Abel Sánchez-Jiménez¹, Gonzalo Aparicio-Rodríguez¹, Paloma Manubens¹, Carlos Calvo-Tapia¹, and José Antonio Villacorta Atienza¹,²

¹Unit of Biomathematics. Department of Biodiversity, Ecology and Evolution. Faculty of Biological Sciences, Complutense University of Madrid
²Unit of Applied Mathematics. School of Optics and Optometry, Complutense University of Madrid

May 09, 2024
THE SIMILARITY STRUCTURE: A NOVEL CHARACTERIZATION OF EFFECT SIZE FOR LARGE SAMPLES

Abel Sánchez-Jiménez¹, Gonzalo Aparicio-Rodríguez¹, Paloma Manubens¹, Carlos Calvo-Tapia¹, José Antonio Villacorta-Atienza¹²,*

1 Unit of Biomathematics. Department of Biodiversity, Ecology and Evolution. Faculty of Biological Sciences. Complutense University of Madrid, Spain
2 Unit of Applied Mathematics. School of Optics and Optometry. Complutense University of Madrid, Spain

* Corresponding author: josea@ucm.es

Abstract

Statistical inference traditionally relies on $p$-values, assessing the alignment between data and the absence of effect. Nevertheless, in large datasets, $p$-values lose relevance, marking minor differences as statistically significant. Thus, it becomes imperative in large data to evaluate the practical, clinical, or biological effect's magnitude. Non-dimensional metrics like Cohen's $d$, allow for general comparisons, but they can obscure practical meaning. Dimensional metrics, such as confidence intervals, lack standardization and may complicate practical interpretation. We propose a novel approach termed the similarity structure for characterizing differences in large samples focused on the probability distribution of two subsamples being of size $N$, given that they are similar (statistically non-different). This quantifies the effect size as the expected sample size when similarity exists, irrespective of data nature, dimensionality, or hypothesis testing. Additionally, it can be translated into common measures like Cohen's $d$ and required sample sizes for a statistical power of 0.9. Furthermore, the similarity structure allows to statistically compare effect sizes, assessing the importance of the factors involved in sample differences. The similarity structure allows for the transparent and versatile assessment, interpretation, and comparison of effect sizes, contributing to more comprehensible and reproducible scientific research. This approach is demonstrated with real-world examples.

Key words: effect size, large samples, hypothesis testing, $p$-value, power analysis

Introduction

Modern science cannot be understood without statistics, a key field to provide objective, rigorous, and reproducible conclusions about the experimental data and, eventually, the findings supporting scientific progress. From the very foundation of modern statistics to the present day, the concept of $p$-value has been one of the cornerstones of inference. The conceptual background of $p$ follows Popper’s thesis claiming that scientific theories can only be falsified [1], as it quantifies the probability of being wrong based on the experimental evidence [2,3]. Thus, $p$ provides a well-defined number measuring the evidence against the so-called null hypothesis, the opposite to the proposed hypothesis to be tested, and allows an objective decision to be made about the validity of the hypothesis of interest. Nonetheless, despite these attractive properties, the $p$-value is far from being exempt from criticism.
Currently, a growing number of scientists are aware of the limitations of the $p$-value, the necessity of extending its correct use, and the need to accompany it with complementary measures or even replace it [4]. From a conceptual point of view, the $p$-value is focused on the null hypothesis $H_0$ so it does not provide any information about the alternative hypothesis, usually the one we want to test as it concerns the presence of the phenomenon of interest [5,6]. This forces to validate the alternative hypothesis through a proper experimental design, which must discard other plausible explanations [7]. To mitigate these limitations, different approaches have been proposed, some of them radically different such as Bayesian inference and $p$-based minimum Bayes factors, to estimate the odds of the null hypothesis relative to the alternative hypothesis [8]. However, beyond the theoretical perspective, both the conspicuous variability of $p$ and its sensitivity to the sample size are some of its major issues when it supports binary decisions after comparing it with the decision threshold $\alpha$. These shortcomings, among others, of $p$-based decisions and their dichotomic nature have been criticized specially for their implications in experiment replicability and so in scientific reproducibility [9].

This situation has worsened over the last decades as the improvement in data management and computation capabilities has made studies based on large samples commonplace, which has extended the misuse of the $p$-value. It is largely known that the $p$-value is sensitive to the sample size, so, in practice, the larger the sample data, the smaller the $p$-value even for similar samples. Thus, hypothesis test inference on large samples will likely lead to systematic rejection of the null hypothesis, which is frequently interpreted as a validation of the alternative hypothesis. Leaving aside the interpretation issues discussed above, far from being a flaw of the $p$-value, this is just due to how $p$ works in the data: any slight difference between populations/samples will be revealed by $p$-based inference on large enough data samples [10]. Therefore, scientific conclusions relying solely on the $p$-value must be avoided and should be supported by complementary measures quantifying in which degree the phenomenon of interest is manifested.

Effect size measures are data-driven assessments of the presence of a certain phenomenon in the population [9]. However, when reporting the magnitude of an effect, it should be distinguished statistical from practical/biological/clinical importance [11]. In other words, with independence of the size of the effect provided by the specific measure used in the study, the researcher must interpret it in the light of the scientific context to conclude about its validity and impact. Nevertheless, identifying such ‘validity values’ for the effect size could be an issue. Research focused on phenomena described by multiple and/or complex descriptors, experimental exploration of recent findings or findings supported by scarce literature, theory-testing research, where small effect sizes may be meaningful if they were predicted by the theory [11], etc. are circumstances in which interpreting the results and so the effect size, can be a challenge. As an example, consider a novel blood testing to detect the presence of a certain molecule as an early marker for Alzheimer’s disease. After testing this potential marker in a large cohort of control and mild cognitive impairment volunteers, the $p$-value provided by the pertinent hypothesis tests is $2 \times 10^{-6}$ and the reported effect size is 12.3 pg/ml. Given the large size of the samples, the decision about the validity of the biomarker cannot be based on the $p$-value and should be grounded on the effect size. However, unless the impact of such a molecule has been extensively described (e.g., previous studies, existing literature, complementary behavioural experiments, etc.), the researcher may not be able to interpret what 12.3 pg/ml biologically means and thus conclude on the clinical importance of the found effect. On the contrary, if the conducted study concerns well-established phenomena described through easily interpretable statistics, the researcher could delimit a priori the validity values for the applied statistic and introduce the accepted value for the effect size into the own hypothesis test by defining the null hypothesis according to this estimated effect [12]. In this case, the $p$-value approach is fully functional, and no complementary measures are necessary to accept...
(with the assumed Type-I error) the $p$-based decision. For example, if we declare practical importance when the mean difference of two compared populations, $\mu_1 - \mu_2$, is above 0.34 (lower bound for the effect size), defining $H_0$: $\mu_1 - \mu_2 \leq 0.34$ will provide both statistical and practical significance through the resulting $p$-value. It should be remarked that, under this approach, there is no issue with large samples, i.e., the $p$-value will be under the significance threshold $\alpha$ regardless of the size of the samples (see Supplementary).

In addition to these conceptual limitations regarding the interpretation of the effect size, its evaluation entails diverse issues that must be considered. Many of the effect size measures are dimensionless (standardized), which makes it possible to compare the effect magnitude among different studies but making it difficult to interpret the actual impact of the phenomenon in the specific research context [13]. Thus, it is advised to complement the effect size with the confidence interval (CI), which provides an envelope within which the actual value (in actual units) of the property of interest is likely to lie [14]. There is a plethora of effect size measures, but some of the most widely used are the Cohen’s $d$ (standardized difference of means) [15] and the Pearson’s $r$ (quantifying the linear association) [16], to be applied when the dependent variable is continuous and the predictor (independent) variable is categorical and continuous, respectively. These, and many others effect size measures, require that the involved distributions fulfil specific assumptions (e.g., normality). However, frequently these assumptions are not applicable to the analysed data, and non-parametric approaches are mandatory. Non-parametric effect size estimation is an open problem; there exist multiple methods and approaches, as assessing the effect size between two distributions as in Cohen’s $d$ [17,18] and transforming the original data into normal data where the above-discussed methods are applicable [13]. However, there is no consensus about their use because of their limitations and, more importantly, because the distinctions between these multiple approaches are hardly accessible to non-experts, which severely limits their use as statistic tool in scientific research.

The issue of assessing the effect size under a non-parametric approach is extensible to CIs. Even when massive datasets do not fulfil normality requirements, their large size is an advantage, since the CIs become narrower as the data sample grows and the estimate of the effect size will be more accurate under the assumption that the mean is a good statistic for the study [18]. Nevertheless, non-parametric data usually require other statistical descriptors (e.g., the median) and, though there are different approaches to calculate CIs for non-parametric less-extended statistics [19], there is no consensus on the closed-form expressions for CIs linked to the most used non-parametric tests. The most extended solution to calculate non-parametric CIs is bootstrapping, in which the CIs come from a distribution of estimated values of a specific statistic obtained from resampling 10,000-15,000 times the original dataset [20]. Despite its flexibility, bootstrapping is a brute force method, so it is computationally expensive, especially for large samples (as each resampled set and the original dataset must have the same size). Actually, what we meant as ‘large sample’ is pertinent here. The problem of enhancing non-relevant differences in large samples when hypothesis tests are used involves dataset sizes around a few hundred [18]. However, accurate CI assessment by bootstrapping may require sample sizes ~ 2000-8000, so in many cases involving large samples, application of bootstrapping methods may lead to inaccurate CI estimations [20]. Finally, though basic bootstrapping method does not require assumptions about the data distribution, it has its own assumptions (e.g., independence in the data), so if such premises are not fulfilled other specialized bootstrapping approaches are mandatory, whose use demands knowledge and experience on these technics. All these limitations make it difficult to rigorously estimate the CIs, and thus the effect size, when assumptions about the structure of data are not or cannot be made.
Different alternatives have been proposed to make inference on large samples via hypothesis test focused on limiting the impact of the sample size on the \( p \)-value instead of reinforcing the interpretation of the obtained \( p \)-value. In this regard, it has been suggested to decrease the significance threshold \( \alpha \) according to the increasing of the sample size, modulating the significance of the findings [21,22]. Another approach has been recently proposed to differentiate between practical and statistical differences by analysing the relationship between the \( p \)-value and the sample size in large datasets [23]. This methodology avoids some limitations of \( p \) by working with its distribution for many random subsamples of different sizes \( n \), which leads to a function \( p(n) \) estimating the dependency of \( p \) on \( n \). However, \( p(n) \) is eventually obtained from the mean \( p \)-value for each size \( n \), \( \bar{p} \), which is not a \( p \)-value itself unless it is scaled by 2 [24], a factor that cannot be improved in general [25]. Thus, when averaging \( p \)-values to compare with a significance threshold \( \alpha \), either \( 2\bar{p} \) must be used [24] or, if \( \bar{p} \) is used, the significance threshold must be set accordingly to \( \frac{\alpha}{2} \) [26]. Moreover, \( \bar{p} \)-based inference is considered a poor method, as, when averaging, one large \( p \)-value can overwhelm many small \( p \)-values [27].

