REVIJA ZA ELEMENTARNO IZOBRAŽEVANJE
JOURNAL OF ELEMENTARY EDUCATION
Vol. 18, No. 1, pp. 107–124, March 2025

NEW POLYSTOCHASTIC STATISTICAL INFERENCE IN SOCIAL SCIENCES - DEFINING NEW RULES AND THRESHOLDS

SINIŠA OPIĆ
Faculty of Teacher Education, University of Zagreb, Zagreb, Croatia

KORESPONDENČNI AVTOR/CORRESPONDING AUTHOR
sinisa.opic@ufzg.hr

Potrjeno/Accepted
25. 2. 2025

Objavljeno/Published
31. 3. 2025

Abstract/Izvleček

The Null Hypothesis Significance Testing (NHST) framework has sparked considerable debate within the scientific community, leading to numerous studies advocating for a re-evaluation of the current system. New polystochastic statistical inference defines methods of statistical inference that integrate rules and thresholds for both rejecting the null hypothesis and confirming the alternative hypothesis. This approach unifies the control of respondents' influence on statistical significance and introduces criteria such as effect size and Bayesian inference for confirming the alternative hypothesis. Unlike NHST, polystochastic statistical inference controls Type I error (p-value) and aims to optimize the confirmation of evidence without increasing the risk of Type II errors.

Keywords:
Bayesian, effect size, NHST, p-value, polystochastic, social science, statistical inference.

Novo polistohastično statistično sklepanje v družboslovju – Določitev novih pravil in mejnih vrednosti

Okvir testiranja pomembnosti ničelne hipoteze (angl. Null Hypothesis Significance Testing – NHST) je sprožil precejšnjo razpravo v znanstveni skupnosti. To je vodilo do številnih študij, ki zagovarjajo ponovno oceno sedanjega sistema. Novo polistohastično statistično sklepanje definira metode statističnega sklepanja, ki združujejo pravila in pragove tako za zavračanje ničelne hipoteze kot za potrditev alternativne hipoteze. Ta pristop poenoti nadzor nad vplivom anketirancev na statistično pomembnost in uvede merila, kot sta velikost učinka in Bayesov sklep za potrditev alternativne hipoteze. Za razliko od NHST polistohastično statistično sklepanje nadzoruje napako tipa I (p-vrednost) in želi optimizirati potrditev dokazov brez povečanja tveganja napak tipa II.

Ključne besede:
Bayesov sklep, velikost učinka, NHST, p-vrednost, polistohastično, družboslovje, statistično sklepanje.

UDK/UDC:
303:311.21

DOI https://doi.org/10.18690/rei.4907

University of Maribor Press

Introduction: A 100-Year-Old Problem (Fisher 1925 - today)

Statistical significance has been a topic of intense debate in many scientific disciplines for a long time, particularly regarding its proper use and potential misuse. According to Rovetta (2024), it is one of the most controversial issues in contemporary science. The binary choice between statistically significant and insignificant results not only reflects a mathematical error but also fails to capture the complexity of statistical methods needed to communicate findings to the public, especially in fields like healthcare. This issue is not limited to medical research; it also affects most other scientific fields. Social sciences face significant challenges in statistical inference, which are compounded by the complexity of the phenomena being studied. Factors such as latent variables, issues of causality, inappropriate scales
for statistical analysis (parametric tests), implausibility, incoherence, hard-to-control extraneous factors, lack of objectivity, and reliability problems all contribute to these challenges. As one author notes, any scientific discipline that grapples with such challenges will achieve long-lasting relevance. To put this issue in historical context, the concept of statistical significance was first introduced by Ronald Fisher in 1925.

In fact, the concept began earlier with the work of Francis Edgeworth (1845–1926), who created a procedure for testing two arithmetic means (subsamples) that was later extended by Pearson to the Chi-square test (Pearson, 1900). Edgeworth's pioneering contribution lies at the beginnings of the development of statistical inference in testing arithmetic means and specificities such as skewness and kurtosis. Ronald Fisher further advanced these ideas in 1925, laying the groundwork for hypothesis testing in inferential statistics. His influential work, *Statistical Methods for Research Workers*, helped define the concept of statistical significance (p-value) as we understand it today.

The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty. (Fisher, 1925, 45)

Fisher's work laid the foundation for inferential statistics and initiated the field of hypothesis testing. Later, Newman and Pearson built upon Fisher's methods, introducing the concepts of Type I error (rejecting the null hypothesis, H0, when it is true) and Type II error (failing to reject H0 when it is false) (Perezgonzalez, 2015).

The contributions of Newman and Pearson are significant, particularly in the context of enhancing statistical power for repeated sampling while considering Type I and II errors, as well as effect size (Holmberg, 2024). However, a notable drawback of the Newman and Pearson approach is its rigidity; it relies on a series of predetermined steps and lacks the flexibility found in Fisher's method. McShane et al. (2019) emphasize the need to abandon the NHST approach (null hypothesis significance testing) in all areas of scientific activity in the biomedical and social sciences, i.e. they offer a broader concept (but one that is unclear): "Results need not first have a p-value or some other purely statistical measure that attains some threshold before consideration is given to the currently subordinate factors. Instead, treated continuously, statistical measures should be considered along with the currently subordinate factors as just one among many pieces of evidence and should not take priority thereby yielding a more holistic view of the evidence" (p. 25).

