BMC Medical Research
Methodology

RESEARCH

Open Access

Agreement test of *P* value versus Bayes factor for sample means comparison: analysis of articles from the Angle Orthodontist journal

Natchalee Srimaneekarn[1], Pattamon Leelachaikul[2], Sasipa Thiradilok[3] and Somchai Manopatanakul[3*]

Abstract

Background Researchers are cautioned against misinterpreting the conventional *P* value, especially while implementing the popular t test. Therefore, this study evaluated the agreement between the *P* value and Bayes factor (BF_{01}) results obtained from a comparison of sample means in published orthodontic articles.

Methods Data pooling was undertaken using the modified PRISMA flow diagram. Per the inclusion criteria applied to The Angle Orthodontist journal for a two-year period (November 2016 to September 2018), all articles that utilised the t test for statistical analysis were selected. The agreement was evaluated between the *P* value and Bayes factor set at 0.05 and 1, respectively. The percentage of agreement and Kappa coefficient were calculated. Plotting of effect size against *P* value and BF_{01} was analysed.

Results From 265 articles, 82 utilised the t test. Of these, only 37 articles met the inclusion criteria. The study identified 793 justifiable t tests (438 independent-sample and 355 dependent-sample t tests) for which the agreement percentage and Kappa coefficient were found to be 93.57% and 0.87, respectively. However, when anecdotal evidence ($1/3 < BF_{01} < 3$) was considered, almost half of the studies missed statistical significance. Furthermore, two-thirds of the significantly reported *P* values ($0.01 < P < 0.05$; 30 independent-sample and 20 dependent-sample t tests) showed only anecdotal evidence ($1/3 < BF_{01} < 1$). Moreover, BF_{01} indicated moderate evidence ($BF_{01} > 3$) for approximately one-third of the total studies, with nonsignificant *P* values ($P > 0.05$). Furthermore, accompanying the *P* values, the effect sizes, especially for studies with independent-sample t tests, were very high with a strong potential to show substantive significance. Although it is best to extend the statistical calculation of a doubted *P* value (just below 0.05), especially for orthodontic innovation, orthodontists may reach a balanced decision relying on cephalometric measurements.

Conclusions The Kappa coefficient indicated perfect agreement between the two methods. BF_{01} restricted this judgement to approximately half of them, with two-thirds of these studies showing nonsignificant *P* values. Simple extensions of statistical calculations, especially effect size and $BF_{01,}$ can be useful and should be considered when finalising statistical analyses, especially for orthodontic studies without cephalometric analysis.

Keywords Agreement test, Bayes factor, *P* value, Effect side, Orthodontics

*Correspondence:
Somchai Manopatanakul
msomchai@rocketmail.com
[1] Department of Anatomy, Faculty of Dentistry, Mahidol University, Bangkok, Thailand
[2] Pu-Kradueng Hospital, Pu-Kradueng, Loie, Thailand
[3] Department of Advanced General Dentistry, Faculty of Dentistry, Mahidol University, 6 Yothi Street, Rachtewi, 10400 Bangkok, Thailand

Background

Statistics can be defined as the science of analysing data and drawing conclusions from situations that involve uncertainty. In most cases, uncertainty results from the impossibility or impracticality of studying the entire population [1]. Logically obtaining evidence resulting from the uncertainty of the experiment should depend only on the likelihood principle [2]. Herein, two philosophies of statistical analysis are compared: frequentist and Bayesian. In frequentist statistical inference, the *P* value is the probability of obtaining a test result that is at least as extreme as the observed results of the statistical hypothesis test (particularly in the scope of this study, the t test) assuming the null hypothesis is true [3] (see Eqs. 1 and 2 in the supplemental file). In other words, it is the probability of obtaining a false-positive (Type I error) from the observed data. Fisher originally showed the computation of significance via a continuous quantification of the *P* value. However, hypothesis testing was proposed by Neyman and Pearson [4, 5], where the outcome of a test was based on a dichotomous decision to show evidence in favour of only one hypothesis. The *P* value has gained widespread acceptance for comparison with the significance level (α). This α, set before the study, is the level of the acceptable false-positive rate, while the false negative (Type II error) is minimised [6].

In summary, the *P* value proposed by Fisher, although incompatible with hypothesis testing, was deeply intertwined with the method, as it reveals the probability of errors. Two types of errors exist: a false-positive, which considers two treatment results differently when they are the same (aforementioned as a Type I error or an α error); and a false negative, which considers two treatment results as the same when in fact they are different (Type II error or a β error) [4]. Both the *P* value and hypothesis testing do not measure the evidence but only their statistical significance. Fisher [4] proposed the term ‘significant’ to convey a meaning quite close to the word’s common language interpretation—something worthy of notice. Thus, ‘significant’ is merely worthy of attention in the form of meriting further experimentation but not proof in and of itself [7].

Many studies have warned about the misuse and misinterpretation of the *P* value, emphasising how little information this concept conveys [4, 6–8]. The misinterpretation of the *P* value was documented in the works of psychologists [1], statistical instructors [9], statisticians [10], medicine residents [11], and dentists [12]. Unquestionably, orthodontists are no exception [13]. Although statisticians have long questioned the *P*-value fallacy [14, 15], the t test (with *P* value) and hypothesis testing have been the most popular techniques of inferential statistics used in orthodontic research. Orthodontists routinely perform cephalometric analysis. This analysis comprises approximately 12 to 35 measurements per radiograph. Accordingly, the assessment of treatment results is a comparison of these measurements. In 2017 alone, among the 923 pages of The Angle Orthodontists journal, there were more than 430 t tests. Therefore, for every two pages, one t test (with *P* value) was reported.

Recently, statisticians have signed a petition to bar statistical significance [16–20]. Goodman was concerned about the current lack of evidence-based statistical inference and widespread error in drawing conclusions [6]. He also cited *P* value misconceptions and the possible consequences of this improper understanding [7]. Bayesian estimation was ultimately proposed as a prominent alternative to validate the classical single-number statistical report of the *P* value [10–22].

Bayes’ theorem was first introduced by Thomas Bayes and has been further developed for more than 200 years [23]. Bayes’ theorem is expressed mathematically. (Please see Eqs. 3 and 4 in the supplemental file.)

