


S-values and Surprisal Intervals to Replace P-values and Confidence Intervals

Authors: ALESSANDRO ROVETTA  
– Technological and Scientific Research, Redeev SRL,
Naples, Italy
alessandrorovetta@redevv.com

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- Misuse of statistical significance continues to be prevalent in science. The absence of intuitive explanations of this concept often leads researchers to incorrect conclusions. For this reason, some statisticians suggest adopting S-values (surprisals) instead of P-values, as they relate the statistical relevance of an event to the number of consecutive heads when flipping an unbiased coin. This paper introduces the concept of surprisal intervals (S-intervals) as extensions of confidence/compatibility intervals. The proposed approach imposes the assessment of outcomes in terms of more and less surprising than some values, instead of statistically significant and statistically non-significant. Moreover, a novel methodology for presenting multiple consecutive S-intervals (or compatibility intervals as well) in order to evaluate the variation in surprise (or compatibility) with various target hypotheses is discussed.

Keywords:

- *confidence intervals; epidemiology; hypothesis testing; public health; significance; surprisal.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

The proper interpretation of P-values in science has been debated for decades [1, 6]. Widespread misinterpretation of this measure has even led some academic journals to abandon its use [13]. However, Greenland et al. emphasize that P-values can still provide valid information for making sound scientific decisions if used as a measure of statistical compatibility instead of statistical significance [10]. In this regard, there are some considerations to be made. Let's suppose we choose a specific statistical test and set a certain target assumption (e.g., also called "target hypothesis"). Every statistical test is mathematically built on the condition that all assumptions, including the target, are true. Then, in a Fisherian sense, the P-value measures the degree of compatibility of the statistical result (the test statistic) with the target and all the background assumptions (e.g., linearity, normality, properly functioning measurement devices). P-values close to 1 indicate high compatibility, while P-values close to 0 indicate low compatibility. Thus, although we may be interested solely in the target hypothesis, it is important to understand that the P-value does not privilege said hypothesis over any other. Indeed, violating the background assumptions can strongly influence P-values, making them uninformative for the fixed scientific goal. Moreover, the reliability of the statistical outcome depends on the scientist's ability to conduct the whole experimental procedure (which cannot be carried out without uncertainties). This means selecting a model capable of providing useful information to analyze the scientific phenomenon (which includes choosing proper data collection methods, estimators or parameters, and hypotheses) as well as guaranteeing human attributes like competence, honesty, transparency, and collaboration [8]. Thus, in light of the interpretative uncertainties that P-values entail, the practice of sharply distinguishing arbitrarily close values (e.g., $P = 0.049$ and $P = 0.051$) is meaningless. According to this, from now on, we will refer to the condition "all background assumptions are true" using the expression "utopian scenario" (emphasizing the practical impossibility of achieving it). Even in the utopian scenario, the P-value is mathematically precluded from providing information about the investigated scientific phenomenon: at best, it can be understood as the probability that chance alone would produce a discrepancy from the target hypothesis prediction as or more extreme than that obtained in our experiment according to the performed test.¹ The key point is that the model assumes that chance is the sole factor at play. In other words, under the target null assumption of zero effect, the statistical model mathematically excludes the occurrence of any scientific phenomenon other than chance (e.g., if our objective is to investigate a drug's effectiveness, under the target null assumption of zero effect, the statistical model we implement mathematically excludes the existence of any pharmacological effect). Indeed, a statistical model takes numbers and yields numbers; it is up to the scientist to interpret these

¹The phrase "chance alone produces" is incomplete as it does not encompass fields of science where randomness is the absence of any cause. Nevertheless, this expression has been chosen because it was considered clearer and suitable for the context of public health.