Although the p-value is considered the "scientific default" in inferential statistics, it is frequently misused and misinterpreted. Many papers in the literature emphasize the need to redefine p-values, supplement them with new methods, or even abolish them completely, leading to confusion across various scientific disciplines. Additionally, some scientific journals discourage the use of p-values. Considering the ongoing concerns surrounding statistical significance and p-values, the American Statistical Association (ASA) published the Statement on Statistical Significance and P-Values. This document includes several important statements, as noted by Wasserstein and Lazar (2016):

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency p-value; debate.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Consequently, the ASA presents a significant challenge in the realm of inferential statistics and clearly defines the meaning of the p-value, offering a more comprehensive approach to statistical inference.

The conclusion suggests the incorporation of new methods "but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct" (Ibid. p. 132).

The limit of statistical significance, or the null hypothesis, has been a topic of considerable debate in the literature for many years. Opinions range from calls for its complete abolition to suggestions like those from Benjamin et al. (2017), who propose reducing the threshold from 0.05 to 0.005. According to the authors, this change would be a significant step forward that could enhance reproducibility in research. They emphasize that this redefinition of the p-value, with its new standard, pertains specifically to research records and not to scientific publications. The aim is to observe how scientists behave under stricter criteria. I believe such a dual approach is unnecessary since we often select levels of statistical significance (0.05 and 0.01) for hypothesis testing in the social sciences (inferential statistics). Of course, the 0.005 level is quite stringent, and the authors highlight this viewpoint in their paper. They justify their approach in relation to Bayesian analysis, as it corresponds to Bayes factors ranging from approximately 14 to 26 in favour of the alternative hypothesis (H1).

Lakens et al. (2018) suggest that instead of lowering the significance threshold from 0.05 to 0.0005, researchers should abandon the term "statistical significance" altogether. They recommend that scientists focus on controlling error rates with an alpha level that is determined by the researcher. Similarly, de Ruiter (2019), while critiquing the proposal to lower the significance level to 0.005, argues that setting an alpha level of $p \leq 0.005$ does not enhance replicability. He believes that the rationale for adopting a new alpha level of 0.005 is weak and that such a change could potentially harm scientific practice.

I agree with the recommendations made by the ASA, emphasizing that it is the responsibility of educators to ensure that scientists understand the term "statistically significant." However, without clear criteria, we risk entering a realm of scientific "sfumato", a form of voluntarism lacking defined standards. The p-value criterion ($p < 0.05$) is inadequate because it is influenced by sample size and fails to accurately reflect the true magnitude of differences or relationships (e. g. in subsamples). Moreover, rejecting the null hypothesis does not provide evidence for confirming the alternative hypothesis. Null Hypothesis Significance Testing (NHST) is a statistical procedure that involves establishing a null hypothesis, generating data related to it, and assessing how much the outcome disagrees with the null hypothesis, using statistical estimates.

While it may be a questionable criterion, having some standard is certainly better than having none at all, or focusing solely on the individual scientist and their choice of methods and procedures.

It is important to define when something exists or does not exist, and the criteria used to reject the null hypothesis when confirming an alternative hypothesis. When Fisher established the $p<0.05$ threshold, he acknowledged that it was not necessarily the best option available. The absence of any criteria is indeed worse than having an arbitrary yet compromise-based standard. Now, 100 years after Fisher introduced this threshold, we still struggle to reach a scientific consensus on how to conduct statistical inference. Authors tend to concentrate too much on identifying what is wrong, what requires change, and proving the "pollution" of our current model instead of seeking an effective solution. This solution will not be perfect; after all, determinism is increasingly less relevant as a scientific postulate, making way for a focus on probabilism.

Polystochastic Statistical Inference in the Social Sciences

Polystochastic Statistical Inference in the Social Sciences is a concept that combines a revised form of the Null Hypothesis Significance Testing (NHST) system, dependent on the sample size (n), with Bayesian inference and effect size, within certain limits. The framework of polystochastic statistical inference consists of two main components:
1. Rejection of the null hypothesis (H0).
2. The confirmation or rejection of the alternative hypothesis (Hn).

When the sample size is >120 (n >120):

- Use a significance level of $p \leq 0.01$ (or smaller). For large samples ($n > 100$, i.e., 120), the sample mean reliably approximates a normal distribution, particularly in populations with pronounced skewness. A stricter criterion is necessary because sample size has a significant effect on statistical significance, as noted by Opić and Rijavec (2022). This leads to an increased risk of Type I error. With larger samples, even minor differences can appear statistically significant. Additionally, as the sample size increases, the standard error decreases. This is particularly important when dealing with pronounced asymmetry (skewness) or variability (variance), since sample size greatly influences the normalisation of the distribution. One valuable and effective method is bootstrapping (resampling), where the sample is treated as a population.

In Bayesian inference, sample size is less critical than in the Frequentist approach; consequently, a smaller sample size is sufficient to achieve the same level of efficiency, as discussed by Ali, Waheed, Shah, and Raza (2023).

Although it is generally considered that the Central Limit Theorem (CLT) begins to apply when the sample size (n) is greater than 30, this is conditional. The assumption is that the population does not exhibit significant skewness or kurtosis. For populations with pronounced skewness or heavy tails (platykurtic distributions, such as the t-distribution with low degrees of freedom), a much larger sample size is required to meet the prerequisite of normal distribution. However, this does not apply to distributions like the Pareto distribution, which is not influenced by the CLT since it has unlimited variance. For sample sizes greater than 120, the t-distribution closely resembles a normal distribution. This is why a sample size limit of 120 is defined.