Bayesian two-sample tests yield slightly improved Type I error rates at the cost of marginally higher Type II error rates. As it does not violate the likelihood principle, Bayesian inference is essentially considered superior to frequentist procedures. Furthermore, the frequentist theory (grounded in average performance) is deemed unrealistic from the Bayesian perspective. Upon scrutiny, Bayesian philosophy provides an abundance of information. The growing number of studies and applications, such as Jeffery’s Bayes factor (BF_{01}) and Krushke’s region of practical equivalence (ROPE), shows a promising future [22, 24]. In addition, the Bayes factor on its own quantifies the evidence of statistical judgements in both directions. However, prior evocation clarification is required to bolster impartiality. A detailed comparison has been cogently documented [25].

To provide an alternative to the *P* value, Jeffreys [24] developed the Bayes factor (or, more specifically, BF_{10}). It is used to designate the relative strength of evidence for two theories: H_1 (alternative hypothesis) and H_0 (null hypothesis). The subscripts one and zero in BF_{10} indicate the alternative hypothesis over the null hypothesis. In contrast, BF_{01} indicates its inverse ratio ($BF_{01} = 1/BF_{10}$), which is in favour of the null hypothesis. As advocated by Kass and Raferty, the Bayes factor is not just a number that represents the evidential result of one experiment that is often used as a ‘scientific label’ to sway the interpretation to suit the external evidence or author’s belief [26]. It integrates background knowledge (biological understanding and previous research, a priori or often termed the *prior*) into its analysis. The Bayes factor was also supported by Goodman for its uncomplicated interpretability [6]. In this study, ‘Cauchy prior’ to the effect size (δ ~ Cauchy), as described by Rouder and colleagues [21], was calculated.

While Fisher defined the *P* value as the probability of obtaining a result equal to or more extreme than the observed results of a statistical hypothesis test (under the assumption of no effect or no difference), he also stated that it should be exercised as a nonquantifiable process of drawing conclusions from observations. As previously mentioned, the *P* value does not consider the size of the observed effect. Wasserman attempted to grade the level of interpretation of the *P* value, as shown in Table 1 [27]. However, studies with large sample sizes may yield significant *P* values with only a small effect. Here, effect size comes into play. While the Bayes factor can quantify evidence in favour of the null hypothesis, the *P* value requires an effect size to quantify the magnitude of differences (within the scope of this study). Therefore, exploring these three parameters—*P* value, effect size, and BF_{01}—in orthodontic studies is interesting.

Methods

In recent studies, statisticians have advocated the Bayes factor hypothesis test for t test confirmation. To avoid the misinterpretation fallacy of *P* values, orthodontic researchers can consider using BF_{01} as a quick shortcut to richer (including *prior*) information. Moreover, the prior may facilitate a better understanding of orthodontic research samples and aid in the validation of *P* values. Among the three major orthodontic journals, The Angle Orthodontist journal is the only noncommercial and open-access journal that publishes statistical evaluations of its own articles [30]. Therefore, this study reexamined articles from The Angle Orthodontist journal that used *P* value with a t test by BF_{01}. The test agreement and properties of both the *P* value and BF_{01} were evaluated in detail. The effect size was calculated, and its relationships with the *P* value and BF_{01} are shown.

Table 1 Evidence categorisation for *P* value, effect size, and Bayes factor. Modified Jeffreys' Bayes factor cutoffs indicating evidence categorised for Bayes factor BF_{01} [27–29]

Measures	Description
P value	
< 0.001	Decisive evidence against the null hypothesis
0.001 – 0.01	Substantive evidence against the null hypothesis
0.01 – 0.05	Positive evidence against the null hypothesis
> 0.05	No evidence against the null hypothesis
Effect size	
< 0.2	Small effect size
0.2 – 0.5	Small to medium effect size
0.5 – 0.8	Medium to large effect size
0.8	Large to very large effect size
Bayes $factor_{01}$	
< 1/100	Extreme evidence for the alternative hypothesis
1/30 – 1/100	Very strong evidence for the alternative hypothesis
1/10 – 1/30	Strong evidence for the alternative hypothesis
1/3 – 1/10	Moderate evidence for the alternative hypothesis
1 – 1/3	Anecdotal evidence for the alternative hypothesis
1	No evidence
1 – 3	Anecdotal evidence for the null hypothesis
3 – 10	Moderate evidence for the null hypothesis
10 – 30	Strong evidence for the null hypothesis
30 – 100	Very strong evidence for the null hypothesis
> 100	Extreme evidence for the null hypothesis

Upon receiving ethical approval from the Institutional Review Board of the Faculty of Dentistry/Faculty of Pharmacy, Mahidol University (MU-DT/PY-IRB 2019/DT014.2703), data pooling was undertaken using the modified Preferred Reporting Items for Systematic Reviews (PRISMA) flow diagram [31]. As per the inclusion criteria applied to the published studies of The Angle Orthodontist journal for a period of two years (November 2016 to September 2018), all articles that utilised the t test for statistical analysis were selected. The exclusion criteria filtered out case reports, editorials, and review papers. Studies that employed statistical analyses other than the t tests were also excluded (Fig. 1).

While the two authors (PL and SM) set the criteria jointly, they independently undertook the tasks of identification, screening, and assessment of the articles' eligibility for inclusion in the sample. The identified key variables were the mean, standard deviation (SD), *P* value, sample size, and type of t test. All articles that met the requirements of the inclusion criteria, the key variables were evaluated, and the information was input into the pooled data table. The first author (NS), a statistician, then reviewed all articles in the table to resolve any disagreements related to article identification and screening. Parameters from the pooled data table were then used for further analysis, following the steps described below.

P value verification

Initially, the *P* values were recalculated for each study to verify whether they supported the null hypothesis. Studies with incongruent results were excluded from the analysis.

