct}} = 0.9$. The horizontal red dotted line indicates the statistical threshold, $\alpha = 0.05$. B) Enlarged view of the region indicated by the red square in Panel A. Following this figure, it can be easily found that $L$ is smaller than $\alpha$ when and only when $a_{ij}$ is higher than 0.481 ($T = 0.481$).

Fig. A1. Calculation of $T$ for $\alpha = 0.05$, $\gamma_0 = 0.5$, $N = 50$, $i = 1$, $N_{\text{trial}} = 1000$, $P_{\text{correct}} = 0.9$, and $\gamma = 0.8$. A) $L$ is plotted against $a_{ij}$. Note that $a_{ij}$ can take discrete values from 0 to 1 with 0.001 ($1/N_{\text{trial}}$) increment. Each black circle corresponds to the value of $a_{ij}$. The horizontal red dotted line indicates the statistical threshold, $\alpha = 0.05$. B) Enlarged view of the region indicated by the red square in Panel A. Following this figure, it can be easily found that $L$ is smaller than $\alpha$ when and only when $a_{ij}$ is higher than 0.481 ($T = 0.481$).

Appendix B: Selection of optimal $i$

For the selection of parameter $i$, the optimal value ($i_{\text{opt}}$) that maximizes the expected statistical power for a given $i$ ($\text{Power}(i)$) is determined.

B.1. General formulation

For the calculation of $\text{Power}(i)$, first, the set of parameters of the true distribution is defined as $\Theta$, such that the true distributions of D-Acc both with and without label information ($p(a_i|n \in \Omega_1)$ and $p(a_i|n \in \Omega_2)$) are uniquely determined at fixed $\Theta$. Consequently, the true distribution can be expressed as

$$p(a_i|\gamma, \Theta) = (1-\gamma)p(a_i|n \in \Omega_2, \Theta) + \gamma p(a_i|n \in \Omega_1, \Theta)$$

(B.1)

Then, the expected statistical power ($\text{Power}(i)$) is defined as

$$\text{Power}(i) = \int_{\gamma = \gamma_0}^{1} \int_{\Theta} P(\text{Significant}|i, \gamma, \Theta) \text{d}\Theta \text{d}\gamma$$

(B.2)

Here, $\Theta$ is the parameter space of $\Theta$, and $\gamma, \Theta$ is the joint prior distribution. $P(\text{Significant}|i, \gamma, \Theta)$ can be numerically calculated with the procedure proposed in Appendix A. The integration range for $\gamma$ is $(\gamma_0, 1)$ when the alternative hypothesis is true.

Then, $i_{\text{opt}}$ is defined as the value of $i$ that maximizes the expected statistical power;

$$i_{\text{opt}} = \arg \max_i \text{Power}(i)$$

(B.3)

B.2. Calculation of $i_{\text{opt}}$ for the i-test-unif-bino ($i_{\text{unif-bino}}$)

In the implementation of i-test-unif-bino, the binomial distribution of D-Acc is assumed, so that $\Theta$ includes one unknown parameter ($P_{\text{correct}}$) and all other parameters determining the distribution ($P_{\text{correct}}$ and $N_{\text{trial}}$) are known. As for the prior distribution $p(\gamma, \Theta)$, the noninformative prior (uniform distribution) is used. Under these assumptions, Eq. (B.2) can be expressed as

$$\text{Power}(i) = \int_{\gamma = \gamma_0}^{1} \int_{P_{\text{correct}} = P_{\text{correct}}}^{1} P(\text{Significant}|i, P_{\text{correct}}+\gamma) \text{d}P_{\text{correct}} \text{d}\gamma$$

(B.4)

The integration range of $P_{\text{correct}}$, is $[P_{\text{correct}}, 1]$ by definition of $P_{\text{correct}}$. Substituting Eq. (B.4) into Eq. (B.3) and ignoring constant terms, Eq. (B.3) transforms to

$$i_{\text{unif-bino}} = \arg \max_i \int_{\gamma = \gamma_0}^{1} \int_{P_{\text{correct}} = P_{\text{correct}}}^{1} P(\text{Significant}|i, P_{\text{correct}}+\gamma) \text{d}P_{\text{correct}} \text{d}\gamma$$

(B.5)
Appendix C: Implementation of the i-test-unif-bino

Figure C.1 shows the pseudocode of the i-test-unif-bino implementation.