So, if we have a category of large samples ($n>120$), CLT also works in the case when asymmetries (skewness) and kurtosis (heavy-tailed) are expressed, which indicates that we meet the main prerequisite for normal distribution, which is required for parametric statistics. However, we then need to reduce the level of statistical significance to $p\leq0.01$ because the size of the sample affects the statistical significance; i.e., we will reject the null hypothesis sooner on large samples than on small samples. The 0.01 criterion is not too strict, and it is already used in medical research (often a much stricter criterion), and in research with high stakes; accordingly, it should become the default for social sciences as well. Therefore, stronger evidence is needed to reject the null hypothesis, but the type 2 error does not increase significantly (as in the case of proposals, it is reduced to a very strict criterion, e. g., 0.0005; (Benjamin et al. 2017).

Rationale for application of the criteria - there are no restrictions on the application of the change of criteria. The advantage is the fact that stronger evidence against the null hypothesis will be needed, thus reducing the type one error. Reducing the p-value criterion to $p\leq0.01$ controls the influence of the sample size (n) on statistical significance when it comes to rejecting the null hypothesis, but it still does not solve the confirmation of the alternative hypothesis.

When the sample is <120 ($n<120$)

- Use the significance level $p\leq0.05$ (or smaller). In smaller samples, the influence on statistical significance is not so pronounced.

Both basic conditions refer only to confirming/rejecting the null hypothesis.

To confirm the alternative hypothesis, at least 2 out of 3 criteria must be met:

1. ****Statistical Significance**:** Confirmed level of statistical significance (n>120; p≤0.01 (or smaller), i.e. n<120, p≤0.05 or smaller) to reject the null hypothesis. The null hypothesis must be rejected (Mandatory).

2. ****Effect Size**:** The effect size should be at least moderate. The effect size (d) provides insight into the actual magnitude of differences in differential designs (Sullivan and Feinn, 2012; Balow, 2017). While statistical significance indicates that the result is unlikely to occur by chance, effect size quantifies how substantial the differences are. The most commonly used effect size in differential designs is Cohen's d (Cohen, 1968), where 0.2 is considered small, 0.5 medium, and 0.8 large. Therefore, to provide evidence in favour of the alternative hypothesis (Hn), it is necessary to meet the criterion of a medium (moderate) effect size for a given test.

Table 1

Shows the most commonly used effect sizes with reference values.

Effect size	**small**	**Medium**	**large**
Cohen`s d (t test)	0.2	0.5	0.8
Eta squared η^2 (ANOVA)	0.01	0.06	0.14
Cohen's f (one way ANOVA/ANCOVA)	0.1	0.25	0.4
Omega squared $\omega 2$ (ANOVA)	0.01	0.06	0.14
Multivariate Omega squared $\omega 2$ (one way ANOVA, MANOVA)	0.01	0.06	0.14
F-Squared $f2$ (multiple nad partial corr)	0.02	0.15	0.35
r Pearson	0.1	0.3	0.5
Odds Ratio (OR)	close to 1	Around 2	3or more
Odds ratio (2*2)	1.5	3.5	9.0
$\eta 2$ (multiple regression)	0.02	0.13	0.26
Cohen's ω (chi square)	0.1	0.3	0.5
Spearman rho (Friedman)	0.1	0.3	0.5
Cramer V (r x c frequency tables)	0.1 (Min(r-1,c-1)=1), 0.07 (Min(r-1,c-1)=2), 0.06 (Min(r-1,c-1)=3)	0.3 (Min(r-1,c-1)=1), 0.21 (Min(r-1,c-1)=2), 0.17 (Min(r-1,c-1)=3)	0.5 (Min(r-1,c-1)=1), 0.35(Min(r-1,c-1)=2), 0.29 (Min(r-1,c-1)=3)

(Source; Vacha-Haase and Thomson, 2004; Cohen, 1992; Cohen, 2008); https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize)

Of course, as with any other statistical indicator, there are limitations. For example, McGrath and Meyer, 2006) show that rpb (point-biserial correlation) is sensitive to sample size, but when it comes to unequal variances, this is also the case with Cohen's d. (compare Ruscio, 2008), and a correction was proposed; they suggested larger values to represent effects (small-medium-large) as the group sizes become more unequal. Calculating effect size is an arbitrary procedure, similar to what Fisher noted about $p<0.05$. Therefore, it is recommended for use only when no better basis for estimating the index is available (Cohen, 1988, p. 25). However, there is no ideal statistical procedure, and there are no certain limitations, but the Effect size is very little influenced by the sample size and shows the real relationship between the variables (shown in the empirical part of the paper) and a very useful indicator in favour of the alternative hypothesis (Hn).

Rationale for the application of the criteria – the list of effect sizes is large, and the author selects a specific one that corresponds to the test used to test the hypotheses. Table 1 shows the most used ones, but this does not mean that the list is not expanding. However, the author chooses a certain and calculated value that should have at least a medium effect to fulfil this criterion - in favor of the alternative hypothesis (Hn).

