Frontiers in Neuroendocrinology 73 (2024) 101133

Contents lists available at ScienceDirect

Frontiers in Neuroendocrinology

journal homepage: www.elsevier.com/locate/yfrne

Review article

From means to meaning in the study of sex/gender differences and similarities

Carla Sanchis-Segura [a,*], Rand R. Wilcox [b]

[a] Departament de Psicologia bàsica, Clinica i Psicobiologia, Universitat Jaume I, Castelló, Spain
[b] Department of Psychology, University of Southern California, Los Angeles, USA

ARTICLE INFO

Keywords:
Sex
Gender
Means
Robust statistics
Precision medicine

ABSTRACT

The incorporation of sex and gender (S/G) related factors is commonly acknowledged as a necessary step to advance towards more personalized diagnoses and treatments for somatic, psychiatric, and neurological diseases. Until now, most attempts to integrate S/G-related factors have been reduced to identifying average differences between females and males in behavioral/ biological variables. The present commentary questions this traditional approach by highlighting three main sets of limitations: 1) Issues stemming from the use of classic parametric methods to compare means; 2) challenges related to the ability of means to accurately represent the data within groups and differences between groups; 3) mean comparisons impose a results' binarization and a binary theoretical framework that precludes advancing towards precision medicine. Alternative methods free of these limitations are also discussed. We hope these arguments will contribute to reflecting on how research on S/G factors is conducted and could be improved.

1. Introduction

> *«It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail».*
> Maslow, AH. *The Psychology of science: A reconnaissance. 1966*
> *«Statistics offers a toolbox of methods, not just a single hammer […] Statistical thinking involves analyzing the problem at hand and then selecting the best tool in the statistical toolbox or even constructing such a tool».*
> Gigerenzer et al. *The null ritual, 2004*

Sex and gender (S/G)[1]-related factors contribute to individual variability in physiology and behavior, and a S/G-biased prevalence, manifestation, and progression for many somatic, psychiatric, and neurological diseases and disorders has been reported (Altemus et al., 2014; Mauvais-Jarvis et al., 2020; Pinares-Garcia et al., 2018). Therefore, it is currently thought that incorporating S/G-related factors in research and data analysis may be crucial for the understanding of these diseases and for advancing precision medicine and refining diagnostic and treatment strategies in healthcare (Bartz et al., 2020; Stachenfeld and Mazure, 2022). Accordingly, funding agencies in the EU, US, Canada, and other geographical regions have implemented recommendations and mandates to promote or ensure the incorporation of S/G-related factors in biomedical preclinical and clinical research (White et al., 2021).

However, the effectiveness of these policies to improve current knowledge about the role of S/G-related factors in health and disease will critically depend not only on the number of studies addressing these factors, but also of their quality and methodological soundness (Rich-

[*] Corresponding autor at: Departament de Psicologia bàsica, clinica i psicobiologia, Facultat de Ciències de la Salut, Universitat Jaume I, Avda Sos Baynat, SN 12071, Castelló, Spain.
E-mail address: csanchis@uji.es (C. Sanchis-Segura).

[1] In this commentary, we use the term "sex/gender" (abbreviated as S/G) to collectively encompass aspects typically associated with "sex" and "gender". This choice allows us to avoid the relevant but unresolved debates regarding the precise definition of these constructs. It also serves us to acknowledge that, although they can be theoretically distinguishable, the elements traditionally linked to sex and gender are inherently intertwined and rarely (if ever) separable, especially in human studies.For practical convenience, we operationalize S/G as consisting of two *S/G-related categories*, which we refer to as "females" and "males". These terms were adopted from the 1200 Subject Release of the Human Connectome Project, from which the dataset used in this commentary to provide specific examples was extracted. It is important to note that these categories and their labels should be viewed as pragmatic place-holders without implying any specific definition or connotation. In fact, the term "S/G-related categories" is intentionally used to acknowledge the diversity of possible definitions for these categories, depending on the specific research context in which they are defined (See Richardson, 2022).

https://doi.org/10.1016/j.yfrne.2024.101133
Received 10 December 2023; Received in revised form 12 March 2024; Accepted 7 April 2024
Available online 9 April 2024

Edwards et al., 2018; Rich-Edwards and Maney, 2023). Specifically, ensuring rigorous research practices is imperative for drawing reliable conclusions about the exact role and quantitative contribution of S/G-related factors. In this regard, several recent studies (Galea et al., 2020; Garcia-Sifuentes and Maney, 2021; Rechlin et al., 2022) have confirmed a progressive increase in the number of studies including both male and female research subjects, but they have also identified some important methodological deficiencies, such as the omission of sample size, an imbalanced use of males and females, or failing to test formally for S/G effects. This commentary tries to bring attention to another methodological concern of these studies: the overreliance on mean comparisons using classic parametric tests (e.g., Student's t-tests and ANOVAs).

The overreliance on mean comparisons is not exclusive of S/G-related studies, but observed across most domains of social, behavioral, and biological sciences. For example, recent systematic reviews indicate that t-tests and/or ANOVAs are used in 84.5 % of physiology studies (Weissgerber et al., 2018) and up to 12.8 times more frequently than their nonparametric counterparts in psychological research (Blanca et al., 2018). These classic parametric methods are widely used because they are the methods most frequently taught (Aiken et al., 2008; Cobb, 2007; Kline, 2013), and the reason why they are so frequently taught is because they are the most commonly used. This is problematic for at least four reasons: 1) These methods are usually taught, learned and put in practice dogmatically, hence replacing statistical thinking by an automatized testing strategy that pays little attention to the tests' assumptions and that frequently misinterprets the tests' results (Gigerenzer et al., 2004; Hoekstra et al., 2012; Kline, 2013); 2) ANOVAs and t-tests s operate under assumptions that are rarely met, exhibiting low power and providing unsatisfactory/ misleading results when these assumptions are violated (Rousselet et al., 2017; Wilcox, 1998); 3) Even when their assumptions are met, parametric methods comparing means can be of limited informative value (Rousselet et al., 2017; Wilcox and Keselman, 2003); 4) Statistics have much more to offer to researchers than simple average comparisons (Gigerenzer et al., 2004; Wilcox, 2023, 2022), but these new methods have not been incorporated to the statistics curriculum of most researchers (Cobb, 2007).

While the overreliance on mean comparisons pervades scientific research, its impact is particularly pronounced in S/G-related studies. This commentary highlights three levels of limitations associated to mean comparisons in this research domain: first, general issues stemming from assumptions and misinterpretations of classic methods comparing means (Sections 2.1, 2.2, and 2.3.1); second, challenges related to the representativeness of means, which are especially pertinent in the case of large, non-randomly-assigned groups such as S/G-related categories (Section 2.3.2); and third, the categorical model imposed by means and mean comparisons that hinders the goal of incorporating S/G-related factors for the understanding of disorders and diseases and the development of individualized treatments (Section 3.1). In response to these problems and limitations, alternative analytical strategies are also briefly introduced.[2] Particular attention is paid to a statistical method (the shift function; section 3.2) that allows a non-binary treatment of S/G-related information, even when this information is collected as obtained from two categories, and that seems more promising to achieve the goals of S/G-related biobehavioral research.

In conclusion, by unveiling the methodological and conceptual limitations of mean comparisons and proposing alternative strategies, this commentary aims to inspire a more nuanced statistical approach in S/G-related biomedical and behavioral studies. From our viewpoint, embracing appropriate, diverse, and informatively rich analytical tools is a key step to unlocking the full potential of S/G-related factors in disease understanding, treatment refinement, and individualized healthcare.

2. Problems with means and mean comparisons

2.1. Normality and the mean

> *«Let him know how to choose the mean and avoid the extremes on either side, as far as possible, not only in this life but in all that which is to come. For this is the way of happiness»*
>
> Plato. *The republic (circa 427 – 347B.C.E.).*
>
> «The average man, the type of our species, is also the model of beauty [...] The margins of variation (higher or lower) are more restricted in a population the closer it gets to perfection»
>
> Adolphe Quetelet. *Du systeme social et des lois qui le régissent (1848)*

The connection between the midpoint and some concept of virtue, goodness, or truth can be traced back to the philosophies of ancient Greece and the earliest Buddhist writings. In 19th century, Adolphe Quetelet introduced his concept of the *'homme moyen'* ('the average man'; (Caponi, 2013; Grue and Heiberg, 2006)), and proposed that human traits follow a normal distribution, with a tendency to cluster around a central value (the mean). He believed that the mean of any human trait represents the nature's ideal value for that trait, while values on either side of the mean were, for excess or defect, deviations from this natural ideal. Quetelet's work played a pivotal role in popularizing the use of the mean and the normal distribution in the social and behavioral sciences. Today, the normal distribution and the mean are central in statistics, and the idea that the mean reflects the ideal or the 'true' value of any variable remains deeply ingrained in our minds.

The centrality of the mean is not just a metaphor about its importance in statistics, but the main reason of its importance and predominant usage. When data are normally distributed, the mean sits exactly in the middle of the distribution, so the mean is the most centered and also the most frequently expected value (i.e., its value coincides with that of the median and the mode). In such cases, the standard deviation accurately accounts for the values' spread at each side of the mean, and these two statistics suffice to properly describe the variable under consideration (i.e., how common or uncommon is each of its possible values). Accordingly, when data are normally distributed, classic parametric tests (i.e., Student's *t* test, ANOVAs) and effect sizes (e.g., Cohen's d) rooted on means and standard deviations provide a suitable strategy to compare two or more groups and quantify the size of their differences.

Among other reasons (introduced in section 2.3.2), methods such as ANOVAs and t-tests are problematic because data are almost never normally distributed (Blanca et al., 2013; Micceri, 1989). Moreover, neither the mean or the standard deviation is *robust* (i.e., their values and those of their confidence interval can be very much affected by a few outlying values; (Hampel et al., 1986; Högel et al., 1994; Huber and Ronchetti, 2009)). Thus, even relatively small deviations from normality can make the sample mean and standard deviation to provide distorted estimates of the typical value of a group and of the group's dispersion (Tukey, 1960). This, in turn, results in a substantial reduction of the power of classical parametric tests to detect differences between groups when they actually exist (false negative or Type II errors; (Tukey, 1960)), but can also inflate the chances of finding a statistically significant difference when there is none (i.e., false positive findings or Type I errors; (Wilcox and Serang, 2017)), and make classic effect sizes as Cohen's d to yield an inaccurate estimation of the size of these differences (Algina et al., 2005; Wilcox and Serang, 2017).

When researchers are confronted with these warnings about the normality assumption and how its violation distorts the output of classic statistical methods, they may react with disbelief and resort to some

[2] The newer and more robust methods we aim to introduce have not yet been incorporated into commercial statistical software packages, but one of the authors (RRW) has developed functions to implement them using the free software R. These (and other) methods are described in Wilcox (2022, 2023) and their corresponding functions included in the Rallfun-v41 file, which can be freely downloaded from https://osf.io/xhe8u/.

common statistical misconceptions to justify their use of t-tests and ANOVAs.[3] Thus, for example, many researchers think that distributions are "normal enough" when their samples are "large enough", so they feel confident applying parametric methods. This belief is also endorsed in some statistics textbooks and specialized articles (e.g., "*The basic assumptions for ANOVA are independence (*i.e.*, independent experimental units and not repeated assessments of the same unit), normally distributed outcomes, and homogeneity of variances across comparison groups. With large samples (n>30 per group), normality is typically ensured by the central limit theorem; however, with small sample sizes in many basic science experiments, normality must be specifically examined. This can be done with graphic displays […] There are also specific statistical tests of normality (*e.g.*, Kolmogorov-Smirnov, Shapiro-Wilk) …*"; (Sullivan et al., 2016)). However, this rationale is based on the results of some old studies that overlooked the problems associated to the presence of outliers, heavy-tailed distributions, and skewness (Field and Wilcox, 2017; Wilcox and Rousselet, 2018). Indeed, there are formal proofs that even a slight departure from a normal distribution can render methods based on means and variances highly misleading (Hampel et al., 1986; Huber and Ronchetti, 2009; Staudte and Sheather, 1990). Furthermore, trying to establish the normality of a distribution from the visual inspection of a distribution is a fraught and error-prone strategy (Thompson, 2008), as is relying the results of the Kolmogorov-Smirnov, Shapiro-Wilk, and other similar tests, especially when samples are not truly large (n>200 per group or even more).

A simple example might be useful to illustrate these problems and their consequences. Therefore, throughout this manuscript, the points to be made will be illustrated using a dataset that contains the information about self-reported S/G-related categories (male/ female) and the body mass index (BMI) scores of 800 individuals (400 males and 400 females). The BMI was developed by Adolphe Quetelet, and we focus on this variable because it is known that it is non-normally distributed (Silverman and Lipscombe, 2022; Tsang et al., 2018), but nevertheless many studies employ t-tests to compare the BMI scores of females and males (e.g., (Friedmann et al., 2001; Mastorci et al., 2020; Vijayalakshmi et al., 2017)).

2.2. *When mean comparisons are meaningless (an example and a brief description of some robust alternatives)*

According to the current practices (Blanca et al., 2018; Weissgerber et al., 2018), the most typical method to compare the BMI scores of females and males would be to compare their means with a Student's *t*-test for independent samples.[4] The values of these means are 26.81 and 26.19 and, when compared with a *t*-test, there is no sufficient evidence to conclude that there is a statistically significant difference between them (t_{798}=1.75, p=0.080). The crude means' difference (0.62) corresponds to a Cohen's d of 0.12 [-0.03, 0.27] standard deviation units (sd), which according the commonly used benchmarks (Cohen, 1988) could be considered as a "negligible" effect. Therefore, it would be ordinarily concluded that "*males and females do not differ in BMI*" (Hoekstra et al., 2006) and many researchers would probably stop their inquiries at this point. However, that would be a wrong decision based on an erroneous conclusion that stems from misleading results obtained with an inadequate analytical strategy.

