International Journal of Statistics and Probability; Vol. 12, No. 2; March 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Importance of Type II Error in Hypothesis Testing

Inmaculada Jiménez-Gamero[1] & Mohamed Analla[2]

[1.] Department of Biology, I.E.S. Siete Colinas, 51002 Ceuta, Spain

[2.] Department of Biology, Abdelmalek Essaadi University, 93030 Tetouan, Morocco

Correspondence: Mohamed Analla, Department of Biology, Faculty of Sciences, Abdelmalek Essaadi University, Sebta Avenue, Mhannech 2, PO Box: 2121, 93030 Tetouan, Morocco

Received: December 19, 2022 Accepted: March 22, 2023 Online Published: March 30, 2023

doi:10.5539/ijsp.v12n2p42 URL: https://doi.org/10.5539/ijsp.v12n2p42

Abstract

Statistical tests of significance theoretically deal with two mutually exclusive hypotheses: the null and the alternative. However, at least in biomedical assays, only the null hypothesis is taken into account through type I error evaluation. But, basing these tests solely on type I error has two drawbacks: first, the probability limits (5%, 1% and 0.1%) arbitrarily set to the significance levels have no scientific justification. Second, acceptation of the null hypothesis is just a matter of chance, as it is mainly conditioned by the sample size due to its direct effect on the power of the test. In this sense, while the alternative hypothesis should be accepted due to its higher likelihood, the inference based on type I error alone would lead erroneously to accepting the null one. A numerical example illustrates how considering type I error alone, a same difference was declared non-significant first but turned out to significant thereafter when the sample size was increased. Therefore, the same null hypothesis was initially accepted and rejected afterwards. However when type II error was included in the test, the same decision was adopted no matter what the sample size was. This was possible through a reformulation of the alternative hypothesis. On the other hand, type II error may, in many cases have more far-reaching consequences than type I, and then should never be ignored, especially in assays dealing with human health, food, toxicity, etc.

Keywords: alternative hypothesis, inference, null hypothesis, sample size, type I error

1. Introduction

Statistical inference is fundamental in applied science studies. In fact, an experiment is almost useless if it is not accompanied with its inference about the population from which the sample was drawn. Statistical tests of significance, and specifically parametric tests, are very effective in making inference. But the way they are interpreted, especially by people with little statistical training, can ruin the advantage of such tests. This has also been magnified by the plethora of software available for statistical analysis. Their user-friendly interface encourages anyone to start performing analyses, sometimes very complicated, without having a clear idea about the scope and limitations of what is being done. On the other hand, the methodology behind the null hypothesis testing was intensively criticized especially in psychological research and related fields. A most recent review can be found in Szucs and Loannidis (2017). Several alternative solutions were proposed from using confidence intervals (Tukey, 1991) passing by figures with error bars (Fidler and Loftus, 2009) to a radical Bayesian approach (Ketler, 2021). All statistical tests of significance always deal with two hypotheses: H_0 the null hypothesis and H_A the alternative hypothesis, also called H_1 by some authors. Theoretically, the experimenter should evaluate the likelihood of veracity of each one of the two hypotheses and retain the one with the higher probability to be true. However, in most of experiments in biomedical research, H_A is usually discarded and all the light is shed on H_0. In practice, the experimenter starts setting a probability limit called α to the chances of rejecting a true H_0, what is known as committing type I error; and rejects H_0 if these chances are smaller than the fixed limit, usually set to 5%. But, basing a significance test on type I error solely, will give an incomplete picture of what is actually happening in the experiment, leading to biased conclusions. Moreover, when rejecting H_0 the experimenter can evaluate the risk to be mistaken by type I error probability; but, he has no idea about the risk of error he is bearing when H_A is rejected, i.e. H_0 is accepted. This is an important issue for any test of significance. But for simplicity, let us focus on the test of comparing two populations' means μ_1 and μ_2, for a variable called X. In such a test, the objective is to assess whether the null hypothesis (H_0: $\mu_1 = \mu_2$) could be true. Assuming that σ_X^2 is the variance and n is the sample size in both cases, the sensitivity of the test will depend on n, on σ_X^2 and also on Z, the value of the Standard Normal probability function. This latter will depend on the chosen significance level (α). In fact, H_0 must be rejected when the observed difference (d_0) between the sample means (say m_1 and m_2), will exceed a threshold value (d_L)