3. ****Bayes factor****

It should be BF(10) > 3, that is, indicating Moderate evidence for H1.

Bayes factor and Bayesian inference are highly useful statistical approaches, and many papers indicate the advantages of using them over p-values (Stern, 2016., Hoijtink, van Kooten, Hulsker, 2016., Jarosz and Wiley, 2014., Assaf and Tsionas, 2018, Goodman, 2008, 2005., Lavine and Schervish, 1999., Morey, Romeijn and Rouder, 2016, Andraszewicz et al, 2015).

Bayes factor defined; $\frac{Pr\,(Data \mid \mathbf{H1})}{Pr(Data \mid \mathbf{H0})}$, where is the posterior probability.

Pr (H0|Data) = $\frac{\Pr\,(Data) \mid \mathrm{H0}) \cdot \Pr\ (\mathrm{H0})}{Pr(Data)}$, analogously

Pr (H1|Data) = $\frac{\Pr\,(Data) \mid \mathrm{H1}) \cdot \Pr\ (\mathrm{H1})}{Pr(Data)}$ (Bayes theorem)

The Bayes factor is a significant step forward in statistical inference, especially because it allows insight into the probability of an alternative hypothesis (which is not the case with NHST), but like all approaches in statistics, it has limitations.

One of these is the Jeffreys-Lindley paradox (Bartlett, 1957; Lindley, 1957). This concerns the influence of sample size on the BF value. In the frequentist approach, large samples affect lower p values, i. e., in favour of the Alternative hypothesis (Hn - or rejecting an H0), while in the Bayesian approach, large samples affect higher values of BF(01), i.e., in favour of H0. So, we have a paradox because the sample size in the frequentist approach significantly affects the probability of rejecting H0, but at the same time in the Bayesian approach, it can affect a higher probability in favour of H0 (Huisman, 2023).

In the literature and machine learning, the interpretation of BF is confusing: the interpretation of BF can be B10, i. e., alternative vs null hypothesis, or BF01, null vs alternative. Most often, when the label is not used, it means BF01.

BF10>1: Evidence favours H1. BF10<1: Evidence Favors H0.
BF01>1: Evidence favours H0. BF01<1: Evidence Favors H1.
Bayes interpretation table (Adjusted to BF10; From Jeffreys, 1961)

Table 2
Bayes factor (BF10)

>	100	Extreme evidence for H1
30	100	Very strong evidence for H1
10	30	Strong evidence for H1
3	10	Moderate evidence for H1
1	3	Anecdotal evidence for H1
	1	No evidence
1/3	1	Anecdotal evidence for H0
1/10	1/3	Moderate evidence for H0
1/30	1/10	Strong evidence for H0
1/100	1/30	Very strong evidence for H0
<	1/100	Extreme evidence for H0

This condition stipulates that the Bayes Factor (BF10) must be at least 3 - indicating moderate evidence for the alternative hypothesis (H1). While a stricter criterion could have been applied, it would likely have caused more issues than benefits, particularly when considering certain limitations of Bayesian inference, such as the Jeffreys-Lindley paradox. This is especially relevant in cases of specific definition prior probability (P(θ)), particularly when using a non-informative prior (uniform).

Rationale for the application of the criterion - the application of this criterion may be limited because e.g. in multivariate tests, the application of Bayesian inference (BF) is limited and under development. Moreover, in complex models, there are a number of limitations (challenges) in the application of BF (Bollen, Harden, Ray, and Zavisca, 2014), including the problem of using an ordinal scale, and the problem of using BF in non-parametric tests (Yuan, and Johnson, 2008). Of course, there are always challenges, but at the same time, most of the works in the univariate approach have BF calculations in statistical programs for data processing, and further development and application are expected.

We can therefore show Polystochastic Statistical Inference in the Social Sciences graphically (Table 3):

Table 3

Polystochastic Statistical Inference in the Social Sciences

Rejecting null hypothesis (H0)	In case n<120 In case n>120	p≤0,05 (or smaller) p≤0,01 (or smaller)
Proving the alternative hypothesis (Hn)	Rejected H0 when n<120; p≤0,05 or smaller when n>120; p≤0,01 or smaller	Condition 1 (mandatory)
	Bayes factor BF10 > 3	Condition 2
	Effect size - medium	Condition 3
		A total of 2 out of 3 conditions must be met

An empirical example

For the simulations (scenarios X1, X2, and X3), a matrix was utilized with the independent variable being study type (undergraduate, graduate, integrated study; ∑n=75) and the dependent variable measured on an ordinal scale using a Likert scale with 5 points.

Differences between sub-groups were tested using ANOVA;

Scenario X1, n=75

$$\begin{matrix} n1 = 25 \\ \text{X1 } n2 = 21 \\ n3 = 29 \end{matrix} \xrightarrow{yields} \text{F(2.72) =1.451, p=0.241; MSB=1.672; MSW=1,152} \quad \begin{matrix} \bar{x}1 = 3.40;\ \sigma1 = 1.258;\ \text{stdErorr} = 0{,}252 \\ \bar{x}2 = 3.71;\ \sigma2 = 1.056;\ \text{stdErorr} = 0{,}230 \\ \bar{x}3 = 3.90;\ \sigma3 = 0.900;\ \text{stdErorr} = 0{,}167 \end{matrix}$$

Effect size; $\eta^2 = \mathbf{0.039}$; CI (95%) = 0.000_{lower} to 0.138_{upper}
ε^2=0.012; CI (95%) =-0.028_{lower} to 0.114_{upper}

BF(10) =0.052 (JZS)

In a sample of n=75, the null hypothesis, which posits that there are no differences between the subsamples concerning the dependent variable, is confirmed. The effect size, measured by eta squared, indicates a very weak real difference. Moreover, Bayesian inference shows a Bayes Factor of BF(10) =0.052, which does not lend support to the alternative hypothesis (Hn).