C.1. Inputs

The i-test-unif-bino requires the following 6 inputs: the experimental result \( \{ a_i \} \), the number of trials \( N_{\text{trial}} \), the chance-level D-Acc \( P_{\text{correct}} \), the prevalence threshold \( \gamma_0 \), the significance threshold \( \alpha \), and the precision parameter \( h \).

C.2. Identifying \( i_{\text{unif-bino}} \)

First, the largest possible value of \( i \) (\( i_{\text{max}} \)) is found with Eq. (2.10) (Figure C.1 Lines 1-2). Then, the expectation of the statistical power for each \( i \) (\( \text{Power}(i) \)) is calculated (Lines 3-10) to identify \( i_{\text{unif-bino}} \) that provides the highest \( \text{Power}(i) \) (Line 11).

The calculation of \( \text{Power}(i) \) is performed with the following two steps. The first step is the calculation of \( T \) (Lines 4-6; Appendix A.1). This was made by performing the \( i \)-test (Eqs. (A.1) and (A.2)) for each possible value of order statistic with the assumption of the binomial distribution for \( \hat{p}(a_i | n \in \Omega) \) (Assumption 4, see below) and identifying the largest value that does not lead significant result. The second step is the calculation of the marginal statistical power (Lines 7-9; Appendix A.2 and Appendix B.2). Assuming that \( P_{\text{correct}} \) and \( \gamma \) are uniformly distributed (Assumption 1), it is approximated with the sum of the probability that \( i \)-test reports the significant result for each combination of \( P_{\text{correct}} \) and \( \gamma \) (Eq. (B.6); Line 7). The probability for each combination of \( P_{\text{correct}} \) and \( \gamma \) is calculated with Eqs. (A.5) (Line 8), (A.7) (Line 9), and calculated value of \( T \). The binomial distribution for true D-Acc distribution with and without label information is assumed in the calculation (Assumption 2 and 3).

C.3. Performing \( i \)-test

Finally, the \( i \)-test with \( i_{\text{unif-bino}} \) was applied to the experimental results (Eqs. (A.1) and (A.2), Lines 12-15) with the assumption of the binomial distribution for the estimation of D-Acc without label information (Assumption 4).

C.4. Assumptions

The four assumptions for the above implementation are listed below:

Assumption 1. \( p(P_{\text{correct}}, \gamma) \) is the uniform distribution (Line 7)

Assumption 2. \( p(a_i | n \in \Omega) = \text{BPDR}(N_{\text{trial}}, a_i, N_{\text{trial}}, P_{\text{correct}}) \) (Line 8)

Assumption 3. \( p(a_i | n \in \Omega) = \text{BPDR}(N_{\text{trial}}, a_i, N_{\text{trial}}, P_{\text{correct}}) \) (Line 8)

Assumption 4 \( p(a_i | n \in \Omega) = \text{BPDR}(N_{\text{trial}}, a_i, N_{\text{trial}}, P_{\text{correct}}) \) (Lines 5 and 13)

By using different assumptions, one may implement different versions of the \( i \)-test. Particularly, the following two implementations are introduced in Supplementary Material S3. In the i-test-unif-perm (Supplementary Material S3.1), empirical estimation procedure is used for the Assumptions 3 and 4, i.e. they are replaced with an empirically estimated distribution using the permutation test. In the i-test-mf-bino (Supplementary Material S3.2), Assumption 1 is replaced with the data-driven point maximum likelihood estimate.
C.5. Error handling

The $t$-test-unif-bino yields an error when there is no $i$ that satisfies

$BCDF(\frac{i - 1}{2}, N, 1 - \gamma_0) < \alpha$ (Line 2), indicating that the $t$-test cannot be

applied to the inputs because of the constraint imposed by Eq. (2.10).