similar to the Least Significant Difference (LSD) used by Sir R. Fisher in his pairwise procedure to compare several means, e.g. Villars (1951), with:

$$d_L = Z_{(\alpha/2)}(2\sigma_X^2/n)^{1/2} \quad (1)$$

In the case of a two-sided test, while in the case of a single-sided test, d_L will be:

$$d_L = Z_{(\alpha)}(2\sigma_X^2/n)^{1/2} \quad (2)$$

Where:

σ_X^2 is the variance in both populations,

n is the sample size in both samples,

α is the probability limit to reject a true H_0, and

Z is the value of the Normal Standard function for that α.

As the test is based on d_0, fixing the value of α will set the value of d_L as the upper limit for a non-significant difference. In fact, α is the lower limit of p the probability to get, by random sampling, a difference at least as large as the one observed assuming a true H_0. Without any scientific justification, it has been agreed that setting the figures 5%, 1% and 0.1%, to α allows concluding that the observed difference is significant if it is higher than the value of d_L for the first figure, highly significant when it is higher than that for the second, and very highly significant when it surpasses the value for the third, e.g. Devore (2012). Then if the observed difference (d_0) is higher than the threshold value (d_L) for α equal to 5%, the probability p will be smaller than 5%. Therefore, it will be concluded that the observed difference is relevant and H_0 will be rejected, assuming a risk to be mistaken equal to at most 5%. Is it reasonable to assume that a 5% chance to be mistaken is low or even bearable? The answer will depend on the consequences of that mistake. If these consequences are severe, an even smaller value for α would be more suitable, e.g. 0.1%. This would minimise the risk of suffering such consequences. Conversely, if the consequences are not important, and there is a possible valuable benefit to be gained by rejecting H_0, it would be worthwhile to choose a higher α-value, for example 10%. This will increase the chances of benefiting if no mistake is made when rejecting the null hypothesis.

2. The Problem

Returning to the expressions (1 or 2) of d_L, since the variance is given by the population, the experimenter can only intervene on Z (which depends on α) and on n. Higher values of n, α or both will increase the sensitivity of the test, and vice-versa. At this point, is crystal clear that one cannot discuss, without some caution the results among similar experiments with different variances, sample sizes or probabilities of significance, because different values of d_L are in use. However, it is very common to see authors discussing own results with published others, without taking into account the value of d_L. In fact, the same difference may be significant in one study and not in another. Effectively, depending on the sample size, a large difference may be dismissed as non-significant because there are too few data; while a ridiculously small difference may be judged significant because there are too many observations in the sample. On the other hand, it is very common to misunderstand the expression 'The observed difference is not significant'. The right meaning is that there is not enough evidence in the data to negate that the observed difference could be due to sampling error; i.e. to declare that H_0 is false. However, many people understand that the observed difference is strictly due to sampling error since H_0 is accepted. Actually, there will always be differences as pointed out by Tukey (1991). A non-significant difference will become significant with higher values of n or α. In the expressions (1 or 2) of d_L, as the value of n increases the threshold of significance (d_L) will become smaller and smaller. To infinity, the threshold will tend to zero and then every difference, be it even little, will automatically be declared significant. The value of n can only be manipulated before conducting the experiment, while the value of α may be changed whenever desired. In two similar experiments with the same variance but with different sample sizes, the experimenter with the smaller sample sizes must increase the risk of being wrong rejecting the null hypothesis (type I error) to match the d_L of the experimenter with more data. This increased risk incurred by the first experimenter, is the price he must pay for having less data. Consider two experiments for a given variable (X):

Experiment A: $\sigma_X^2 = 50$ and $n = 64$ in both samples.