When the results are multiplied in a larger matrix sample of n=150, the findings are as follows (scenario X2):

Scenario X2; n=150

$$\begin{matrix} n1 = 50 \\ \text{X2 } n2 = 42 \\ n3 = 58 \end{matrix} \xrightarrow{yields} \text{F(2.147) =2.963, p=0.055; MSB=3.345; MSW=1.129} \quad \begin{matrix} \bar{x}1 = 3.40;\ \sigma1 = 1.245;\ \text{stdErorr} = 0.176 \\ \bar{x}2 = 3.71;\ \sigma2 = 1.043;\ \text{stdErorr} = 0.161 \\ \bar{x}3 = 3.90;\ \sigma3 = 0.892;\ \text{stdErorr} = 0.117 \end{matrix}$$

Effect size; $\eta^2 =$ 0.039; CI(95%)=0.000_{lower} – 0.108_{upper}
ε^2=0.026; CI(95%)=-0.014_{lower} – 0.096_{upper}

BF(10) =0.117 (JZS)

By duplicating the results in the matrix, the arithmetic mean remained unchanged (n1, n2, n3). However, the standard errors decreased because the denominator includes √n. The results of the F ratio suggest that we are nearing the threshold for rejecting the null hypothesis at a statistical significance level of $p < 0.05$. Nonetheless, the effect size values remained the same ($\eta^2 = 0.039$). Additionally, the Bayesian inference results (BF10 = 0.117) do not support the alternative hypothesis.

Then, multiplying the results in the matrix (N=300), the results are as follows (scenario X3):

Scenario X3; n=300

$$X3\ \begin{matrix} n1 = 100 \\ n2 = 84 \\ n3 = 116 \end{matrix} \xrightarrow{yields} \begin{matrix} \text{F(2.297)=5.896;p=0.003;} \\ \text{MSB=6.689; MSW=1.118} \end{matrix} \begin{matrix} \bar{x}1 = 3.40;\ \sigma1 = 1.239;\ \text{stdErorr} = 0.124 \\ \bar{x}2 = 3.71;\ \sigma2 = 1.036;\ \text{stdErorr} = 0.113 \\ \bar{x}3 = 3.90;\ \sigma3 = 0.888;\ \text{stdErorr} = 0.082 \end{matrix}$$

Effect size; $\eta^2 = 0.039$; CI (95%) = 0.005_{lower} to 0.086_{upper}
$\varepsilon^2 = 0.032$; CI (95%) = - 0.002_{lower} to 0.080_{upper}

BF(10) =1.146 (JZS)

For a sample of 300 respondents, the results indicate a rejection of the null hypothesis, with a p-value of 0.003. In scenario X3 (n=300), the arithmetic means remained unchanged since the data set was identical. However, the null hypothesis significance testing (NHST) still led to a rejection of the null hypothesis (p=0.003). The effect size, measured by η^2, was consistent at 0.039 across the x1, x2, and x3 models, indicating a very weak real difference. Additionally, the Bayesian inference showed a Bayes Factor (BF10) of 1.146, which does not support the alternative hypothesis (Hn).

The Jeffreys-Lindley paradox is not evident in this case, despite using a non-informative prior (a uniform prior over the mean). This is because increasing the sample size (n) did not result in a decrease in the Bayes factor in support of the null hypothesis. Clearly, the sample size has a stronger impact on the p-value than on the Bayes factor. In this case, the null hypothesis is rejected in the X3 model. However, there is not enough evidence to confirm the differences between the subsamples, as neither Bayesian analysis nor effect size support such a conclusion. According to the polystochastic inference approach, to validate the alternative (affirmative) hypothesis, at least two out of three conditions must be met. In this instance, only one condition has been satisfied. Therefore, even though the null hypothesis was rejected in scenario 3, the alternative hypothesis (Hn) regarding the existence of differences between the subsamples is not confirmed.

In the subsequent section, a new simulation is introduced (y1; y2; y3; y4). The differences between male students (n=32) and female students (n=40) regarding the dependent variable (*The online classes were well organized*) using a sample size of n=72 were tested. A T-test for independent samples was conducted. The results are as follows:

Scenario Y1; n = 72

$$Y1\begin{matrix} n1 = 32 \\ n2 = 40 \end{matrix} \xrightarrow{yields} t(70) = -0.956,\ \mathbf{p=0.342} \qquad \begin{matrix} \bar{x}1 = 3.41 & \sigma 1 = 1.012; \text{stdError} = 0.179 \\ \bar{x}2 = 3.65 & \sigma 2 = 1.122; \text{stdError} = 0.177 \end{matrix}$$

Although the disproportion of the subsample is partially expressed, the group is homogeneous; F (70;68.957) =0.126, p=0.724. Also, when sampling the distribution, there is no significant asymmetry, skew=-0.813, nor is it a pronounced significant leptokurtic distribution (Kurtosis=0.272). Analogously, in the case of subsamples, the sampling distribution is not markedly asymmetric, nor is significant kurtosis pronounced.