In summary, hypothesis inference testing on large samples entails difficulties in characterizing and interpreting the practical/clinical/biological importance of the likely-obtained significance. In general, when no assumptions about the structure of the data can be made or there is no a clear knowledge about the practical impact of the reported effect sizes, support the \( p \)-value by quantifying the real significance of the differences found may require advanced skills in statistics, and, even in this case, there is no consensus on many of the existing approaches that can be applied. Therefore, statistical analysis of large samples is especially challenging for most researchers, who are non-experts in statistics and require accessible tools to transparently explore the potential effects present in their experimental data.

The present work proposes a novel approach to characterize the differences between two large samples named similarity structure, which estimates the probability distribution of finding two similar (statistically non-different) subsamples of size \( N \), where each subsample comes from one of the large subsamples. Thus, the sample differences are assessed not as a single value (the \( p \)-value in traditional frequentist inference) but as a function of the sample size, whose structure, statistical properties, and related statistical measures allows to characterize the effect size in different and complementary ways. Besides the qualitative description of the sample differences provided by the dependency on \( N \) of the similarity structure, the expected sample size of the similar subsamples, named similarity size, provides an effect size measure in transparent and general units. The similarity size can also be interpreted in terms of other widely extended effect size measures as the Cohen’s \( d \), and estimates the sample size for a statistical power of 0.9; this interpretability can contribute to its practical use. Finally, the similarity structure allows comparison of effect sizes to conclude whether they are significantly different, characterizing the magnitude of the effect relative to a reference effect size. The similarity of subsamples of different sizes is evaluated by introducing them into the proper hypothesis test according to the nature of the data. This makes the similarity structure a flexible, assumption-free, and user-friendly approach for comparing large one and high-dimensional samples. To illustrate the application of the similarity structure, four cases have been studied. The first two cases illustrate the comparison of numerically generated normal distributions with different means and standard deviations, respectively. The remaining cases show how the similarity structure serves to analyse and discuss experimental large and multidimensional datasets of disparate nature (eye tracking data and distribution of protist population) in a common framework.
Materials and methods

The similarity structure

To introduce the similarity structure, let us consider two large samples $S_1$ and $S_2$, the experimental samples that will be analyzed to characterize and assess the size of the differences between them, i.e., the effect size. Due to their large size, we can consider $S_1$ and $S_2$ as ‘populations’ in the sense that their frequency distributions will serve to generate the corresponding density functions for continuous variables and probability distributions for discrete variables. This will allow us to simulate the original samples $S_1$ and $S_2$ but with arbitrary size $N$ by generating random subsamples of size $N$, $s_1$ and $s_2$ (from $S_1$ and $S_2$, respectively) drawn with replacement. This will be denoted as $s_j \sim S_j$, where $j = 1, 2$.

Therefore, to statistically analyze the samples $S_1$ and $S_2$ according to their size, we introduce sets of $r$ pairs of random subsamples of size $N$, defined as:

$$E_N \equiv \left\{ (s_{1,i}, s_{2,i}) \mid s_{1,i} \sim S_1, s_{2,i} \sim S_2 \text{ and size}(s_{1,i}) = \text{size}(s_{2,i}) = N \right\}_{i=1}^r,$$

(1)

where $N$ takes $q$ equidistant natural numbers between $N_0$ and $N_f$; this will be denoted as $N \in \{N_0, N_f; q\}$. The final subsample size $N_f$ satisfies that $N_f \leq N_{\text{exp}} \equiv \min\{\text{size}(S_1), \text{size}(S_2)\}$, but it may be much smaller than $N_{\text{exp}}$ when the samples $S_1$ and $S_2$ are quite different (see Computation of the similarity structure section).

For the sake of clarity, we will consider hereafter that any subsample denoted by $s_1$, $s_2$ or similar satisfies that $s_1 \sim S_1$ and $s_2 \sim S_2$, respectively, and will belong to $E_N$, where $N$ is its size.

The similarity structure will be based on the probability of finding non-different samples of size $N$. This probability will be estimated by applying to each pair $(s_{1,i}, s_{2,i})$ the proper hypothesis test $H(s_{1,i}, s_{2,i})$ with a significance level $\alpha$. For instance, if we consider continuous variables and no assumptions about the structure of the samples beyond the independence of the observations, the Kolmogorov-Smirnov test could be applied, so:

$$H(s_{1,i}, s_{2,i}) = \begin{cases} H_0: s_{1,i} \text{ and } s_{2,i} \text{ come from the same distribution} \\ H_1: s_{1,i} \text{ and } s_{2,i} \text{ come from different distributions} \end{cases}$$

where $H_0$ and $H_1$ are the null and alternative hypothesis, respectively. After the $r$ pairs of subsamples in $E_N$ have been compared through the hypothesis test, for $r_0$ of those $r$ hypothesis tests the null hypothesis $H_0$ cannot be rejected. Hereinafter, when $H(s_1, s_2) = H_0$ we will say that $s_1$ and $s_2$ are similar. Thus, we estimate the probability of getting a pair of similar subsamples $s_1$ and $s_2$ of size $N$ as:

$$p(>\alpha|N) \equiv p(H(s_1, s_2) = H_0 \mid \text{size}(s_1) = \text{size}(s_2) = N) \equiv \frac{r_0}{r}.$$

(2)

The large sizes of samples $S_1$ and $S_2$, will make evident the differences between them, so we can assume as true the alternative hypothesis that populational differences are at least as large as those observed in the samples when calculating the probability that $H(s_1, s_2) = H_0$ (type-II error). Thus, the estimation in Eq. (2) is the numerical approach to the $\beta$-curve vs. the sample size $N$, where $\beta$ or type-II error is related to the statistical power as $\beta = 1 - \text{power}$. This Monte Carlo approach to
power (β) estimation is often the only solution for dealing with real data, which do not usually fit those designs that admit analytical expressions of the power [28].

The main assumption of this proposal is that the similarity structure of the samples \( S_1 \) and \( S_2 \) is accurately represented by the set of pairs of subsamples \( \{ E_N \}_{N \in [N_0, N_f, q]} \). This accuracy is checked by comparing the probability \( p(\alpha > N) \) obtained from the Eq. (2) and the posterior fitting (see below) with the theoretical \( \beta \)-curve for the comparison of large normal samples with different means and equal standard deviations (see Results, section **Samples from normal distributions with different means**). Such a comparison reveals a difference around 1% and 2% between the \( \beta \)-curve and the numerical approach \( p(\alpha > N) \) in Eq. (2). This demonstrates the suitability of such numerical approach and, therefore, of the characterization of the comparison between \( S_1 \) and \( S_2 \) by the set \( \{ E_N \}_{N \in [N_0, N_f, q]} \).

The similarity structure of two samples \( S_1 \) and \( S_2 \) is intended to characterize the effect size by assessing how the similarity of their subsamples in \( \{ E_N \}_{N \in [N_0, N_f, q]} \) depends on their size or, more specifically, what is the probability of obtaining subsamples of size \( N \) from those in \( \{ E_N \}_{N \in [N_0, N_f, q]} \) given the similarity of such subsamples. Thus, the larger the differences between \( S_1 \) and \( S_2 \), the more rapidly the probability of finding similar subsamples decreases as their size \( N \) increases, since the differences between the samples become more evident as the size of the subsamples increases.

Therefore, the similarity structure will be given by the probability:

\[
p_{\alpha}(N) \equiv p(N \mid \alpha) \equiv p(\text{size}(s_1) = \text{size}(s_2) = N \mid H(s_1, s_2) = H_0),
\]

(note the shorter notation \( p_{\alpha}(N) \) for \( p(N \mid \alpha) \) to remark its dependency on \( N \) which will be obtained by applying the Bayes' theorem:

\[
p_{\alpha}(N) = \frac{p(\alpha > N)p(N)}{p(\alpha > N)} = \frac{p(\alpha > N)p(N)}{\sum_{N \in [N_0, N_f, q]} p(\alpha > N)p(N)},
\]

where \( p(N) \) denotes the probability of having two subsamples of size \( N \) among the pairs of subsamples in \( \{ E_N \}_{N \in [N_0, N_f, q]} \).