Experiment B: $\sigma_X^2 = 50$ and $n = 25$ in both samples.

If the experimenter A uses an α equal to 5%, his d_L is $1.96 \times (2 \times 50 / 64)^{1/2}$, which is equal to 2.45 for a one-sided test calculus. To discuss the results of the two experiments, the same d_L must be used in both. Therefore, the experimenter B (with less data) must use an α of about 11%, because his corresponding Z-value should be $2.45 / (2 \times 50 / 25)^{1/2}$ equal to

1.225, in order to get the same d_L. In this scenario, the two experiments handle the same value of d_L, and are therefore perfectly comparable. Unfortunately, one starts discussing own results once the experiment is over and the analysis is done. Moreover, the discussion usually addresses several papers and it is not a matter to go adapting the discussion to each one. A fundamental question is 'Who should set the value of d_L in a given experiment?'. The obvious answer should be 'The experimenter!'. However, in real life, it is decided simply by chance. This is so because the population variance is generally unknown and must be estimated from the sample at hand. The sample size n is often what is available, and more data could be obtained for reasons of cost, time, etc. This is, in fact, the fundamental controversy about the concept of statistical tests of significance, between mathematical theorists and practitioners, at least in biomedical assays. Theorists state that a first preliminary, study with some amount of data, should be carried out to have an idea on variance in order to set the power of the test (controlling type II error). The practitioner, in contrast, works with all the data available and obviates the power of the test. Several reference books on statistics propose, deducing the adequate sample size from the expression of d_L, (e.g., Dagnelie, 2013). Others prone using the effect size as defined by Cohen in 1968 (e.g., Hulley, Cummings, Browner, Grady, & Newman, 2007). All these proposals assume that some information should exist, about the population under study, before performing the experiment. But, getting such information is costly in time and funds, which makes it seldom available. Furthermore, tackling the problem through the power of the test is a rather twisting way to account for the alternative hypothesis in the testing procedure. A straightforward solution is presented in the next paragraphs.

The argument of the test is: assuming a true H_0, what is the probability p to get a value of d_0 equal or larger than the fixed d_L. Usually, when p exceeds 5%, it is considered a high figure and H_0 is accepted as true. It does not matter whether this probability is equal 6% or 99%; which is not reasonable, since a distinction must be made between two very distant situations. What about a p equal to 0.045 or 0.055? Applying literally the test of significance, should H_0 be rejected in the first figure and accepted in the second? Moreover, the observed difference d_0 may also be observed when H_A is true. Rejecting a true H_A is the other error (type II error) never considered in the testing, and whose probability is linked to that of type I error. Effectively, when the probability of type I error decreases, that of type II error increases, and vice-versa. Moreover, there are situations where type II error may have more far-reaching consequences than type I; and therefore, should receive more consideration. Let us look at the following example: A pharmaceutical laboratory has identified a substance that lowers blood cholesterol levels. Suppose the concern is to ensure that this drug does not affect the number of platelets (for example), which could increase the risk of blood clotting. To address this question, an assay is conducted on two groups of patients. The first group will receive the drug and second will receive a placebo. The null hypothesis (H_0) is then: 'The cholesterol-lowering drug has no effect on the number of platelets'. The alternative hypothesis (H_A) will be 'The cholesterol-lowering drug does have this effect'. In this assay, committing type I error does not have serious consequences, except losing the opportunity to market the drug. One could consider increasing the significance level, as recommended by some literature (e.g., Steel and Torrie, 1986) to be more certain that the drug has no effect. However, this would increase type II error probability, whose consequences are much more serious, i.e. putting on the market a very harmful product for human life (see detailed discussion in Sowey and Petrocz, 2017). The solution, therefore, relies on considering both types of error and choosing the hypothesis with the higher likelihood to be true. But at the same time, maximising profit and lowering the risk of harmful consequences. In this sense, in the example above, one could almost do without type I error, and consider only type II error due to its severity. Nevertheless, if type II error should be included in the decision-making, the alternative hypothesis should be defined more accurately.