Effect size; Cohen's d **= - 0.227**; CI (95%) = -0.692_{lower} to 0.240_{upper}

Hedges´ correction = - 0.224; CI (95%) = -0.685_{lower} to 0.238_{upper}

Glass delta = - 0.217; CI (95%) = -0.683_{lower} to 0.251_{upper}

BF(01) =**3.650**, posterior distribution in intervals is shown in Figure 1

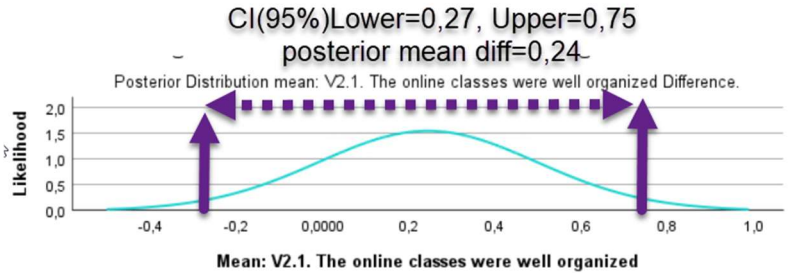

Figure 1 - posterior distribution (Credible Intervals)

Thus, on a sample of 72 subjects, the NHTS approach is confirmed by the null hypothesis of no difference between subsamples, Cohen's d indicates a very low difference between subsamples, nor does and BF(01) favour the alternative (affirmative) hypothesis.

In the further simulation (Y2), the results in the matrix were multiplied; (n=144), the results are as follows:

Scenario Y2; n = 144

$$\text{Y2}\ \begin{matrix} n1 = 64 \\ n2 = 80 \end{matrix} \xrightarrow{yields} \text{t (142)= - 1,362; p= 0,175;}\ \begin{matrix} \bar{x}1 = 3.41 & \sigma 1 = 1.003; \text{stdError} = 0.125 \\ \bar{x}2 = 3.65 & \sigma 2 = 1.115; \text{stdError} = 0.125 \end{matrix}$$

The set is homogeneous; F (142;139,994)=0.256, p=0.614
Effect size; Cohen's d = - 0.228; CI (95%) = - $0.558_{\text{lower to}}$ 0.102_{upper}
Hedges´ correction = -0.227; CI (95%) = - 0.555_{lower} to 0.101 $_{\text{upper}}$
Glass delta = - 0.219; CI (95%) = - 0.548_{ower} to 0.113 $_{\text{upper}}$

BF(01) =3.173

There was a decrease in the p-value (0.342 to 0.175), which still does not indicate the rejection of the null hypothesis, and at the same time, the BF and the effect size are not in favor of H1. In the further simulation (Y3), the results in the matrix are multiplied; (n=288), the results are as follows:

Scenario Y3, n = 288

$$\text{Y3}\ \begin{matrix} n1 = 128 \\ n2 = 160 \end{matrix} \xrightarrow{yields} \text{t(286) = -1,933; } \mathbf{p= 0,054}\ \begin{matrix} \bar{x}1 = 3.41 & \sigma 1 = 1.000; \text{stdError} = 0.088 \\ \bar{x}2 = 3.65 & \sigma 2 = 1.111; \text{stdError} = 0.088 \end{matrix}$$

The set is homogeneous; F (286;282,067) = 0.515, p=0.474 and t value is used: equal variance assumed.
Effect size; Cohen´s d = -0.229; CI(95%)= - 0.462_{lower} to 0.004_{upper}
Hedges´ correction= -0.229; CI(95%)= -0.461_{lower} to 0.004 $_{\text{upper}}$
Glass delta =-0.219; CI(95%)= -0.453_{ower} to 0.015 $_{\text{upper}}$

BF(01) =1.745

In the Y3 simulation, the impact of sample size on statistical significance is evident. At the $p < 0.05$ level, the null hypothesis (H0) can be rejected since it is at the threshold value. However, it is not rejected at the $p \leq 0.01$ level. The arithmetic means, Cohen's d (effect size), Hedges' g correction, and Glass delta all remain unchanged (very small differences) and indicate a small effect. Additionally, the Bayes Factor BF(01) does not support the alternative hypothesis.
And finally, we have the Y4 simulation (n=576)

Scenario Y4, n=576

$$\text{Y4}\ \begin{matrix} n1 = 256 \\ n2 = 320 \end{matrix} \xrightarrow{yields} \text{t (574)=-2.739; p = 0.006}\ \begin{matrix} \bar{x}1 = 3.41 & \sigma 1 = 0.998; \text{stdError} = 0.062 \\ \bar{x}2 = 3.65 & \sigma 2 = 1.110; \text{stdError} = 0.062 \end{matrix}$$

Effect size; Cohen's d = - 0.230; CI (95%) = - 0.394_{lower} to - 0.065_{upper}
Hedges´ correction = - 0.229; CI (95%)= - 0.394_{lower} to -0.065 $_{\text{upper}}$
Glass delta = - 0.220; CI (95%) = - 0.385_{ower} to - 0.054 $_{\text{upper}}$

BF(01) = 0.378, or BF(10) =1/BF(01) =2.64. The posterior distribution in the intervals is shown in Figure 2 (prior is flat; noninformative)