based on the research context. For this reason, it never makes sense to state that the P-value is the probability that chance produced or would produce the observed scientific effect. Even in the utopian scenario, the P-value never allows the researcher to reject a target hypothesis (since absence of evidence is not evidence of absence) or to confirm it (since even a P-value of .99 does not exclude the presence of many other models with equal compatibility). In this regard, P-values are not absolute measures of compatibility, as the data consistency with a certain hypothesis could change drastically depending on the adopted test (e.g., the data may be highly consistent with the normality hypothesis via Shapiro-Wilk, but not Kolmogorov-Smirnov). Alongside this, degrees of compatibility that appear markedly different could be highly compatible with each other. As shown by McShane et al., an original study with $P = 0.005$ and a replication study with $P = 0.194$ were highly compatible with one another in the sense that the P-value of the chosen comparison test, assuming no difference between them, was $P = 0.289$ [15]. Therefore, the difference between “statistically significant” and “statistically not significant” would be “statistically not significant” at the 0.05 level [5]. Nonetheless, the absurdity of adopting dichotomous thresholds goes beyond this example, as it blends two incompatible approaches: the (neo) Fisherian one, as described above, and the decision-theoretic Neyman-Pearson one. The first is mathematically structured to provide information on individual studies under the conditions mentioned above, while the second is mathematically structured to provide information on groups of studies (but never on individuals within that group) in numerous repetitions under the same scientific conditions (utopian scenario). This even leads to two distinct mathematical definitions of the P-value, which the reader can delve into by consulting other literature [8, 9]. Given that the overall goal of public health statistics is to inform decisions based on individual studies (e.g., randomized control trials, systematic reviews with meta-analysis, etc.), the (neo) Fisherian approach must be preferred. Nevertheless, in addition to what has already been discussed, there are further inherent difficulties in the use of the P-value that could be addressed by adopting some valid alternatives.

2. SURPRISAL AS AN ALTERNATIVE TO STATISTICAL SIGNIFICANCE AND COMPATIBILITY

2.1. Relationship between P-values and S-values

P-values exhibit some counterintuitive behaviors. For instance, even though the pairs $(P_1 = 0.05, P_2 = 0.10)$ and $(P_3 = 0.90, P_4 = 0.95)$ are formed by P-values that differ by the same amount, $\Delta P = 0.05$, the information contained in the regions identified by these two pairs differs substantially. This happens because the area of the corresponding curve is geometrically distributed differently along the bell curve. For this reason, the use of Shannon information (also

known as “surprisal” or “S-value”) has been proposed based on the following reasoning: given the probability P of an event, this can be related to the probability of obtaining S consecutive heads by flipping an unbiased coin using the formula $P = 0.5^S = 2^{-S}$ [17]. It follows that $S = -\log_2 P$. In the utopian scenario, the S-value measures the degree of surprise of the test result (e.g., the “t” statistic or the chi-squared statistic) compared to the target assumption. The aim is to compare “statistical significance” with a phenomenon that we are familiar with in everyday life. However, mathematically speaking, the S-value measures continuous information (bits). It is up to the reader to interpret that information in relation to the context. Values such as $S = 4.3$ cannot be understood as “4.3 consecutive heads”; however, this writing can be interpreted as “in the utopian scenario, the statistical result is approximately as surprising as 4 consecutive heads - or slightly more than 4 consecutive heads - when tossing a fair coin.” At the conventional threshold $P = .05$, $S = 4.3$ bits correspond. Thus, when we evaluate the difference between $P_1 = .05$ and $P_2 = .10$, we obtain $\Delta S = |S_2 - S_1| = \log_2(0.10/0.05) = 1$ bit, while between $P_3 = .90$ and $P_4 = .95$, we obtain $\Delta S = |S_3 - S_4| = \log_2(0.95/0.90) = 0.08$ bits. Hence, the difference in statistical surprise now becomes evident. However, the philosophy underlying the S-value goes beyond this simplification: the goal is to evaluate results in classes of practical equivalence. Considering the uncertainties mentioned above, there is no practical difference between $P = .05$ ($S = 4.3$) and $P = .0625$ ($S = 4$), since both results are surprising by about as much as 4 consecutive heads. This is why it is good practice to round S values to the nearest integer (although more precise values should always be reported as supplementary material to allow for multi-comparison adjustments or meta-analyses).