3. The Solution

The main problem in fact is a bad formulation of the alternative hypothesis. While the null hypothesis H_0 is '$\mu_1 = \mu_2$', it is not sufficient to say that the alternative hypothesis H_A will be '$\mu_1 \neq \mu_2$'. The null hypothesis is precisely defined, but H_A seems vague. It is necessary to specify further and express H_A as '$\mu_1 = \mu_2 \pm \varepsilon$' in the case of a two-sided test, and '$\mu_1 = \mu_2 + \varepsilon$' or '$\mu_1 = \mu_2 - \varepsilon$' in the case of a single-sided test. The value of ε must be defined by the experimenter, allowing him full control over the statistical test. Think of this value of ε as a tolerance range for accepting the null hypothesis. This may accomplish the same task as the d developed by Cohen (); But without a need of previous information on the experiment. In the example above, let's say that nothing would happen if the number of platelets deviates by less than 200 units between the two population means (for example). Figure 1 represents graphically the case of a two-sided test; while figure 2 illustrates the case of a single-sided test, when the value of ε is positive (a) and when it is negative (b).

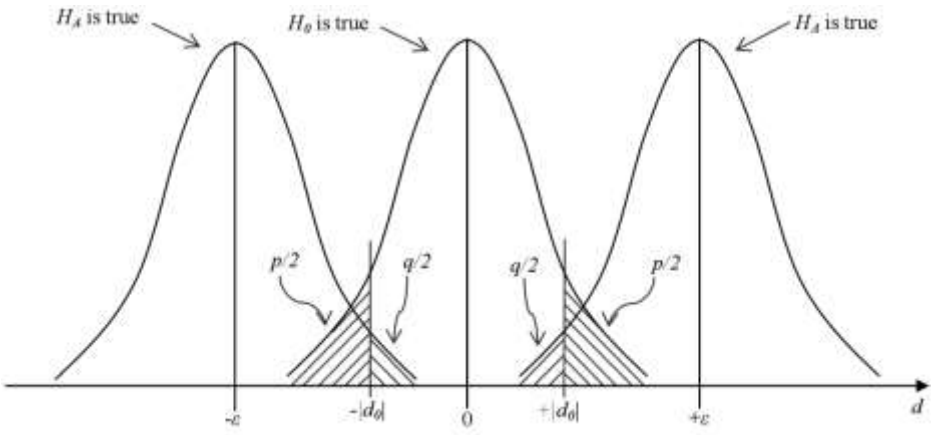
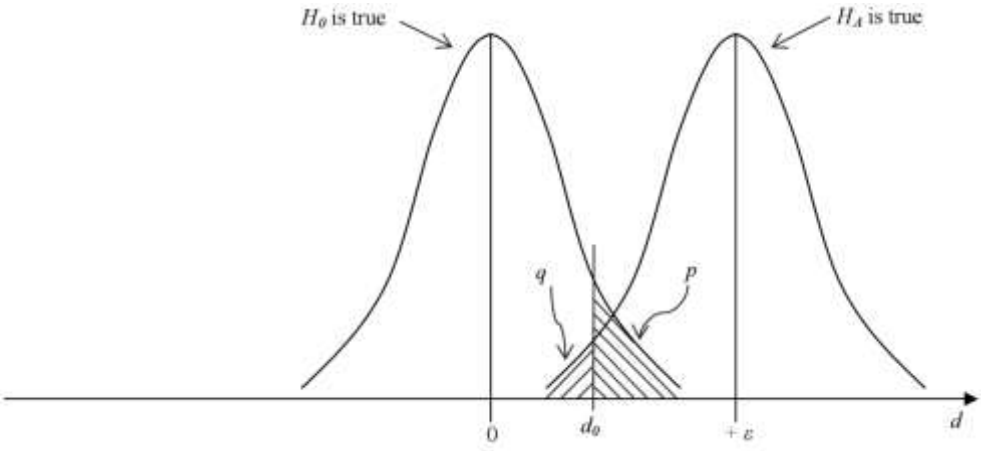