Since the subsamples considered for generating the similarity structure are arranged according to their sizes in the sets \( \{ E_N \}_{N \in [N_0, N_f, q]} \), the total number of subsamples is \( rq \) and the number of subsamples of size \( N \) (i.e., those subsamples in \( E_N \)) is \( r \). Thus, assuming no bias in the subsample sizes:

\[
p(N) = \frac{r}{rq} = \frac{1}{q}.
\]

Substituting this into Eq. (4) leads to:
\[
p_\alpha(N) = \frac{\sum_{N \in [N_0, N_f]} p(\alpha | N)}{q} = \frac{\sum_{N \in [N_0, N_f]} p(\alpha | N)}{\sum_{N \in [N_0, N_f]} p(\alpha | N)} = \frac{p(\alpha | N)}{\sum_{N=N_0}^\infty p(\alpha | N)}.
\]

The ideal case is the asymptotic configuration, where \( N_f \to \infty, q = N_f - N_0 + 1, \) and \( r \to \infty \) (since \( r = \frac{(N_f+1)N_f}{2} \) as maximum possible value), so the similarity structure is finally defined as:

\[
p_\alpha(N) = \frac{p(\alpha | N)}{\sum_{N=N_0}^\infty p(\alpha | N)}.
\]

Note that 1) the similarity structure \( p_\alpha(N) \) is a probability distribution, since \( \sum_{N=N_0}^\infty p_\alpha(N) = 1 \), while \( p(\alpha | N) \) is not, and 2) \( N_0 \geq 2 \) because statistical tests require a minimum sample size.

To balance the theoretical definition of the similarity structure in Eq. (7) with a practical computational implementation (represented by the values of parameters \( r, N_0, N_f, \) and \( q; \) see below Computation of the similarity structure), the proper function will be fitted to the probabilities \( \{p(\alpha | N)\}_{N \in [N_0, N_f]} \) given by Eq. (2). We propose that \( p(\alpha | N) \) can be described by the relationship:

\[
p(\alpha | N) = (a + bN + c N^2 + d N^3)e^{-dN},
\]

where the parameters \( \{a, b, c, d, e\} \) will be adjusted after fitting the right term of Eq. (8) to the data \( \{p(\alpha | N)\}_{N \in [N_0, N_f]} \). The suitability of Eq. (8) is given by two factors. On the one hand, it asymptotically approaches to zero as \( N \to \infty \). On the other hand, under Eq. (8) the normalization term \( \sum_{N=1}^\infty p(\alpha | N) \) admits a closed form, which allows an exact calculation of the similarity structure. In particular:

\[
\sum_{N=N_0}^\infty p(\alpha | N) = \sum_{N=N_0}^\infty (a + bN + c N^2 + d N^3)e^{-dN} = a \sum_{N=N_0}^\infty e^{-dN} + b \sum_{N=N_0}^\infty Ne^{-dN} + c \sum_{N=N_0}^\infty N^2 e^{-dN} + e \sum_{N=N_0}^\infty N^3 e^{-dN},
\]

where each term can be written as:

\[
\sum_{N=N_0}^\infty e^{-dN} = \frac{e^{-d(N_0-1)}}{e^d - 1}
\]

\[
\sum_{N=N_0}^\infty Ne^{-dN} = \left( \frac{N_0 e^d - N_0 + 1}{(e^d - 1)^2} \right) e^{-d(N_0-1)}
\]
\[
\sum_{N=N_0}^{\infty} N^2 e^{-dN} = \left( -\frac{N_0(N_0(1 - e^d) - 1) + N_0 - 1}{(e^{-d} - 1)^2} e^{-dN_0} + 2 \frac{(N_0 - 1)(N_0 e^{-d} - 1) e^{-dN_0}}{(e^{-d} - 1)^3} \right) e^{-d}
\]  

\[
\sum_{N=N_0}^{\infty} N^3 e^{-dN} = \left( -\frac{N_0(-N_0 e^{-d} + N_0 - 1) + N_0 - 1}{(e^{-d} - 1)^2} + 2 \frac{(-N_0 + (N_0 - 1)e^{-d}) e^{-dN_0}}{(e^{-d} - 1)^3} \right.
+ \left. \frac{(N_0^2 - N_0)(-N_0 e^{-d} + N_0 - 1) + 2N_0(N_0 - 1)}{(e^{-d} - 1)^2} e^d e^{-dN_0} \right.
+ \left. 4 \frac{(N_0 - N_0 e^{-d} + N_0 - 1) e^{-dN_0}}{(e^{-d} - 1)^3} \right)
+ \left. \frac{6(-N_0 + N_0 e^{-d} - e^{-d}) e^{-dN_0}}{(e^{-d} - 1)^4} \right) e^{-d}
\]  

Therefore, the similarity structure will finally be written as:

\[
p_\alpha(N) = (\hat{a} + \hat{b}N + \hat{c}N^2 + \hat{d}N^3)e^{-dN},
\]  

where

\[
\hat{x} = \frac{\chi}{a \sum_{N=N_0}^{\infty} e^{-dN} + b \sum_{N=N_0}^{\infty} Ne^{-dN} + c \sum_{N=N_0}^{\infty} N^2 e^{-dN} + e \sum_{N=N_0}^{\infty} N^3 e^{-dN}}
\]  

and \(x \in \{a, b, c, d, e\}\).

In summary, the steps to obtain the similarity structure of two samples \(S_1\) and \(S_2\) are:

1. Generate the sets \(\{E_N\}_{N \in [N_0, N_f; q]}\) with the \(r\) pairs of subsamples randomly drawn with replacement from \(S_1\) and \(S_2\) for each \(N\) (Eq. (1)).

2. Apply the appropriate hypothesis test to compare the pairs of subsamples in the sets \(\{E_N\}_{N \in [N_0, N_f; q]}\) to obtain \(\{p(\alpha \mid N)\}_{N \in [N_0, N_f; q]}\), i.e., the estimation of the probability \(p(\alpha \mid N)\) as a function of \(N\) (Eq. (2)).

3. Fit the function \((a + bN + cN^2 + eN^3)e^{-dN}\) to \(\{p(\alpha \mid N)\}_{N \in [N_0, N_f; q]}\) and get the parameters \(\{a, b, c, d, e\}\).

4. Calculate the parameters \(\{\hat{a}, \hat{b}, \hat{c}, \hat{d}\}\) according to Eq. (15) and Eqs. (10)-(13).
5. Finally, obtain the similarity structure \( p_\alpha(N) \) by substituting the parameters \( \{\hat{a}, \hat{b}, \hat{c}, \hat{d}\} \) into Eq. (14).

Assessing and interpreting the effect size: the similarity size

The similarity structure of two samples characterizes how likely is that two similar subsamples have a certain size. Since the differences between the samples will be more salient as the probability of finding similar subsamples decreases with their size, the steeper the \( p_\alpha(N) \) curve, the greater the differences between the compared samples and, therefore, the corresponding effect size. This is illustrated in Fig. 1.A, which shows the similarity structures of two comparisons: \( S_0 \sim N(0,1) \) vs. \( S_1 \sim N(0.05,1) \) (blue curve), and \( S_0 \sim N(0,1) \) vs. \( S_2 \sim N(0.07,1) \) (red curve), where \( N_{\text{exp}} = 20000 \) for all samples \( S_0, S_1, \) and \( S_2 \). Denoting as \( p_{\alpha,i}(N) \) the similarity structure for the comparisons \( S_0-S_i \), with \( i = 1,2 \), \( p_{\alpha,1}(N) \) (red curve) is steeper than \( p_{\alpha,1}(N) \) (blue curve) since the mean of the normal distribution for \( S_2 \) is greater than for \( S_1 \).

This way, the similarity structure could be used to characterize the effect size at different levels. On the one hand, we propose to assess the effect size by means of the expected size under the probability distribution given by the similarity structure, denoted as similarity size or \( N_S \):

\[
N_S \equiv \sum_{N=N_0}^{\infty} N \, p_\alpha(N). \tag{16}
\]

Introducing the fitted function in Eq. (14):

\[
N_S = \sum_{N=N_0}^{\infty} N \left( \hat{a} + \hat{b} \, N + \hat{c} \, N^2 + \hat{d} \, N^3 \right) e^{-dN} = \\
= \hat{a} \sum_{N=N_0}^{\infty} N e^{-dN} + \hat{b} \sum_{N=N_0}^{\infty} N^2 e^{-dN} + \hat{c} \sum_{N=N_0}^{\infty} N^3 e^{-dN} + \hat{d} \sum_{N=N_0}^{\infty} N^4 e^{-dN}, \tag{16}
\]

so the similarity size is exactly calculated from Eqs. (11)-(13). A closed expression for \( \sum_{N=N_0}^{\infty} N^4 e^{-dN} \) can be found in the Supplementary.

Figure 1.A shows the corresponding similarity sizes \( N_{S,1} \) and \( N_{S,2} \) for the sample comparisons \( S_0-S_1 \) (blue) and \( S_0-S_2 \) (red), respectively. By definition, the similarity size is the expected size of similar samples. Following the previous reasoning, the smaller the similarity size, the larger the differences between the compared samples, since the similarity is lost more rapidly with increasing the sample size as the compared samples become more and more different. In other words, a small \( N_S \) means that the set of similar subsamples is small, so it would be unlikely to find similarities, implying that the original samples are quite different, and the effect size is large. On the contrary, a large \( N_S \) means that the set of similar subsamples is large, so similarity is common, and the samples should be slightly different, implying a small effect size.

The similarity size \( N_S \) allows to measure and interpret the effect size in a general and transparent way, characterizing it as observations regardless of the sample units. More in detail and considering the above discussion about interpreting the subsample \( s \in E_N \) as the sample \( S \) if it were of size \( N \), the effect size measured by \( N_S \) will be the expected size of the compared samples given that they
were similar. Moreover, our results suggest that $N_S$ serves as estimator of the sample size required to attain a statistical power of 90% assuming an effect such as that observed in the samples, which increases its interpretability in more practical terms (see Results).