To understand this chain of errors, it is important to start by depicting the distributions of BMI scores of females and males (something crucial, but seldomly done (Weissgerber et al., 2015); see also section 2.3.2). Panels A and B of Fig. 1 show these empirical distributions, and also depict the expected ones if cases had been sampled from normal distributions with the same means and standard deviations.

When looking at the empirical BMI distributions, it would be hard to conclude that the females' data are "normal enough", but more doubts could arise when judging the normality of the males' data. However, none of these distributions is really normal, both are right-skewed ($\text{skewness}_{\text{males}}$= 0.71, $\text{skewness}_{\text{females}}$=1.06) and leptokurtic ($\text{kurtosis}_{\text{males}}$=3.27, $\text{kurtosis}_{\text{females}}$=3.89).[5] But what would normality tests say about them? To address this question, the outcomes of three normality tests on 1,000 random samples of various sizes (n=10, 20, 40, 80, 160, 320, and 400 per group) drawn from both empirical and expected distributions were evaluated (panels C and D of Fig. 1). As can be readily observed, when sampling from truly normal distributions, the likelihood of incorrectly classifying a sample as "non-normal" remained consistent at the expected 5%, irrespective of the test or sample size. However, when sampling from the non-normally distributed empirical datasets, these tests were prone to misclassifying as 'normal' distributions that are not when n < 80, a situation commonly encountered in experimental studies. With n > 80, the females' distribution's non-normality was reliably detected at the expected 95% rate. In contrast, the same did not hold true for the males' distribution. Even with a substantial sample size, such as n=400, the Kolmogorov-Smirnov test struggled to detect 'non-normality' in the males' BMI distribution, with a success rate below 75%.

From Fig. 1, it can be appreciated that, in addition to not being normal or even symmetrical, the males' and females' distributions differ

[3] In other cases, researchers may be aware that their data are not normally distributed and may try to achieve normality by transforming their original data (e.g., by applying logarithmic or Box-Cox transformations). This approach can be adequate and useful in some scenarios, and it has been traditionally favored in introductory statistics' books due to the historical neglect of classic non-parametric methods. However it should be noted that: 1) Data transformation does not always solve the problems associated with outliers and skewed distributions; 2) When comparing two groups, it is necessary to find a transformation that works equally well for both groups, but situations can be encountered in which the best transformation for the data group 1 is not the same than for group 2; 3) Transformations can significantly complicate the interpretation of results. For instance, once a transformation such as the square root is applied, inferences about the means of the original data become impractical. Additionally, back-transforming data from the square root scale typically does not yield satisfactory estimates of the original population parameters.; 4) In contrast to their classic predecessors, recently developed non-parametric methods are just as flexible and powerful as t-tests and other popular parametric methods. Therefore, although this issue remains unresolved, we suggest that, in most cases, it may be safer to avoid transformations and instead conduct group comparisons using these more recently developed methods. For a more comprehensive discussion, refer to (Grayson, 2004; Keselman et al., 2002; Pek et al., 2018; Wilcox, 2022)).

[4] For simplicity, we focus our comment on the simplest case of comparing two means with t-tests for independent samples. However, the same limitations and criticisms apply to all variations of this test and to the situations in which more than two means are compared with ANOVAs. Similarly, although here we only describe suitable robust alternatives to t-tests, robust alternatives to all kinds of ANOVAs and ANCOVA methods have been developed. An extensive description of these methods as well as of the description of the R functions to apply them can be found in (Wilcox, 2023, 2022) and some worked case-examples are described in (Field and Wilcox, 2017; Wilcox and Rousselet, 2023a, 2023b).

[5] In a normal distribution, skewness is zero because this distribution is perfectly symmetrical. A commonly rule of thumb considers distributions as "approximately symmetric" if skewness is between −0.5 and 0.5, "moderately skewed" if skewness is between −1 and −0.5 or between 0.5 and 1, and "highly skewed" when skewness is less than −1 or greater than 1.The normal distribution has kurtosis of 3 and it is considered *mesokurtic*. Thus, when the value of kurtosis is >3 distributions are considered *leptokurtic* and this indicates that they have "fat" tails (that is, the distribution has more values at the extremes than a normal distribution has). Conversely, when kurtosis <3, the distribution is *platykurtic*, meaning it has less values in the extremes than those found in a normal distribution. Note that some statistical software programs do not calculate kurtosis but "excess kurtosis" (kurtosis −3) to provide a simple and direct comparison to the normal distribution.

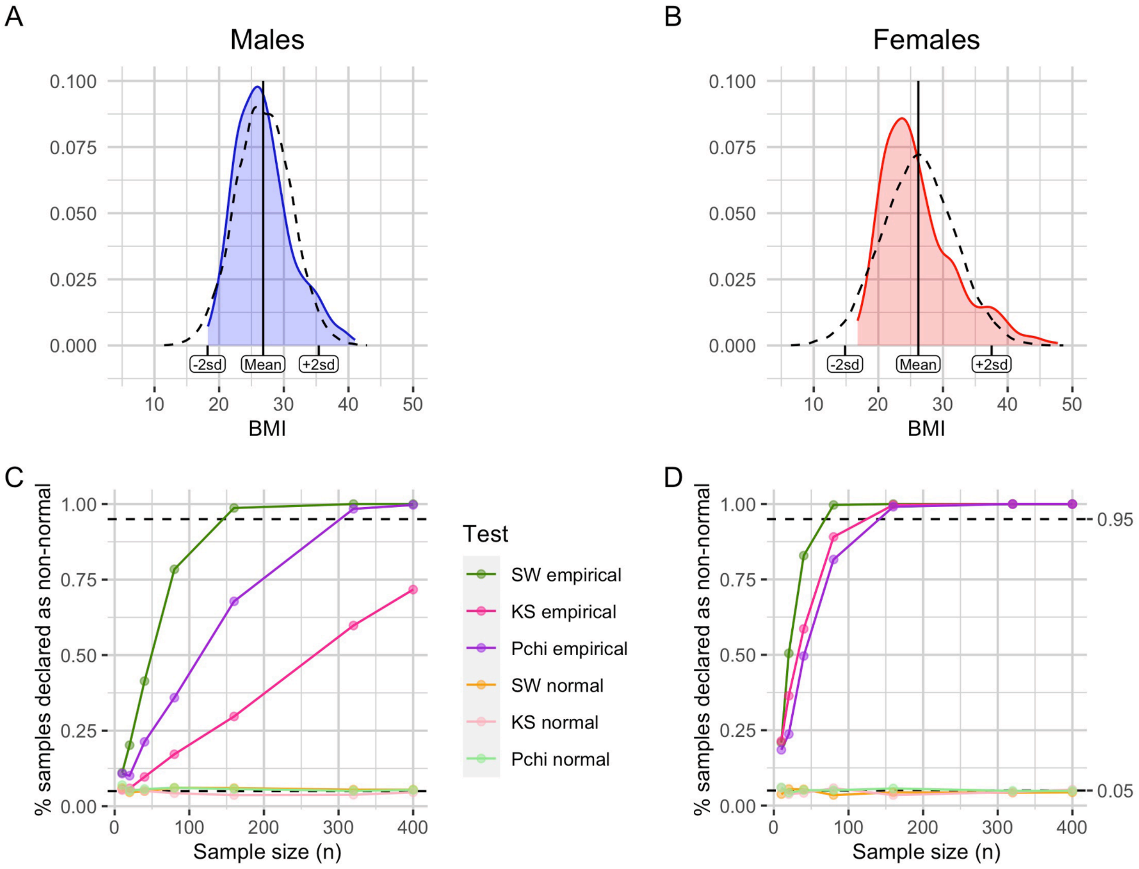


Fig. 1. The distribution of Body Mass Index (BMI) scores. Panels A and B depict the empirical distribution of the males' (blue) and females' (red) BMI scores observed in sample with n = 400, respectively. The panels also include the expected distribution (black dashed curves) if cases would have been sampled from a normal distribution with the same mean and standard deviation. Panels C and D depict the percent of samples declared as "non-normal" by different normality tests in a simulation performed on 1,000 iterations over random samples with n = 10, 20, 40, 80, 160, 320, or 400 extracted from the empirical and expected distributions depicted in panels A and B. (Abbreviations: SW, Shapiro-Wilks' test; KS, Kogolmorov-Smirnov test, Pchi, Pearson-Chi squared normality test; Emp = empirical, norm = expected when sampling from a normal distribution).

in skewness and, also, in spread. In this scenario, the standard error of each mean grows very large and the T-statistic does not follow its expected distribution, so the power of t-tests to detect a difference between two means is severely reduced (for a complete and technical explanation, see (Wilcox, 2022)). Furthermore, disparities in skewness among distributions can profoundly disrupt methodologies grounded in means (Ozdemir et al., 2013; Pratt, 1964; Wilcox and Rousselet, 2023a), and because kurtosis increases the value of the standard deviation, Cohen's d values are artefactually reduced, hence suggesting that effects are smaller than they really are (see Fig. 2 and (Algina et al., 2005; Wilcox, 2023, 2022)).

From the panels A and B of Fig. 1 it can be also appreciated that, when distributions are not normal, the means no longer sit at the middle nor at the peak of the distribution. In fact, when distributions of continuous variables are right-skewed, the mean value is larger than the median's value and the median is larger than that of the mode, so the mean does not represent the most central nor the most frequent value and this discrepancy grows with the degree of skewness. In other words, when distributions depart from normality toward a skewed distribution, the mean does not reflect a typical response and comparing two groups through them becomes not only inaccurate but largely arbitrary and potentially misleading.

At this point it is worth remembering that the mean is only one of many existing central location measures, and it is not a robust one. The median and the trimmed means[6] are also central location measures, and they are much more robust than the mean (for an overview of these and other robust central location measures, see (Wilcox, 2022; Wilcox and Keselman, 2003)). Similarly, there are several statistics more robust than the standard deviation to quantify dispersion (for an overview, see (Högel et al., 1994; Wilcox, 2022)). The use of robust measures of central location coupled with robust measures of spread provides a series of robust, non-parametric analogs of the classic Student's *t*-test and Cohen's d (for an overview, see (Algina et al., 2005; Wilcox, 2023, 2022; Wilcox and Keselman, 2003)). These alternative methods are almost as

[6] Trimmed means is the name given to the means calculated after eliminating ("trimming") a percent (usually, around 10%) of values of each tail of the original distribution. Although ignoring part of the sample might seem somehow "wrong", it is not when the removed values are suspected to introduce some sort of "contamination" to the distribution of interest (that is, when these values may represent individuals of a different population or measurements under different, often unnoticed and/or unusual, circumstances). Eliminating truly contaminating values is necessary to obtain robust estimators, but problems arise when using incorrect methods that may eliminate "proper" values or failing to eliminating all the contaminating values (for an ampler discussion, see (Hampel et al., 1986; Rousseeuw and Stahel, 2011) In this respect, it should be noted that common practices of eliminating cases ±2 standard deviations are inadequate (see (Leys et al., 2013; Wilcox and Keselman, 2003)).

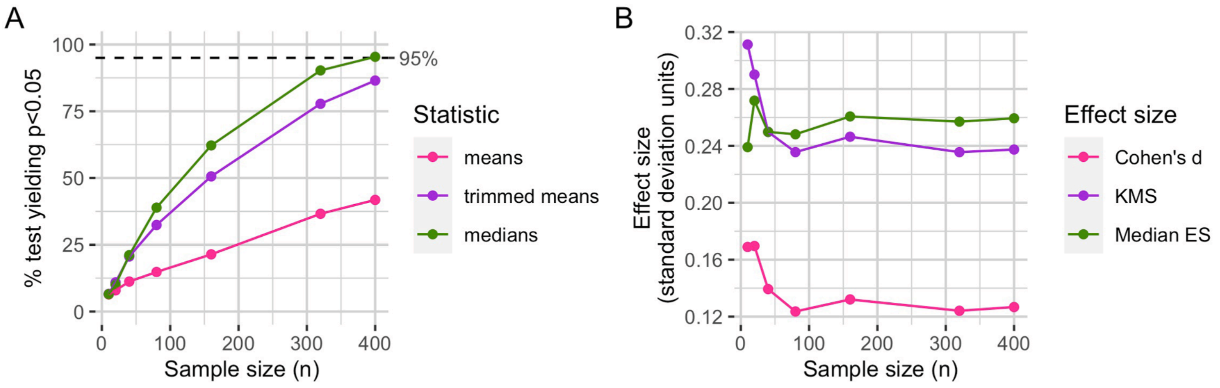


Fig. 2. Comparison of the Student's *t*-test and Cohen's d with some of their robust analogs. Panel A depicts the percentage of occasions on which the performed tests comparing the means, the 20 % trimmed means, or the medians of the BMI scores of females and males yield a p-value lower than 0.05. These results are based on a simulation involving 1,000 iterations on random samples with sizes n = 10, 20, 40, 80, 160, 320, or 400. Panel B depicts the average of the estimated values of different effect size indexes across the same random samples. The effect size indexes compared were the classic Cohen's d and two robust analogs of this index, one based on trimmed means and winsorized standard deviations (KMS) and another based on medians and the median absolute deviation ("median ES"). Note that, although robust effect size indexes are standardized with spread measures other than the standard deviation, they are usually re-scaled to standard deviation units, so their values have the same scale and interpretation than Cohen's d.