T-statistic calculation

For the independent-sample t test, an F test was implemented to determine the similarity of variances on a test of $\mathbb{H}_0 : \sigma_1^2 = \sigma_2^2$ versus $\mathbb{H}_1 : \sigma_1^2 \neq \sigma_2^2$ using the formula as described in the supplemental file (Eq. 6 in the supplemental file) [32]:

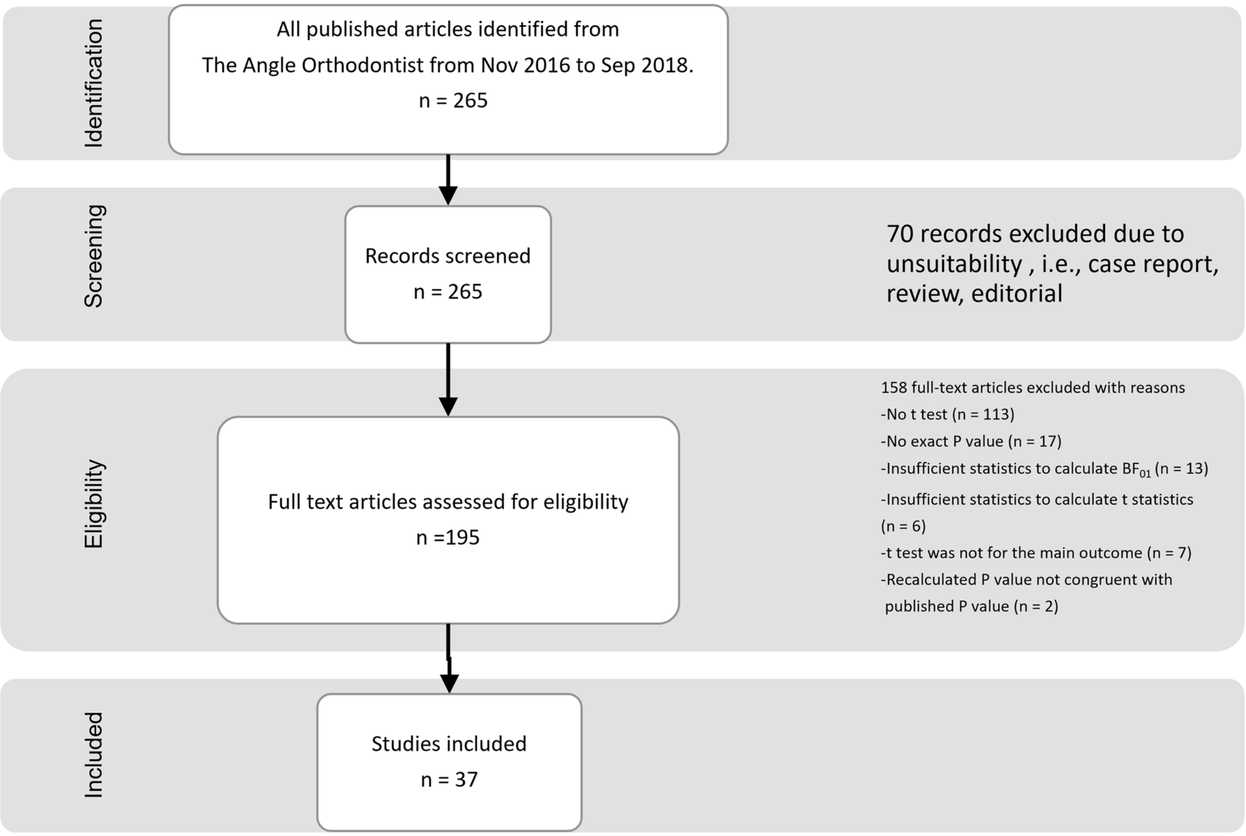
Identification
All published articles identified from
The Angle Orthodontist from Nov 2016 to Sep 2018.
n = 265
Screening
Records screened
n = 265
70 records excluded due to unsuitability , i.e., case report, review, editorial
Eligibility
Full text articles assessed for eligibility
n =195
158 full-text articles excluded with reasons
-No t test (n = 113)
-No exact P value (n = 17)
-Insufficient statistics to calculate BF01 (n = 13)
-Insufficient statistics to calculate t statistics (n = 6)
-t test was not for the main outcome (n = 7)
-Recalculated P value not congruent with published P value (n = 2)
Included
Studies included
n = 37

Fig. 1 Modified PRISMA 2009 flow diagram showing the process of articles selection

Subsequently, the mean, SD, and sample size were input to compute t-statistics using formulae according to the equality (see Eq. 7 in the supplemental file) and inequality (see Eq. 8 in the supplemental file) of the variances.

For the dependent-sample t test, the t-statistic was calculated using the mean, SD, and sample size as inputs into the formula included in the supplemental file (Eqs. 11, 12, and 13 in the supplemental file).

Computation of *P* value for t-distribution

The calculated t from Eqs. 7 to 13 in the supplemental file with their corresponding degree of freedom (df) were then calculated using the formula shown in the supplemental file (Eq. 14 in the supplemental file) [33].

BF_{01} calculation

A retest was then carried out using the t-statistic calculated per the aforementioned method (see Eq. 5 in the supplemental file). The computed t-statistic and sample size were then input into R Studio (Package: BayesFactor) [34] using their formula [35]. This formula (see Eq. 5 in the supplemental file) defines the effect size using a default scale setting at 0.707 ($\delta \sim$ Cauchy) [3]. To interpret BF_{01}, modified Jeffreys' Bayes factor cutoffs [23, 28] were used in this study. Similarly, the *P* value and effect size were interpreted using the cutoffs given by Wasserman [27] and Cohen [28], respectively. The values of the three cutoffs are shown in Table 1.

Calculation of effect size

The effect sizes were computed separately for the dependent- and independent-sample t tests. The following formulae denote the employed mathematical calculations for the dependent- (see Eq. 15 in the supplemental file) and independent-sample t tests [36] (see Eq. 16 in the supplemental file).

Agreement test

The results were then categorised, with the *P* value set to $\alpha = 0.05$ and BF_{01} set to 1. In other words, the results were classified based on whether they rejected the null hypothesis in the condition set. When BF_{01} was greater than 1 and the *P* value was greater than 0.05, it was considered as no evidence against the null hypothesis. However, the null hypothesis was rejected when BF_{01} was less than 1 and the *P* value was less than 0.05. A BF_{01} value of 1 indicated a lack of evidence to either support or reject the null hypothesis. Finally, the agreement percentage and Cohen kappa coefficient were computed.

Performance evaluation of *P* value, BF_{01}, and effect size

Finally, all test and retest results showing the *P* value, BF_{01}, and effect size were plotted on a scattergram and analysed in detail.