On the other hand, the generality of the similarity size makes it possible to interpret it in terms of the Cohen’s $d$ even in situations where $d$ cannot be directly applied. This is developed in the Results section. First, a set of large normal samples is considered: $S_0 \sim N(0,1) \text{ and } S_m \sim N(m, 1)$, with $0.05 \leq m \leq 1.2$. The similarity structures of $S_0$ vs. $S_m$ are then calculated by applying the Kolmogorov-Smirnov test, leading to a relationship between the similarity size $N_S$ and the sample mean $m$. Due to the large sample size, $m$ will be practically equal to Cohen’s $d$, so the relation between both effect size measures, $N_S$ and $d$, is provided. Then, a set of large normal samples with different standard deviations $S_{std} \sim N(0,\text{std})$ with $1 \leq m \leq 2$, is considered and the similarity structures of $S_0$ vs. $S_{std}$ are obtained by applying the Kolmogorov-Smirnov test. Since $N_S$ only depends on the hypothesis test and $\alpha$ (equal in both cases), the previous relationship between $N_S$ and $d$, allows to characterize the effect size due to different standard deviations in terms of Cohen’s $d$. Therefore, the similarity size serves both as an interpreter of the effect size by itself and as a ‘translator’ to widely-stablished effect size measures.

Comparing effect sizes: the similar effect probability

The effect size is a central quantification of the real impact of scientific findings. However, its interpretation, regardless of the measure used, should be made in the particular context of the study. In the present work, the similarity size characterizes the effect size as number of observations, e.g., as number of participants in a study. However, despite of its transparency, concluding that a certain number of individuals is acceptable or not depends on the problem addressed (it is not the same in a study on a fatal disease as if headache is studied). Therefore, beyond assessing and interpreting the effect size, it is often interesting to quantify the difference between effect sizes.

Let us consider, for example, the comparison of the effect of two drugs versus a placebo. In case the effect of one of the drugs is already known and accepted as relevant, the objective will be to assess the difference between the two drugs to conclude whether the second compound has more or less effect than the first one. The similarity structure, as a probability distribution, offers a statistical characterization of the differences between the effect sizes of each drug vs. the placebo. Let us consider again the situation in Fig. 1. Since the similarity size $N_{S,i}$ is a measure of the effect size for the comparisons $S_0$ vs. $S_i$, with $i = 1,2$, we can ask ourselves what the probability of one of the effect sizes ($N_{S,i}$) will be when the other comparison is made. Thus, we evaluate $p_{\alpha,1}(N_{S,2})$ and $p_{\alpha,2}(N_{S,1})$, and define the similar effect probability $p_S$ as the minimum of both probabilities, i.e.,

$$p_S(N_{S,1}, N_{S,2}) \equiv \min \left( p_{\alpha,1}(N_{S,2}), p_{\alpha,2}(N_{S,1}) \right).$$

(17)

This is illustrated in Fig. 1 B, which depicts $p_{\alpha,1}(N_{S,2})$ (blue area) and $p_{\alpha,2}(N_{S,1})$ (red area) for the comparisons $S_0$ vs. $S_1$ and $S_0$ vs. $S_2$, respectively. Note from the Fig. 1 that the similarity structures of both comparisons must have been calculated with the same value for the initial size $N_0$. The smallest one, $p_{\alpha,2}(N_{S,1})$, is equal to the similar effect probability $p_S$. Thus, $p_S$ gives the probability that, being $S_0$ and $S_2$ similar, they are of size $N_{S,1}$. Thus, if $p_S$ is small enough, a sample size of $N_{S,1}$ is unlikely when $S_0$ and $S_2$ are compared, contrary to what happens under the effect $S_0$ vs. $S_1$, for which $N_{S,1}$ is the expected size. We conclude, therefore, that both effects are different. Specifically,
following the standard criterion for statistical significance, we will declare two effect sizes to be significantly different whenever $p_S < 0.05$.

The similar effect probability allows not only to statistically contrast differences in the same problem, but also to deal with effect size comparisons between problems of different nature, provided that, as mentioned above, the corresponding similarity structures have been calculated by applying the same hypothesis test and significance level. For instance, in the previous example, the similar effect probability would not only assess the differences in the effects induced by the drugs tested but would also allow comparing the effect of both compounds with other solutions not necessarily pharmacological (surgical, other therapeutic approaches, lifestyle changes, etc.), as long as the units of the variables of interest are the same as well as the hypothesis test and the significance level for calculating the similarity structures.

![Figure 1](image)

**Figure 1.** Similarity structure, similarity size and similar effect probability. A. Three samples of size $2 \times 10^4$ are considered: $S_0 \sim N(0,1)$, $S_1 \sim N(0.05,1)$, and $S_2 \sim N(0.07,1)$. The similarity structures $p_{\alpha,1}(N)$ (blue curve) and $p_{\alpha,2}(N)$ (red curve) for the comparisons $S_0$ vs. $S_1$ and $S_0$ vs. $S_2$, respectively, are depicted with their corresponding similarity sizes $N_{S,1}$ and $N_{S,2}$. It shows how the larger the effect size (the value of the sample mean), the smaller the similarity size. B. The shaded areas give the probability of each similarity size under the other similarity structure, namely $p_{\alpha,1}(N_{S,2})$ (blue area) and $p_{\alpha,2}(N_{S,1})$ (red area). The smallest one is the similar effect probability $p_S$.

**Computation of the similarity structure**

*Parameter values.* The similarity structure requires to define the hypothesis test and the significance level required to properly address the analysis of data. However, the numerical calculation of the probabilities $\{p(> \alpha|N)\}_{N \in \{N_0, N_f, q\}}$ (Eq. (2)) critically depends on the parameters $N_0, N_f, q$, and $r$ to balance suitable estimations with practical computation times (see below). We briefly discuss the role of these parameters, their optimal values and their default values defined in the ready-to-use R code provided.

- $N_0$. The initial subsample size should be as small as possible according to the hypothesis test applied. The default value is $N_0 = 2$.

- $N_f$. The final subsample size should be such that $p(> \alpha|N)$ approaches zero at the final part of the interval $[N_0, N_f]$ (about its final third). This way, all relevant information about $p(> \alpha|N)$ is estimated numerically with performing no spurious calculations, since computation for sufficiently large sample sizes would not contribute to the estimation (the probabilities would be close to
zero). The default value is $N_f = N_{exp}$ although it is advised to estimate it by calculating a fast approach to $p(\alpha \mid N)$ with small values for $q$ and $r$.

- $q$. The number of sample sizes equidistant between $N_0$ and $N_f$ should be chosen to properly cover the shape of $p(\alpha \mid N)$. It should be considered that the differences between the analysed samples modulates the optimum value for $q$. If the samples are nearly equal, $p(\alpha \mid N)$ could hardly vary even for $N_f = N_{exp}$, so a large $q$ would imply spurious calculations. The default value is $q = 60$.

- $r$. The number of subsample pairs in each set $E_N$ (Eq. (1)) should be sufficiently large for an acceptable variability of the numerical data of the probability $p(\alpha \mid N)$. However, this variability depends on the nature of the data so, for the selection of the values for $N_f$, $r$, and $q$, a quick initial calculation of $p(\alpha \mid N)_{NE \in [N_0, N_f, q]}$ is recommended to estimate its variability using the default parameter values, and refine them if necessary.

**Fitting.** The fitting of the function in Eq. (8) to the data of the numerical estimation $p(\alpha \mid N)_{NE \in [N_0, N_f, q]}$ is done by adjusting the parameters in three approaches: $(a + bN)e^{-dN}, (a + bN + c N^2)e^{-dN}$, and $(a + bN + c N^2 + e N^3)e^{-dN}$. Each approach is adjusted by using a fixed set of initial parameter values. From the adjusted models, we discard those below zero, above one, and those with a local minimum.

Low variability of the numerical estimation $p(\alpha \mid N)_{NE \in [N_0, N_f, q]}$ of the probability $p(\alpha \mid N)$ improves the fitting and thus the accuracy of the similarity structure. However, this variability depends on the nature of the data so, for the selection of the values for $N_f$, $r$, and $q$, a quick initial calculation of $p(\alpha \mid N)_{NE \in [N_0, N_f, q]}$ is recommended to estimate its variability using the default parameter values, and refine them if necessary.

Fitting. The fitting of the function in Eq. (8) to the data of the numerical estimation $p(\alpha \mid N)_{NE \in [N_0, N_f, q]}$ is done by adjusting the parameters in three approaches: $(a + bN)e^{-dN}, (a + bN + c N^2)e^{-dN}$, and $(a + bN + c N^2 + e N^3)e^{-dN}$. Each approach is adjusted by using a fixed set of initial parameter values. From the adjusted models, we discard those below zero, above one, and those with a local minimum.

The calculation of the similarity structure comprises two main stages: the numerical estimation of $p(\alpha \mid N)$ and the fitting to obtain the similarity structure $p_a(N)$. The Monte Carlo simulation in the first stage is the most time-consuming process but it is suitable to be distributed. Thus, the ready-to-use R code is parallelized to speed up the calculations.

**Computation times.** The calculation of the similarity structure comprises two main stages: the numerical estimation of $p(\alpha \mid N)$ and the fitting to obtain the similarity structure $p_a(N)$. The Monte Carlo simulation in the first stage is the most time-consuming process but it is suitable to be distributed. Thus, the ready-to-use R code is parallelized to speed up the calculations.

We illustrate the computation times with the first case from the Results section. The similarity structure for samples $S_0$ vs. $S_m$, where $S_0 \sim N(0,1), S_m \sim N(m, 1)$, and $m \in \{0.1, 0.5, 1\}$ is calculated. Bilateral $t$-Student test with $\alpha = 0.05$ is applied, with $N_0 = 2, N_{exp} = 2 \times 10^4$, $q = 60$, and $r = 300$. The resulting computation times in a MacBook Pro, Apple M2 Pro, 16 Gb RAM, macOS Ventura 13.4.1 with eight cores were:
\[
m = 0.1, N_f = 8000 \rightarrow \text{cpu time} = 13.04 \pm 0.12 \text{ s}
\]
\[
m = 0.5, N_f = 300 \rightarrow \text{cpu time} = 8.05 \pm 0.09 \text{ s}
\]
\[
m = 1, N_f = 80 \rightarrow \text{cpu time} = 7.94 \pm 0.06 \text{ s}
\]

The same calculations in a PC, 16 Gb RAM, Windows 10 with 4 cores and 8 logical processors took about four seconds longer. The decrease in the computation time with \(m\) is due to the decrease in the size of the subsamples (small values of \(N_f\)) and, therefore, in the computational resources required.

