METHOD ARTICLE

REVISED A new statistical method to analyze Morris Water Maze data using Dirichlet distribution [version 2; peer review: 2 approved]

Marianne Maugard[1,2*], Cyrille Doux[3,4*], Gilles Bonvento[1,2]

[1]Commissariat à l'Energie Atomique et aux Energies Alternatives, Département de la Recherche Fondamentale, Institut de Biologie François Jacob, Fontenay-aux-Roses, 92260, France
[2]Neurodegenerative Disease Laboratory, Centre National de la Recherche Scientifique, Université Paris-Sud, Université Paris-Saclay, Fontenay-aux-Roses, 92260, France
[3]AstroParticule et Cosmologie, Université Paris Diderot, Paris, 75205, France
[4]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, 19104, USA

[*] Equal contributors

v2 **First published:** 06 Sep 2019, **8**:1601 (https://doi.org/10.12688/f1000research.20072.1)
Latest published: 28 Oct 2019, **8**:1601 (https://doi.org/10.12688/f1000research.20072.2)

Abstract
The Morris Water Maze (MWM) is a behavioral test widely used in the field of neuroscience to evaluate spatial learning memory of rodents. However, the interpretation of results is often impaired by the common use of statistical tests based on independence and normal distributions that do not reflect basic properties of the test data, such as the constant-sum constraint. In this work, we propose to analyze MWM data with the Dirichlet distribution, which describes constant-sum data with minimal hypotheses, and we introduce a statistical test based on uniformity (equal amount of time spent in each quadrant of the maze) that evaluates memory impairments. We demonstrate that this test better represents MWM data and show its efficiency on simulated as well as *in vivo* data. Based on Dirichlet distribution, we also propose a new way to plot MWM data, showing mean values and inter-individual variability at the same time, on an easily interpretable chart. Finally, we conclude with a perspective on using Bayesian analysis for MWM data.

Keywords
Morris Water Maze, Statistical analysis, Dirichlet distribution

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 28 Oct 2019		✓ report
version 1 published 06 Sep 2019	✓ report	? report

1 **Richard GM Morris**, University of Edinburgh, Edinburgh, UK

2 **Avgoustinos Vouros**, University of Sheffield , Sheffield, UK
Luca Manneschi, University of Sheffield, Sheffield, UK
Eleni Vasilaki, University of Sheffield, Sheffield, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Gilles Bonvento (gilles.bonvento@cea.fr)

Author roles: Maugard M: Conceptualization, Data Curation, Formal Analysis, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Doux C**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Bonvento G**: Conceptualization, Funding Acquisition, Project Administration, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was part of a project supported by Association France Alzheimer and Fondation de France (Prix Spécial 2012 to G.B. and collaborators) and Fondation Plan Alzheimer (G.B.).

How to cite this article: Maugard M, Doux C and Bonvento G. **A new statistical method to analyze Morris Water Maze data using Dirichlet distribution [version 2; peer review: 2 approved]** F1000Research 2019, **8**:1601 (https://doi.org/10.12688/f1000research.20072.2)

First published: 06 Sep 2019, **8**:1601 (https://doi.org/10.12688/f1000research.20072.1)

REVISED **Amendments from Version 1**

In this revised version, we have (1) clarified why there is no alternative to using the t-test and indicated (2) that our statistical analysis can be used each time a probe test is performed with the sum spent in the 4 quadrants totalling 100%, (including reversal learning for example).

Any further responses from the reviewers can be found at the end of the article

1 Introduction

The Morris Water Maze (MWM) was first described by Richard Morris in the 80's 1 and is still one of the most commonly used tasks to evaluate spatial learning in rodents, including normal and genetically modified mice. While the standard reference memory task is mostly used and is validated as an assay for hippocampus-dependent spatial navigation and reference memory, modifications of the basic protocol allow to also evaluate reversal learning, the delayed match to place task and procedures for dissociating encoding and retrieval. At least for the standard reference memory and reversal learning tasks, these procedures require probe test data that display a constant-sum constraint. The maze consists of a large circular tank filled with opaque water in which rodents can escape onto a platform hidden just beneath the surface. During a training phase animals perform repeated blocks of 60 second-long trials to find the location of a fixed platform using distant visual cues from semi-random start locations and the escape latency is recorded. Since data are right-truncated at 60 seconds, in contradiction with a normal distribution and causing potentially biased results, statistical guidelines have been published to properly characterize learning behaviors using survival data 2.

During a probe test session, the platform is removed and animals freely navigate into the pool from the same start location and for the same fixed amount of time (*e.g.* 60 seconds). The path of the animal is recorded using a video camera and an automatic tracking system. Data collected during the probe test session can be classified into three categories: time spent per zone, which can be theoretical quadrants defined on the pool or a theoretical annulus drawn around the platform location; number of crossings of the platform area; or total proximity to the platform center 3. Creating a large database using several published tests from their institute and simulated data, Maei et al. have shown that total proximity allows the best detection for small samples, whereas time spent in quadrants is of great interest for bad performers 4. Since this test is often used to characterize memory loss, time spent in theoretical quadrants is mostly found in the literature.

Several hypotheses can be tested using data obtained from time spent in quadrants: 'Can one group of rodent remember the platform location?' or 'Is there any difference of memory abilities between several groups of rodents?'. In both cases, the statistical analysis of the data often focuses on the target quadrant (*e.g.* that where the platform was placed during the learning phase) using parametric tests like ANOVAs and *t*-tests[1]. These tests are based on normal distributions and independence, which cannot be accurately assumed in this context since 1) variables are defined on a finite interval and 2) variables corresponding to the time spent in the four quadrants are necessarily anti-correlated. Moreover, these tests neglect the time spent in the three other quadrants inducing a loss of information and hiding the aspect of preference for one quadrant that is supposed to reflect efficient spatial memory. Some authors used non-parametric alternatives, but even if their use may be preferable with the sample size of behavioral studies, they still do not fully describe the experiment. These observations suggest that a better characterization and a more suitable statistical analysis of data obtained through the MWM could significantly improve the accuracy of the results.