Results

Of 265 articles, 195 showed statistical analyses, and 82 used a student's t test (42%). Note that only 37 articles satisfied the criteria for the retest method (Fig. 1). The selected articles contained 793 t tests, including 438 independent-sample t tests and 355 dependent-sample t tests. Therefore, when an orthodontic researcher opted for a t test, they reported approximately 21.4 t tests in one article. The evaluation of agreement was performed with the *P* value set at 0.05 and Bayes factor (BF_{01}) set at 1. The results showed that most retests (742 retests, 93.57%) produced congruent results. The Cohen kappa coefficient was 0.87, indicating perfect agreement between these two tests. For the studies where the Bayes factor suggested anecdotal evidence ($1/3 < BF_{01} < 3$) either in favour of or against the effect, most of these studies showed nonsignificant *P* value (322/372 t tests; Table 2). More important, for approximately two-thirds of the reported significant *P* values between 0.01 and 0.05, BF_{01} was between 1/3 and 1 (50/82 t tests). These comprised both types of t tests. The number of independent-sample t tests was approximately 1.5 times that of the dependent-sample t tests (30/20 t tests). Furthermore, BF_{01} quantified the evidence in favour of the null hypothesis, indicating moderate evidence ($BF_{01} > 3$) for approximately one-third of the total studies with a nonsignificant *P* value (15.51%/56.11%; *P* value > 0.05; Table 2 and Fig. 2).

Scattergrams of three parameters—*P* value, effect size, and BF_{01}—were plotted and analysed (Figs. 2, 3, and 4). The scale of the axes of all scattergrams mostly followed the categorisation described in Table 1. However, some were extended to distinguish the scatter of the plots. Accordingly, the scattergram of each pane shows different gradations.

Approximately half of the studies showed very large effect sizes (more than 0.8). At the *P* value of 0.05, effect sizes for dependent- and independent-sample t tests were approximately 0.5 and 2, respectively, whereas at the *P* value of 0.01, effect sizes for dependent- and independent-sample t tests were approximately 0.5–1 and 2.5–3, respectively. Moreover, at an effect size of 2, the corresponding *P* values ranged from 0.1 to 0.01, particularly for the independent-sample t test. Interestingly, at an effect size of 0.5, *P* values ranged from 0.5 to 0.001 for the dependent-sample t test. These findings emphasised a moderate relationship between the *P* values and effect sizes but not a strong correlation.

As the decision values of BF_{01} spanned $0.3 < BF_{01} < 3$, the interpretation was simpler. The dependent- and independent-sample t test plots of BF_{01} and effect size were still gathered around two separate lines, with a smaller effect size for the dependent-sample t test. The Bayes factor with evidence for the null and alternative hypotheses showed an effect size of a very wide range (0 to 0.8 and 0.5 to 22, respectively).

Considering the three scattergrams, the *P* value and BF_{01} parameters agreed with each other; however, the effect sizes did not. The results began with the extreme side of the decision. The dependent-sample t test revealed a small to medium effect size for the decisive side of no evidence against the null hypothesis as measured by *P* value and decisive evidence for the null hypothesis as measured by BF_{01}. The independent-sample t

Table 2 Agreement test results

Type of t test	Number of t tests that P indicated evidence						Total
	BF_{01} only indicated anecdotal evidence			BF_{01} could indicate evidence			
	Nonsignificant *P* value	Significant *P* value	Total	BF_{01} indicated evidence for H_0 with nonsignificant *P* value	BF_{01} indicated evidence for H_1, with significant *P* value	Total	
Independent-sample t test	232 (29.26%)	30 (3.78%)	262 (33.04%)	65 (8.20%)	111 (14.00%)	176 (22.99%)	438 (55.23%)
Dependent- sample t test	90 (11.35%)	20 (2.52%)	110 (13.87%)	58 (7.31%)	187 (23.58%)	245 (30.90%)	355 (44.77%)
Total	322 (40.61%)	50 (6.30%)	372 (46.91%)	123 (15.51%)	298 (37.58%)	421 (53.09%)	793 (100%)

P value was set to $\alpha = 0.05$, BF_{01} of 3 to 1/3. Numbers in brackets show percentage of t tests of total studies (793). H_0: Null hypothesis; H_1: Alternative hypothesis; BF_{01}: Bayes factor$_{01}$

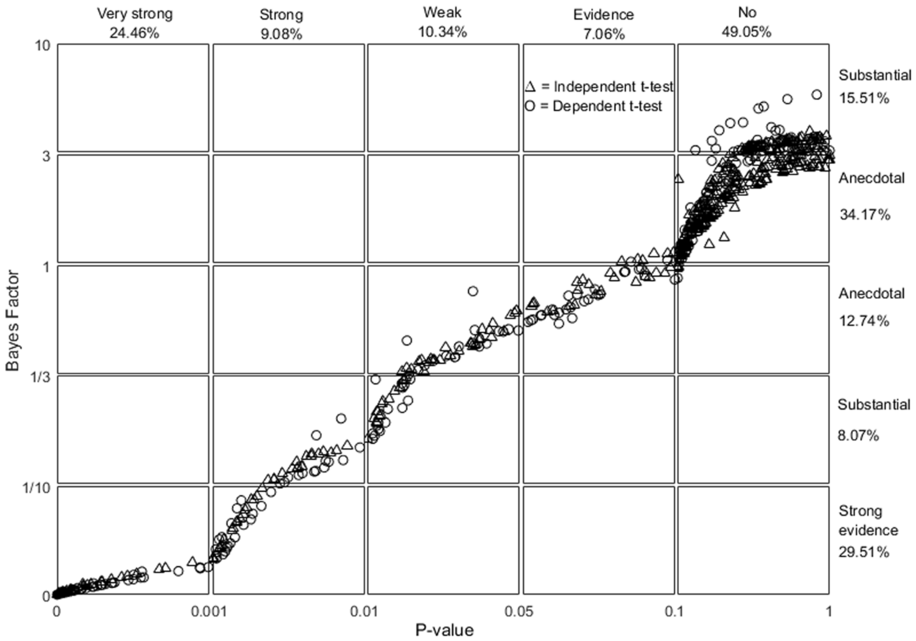

Fig. 2 Scattergram of Bayes factor (BF_{01}) against *P* value. The triangles denote the independent-sample t test, and circles represent the dependent-sample t test. Some panes are extended to distinguish plot scatter. Accordingly, the scattergram of each pane shows different gradations. This scattergram was created using MATLAB software with a Mahidol University licence

