

# Multiple Confidence Intervals and Surprisal Intervals to Avoid Significance Fallacy

Alessandro Rovetta<sup>1, 2</sup>

Review began 12/31/2023

Review ended 01/05/2024

Published 01/09/2024

© Copyright 2024

Rovetta. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. Research and Disclosure Division, R&C Research, Bovezzo (BS), ITA 2. Technological and Scientific Research, Redeev Srl, Naples, ITA

**Corresponding author:** Alessandro Rovetta, rovetta.mresearch@gmail.com

## Abstract

Overconfidence in statistical results in medicine is fueled by improper practices and historical biases afflicting the concept of statistical significance. In particular, the dichotomization of significance (i.e., significant vs. not significant), blending of Fisherian and Neyman-Pearson approaches, magnitude and nullification fallacies, and other fundamental misunderstandings distort the purpose of statistical investigations entirely, impacting their ability to inform public health decisions or other fields of science in general. For these reasons, the international statistical community has attempted to propose various alternatives or different interpretative modes. However, as of today, such misuses still prevail. In this regard, the present paper discusses the use of multiple confidence (or, more aptly, compatibility) intervals to address these issues at their core. Additionally, an extension of the concept of confidence interval, called surprisal interval (S-interval), is proposed in the realm of statistical surprisal. The aforementioned is based on comparing the statistical surprise to an easily interpretable phenomenon, such as obtaining S consecutive heads when flipping a fair coin. This allows for a complete departure from the notions of statistical significance and confidence, which carry with them longstanding misconceptions.