Focusing on the question 'Can one group of rodents remember the platform location?' to evaluate memory abilities, we suggest to use the Dirichlet distribution, a distribution that describes several variables with a constant sum, to collectively describe the fraction of time spent in the four quadrants of the maze. This test would provide a unique *p*-value allowing determination of whether the rodents spent the same amount of time in the four quadrants or not, a primary indication of significant spatial memory. In the case of differences between the four quadrants, this test can be followed by four *post-hoc* Student *t*-tests to identify preference or aversion for some quadrants. In comparison, the currently used method (*i.e.* directly using the Student *t*-test on the target quadrant) does not allow to identify memory loss and may hide some bias (Figure 1).

We will first describe the methodology we developed with the Dirichlet distribution and the correction required to fit with the sample size of behavioral experiments. Using simulated data we will show that beyond the better description of the results, using Dirichlet distribution allows reducing the number of false positives and false negatives, significantly improving the reliability of the analysis. We then applied this test on *in vivo* data to validate its use in experimental conditions, also providing a way to graphically present the data that takes into account interindividual variability. Finally, we will discuss the advantages and limits of the application of the Dirichlet distribution on behavioral studies, broaching the major inputs that using Bayesian inferences could bring in this field of research.

2 Methods

2.1 The Dirichlet distribution

The Dirichlet distribution is a multivariate generalization of beta distributions. It describes the distribution of K-dimensional vectors p for which the sum of all the coordinates is fixed, *i.e.* $\sum_{k=1}^{K} p_k = 1$. It is parametrized by a K-dimensional vector $\boldsymbol{\alpha}$

[1]Among the 30 most recent articles using the Morris Water Maze test and published at the end of October 2018, 25 used time spent in quadrants as a criterion. 23 used a parametric test (Student *t*-test or ANOVA) to analyze their data, whereas only 2 used a non parametric alternative. 24 out of 25 articles presented data as bar charts without presenting inter-individual variability.

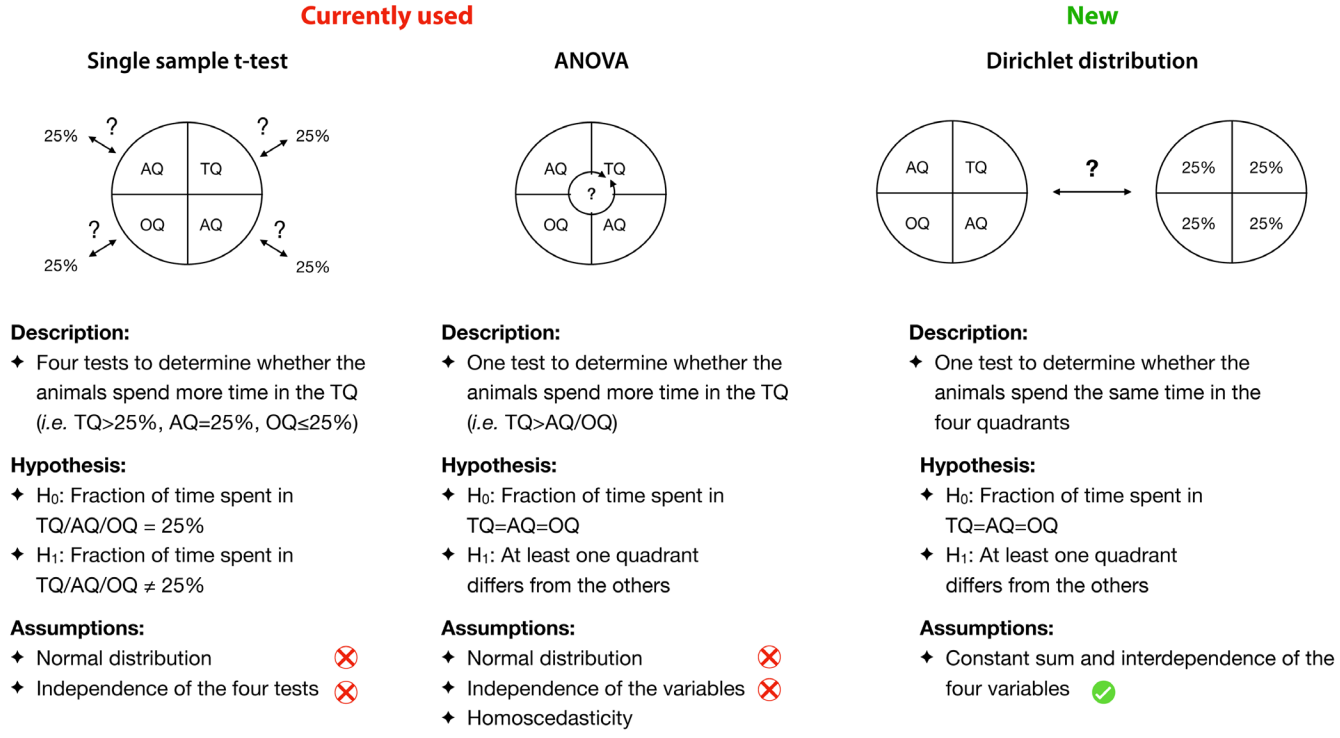

Figure 1. Summary of the analyses of the MWM that are currently used in comparison with Dirichlet distribution. To determine whether one group of animals has a preference for a quadrant one usually uses *t*-tests or ANOVAs assuming normal distribution and independence. Those assumptions do not describe correctly the dataset obtained from a MWM. In comparison, the Dirichlet distribution allows to answer the same question but describes properly the constant sum constraints and interedependence of the variables.