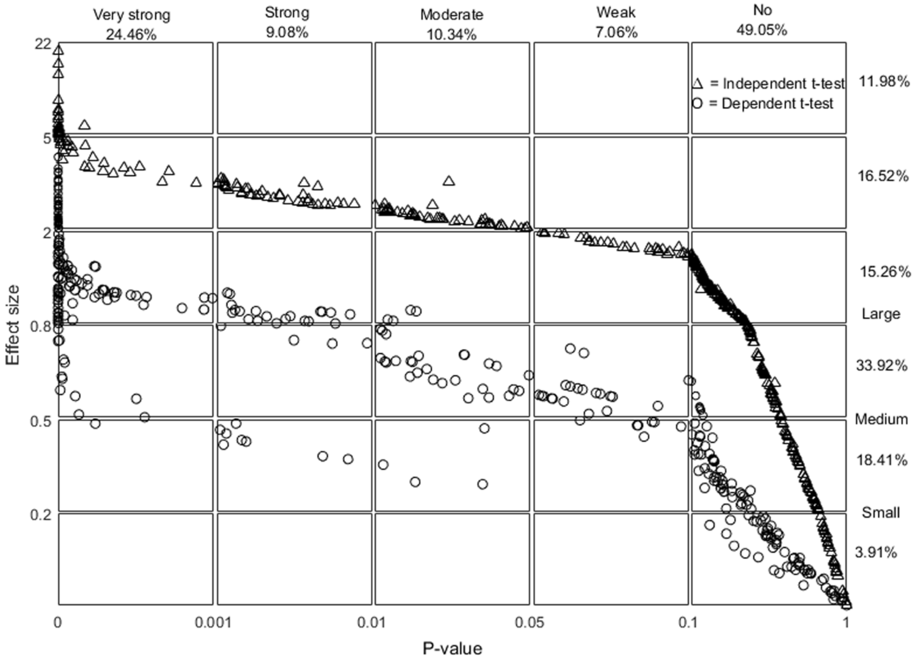

Fig. 3 Scattergram of effect size against *P*-value. Plots of the dependent- and independent-sample t tests are observed distinctively. This scattergram was created using MATLAB software with a Mahidol University licence

test, by contrast, revealed a small-to-large effect size. More important, on the decisive side of strong evidence against the null hypothesis by *P* value and BF_{01}, the dependent-sample t test showed medium to very large effect sizes in most cases. Additionally, the independent-sample t test revealed an extremely large effect size.

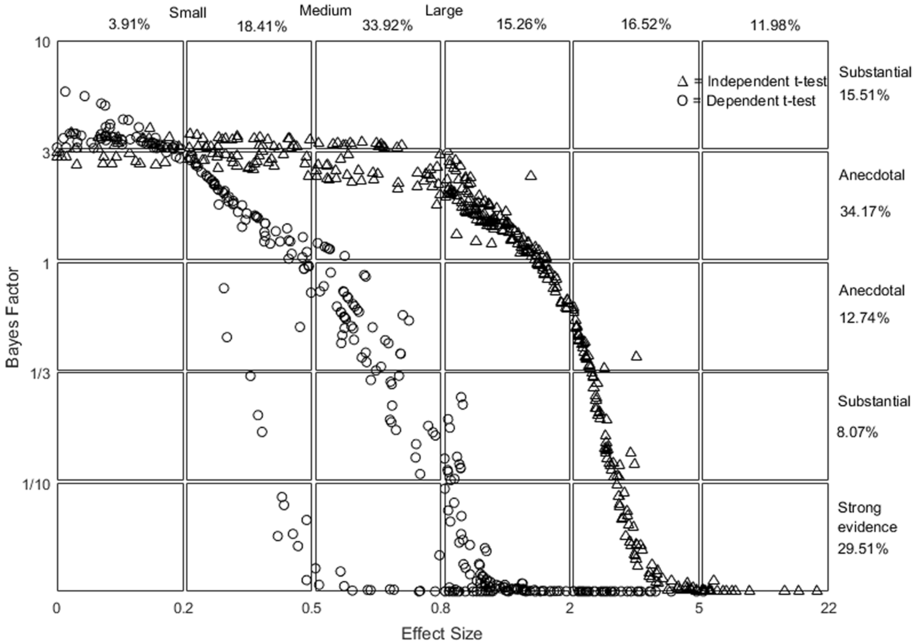

Fig. 4 Scattergram of effect size against BF_{01}. Distinct plots are observed for the dependent- and independent-sample t tests. This scattergram was created using MATLAB software with a Mahidol University licence

Another observation pertained to the *P* values lying on the threshold of a dichotomous decision between the null and alternative hypotheses. While BF_{01} withheld the decision, effect sizes for the dependent-sample t test ranged from low to high. Nevertheless, the independent-sample t test showed very high effect sizes in all cases.

Discussion

There was agreement on the retest results in most of the included studies. It is well accepted that anecdotal evidence obtained from Bayes factor estimation garners only a bare mention. This study used this area of concern ($1/3 < BF_{01} < 3$) to evaluate the strength of the *P* value and examine its agreement with BF_{01}. When anecdotal evidence was considered ($1/3 < BF_{01} < 3$), BF_{01} reserved judgement for almost half of the studies. Under the medium Cauchy prior, two-thirds of the frequentist results ($0.01 < P \text{ value} < 0.05$) would thus be deemed anecdotal evidence ($1/3 < BF_{01} < 1$) for the alternative hypotheses. Interestingly, in this context of anecdotal evidence, the number of independent-sample t tests was approximately 1.5 times that of the dependent-sample t tests (30/20 t tests). Therefore, further investigation involving the effect sizes using scattergrams is warranted.

One of the most popular advantages of Bayesian statistics is its ability to quantify evidence in favour of the null hypothesis. In this study, BF_{01} provided at least moderate evidence (more than 3) for the null hypothesis in 15.51% of the cases with nonsignificant *P* values. This advantage of BF_{01} over *P* value estimation is crucial and may reduce publication bias, as emphasised by multiple studies [22, 25, 37, 38].

The effect size showed moderate to high agreement but did not agree perfectly with the *P* value and BF_{01}. Mostly it facilitated statistical interpretation, as shown in the scatterplot with *P* values and BF_{01}. Furthermore, the effect sizes of the selected articles were remarkably large, considering Cohen's classification [29]. Additionally, the scattergrams showed a distinct separation between the dependent- and independent-sample t test plots. The effect sizes of the dependent-sample t test were mostly smaller than those of the independent-sample t test. Moreover, in the context of the independent-sample t test, when BF_{01} gave the decision for the alternative hypothesis, significant *P* values were reported for two-thirds of the studies. These test results all showed very high effect sizes (greater than 2), complementing the significant *P* values. Orthodontic researchers may be particularly interested in these findings on effect size.