2.2. S-values don’t address the magnitude fallacy

Surprisals can be effective in properly evaluating statistical surprise, but they cannot address the common confusion about the difference between statistical surprise or compatibility and magnitude [11, 12]. Therefore, this paragraph will address the relationship between statistical compatibility and effect size, allowing for a proper introduction of the relationship between surprisal and effect size. A statistical phenomenon can be rare and unexpected (high surprise) but weak (low magnitude), meaning it may have little practical impact. For example, while following a weight-loss diet in accordance with the health recommendations of their primary care physician, one may consistently lose about one gram per day for 100 days (a scientific effect that is unlikely to be due to chance) but still be far from their target weight (indeed, losing 100 grams in 100 days has a negligible impact on physical health). If we were to statistically model such a real-world situation using linear regression, we would obtain a very low P-value (indicating a surprising result), but also a very low slope coefficient (indicating that the trend’s intensity would be low compared to the predetermined objective) [18]. Nevertheless, the P-value is inherently linked to the concept of effect

size (ES), as it can be expressed, at least, as a function $P = f(ES, N)$ where N is the sample size. Considering a fixed $N = N_0$, the P-value could be exclusively linked to the effect size. The latter can be examined through the effect size parameter (which can provide information on the intensity of the statistical phenomenon) but also through the width of confidence intervals (which can provide information on how the statistical effect size changes in relation to the P-value). However, the concept of confidence is commonly (mis)understood within an inferential fashion, i.e., it does not probabilistically concern our single already completed experiment. Indeed, a confidence interval can be obtained by selecting an arbitrary threshold α and performing the operation $100 \cdot (1 - \alpha)\%$; the most well-known case is $\alpha = .05$ with a 95% confidence interval. When testing continuous data, by calculating 95% confidence intervals in infinite utopian applications, 95% of these intervals contain the true value (coverage probability) [2]. However, even assuming that a sufficiently large number of repetitions of the experiment is enough and assuming to work in the utopian scenario in each of these, such an approach cannot mathematically tell us which intervals contain the true value. Furthermore, the above definition of confidence interval conflicts with the abandonment of statistical significance thresholds. To solve these dilemmas, Rafi et al. propose a terminological modification: give up the term “confidence” in favor of the Fisherian term “compatibility” [17]. In this framework, considering the utopian scenario, a compatibility interval contains all the target assumption predictions that, compared to certain threshold hypotheses and according to the performed test, are more compatible with the calculated experimental result (e.g., the difference between two sample estimators or population parameters). In other words, any model prediction that lies inside (resp. outside) the obtained compatibility interval will result in a P-value higher (resp. lower) than the selected threshold. In order to address the problem of the arbitrary threshold choice, it has been proposed to provide tables that relate various P-values to their respective compatibility intervals or to present multiple compatibility intervals (e.g., 50%, 75%, 90%). A particularly interesting and information-rich solution is to graphically represent all compatibility intervals from 1% to 99% [17]. However, this may greatly increase the reading load or even be confusing in the case of multiple results. Besides, there is currently a clear asymmetry in the definition and application of the concepts of compatibility interval and statistical surprise (S-value) since the former remains confined by definition within the scope of statistical significance (P-value). For these reasons, the present manuscript proposes and discusses two points: 1) the concept of surprisal interval, which can address the issues related to the obscured relationship between compatibility interval and surprise, and 2) a novel convention to compress information on the relationship between P/S-value and compatibility/surprisal intervals (based on the work of Xie et al. concerning confidence distribution [21]) and allows the presentation of these results in a single compact form.