of positive reals $\alpha_k > 0$, $1 \le k \le K$, such that its probability density function is given by

$$\mathscr{P}(\mathbf{p}\,|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} p_k^{\alpha_k - 1}, \qquad (1)$$

where $B(\boldsymbol{\alpha}) = \prod_{k=1}^{K}\Gamma(\alpha_k)/\Gamma(s)$ is the multivariate beta function and $s = \sum_{k=1}^{K}\alpha_k$ is the *precision*. The marginal distributions are beta distributions with parameters $(\alpha_k, s - \alpha_k)$ with expectation values $m_k = \alpha_k/s$, variance $m_k(1 - m_k)/(s + 1)$ and covariance between coordinates p_i and p_j given by $-m_i m_j/(s + 1)$. Therefore, the higher the precision s, the less diffuse coordinates are around their means. The Dirichlet distribution is the most general distribution for fixed-sum variables, motivating its use to describe compositional or fractional data such as MWM data, where $K = 4$. The likelihood of a sample of N independent observations $\mathrm{D} = \{\mathrm{p}_1, \ldots, \mathrm{p}_N\}$ is given by

$$\mathscr{P}(\mathbf{D}\,|\boldsymbol{\alpha}) = \prod_{i=1}^{N}\mathscr{P}(\mathbf{p}_i\,|\boldsymbol{\alpha}). \qquad (2)$$

2.2 Likelihood-ratio test based on the Dirichlet distribution

Description of the test To reflect memory abilities, we would like to test whether the fraction of time spent in the four quadrants significantly differs from a uniform distribution, thus showing preference for one or several quadrants. To do so, we propose a likelihood-ratio test based on the Dirichlet distribution to distinguish between the null hypothesis of uniformity $H_0 : \{\exists \alpha > 0, \forall\, k, \alpha_k = \alpha\}$ (implying that all means m_k are equal to $1/K$ but the precision is not constrained), and the general hypothesis H_1 where the α_k's are unconstrained. The likelihood-ratio statistic reads

$$\Lambda = 2 \times \left[\sup_{\alpha \in H_1} \ln \mathscr{P}(\mathbf{D}\,|\boldsymbol{\alpha}) - \sup_{\alpha \in H_0} \ln \mathscr{P}(\mathbf{D}\,|\boldsymbol{\alpha})\right]. \qquad (3)$$

In order to fit the distribution parameters to their maximum likelihood values, we refer to the numerical schemes developed in 5 and we used the open-source Python module `dirichlet` implemented by Eric Suh[2] and run with Python 3.6. In particular 5, proposes a technique to alternatively fit the means m_k and precision s, faster than fitting directly the α_k's. The maximum likelihood parameters under the null hypothesis are thus estimated by setting the means to their uniform value, $m_k = 1/K$, and fitting the precision s. We provide a slightly modified version of the `dirichlet` package, forked from that of Eric Suh, which is publicly available[3]. Under the null hypothesis,

[2]https://github.com/ericsuh/dirichlet

[3]https://github.com/xuod/dirichlet

the likelihood-ratio statistic Λ asymptotically follows a χ^2-distribution with $K-1$ degrees of freedom.

Bartlett correction Biological samples are usually limited and for small samples the statistic's distribution deviates from a χ^2_{K-1}, as can be seen in Figure 2. We propose an approximate Bartlett-type correction 6 for small samples, which amounts to rescale the likelihood-ratio statistic to match its asymptotic mean, which is $K-1$ in this test. Such a correction has been shown to correctly reproduce the first three moments of the asymptotic χ^2-distribution 6. In order to derive the scaling factor, we needed to compute the expected value of the statistic as a function of the number of samples N and the precision s. To do so, we drew random samples of N observations from uniform Dirichlet distributions with precision s, varying N between 2 and 100 and s between 1 and 1000 (with logarithmically-spaced values), and measured the mean of the statistic Λ[4]. We found that the mean value of the likelihood-ratio statistic Λ depends very little on s in the probed range, and that the difference to the asymptotic value of $K-1$ is well-fitted with a power law in N, *i.e.* $\langle\Lambda\rangle - (K-1) \sim a_K N^{b}_K$ (data not shown). We found the approximate values $a_K = 5.9$ and $b_K = -1.4$ for $K = 4$. We therefore propose to use a corrected statistic

$$\tilde{\Lambda} \equiv \Lambda \times \frac{(K-1)}{(K-1) + a_K N^{b_K}}, \tag{4}$$

and to compare its observed value to the χ^2 expected value corresponding to the desired statistical significance.

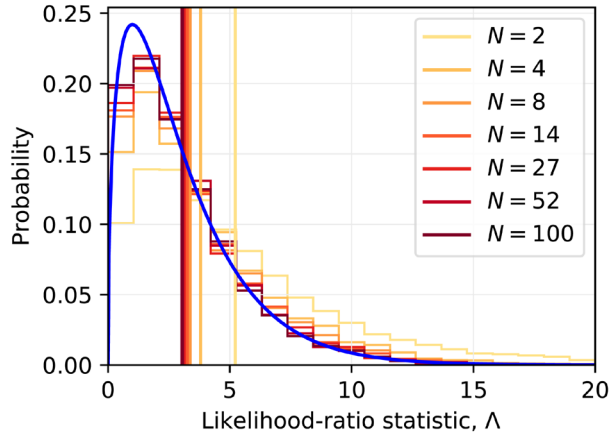

Figure 2. Histogram of the likelihood-ratio statistic Λ for s = 10 for various sample sizes N. Histograms of the likelihood-ratio statistic Λ are represented in different colors according to the N value. Means are represented by vertical lines of the same color. The χ^2_3-distribution is represented in blue. For small samples, the distribution of the likelihood ratio slightly deviates from a χ^2_3 distribution and the mean is significantly greater than the theoretical value of 3.