**Scripts.** The ready-to-use R-scripts and data to compute the similarity structure are available in the OSF repository: [https://osf.io/udx7e/?view_only=1c2a97208bd94e2a8398c4864211693c](https://osf.io/udx7e/?view_only=1c2a97208bd94e2a8398c4864211693c). These scripts and data produce the result of the analysis described above. The figures are reproducible in Matlab scripts, that are available upon request from the authors.

**Results**

In order to illustrate the application of the similarity structure, the similarity size, and the similar effect probability to characterize the effect size in large samples, we consider four cases. The first one is focused on large normal samples with different means, to compare and link the proposed effect size characterization with Cohen’s \(d\), as one of the most well-known effect size measures. The second case presents large normal samples with different standard deviations, showing how to measure the effect size in general contexts through the similarity structure, as well as illustrating the interpretation of the effect size in terms of Cohen’s \(d\) even though it cannot be directly applied. The last two cases involve large experimental datasets and illustrate the application of the proposed approach to diverse real samples. They correspond to human gaze fixation during the presentation of dynamic stimuli and to distribution of protist populations in natural ecosystems. To illustrate the generality of the similarity structure, different hypothesis tests have been applied in each case, showing the convergence in the assessment of the effect sizes, their interpretation, and their comparison despite of the disparate nature of the data.

**Samples from normal distributions with different means**

The similarity structure is first applied to the comparison of two large samples coming from normal distributions. The reference sample \(S_0\) comes from a distribution \(N(0,1)\), whereas the other sample \(S_m \sim N(m, 1)\), where the parameter \(m\) varies between 0.05 and 1.2 at intervals of 0.05. All samples consist of \(2 \times 10^4\) data. To isolate the effect of the differences in the sample means from the random nature of the own samples, the samples \(S_m\) are defined as \(S_m = S_0 + m\), so each \(S_m\) is equal to \(S_0\) shifted by \(m\).

Our goal is to study how the similarity structure characterizes the difference of means between \(S_0\) and \(S_m\), i.e., the value of \(m\), focusing on the assessment of the effect sizes (according to \(m\)), their interpretation, and their comparison. Given the large sample size, the mean \(m\) is almost equal to Cohen’s \(d\) (differences in the third-fourth decimal place; R function ‘cohen.d’ in the package ‘effsize’ [30]). Thus, this case allows us to analyse the similarity structure as a measure of the effect size in comparison with one of the most extended effect size measures.

To validate the calculation of the similarity structure, we compare the numerical estimation of \(p(> \alpha \mid N)\) (obtained by using the t-test), its corresponding fitting, and the \(\beta\) curve (vs. the sample
size $N$), which can be theoretically obtained from the two-tailed t-Student test to analyse the difference between two normal samples. Figure 2 illustrate this by showing two sample comparisons: $S_0$ vs. $S_{0.25}$ and $S_0$ vs. $S_{0.35}$. In both cases, the numerical estimation of $p(> \alpha \mid N)$ (i.e., $\{p(> \alpha \mid N)\}_{N \in \{N_{0.25}, N_{0.35}\}}$) is depicted by the blue line, with the corresponding fitting (Eq. (8)) in red and $\beta$ curve in green. To confirm the good fit suggested by these two comparisons, the inset shows the gof for the set of comparisons $\{S_0$ vs. $S_m\}$ with $m \in \{0.05, 0.1, \ldots, 1.15, 1.2\}$, verifying that the gof oscillates between 1% and 2% of the maximum of $p(> \alpha \mid N)$.

**Figure 2.** Accuracy of the estimation of the probability $p(> \alpha \mid N)$. A. Comparison of the numerical estimation of the probability $p(> \alpha \mid N)$ that two subsamples of size $N$ are similar (two-tailed t-Student test applied; blue line) with its corresponding fitting (Eq. (8); red curve) and the theoretical $\beta$ curve (1-power with the two-tailed t-Student test; green curve). The depicted curves correspond to pairs of subsamples of size $N$ coming from large normal samples (size = $2 \times 10^4$): $S_0$ vs. $S_{0.25}$ (flattest curves) and $S_0$ vs. $S_{0.35}$ (steepest curves). B. Goodness of fit of $p(> \alpha \mid N)$ for the comparison of subsamples of size $N$ coming from large normal samples (size = $2 \times 10^4$) $S_0$ vs. $S_m$, with $m \in \{0.05, 0.1, \ldots, 1.15, 1.2\}$.

**Assessment of the effect size.** The similarity structure of samples $S_0$ vs. $S_m$ according to the difference of their means $m$ is shown in Figure 3.A. The main panel shows the similarity structures $p_{\alpha}(N)$ obtained from the two-tailed t-Student test. The inset depicts the curves $p_{\alpha}(N)$ when the two-sided Kolmogorov-Smirnov test is applied ($\alpha = 0.05$ in both cases). As expected, the higher the value of $m$, the steeper the curve and the smaller the probability $p_{\alpha}(N)$ of similar subsamples of size $N$. The assessment of the effect sizes through the similarity size is illustrated in Figure 3.B, showing the relationship between $N_S$ and the mean $m \in \{0.05, 0.1, \ldots, 1.15, 1.2\}$ for t-Student test (circles) and Kolmogorov-Smirnov test (squares). As commented above, $m$ is almost identical to Cohen’s $d$, which is encoded by the curve colour. Therefore, the obtained $N_S$ characterize the effect size of comparing two large normal samples as the expected size of similar subsamples (e.g., for the mean difference $m = 0.5$, two similar subsamples will have an expected size of 35). As expected, the similarity sizes obtained from applying the t-Student test are smaller than those obtained by applying the Kolmogorov-Smirnov test, as expected from the lower statistical power of nonparametric tests.
Figure 3. Comparison of normal samples with different means and equal standard deviations based on their similarity structure. A. Similarity structure of two normal samples $S_0$ and $S_m$, with $m \in \{0.05, 0.1, ..., 1.15, 1.2\}$ according to the difference of their means $m$ (colored scale) obtained by applying the t-Student test (Main panel) and Kolmogorov-Smirnov test (inset), respectively. B. Similarity size $N_S$ vs. $m$ colored according to the corresponding Cohen’s $d$ (circles/squares for the t-Student/Kolmogorov-Smirnov test, respectively).

**Interpretation of the effect size.** To enhance the interpretability of the similarity size, we study the relationship between $N_S$ and the sample size required for a statistical power of 0.9, which will be denoted as $N_1^*$. Thus, in terms of $p(> \alpha \mid N)$, $N_1^*$ will satisfy that $p(> \alpha \mid N_1^*) = 0.1$. Figure 4 shows $N_1^*$ vs. $N_S \in \{N_S \mid m \in \{0.05, 0.1, ..., 1.15, 1.2\}\}$, where this set contains the $N_S$ obtained from the comparison of normal samples $S_0$ vs. $S_m$ with $m \in \{0.05, 0.1, ..., 1.15, 1.2\}$ (Fig. 3B). The $N_S$ and $p(> \alpha \mid N)$ to obtain $N_1^*$ were calculated by applying the t-Student and Kolmogorov-Smirnov tests (Fig. 4 panels A and B, respectively). The corresponding linear models estimating $N_1^*$ from $N_S$ (red lines and red equations) propose the simple approximation $N_1^* \approx 2.5 N_S$, with the error below 12%-8% (insets in Fig. 4; percentage calculated on the exact value of $N_1^*$). The largest errors are for small values of $N_S$, where the intercept of the linear relationship, not considered in the approximation, is not negligible. In any case, when the sample size given by the linear model deviates from $N_1^*$, it keeps the power between 0.8 and 0.9 (black lines for $N_2^*$ such that $p(> \alpha \mid N_2^*) = 0.2$). Consequently, a practical interpretation of $N_S$ is that 2.5 times its value is approximately the sample size required to obtain a statistical power of 0.9.
Figure 4. Interpreting the similarity size as sample size required for attaining a statistical power of 0.9. A. \( N_1^* \) denotes the sample size such that \( p(> \alpha \mid N_1^*) = 0.1 \), related with a power = 0.9. The values of \( N_1^* \) (blue circles) were obtained by numerically estimating the probabilities \( p(> \alpha \mid N) \) by applying the t-Student test to the similarity structure of two large normal samples \( \{S_0 \text{ vs. } S_m \text{ with } m \in \{0.05, 0.1, ..., 1.15, 1.2\}\}. The linear model adjusted to \( N_1^* \) vs. \( N_S \) (red line) is detailed (red equation), with the sample sizes \( N_2^* \) corresponding to a power level of 0.8 \( p(> \alpha \mid N_2^*) = 0.2 \); black line. The error of estimating \( N_1^* \) as \( 2.5 N_S \) is shown in the inset. B. Same as A but applying the Kolmogorov-Smirnov test.

On the other hand, the relationship between the similarity size \( N_S \) and \( m \) in Fig. 3.B opens an alternative venue to interpret the effect size given by \( N_S \) in terms of Cohen’s \( d \) regardless of whether the standardized mean can be directly applied. The similarity structure only depends on the applied statistical test and the significance level \( \alpha \). Hence, two pairs of compared samples exhibiting closely related similarity structures obtained from the same statistical test and equal \( \alpha \), will have almost the same effect size in terms of \( N_S \), no matter their disparate nature and statistical structure. Therefore, let us consider normal samples of different means and data samples whose structure does not allow to apply \( d \), and assume that, in both cases, their similarity structures have been obtained by applying the same test and \( \alpha \) value. Since similar \( N_S \) values indicate similar effect sizes and for the normal samples the effect size can be described by the Cohen’s \( d \), then the effect size for the later data samples can also be interpreted in terms of \( d \) through \( N_S \).