Table 1
Misinterpretations (fallacies) of p-values commonly observed in scientific studies. Items marked with (*) are specifically discussed in the main text. For a more comprehensive list of these and other fallacies, refer to Kline (2013) and the references included herein.

Fallacies generally affecting p-values interpretation.	
"Odds against chance fallacy"	The false (and almost ubiquitous) belief that p-values indicate the probability that a result happened by "chance" or sampling error.
"Inverse probability fallacy" (*"Bayesian Id's wishful thinking"*)	The erroneous belief that confounds the likelihood of observing evidence if a hypothesis is true with the likelihood of the hypothesis being true if the evidence is observed. This results in false statements such as saying that, when $p < 0.05$, the probability that the null hypothesis is true is < 0.05.
"Local Type I error fallacy"	The mistaken belief that p-values inform about the likelihood of erroneously rejecting the null hypothesis in a particular study (e.g., stating that, when $p < 0.05$, the likelihood of being erroneously rejecting the null hypothesis is < 5 %). It often arises from the "*inverse probability fallacy*" and it implies forgetting that, for any particular study, the probability of erroneously rejecting (or failing to reject) the null hypothesis can only be 0 or 1 and that, therefore, p-values do not inform of the correctness of the decision of rejecting the null hypothesis in any particular study.
"The sanctification fallacy"	The mistaken belief regarding p-values, wherein an effect (or its significance) is often assumed when $p < 0.05$, even if it's by a narrow margin, while the absence of an effect is concluded when p is just slightly above 0.05
Fallacies commonly associated to p > 0.05	
"Zero fallacy"(*) (*"the slippery slope of non-significance"*)	The erroneous belief that the failure to reject the null hypothesis reveals that the difference between the means of two populations equal to zero.
"Equivalence fallacy" (*)	The mistaken belief that the failure to reject the null hypothesis in a means-based comparison reveals that two populations are equivalent.
Fallacies commonly associated to p < 0.05	
"Valid research hypothesis fallacy" (*)	The false belief that, when $p < 0.05$, the probability that the alternative/ research hypothesis is > 0.95.
"Replicability fallacy" (*)	The mistaken belief that the complement of p indicates the probability of obtaining a "significant" result in a replication study.
"Causality fallacy" (*)	The erroneous belief that statistical significance proves that the tested independent variable is the underlying causal agent (literally, "the factor" or "doer") of the phenomenon under investigation.
"Magnitude fallacy" (*) (*"the slippery slope of significance"*)	The use of the term "significant" (without the qualifier "statistically"), when describing results, and their automatic interpretation as "large" or "important".

powerful and accurate as these classic statistics when the assumptions of normality and equal variances are met. However, because they do not make any assumption about the distributions from which they are calculated, they are much more accurate and powerful to unravel between-group differences when the normality and homoscedasticity assumptions are violated. This can be illustrated with the data of our example. The values of the 20% trimmed means for males and females are 26.31 and 25.18, and their difference is statistically significant (p= 0.003) and no longer "negligible" but "small" in size (0.24 [0.08, 0.39]). Similarly, the values of the medians are 26.25 and 24.83, the p-value associated to their comparison with a percentile bootstrap method is <0.001, and the size of their difference is 0.32 [0.15, 0.47]. These

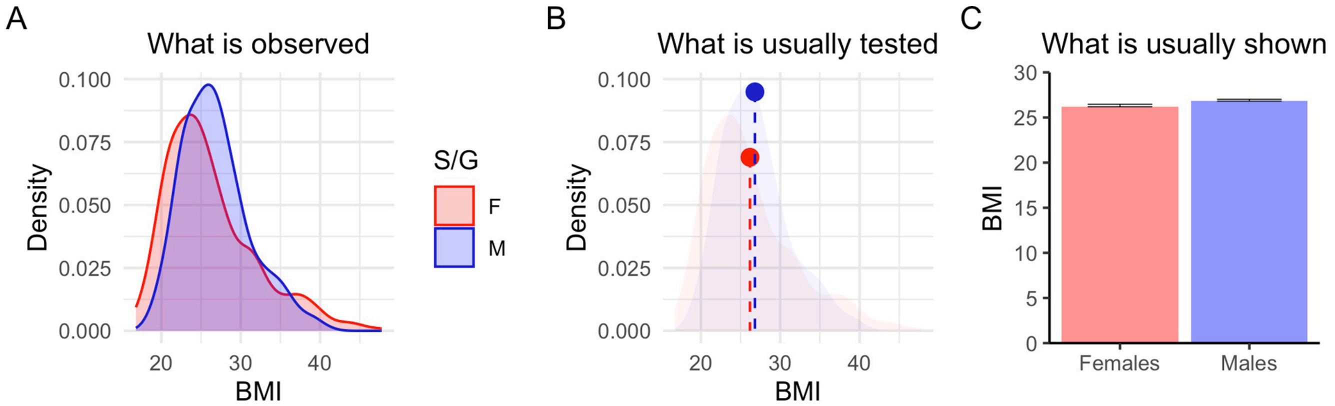


Fig. 3. Mean comparisons often hide more than they reveal. Panel A illustrates the distributions of the BMI scores of the females and males included in our sample. Panel B depicts what is tested (and how much is obviated) when conducting a Student *t*-test for independent samples. Panel C depicts how inappropriate graphical depictions impede judging the capability of means and mean comparisons to represent the compared S/G-related categories and their possible differences.

results are quite similar between them, but quite different than those obtained from means (26.81 and 26.19) and a *t*-test (p=0.080; 0.12 [-0.03, 0.27]) on the same subjects.

To better illustrate the behavior of these robust methods and compare their results with those yielded by t-tests and Cohen's d, we evaluated their performance across several sample sizes. The results of these simulations based on BMI data are illustrated in Fig. 2. As it can be readily observed, rejecting the null hypothesis required large samples because the estimated between group differences are truly small, but robust methods have consistently more (in this case, around twice the) power to detect these differences than t-tests. Thus, for example, the method based on median comparisons rejects the null hypothesis at rates close to the expected 95% when n≥ 320, whereas under the same conditions the classic *t*-test based on means only rejects the null hypothesis in 37–40% of the cases. Panel B of Fig. 2 illustrates that the size of the between groups standardized difference is substantially (in this case, 2-fold) larger when based on robust statistics of location and spread (such as medians and median absolute deviation) than when based in the means and standard deviations that are much more affected by extreme values.

In summary, there are many situations in which the data are not normally distributed and normality tests may fail to detect this situation, especially when sample size is limited. In this scenario, the mean and the standard deviation are invalid estimators of central location and spread. Consequently, t-tests and other similar parametric statistics are largely insensitive to between-group differences and Cohen's d underestimates their magnitude. However, there are more sensitive and accurate testing methods and effect size indexes to identify and describe these differences.

2.3. What means and mean comparisons mean?

Just as the assumptions of commonly used parametric tests are often ignored because of existing statistical myths about their robustness against normality violations, the interpretation of their results is usually flawed because of commonly shared misconceptions about p-values and means. Most of these false beliefs seem to reveal a wishful attempt to bridge the gap between what these comparisons can provide and what it is wanted to know. Because the fallacies commonly associated to the interpretation of p-values in the context of significance testing have been extensively debated (Cohen, 1994; Hirschauer et al., 2022; Kline, 2013), we will not discuss all of them here (but see Table 1). Rather, we will briefly discuss some misconceptions about means and mean-based comparisons, an issue that has received far less attention despite of being similarly important. In this last regard, note that, although for simplicity we will mainly refer to means and their comparisons, some of the criticisms and limitations also apply to their just described robust analogs.

2.3.1. Common misinterpretations of p-values in the context of mean comparisons

> «*What's wrong with NHST [null hypothesis significance testing]? Well, among many other things, it does not tell us what we want to know and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!*»
> Cohen, J. (1994)

As hinted at the beginning of section 2.2, concluding that "*males and females do not differ in BMI*" because the p-value associated to this comparison was higher than 0.05 would not be unusual, but it is erroneous for several reasons. Thus, this conclusion stems from a dichotomous interpretation of p-values and it incurs in the so called *zero fallacy* (also known as *the slippery slope of non-significance;* (Cumming, 2012)). This common misinterpretation of p-values> 0.05 takes the absence of evidence for a statistically significant difference between two (or more) groups as evidence of the absence of a difference between these groups (Altman and Bland, 1995; Cumming, 2012). However, significance tests cannot prove the absence of a difference.[7] Moreover, p-values larger than 0.05 can be obtained even if two groups actually differ when statistical power is insufficient. It is well known that insufficient power can be due to the use of small samples (which was clearly not the case in our example), but it is generally less known that power is also substantially

[7] Although it is often overlooked, significance testing only allows us to either reject or fail to reject the null hypothesis of no differences. That is, the null hypothesis can never be accepted and, therefore, hypotheses can never be falsified with significance testing methods. It has been argued (Heene and Ferguson, 2017) that this implies an inversion of the proper process of hypotheses falsification that characterizes science (which assumes that there are not "true" hypotheses but only hypotheses proved false or yet to be been proven false).For the study of S/G-related effects, this inverted logic implies the impossibility of showing that a difference between females and males (or whatever other S/G categories) does not exist or to statistically substantiate S/G-similarities with the p-values obtained from hypothesis testing methods. This fact, together with the current praise of significance testing methods and p-values, promote an asymmetrical framework that equates a gain of knowledge with the identification of statistically significant differences between S/G-related categories (see also note #8). Therefore, is worth reminding here that, as eloquently put by (McCarthy and Konkle, 2005) almost twenty years ago, "*Understanding how the sexes are the same is just as important as how they differ, but the latter receives far less attention and little value as a genuine scientific finding*".

reduced when comparing groups of unequal size (which, again, is not the case in our example but it seems to occur in around a third of neuroscience and psychiatry S/G studies (Rechlin et al., 2022)), or when employing statistical methods that are not really suited for the data at hand (as it was the case in our example). In a related but yet different sense, this conclusion incurs also in the so-called *equivalence fallacy* (Kline, 2013) as it erroneously interprets the failure to reject a null hypothesis of no differences between means as indicative that the compared populations are equivalent or "not different". However, means are not populations, and, as illustrated in section 2.2, the fact that there is not a statistically significant difference between their means does not necessarily entail that there is not a difference between males and females (see also section 3).

In a similar vein, concluding that "*males and females significantly differ in their BMI scores*" because the p-value associated with a mean (or other average-based) comparison is less than 0.05 is also very common but incorrect. This conclusion also stems from a dichotomous interpretation of p-values and, in this case, it falls in the so-called *valid research hypothesis fallacy* (Carver, 1993, 1978), which takes the rejection of the null hypothesis as evidence supporting the researcher's hypothesis. However, the researchers' hypothesis is not under evaluation, only the null hypothesis is, and it is known beforehand that the null hypothesis of no differences is false[8] (though statistical power may not always be sufficient to demonstrate this; for an ampler discussion, see (Cohen, 1994; Hirschauer et al., 2022; Tukey, 1991)). This should not be interpreted as if p-values would have no meaning or utility, but rather that they should be properly understood and used. In the words of Wassertein et al (2019), "*we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (*e.g.*, p=0.08). They should also be described in language stating what the value means in the scientific context. […] we must recognize afresh that statistical inference is not—and never has been—equivalent to scientific inference*" and, therefore, when interpreting p-values, it is necessary to "*Accept uncertainty. Be thoughtful, open, and modest*".

Unfortunately, these calls are often unattended and the allure of "significant results" is such that p-values< 0.05 ordinarily trigger many other sophisms. For example, it is commonly assumed that, when p<0.05, the probability of replicating the same result in future studies must be >0.95 (*replicability fallacy*;(Carver, 1978)), and that the factor tested is "the cause" of the observed effect (*causality fallacy*;(Kline, 2013)). In terms of replicability, it's important to highlight that if the effect size in the entire population matches that observed in a sample where a *t*-test produces a p-value below 0.05, the likelihood of obtaining a p-value <0.05 in a replication study is 0.5, not 0.95 (Greenwald et al., 1996). On the other hand, establishing causality requires more than just statistical testing – it also involves careful consideration of the research design, potential confounding variables, and alternative explanations for the observed results. The challenges in establishing causation become even much more pronounced when the factor tested cannot be directly manipulated and random assignment is not possible, as is the case when comparing S/G-related categories (Cox, 2006; Jacklin, 1981). Thus, to avoid the *causality fallacy* (and the essentialist statements about S/G categories frequently associated with it, see below), it is worth remembering that "*Sex is not a force that produces these contrasts; it is merely a name for our total impression of the differences*" (Lillie, 1939) and that S/G, especially if operationalized in two (or more) categories, is more often a moderator than a "true" factor (for an ampler discussion, see (Jacklin, 1981; Krieger, 2003; Maney, 2016; Richardson, 2022; Springer et al., 2012)).

Additional problems with the conclusion "*males and females significantly differ in their BMI scores*" arise from its omission of two highly relevant words: "statistically" and "means". Firstly, bereft of its necessary complement (statistically), the term "significant" acquires its ordinary meaning, incorporating additional connotations such as "large," "important," or even "fundamental." This fallacy is referred to as the *slippery slope of significance* (Cumming, 2012). It is more likely to occur when there are pre-existing notions about the compared categories as being intrinsically different or even "opposite" (as is commonly the case for S/G-related categories), and its consequences are aggravated when not including any effect size index. Secondly, omitting the word "means" when describing the obtained results allows 'females' and 'males' to become the subjects of the sentence. This omission is also more likely to occur when comparing categories perceived as intrinsically or essentially different (e.g., S/G-related categories), and makes the conclusion to become a generic statement that does not properly represent the analysis performed. Furthermore, the omission of the term "means" may incorrectly imply that all males' BMI scores are different from all females, even when this has not been actually tested and is unlikely to be true. The following section delves deeper into this issue.