Validation of the test To validate this correction, we compared the distribution of the uncorrected statistic Λ and the corrected statistic $\tilde{\Lambda}$ from our simulated samples to a χ^2_{K-1}-distribution with probability-probability plots. As shown in Figure 3, the uncorrected statistic yields p-values significantly different from the theoretical ones while the corrected p-values are in

[4]For each tuple (N, s), the number of samples is increased until the means is measured with relative error below 0.01.

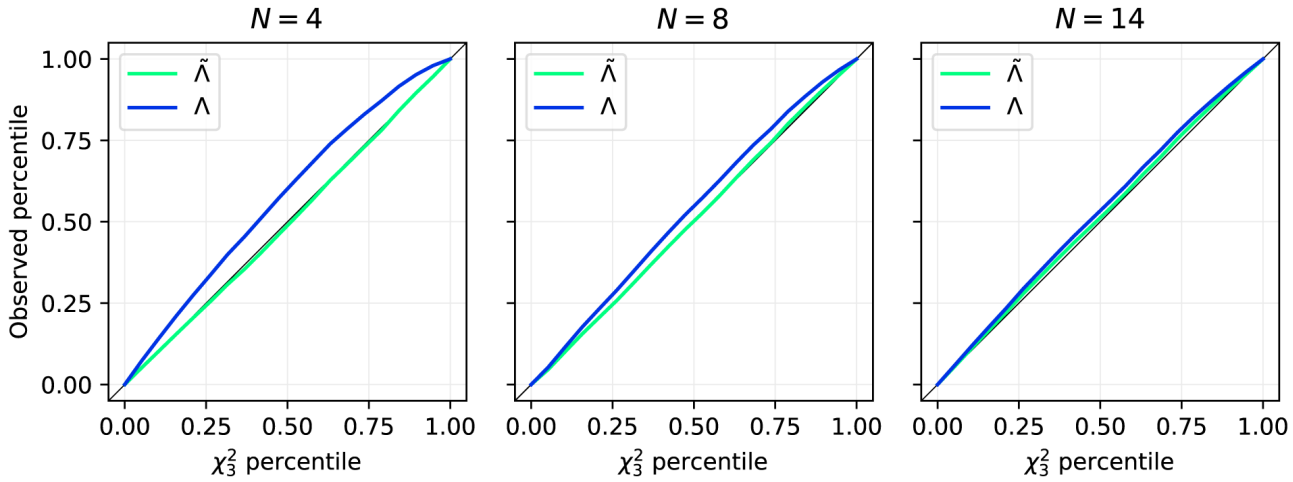

Figure 3. Probability-probability plots for uncorrected and corrected statistics from our simulated data. The uncorrected and corrected statistics are compared with a χ^2_{K-1} distribution for several sample sizes N. Grey lines represent equality between Λ percentiles and χ^2_{K-1} percentiles. Blue lines correspond to the uncorrected statistics and green lines to the corrected one. There is a difference between the p-values from the uncorrected statistics and the theoretical ones that disappears after correction, especially for small sample size.

perfect agreement with the χ^2_3-distribution. Therefore, this correction significantly improves the reliability of the test for small samples.

We also computed the number of false negatives on the simulated data to evaluate the rate of type 1 error (Figure 4). We found that using the *p*-value from the corrected statistic leads to a consistent rate of type 1 error (*i.e.* α =5%), independent on the number of samples *N* or the precision s. On the contrary, using the non-corrected statistic leads to more false negatives, especially when the number of samples is small.

Comparison with the one-sample Student *t*-test In order to compare the type 2 error obtained with the Dirichlet distribution with the results obtained using a one-sample *t*-test on the target quadrant as often done in the literature, we simulated data from a non-uniform Dirichlet distribution with the parameters $\boldsymbol{\alpha}$ = (40; 20; 20; 20). We found that the *p*-value from the corrected statistic of the Dirichlet distribution is mainly lower than the one obtained with a single *t*-test on the target quadrant (75% of the *p*-values are lower for s = 30). This means that for some cases where the target quadrant is preferred, using a one-sample *t*-test on the target quadrant would not detect this preference whereas the test based on Dirichlet distribution would detect the divergence from uniformity. Beyond improving the description and the interpretation of data from the MWM, the test we propose extracts more information from the same experiment as it is based on a larger dataset and then decreases the number of false-positives.

Post-hoc analysis Using the test based on Dirichlet distribution, we can determine whether the fraction of time spent in the quadrants is uniformly distributed. In the case of a divergence from uniformity, we would like to evaluate what are the quadrants responsible for this divergence as a *post-hoc* analysis.

This can be performed by comparing the marginal distributions of each quadrant, that are Beta distributions, to a theoretical Beta distribution with parameters $\alpha = 0.25s$ and $\beta = 0.75s$. The only simple way to compare one distribution with a theoretical one is to apply a single sample t-test that compares a normal distribution with a theoretical normal distribution. However, we noticed that in the range of inter-individual variability we have in this kind of study (given by the parameter s of the Dirichlet distribution, usually found between 20 and 50), the marginal distribution are fairly close to a normal distribution. Seeking for simplification, we advise to apply single sample *t*-tests for a *post-hoc* characterization of the preference for a quadrant in groups showing a divergence from uniformity.

