pISSN, eISSN 0125-5614
M Dent J 2020; 40 (2) : 107-112

Agreement test of statistical analysis results between online bayes estimation versus t-test in Mahidol Dental Journal.

Autchariya Kho-Srisut[1], Natchalee Srimaneekarn[2], Sittichoke Osiri[3], Sasipa Thiradilok[4], Pornpoj Fuangtharnthip[4], Binit Shrestha[5], **Somchai Manopatanakul**[4]

[1] *Sawaeng-Ha hospital, Angthong*
[2] *Department of Anatomy, Faculty of Dentistry, Mahidol University*
[3] *Department of Operative Dentistry and Endodontics, Faculty of Dentistry, Mahidol University*
[4] *Department of Advanced General Dentistry, Faculty of Dentistry, Mahidol University*
[5] *Department of Prosthodontics, Faculty of Dentistry, Mahidol University*

Objective: This research aimed to assess agreement between t-test and Bayesian estimation using the results from Mahidol Dental Journal. In general, to reveal the difference of means between two sample groups, Inferential statistics using Null Hypothesis Significant Testing (NHST), particularly t-test, has long been accepted. However, statistical analysis revolutionized by the Journal of the Basic and Applied Social Psychology (BASP), almost dismissed papers with NHST. Later, American and Thai Statistical Association also published articles that explained limitation of P-value. Alternatively, Bayesian estimation which has been developed for more than 200 years, has been recommended as a substitution for t-test.

Materials and methods: Upon completion of ethical approval, data were pooled from the articles using t-test in Mahidol Dental Journal from 2007-2017. Then the mean, standard deviation and sample size published in these articles were used to calculate the t-value. Online Bayesian estimation program (http://pcl.missouri.edu/bayesfactor) was applied utilizing the aforementioned calculated t-values. Agreement percentage and Cohen Kappa Coefficient were also computed.

Results: From the overall 274 articles, 21 articles adopted independent sample t-test and 2 articles adopted one sample t-test statistical analyses. Eighty-seven percent of the articles published in Mahidol Dental Journal showed agreement of research results between t-test and Bayesian estimation. The Cohen Kappa Coefficient was 0.73 indicating substantial agreement between these two tests. Further, the tendency of disagreements occurred with P-values starting from 0.05 to 0.085.

Conclusion: Mahidol Dental Journal showed substantial agreement for both statistical analyses. Future study will suggest the detail investigation on how Bayes theorem clarifies the disagreement between these two statistical test results and the situation when Bayes may perform better.

Keywords: agreement, Bayes, t-test

How to cite: Kho-Srisut A, Srimaneekarn N, Osiri S, Thiradilok S, Fuangtharnthip P, Shrestha B, Manopatanakul S. Agreement test of statistical analysis results between online bayes estimation versus t-test in Mahidol Dental Journal. M Dent J 2020; 40: 107-112.

Introduction

Frequentist or inferential statistics using Null Hypothesis Significant Testing (NHST) have long been accepted as well as regularly used to observe the difference between two means, especially the t-test signifying P-value The P-value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the Null hypothesis is correct. Its interpretation depends on the value set, sample size and the explanation of its definition set by Fisher [1]. Dichotomous decision based on just P-value with vague understanding

Correspondence author: Somchai Manopatanakul
Department of Advanced General Dentistry, Faculty of Dentistry, Mahidol University.
6 Yothi Road, Ratchathewi District, Bangkok 10400, Thailand
Email: msomchai@rocketmail.com, Tel: +662-200-7853. Fax: +662-200-7852

Received : 24 October 2020 Accepted : 20 May 2020

can easily lead to the misinterpretation of the result [2, 3]. This P-value misinterpretation among dentists was also evidently documented [4]. Further, misconception and misperception of P-value were not uncommon even for statisticians [5]. While Goodman detailed twelve common misconceptions of P-value, the author specified and forewarned the non "evidence-based" statistical inference and a widespread misperception in drawing conclusion [6]. Moreover, this P-value interpretation dictates whether treatment, operation, or medication will be prescribed or not. In 2014, a statistical evolution was prompted by the Journal of the Basic and Applied Social Psychology (BASP) [7, 8]. In order to be responsible to the readers of all published articles, BASP journal launched a stringent rule regarding the acceptance of papers that used NHST. Later, Thai and American Statistical Associations also published articles explaining the limitations of P-value [2,9].