2.3.2. *Can mean differences be really considered group differences?*

> *«The over-reliance on the mean expresses a way of thinking about distributions and variability that we believe poses potentially grave problems for our science».*
> Speelman, CP and McGann, M (2013)

At the beginning of section 2, we mentioned that Adolphe Quetelet popularized the use of the mean, which he believed represented the ideal or true value of a trait as opposed to those values lying to either side, which he regarded as undesirable deviations or errors. While contemporary researchers do not endorse Quetelet's ideas, the statistical approaches most commonly used still rely on the mean as the best value to summarize a set of scores and treat the variation as "error" or "noise".

(Speelman and McGann, 2013). In this regard, it is revealing to observe that measures of spread are ordinarily replaced with the standard error of the mean (SEM) that tell us nothing about the scores' distribution (Andrade, 2020; Davies, 1998) and that bar and line graphs depicting means and SEMs –but hiding the data distribution- are commonly used for presenting continuous data (Lane and Sándor, 2009; Weissgerber et al., 2015). The brain pictures typically found in neuroimaging studies have the same concealing effects (Allen et al., 2012; Roskies, 2007). It is very unlikely that these generalized practices stem from a shared, thought-out model or from deliberate attempts to conceal variation. Rather, they probably are a mere reproduction of a learned tradition about the usage of means that is largely blind to its own assumptions, limitations, and implications.

Reducing a set of scores to its mean, or any similar statistic, carries with it the assumption that the information inherent in these data can be accurately encapsulated by a single number. While means (and other summary statistics) are convenient for communicating complex or

[8] As Tukey made clear "*It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place*" (Tukey, 1991). This fact implies that, with a sample sufficiently large, any difference will achieve statistical significance. Therefore, when working with very large datasets, statistical significance may be largely irrelevant. To deal with this situation, Tukey proposed to interpret p-values as a continuous index quantifying the strength of the empirical evidence that a decision about which group has a larger value in a location measure can be made, *and not as quantifying the evidence supporting the researchers' hypotheses* (Tukey's "three-decisions rule"; (Jones and Tukey, 2000; Rice and Krakauer, 2023)). As discussed by Cohen (1994), this implies recognizing that p-values are solely informative about the possible direction of an effect and, consequently, the need to establishing the effect's size (and its confidence intervals) as well as its practical significance (the "so what" question). Other authors have proposed testing non-nil null hypotheses through equivalence testing procedures (Lakens et al., 2018; Seaman and Serlin, 1998)) or to include some alternatives/ complements to p-values (e.g., s-values; (Greenland, 2019); SGPV, (Blume et al., 2019); "analysis of credibility" (Matthews, 2019)). For a recent and comprehensive overview of how p-values should (and should not) be used, see Wassertein (2019).

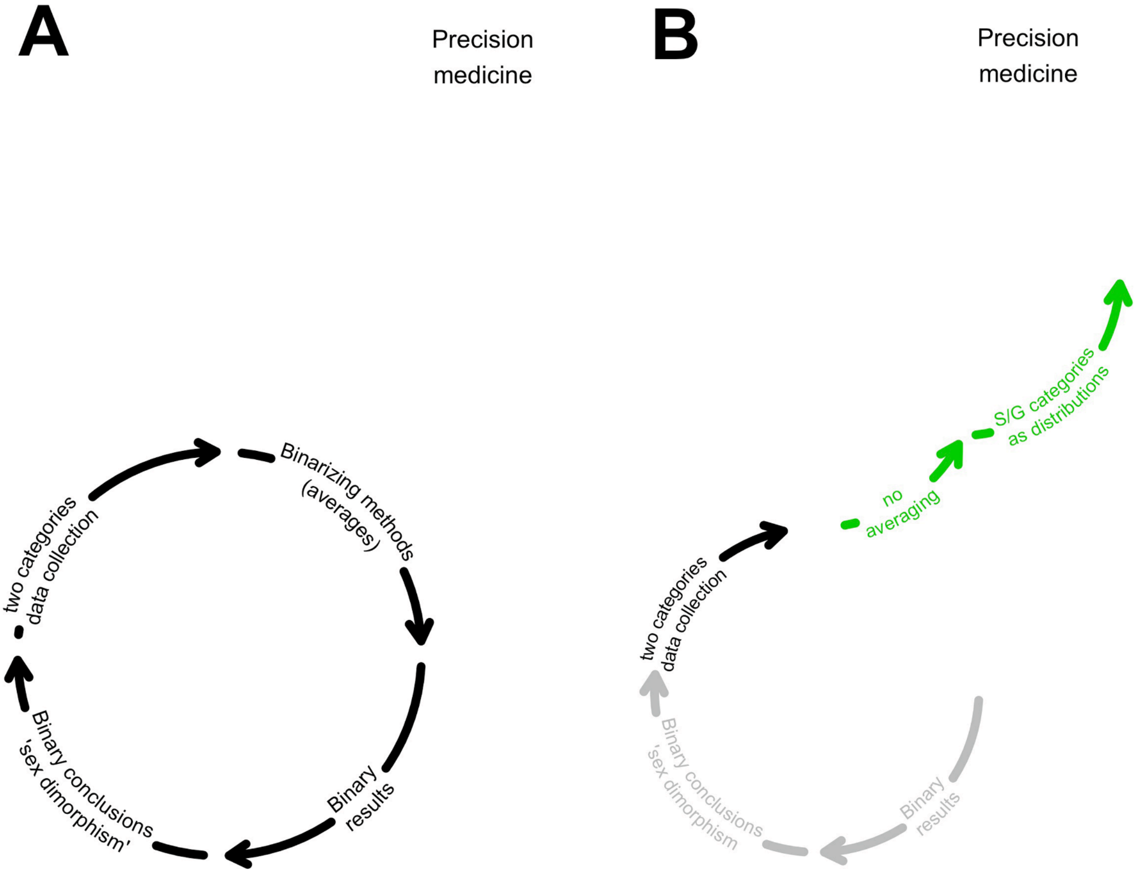

A
Precision medicine
Binarizing methods (averages)
Binary results
Binary conclusions 'sex dimorphism'
two categories data collection
B
Precision medicine
two categories data collection
no averaging
S/G categories as distributions
Binary results
Binary conclusions 'sex dimorphism

Fig. 4. Averaging imposes a binary analytical and interpretative framework. Panel A illustrates how, by producing a categorization (in most cases, a binarization) of the information collected about S/G-related categories, averaging prompts binary conclusions (often mispresented in terms of "sex dimorphism") that feedback the use of these categories and averages and keeps S/G-related research away from its goal of advancing toward precision medicine. Panel B shows that this self-reinforcing but counterproductive loop can be broken when averages are replaced with other analytical strategies that allow for the treatment of S/G-related categories as distributions (see section 3.2), hence providing more detailed and nuanced information that may potentially be more useful when incorporating S/G-related information into the design of individualized health interventions.

extensive datasets, their utility hinges on their representativeness. However, *means are trees that can make us miss the forest* and, therefore, we must ask whether means faithfully represent the data or whether they conceal important nuances. In this regard, it's worth remembering that, as illustrated in section 2.2, when improperly applied (i.e., when data are non-normally distributed), means may not accurately reflect the data and can prompt misleading conclusions. However, even when dealing with normally distributed data, solely reporting means results in a major loss of information that can make data to appear more uniform or stable than they really are (Speelman and McGann, 2013). Unfortunately, because –as already mentioned- scores' distributions are rarely shown, it is often impossible to assess to which extent means actually provide an adequate summary of the data (Lane and Sándor, 2009; Weissgerber et al., 2015).

Mean comparisons share the assumptions, limitations, and potential pitfalls of averaging but also introduce a new one –that the differences among members within the compared group are constant, or at the very least, adequately represented by the results of a single comparison. When this assumption is not satisfied, the identified means' difference (or absence thereof) does not represent the distinction between the compared groups. At best, it just illustrates the distinction existing between a subset of members from one group and a subset of members from the other group (Anastasi, 1981; Jacklin, 1981; Speelman and McGann, 2013) but, in the worse scenario, it may just provide a misleading summary of several effects running in the same or in opposite directions (see section 3). These distortions may occur when comparing any two groups, but they are more likely to happen when comparing large, non-randomly-assigned groups (as it is the case of S/G-related categories;(Anastasi, 1981; Jacklin, 1981)).

The validity and representativeness of mean comparisons are rarely questioned, but there is a large gap between the complexity and informational richness of what is measured and the scarcity of what it is tested (although improper graphical representations impede both to realize this gap and to evaluate its extent; see Fig. 3). Similarly, there is also a gap between what is tested and what is concluded. As already mentioned, the identification between means and the populations they aim to speak for is so strong that the word "means" is often omitted, hence leading to generic and misleading statements such as "*males and females do not differ in BMI*" or "*males and females significantly differ in their BMI scores*". These generic statements often make us to forget about within-group differences and also that, when these differences are larger than those observed between the means, an individuals' membership in a given group provides little or no information about its status in the considered trait (Anastasi, 1981).

It should be emphasized that, even when the average of one group exceeds that of the other by a large amount, some individuals in the lower-scoring group surpass some individuals in the higher-scoring group and that this should be quantified with appropriate effect size indexes (i.e., probability of superiority; PS).[9] In our sample the PS estimate is 0.57, which again indicates that females and males seem to not differ much in their BMI scores. However, PS and other probabilistic effect size indexes are more apprehensible and meaningful than Cohen's d,[10] provide a complementary perspective about how the groups differ (Grissom & Kim, 2005; Wilcox & Rousselet, 2017, 2023), and also show how misleading the generic statements that often stem from mean comparisons can be.

3. Rethinking analytical strategies for s/g research

3.1. Are mean comparisons useful, useless, or counterproductive in studying s/g research?

> *«It is difficult to understand why statisticians [and endocrinologists/physiologists] commonly limit their inquiries to Averages, and do not revel in more comprehensive views»*
>
> Galton, F.R.S. Natural inheritance (1889)

It is often overlooked that the statistical methods used, rather than the theoretical models one may subscribe, shape how research aims are translated into specific questions and determine the nature and quality of the answers obtained. In this section, we discuss how the overreliance on means and mean comparisons not only reduces S/G-related research to the search of average differences between S/G-related categories but also introduces a categorical model that hinders the goal of increasing current knowledge about the role of S/G-related factors in health and disease.

It might seem that when sample size is "small" (e.g., when n< 10–20 cases per group), comparing S/G-related categories through their means is probably *all what can be done*. However, these comparisons can be improved by: 1) Depicting the data distribution with scatterplots, box plots, or violin plots; 2) Comparing medians or trimmed means, which (as shown in section 2.2) almost always provide more accurate and more powerful comparisons than those based on means when sample size is small; 3) Incorporating probabilistic effect sizes such as the PS. These effect size indexes provide a complementary perspective to those informing about the magnitude of the averages' difference and should prevent generic statements about "(all) males and (all) females" when drawing conclusions; 4) Strictly restricting conclusions to what has been compared (averages) and explicitly naming in these conclusions the kind of average that has been compared. Nevertheless, even when all these improvements are incorporated, the results of studies conducted with small sample sizes and comparing S/G categories through a single central location measure should be regarded with caution and may contribute little to our understanding.

When samples are large, comparing S/G-related categories through their averages should be regarded as an inefficient allocation of resources and a missed opportunity to get deeper insights about the data. Think about this common situation: Substantial efforts and resources are devoted to collect the largest possible sample (in some cases including hundreds or thousands of females and males), but this information is reduced to a single number, the difference between their respective means (or, worse even, the p-value associated to this difference). This single value is then used to draw conclusions not only about the differences between the samples but also about the samples themselves, and even about the broader populations they aim to represent. Is this approach truly sensible? Is it really the best we can do?

Probably not, but it is what is most frequently done in S/G-related research. In fact, this analytical strategy is so common (and not only among S/G-related studies) that it does not have a name, but it could very well be called "*the hourglass fallacy*". The hourglass metaphor is aimed to emphasize the significant gap between the resources invested and the information obtained, as well as the exaggerated conclusions drawn, while highlighting the inherent fragility of situating average differences as the epicenter of this approach. Referring to it as a fallacy may seem too severe, but the term seems appropriate when considering that this analytical strategy does not only imply a suboptimal allocation of resources but also imposes a categorical (often binary) framework. This categorical/ binary framework seems particularly inappropriate for biobehavioral S/G-related research as it hinders its main goal −describing S/G-related variation to promote more personalized health interventions.