2.3 Bayesian inference of the parameters of the Dirichlet distribution

Bayesian analysis can be used to infer constraints on the parameters α_i's of the Dirichlet distribution used to model the data (and subsequently the means m_i's). Specifically, we performed nested sampling of the parameter space of the Dirichlet distribution using the `PyMC3` package with Python 3.6. For simplicity, we used the Jeffreys prior[5] $\pi(\boldsymbol{\alpha})$ which does not depend on the model parametrization (*e.g.*, sampling over $\boldsymbol{\alpha}$ or ($\mathbf{m}$, s)) and leave the discussion about this choice for future work. The output is a sample of vectors $\boldsymbol{\alpha}$ distributed as the posterior given the data, *i.e.* $\mathscr{P}(\boldsymbol{\alpha} \mid \mathbf{D}) \propto \mathscr{P}(\mathbf{D} \mid \boldsymbol{\alpha})\pi(\boldsymbol{\alpha})$, which enables us to compare confidence regions for different groups and visualize the consistency with uniformity.

3 Results

We used a dataset obtained comparing memory abilities of female *wild-type* mice to female 3xTg AD mice, a model for

[5]For the K-dimensional Dirichlet distribution, the Fisher information is $I(\boldsymbol{\alpha}) = \mathrm{diag}(\Psi_1(\boldsymbol{\alpha})) - \Psi_1(s)J_4$, where J_4 is a $K \times K$ matrix of ones, and the Jeffreys prior is $\pi(\boldsymbol{\alpha}) \propto \sqrt{|I(\boldsymbol{\alpha})|}$.

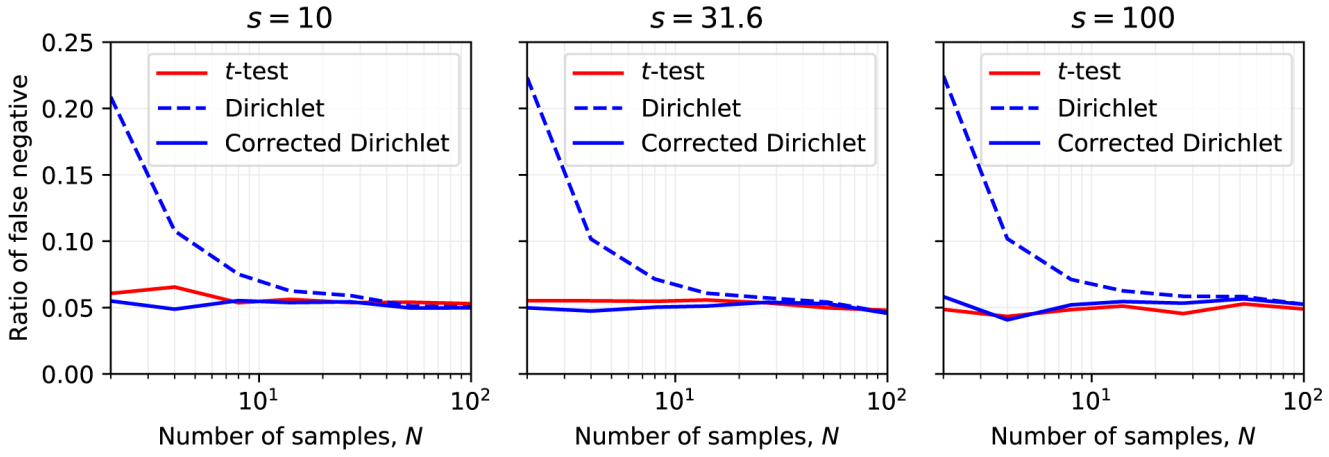

Figure 4. Rate of false negatives depending on the sample size *N* for different values of the precision *s* for a *p*-value of 0.05. The rate of false-negative (*i.e.* type 1 error) using the corrected version of the statistic is represented by the blue line whereas the rate of false-negative using the uncorrected version of the statistic is represented by the blue dotted line. We found that the correction results in a consistent rate of false negatives, independent of the sample size *N* and precision *s* in the range tested. The red line represents the rate of false-negative using a single-sample *t*-test on the target quadrant.

Alzheimer's Disease 7. All information related to experimental and ethical procedures are available in 8.

3.1 Application of the likelihood-ratio test based on the Dirichlet distribution

We compared the distribution obtained for each group to a uniform distribution in the probe test of the standard reference memory task and we found that the Dirichlet distribution obtained for *wild-type* mice was significantly different from a uniform distribution (p = 0.0021), whereas the one obtained for 3xTg mice did not differ from a uniform distribution (p = 0.26). We also propose a module, included in the Dirichlet package, to draw charts showing at the same time mean values with uncertainties[6] and inter-individual variability according to Dirichlet distribution (Figure 5). This result shows that 3xTg AD mice display long term memory deficits, which is in accordance with previous observations 9.

To better characterize long term memory in *wild-type* mice, we applied single sample t-tests on the four quadrants as a *post-hoc* analysis. We performed a one-tailed single sample t-test to assess whether the fraction of time spent in the target quadrant by *wild-type* mice is greater than the theoretical value 25%. Conversely, we performed a one-tailed single sample t-test to assess whether the fraction of time spent in the opposite quadrant by *wildtype* mice is lower than the theoretical value 25%. For adjacent quadrants we performed two-tailed single sample t-tests. We observed that the fraction of time spent in the target and opposite quadrants were respectively significantly higher (p = 0.025) and lower (p = 0.013) than 25%. The fraction of time spent in the adjacent quadrants did not differ from 25%.

Using this dataset with usual sample sizes for behavioral studies (N=7), we confirmed that our test is able to discriminate efficient and deficient memory abilities on real data.

3.2 Perspectives using Bayesian inference

We inferred constraints on the parameters of the Dirichlet distribution for *wild-type* and 3xTg mice. Figure 6 indicates compatibility of the data with uniformity for 3xTg mice and shows a clear preference for the target quadrant for *wild-type* mice suggesting memory deficits in 3xTg mice compared with *wild-type*.