**Categories:** Public Health, Epidemiology/Public Health, Medical Education

**Keywords:** surprisal, confidence intervals, public health, testing, significance

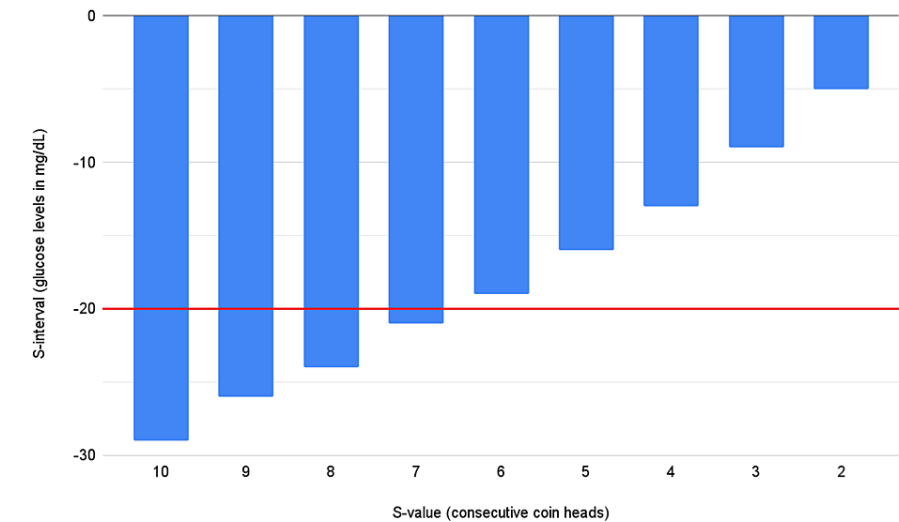
## Editorial

For decades, the scientific community has been plagued by dangerous misuses and abuses of statistical significance [1]. Especially in the field of public health, these errors can lead to severe consequences, such as the approval or endorsement by certain members of the scientific community for practices that have weak supporting evidence. In a time when information overload is a global issue, even within the mere research field, it is imperative that scientific standards become even more rigorous. However, the historical efforts of leading figures and organizations in the international statistical field have not been sufficient to address such longstanding problems to date, although they have been essential in limiting them. Specifically, the tendency to classify results as non-significant (e.g.,  $P \geq .05$ ) and significant (e.g.,  $P < .05$ ) is mathematically incorrect and absurd, given that i) it improperly blends two mathematically incompatible approaches, namely Fisher's (which allows for the assessment of individual studies, most useful in the medical field) and Neyman-Pearson's (which assesses groups of studies and never individual ones), ii) as demonstrated by McShane et al., two results with very distant P-values (i.e.,  $P = .005$  and  $P = .194$ ) can be extremely compatible with each other according to a certain test (i.e.,  $P = .289$ ), iii) statistical testing cannot prove or disprove a specific target hypothesis but can, at most, show the compatibility of such a hypothesis with experimental data through a specific statistical test, iv) statistical testing is mathematically disconnected from the scientific phenomenon, as it assumes that there is no other phenomenon outside chance (it is up to the researcher to interpret the numerical outcome considering the whole scientific scenario) [2, 3]. In addition, the magnitude fallacy (i.e., the inability to distinguish between statistical and practical relevance) and nullification fallacy (i.e., an obsession with the null hypothesis over other relevant hypotheses) consistently distort the vast majority of scientific analyses, as they have roots tied to the teaching of statistics in universities [1, 4]. In order to address these issues, this paper proposes an enrichment and expansion of some methodologies recently discussed by the statistical community. The first approach concerns traditional confidence intervals, which, in accordance with Greenland et al., should be termed compatibility intervals. Indeed, if and only if all assumptions of the statistical test have been properly verified, the P-value is an approximate index of the compatibility of experimental data with the target hypothesis according to the statistical test. P-values close to 1 indicate high compatibility, while P-values close to 0 indicate low compatibility. Thus, for example, assuming that all experimental and statistical procedures have been executed correctly, a 95% confidence interval of the form 95% CI = (a, b) states that all predictions of the target hypothesis between values a and b have approximately a statistical compatibility level with the experimental results greater than  $P = .05$  (i.e., they all have  $P > .05$ ). Let's consider a practical example: suppose that, after a certain treatment to reduce blood glucose levels, the test group recorded a mean of 120 mg/dL (SD 15 mg/dL) compared to a mean of 140 mg/dL (SD 20 mg/dL) in the placebo group. Both groups were of 10 patients. After verifying that our data were sufficiently compatible with the assumptions of the test (a vital aspect to which the researcher must give utmost importance, although, in this case, it goes beyond the scope of this paper), we apply a one-tailed Welch t-test to assess the situation statistically.

### How to cite this article

Rovetta A (January 09, 2024) Multiple Confidence Intervals and Surprisal Intervals to Avoid Significance Fallacy. Cureus 16(1): e51964. DOI 10.7759/cureus.51964

Nonetheless, instead of solely assessing the null hypothesis, we evaluate degrees of compatibility that interest us. For instance, we calculate the following compatibility intervals: 99%, 95%, and 90% (that is, we seek all predictions that, in the aforementioned ways, result in P-values greater than .01, .03, and .05, respectively). To present these results, we adopt a new notation that is very convenient to read, namely  $99|97|95\% \text{ CI} = (-20, +\infty|-16, +\infty|-14, +\infty)$ . This notation provides various pieces of information: e.g., if the experimental difference between the two groups mean values falls between -16 and -14, the P-value would be between .03 and .05. The observed experimental difference between the two groups (120-140 mg/dL = -20 mg/dL) has a compatibility of above .01. Additionally, we can observe how compatibility changes based on various hypothetical differences. In this case, since the P-value varies substantially even for small changes in the outcome (e.g., the clinical difference between -20 and -16 mg/dL or between -16 and -14 mg/dL is negligible and could be caused by daily random oscillations), these results are uncertain and require further investigation. Nevertheless, even if we were to be satisfied at a statistical level, could we conclude the effectiveness of the treatment? No. In fact, we have not found evidence supporting the treatment's effectiveness but rather evidence compatible with it. There could be other equally plausible hypotheses for this result (e.g., an error in the random sampling procedure, a statistical coincidence or anomaly, etc.). Therefore, not only would further statistical confirmation be necessary, but there should be reasons of a different nature (e.g., clinical, biochemical, etc.) supporting the hypothesis of effectiveness. Furthermore, we have not assessed the cost-risk-benefit ratios, which are essential for making a public health decision (e.g., is the therapy invasive? Do the benefits justify the treatment?). However, compatibility intervals can create confusion since the amount of information in different P-value intervals is not constant (e.g., the difference between  $P_1=.01$  and  $P_2=.05$  contains much more information than the difference between  $P_3=.95$  and  $P_4=.99$  even if both  $\Delta P=.04$ ). This can pose problems with the interpretation of multiple compatibility intervals. In this regard, we introduce the concept of S-value [5]. Suppose we calculate the probability of obtaining S consecutive successes (heads) by flipping a fair (unbiased) coin. Since the probability of success in a single toss is  $P=1/2$ , for S flips, we have  $P=(1/2)^S$ . It follows that  $S=-\log_2 P$  (i.e., the base-2 logarithm of P). The main advantage is the comparison of the surprise of a statistical event to an everyday event that we have a very intuitive perception of. This way, it is easy to observe that  $P_1=.01$  implies  $S_1=6.6$ ,  $P_2=.05$  implies  $S_2=4.3$ ,  $P_3=.95$  implies  $S_3=0.07$ , and  $P_4=.99$  implies  $S_4=0.01$ . Thus, the difference between  $P_2$  and  $P_1$  is  $\Delta S_{1,2}=6.6-4.3=2.3$ , while the difference between  $P_4$  and  $P_3$  is  $\Delta S_{3,4}=0.07-0.01=0.06$ . The purpose and role of S-values are also to model the implicit uncertainty in each statistical analysis. In fact, obtaining exactly 6.6 consecutive heads is not possible, which is why, in decision logic, it is necessary to consider such a result as "a little more than 6 consecutive heads" or "a little less than 7 consecutive heads" or "between 6 and 7 consecutive heads." The relationship between S-values and compatibility intervals is not entirely clear, as it still has to go through the P-value. For this reason, the concept of S-interval has been introduced. In particular, assuming that all background assumptions have been verified, the S-interval contains all and only the target assumption predictions that are approximately less surprising than S consecutive heads - by flipping a fair coin - compared to the experimental result according to the statistical test. For example, a 4-I, under the conditions mentioned above, contains all the target assumption predictions less surprising than 4 consecutive heads. Calculating an S-I is straightforward; for instance,  $S=4$  implies  $P=(0.5)^4=.063$ , i.e., a compatibility interval of  $100(1-P)\%=100(1-0.063)\%=93.7\%$ . Resuming our example, we show the following S-intervals:  $7|5|4\text{-I} = (-21, +\infty|-16, +\infty|-13, +\infty)$ . We can thus see that, to transition from a surprise level of 5 consecutive to 4 consecutive, only a variation of 3 mg/dL is required (fully compatible with a daily random fluctuation). The variation of the effect size as a function of statistical surprise (and vice versa) can also be graphically represented in order to be better evaluated (Figure 1). In general, it is observed that clinically small changes in effect size correspond to large changes in statistical surprise.