3. SURPRISAL-BASED APPROACH

3.1. Surprisal interval

The definition of surprisal interval is based on a specific partition of the probability density function. Specifically, it consists of associating the natural values $S = 1, 2, \dots, n$ with their respective areas by exploiting the relationship $P = 2^{-S}$. The first ten results are shown in Table 1. Through this operation, it is now possible to define surprisal intervals (S-I), analogous to confidence intervals (CI). We consider the case of a normal distribution (Figure 1). Let's assume we want to find a 4-I ($S = 4$). The corresponding exact P-value is $P = 2^{-4} = 0.0625 \sim 0.063$. Therefore, the corresponding compatibility interval is $(1 - 0.063) \cdot 100\% \text{ CI} = 93.7\% \text{ CI}$. To calculate it in practice, we need to ask ourselves: what is the value of z for which the area under the Gaussian curve between $-z$ and z is equal to 0.937 (i.e., 93.7% of the total unit area)? The answer is reported in Table 1. Afterward, it is sufficient to calculate $93.7\% \text{ CI} = (r - z \cdot \bar{\sigma}, r + z \cdot \bar{\sigma}) = (r - 1.86 \cdot \bar{\sigma}, r + 1.86 \cdot \bar{\sigma})$, where r is the calculated experimental result and $\bar{\sigma}$ is the standard error.

<i>S</i> -value	<i>P</i> -value	<i>z</i> -value***	$100 \cdot (1 - P)\% \text{ CI}$
1	0.500	0.67	50%
2	0.250	1.15	75%
3	0.125	1.53	87.5%
4	0.063	1.86	93.7%
5	0.031	2.16	96.9%
6	0.016	2.42	98.4%
7	0.008	2.66	99.2%
8	0.004	2.89	99.6%
9	0.002	3.10	99.8%
10	0.001	3.30	99.9%

Table 1: Association between surprisal (S-I) and compatibility intervals (CI). *** the shown *z*-values are valid only for the Gaussian distribution; conversely, the relationships between *S*-values, *P*-values, and compatibility intervals are general.

The interpretation of our 4-I is as follows: in the utopian scenario, all target assumption predictions that lie inside (resp. outside) the 4-interval are less (resp. more) surprising than getting 4 consecutive heads - when flipping a fair coin - compared to the calculated experimental result according to the statistical test. In other words, let's suppose we choose a specific statistical test and consider a target assumption predicting an effect h . Let's also suppose we calculate an experimental result r (e.g., the difference between two population mean values)

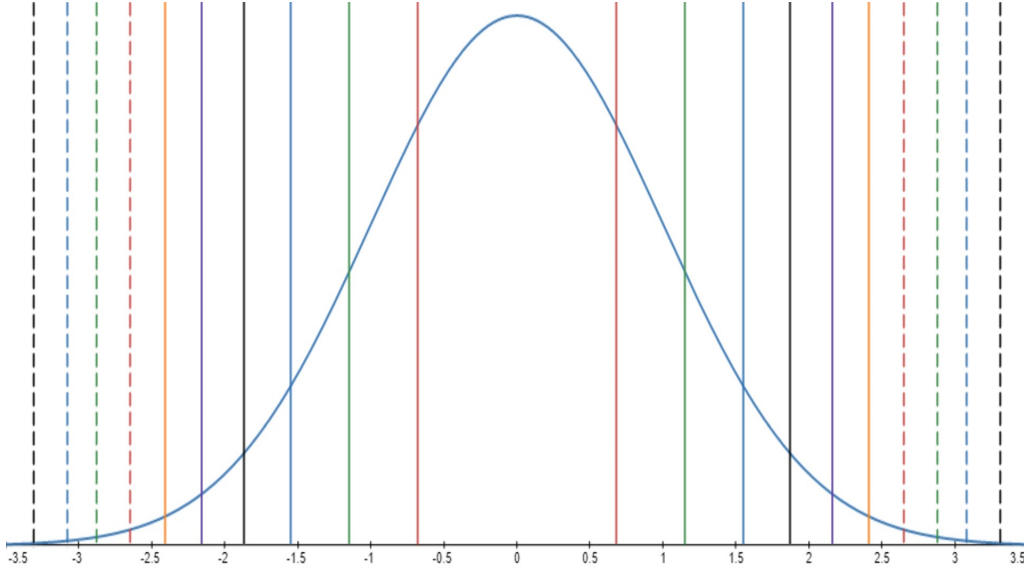


Figure 1: Surprisal intervals for integer values of S from 1 to 10.