Discussion

We proposed a statistical approach for the analysis of MWM probe test data based on the Dirichlet distribution as a model for the fraction of time spent by rodents in the quadrants of the maze. In the context of behavioral experiments that usually generate a lot of data with high inter-individual variability, a lot of parameters can be taken into account to extract evidence of memory abilities 3. In the literature, the time spent in quadrants – the target quadrant, but sometimes also the opposite quadrant 9,10 – is commonly used to assess long-term memory. Even if the focus on the time spent in quadrants is broadly accepted as a good

[6]We approximate the variance by the inverse Fisher information, which is a lower bound, given by $\hat{\sigma}^2(\alpha_i) = 1/(\psi_1(\alpha_i) - \psi_1(s))^n$ where ψ_1 is the trigamma function. A full bayesian analysis can be performed to obtain those error bars, as suggested in section 2.3.

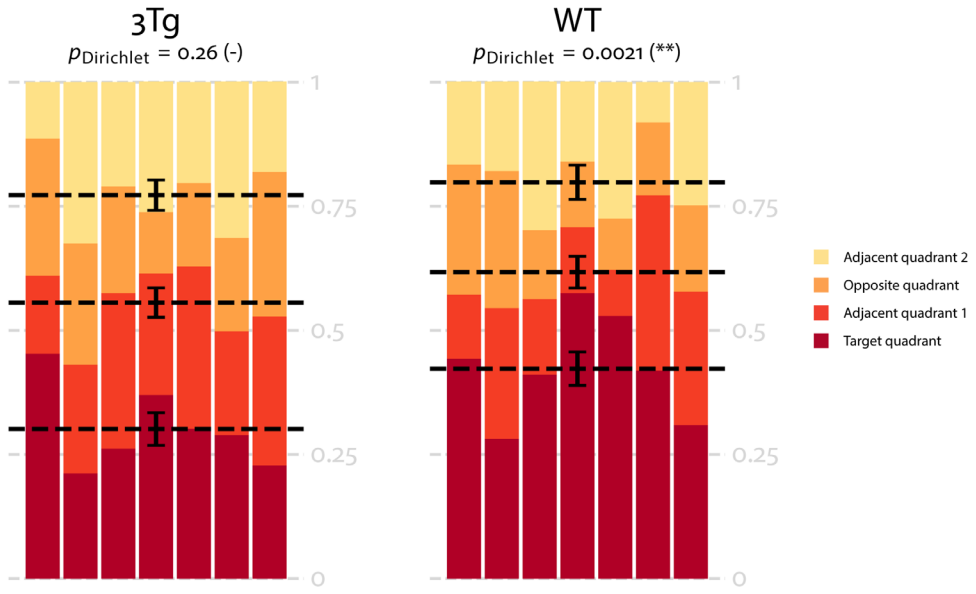

Figure 5. Time spent in the four quadrants by *wild-type* and 3xTg AD mice. In this plot, each column represents a sample and each color represents a quadrant. Mean values for the fraction of time spent in each quadrant is represented by a dotted line and the error bars on the means are approximated with the inverse Fisher information. For 3xTg mice the fraction of time spent in each quadrant is approximately similar leading to a uniform distribution (p = 0.26) whereas for *wild-type* mice the time spent in the target quadrant is significantly higher leading to a non-uniform distribution (p = 0.0021).

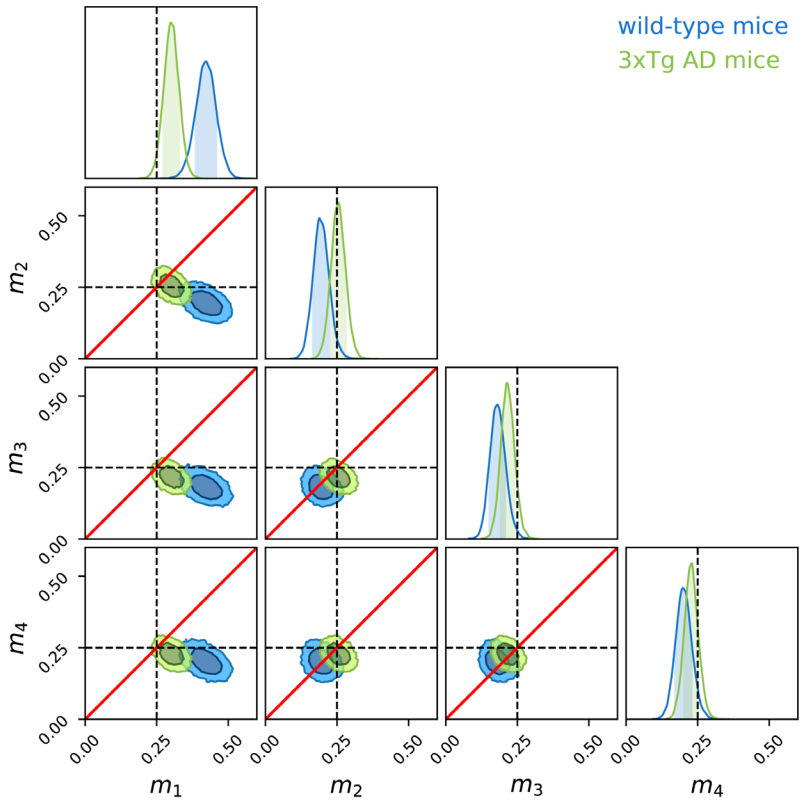

Figure 6. Fraction of time spent in the four quadrants for *wild-type* and 3xTg AD mice in the case of Bayesian inference. Corner plot representing constraints on the mean fractions of time m_i's for the two data sets *wild-type* (blue) and 3xTg (green). The diagonal plots show the marginal distributions of m_i's (with shaded 68% confidence interval) and off-diagonal plots show the two-dimensional distributions of pairs of these variables (inner and outer contours represent the 68% and 95% confidence levels). The black dashed lines represent the case of uniformity (25%) and the red lines correspond to equal time spent in both considered quadrant. Constraints on m_1 (leftmost column) indicate that *wild-type* mice favor the target quadrant.

index to evaluate reference memory, there is no consensus about the processing of these data. In this context, the Dirichlet distribution has the great advantage to simultaneously take into account the four quadrants and to correctly account for the constant-sum constraint of such data, which implies both deviation from the normal distribution and interdependence. That way, it gives a correct description of the data obtained from MWM probe tests and provides meaningful plots representing mean performances and inter-individual variability.