Another error can occur when there is no $T$ that satisfies

$L \leq \alpha$ with particular values of $i$ (Line 4). This indicates that the $t$-test never reports

a significant result with the value of $i$. In this case, such values should

be omitted in the search of $t_{unif-bino}$.

Supplementary materials

Supplementary material associated with this article can be found, in

the online version, at doi:10.1016/j.neuroimage.2021.118456.

References

Allefeld, C., Gieger, K., Haynes, J.D., 2016. Valid population inference for information-based imaging: From the second-level $t$-test to prevalence inference. Neuroimage 141, 378–392. doi:10.1016/j.neuroimage.2016.07.040.

Allefeld, C., Haynes, J.D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage 89, 345–357. doi:10.1016/j.neuroimage.2013.11.043.

Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences. Academic press.

Combrinck, E., Jerbi, K., 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods 250, 126–136.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S., Turner, R., 1995. Analysis of fMRI time-series revisited. Neuroimage 2, 45–53. doi:10.1006/nimg.1995.1023.

Friston, K.J., Holmes, A.P., Price, C.J., Büchel, C., Worsley, K.J., 1999. Multisubject fMRI studies and conjunction analyses. Neuroimage 10, 385–396. doi:10.1006/nimg.1999.0484.

Gallivan, J.P., McLean, D.A., Valyear, K.F., Pettypiece, C.E., Culham, J.C., 2011. Decoding action intentions from preparatory brain activity in human parieto-frontal networks. J. Neurosci. 31, 9599–9610. doi:10.1523/JNEUROSCI.0080-11.2011.

Gallivan, J.P., McLean, D.A., Flanagan, J.R., Culham, J.C., 2013. Where one hand meets the other: limb-specific and action-dependent movement plans decoded from preparatory signals in single human frontoparietal brain areas. J. Neurosci. 33, 1991–2008. doi:10.1523/JNEUROSCI.0541-12.2013.

Gilbert, S.J., Fung, H., 2018. Decoding intentions of self and others from fMRI activity patterns. Neuroimage 172, 278–290. doi:10.1016/j.neuroimage.2017.12.090.

Good, P., 2000. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer-Verlag, New York.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456. doi:10.1146/annurev-neuro-062012-170325.

Hirose, S., Nambu, I., Naito, E., 2015. An empirical solution for over-pruning with a novel ensemble-learning method for fMRI decoding. J. Neurosci. Methods. 239, 238–245. doi:10.1016/j.jneumeth.2014.10.023.

Hirose, S., Nambu, I., Naito, E., 2018. Cortical activation associated with motor preparation can be used to predict the freely chosen effector of an upcoming movement and reflects response time: An fMRI decoding study. Neuroimage 183, 584–596. doi:10.1016/j.neuroimage.2018.08.060.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103, 3863–3868. doi:10.1073/pnas.0600244103.

Nambu, I., Hagara, N., Hirose, S., Wada, Y., Kawato, M., Naito, E., 2015. Decoding sequential finger movements from preparatory activity in higher-order motor regions: a functional magnetic resonance imaging multi-voxel pattern analysis. Eur. J. Neurosci. 42, 2851–2859. doi:10.1111/ejn.13063.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553. doi:10.1371/journal.pcbi.1003553.

Nishida, S., Nishimoto, S., 2018. Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage 180, 232–242. doi:10.1016/j.neuroimage.2017.08.017.

Rosenblatt, J.D., Vink, M., Benjamini, Y., 2014. Revisiting multi-subject random effects in fMRI: advocating prevalence estimation. Neuroimage 84, 113–121. doi:10.1016/j.neuroimage.2013.08.025.

Spirdon, M., Kanwisher, N., 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron 35, 1157–1165. doi:10.1016/S0896-6273(02)00877-2.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage 65, 69–82. doi:10.1016/j.neuroimage.2012.09.063.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017. doi:10.1016/j.neuroimage.2009.03.025.

Yamashita, O, Sato, MA, Yoshikawa, T, Tong, F, Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. Neuroimage 42, 1414–1429. doi:10.1016/j.neuroimage.2008.05.050.