Averaging and average comparisons may seem justified (or even the only possible analytical strategy) in S/G-related research because data are ordinarily collected as belonging to two distinct and mutually-exclusive categories (males and females). However, it is the process of averaging (and not the way data are collected) that generates a categorization of the outcome, imposing a dichotomization that may not actually exist in the data. In fact, even when information about S/G-related factors is collected as a categorical variable with only two possible and mutually-excluding values (males or females), their measurements in almost any trait are rarely dichotomous (Hyde, 2014; Joel, 2011; Maney, 2016; Reis and Carothers, 2014). That is, once a trait is measured, females and males no longer form two categories but two empirical distributions that spread at different probability levels within particular ranges of the outcome's continuum. Therefore, it is only when all this information is forcefully simplified into a single number, typically the mean, that S/G-related data becomes actually binary (Fig. 4A). These averages not only contribute little to the goals of biomedical S/G research but may also discard crucial aspects that this research could or should provide. In fact, the overreliance on mean comparisons has led to

[9] Probability of Superiority (PS) is one of the many names used to designate a series of estimators of the probability that the values of a group "A" are higher than those coming from another group "B" (Grissom and Kim, 2012). The PS can be derived from the value of Cohen's d using some established formulas (and then it is commonly referred as "common language effect size" or CLES (McGraw and Wong, 1992)) but these calculations (as Cohen's d itself) assume that scores follow a normal distribution. Consequently, their results can be misleading when the normality assumption is violated and it is safer (and often, more accurate) to estimate this probability from non-parametric methods (Ruscio, 2008; Wilcox, 2022). Thus, for example, in our sample the probability of a male having a larger BMI score than a female is 0.53 when directly derived from Cohen's d, but 0.57 when non-parametrically estimated.From Cohen's d other probabilistic effect sizes can be derived, such as U1 (the percent of overlap between two distributions), U2 (the percentage of cases of group A that exceeds the same percentage in group B), and U3 (the percentage of cases of group A that exceeds the median of group B). These statistics also assume normality (Cohen, 1988), but non-parametric estimators for U1 (Pastore and Calcagnì, 2019) and a robust extension of U3 based on the median ("quantile shift"; Wilcox, 2021) have been developed.A final worth mentioning variation of PS is Cliff's delta (Cliff, 2014). This statistic estimates the probability that a randomly selected observation from one group is larger than a randomly selected observation from another group, minus the reverse probability (i.e., P (A>B) minus P(B>A)). Cliff's delta is both an informative effect size and a robust non-parametric test of the differences between two (or more) groups that does not relies on any location measure (Wilcox, 2022, 2021, 2006).

[10] Cohen's d is commonly used but difficult to understand and, often, poorly interpreted (Acion et al., 2006; Hanel and Mehler, 2019; Ruscio, 2008). In fact, most people find difficult to grab what 0.2 or any other number of standard deviations actually implies or means. Some "translations" of Cohen's d have been developed to make it more readily comprehensible and meaningful (e.g., the "common language effect size" and Cohen's U statistics; see note #9). Yet, Cohen's d is very frequently interpreted resorting to some cut-offs that classify effects as "negligible", "small", "moderate" and "large". This has been very much criticized (Glass et al., 1982; Thompson, 2011) and, in fact, Cohen himself warned about using those boundaries and explicitly noted that they might be especially inappropriate in biological research and other similar experimental contexts (Cohen, 1988).

a situation where more is known about whether the means of S/G-related categories differ than about how the members of these categories relate to specific traits.

However, since the goal is not to draw conclusions about mean differences but about males and females themselves, simplistic comparisons between ’the average male’ and ’the average female’ are often used to make unwarranted generic statements and conclusions about all females and all males. As already mentioned, these generic statements omit key words (such as “statistically” and “means”) then hiding the reductionist nature of average comparisons and making differences between S/G-related categories appear as “large” and “universal”. This terminological and conceptual drift reaches its peak (and is bolstered by) the currently common misuse of the term *sexual dimorphism* in biomedical studies. Although this term literally means “two-forms”, it is often employed to refer to any statistically significant difference found between females and males, hence ignoring that, except for a few aspects related to reproductive functions, these differences rarely (if ever) take two distinct forms. To the contrary, when effect sizes are calculated, females and males show a high degree of overlap in most traits, and individual differences within the members of each of these categories can be as large or even larger than that existing between their averages (e.g., (Hyde, 2014; Maney, 2016; Reis and Carothers, 2014; Ritchie et al., 2018; Zell et al., 2015). There have been repeated calls to cease this misleading and uniformizing use of the term “sexual dimorphism” and to classify female-male differences according to their statistical characteristics and other criteria (DeCasien et al., 2022; Eliot et al., 2023, 2021; Joel, 2011; Joel and McCarthy, 2017; McCarthy et al., 2012). Nevertheless, these claims have had little effect on how researchers ordinarily report their findings, and the misuse of the term “sexual dimorphism” continues feeding back the same binary framework based on averages that initially motivated its use (Fig. 4A).

At this point, we should probably ask ourselves: Is averaging −and the theoretical model it imposes- the best strategy to incorporate S/G-related factors in the description of biological and behavioral traits? Is averaging the best strategy to advance precision medicine and more individualized treatments? If the answer to both questions is “no”, we should also ask ourselves what exactly studies comparing the averages of males and females provide and whether they should be continued. Even more important, we should think which other analytical strategies could be more suitable for attaining the intended goals. In the next section, a promising alternative to mean comparisons is illustrated.

3.2. A shift away from average comparisons (the shift-function)

> *«An Average is but a solitary fact, whereas if a single other fact be added to it, an entire normal scheme, which nearly corresponds to the observed one, starts potentially into existence. […]*
> *So, in respect to the distribution of any human quality or faculty, a knowledge of mere averages tells but little; we want to learn how the quality is distributed among the various members of the Fraternity or of the Population, and to express what we know in so compact a form that it can be easily grasped and dealt with […] A knowledge of the distribution of any quality enables us to ascertain the Rank that each man holds among his fellows in respect to that quality. This is a valuable piece of knowledge»*
> Galton, F.R.S. Natural inheritance (1889)

To move away from the intrinsic limitations of average comparisons −and the theoretical model they impose- requires stop treating males and females as if they were homogeneous categories summarizable by a single number and to start treating them as distributions with appropriate analytical strategies (Fig. 4B). These methods should provide a complete assessment of differences and similarities in location, spread, and shape without making distribution assumptions or requiring large sample sizes. Moreover, they should not only be comparative but also descriptive, providing both numerical and graphical insights about the compared groups and not only about their differences.

A method able to satisfy all these requirements is the so-called shift function (Doksum and Sievers, 1976; Wilcox, 2021, 2006). The shift function is both an inferential method and an informative graphical display. Without getting into its statistical details (which can be found in (Wilcox, 2022; Wilcox and Rousselet, 2023a)), the inferential method can be described as a non-parametric test that extends the robust median comparisons illustrated in section 2.2 to simultaneously compare several quantiles.[11] This method can be applied when dealing with two related groups (Rousselet et al., 2017), two independent groups (Rousselet et al., 2017), or four groups in a 2x2 design (Wilcox and Rousselet, 2023b), all without assuming normality or homoscedasticity. In addition, this procedure allows customizing which quantiles to compare (a decision that should take into account the available sample size; see below), and it provides p-values adjusted to the number of performed comparisons as well as the differences between these quantiles, along with their 95% confidence intervals as non-standardized effect sizes. The classic graphical display of the shift-function represents the values of the quantiles of one of the compared groups on the x-axis and the between-group quantile differences on the y-axis, hence revealing how and by how much one distribution must be adjusted or ’shifted’ to match the other.

To illustrate this method, let’s return to our BMI example and compare the deciles of the males’ and females’ distributions (Fig. 5). From this depiction, a new finding immediately emerges: the differences between males and females are not consistent across the entire BMI range. Specifically, differences favoring males are observed for deciles 1–5 (that is, among individuals with low to intermediate BMI, females have BMI scores that are consistently around 1.5 points lower). These differences appear to become less reliable and progressively smaller at deciles 6–7, ultimately reversing their direction in deciles 8 and 9, although evidence for this latter effect is inconclusive (that is, among individuals with the largest BMI scores, females seem to have larger BMI scores, but there is large individual variation). This in-depth analysis provides a comprehensive and informative view of how these two groups differ in location. It also explains why the mean differences were deemed ’negligible’ and failed to achieve statistical significance (i.e., because opposing effects canceled each other out when calculating the averages). However, the shift function does not only inform about location differences. Thus, the fact that the line is not parallel to the x-axis reveals that the groups also differ in spread, and its positive slope indicates that the spread is larger in the represented group (in this case, the females). Moreover, the non-linear nature of the shift-function indicates that the distributions are skewed and, because the spread is larger on the right side of the divide created by the median, it can be deduced that the distributions are right-skewed.

The informativeness of the shift-function becomes self-evident when comparing Figs. 3 and 5. Moreover, this method is powerful and does not require prohibitive sample sizes, allowing a reliable comparison of all deciles when n≥30 per group (and the comparison of the three quartiles when n≥20 (Wilcox and Rousselet, 2023a)). However, the classic graphical display of the shift function has some potential drawbacks.

[11] Quantiles are cut points that divide an ordered variable (x) in intervals each of them containing the same number of cases. The most commonly used quantiles include quartiles (Q1, Q2, and Q3), deciles (D1 to D9), and percentiles (P1 to P99) and the most commonly known quantile is the median (also denoted as Q2, D5, and P50). Quantiles are cut points, and they should not be confounded with the intervals they produce (e.g., quartiles are 3 values that split a variable into its quarters, deciles are 9 values that split a variable into its tenths, etc.).Quantiles provide the basis for some robust measures of location and spread (Wilcox, 2023, 2022), and there are different procedures to compare quantiles from independent or related groups and quantify the magnitude of these differenced in a similar way to Cohen’s d (Wilcox, 2023, 2022). Moreover, as hinted in the footnote #9, quantiles provide the basis to calculate some probabilistic effect sizes (see also footnote #12).

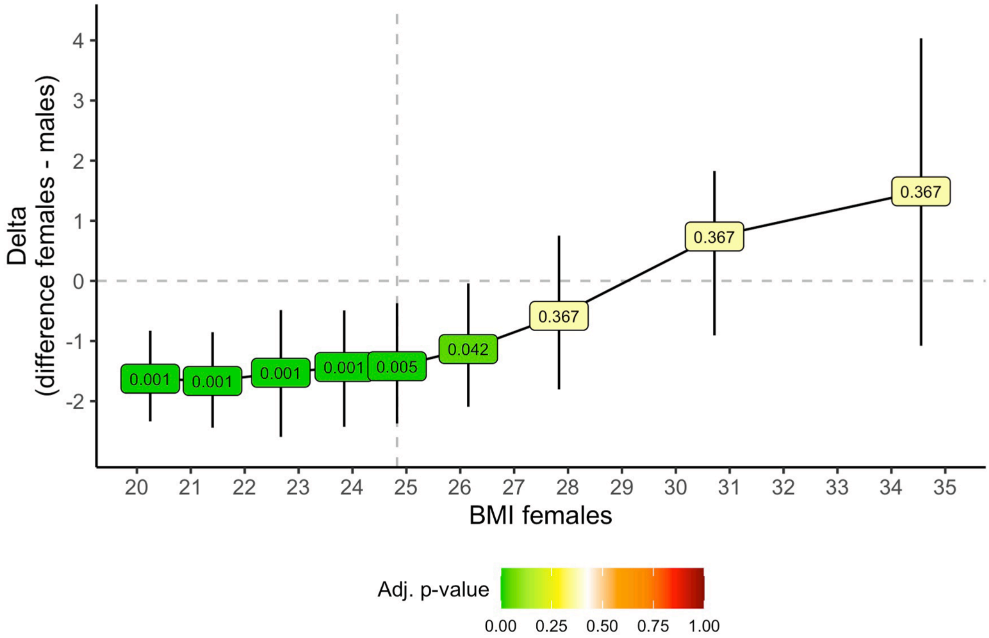


Fig. 5. The shift function of BMI data. The x-axis displays the deciles of the females' BMI scores. The y-axis is the estimated difference between the corresponding deciles of females and males. For each decile (dots) the bootstrap-estimated 95% confidence interval is also depicted (vertical lines). To enhance readability, this graph also includes: 1) A horizontal dashed line highlights the height at which the between-group differences equal to 0; 2) A vertical dashed line marks the median of the females' scores; and, 3) Dots are color-coded depending on the p-values associated to each comparison. P-values were adjusted for multiple comparisons using the strict Hochberg's method.