We showed that the corrected test based on the Dirichlet distribution gives a consistent rate of false-negative, even for small sample size. This indicates that this test can be safely used even in the context of behavioral studies with sample size smaller than 10 individuals, as we confirmed using the results previously obtained on *wild-type* and 3xTg AD mice.

Beyond the great improvement in the description of MWM probe test data, we also showed that this test gives less false-negatives than its inaccurate but commonly used alternative, the Student *t*-test. Therefore, using Dirichlet distribution is the best option to extract reliable information from time spent in quadrants during a MWM probe test. Combination of this analysis with results based on other measures of performance will give a comprehensive and accurate description of rodent memory abilities.

However, there are two main limitations in the use of the likelihood-ratio test based on the Dirichlet distribution: 1) it cannot directly identify the preferred quadrant and 2) it cannot compare memory abilities between several groups of animals. We proposed to overtake the first limitation by performing a *post-hoc* analysis to determine which quadrants are responsible for divergence from uniformity. We showed that performing single-sample *t*-tests, the only existing statistical test comparing one distribution with a theoretical one, as a *post-hoc* analysis (instead of *ad-hoc*) is satisfying. However, more interesting results can be obtained using Bayesian statistics, a method that can also permit comparison between groups. Deriving informative *p*-values on binary tests from such analysis remains challenging but represents an active field of research that could soon provide a great opportunity to improve MWM statistical analyses.

Conclusion

We propose here a new way to analyze MWM probe test data from the standard reference memory task of the MWM that accurately and simultaneously describes the four variables of time spent in the quadrants and allows to extract more information from the same experiments than the currently used method. All the packages required to perform the statistical test and to draw the corresponding chart are publicly available[7] and can be easily run with R using the `reticulate` package 11. Minor modifications of this test would allow to apply the same methodology on other behavioural tests facing the same constraints like H-Maze or Y-Maze.

Data availability

Underlying data

A Python notebook with the code to reproduce the simulations and figures is available at https://github.com/xuod/dirichlet 12.

In vivo dataset available from: https://github.com/xuod/dirichlet/tree/master/example 12. (Experimental procedures and data acquisition are detailed in 8).

Software availability

Dirichlet package from Eric Suh (Fitting the parameters of a Dirichlet distribution) available from: https://github.com/ericsuh/dirichlet

License: MIT

Dirichlet package used in the present study (Likelihood-ratio test based on Dirichlet distribution) available from: https://github.com/xuod/dirichlet/tree/master/dirichlet

Archived package as at time of publication: http://doi.org/10.5281/zenodo.3373955 12

License: MIT

[7]https://github.com/xuod/dirichlet

References

1. Morris R: **Developments of a water-maze procedure for studying spatial learning in the rat.** *J Neurosci Methods.* 1984; **11**(1): 47–60.
PubMed Abstract | Publisher Full Text
2. Jahn-Eimermacher A, Lasarzik I, Raber J: **Statistical analysis of latency outcomes in behavioral experiments.** *Behav Brain Res.* 2011; **221**(1): 271–275.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Vorhees CV, Williams MT: **Morris water maze: procedures for assessing spatial and related forms of learning and memory.** *Nat Protoc.* 2006; **1**(2): 848–858.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Maei HR, Zaslavsky K, Teixeira CM, *et al.*: **What is the Most Sensitive Measure of Water Maze Probe Test Performance?** *Front Integr Neurosci.* 2009; **3**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Minka TP: **Estimating a Dirichlet distribution.** *Annals of Physics.* 2003; **2000**(8): 1–13.
Reference Source
6. Barndorff-Nielsen OE, Cox DR: **Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator.** *J R Stat Soc Series B Stat Methodol.* 1984; **46**(3): 483–495.
Publisher Full Text
7. Oddo S, Caccamo A, Shepherd JD, *et al.*: **Triple-transgenic model of Alzheimer's disease with plaques and tangles: intracellular Abeta and synaptic dysfunction.** *Neuron.* 2003; **39**(3): 409–421.
PubMed Abstract | Publisher Full Text
8. Maugard M: **Role of Astrocytic Serine in Learning and Memory and its Implications in Alzheimer's Disease.** PhD thesis, Université Paris Saclay. 2018; 6.
Reference Source
9. Clinton LK, Billings LM, Green KN, *et al.*: **Age-dependent sexual dimorphism in cognition and stress response in the 3xTg-AD mice.** *Neurobiol Dis.* 2007; **28**(1): 76–82.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Billings LM, Oddo S, Green KN, *et al.*: **Intraneuronal Abeta causes the onset of early Alzheimer's disease-related cognitive deficits in transgenic mice.** *Neuron.* 2005; **45**(5): 675–688.
PubMed Abstract | Publisher Full Text
11. Allaire JJ, Ushey K, Tang Y, *et al.*: **reticulate: R Interface to Python.** 2017.
Reference Source
12. Suh EJ, Doux C, Braem N: **xuod/dirichlet 0.8 (Version 0.8).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3373955

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 05 November 2019

https://doi.org/10.5256/f1000research.23263.r55813

✔ **Avgoustinos Vouros**
University of Sheffield , Sheffield, UK
Luca Manneschi
Department of Computer Science, University of Sheffield, Sheffield, UK
Eleni Vasilaki
Department of Computer Science, University of Sheffield, Sheffield, UK

Thank you for sharing with us a revised version of the manuscript. First, we would like to clarify that approval with reservations intended to be approval with minor corrections.