To elaborate this idea, we focus on the relationship shown in Fig. 3.B between \( d \) and the similarity size \( N_S \) of normal samples of different means, obtained from the Kolmogorov-Smirnov test \( (\alpha = 0.05) \). This dependency is illustrated in Fig. 5 and can be modelled as:

\[
N_S = 1 + \frac{12}{d_{1.95}}. \tag{19}
\]

Thus, the inverse of this function leads to express the standardized mean \( d \) in terms of the similarity size \( N_S \):

\[
d = \frac{3.58}{(N_S - 1)^{1.95}}. \tag{20}
\]

Therefore, regardless of whether the assumptions to apply \( d \) are not met (normality and homoscedasticity), if two samples are compared through the similarity structure obtained by applying the Kolmogorov-Smirnov test and \( \alpha = 0.05 \), the effect size could be interpreted in terms of \( d \) by introducing the corresponding \( N_S \) in Eq. (20). This will be exploited in the next case of study.
Comparison of effect sizes. The similar effect probability $p_S$ allows to compare two effect sizes $N_S$, assessing how likely is to get one of the effect sizes in the statistical conditions of the other one. This is shown in Fig. 6.A as a matrix where each element contains the effect size comparison between $S_0$ vs. $S_{m_1}$ and $S_0$ vs. $S_{m_2}$ (with $m_1, m_2 \in \{0.05, 0.1, ..., 1.15, 1.2\}$). The lower/upper triangular part of the matrix corresponds to the similar effect probability $p_S$ obtained from the t-Student/Kolmogorov-Smirnov test (the diagonal contains the averaged $p_S$ for both tests). The dark blue regions correspond to those effect sizes significantly different ($p_S < 0.05$); for instance, $S_0$ vs. $S_{0.25}$ will be significantly different from $S_0$ vs. $S_m$ when $m \leq 0.15$.

On the other hand, since $m$ and $d$ are almost equal, the similar effect probability reveals a connection with Cohen’s $d$ that offers a justification of the widely used equivalence between its value and the interpretation of the effect size. The descriptors of the effect size based on the value of the standardized difference $d$ were initially suggested by Cohen [15] and extended by Sawilowsky [31] as a rule of thumb to characterize the effect size in a comprehensible way. Thus, $d \sim 0.2$ corresponds to a small effect size, $d \sim 0.5$ to a medium effect size, $d \sim 0.8$ to a large effect size, and $d \sim 1.2$ to a very large effect size. These correspondences imply that there should be relevant differences between the consecutive grid $d$ values 0.2-0.5, 0.5-0.8, and 0.8-1.2, differences that are explained by the similar effect probability. Figure 6.B offers a simplified scheme of the matrix for $p_S$ by depicting in blue the area in Fig. 6.A with $p_S < 0.05$, i.e., that corresponding to significant differences between the effect sizes characterized by $N_S$ for different pairs of $d$ (i.e., pairs of $m$). Thus, when the effect size $N_S$ for $d = 0.5$ is compared with those of smaller $d$ by means of the similar effect probability, the effect sizes are significantly different for $d \leq 0.25$. This is consistent with the above-mentioned distinction between $d \sim 0.2$ and $d \sim 0.5$ for small and medium effect sizes. Similarly, significant effect size differences are found between $d \leq 0.45$ and $d = 0.8$, and between $d \leq 0.7$ and $d = 1.2$, which are aligned with the separation between $d \sim 0.5$ and $d \sim 0.8$ for medium and large effect sizes, and between $d \sim 0.8$ and $d \sim 1.2$ for large and very large effect sizes, respectively. Therefore, the similar effect probability identifies and, more importantly, justifies widely extended relationships to characterize effect sizes. In any case, it should be remarked that the effect size must be interpreted in the specific scientific context, so these ‘one-for-all’ rules of thumb should be used with caution and always as a rough approach to the understanding of the results.
Figure 6. Similar effect probability for the comparison of normal samples with different means and equal standard deviations. A. Matrix with the similar effect probability $p_S$ for $0.1 \leq m \leq 1.2$. The triangular lower and upper parts correspond to the $p_S$ obtained from applying the $t$-Student and Kolmogorov-Smirnov tests, respectively. Dark blue areas correspond to $p_S < 0.05$ (significant differences between the effect sizes). B. Scheme of the lower triangular part of the matrix in A in terms of the standardized difference $d$. The blue area illustrates those effect size comparisons with $p_S < 0.05$, i.e., those effect sizes significantly different for the $t$-Student test. Light vertical bars represent the classical Cohen’s effect size classification. For instance, the height of the first vertical bar denotes the values $d \leq 0.2$ for $d = 0.5$ (in the horizontal axis), which means the distinction between $d \sim 0.2$ and $d \sim 0.5$ for characterizing small and medium effect sizes, respectively. The other bars denote the separation between $d \sim 0.5$ and $d \sim 0.8$ for medium and large effect sizes, and $d \sim 0.8$ and $d \sim 1.2$ for large and very large effect sizes. The bars match with the region $p_S < 0.05$, so those significantly different effect sizes are coherent with Cohen’s classification.

Samples from normal distributions with different standard deviations

This case addresses the comparison between normal samples with different standard deviations through their similarity structure. Following the previous case, setting the mean to zero and tuning the standard deviation $std$, we change the nature of the samples while maintaining the simplicity of the assumptions. This example shows that: 1) the effect size associated to differences in the standard deviation can be interpreted in terms of the sample size $N_s$, 2) the effect size can be interpreted in terms of Cohen’s $d$, although this measure is not directly applicable in this case, 3) $N_s$ estimates the sample size $N_s^*$, and 4) the effect sizes are also characterized by their differences, obtained thought the similar effect probability.

The samples to be compared are denoted as $S_1 \sim N(0,1)$ and $S_{std} \sim N(0, std)$ and are made up of $2 \times 10^4$ data. The parameter $std$ is the standard deviation, controlling the difference between $S_1$ and $S_{std}$, and it varies between 1 and 2 in intervals of 0.05. As in the previous case of normal samples of different means, the potential differences between $S_1$ and $S_{std}$ due to their random generation are mitigated by obtaining $S_{std}$ for each $std$ value as $S_{std} = std \cdot S_1$.

The similarity structures of the comparisons $S_1$ vs. $S_{std}$ with $1 \leq std \leq 2$ were calculated by applying the Kolmogorov-Smirnov test with $\alpha = 0.05$ (Fig. 7.A). To characterize the effect size from the similarity structures, the similarity size $N_s$ is obtained for each $std$ value (Fig. 7.B). Since they have been calculated in the same conditions as the $N_s$ in Fig. 5, the effect size when comparing two large normal samples with different standard deviations can be interpreted in terms of the standardized mean $d$ via Eq. (20), encoded in the symbol colours in Fig. 7.B. In particular, $std \sim 1.5$ and $std \sim 2$ can be assimilated to $d \sim 0.2$ and $d \sim 0.5$, so they can be roughly interpreted as small and medium effect sizes, respectively.
The previous example shown that the effect size given by $N_S$ can be interpreted in more practical terms as an estimator of the sample size required for a power of 0.9. Figure 7.C illustrates the same for the present case, showing the relationship between $N_S$ and $N_1^*$ (blue circles) and the linear model (red line and red equation; the black line corresponds to $N_2^*$ vs. $N_S$). From the coefficient of the linear model and following the estimation in the previous case, we approximate $N_1^* \approx 2.5 N_S$, whose error, around 4%, is depicted in the inset. This result suggests that this estimation can be a general rule to interpret the similarity size, which will be confirmed in the next cases.

Finally, the similar effect probability is shown, for the sake of clarity, as a symmetric matrix (Fig. 7.D), and characterizes the effect sizes through their differences. Thus, for instance, the effect size when comparing $S_1$ vs. $S_2$ is significantly different from the effect sizes when comparing $S_1$ vs. $S_{std}$ with $1 \leq \text{std} \leq 1.5$. From the previous considerations we consistently conclude, once again, that small and medium effect sizes are significantly different.

**Figure 7.** Similarity structure analysis of normal samples with equal means and different standard deviations.  
A. Similarity structure of two normal samples $S_1 \sim N(0,1)$ and $S_{std} \sim N(0,\text{std})$, where $1 \leq \text{std} \leq 2$, from applying the Kolmogorov-Smirnov test and with $\alpha = 0.05$. B. Similarity size $N_S$ vs. $\text{std}$. The colours indicate the corresponding $d$ values according to Eq. (20). C. Estimation of $N_1^*$ as 2.5 $N_S$ (blue circles for $N_1^*$ vs. $N_S$) suggested by the linear model (red line and red equation) and from the previous example. Inset: estimation error. D. Similar effect probability $p_S$. For the sake of clarity, the triangular upper and lower parts are equal. Dark blue areas correspond to $p_S < 0.05$ (significant differences between the effect sizes).

**Gaze fixation during visual dynamic stimuli**

This example illustrates the analysis of large real experimental samples. The experiment involves eye tracking of moving visual stimuli consisting of two spheres moving forward and backward in collision trajectories and disappearing before collision. A random sequence of 16 stimuli with the
spheres randomly oriented (8 moving forward and 8 moving backward), was displayed to 17 women and 17 men, obtaining the corresponding gaze fixation coordinates (see Supplementary for details). After transforming such coordinates into a common reference system, the eye-tracking data can be displayed as shown in Fig. 8.A, where the contour maps represent the spatial distribution of the gaze fixations, and the red circles depict the sequence of normalized positions of the spheres during their movement.