Besides inferential statistics, Bayesian estimation was first developed by Thomas Bayes and popularized for more than 200 years [10]. Bayes' theorem is defined mathematically as the following equation:

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

where: $P(A)$ = The probability of A occurring
$P(B)$ = The probability of B occurring
$P(A|B)$ = The probability of A given B
$P(B|A)$ = probability of B given A

In short, Bayes' theorem gives the probability of an event based on new information that is, or may be related, to that event. Schoot and colleagues also clearly described and summarized the advantage of Bayes over P-value in the table 1 [11]. Briefly, they synopsized four main reasons supporting Bayesian statistics. First, conventional methods may lack the potential to estimate some complex models. Second, the definition of Bayes' probability is more preferable than one number (P-value). Third, Bayes not only conducts statistical analysis, but also takes background knowledge (a priori) into account. Last, large sample size is not necessarily required [12]. Bayes theorem was then ultimately recommended as an Alternative for t-test [13].

Instead of P-value, Bayesian estimation applies conditioning probability to explain a chance that Null hypothesis might occur. To ease the understanding of Bayes theorem, Bayes factor (K), the most fundamental calculation was proposed to be very similar to NHST and P-value [14]. Kass and Raftery further documented and advocated Bayes factor in 1995 [15]. Bayes factor is defined as the likelihood ratio comparing two competing hypotheses. In this t-test context, they

Table 1 Overview of the similarities and differences between frequentist and Bayesian statistics [11]

	Frequentist statistics	Bayesian statistics
Definition of the P- value	The probability of observing the same or more extreme data assuming that the Null hypothesis is true in the population	The probability of the (Null) hypothesis
Large samples needed?	Usually, when normal theory-based methods are used	Not necessarily
Inclusion of prior knowledge possible?	No	Yes
Nature of the parameters in the model	Unknown but fixed	Unknown and therefore random
Population parameter	One true value	A distribution of values reflecting uncertainty
Uncertainty is defined by	The sampling distribution based on the idea of infinite repeated sampling	Probability distribution for the population parameter
Estimated intervals	Confidence interval: Over an infinity of samples taken from the population, 95% of these contain the true population value	Credibility interval: A 95% probability that the population value is within the limits of the interval

are Null and Alternative hypotheses and Bayes factor supports one hypothesis or another. When P-value and its Bayesian counterpart (Bayes' factor) were compared, Bayes' explanation of its result was also preferred by Goodman for more straightforward interpretability [6]. Bayes Factor cut-offs were also proposed by Jeffreys to ease the interpretation (table 2) [14].

Since calculation of Bayes' factor is very simple. There is no need of any special computerized algorithm. Moreover, with modern computer technology, commercial software, and even freeware statistical analysis program, it has become more accessible and user friendly [16,17]. Although Bayesian estimation was introduced to Thai dental research by several pioneering groups [18,19], there are only a few Thai dental researches using Bayesian estimation. Therefore, it may be important to direct the Thai researchers towards this Bayesian estimation. Moreover, a comparison of the results of Thai dental researches using t-test and Bayesian estimation has not been made so far and would serve to further validate its efficacy. Further, to our knowledge, no journal to date pioneers to verify their published data. Mahidol Dental Journal would serve as the first to validate and declare itself. Therefore, this research aimed to compare the research results from Mahidol Dental Journal using t-test versus Bayesian estimation.

Material and methods

At the outset, ethical approval was obtained from Institutional Review Board of the Faculty of Dentistry / Faculty of Pharmacy, Mahidol University (MU-DT/PY-IRB-2017/DT122). Studies published from 2007 to 2017 in Mahidol Dental Journal were manually searched separately by two calibrated reviewers (NS and SM). This search included articles with independent or one sample t-test statistical analyses. Exclusion criteria were articles with dependent sample t-test. This is due to the fact that if only published data are used, dependent sample t-test lacks raw data to calculate the t-value.