Figure 1. Graphical representation in the case of a two-sided test, where H_A is '$\mu_1 = \mu_2 \pm \varepsilon$'

The observed difference in the sample at hands is d_0. When H_0 is true, the distribution of d-values is centred on zero; while it is centred on the value of ε when H_A is true. The probability to get d-values equal to or greater than d_0, when H_0 is true, is equal to p (Called α when H_0 is rejected). The probability to get d-values equal to or less than d_0, when H_A is true, is equal to q (Called β when H_A is rejected). The probability p quantifies indirectly the likelihood of H_0 to be true, is therefore an indirect measure of its veracity. Likewise, the probability q quantifies indirectly the likelihood of H_A to be true, is therefore an indirect measure of its veracity. Thus, joggling with these two probabilities will allow deciding which hypothesis is more credible, and then which one should be retained as true. Also, the range of the difference between them will give an idea about the extent of uncertainty about the final conclusion of the test.

a



b

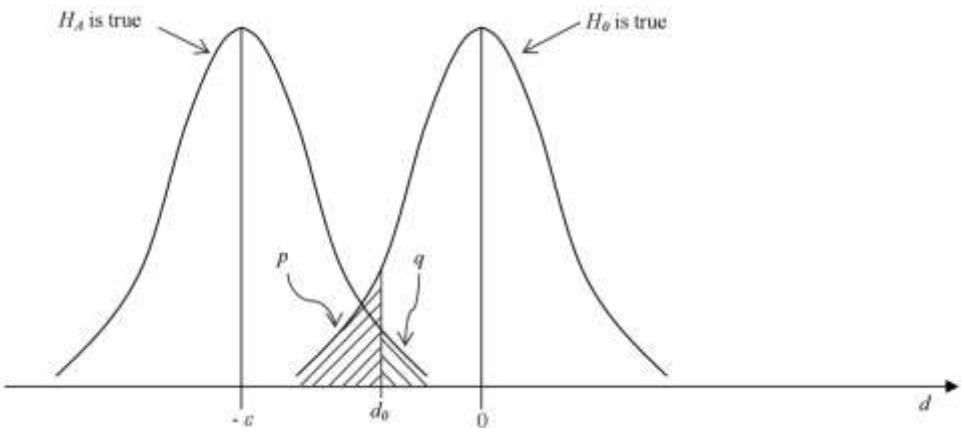


Figure 2. Graphical representation in the case of a single-sided test: (*a*) when H_A is '$\mu_1 = \mu_2 + \varepsilon$' and (*b*) when H_A is '$\mu_1 = \mu_2 - \varepsilon$'

4. Numerical Example

Suppose an experiment investigating whether workers working on night shifts, with disturbed sleep, eventually will develop high blood pressure. Forty workers are randomly chosen: twenty among the day shifts and twenty among the night shifts; and their resting systolic blood pressure is measured. Suppose that the assay gave the following sample means: m_1 of the day group equal to 12.00, while m_2 of the night group equal to 12.88. Let us assume for simplicity, that a previous large-scale study has indicated that the variance is equal to 4 in the two populations. Now let take the most important decision of the experiment. Let consider that if the difference in blood pressure is less than 1 cm Hg between the two populations, it will not be a relevant difference, which is quite reasonable. This decision will be the responsibility of the experimenter, and will give him full control over the experiment. The experimenter decides which difference is relevant and which is not.