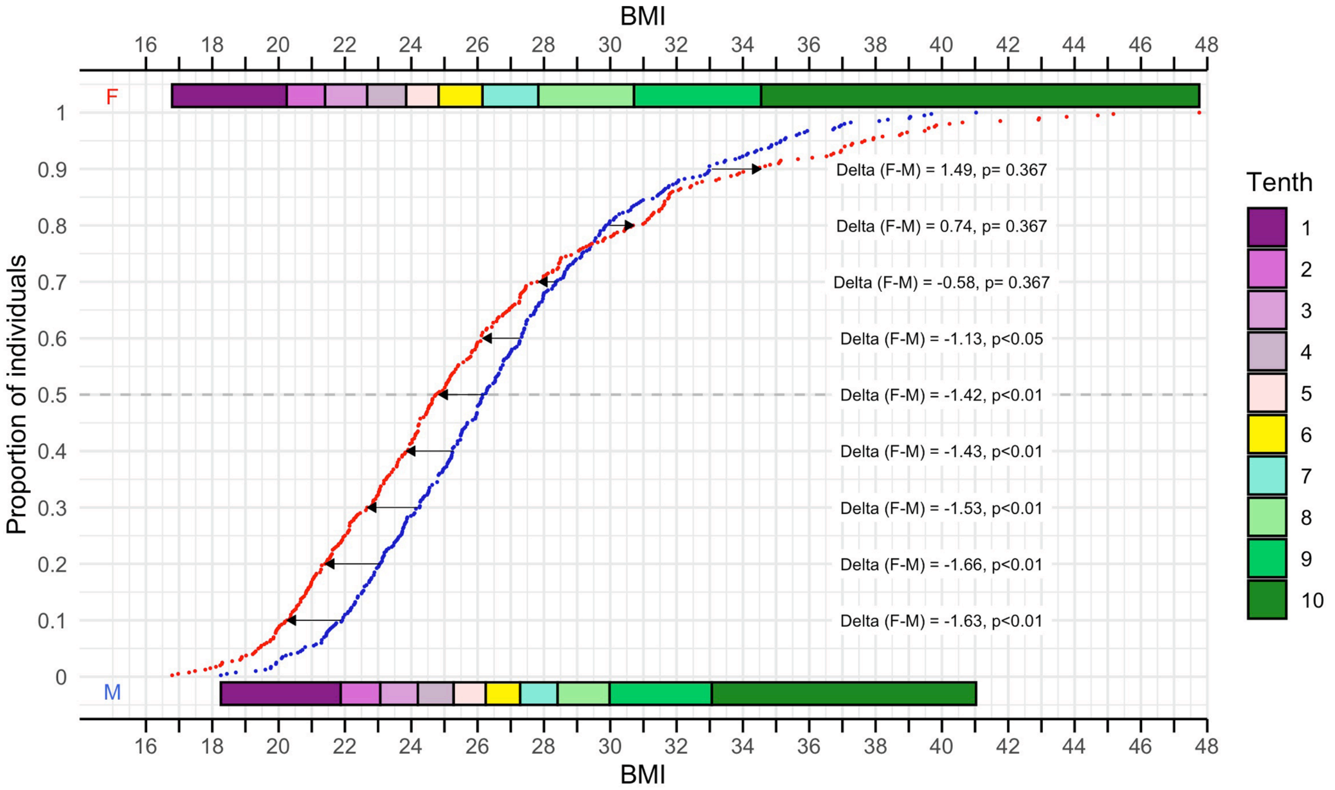


Fig. 6. Alternative graphical display for the shift function based on cumulative density functions (CDFs). BMI scores, organized in ascending order, are depicted for males (blue) and females (red). The x-axes show the BMI score range, while the y-axis indicates the proportion of cases with a BMI score equal to or lower than any given value. The distance between the resulting CDFs reflects male–female differences, akin to the classical shift-function. Arrows highlight decile-based comparisons, but the graph's grid allows for additional user-defined comparisons (see details and examples in the main text). Colored rectangles near the x-axis represent tenths of the BMI distributions, showcasing male–female differences in spread and skewness.

Firstly, without prior exposure or guidance, this visualization might be challenging to interpret, especially for non-expert audiences. Secondly, the outlook of the shift function (e.g., the sign of observed differences and of the slope) depends on which group is designated as the “reference” group, so this graphical representation may be more suitable for comparisons in which there is true control or reference group. Finally, this depiction is very informative of the between-groups differences, but it does not describe the groups themselves.

Fig. 6 presents an alternative graphical display for the shift function that may be more appropriate in the context of S/G-related research because it does not require designing a reference group and because shows all the individuals’ scores. Specifically, Fig. 6 illustrates the cumulative density functions (CDFs)[12] of BMI scores for males and females at the center of the plot. The distance between these distributions provides the same information than the classical depiction of the shift-function, the magnitude of the between group differences across the range of BMI scores. The deciles’ comparisons are graphically (arrows) and numerically summarized, whereas the values of the deciles themselves can be easily approximated from the colored rectangles included in the figure (that represent a breakdown of females’ and males’ distribution into its tenths). From these rectangles, the right-skewness of these distributions and the female-male differences in spread are also easily grasped.

However, the main strength of this figure resides on the fact that it does not only serve to describe the distributions or illustrate the results of the experimenter-chosen comparisons (in this case, the deciles’ comparisons). It can also be used as a computational device that enables readers to compare the BMI scores of females and males without other restrictions than those imposed by the axes’ scales.

To conduct within- or between-groups comparisons with this graph, one simply needs to use its grid as in the classic “battleship” game. Thus, for example, one can reproduce the between-groups decile comparisons (or extend them to other quantiles) by first projecting a horizontal line from the selected proportion of individuals in the y-axis (e.g., 0.5) to each of the two CDFs and then projecting the corresponding vertical lines to the x-axes to determine and compare the approximate BMI value in each group (in this case, 24.8 in females, 26.2 in males). Following the same procedure but projecting the vertical lines till the colored rectangles instead to the BMI axes, it can be found that around 65% of the males show BMI scores larger than the females’ median BMI then obtaining a non-parametric effect size similar to Cohen’s U3 (see footnote #8).

Comparisons can be performed the other way around too. That is, one can choose any BMI value from any of the two x-axes and project a vertical line to each of the CDF curves and then project horizontal lines to the y-axis to ascertain which proportion of females and males have BMI scores that are lower or equal to the selected BMI value. In this particular example, this second type of comparisons may have especial value because BMI scores are usually broken in health-relevant categories (underweight, BMI<18.5; normal weight, 18.5 to 24.9; overweight, 25 to 29.9; obese>=30). Thus, it can be interesting to know how many males and females fall within these BMI categories and whether their relative frequencies at any of these categories differ. Such estimations and can be performed by taking the limits of a category in the x-axis and projecting lines to the CDFs, and then project the lines to the y-axis. Despite the scale of the axes, it is easily seen that that there seem to be a larger proportion of females in the underweight and normal weight categories but probably not in the other two categories. In fact, when these proportions are compared (Newcombe, 1998), it is found that females fall more frequently than males in the categories of underweight (0.025 vs. 0.005, p=0.042) and normal weight (0.48 vs. 0.36, p<0.001) categories, less frequently in the overweight category (0.27 vs. 0.44, p<0.001), and with similar frequency in the obese category (0.22 vs. 0.19, p=0.338).

Overall, as shown by this example and some recent studies, the simultaneous comparison of several quantiles allows treating S/G-related categories as distributions, both within simple (Fig. 6, (Sanchis-Segura et al., 2023, 2022)) and factorial designs (Fig. 7, (Sebastián-Tirado et al., 2023)). This approach coupled with the inclusion of additional effect sizes and appropriate graphical depictions, provides a complete description of how the individuals included in these categories are, and also a nuanced and informative analysis of how and by how much they resemble and differ. It seems reasonable to propose that such detailed information can probably be more useful than that provided by averages when trying to incorporate S/G-related information to the comprehension of biological and behavioral phenomena and/or when searching and designing more individualized health-related interventions. Furthermore, the ability of the shift function to unveil complex patterns of relationships, which would be unnoticed or even masked when solely relying on average comparisons, should prompt a critical reflection on the potential limitations of mean comparisons in S/G-related research. Specifically, we should reflect about how many relevant findings may have been overlooked or mischaracterized due to overly narrow statistical approaches, and also whether comparing averages alone should be explicitly acknowledged as a limitation in bio-behavioral studies aimed at incorporating S/G-related information.

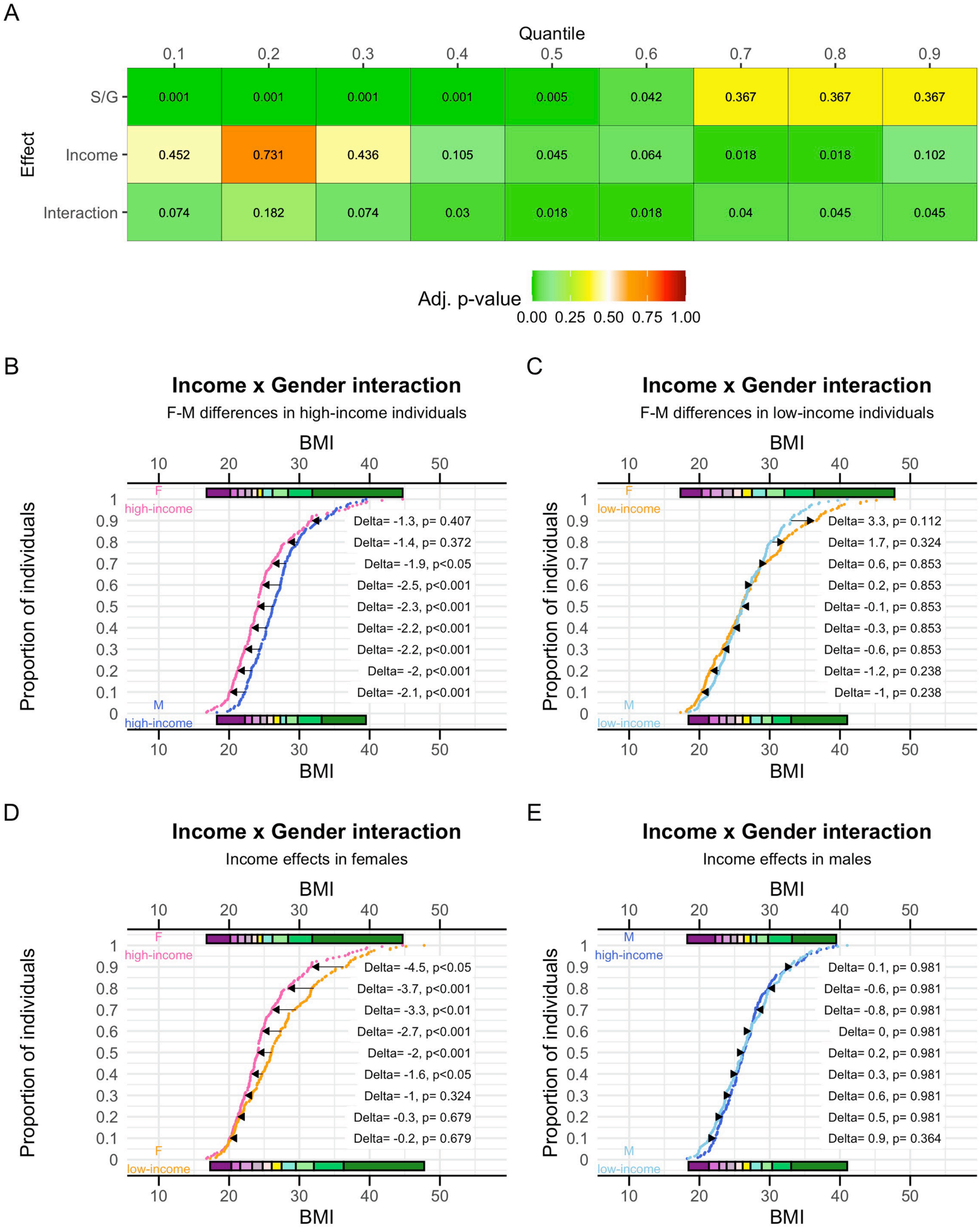
4. Conclusion

> *«Whenever I read statistical reports, I try to imagine my unfortunate contemporary, the Average Person, who, according to these reports, has 0.66 children, 0.032 cars, and 0.046 TVs».*
>
> Kató Lomb. Polyglot (2008)
>
> *«Far better an approximate answer to the right question, which is often vague, that an exact answer to the wrong question, which can always be made precise»*
>
> John Tukey. The future of data analysis (1962)

In this commentary, we have shortly reviewed the –often overlooked- statistical shortcomings of mean comparisons. We have shown that means can be misleading, but also that any average is an attempt to reduce complex data into a single number that –more often than not- may be unnecessary and inappropriate. Moreover, we have shown that average comparisons often lead to unwarranted generic statements about males and females. These generic statements are not only misleading but also potentially dangerous or harmful. As research in cognitive psychology has shown, generic statements require little evidence to be accepted as true, especially if coming from sources judged as trust worthy and/ or referring to group differences (Cimpian et al., 2010; Prasada and Dillingham, 2006). Furthermore, these statements are likely to be seen as characterizing the entire kind, prompt stereotypical views about the subjects of those statements (Cimpian et al., 2010; Gelman, 2004), and, when referring to biological properties, they evoke essentialist thinking (Noyes and Keil, 2019).

We have also proposed that mean comparisons might be especially inappropriate in the context of S/G-related studies. This is not only because of the perils of generic statements about the members of S/G-related categories, but also because mean comparisons do not approach, and may even separate, S/G-related biomedical research from

[12] *Cumulative distribution functions* (CDFs) are graphical representations that display cumulative proportions on the y-axis across the values of the variable of interest (x-axis). A cumulative proportion represents the proportion of scores that are equal lower than a given value of the variable of interest (in this case BMI). Cumulative proportions and quantiles are related but should not be confused. Quantiles are cut points that divide a variable in equally populated intervals (see note #11), while cumulative proportions indicate the proportion of data points below a given value within the dataset. Thus, for example, the cumulative proportion of the first quartile (Q1 or P25 if expressed as a percentile) is 0.25, meaning that 25% of the data lies below the value of Q1 and 75% above it.



(caption on next page)

Fig. 7. Application of the shift function in factorial designs. This figure employs the BMI data to illustrate the use of the shift function to analyze main effects and interactions within factorial designs, presenting an alternative approach to two-way ANOVAs that focuses on multiple quantiles rather than just mean comparisons. For these analyses, we introduced 'income' as an additional binary factor ('low income' vs. 'high income'). Thus, within this 2x2 design, decile comparisons were conducted between 1) females and males (with high- and low-income individuals combined within each S/G category); 2) high- and low-income groups (with females and males combined within each income level); and 3) high vs. low-income individuals within each S/G category. Panel A reports the p-values from these analyses, which yielded evidence enough to suggest a between factors' interaction at least at some quantiles. Panels B-E further detail this interaction by showing planed dyadic comparisons. Specifically, panels B and C confirm that S/G differences (M > F) are predominantly observed at low/ intermediate BMI scores and reveal that these S/G differences occur among high-income (panel B), but not among low-income, individuals (panel C). Panels D and E illustrate that this discrepancy occurs because 1) low income is associated with enhanced BMI scores in females (but not in males); and 2) This female-selective effect reduces the overall S/G-related effects, and at the highest BMI scores, it counteracts them. Taken together, these results allow concluding that income level moderates S/G differences in BMI and show again that S/G differences are not uniform across the entire BMI range, then reinforcing the notion that neither the BMI scores of males/ females nor their differences in this variable can be properly summarized by using a single number. Instead, more comprehensive and nuanced statistical approaches such as those illustrated here seem to be required. For methodological details and the R functions used in these comparisons refer to (Wilcox and Rousselet, 2023b) and (Wilcox, 2022), chapters 7 and 8.

its main goal –understanding the S/G-biased prevalence and manifestations of somatic, psychiatric, and neurological diseases to advance towards more tailored diagnostics and therapeutics. In fact, averages (and specially those obtained from broadly defined and non-randomly-assigned categories) come at the expense of the personalization which precision medicine specifically aims for: Precision medicine aims to account for individual variability, while averages do the opposite (mask or ignore variability). Therefore, just documenting differences between the averages of S/G-categories risks producing findings with little relevance to human health, but also of treating the individuals included in each of these categories as if they were equal or highly similar in the trait (s) considered, which often may not be the case (for an ampler discussion, see (Epstein, 2007; Galea and Lee, 2023; Richardson et al., 2015)).