All studies that met the inclusion criteria were collected. While the first author (AC) and corresponding author (SM) collected related information from the included articles, the second author (NS), a statistician, reviewed all the collected information. These variables contained mean and standard deviation (SD) that were mentioned in the articles. In addition, sample size and type of t-test were also identified. These data were inserted to the data pooling table. It should be noted here that in this preliminary study, only relevant published information of the most distinctive one sample or independent sample t-test mentioned in the abstract from each study was included and there was no attempt to contact the corresponding authors. Further, to resolve any argument related

Table 2 Jeffrey's Bayes Factor cut-offs demonstrating its interpretation related to Alternative hypothesis or Null hypothesis [14].

Alternative

Bayes factor$_{10}$	Description
≤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 – 1/3	Moderate 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

to the articles searching and analysis, these arguments were re-analyzed and resolved by the statistician (NS). Mean, SD and sample size of each group from the data pooling table were then used to calculate the t-value using the following methodology:

For 1-sample t-test, statistic was calculated using the reported mean and SD. Similarly, for 2-sample t-test, statistic was obtained using a formula for unequal variance, the reported means and SDs. These t statistics and sample size were then inputted to the online Bayesian estimation program. The Bayes factor was computed with version 0.9.8 of the BayesFactor package [R version 3.3.2 (2016-10-31)]. The website (http://pcl.missouri.edu/bayesfactor) *per se* describes in detail about this Bayesian factor calculation. These calculations were performed separately between independent sample and one sample t-test. The results designated the conditioning probability whether it favored Alternative or Null hypothesis. Accordingly, the agreement percentage and Cohen Kappa coefficient between research results of t-test and Bayesian estimation program were reported.

Bayes estimation explains the probability of an event, based on prior knowledge of conditions that may be associated with its occurrence. Bayes estimation mimicking t-test also offers more comprehensive calculation including the effect size of this prior. To define this effect size, the aforementioned online program uses the default setting of the scale at 0.707 $\{(\sqrt{2}) / 2\}$. Rationally, this scale should be defined by the researcher to suit the logic of their samples. However, this study simply provided the Bayes estimation only at this default setting [20].

Results

From the overall 274 articles, 21 articles adopted independent sample t-test and 2 articles adopted one sample t-test statistical analyses. Eighty-seven percent of articles that were published in Mahidol Dental Journal showed the result agreement for both t-test and online Bayesian estimation. The Cohen Kappa Coefficient was 0.73 denoting substantial agreement between these two tests. The detailed information of all articles evaluated was presented in Table 3 and 4.

Discussion

From the statistical analysis using both classical Inferential statistical analysis and Bayesian estimation recommended by modern statisticians, the Cohen Kappa Coefficient designated substantial agreement between these two tests when evaluating articles from Mahidol Dental Journal. Further, the disagreement showed the tendency to occur starting from P-values of 0.05 to 0.085. This was supported by Berger and Mortera who advocated the weakness of P-value of 0.05 of not being able to represent much evidence against the Null [21]. Therefore, there might be a tendency that Bayes estimation may perform better at this P-value of 0.05.