Here, a single-sided test should be used, where H_A is '$\mu_1 = \mu_2 + \varepsilon$' and H_0 is '$\mu_1 = \mu_2$', where $\varepsilon = 1$. Applying the test of comparing two population means with known variance, the results of the statistical analysis are compiled in second row of table 1. The observed difference d_0 between the two sample-means is equal to 0.88. This difference will not be significant given the values of d_L shown in table 1. Effectively, d_0 is lower than the smallest value (1.041) corresponding to α equal to 5%, and therefore H_0 should be accepted. Accordingly, the probability p to get d-values equal to or greater than 0.88, when H_0 is true, is equal to 0.082. This p-value, greater than 5%, confirms that H_0 should be accepted. However, the probability q to get d-values equal to or less than 0.88, when H_A is true, is equal to 0.425. Surprisingly, H_A is more likely to be true than H_0 (0.425 versus 0.082), and therefore it is more reasonable to accept H_A and reject H_0. In other words, fulfilling the test in its classical form, the null hypothesis (H_0) is accepted. But, taking into account both types of error makes it wiser to accept H_A and reject H_0.

Let us increase the sample size to 200 workers, keeping the same observed difference d_0 (0.88). The results of this new analysis are reported in the third row of table 1. Now, re-running the calculation, d_0 surpasses the highest value of the significance threshold (0.618). It must be declared very highly significant, and then H_0 should be rejected. Accordingly, the probability p to get d-values equal to or greater than 0.88, when H_0 is true, is less than 0.001. As this value is now less than 5%, it confirms that H_0 must be rejected. However, the observed difference (0.88) is the same as in the case of 20 workers. But, as the sample size has increased, the values of d_L have decreased, and now the conclusion of the test is just the opposite. On the other hand, the probability q to get d-values equal to or less than 0.88, when H_A is true, is 0.274. Taking into account both types of error, the conclusion of the test remains the same: H_A is more likely to be true than H_0 (0.274 versus < 0.001). That is, the final decision about the experiment has changed working only with type I error; while that decision remains unchanged when both errors are considered. Now, suppose that the two cases above correspond to two assays conducted by two different authors. It is absolutely clear that some caution must be taken to ensure a correct discussion between their results. This is one of the negative effects of basing the decision of a test of significance on type I error alone. On the other hand, very large sample sizes may complicate the analysis and lead to absurd conclusions. This is the other negative effect of basing the test of significance on type I error alone.

Table 1. Results of the analysis of blood pressure assay for sample sizes of 20 or 200 workers in each group

	d_0	*SE*	Values of d_L for α at *5%*, *1%* or *0.1%**			*p***	*q***
20 workers	0.88	0.633	1.041	1.472	1.956	0.082	0.425
200 workers	0.88	0.200	0.329	0.465	0.618	<0.001	0.274

* Values of d_L were calculated using the formulae in (2), where $Z_{(\alpha)}$ is: 1.645 for 5%, 2.326 for 1% and 3.090 for 0.1%.

** Probabilities were calculated using the tool 'Probability Distribution Calculator' of Statistica (2014).

Let us now assume that the sample mean of the night group is 12.44 instead of 12.88; then d_0 is now equal to 0.44 (table 2). For 20 workers, this difference will not be significant given the values of d_L shown in table 1. Effectively, d_0 is lower than the smallest value (1.041) corresponding to α equal to 5%, and therefore H_0 should be accepted. Likewise, the probability p to get d-values equal to or greater than 0.44, if H_0 was true, would, this time, be equal to 0.245. This value, greater than 5%, will confirm that H_0 must be accepted. Accordingly, the probability q to get d-values equal to or less than 0.44, if H_A was true, would be 0.188. Since H_A is less likely to be true than H_0 (0.188 versus 0.245) the null hypothesis will be accepted. In the case of 200 workers, d_0 surpasses the first value of the significance threshold (0.618). It must be declared significant, and then H_0 should be rejected. Likewise, the probability p to get d-values equal to or greater than 0.44, if H_0 was true, will be equal to 0.014. This value, less than 5%, will confirm that H_0 must be rejected. Also, the probability q to get d-values equal to or less than 0.44, if H_A was true, will be equal to 0.003. That is, the probability of H_A to be true is smaller than that of H_0 (0.003 versus 0.014). Obviously, both probabilities are too small;

but if we were to accept one of the two hypotheses, it is clear that the one to be accepted should undoubtedly be H_0; because it is more than four times more likely to be true than H_A. This is a typical case where an excess of data may lead to erroneous conclusion about an experiment, when only type I error is accounted for in the statistical analysis. Again, the final decision on the experiment has changed using the classical form of the test, while that decision remains unchanged when both errors are considered.