The strength of this study can be derived from the utility of the detailed report on t test parameters published in The Angle Orthodontist journal. First, it should be noted that all t tests from The Angle Orthodontist journal showed standard deviations, and most also displayed exact *P* values. A few studies have indicated additional statistics including confidence intervals, effect sizes, and standard errors. Comprehensive reports on t-test parameters enabled the re-evaluation of the t-test results for the purpose of this study. Second, even though the effect size and BF_{01} calculations are simple, they greatly facilitate statistical interpretation. In addition, to convey the objectivity of the results in orthodontic research, researchers can carry out effect size and BF_{01} estimations. For nonstatisticians, it can be useful to validate statistical interpretation. Furthermore, this one-click away statistical validation is an easy shortcut that requires only a few statistical parameters to make calculations [39–41]. Third, the strength of Bayes' theorem is that it guides the user toward an understanding of statistical thinking. For nonstatisticians, BF_{01} can serve as a stepping stone to many applications of the Bayes factors. Compared to the frequentist procedure, BF_{01} yields better Type I error control at the expense of an increased Type II error [25]. The practical differences between the varied prior distributions used to calculate BF_{01} are well understood [42]. When the prior is cogently set, BF_{01} provides straightforward and rich information for validation and interpretation. For example, the *prior* incorporated into BF_{01} drives orthodontic researchers to understand the properties of their research samples. For instance, analysing functional magnetic resonance imaging (fMRI), Han recently showed the methodological implication of the adjustment of informative prior distribution that is more suitable to medical imaging studies [43]. Specifically, the adjustment of the parameter prior, the effect size, could be considered for both the centre of the prior distribution to a particular value (depending on theoretical considerations, meta-analysis, or prior research). Moreover, its scale of how wide and narrow the distribution can also be adjusted. Ultimately, it is very challenging for orthodontic researchers to start considering the exclusive prior for orthodontic research and gain these benefits from these innovative and trustworthy studies [43, 44].

Statisticians even recommend that Bayes' theorem be included in introductory statistics courses as a substitute for inferential statistics [45, 46]. Dienes stated that researchers who foresee what the theory predicts know how much evidence supports the theory [38]. Hence, orthodontic researchers may understand and benefit much more from Bayesian thinking if they determine which evidence from their research supports their hypothesis and their own proposed idea. Moreover, statistical analyses that use Bayesian estimations specifically to mimic t-test evaluations are now widely propounded [21, 46–48]. Basic online Bayes freeware, with a simple interface to allow easy understanding and proper use, has also become available [39–41, 49]. Such software is designed to offer researchers and clinicians more algorithms and parameters to critique before attempting a statistical judgement. These websites offer advanced algorithms and essential parameters that can be used before judging the difference between two sample means. Specifically, Jeffreys's Amazing Statistics Program (JASP) provides a user-friendly stats module with supportive guidance for Bayes factor computation for t test, analysis of variances (ANOVAs), and correlation coefficient without using Bayes factor in Microsoft Excel sheet [41, 50, 51]. These advantages and the easy availability of statistical packages may encourage orthodontists to validate t-test results using the Bayes factor hypothesis test and inspire Bayesian thinking.

This study also has several limitations. First, before the repeated t-test statistical calculations were attempted, an F test was conducted to determine the similarity of variances. Second, during the screening of the articles, a few were found for which multiple t tests were conducted. Thus, the tests were not exactly independent. Third, although studies with analysis of variances were excluded, multiple comparisons of the means offering similar measurements of the same subject might affect the results of this study. Fourth, the retests in this study were set to be very easy by using just the default BF_{01} with a medium Cauchy prior to effect size. (The effect size followed a Cauchy distribution centred on zero with a scale parameter of 0.707 for the alternative hypothesis.) Furthermore, no sensitivity test was conducted in this study. Considering the aforementioned factors, care should be taken when generalising the results of this study or comparing them with those from similar works.

Because of the lack of raw data, additional Bayesian calculations could not be undertaken. This poses another limitation to the present study. To the best of our efforts, BF_{01} was calculated solely from the reported t-test parameters. For future research on the subject, access to raw data sources will be greatly beneficial, allowing published results to be reevaluated using more logical and innovative applications of Bayesian statistics.

Comparison to similar reports and implications on orthodontic research

Interesting findings lie in the fact that orthodontic studies, while mostly conducted on a small sample size, still show a very high effect. This is important since when the significant *P* value shows the direction, presuming that the treatment effect is present, with this small sample size and high effect side, the significant *P* value from the orthodontic data is more likely to be informative. Second, other fields reported 21% to 31% nonsignificant *P* values [37, 52], and this study showed more nonsignificant *P* values (56.12%). The Bayes factor quantification also revealed that 15.51% showed a BF_{01} of more than 3, indicating moderate evidence for the null hypothesis. Hence, orthodontists might benefit more from Bayes factor quantification than from other fields.

This study emphasises the validation of the t test. Readers can revalidate the summary when encountering the incongruent result of the two parameters, *P* value and Bayes factor, or even with a nonconverging effect size on significant P. It is also crucial to retest the significant P of the innovative orthodontic material or treatment intervention, especially when the *P* value is slightly below 0.05. The supplementary calculation of the effect side and Bayes factor facilitates the analysis of the correctness of the conclusion [50, 51]. From our published research [53], the difference between the ratio of teeth (Cumulative Percentage Ratio7) between Australian and Thai was significantly different ($P<0.05$). Therefore, it was reasonable to recommend a new ratio to specifically suit Thai patients [one sample t test revealed that $t(73)=-2.274$; $P=0.029$; $BF_{01}=0.7$; effect side $=0.378$]. At present, scrutinised BF_{01} shows anecdotal evidence by the graphical presentation by JASP summary statistics; it is probably best to restate this conclusion. Furthermore, since this is the only study of its kind and there are no other similar studies, this test should be reconfirmed. Moreover, this simple extension of statistical validation is largely rational, even outside orthodontic research, to dentistry and medical fields.