Further, we are pleased with the way you have addressed the second point we raised, which was to comment on the applicability of your methods on other types of trials.

However, we feel that the reply to the first point we raised didn't add to the manuscript, in fact, it might have even deteriorated the flow of the text. The comment was not meant to be interpreted such as "why you do a t-test", this is very well justified in the original version. Rather we are asking: "Could you comment whether in the case that the data are not following a Gaussian distribution you can still apply your method?"

We will not provide any further reviews, but rather leave it to you whether you would like to improve the manuscript based on the feedback or not.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, Behavioural Neuroscience

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 30 September 2019

https://doi.org/10.5256/f1000research.22038.r53542

?

Avgoustinos Vouros
University of Sheffield , Sheffield, UK
Luca Manneschi
Department of Computer Science, University of Sheffield, Sheffield, UK
Eleni Vasilaki
Department of Computer Science, University of Sheffield, Sheffield, UK

The authors suggest a new statistical procedure to analyze in more depth a common performance measurement in the Morris Water Maze experimental procedure.

The manuscript is well-written and all the methods explained in detail. The new proposed method is well-justified and shows much potential in the field of data analytics for behavioural procedures similar to the Morris Water Maze. Code of all the methods that the author describe is open source and publicly available. Instructions on how to run the code are clear.

I would like to point out some aspects that could have been addressed/discussed more:

1. For the post-hoc analysis the authors suggest the usage of t-tests on the four quadrants. Have the authors consider, if applicable, alternative tests of hypothesis in case the assumptions of the t-test (e.g. the data are not normally distributed) are not fulfilled?

2. The authors focus on the probe trials but couldn't analysis on other types of trials (e.g. reversal learning) also been benefit from their method? Can the authors comment if their method is also applicable to other scenarios where there are mutually exclusive possibilities?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, Behavioural Neuroscience

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 24 Oct 2019

Gilles Bonvento, Centre National de la Recherche Scientifique, Université Paris-Sud, Université Paris-Saclay, Fontenay-aux-Roses, France

We thank the reviewers for their positive comments regarding our statistical test.

1. For the post-hoc analysis the authors suggest the usage of t-tests on the four quadrants. Have the authors consider, if applicable, alternative tests of hypothesis in case the assumptions of the t-test (e.g. the data are not normally distributed) are not fulfilled?

As far as we know, there is no test available to compare non-normal distribution with a theoretical one.

1. The authors focus on the probe trials but couldn't analysis on other types of trials (e.g. reversal learning) also been benefit from their method? Can the authors comment if their method is also applicable to other scenarios where there are mutually exclusive possibilities?

Our method can be used each time a probe test is performed with the sum spent in the 4 quadrants totalling 100%, such as for reversal learning.

Competing Interests: No competing interests were disclosed.

Reviewer Report 23 September 2019

https://doi.org/10.5256/f1000research.22038.r53541

Richard GM Morris
Centre for Cognitive and Neural Systems, University of Edinburgh, Edinburgh, EH8 9JZ, UK

Maugard *et al.* identify a major difficulty in analyzing probe test data from the watermaze which typically consists of time (in sec) or proportions of time (%) spent in each of four quadrants of the pool. While it is straightforward to suppose that more time should be spent in the training quadrant, leading some to analyze only that quadrant, a better method would be to use a mathematical approach that recognizes

that the time in all 4 quadrants must add to 100% but their distribution is of interest. Some authors have attempted (upon advice) to address the quadrants problem by reducing the numerator degrees of freedom by 1 (e.g. Morris), but this is unlikely to be statistically adequate, and anyway suffers from issues associated with the normality of the data etc. What is proposed here is something called the Dirichlet distribution which explicitly recognizes the summation to 100% problem and provides a mathematical way of looking at more than just the training quadrant.

This innovation seems valuable, but I add a qualification. This is that the watermaze is, essentially, no more than a pool of water with a hidden platform in which a large variety of different tasks can be run. That most users of the watermaze use only the standard reference memory task does not mean that other variants are not of interest - reversal learning (Hans-Peter Lipp and David Wolfer, Univ Zurich), the delayed match to place task (Morris, e.g. Steele and Morris Hippocampus 1999[1]), and procedures for dissociating encoding and retrieval (e.g. Rossato *et al.* Current Biology 2018[2]). These other procedures are of analytical interest. However, they may still require probe test data (e.g. Rossato *et al.* 2018[2]).

My recommendation is that the paper be provisionally accepted - it makes a very valuable point - but attention must be paid to the fact there is not just one single task that can be run in a watermaze. There is greater potential.

To add also, there are a large variety of measures of performance including latency, path-length, directionality, proximity index and all the measures associated with a probe test. Given this, the thrust of the paper is arguably less novel than the authors surmise, albeit this being a very clever solution to an unsolved problem associated with probe tests.

Recommendation: index subject to small revisions (as advised above).

References

1. Steele R, Morris R: Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus*. 1999; **9** (2): 118-136 <118::AID-HIPO4>3.0.CO;2-8">Publisher Full Text
2. Rossato JI, Moreno A, Genzel L, Yamasaki M, Takeuchi T, Canals S, Morris RGM: Silent Learning.*Curr Biol*. 2018; **28** (21): 3508-3515.e5 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neurobiology of learning and memory.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 24 Oct 2019

Gilles Bonvento, Centre National de la Recherche Scientifique, Université Paris-Sud, Université Paris-Saclay, Fontenay-aux-Roses, France

We would like to thank the reviewer for his positive comments. We appreciate the explicit acknowledgment regarding our valuable test. We have now mentioned in the text that a large variety of different tasks can be run using modifications of the basic protocol.

Competing Interests: No competing interests were disclosed.