Finally, we have tried to briefly illustrate that there are other analytical strategies that are informationally richer and that attend to both within- and between-groups variation by treating categorically collected information as continuous distributions. This approach is descriptive and not only comparative, and it conceives S/G-related factors as sources of individual variation and not just of between-group differences, hence making it more promising for advancing toward precision medicine. However, we do not think switching the way data are analyzed suffice to achieve this goal. It is just one step that must be integrated with other necessary improvements. In this regard, showing that males and females (and that all males and all females) are not the same is not the same than knowing what makes them to differ. Thus, as others have proposed (DiMarco et al., 2022; Jacklin, 1981; Richardson, 2022) and it seems to be increasingly recognized (Massa et al., 2023; Wierenga et al., 2023; Dubois et al., in press), S/G-related categories need to be replaced by the specific S/G-related factors which are suspected to operate –often in interaction with other variables- in each particular case (e.g., chromosomal complement, differential access to health resources, etc.). This dual commitment to replacing S/G categories with clearly defined S/G-related factors and adopting methodological approaches that transcend mere average comparisons seems a promising strategy to navigate the complexities of precision medicine and uncover the intricate variability within and between diverse populations.

5. Data availability statement

This study employed data from the open source 1200 Subject Release (S1200) of the Human Connectome Project (HCP). The access to this sample should be directly requested to the Washington University – University of Minnesota Consortium of the Human Connectome Project (WU-Minn HCP).

CRediT authorship contribution statement

Carla Sanchis-Segura: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Rand R. Wilcox:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Acion, L., Peterson, J.J., Temple, S., Arndt, S., 2006. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. Stat. Med. 25 https://doi.org/10.1002/sim.2256.

Aiken, L.S., West, S.G., Millsap, R.E., 2008. Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, west, sechrest, and Reno's (1990) survey of PhD programs in North America. Am. Psychol. 63 https://doi.org/10.1037/0003-066X.63.1.32.

Algina, J., Keselman, H.J., Penfield, R.D., 2005. An alternative to cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. Psychol. Methods 10. https://doi.org/10.1037/1082-989X.10.3.317.

Allen, E.A., Erhardt, E.B., Calhoun, V.D., 2012. Data visualization in the neurosciences: overcoming the curse of dimensionality. Neuron. https://doi.org/10.1016/j.neuron.2012.05.001.

Altemus, M., Sarvaiya, N., Neill Epperson, C., 2014. Sex differences in anxiety and depression clinical perspectives. Front. Neuroendocrinol. https://doi.org/10.1016/j.yfrne.2014.05.004.

Altman, D.G., Bland, J.M., 1995. Statistics notes: absence of evidence is not evidence of absence. BMJ. https://doi.org/10.1136/bmj.311.7003.485.

Anastasi, A., 1981. Sex differences: historical perspectives and methodological implications. Dev. Rev. https://doi.org/10.1016/0273-2297(81)90017-4.

Andrade, C., 2020. Understanding the difference between Standard Deviation and Standard error of the mean, and knowing when to use which. Indian J. Psychol. Med. 42 https://doi.org/10.1177/0253717620933419.

Bartz, D., Chitnis, T., Kaiser, U.B., Rich-Edwards, J.W., Rexrode, K.M., Pennell, P.B., Goldstein, J.M., O'Neal, M.A., Leboff, M., Behn, M., Seely, E.W., Joffe, H., Manson, J. E., 2020. Clinical Advances in Sex- and Gender-Informed Medicine to Improve the Health of All: A Review. JAMA Intern. Med. https://doi.org/10.1001/jamainternmed.2019.7194.

Blanca, M.J., Arnau, J., López-Montiel, D., Bono, R., Bendayan, R., 2013. Skewness and kurtosis in real data samples. Methodology 9. https://doi.org/10.1027/1614-2241/a000057.

Blanca, M.J., Alarcón, R., Bono, R., 2018. Current practices in data analysis procedures in psychology: what has changed? Front. Psychol. 9 https://doi.org/10.3389/fpsyg.2018.02558.

Blume, J.D., Greevy, R.A., Welty, V.F., Smith, J.R., Dupont, W.D., 2019. An introduction to second-generation p-values. Am. Stat. 73 https://doi.org/10.1080/00031305.2018.1537893.

Caponi, S., 2013. Quetelet, the average man and medical knowledge. Hist. Ciencias, Saude - Manguinhos 20. https://doi.org/10.1590/S0104-597020130003000006.

Carver, R.P., 1978. The case against satistical significance testing. Havard Educ. Rev. https://doi.org/10.4097/kjae.2017.70.2.144.

Carver, R.P., 1993. The case against statistical significance testing, revisited. J. Exp. Educ. https://doi.org/10.1080/00220973.1993.10806591.

Cimpian, A., Brandone, A.C., Gelman, S.A., 2010. Generic statements require little evidence for acceptance but have powerful implications. Cogn. Sci. 34, 1452. https://doi.org/10.1111/J.1551-6709.2010.01126.X.

Cliff, N., 2014. Ordinal methods for behavioral data analysis. Ordinal Methods for Behavioral Data Analysis. https://doi.org/10.4324/9781315806730.

Cobb, G.W., 2007. The introductory statistics course: a ptolemaic curriculum? Technol. Innov. Stat. Educ. 1 https://doi.org/10.5070/t511000028.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Lawrence Erlbaum, Hillsdale, NJ.

Cohen, J., 1994. The earth is round (p &lt;.05). Am. Psychol. https://doi.org/10.1037/0003-066X.49.12.997.

Cox, D.R., 2006. Causality: some statistical aspects. J. r. Stat. Soc. Ser. A (statistics Soc. https://doi.org/10.2307/2982962.

Cumming, G., 2012. Understanding the new statistics: effect sizes, confidence intervals and meta-analysis, 1st ed. Routledge, New York.

Davies, H.T.O., 1998. Describing and estimating: Use and abuse of standard deviations and standard errors. Hosp. Med. 59.

DeCasien, A.R., Guma, E., Liu, S., Raznahan, A., 2022. Sex differences in the human brain: a roadmap for more careful analysis and interpretation of a biological reality. Biol. Sex Differ. 13 https://doi.org/10.1186/s13293-022-00448-w.

DiMarco, M., Zhao, H., Boulicault, M., Richardson, S.S., 2022. Why "sex as a biological variable" conflicts with precision medicine initiatives. Cell Reports Med. https://doi.org/10.1016/j.xcrm.2022.100550.

Doksum, K.A., Sievers, G.L., 1976. Plotting with confidence: graphical comparisons of two populations. Biometrika 63. https://doi.org/10.1093/biomet/63.3.421.

Eliot, L., Ahmed, A., Khan, H., Patel, J., 2021. Dump the "dimorphism": comprehensive synthesis of human brain studies reveals few male-female differences beyond size. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2021.02.026.

Eliot, L., Beery, A.K., Jacobs, E.G., LeBlanc, H.F., Maney, D.L., McCarthy, M.M., 2023. Why and how to account for sex and gender in brain and behavioral Research. J. Neurosci. 43 https://doi.org/10.1523/JNEUROSCI.0020-23.2023.

Epstein, S., 2007. Inclusion: the politics of difference in Medical Research. University of Chicago Press, Chicago.

Field, A.P., Wilcox, R.R., 2017. Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers. Behav. Res. Ther. 98 https://doi.org/10.1016/j.brat.2017.05.013.

Friedmann, J.M., Elasy, T., Jensen, G.L., 2001. The relationship between body mass index and self-reported functional limitation among older adults: a gender difference. J. Am. Geriatr. Soc. 49 https://doi.org/10.1046/j.1532-5415.2001.49082.x.

Galea, L.A.M., Lee, B.H., de leon, R.G., Rajah, M.N., Einstein, G., 2023. Beyond sex and gender differences: The case for women's health research, in: Principles of Gender-Specific Medicine: Sex and Gender-Specific Biology in the Postgenomic Era. https://doi.org/10.1016/B978-0-323-88534-8.00045-6.

Galea, L.A.M., Choleris, E., Albert, A.Y.K., McCarthy, M.M., Sohrabji, F., 2020. The promises and pitfalls of sex difference research. Front. Neuroendocrinol. https://doi.org/10.1016/j.yfrne.2019.100817.

Garcia-Sifuentes, Y., Maney, D.L., 2021. Reporting and misreporting of sex differences in the biological sciences. Elife 10. https://doi.org/10.7554/eLife.70817.

Gelman, S.A., 2004. Learning words for kinds: generic noun phrases in acquisition. Weav. a Lex.

Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The Null Ritual, in: The Sage Handbook of Quantitative Methodology for the Social Sciences.

Glass, G.V., McGaw, S.B.S., Lee, M., 1982. Meta-analysis in social Research. Beverly Hills Sage Publ. https://doi.org/10.2307/1165349.

Grayson, D., 2004. Some myths and legends in quantitative psychology. Underst. Stat. 3, 101–134. https://doi.org/10.1207/s15328031us0302_3.

Greenland, S., 2019. Valid P-values behave exactly as they should: some misleading Criticisms of P-values and their resolution with S-values. Am. Stat. 73 https://doi.org/10.1080/00031305.2018.1529625.

Greenwald, A.G., Gonzalez, R., Harris, R.J., Guthrie, D., 1996. Effect sizes and p values: what should be reported and what should be replicated? Psychophysiology. https://doi.org/10.1111/j.1469-8986.1996.tb02121.x.

Grissom, R.J., Kim, J.J., 2012. Effect sizes for research: Univariate and multivariate applications, second edition, Effect Sizes for Research: Univariate and Multivariate Applications, Second Edition. Routledge, Multivariate application tests. https://doi.org/10.4324/9780203803233.

Grue, L., Heiberg, A., 2006. Notes on the history of normality – reflections on the work of quetelet and galton. Scand. J. Disabil. Res. 8 https://doi.org/10.1080/15017410600608491.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J., 1986. Robust statistics. Wiley, New York.

Hanel, P.H.P., Mehler, D.M.A., 2019. Beyond reporting statistical significance: identifying informative effect sizes to improve scientific communication. Public Underst. Sci. 28 https://doi.org/10.1177/0963662519834193.

Heene, M., Ferguson, C.J., 2017. Psychological science's aversion to the null, and why many of the things you think are true. Aren't, in: Psychological Science under Scrutiny. https://doi.org/10.1002/9781119095910.ch3.

Hirschauer, N., Grüner, S., Mußhoff, O., 2022. Better inference in the 21st century: a world beyond $p < 0.05$, in. American Statistician. American Statistical Association 113–117. https://doi.org/10.1007/978-3-030-99091-6_8.

Hoekstra, R., Finch, S., Johnson, A., 2006. Probability as certainty: dichotomous thinking and the misuse of p values. Psychon. Bull. Rev. https://doi.org/10.3758/BF03213921.

Hoekstra, R., Kiers, H.A.L., Johnson, A., 2012. Are assumptions of well-known statistical techniques checked, and why (not)? Front. Psychol. 3 https://doi.org/10.3389/fpsyg.2012.00137.

Högel, J., Schmid, W., Gaus, W., 1994. Robustness of the Standard Deviation and other measures of dispersion. Biometrical J. 36 https://doi.org/10.1002/bimj.4710360403.

Huber, P.J., Ronchetti, E., 2009. Robust statistics, 2nd ed. Wiley.

Hyde, J.S., 2014. Gender Similarities and differences. SSRN. https://doi.org/10.1146/annurev-psych-010213-115057.

Jacklin, C.N., 1981. Methodological issues in the study of sex-related differences. Dev. Rev. 1, 266–273.

Joel, D., 2011. Male or Female? Brains are Intersex. Front. Integr. Neurosci. https://doi.org/10.3389/fnint.2011.00057.

Joel, D., McCarthy, M.M., 2017. Incorporating sex as a biological Variable in neuropsychiatric Research: where are we now and where should we be? Neuropsychopharmacology. https://doi.org/10.1038/npp.2016.79.

Jones, L.V., Tukey, J.W., 2000. A sensible formulation of the significance test. Psychol. Methods 5. https://doi.org/10.1037/1082-989X.5.4.411.

Keselman, H.J., Wilcox, R.R., Othman, A.R., Fradette, K., 2002. Trimming, transforming statistics, and bootstrapping: circumventing the biasing effects of heterescedasticity and nonnormality. J. Mod. Appl. Stat. Methods 1. https://doi.org/10.22237/jmasm/1036109820.

Kline, R.B., 2013. Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.)., Beyond significance testing: Statistics reform in the behavioral sciences (2nd ed.). https://doi.org/10.1037/14136-000.