As only published P-value and the most distinct t-test of each study were compared and only the decision indicated from the online program to accept or reject Null hypothesis was showed, Bayes estimations in this study were only in a very early stage of approximation. Although the online program gives the Bayes factor as number which indicates level to support Null or not, most calculations are in default setting. Bayes' theorem also involves a priori. This a priori involvement drives researchers of each discipline to understand more on their research methodology to righteously design the research methodology. For example, Bayes factor may requires specific cutoff point for different research depending on the strength of association of a prior and posterior. Opposed to Jeffreys, in Forensic science, specifically in a criminal trial, it requires posterior odds for H_1 (guilt) against H_0 (innocence) of at least 1,000 rather than 100 to be considered as extreme evidence [22] (table 3). Further, As Dienes advocated, researchers who know what the theory predicts, know how much evidence supports a theory [23]. As such, only dentist scrupulously understand dental researches and dental researchers who understand what evidence from their research support their hypothesis testing and their own proposed idea may understand and benefit a lot more from Bayes theorem. Because all parameters should be completely collected and defined by the researcher to suit the logic of the samples, it is mandatory to collect more detail raw data, avoid the default setting, construct the Bayes factor specialized calculation formula and investigate further the relationship of P-value and Bayes factor in the next study. This further investigation may shed more light to how Bayes will methodically offer advantage over P-value in dental research. Moreover, with this solid fundament in mind, this online Bayesian estimation program may serve as an initial quick optional tool to validate the publicized t-test result.

In addition, statisticians are also keen to provide more of the Bayesian estimation for all inferential statistics including analysis of variances (ANOVA), analysis of co-variances (ANCOVA), correlation, and regression [24, 25]. Without hesitation, accustomed inferential statistics based on NHST are still contentedly utilized by most dental journals. Therefore, it may also be absorbing to conduct more thorough researches comparing all types of statistical analyses from many more journal publications.

Table 3 Data pooling table showing articles with one sample t-test

Article number	Sample size	Mean and SD	μ under H_0	t-value	P-value	H_0	Bayes	Agreement of results
1	67	8.70±0.51	8.65	0.802	0.05	accepted	accepted	Yes
2	56	9.47±1.41	19.20	51.640	0.01	accepted	accepted	Yes

Table 4 Data pooling table depicting articles with independent samples t-test. Grey box depicted the articles with disagreement between t-test and Bayesian estimation.

Article	Group 1		Group 2		t-test			Bayes	Agreement
number	Sample size	Mean and SD	Sample size	Mean and SD	t-value	P	Ho		of results
1	40	41.39±2.17	40	38.62±2.34	5.479	<0.001	rejected	rejected	Yes
2	57	45.22±2.46	66	43.39±2.25	4.280	<0.001	rejected	rejected	Yes
3	16	320.77±50.75	16	291.05±45.81	3.622	0.001	rejected	rejected	Yes
4	200	8.07±0.51	200	7.63±0.62	5.481	<0.01	rejected	rejected	Yes
5	51	57.62±4.65	12	50.82±5.56	3.926	0.01	rejected	rejected	Yes
6	10	0.75±0.06	10	0.69±0.07	2.058	0.03	rejected	rejected	Yes
7	8	43.39±2.94	14	40.77±2.45	2.133	0.04	rejected	rejected	Yes
8	147	157.3±6.2	232	155.9±5.8	2.229	0.05	rejected	rejected	Yes
9	15	2.67±0.62	15	2.67±0.49	0	0.05	accepted	accepted	Yes
10	20	25.64±2.03	20	24.99±2.43	0.918	0.05	accepted	accepted	Yes
11	6	37.43±1.14	6	36.19±2.02	1.310	0.05	accepted	accepted	Yes
12	30	34.31±1.99	30	30.68±1.77	7.465	0.05	rejected	rejected	Yes
13	16	8.85±0.48	51	8.66±0.52	2.652	0.05	accepted	accepted	Yes
14	20	17.23±6.03	20	15.8±4.46	0.853	0.05	accepted	accepted	Yes
15	10	0.6±0.43	10	0.3±0.29	1.829	0.05	rejected	rejected	Yes
16	15	63.8±43.9	15	13.7±5.3	4.388	0.05	rejected	rejected	Yes
17	30	16.48±3.52	34	14.48±2.72	2.519	0.05	rejected	rejected	Yes
18	163	7.99±0.52	154	7.57±0.58	6.775	0.05	rejected	rejected	Yes
19	23	17.59±0.95	21	17.00±0.94	2.069	0.05	accepted	rejected	No
20	138	1.74±0.30	90	0.81±0.40	18.866	0.07	accepted	rejected	No
21	147	28.05±2.83	2	23.00±4.10	1.736	0.085	accepted	rejected	No

Conclusion

In this preliminary study, two statistical analysis methods act as the ruler judging the difference of the arithmetic means between two groups. The outcomes were that articles from Mahidol Dental Journal showed substantial agreement of both statistical analysis methods. These disagreements showed the tendency to occur when P-values were from 0.05 to 0.085. Further study aimed to collect in-depth raw data, formulate specific calculation formula and evaluate the advantage of Bayes factor over P-value. To summarize, this study shows the alternative methods comparing the means of two groups. It also promotes logical thinking of statistical analysis.