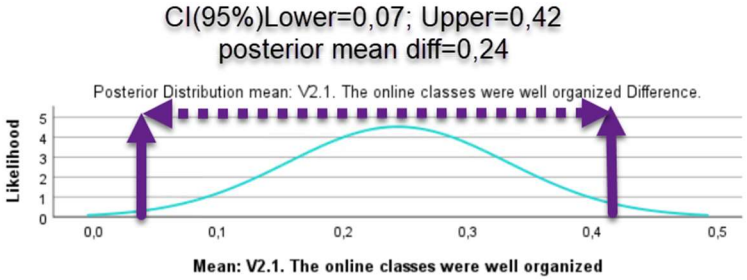

Figure 2- posterior mean difference ($\text{Mean}_{\text{diff}}$ Posterior)

In this case, the null hypothesis is rejected at a significance level of $p < 0.01$ since $p = 0.006$. However, there is still no evidence to support the alternative hypothesis. Cohen's d is -0.230, indicating a low effect size, and the Bayes Factor (BF01) is 0.378, which means that BF10 is 1/BF01, resulting in BF10 = 2.64. Although the value of BF10 (favouring the alternative hypothesis) increased with the sample size, moderate evidence for the alternative hypothesis (Hn) was still not achieved.
In the simulations involving scenarios X1, X2, X3, X4, and Y1, Y2, Y3, Y4, Y5, the influence of the sample size of the respondents is evident. Additionally, the effectiveness of the Polystochastic Statistical Inference in the Social Sciences approach is highlighted, as it controls for type 1 errors ($n > 120$, $p < 0.01$) and the probabilities of confirming the alternative (affirmative) hypothesis (Hn).

Conclusion

Even after 100 years since the significant contributions of Sir Ronald Aylmer Fisher to the field of statistical inference, many papers published today continue to demonstrate that this approach has major flaws. It often leads to misconceptions, incorrect interpretations, wrong conclusions, and generalizations.

Furthermore, it is estimated that a substantial percentage of papers—approximately 80%—in the social sciences arrive at erroneous conclusions based on the null hypothesis significance testing (NHST) approach. This increasingly resembles *Gödel's Incompleteness Theorem,* which, when interpreted, relates to the idea of proving something that cannot be proven. However, as early as 1925, Fisher acknowledged that this approach was not the best solution. Today, numerous papers highlight the shortcomings of the existing NHST system and the limitations of other methodologies. Polystochastic statistical inference in the social sciences introduces a new approach that clearly defines the boundaries of statistical inference. By lowering the p-value threshold from 0.05 to 0.01 (or smaller) for large samples ($n > 120$), we can better control the influence of sample size on statistical significance, effectively reducing the risk of a Type I error. While some research suggests that an even stricter criterion may be necessary, this can lead to an increased risk of a Type II error. There is no universally ideal threshold. However, the significant advantage of the polystochastic statistical inference approach in the social sciences lies in its ability to support an alternative (affirmative) hypothesis when the null hypothesis is rejected.

To confirm the alternative hypothesis, 2 of 3 conditions must be met; the compulsory condition is that the null hypothesis is rejected, then the Effect size is at least medium, and BF (10) > 3. We could see this as the need to introduce a stricter criterion (e.g., BF (10) >10 or more, indicate the limitations of Bayesian inference for complex models (which is correct), or indicate the operation of the Jeffreys-Lindley paradox, the problematic nature of the non-informative prior. However, Polystochastic Statistical Inference in the Social Sciences offers a framework that provides clear rules (thresholds) for statistical inference in the social sciences. The approach is set to allow the author to control the influence of sample size on the probability of rejection of the null hypothesis, but what is more important is that it has a framework for confirming the alternative hypothesis. The author has the option of choosing conditions (2/3) because it is assumed that for certain multivariate tests, statistical programs still do not offer Bayesian, or, for example, with certain non-parametric tests, Bayesian is not yet often being used (or is controversial).

The new approach, Polystochastic Statistical Inference in the Social Sciences, represents a significant advance in statistical inference within this field, providing clear rules and thresholds. It maintains flexibility in its application, avoiding a substantial increase in Type II error, even if we were to pursue a further reduction in p-values. Additionally, it offers a balanced method for confirming alternative hypotheses.

Authors are encouraged to specify which approach they have chosen in their work, whether it be NHST or Polystochastic Statistical Inference (PSSI). Beyond this, the approach provides valuable statistical insights, such as confidence intervals and credible intervals, aimed at enhancing our understanding of the data. Ultimately, PSSI establishes a clear framework and threshold for statistical inference in the social sciences.

Acknowledgment - I would like to thank my colleague Irena Tadic for the matrix used for the empirical part and the reviewers for their valuable suggestions. I am grateful to the Centre for Educational Measurement and Assessment (CEMA) for its support and to Fisher Library the University of Sydney for the space to work (100 years after Fisher).

References

Ali, S., Waheed, M., Shah, I., and Raza, S. M. M. (2023). Bayesian sample size determination for coefficient of variation of normal distribution. *Journal of Applied Statistics*, *51*(7), 1271–1286. https://doi.o–rg/10.1080/02664763.2023.2197571

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., and Wagenmakers, E. J. (2015). An Introduction to Bayesian Hypothesis Testing for Management Research. *Journal of Management, 41*(2), 521-54. https://doi.org/10.1177/0149206314560412

Assaf, G. A., and Tsionas, M. (2018). Bayes factors vs. P-values. *Tourism Management, 67*,17-31.