It should also be mentioned that most orthodontic reports show treatment effects utilising cephalometric analysis. Therefore, the fallacy of only one reported parameter (*P* value) can become less crucial. From the study of soft tissue change by treatment intervention, one parameter was misguided by the *P*-value fallacy [54]. Among all measurements, one of them—the distance from the vertical reference plane (VRP) to Sella-Nasion (Sn) in millimetres—was analysed using a t test, and the misdirected significant P was reported. However, this can be easily verified using freeware equipped with helpful instruction [50, 51], [$t(27,28)=-2.102$; $P=0.040$; $BF_{01}=0.61$; effect size $=2.107$]. Finally, the verification presents one result with an incongruent significant *P* value (although with a large effect size) versus the anecdotal BF_{01}. However, since the authors concluded knowledgeably from 25 cephalometric parameters, the final conclusion was not affected by one misdirected *P* value. In short, regarding the orthodontic field armed with cephalometric analysis, it is quite challenging for the *P*-value fallacy to affect the main orthodontic treatment conclusion.

Additionally, familiarity with the Bayes framework has numerous advantages for orthodontic, dental, and medical researchers. First, the Bayes framework is a prominent alternative to following the recommendation of the American Statistical Association (ASA) to go beyond the *P* value [19], especially for orthodontic reports of new treatment interventions or materials. Second, for editors, reviewers, and readers, increasing the sample size to hack the *P* value (data dredging) can be easily verified without knowing the raw data [2, 48, 51]. Finally, it is a self-preparation method for modern Bayes statistical software.

Conclusions

When comparisons of sample means were retested, the studies published in The Angle Orthodontist journal showed mostly congruent results (Kappa coefficient $=0.87$) between the two statistical parameters *P* value and Bayes factor (using the *P* value set at 0.05 and Bayes factor [BF_{01}] set at 1). However, when anecdotal evidence ($1/3<BF_{01}<3$) was considered, the Bayes factor reserved judgement for almost half of the studies. However, it quantified moderate evidence favouring the null hypothesis (15.51% nonsignificant *P* values). This advantage of the Bayes factor over the *P* value is crucial for reducing publication bias. In conclusion, statistical judgement should be made with caution. As most clinical orthodontic outcomes were evaluated using cephalometric analysis, drawing conclusions from many cephalometric parameters prevents the P-fallacy effect. Despite this fact, it is recommended that in addition to *P* value computations, other statistical estimations such as effect size or BF_{01} be used to validate judgement and facilitate statistical interpretation. If these parameters show nonconverging results, then the available user-friendly statistical software facilitates this verification. This test is crucial for orthodontic innovative material or treatment modalities.

Abbreviations

α error	Type I error
ANOVAs	Analysis of variances
ASA	American Statistical Association
β error	Type II error
BF	Bayes factor
df	Degree of freedom
fMRI	Functional magnetic resonance imaging
F test	A statistical test based on the F distribution
H_0	Null hypothesis
H_1	Alternative hypothesis
JASP	Jeffreys's Amazing Statistics Program
P	Probability
PRISMA	Preferred Reporting Items for Systematic Reviews
SD	Standard deviation
t	Hypothesis test
VRP to Sn	Vertical reference plane to Sella-Nasion

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01858-z.

Additional file 1.

Acknowledgements

The Angle Orthodontists Journal website, administered by the Edward H. Angle Society of Orthodontists (EHASO), kindly provided the data from its open-access journal website. The constructive criticism of Professor CTC Ho from Queensland Children's Hospital, Australia, and the two reviewers are also truly appreciated. In addition, we are grateful for the English language editing of this manuscript by American Journal Expert (aje.com), Editage.com and Dr Siew Peng Neoh, Orthodontic Lecturer, Mahidol University.

Authors' contributions

All authors have contributed to the writing of the article. PL: retrieve collected and analyzed data from the Angle Orthodontists Journal. SM: corresponding author of the article, collected and analyzed data, conceptualized the methodology. ST: conceptualized the methodology. NS: validated the retrieved data and the statistical analyses. All authors read and approved the final manuscript.

Funding

This research was supported by Mahidol University.

Availability of data and materials

The data supporting this study's findings are available on The Angle Orthodontists Journal website. (https://meridian.allenpress.com/angle-orthodontist). The dataset analysis of the current study to validate all results is available in the Open Science Framework (OSF) repository, https://osf.io/7rxhg/.

Declarations

Ethics approval and consent to participate

Prior to the commencement of this study, approval was obtained from the Ethics of the Faculty of Dentistry/Pharmacy, Mahidol University Institutional Review Board (COE.No. MU-DT/PY-IRB 2019/DT014.2703). All data were retrieved from The Angle Orthodontists Journal, an Open Access Journal.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 January 2022 Accepted: 2 February 2023
Published online: 16 February 2023