Krieger, N., 2003. Genders, sexes, and health: what are the connections - and why does it matter? Int. J. Epidemiol. https://doi.org/10.1093/ije/dyg156.

Lakens, D., Scheel, A.M., Isager, P.M., 2018. Equivalence testing for psychological Research: a tutorial. Adv. Methods Pract. Psychol. Sci. 1 https://doi.org/10.1177/2515245918770963.

Lane, D.M., Sándor, A., 2009. Designing better graphs by including distributional information and integrating words, numbers, and images. Psychol. Methods. https://doi.org/10.1111/j.1708-8208.2011.00364.x.

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49 https://doi.org/10.1016/j.jesp.2013.03.013.

Lillie, F., 1939. General biological introduction. In: Allen, E. (Ed.), Sex and Internal Secretions: A Survey of Recent Research. Williams & Wilkinks, Baltimore, pp. 3–14.

Maney, D.L., 2016. Perils and pitfalls of reporting sex differences. Philos. Trans. R. Soc. B Biol. Sci. https://doi.org/10.1098/rstb.2015.0119.

Mastorci, F., Doveri, C., Trivellini, G., Casu, A., Bastiani, L., Pingitore, A., Vassalle, C., 2020. Sex differences in body mass index, mediterranean diet adherence, and physical activity level among italian adolescents. Heal. Behav. Policy Rev. 7 https://doi.org/10.14485/HBPR.7.6.8.

Matthews, R.A.J., 2019. Moving Towards the post $p < 0.05$ era via the analysis of credibility. Am. Stat. 73 https://doi.org/10.1080/00031305.2018.1543136.

Mauvais-Jarvis, F., Bairey Merz, N., Barnes, P.J., Brinton, R.D., Carrero, J.J., DeMeo, D. L., De Vries, G.J., Epperson, C.N., Govindan, R., Klein, S.L., Lonardo, A., Maki, P.M., McCullough, L.D., Regitz-Zagrosek, V., Regensteiner, J.G., Rubin, J.B., Sandberg, K., Suzuki, A., 2020. Sex and gender: modifiers of health, disease, and medicine. Lancet. https://doi.org/10.1016/S0140-6736(20)31561-0.

McCarthy, M.M., Arnold, A.P., Ball, G.F., Blaustein, J.D., De Vries, G.J., 2012. Sex differences in the brain: the not so inconvenient truth. J. Neurosci. https://doi.org/10.1523/JNEUROSCI.5372-11.2012.

McCarthy, M.M., Konkle, A.T.M., 2005. When is a sex difference not a sex difference? Front. Neuroendocrinol. https://doi.org/10.1016/j.yfrne.2005.06.001.

McGraw, K.O., Wong, S.P., 1992. A common language effect size statistic. Psychol. Bull. https://doi.org/10.1037/0033-2909.111.2.361.

Micceri, T., 1989. The unicorn, the Normal curve, and other improbable creatures. Psychol. Bull. 105 https://doi.org/10.1037/0033-2909.105.1.156.

Newcombe, R.G., 1998. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Stat. Med. 17 https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.

Noyes, A., Keil, F.C., 2019. Generics designate kinds but not always essences. Proc. Natl. Acad. Sci. U. S. A. 116 https://doi.org/10.1073/pnas.1900105116.

Ozdemir, A.F., Wilcox, R.R., Yildiztepe, E., 2013. Comparing measures of location: some small-sample results when distributions differ in skewness and kurtosis under heterogeneity of variances. Commun. Stat. Simul. Comput. 42 https://doi.org/10.1080/03610918.2011.636163.

Pastore, M., Calcagnì, A., 2019. Measuring distribution similarities between samples: a distribution-free overlapping index. Front. Psychol. https://doi.org/10.3389/fpsyg.2019.01089.

Pek, J., Wong, O., Wong, A.C.M., 2018. How to address non-normality: a taxonomy of approaches, reviewed, and illustrated. Front. Psychol. https://doi.org/10.3389/fpsyg.2018.02104.

Pinares-Garcia, P., Stratikopoulos, M., Zagato, A., Loke, H., Lee, J., 2018. Sex: A Significant Risk Factor for Neurodevelopmental and Neurodegenerative Disorders. Brain Sci. 2018, Vol. 8, Page 154 8, 154. https://doi.org/10.3390/BRAINSCI8080154.

Prasada, S., Dillingham, E.M., 2006. Principled and statistical connections in common sense conception. Cognition 99. https://doi.org/10.1016/j.cognition.2005.01.003.

Pratt, J.W., 1964. Robustness of some procedures for the two-sample location problem. J. Am. Stat. Assoc. 59 https://doi.org/10.2307/2283092.

Rechlin, R.K., Splinter, T.F.L., Hodges, T.E., Albert, A.Y., Galea, L.A.M., 2022. An analysis of neuroscience and psychiatry papers published from 2009 and 2019 outlines opportunities for increasing discovery of sex differences. Nat. Commun. 13 https://doi.org/10.1038/s41467-022-29903-3.

Reis, H.T., Carothers, B.J., 2014. Black and white or shades of gray: are gender differences categorical or dimensional? Curr. Dir. Psychol. Sci. 23 https://doi.org/10.1177/0963721413504105.

Rice, K.M., Krakauer, C.A., 2023. Three-decision methods: a sensible formulation of significance tests-and much else. Annu. Rev. Stat. Its Appl. https://doi.org/10.1146/annurev-statistics-033021-111159.

Richardson, S.S., 2022. Sex contextualism. Philos. Theory, Pract. Biol. 14 https://doi.org/10.3998/ptpbio.2096.

Richardson, S.S., Reiches, M., Shattuck-Heidorn, H., Labonte, M.L., Consoli, T., 2015. Opinion: focus on preclinical sex differences will not address women's and men's health disparities. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1516958112.

Rich-Edwards, J.W., Kaiser, U.B., Chen, G.L., Manson, J.A.E., Goldstein, J.M., 2018. Sex and gender differences research design for basic, clinical, and population studies: essentials for investigators. Endocr. Rev. https://doi.org/10.1210/er.2017-00246.

Rich-Edwards, J.W., Maney, D.L., 2023. Best practices to promote rigor and reproducibility in the era of sex-inclusive research. Elife e90623. https://doi.org/10.7554/eLife.90623.

Ritchie, S.J., Cox, S.R., Shen, X., Lombardo, M.V., Reus, L.M., Alloza, C., Harris, M.A., Alderson, H.L., Hunter, S., Neilson, E., Liewald, D.C.M., Auyeung, B., Whalley, H.C., Lawrie, S.M., Gale, C.R., Bastin, M.E., McIntosh, A.M., Deary, I.J., 2018. Sex differences in the adult human brain: evidence from 5216 UK biobank Participants. Cereb. Cortex. https://doi.org/10.1093/cercor/bhy109.

Roskies, A.L., 2007. Are neuroimages like photographs of the brain?, in: Philosophy of Science. https://doi.org/10.1086/525627.

Rousseeuw, P.J., Stahel, W.A., 2011. Robust statistics: the approach based on influence functions. John Wiley & Sons Inc.

Rousselet, G.A., Pernet, C.R., Wilcox, R.R., 2017. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. Eur. J. Neurosci. https://doi.org/10.1111/ejn.13610.

Ruscio, J., 2008. A probability-based measure of effect size: robustness to base rates and other factors. Psychol. Methods. https://doi.org/10.1037/1082-989X.13.1.19.

Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á.J., Félix, S., Forn, C., 2022. Beyond "sex prediction": estimating and interpreting multivariate sex differences and similarities in the brain. Neuroimage 257, 119343. https://doi.org/10.1016/j.neuroimage.2022.119343.

Sanchis-Segura, C., Cruz-Gómez, Á.J., Esbrí, S.F., Tirado, A.S., Arnett, P.A., Forn, C., 2023. Multiple sclerosis and depression: translation and Adaptation of the spanish version of the Chicago multiscale depression inventory and the study of factors associated with depressive symptoms. Arch. Clin. Neuropsychol. 38 https://doi.org/10.1093/arclin/acac096.

Seaman, M.A., Serlin, R.C., 1998. Equivalence confidence intervals for two-group Comparisons of means. Psychol. Methods 3. https://doi.org/10.1037/1082-989X.3.4.403.

Sebastián-Tirado, A., Félix-Esbrí, S., Forn, C., Sanchis-Segura, C., 2023. Are gender-science stereotypes barriers for women in science, technology, engineering, and mathematics? exploring when, how, and to whom in an experimentally-controlled setting. Front. Psychol. 14 https://doi.org/10.3389/fpsyg.2023.1219012.

Silverman, M.P., Lipscombe, T.C., 2022. Exact statistical distribution of the body mass index (BMI): analysis and Experimental confirmation. Open J. Stat. 12 https://doi.org/10.4236/ojs.2022.123022.

Speelman, C.P., McGann, M., 2013. How mean is the mean? Front. Psychol. 4 https://doi.org/10.3389/fpsyg.2013.00451.

Springer, K.W., Mager Stellman, J., Jordan-Young, R.M., 2012. Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. Soc. Sci. Med. https://doi.org/10.1016/j.socscimed.2011.05.033.

Stachenfeld, N.S., Mazure, C.M., 2022. Precision medicine requires understanding how both sex and gender influence health. Cell. https://doi.org/10.1016/j.cell.2022.04.012.

Staudte, R., Sheather, S.J., 1990. Robust estimation and testing. Wiley, New York.

Sullivan, L.M., Weinberg, J., Keaney, J.F., 2016. Common statistical pitfalls in basic science research. J. Am. Heart Assoc. https://doi.org/10.1161/JAHA.116.004142.

Thompson, B., 2008. Foundations of behavioral statistics: an insight-based approach. Guilford Press, New York.

Thompson, B., 2011. Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. Best Practices in Quantitative Methods. https://doi.org/10.4135/9781412995627.d21.

Tsang, S., Duncan, G.E., Dinescu, D., Turkheimer, E., 2018. Differential models of twin correlations in skew for body-mass index (BMI). PLoS One 13. https://doi.org/10.1371/journal.pone.0194968.

Tukey, J.W., 1991. The philosophy of multiple Comparisons. Stat. Sci. https://doi.org/10.1214/ss/1177011945.

Tukey, J.W., 1960. A survey of sampling from contaminated normal distributions, in: Olkin, I., Ghurye, W., Hoeffding, W., Madow, W, Mann, H. (Eds.), Contributions to Probability and Statistics. pp. 448–485.

Vijayalakshmi, P., Thimmaiah, R., Reddy, S.S.N., Kathyayani, B.V., Gandhi, S., BadaMath, S., 2017. Gender differences in body mass index, body weight perception, weight satisfaction, disordered eating and weight control strategies among Indian Medical and nursing undergraduates. Investig. y Educ. En Enferm. 35 https://doi.org/10.17533/udea.iee.v35n3a04.

Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond Bar and line graphs: time for a new data presentation Paradigm. PLoS Biol. https://doi.org/10.1371/journal.pbio.1002128.

Weissgerber, T.L., Garcia-Valencia, O., Garovic, V.D., Milic, N.M., Winham, S.J., 2018. Why we need to report more than 'data were analyzed by t-tests or ANOVA'. Elife 7. https://doi.org/10.7554/eLife.36163.

White, J., Tannenbaum, C., Klinge, I., Schiebinger, L., Clayton, J., 2021. The integration of sex and gender considerations into biomedical research: lessons from international funding agencies. J. Clin. Endocrinol. Metab. 106 https://doi.org/10.1210/clinem/dgab434.

Wilcox, R.R., 1998. How many discoveries have been lost by ignoring modern statistical methods? Am. Psychol. 53 https://doi.org/10.1037/0003-066X.53.3.300.

Wilcox, R.R., 2006. Graphical methods for assessing effect size: some alternatives to cohen's d. J. Exp. Educ. https://doi.org/10.3200/JEXE.74.4.351-367.

Wilcox, R.R., 2021. Inferences about a probabilistic measure of effect size when dealing with more than two groups. J. Data Sci. 9 https://doi.org/10.6339/jds.201107_09(3).0010.

Wilcox, R.R., Rousselet, G.A., 2023b. Preprint: A Quantile Shift Approach To Main Effects And Interactions In A 2-By-2 Design. https://doi.org/https://doi.org/10.48550/arXiv.2305.12366.

Wilcox, R.R., Keselman, H.J., 2003. Modem robust data analysis methods: measures of central tendency. Psychol. Methods 8. https://doi.org/10.1037/1082-989X.8.3.254.

Wilcox, R.R., Rousselet, G.A., 2018. A guide to robust statistical methods in neuroscience. Curr. Protoc. Neurosci. https://doi.org/10.1002/cpns.41.

Wilcox, R.R., Rousselet, G.A., 2023a. An updated guide to robust statistical methods in neuroscience. Curr. Protoc. 3 https://doi.org/10.1002/cpz1.719.

Wilcox, R.R., Serang, S., 2017. Hypothesis testing, p values, confidence intervals, measures of effect size, and bayesian methods in light of modern robust techniques. Educ. Psychol. Meas. 77 https://doi.org/10.1177/0013164416667983.

Wilcox, R.R., 2022. Introduction to Robust Estimation and Hypothesis Testing, 5th. ed, Introduction to Robust Estimation and Hypothesis Testing. Academic Press. https://doi.org/10.1016/C2019-0-01225-3.

Wilcox, R.R., 2023. A Guide to Robust Statistical Methods, 1st ed. Springer Cham. https://doi.org/978-3-031-41712-2.

Zell, E., Krizan, Z., Teeter, S.R., 2015. Evaluating gender similarities and differences using metasynthesis. Am. Psychol. 70 https://doi.org/10.1037/a0038208.