Table 2. Results of the blood pressure assay assuming a sample mean of the night group equal to 12.44 instead of 12.88

	d_0	*SE*	*p*	*q*
20 workers	0.44	0.633	0.245	0.188
200 workers	0.44	0.200	0.014	0.003

5. Conclusion

If type II error is considered together with type I, the experimenter's inference on the outcome of an experiment will always be the same, regardless of the sample size. This is also an elegant way to get rid of the arbitrary values of 5%, 1% and 0.1% usually applied as significance levels. Also, the gap between the probabilities of these two errors will allow an evaluation of the extent of uncertainty about the final conclusion of the test. Moreover, taking into account the consequences of each one of these errors will further refine the final conclusion of the inference. However a better definition of the alternative hypothesis is needed in order to take into account of this hypothesis in the testing. Certainly, it is not always easy to define the alternative hypothesis with sufficient precision. But even a rough definition is better than letting chances decide, especially when the consequences of getting it wrong are costly, i.e. human health, food quality or toxicity, etc. In this sense, experimenters and especially those with little statistical background should keep their experimental design as simple as possible. Also, they should analyse their data using the simplest statistical tests, e.g. comparisons among means. Less is more as pointed out by Cohen (1990). This will simplify the establishment of the right alternative hypothesis. The extension to comparisons among frequencies (or proportions), as they follow Binomial-like probability laws, or among counts, as they follow a Poisson probability law, could be carried out using the approximation to such a test through the De Moivre-Laplace theorem, see Feller (1968) or DasGupta (2010) for a formal presentation. Additional work is needed to assess all the ins and outs of the best way to carry out statistical tests of significance.

Acknowledgements

The authors are indebted to anonymous reviewer who greatly helped improving the first version of the manuscript.

References

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Dagnelie, P. (2013). *Statistique Théorique et Appliquée, Tome 1 : Statistique descriptive et bases de l'inférence statistique* (3ème ed). Bruxelles, Belgium: De Boeck.

DasGupta, A. (2010). *Fundamentals of Probability: A First Course* (1st ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-5780-1_1

Devore, J. L. (2012). *Probability And Statistics For Engineering And The Sciences* (9th ed.). Boston, MA: Cengage Learning.

Feller, W. (1968). *An introduction to probability theory and its Applications: Volume 1* (3rd ed). New York, NY: John Wiley and Sons.

Fidler, F., & Loftus, G. R. (2009). Why Figures with Error Bars Should Replace p Values: Some Conceptual Arguments and Empirical Demonstrations. *Journal of Psychology*, *217*(1), 27–37. https://doi.org/10.1027/0044-3409.217.1.27

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2007). *Designing clinical research* (3rd ed). Philadelphia, PA: Lippincott Williams & Wilkins.

Ketler, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, *36*, 1263–1288. https://doi.org/10.1007/s00180-020-01034-7

Sowey, E., & Petocz, P. (2017). *A Panorama of Statistics, Perspectives, Puzzles and Paradoxes in Statistics* (1st ed).

New York, NY: John Wiley and Sons. https://doi.org/10.1002/9781119335139

Statistica, (2014). *Data analysis software system* (12th ver). Tulsa, OK: StatSoft Inc.

Steel, R. G. D., & Torrie, J. H. (1986). *Principle and procedures of statistics: A biometrical approach* (2nd ed.). New York, USA: McGraw Hills.

Szucs, D., & Loannidis J. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience: Section Cognitive Neuroscience*, *11*, 943–921. https://doi.org/10.3389/fnhum.2017.00390

Tukey, (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116. https://doi.org/10.1214/ss/1177011945

Villars, D. S. (1951). *Statistical Design and analysis of experiments for development research*. Dubuque, IA: WNC Brown Company Publishers.