References

1. Oaks M. Statistical Inference: A Commentary for the Social and Behavioral Sciences. New York: Wiley; 1986.
2. Burger JB, Wolpert RL. The likelihood principle. Hayward CA: Institute of Mathematical Statistics. 1988. https://jstor.org/stable/4355509. Accessed 10 June 2022.
3. Altman D. Practical Statistics for Medical Research. London: Chapman and Hall CRC; 1991.
4. Fisher R. Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd; 1925.
5. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc London Ser A, Contain Pap a Math or Phys Character. 1933;231:289–337.
6. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:1005–13. https://doi.org/10.7326/0003-4819-130-12-199906150-00019.
7. Goodman S. A dirty dozen: Twelve P-value misconceptions. Semin Hematol. 2008;45:135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.
8. Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. Robust misinterpretation of confidence intervals. Psychon Bull Rev. 2014;21:1157–64. https://doi.org/10.3758/s13423-013-0572-3.
9. Haller H, Krauss S. Misinterpretations of significance: A problem students share with their teachers? Methods Psychol Res Online. 2002;7:1–20.
10. Lecoutre MP, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. Int J Psychol. 2003;38:37–45. https://doi.org/10.1080/00207590244000250.
11. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the Biostatistics and results in the medical literature. JAMA. 2007;298:1010–22. https://doi.org/10.1001/jama.298.9.1010.
12. Scheutz F, Anderson B, Wulff HR. What do dentists know about statistics? Eur J Oral Sci. 1988;96:281–7. https://doi.org/10.1111/j.1600-0722.1988.tb01557.x.
13. Pandis N. The P value problem. Am J Orthod Dentofac Orthop. 2013;143:150–1. https://doi.org/10.1016/j.ajodo.2012.10.005.
14. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: What do p values and confidence intervals really represent? Anesth Analg. 2018;126:1068–72. https://doi.org/10.1213/ANE.0000000000002798.
15. Leung W-C. Balancing statistical and clinical significance in evaluating treatment effects. Postgr Med J. 2001;77:201–4. https://doi.org/10.1136/pmj.77.905.201.
16. Trafimow D. Editorial. Basic Appl Soc Psych. 2014;36:1–2. https://doi.org/10.1080/01973533.2014.865505.
17. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych. 2015;37:1–2. https://doi.org/10.1080/01973533.2015.1012991.
18. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05." Am Stat. 2019;73:1–19. https://doi.org/10.1080/00031305.2019.1583913.
19. Wasserstein RL, Lazar NA. The ASA's statement on p -values: Context, process, and purpose. Am Stat. 2016;70:129–33. https://doi.org/10.1080/00031305.2016.1154108.
20. Wagenmakers EJ, Marsman M, Jamil T, et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychon Bull Rev. 2018;25:35–57. https://doi.org/10.3758/s13423-017-1343-3.
21. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev. 2009;16:225–37. https://doi.org/10.3758/PBR.16.2.225.
22. Kruschke JK. Bayesian estimation supersedes the t test. J Exp Psychol Gen. 2013;142:573–603. https://doi.org/10.1037/a0029146.
23. Bayes T, Price M. An essay towards solving a problem in the doctrine of chances. Philos Trans. 1763;1683–1775:370–418.
24. Jeffreys H. Theory of Probability. 3rd ed. New York: The Clarendon Press, Oxford University Press; 1983.
25. Kelter R. Bayesian and frequentist testing for difference between two groups with parametric and nonparametric two-sample tests. Wiley Interdiscip Rev Comput Stat. 2021;13:e1523. https://doi.org/10.1002/wics.1523.
26. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90:773–95. https://doi.org/10.1080/01621459.1995.10476572.
27. Wasserman L. All of statistics: A concise course in statistical inference. New York: Springer; 2004.
28. Lee MD, Wagenmakers EJ. Bayesian cognitive modeling: A practical course. Amsterdam, Netherlands: Cambridge University Press; 2013.
29. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum; 1988.
30. Law S, Chudasama D, Rinchuse D. Evidence-based orthodontics. Angle Orthod. 2010;80:952–6. https://doi.org/10.2319/012110-44.1.
31. Moher D, Liberati A, Tetzlaff J, Altman D, Group TP. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6:1–6. https://doi.org/10.1371/journal.pmed.1000097.
32. Rosner B. Fundamentals of Biostatistics. 8th ed. Boston, MA: Cengage Learning; 2015.
33. Krishnamoorthy K. Handbook of statistical distributions with applications. Boca Raton, FL: Chapman and Hall; 2006.
34. Morey RD, Rouder JN. BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12–4.2. https://cran.r-project.org/package=BayesFactor. Accessed 10 June 2022.
35. Morey RD. Using the "BayesFactor" package, version 0.9.2+. 2015. https://richarddmorey.github.io/BayesFactor/. Accessed 28 Mar 2020.
36. Mussweiler T. Doing is for thinking! Psychol Sci. 2006;17:17–21. https://doi.org/10.1111/j.14679280.2005.01659.x.
37. Wetzel R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. Perspect Psychol Sci. 2011;6:291–8. https://doi.org/10.1177/1745691611406923.
38. Dienes Z. Using Bayes to get the most out of non-significant results. Front Psychol. 2014;5:781–97. https://doi.org/10.3389/fpsyg.2014.00781.

39. Bayes factor for grouped or two-sample t-tests | Perception and cognition Lab. http://pcl.missouri.edu/bf-two-sample. Accessed 29 Apr 2021.
40. Bayesian estimation supersedes the t-test (BEST) - Online. http://sumsar.net/best_online/. Accessed 29 Apr 2021.
41. JASP - A fresh way to do statistics. https://jasp-stats.org/. Accessed 29 Apr 2021.
42. Rawenzwaaiij Dv, Etz A. Simulation studies as a tool to understand bayes factors. AMPPS. 2021;4:1–31.
43. Han H. A method to adjust a prior distribution in Bayesian second-level fMRI analysis. PeerJ. 2021;9:e10861.
44. Zondervan-Zwijnenburg M, Peeters M, Depaoli S, Van de Schoot R. Where do priors come from? Applying guidelines to construct informative priors in small sample research. Res Hum Dev. 2017;14:305–20.
45. Carlin BP, Louis TA. Bayes and Empirical Bayes Methods for Data Analysis Bayesian Theory. Vol 85. 2nd ed. New York: Chapman and Hall CRC; 2000.
46. Wang M, Liu G. A simple two-sample Bayesian t-test for hypothesis testing. Am Stat. 2016;70:195–201. https://doi.org/10.1080/00031305.2015.1093027.
47. Gönen M, Johnson WO, Lu Y, Westfall PH. The Bayesian two-sample t test. Source Am Stat. 2005;59:252–7. https://doi.org/10.1198/000313005X55233.
48. Kruschke JK, Liddell TM. Bayesian data analysis for newcomers. Psychon Bull Rev. 2018;25:155–7. https://doi.org/10.3758/s13423-017-1272-1.
49. Kelter R. Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Med Res Methodol. 2020;22:88. https://doi.org/10.1186/s12874-020-00968-2.
50. JASP http://jasp-stats.org/2018/04/11/teaching-bayesian-estimation-with-the-summary-stats-module/. Accessed 9 Sept 2022.
51. Ly A, Raj A, Etz A. Bayesian reanalyses from summary statistics: A guide for academic consumers. AMPPS. 2018;1:367–74. https://doi.org/10.1177/2515245918779348.
52. Hoekstra R, Monden R, Ravenzwaaij D, Wagenmakers E. Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. PLoS ONE. 2018;13:1–9.
53. Manopatanakul S, Watanawirun N. Comprehensive intermaxillary tooth width proportion of Bangkok residents. Braz Oral Res. 2011;25:21–7. https://doi.org/10.1590/s1806-83242011000200005.
54. Kim K, Choi S, Choi E, Choi Y, Hwang C, Cha J. Unpredictability of soft tissue changes after camouflage treatment of Class II division 1 malocclusion with maximum anterior retraction using miniscrews. Angle Orthodontist. 2017;87:230–8. https://doi.org/10.2319/042516-332.1.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.