The data are analysed by considering two factors of interest: participant’s gender and stimulus type. According to the participant’s gender, the stimuli are denoted by \( w \) and \( m \) for women and men, respectively. Stimulus type refers to whether the spheres move forward or backward in the direction of the collision point and will be denoted by the subindexes \( f \) and \( b \), respectively. Thus, the gaze fixation distributions for women (men) when forward and backward stimuli were displayed will be denoted by \( w_f(m_f) \) and \( w_b(m_b) \), respectively. According to these two factors, the spatial distributions of the gaze fixations have been compared two-by-two keeping constant one factor, either gender or stimulus type (black double arrows in Fig. 8.A). The four samples have between 1200 and 1400 gaze fixation 2D coordinates.

The statistical comparisons between the experimental samples and among their corresponding subsamples have been obtained by applying the bidimensional Kolmogorov-Smirnov test [32] (the algorithm is implemented in R with the ‘fasano.franceschini.test’ package [33]). The large sizes of the datasets provide statistical significance for the four comparisons, with \( p \)-values around 0.003 and 0.006 (see Supplementary), even though the two-dimensional distributions in Fig. 8.A suggest that there should be differences between stimulus type and gender factors. Besides, the characterization of eventual differences through the assessment of the effect size is also an issue in this case. The evaluation of the differences of means would be dependent on external factors such as screen resolution, which can make the results interpretable but hardly generalizable. Besides, the dispersion of the fixation points could be an aspect to be analysed but its apparent variation in horizontal and vertical directions suggest a violation of the homoscedasticity, making difficult to weight the difference of means by the variance. This points out the difficulty to apply effect size measures directly to 2D distributions, beyond indirect strategies such as those based on dimensionality reduction. However, the similarity structure allows to characterize and assess differences in these two-dimensional distributions by means of the same procedure used for one-dimensional data, just applying the proper nuclear statistical test for identifying differences between random bidimensional subsamples.

The similarity structures of the relevant sample comparison (default values for the parameters) indicate that the type of stimulus (forward or backward) is the primary factor affecting gaze fixation, whereas participant’s gender has considerably less effect (Fig. 8.B). This is confirmed by the similarity effect (Fig. 8.C), which shows that \( N^*_S \) for \( w_f - w_b \) and \( m_f - m_b \) are one order of magnitude smaller than the similarity sizes for \( w_f - m_f \) and \( w_b - m_b \). Specifically, when considering different participant’s gender and the same stimulus type, the similar samples are expected to be composed of hundreds of gaze fixations. On the contrary, much smaller similar samples, composed of tens of gaze fixation coordinates, are expected when considering different types of stimuli for the same gender. In addition, the sample size \( N^*_S \) necessary to reach a power equal to 0.9 can also be described by a linear model depending on \( N^*_S \) like those of the previous examples (Fig. 8.D). Hence, the approximation \( N^*_S \approx 2.5 N^*_S \) has an error below 4% (inset in Fig. 8.D).

These conclusions about the effect size of the considered factors are statistically supported by the similar effect probability (Fig. 8.E). On the one hand, the effect size of the stimulus type is significantly larger than the effect size of the gender (\( p_S < 0.05 \) when \( w_f - w_b \) and \( m_f - m_b \) are
compared with $w_f - m_f$ and $w_b - m_b$). On the other hand, there are no differences when comparing the effect size of the stimulus type between women and men ($p_S \sim 0.1$ when $w_f - w_b$ and $m_f - m_b$ are compared) and the effect size of the gender between forward and backward stimuli ($p_S \sim 0.3$ when $w_f - m_f$ and $w_b - m_b$ are compared). Therefore, the use of the similarity structure allows us to make a dichotomic decision on the factors analysed that could not be made by orthodox $p$-based inference, concluding that the immediate visual processing of simple dynamic stimuli represented by the gaze fixation, shows no relevant differences between genders but is highly dependent on the type of stimulus.

![Figure 8](image)

**Figure 8.** Similarity structure analysis of gaze fixations. A. Gaze fixation has been recorded according to two factors: type of stimulus ($f$ and $b$ for the stimuli in which the displayed spheres move forward and backward, respectively) and gender ($w$ and $m$ for women and men, respectively). The gaze fixation distributions (contour density plots) are represented together with the normalized spheres’ positions (red circles) and denoted by the corresponding gender with a subindex for the stimulus type (e.g., $w_f$ stands for gaze fixation for women when the forward stimuli were displayed). These distributions were analysed keeping constant one factor (either stimulus type or gender; black double arrows). B. Similarity structure of the pairwise comparisons of gaze fixation distributions in A (for clarity, the square root of $p_S(N)$ is shown). C. Similarity sizes $N_S$ quantifying the effect size for each pair-wise comparison. D. Linear estimation of $N^*_1$ from $N_S$. The inset shows the error of the approximation $N^*_1 \approx 2.5N_S$. E. Similar effect probability $p_S$. For the sake of clarity, the triangular upper and lower parts are equal. Dark blue areas correspond to $p_S < 0.05$ (significant differences between the effect sizes).

**Protist populations in natural ecosystems**

Similarity structure is applied now to explore multidimensional data consisting of number of individuals belonging to multiple protist species collected in different natural conditions. The similarity structure is then applied to reveal the influence of such conditions on the occurrence of individuals in the reported species. The samples were collected from nineteen rock basins located in a natural preserved area differing in their distance to a reference location and in their volume [34]. After being extracted from the basin sediment, the samples were grouped according to their DNA sequence similarity into Operational Taxonomic Units, OTUs. These clusters can be assimilated to species with some caution and the number of reads for each OTU can be assimilated to number of individuals. Thus, for each sample, the data consist of the OTUs and the corresponding reads.
The samples were collected in twenty basins under two factors whose influence on the protist diversity is studied: proximity of the basin to a reference location and volume of the basin. For each factor, four classes of basins are considered. Regarding proximity, those basins close to the reference less and more than 300 m are denoted by C (closer) and D (distant), respectively. Regarding basin volume, those with a volume smaller and larger than 3000 cm$^3$ are denoted by subindexes v and V, respectively. Thus, the basins are grouped in the four classes C$_v$, D$_v$, C$_V$, and D$_V$, where, for instance, C$_V$ denotes the set of basins distant to the reference less than 300 m and with a volume > 3000 cm$^3$. Finally, each class is represented by a single sample obtained by averaging the reads for each OTU in the class samples. These final samples will also be denoted by C$_v$, D$_v$, C$_V$, and D$_V$, containing 78 OTUs with 57922, 61861, 46877, and 47224 total counts, respectively (see Supplementary for details).

In order to analyse the influence of proximity and volume of basins over the species diversity, four comparisons were performed: C$_v$ vs. C$_V$ and D$_v$ vs. D$_V$ (C$_v$ vs. D$_v$ and C$_V$ vs. D$_V$) will allow to assess the effect of the basin volume (basin location) according to their location (volume). These pairs of datasets are depicted in Fig. 9.A. The two-by-two sample comparisons by means of the Poisson regression give a $p$-value of 0 as expected by the large sample sizes. Therefore, the potential differences between the samples should be characterized through the corresponding effect sizes. One of the most used measures to characterize the similarity of protist populations is the Bray-Curtis distance, so the larger the distance, the more different the populations are. The Bray-Curtis distances for the comparisons are: 0.58 for C$_v$ vs. C$_V$, 0.6 for C$_v$ vs. D$_v$, 0.58 for C$_V$ vs. D$_v$, and 0.43 for D$_v$ vs. D$_V$. Since this distance is bounded between 0 and 1, the largest difference between these distances is 15%, but the remaining differences are 2% and 0%. Thus, despite its popularity, this effect size measure presents disadvantages, such as being a dimensionless distance, which makes difficult to interpret its value in practical terms and can show low sensitivity to discriminate differences.

The similarity structure is then applied to characterize the effect size mitigating these limitations. The similarity structures of the four comparisons were obtained by applying the Poisson regression, illustrating the flexibility and generality of the method ($N_0 = 5$, $N_f = 130, 300, 200,$ and $200$ for C$_v$ vs. C$_V$ and D$_v$ vs. D$_V$, C$_v$ vs. D$_v$, and C$_V$ vs. D$_v$, respectively; default values for the remaining parameters). The similarity structure, i.e., the $p_N (> \alpha)$ vs. $N$ curves shown in Fig. 9.B, suggest that protists diversity is more affected by changes in the volume of those basins close to the reference location than in the distant ones. On the other hand, the variation in population diversity due to spatial proximity seems to be more prominent in those basins with high volume. To verify these interpretations, we quantified the effect sizes by means of the similarity sizes $N_S$. Figure 9.C shows the expected number of observations in similar subsamples for each of the four comparisons, which, as in the previous examples, accurately estimates the sample size $N^*_1$ required for a power of 0.9 as $N^*_1 \cong 2.5 N_S$ (Fig. 9.D). Both measures, $N_S$ and $N^*_1$, provide transparent and practical interpretation of the effect size, in contrast with the Bray-Curtis distance.

Finally, the relative magnitude of the reported effect sizes is quantified by the similar effect probability in Fig. 9.E, confirming the importance of the volume as factor influencing the protists diversity in nearby basins compared with the distant ones. Note that the similar effect probability for this case is $p_S = 0.079$, i.e., though the significance threshold here proposed ($p_S = 0.05$) is not reached, when comparing C$_v$ vs. C$_V$, there is a probability of 7.9% of getting similar samples of size lower or equal to the expected size (the similarity size $N_S$) obtained when comparing D$_v$ vs. D$_V$. At this point, it will be the researcher who should judge if this probability is low enough or more data should be collected to refine the experiment. This example illustrates how the similarity structure can help researchers address decisions in real-life problems beyond the significant/non-significant
dichotomic resolutions, in line of the approaches proposed by different authors to move statistics beyond the p-value [4].