**FIGURE 1: S-intervals (blood glucose levels in mg/dL) as a function of the respective S-values (statistical surprise compared to the number of consecutive heads when flipping a fair coin) according to the one-tailed Welch t-test.**

Red line = observed experimental result.

Hoping that these examples have clarified the potentials and limitations of the statistical approach in public health, the author of this manuscript invites the scientific community to consider such approaches for conducting statistically informative investigations at a scientific level. The methodologies presented here draw on previous literature, although the concepts of S-intervals and the proposed convention for reporting multiple intervals are, to the best of the author's knowledge, new.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

**Concept and design:** Alessandro Rovetta

**Acquisition, analysis, or interpretation of data:** Alessandro Rovetta

**Drafting of the manuscript:** Alessandro Rovetta

**Critical review of the manuscript for important intellectual content:** Alessandro Rovetta

**Supervision:** Alessandro Rovetta

Disclosures

**Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Wasserstein RL, Lazar NA: The ASA's statement on p-values: context, process, and purpose . Am Stat. 2016, 70:129-133. [10.1080/00031305.2016.1154108](https://doi.org/10.1080/00031305.2016.1154108)

2. McShane BB, Bradlow ET, Lynch JG, Meyer RJ: EXPRESS: "Statistical Significance" and statistical reporting: moving beyond binary. J Mark. 2023, 0(ja): [10.1177/00222429231216910](https://doi.org/10.1177/00222429231216910)

3. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG: Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016, 31:337-350. [10.1007/s10654-016-0149-3](https://doi.org/10.1007/s10654-016-0149-3)
4. Kühberger A, Fritz A, Lerner E, Scherndl T: The significance fallacy in inferential statistics . *BMC Res Notes.* 2015, 8:84. [10.1186/s13104-015-1020-4](https://doi.org/10.1186/s13104-015-1020-4)
5. Greenland S, Mansournia MA, Joffe M: To curb research misreporting, replace significance and confidence by compatibility: a preventive medicine golden jubilee article. *Prev Med.* 2022, 164:107127. [10.1016/j.ypmed.2022.107127](https://doi.org/10.1016/j.ypmed.2022.107127)