Balow, C. (2017). The "effect size" in educational research: What is it and how to use it? [sic] *Illuminate Education*. Retrieved from

www.illuminateed.com/blog/2017/06/effect-size-educational-research-use/ on July 14, 2019.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44*, 533–553.

Benjamin, D. J., Berger, J. O., Johannesson, M. et al. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10.doi: 10.1038/s41562-017-0189-z.

Bollen, K. A., Harden, J. J., Ray, S., and Zavisca, J. (2014). BIC and Alternative Bayesian Information Criteria in the Selection of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 21*,1–19.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences* (1st ed.). New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). Power Primer. *Psychological Bulletin* [PsycARTICLES]; 112, 1; PsycARTICLES, pp. 155-159.

De Ruiter, J. (2019). Redefine or justify? Comments on the alpha debate. *Psychonomic Bulletin & Review, 26*, 430–433, https://doi.org/10.3758/s13423-018-1523-9.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd.

Goodman, S. (2005). Introduction to Bayesian methods I: Measuring the strength of evidence. *Clinical Trials*, *2* (4), 282-290.

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology, 45* (3), 135-140.

Hoijtink, H., van Kooten, P., and Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the Bayes Factor. *Multivariate Behavioral Research*, *51*, 2-10. doi: 10.1080/0027317–1.2014.969364

Holmberg, A. (2024). Toward a Better Understanding of Statistical Significance and p Values in Nursing. *Nursing Forum*, Article ID 7263781, 1-11 pageshttps://doi.org/10.1155/2024/7263781

Huisman, L. (2023). Are P-values and Bayes factors valid measures of evidential strength? *Psychonomic Bulletin & Review, 30*, 932–941. https://doi.org/10.3758/s13423-022-02205-x

Jarosz, A., and Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving, 7*, 2-9. doi: 10.7771/1932-6246.1167

Jeffreys H. (1961). *Theory of Probability* (3rd ed.). New York: Oxford University Press.

Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., ... Bradford, D. (2018). Justify your alpha. *Nature Human Behavior, 2,* 168-171.

Lavine, M., and Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician, 53* (2), 119-122.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192.

McGrath, R. E., and Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. *Psychological Methods*, *11*(4), 386-401. doi: 10.1037/1082-989X.11.4.386.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, 73 (1), 235-245.

Morey, R. D., Romeijn, J.W., and Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18.

Opić, S, and Rijavec, M. (2022). Misconceptions of the p-value - let us use new approaches and procedures // 2. Međunarodna znanstvena i umjetnička konferencija Suvremene teme u odgoju i obrazovanju - STOO 2 In memoriam Prof. Emer. Dr sc. Milan Matijević u suradnji sa Zavodom za znanstvenoistraživački i umjetnički rad Hrvatske akademije znanosti i umjetnost / D. Velički and M. Dumančić, (eds.). Zagreb: Sveučilište u Zagrebu Učiteljski fakultet, 2022. pp. 1-21.

Pearson, K. X. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. London, Edinburgh and Dublin *Philosophical Magazine and Journal of Science*, *50* (302), 157–175, https://doi.org/10.1080/14786440009463897

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing, *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.00223-

Rovetta, A. (2024). Abandon Statistical Significance: A Gentle Introduction to S-values and S-intervals. https://doi.org/10.31219/osf.io/ywhu9 (https://osf.io/preprints/osf/ywhu9)

Ruscio, J. (2008). A Probability-Based Measure of Effect Size: Robustness to Base Rates and Other Factors. *Psychological Methods, 13*(1), 19 –30.

Sullivan, G. M., and Feinn, R. (2012). Using effect size—Or why the p-value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Stern, H. S. (2016). A test by any other name: P-values, Bayes Factors, and statistical inference. *Multivariate Behaviour Research*, *51*(1), 23-39. doi:10.1080/00273171.2015.1099032

Vacha-Haase, T., and Thomson, B. (2004). How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology, 51*(4), 473–481. https://doi.org/10.1037/0022-0167.51.4.473

Wasserstein, R. L., and Lazar, N. A. (2016). ASA Statement on Statistical Significance and P-Values - Context, Process, and Purpose. *The American Statistician*, *70* (2), 129-133.

Yuan, Y., Johnson, V. E. (2008). Bayesian Hypothesis Tests using Nonparametric Statistics. *Statistica Sinica, 18* (3), 1185-1200.

Original quote: an erratum (page 110);

Similarly, de Ruiter (2019), while critiquing the proposal to lower the significance level to 0.005, argues that setting an alpha level of $p \leq 0.005$ does not enhance replicability. He believes that the rationale for adopting a new alpha level of 0.005 is weak and that such a change could potentially harm scientific practice

Author

Siniša Opić, PhD

Full Professor, Universtiy of Zagreb, Faculty of Teacher Education, Savska 77, 10000 Zagreb, Croatia, e-mail: sinisa.opic@ufzg.hr

Redni profesor, Univerza v Zagrebu, Pedagoška fakulteta, Savska 77, 1000 Zagreb, Hrvaška, e-pošta: sinisa.opic@ufzg.hr