References

1. Altman D. *Practical statistics for medical research.* Chapman and Hall CRC: London, 1991.
2. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. ***Am Stat*** 2016; 70: 129–133.
3. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". ***Am Stat*** 2019; 73: 1–19.
4. Scheutz F, Anderson B, Wulff HR. What Do Dentists Know About Statistics? ***Eur J Oral Sci*** 1988; 96: 281–287.
5. 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.
6. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. ***Semin Hematol*** 2008; 45: 135–140.
7. Trafimow D. Editorial. ***Basic Appl Soc Psych*** 2014; 36: 1–2.
8. Trafimow D, Marks M. Editorial. ***Basic Appl Soc Psych*** 2015; 37: 1–2.
9. Nguyen H. Why P-values are Banned? ***Thail Stat*** 2016; 14: 448.
10. O'Hagan A. Bayesian statistics: principles and benefits. In: van Boekel M, Stein A, van Bruggen A (eds). *Bayesian Statistics and Quality Modelling in the Agro-Food Production Chain*. Springer: Wageningen, The Netherlands, 2004, pp 31–45.
11. Van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG. A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. ***Child Dev*** 2014; 85: 842–860.
12. van de Schoot R, Depaoli S. Bayesian Analyses: Where to Start and What to Report. ***Eur Helath Psychol*** 2014; 16: 75–84.
13. 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: 225–237.
14. Jeffreys H. *Theory of probability*. Third edition. The Clarendon Press, Oxford University Press: New York, 1983.
15. Kass RE, Raftery AE. Bayes Factors. ***J Am Stat Assoc*** 1995; 90: 773–795.
16. Kruschke JK. Bayesian Estimation Supersedes the T Test. ***J Exp Psychol Gen*** 2013; 142: 573–603.
17. Kruschke JK, Liddell TM. Bayesian Data Analysis for Newcomers. ***Psychon Bull Rev*** 2018; 25: 155–157.
18. Suebnukarn S, Rungcharoenporn N, Sangsuratham S. A Bayesian Decision Support Model for Assessment of Endodontic Treatment Outcome. ***Oral Surg Oral Med Oral Pathol Oral Radiol Endod*** 2008; 106: e48–e58.
19. Thanathornwong B, Suebnukarn S, Songpaisan Y, Ouivirach K. A System for Predicting and Preventing Work-Related Musculoskeletal Disorders among Dentists. ***Comput Methods Biomech Biomed Engin*** 2014; 17: 177–185.
20. Morey RD. Using the 'BayesFactor' Package, version 0.9.2+. 2015. Available from https://richarddmorey.github.io/BayesFactor/ (accessed 28 Mar2020).
21. Berger JO, Mortera J. Interpreting the Stars in Precise Hypothesis Testing. ***Int Stat Rev*** 1991; 59: 337–353.
22. Evett IW. Bayesian Inference and Forensic Science: Problems and Perspectives. ***Stat*** 1987; 36: 99–105.
23. Dienes Z. Using Bayes to Get the Most out of Non-Significant Results. ***Front Psychol*** 2014; 5: 1–17.
24. Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J *et al*. Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications. ***Psychon Bull Rev*** 2018; 25: 35–57.
25. Wagenmakers EJ, Love J, Marsman M, Jamil T, Ly A, Verhagen J *et al*. Bayesian Inference for Psychology. Part II: Example Applications with JASP. ***Psychon Bull Rev*** 2018; 25: 58–76.