**Figure 9.** A. Similarity structure analysis of protist populations. A. Graphical comparison of the analyzed samples grouped according to the factors under study. D_v and D_V (dark and pale blue, respectively) correspond to the averaged samples collected from distant basins of small and large volume, respectively. C_v and C_V (dark and pale red, respectively) are the average of the samples collected from nearby basins of small and large volume, respectively. C_v and D_v (dark and pale light blue, respectively) correspond to the averaged samples gathered from close and distant small volume basins, respectively. The same stands for C_V and D_V (dark and pale orange, respectively) when the basin volume is large. For the sake of clarity, only OTUs with counts > 450, 250, 300, and 300 (respectively) in both samples are displayed. B. Similarity structure of the pairwise sample comparisons detailed in A. C. Similarity size N_s quantifying the effect size in terms of the expected number of OTUs in similar subsamples. D. Linear estimation of N^*_1 from N_s. The inset depicts the error of the approximation N^*_1 \cong 2.5N_s. E. Similar effect probability p_S. To make it clearer, the triangular upper and lower parts are equal. Dark blue areas correspond to p_S < 0.08.

**Discussion**

For decades, the predominant role of the p-value in statistics has been questioned. Even though its use was originally proposed by Fisher as a first step in the inferential analysis of data [9], its attractive features have led to its misuse based on the misunderstanding of its limitations [35]. Many different approaches have been proposed to complement and even replace the p-value as a key to objectively conclude about the existence of effects [4]. The similarity structure here proposed aligns with these alternative ideas and provides complementary methods to conduct analysis of large samples in a single approach. Thus, the similarity structure provides 1) a qualitative description of the structure of the differences between the analysed datasets (p_N (> \alpha) vs. N curve), 2) a transparent quantification of the effect size by the signification size N_S in terms of number of individuals (observations, participants, etc.), 3) a versatile interpretation of the effect size as the standardized mean d and the sample size N^*_1 for a statistical power of 0.9, and 4) a dichotomic decision to conclude whether diverse effect sizes are significantly different.

The p-value is a probability that measures how likely the observations are compatible with the lack of effect (i.e., with the null hypothesis). Thus, as discussed previously, even in the absence of effect, the p-value could eventually be smaller than \alpha by chance, wrongly suggesting statistical significance.
This situation is particularly serious when large samples are studied and multiple hypothesis tests are performed, e.g., when many factors are analysed, as the misuse of the inferential tools could lead to conclude about the meaningfulness of spurious effects obtained from randomly significant \( p \)-values. Whereas in \( p \)-based methods a correction is usually applied to mitigate this effect [36], the similarity structure naturally adopts this \( p \)-value particularity by embedding it into the probability \( p_N(>\alpha) \). As such a probability is assessed from the binary decision about the similarity of two subsamples and not from the \( p \)-value itself, this encapsulates the issue of the variability of \( p \) and the significance threshold \( \alpha \) (type I error) is reflected in the probability \( p_N(>\alpha) \). This encapsulation of the inference provides the method with enough flexibility to easily introduce the proper hypothesis test according to the data requirements. In this sense, the similarity structure is assumption-free, as Kolmogorov-Smirnov or any other non-parametric test can be applied when comparing the subsamples if no assumptions about the structure of the datasets are made. Nevertheless, in the case where the populations are assumed to fulfil certain conditions regarding their distributions (normality, homoscedasticity, etc.), more statistically powerful (parametric) tests can be straightforwardly introduced to obtain the similarity structure (first example in Results).

The similarity structure, i.e., the dependence of the probability \( p_N(>\alpha) \) on \( N \), makes it possible to quantify the sample differences and the corresponding effect size by assessing the similarity size \( N_S \), which allows to rank the factors under study according to their effect in terms of observations/population. The similarity size is the expected size of the similar subsamples, or, in other words, if the size of two subsamples (each from one of the samples under study) is much larger than \( N_S \), the probability that these subsamples are different is high. That leads to the possibility of using \( N_S \) to estimate the sample size \( N_*^0 \) necessary to attain a statistical power equal to 0.9. The cases here studied indicate that the simple estimation \( N_*^0 \cong 2.5 \times N_S \) can be a general approximation regardless of the nature of the datasets, their structure, the hypothesis test applied, etc., as long as their similarity structure covers sufficiently small sample sizes \( N \) (see the last example in Results). Although future work is needed to rigorously prove whether this approximation holds for any similarity structure, our work suggests an interesting connection that may be leveraged to strengthen the practical interpretation of the similarity size.

In this sense, the similarity size allows to interpret the effect size in terms of the Cohen’s \( d \) in cases where \( d \) is not directly applicable because of the structure of the samples (they are not normal) or the effect size does not rely on the difference of means (as in the second example, where normal samples of different standard deviations are compared). This translation requires a common hypothesis test to construct 1) the similarity structures of the normal samples with different means to provide the relationship between \( N_S \) and \( d \), and 2) the similarity structure of the large samples to be analysed, whose \( N_S \) is introduced in the previous relationship to get the effect size of the studied samples as the standardized mean \( d \). In this work we have shown how to interpret \( N_S \) as \( d \) through the Kolmogorov-Smirnov test but the same can be done with any other hypothesis test applicable to the comparison of normal samples. In any case, this connection provides a characterization of the effect size in terms of a widely used measure, offering a practical tool for non-experts to get a rough but familiar approach to the magnitude of the differences. However, it should not be forgotten that, as the Cohen’s \( d \), whose interpretation is sometimes questioned [13], these rules of thumb should be handled with care since the final evaluation of the importance of the effect must be done in the context of the problem.

The difficulty to interpret an effect size could often be overcome by comparing it with a reference effect size whose interpretation is accepted. This possibility is provided by the similar effect probability \( p_S \) when the effect size is characterized by the similarity size \( N_S \). Hence, consider a novel and a reference effect size assessed by their corresponding similarity sizes \( N_S \) and \( N_{S,0} \), respectively,
where the reference effect size is interpreted as ‘medium’ from previous studies, literature, scientific evidence, etc. Imagine that the novel effect size is significantly larger than the reference one, i.e., $N_S < N_{S,0}$ and $p(N_{S,0}, N_S) < 0.05$. Then, it would be reasonable to conclude that the novel effect size could be classified as ‘large’ or even ‘very large’. In other words, the significance in the difference between effect sizes can be used as criterion to delimit the magnitude of the effect.

In addition, the generality of the similarity structure allows the comparison of effect sizes from data of disparate nature, enabling synergic decisions to be made. For instance, imagine two drugs with distinct therapeutical targets: hypertension and tremors. Studies supporting the eventual benefits of both therapeutic strategies would report intra-effect sizes, e.g., effects in terms of blood pressure units and inertial measure units, respectively, which would make it difficult to compare these treatments from a general perspective (social impact, investment/profit balance, etc.). The reporting of the effect sizes in terms of the significance size, however, stands the impact of both therapies on common ground and allows comparison of both scenarios (if the same hypothesis test and significance level have been used in the corresponding similarity structures). Thus, if the studies of these hypertension and tremor treatments provided significance sizes $N_S = 150$ and $N_S = 50$, respectively (in number of individuals) and the effect size probability was significant, we would conclude that the number of patients affected by tremors positively impacted by the therapy would be expected to be greater than the hypertensive patients benefiting from the reducing blood pressure drug, with the similar effect probability providing the statistical significance of this difference. Therefore, this kind of quantitative information together and the similar effect probability, complementary to the usual intra-effect sizes (quantitative reduction of blood pressure and tremors) can contribute to extend the conclusions of studies beyond their scientific scope, supporting more general decisions, as in this example, concerning public and private health policies, corporate strategies, clinical / commercial viability of treatments, etc.

The similarity structure is aligned with bootstrapping-based methods, such as MRPP (multi-response permutation procedures) [37], as it works by extracting information from ‘pieces’ of the available data. However, these methods aim to obtain a single $p$-value, so the limitations discussed above concerning large samples are also applicable. The similarity structure is a computationally intense method; however, it has the advantage of being a user-friendly tool since no expert knowledge is required to use it. On the one hand, the method is highly and easily parallelizable, as the probability $p_N(>\alpha)$ can be calculated independently for the different values of $N$. On the other hand, the method can be straightforwardly customized, just by adapting the hypothesis test to calculate the probability $p(>\alpha|N)$ to the particular problem. In the ready-to-use scripts provided to calculate the similarity structure, this tuning is implemented by properly changing a single module, which gives the method great flexibility.

In summary, this work proposes the similarity structure as a novel method to interrogate experimental dataset, not by assessing their differences with a single $p$-value but quantifying the effect size through the structure of their similarities as their size increases. This approach is agnostic to data nature, dimensionality, and the choice of hypothesis test, and its suitability is particularly remarkable on large samples, where the dataset sizes could lead to the acceptance of non-relevant effects based on artificially small $p$-values. The main features of the similarity structure are its four dimensions for the analysis and interpretation of the results: 1) the qualitative description of the sample differences, 2) the quantitative assessment of the effect size in terms of number of individuals as general, transparent, and easily interpretable units, 3) its interpretability in terms of the sample size for a power equal to 0.9 and widely used effect size measures, and 4) the possibility of dichotomically characterize the findings by statistically comparing effect sizes. Thus, the similarity structure follows the spirit of the ASA statement on $p$-values [35] and alternative proposals to
address its misuse: contextualizing the interpretation of $p$ [38], making the application of the statistical procedures more intuitive and transparent [39], and introducing novel approaches to complement and even replace the hypothesis testing [4]. We hope the similarity structure could contribute to the growing evidence, ideas, and procedures proposing a shift in the statistical analysis with the aim of doing more transparent and reproducible science.

**Acknowledgments.** We thank A. Murciano-Cespedosa for fruitful discussions and M. Martin-Cereceda for her support with the experimental data.

**Conflict of interest:** None declared.

**Funding.** This work has been supported by the National Project of the Ministry of Science and Innovation, Spain, Grant PID2022-138659NB-I00.

**References**


[2] Pearson K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 5:157–175.


[39] Kennedy-Shaffer L. (2019), Before p < 0.05 to Beyond p < 0.05: Using History to Contextualize p-Values and Significance Testing. The American Statistician, 73.