in the utopian scenario. The 4-I contains all and only the values h such that $r - h$ (i.e., the difference between r and h) is less surprising, according to the chosen test, than 4 consecutive heads when tossing a fair coin. Therefore, a general definition of surprisal interval is as follows:

Definition 3.1. If and only if all the background assumptions are true, a surprisal interval (or S-interval) is the interval that contains all and only the target assumption predictions that are less surprising than S consecutive heads - when tossing a fair coin - compared to the calculated experimental result according to the statistical test.

Let's apply this new definition to evaluate a two-tailed one-sample t-test for a sample mean value of $\bar{x} = 10$, with a standard error $\bar{\sigma} = 5$ and a population mean value $\mu = 0$ (such that $r = \bar{x} - \mu = 10 - 0 = 10$). For simplicity, we also assume that the degrees of freedom are greater than 30 (such that $t \sim z$) and all the background assumptions are sufficiently met. Let's then calculate the following S-intervals: 4-I, 5-I, and 6-I. According to Table 1, $S = 4$ implies $t = 1.86$ (this happens because, in this specific example, $t \sim z$). So, we have 4-I = $(r - t \cdot \bar{\sigma}, r + t \cdot \bar{\sigma}) = (10 - 1.86 \cdot 5, 10 + 1.86 \cdot 5) = (1, 19)$. Similarly, we obtain 5-I = $(-1, 21)$ and 6-I = $(-2, 22)$. It is now easy to observe that the difference between $r = 10$ and the null hypothesis prediction ($h = 0$) is more surprising than 4 consecutive heads and less surprising than 5. In fact, the 4-I has a lower bound equal to $h = 1$ while the 5-I has a lower bound equal to $h = -1$ (hence, $h = 0$ must be somewhere in the middle). Considering the 4-I, we can also observe that $r - h = 10 - h$ is less surprising than 4 consecutive heads for all $h \in (1, 19)$ and

more surprising than 4 consecutive heads for all $h < 1$ or $h > 19$.² To provide a compact overview of the variation in the width of our S -intervals as a function of the corresponding S -values, we can adopt the notation $5|4\text{-I} = (-1, 21|1, 19)$, which can be easily extended to three or more surprisal intervals depending on the needs of the authors and stakeholders (e.g., $6|5|4\text{-I} = (-2, 22| -1, 21|1, 19)$). Thus, in general, we can define the convention as follows:

Definition 3.2. An n -tuple of surprisal intervals can be expressed as $S_1|\dots|S_n\text{-I} = (S_1\text{-I}|\dots|S_n\text{-I})$.

Each S_i represents a specific S -value and each $S_i\text{-I}$ represents the corresponding surprisal interval. As an additional convention, it could be suggested to report at least three surprisal intervals: the narrowest S -interval containing the prediction of the null hypothesis, the narrowest S -interval not containing the prediction of the null hypothesis, and the 4-interval. The first two serve to locate the null hypothesis prediction, while the third serves as a general reference for comparison with other surprisal intervals (as it covers about 94% of the area, similarly to the classic 95% CI). In the previous example, the binary formulation $5|4\text{-I} = (-1, 21|1, 19)$ was sufficient.

3.2. Practical advantages of surprisal intervals

In general, the concept of surprisal as a measure of statistical surprise is absolutely unnecessary from a purely mathematical, statistical, or computational point of view. As a matter of fact, a correct use of P-value and compatibility intervals could compensate for any criticality exposed in the introductory section (although this would be much longer, subtle, and uneasy to present). However, the world of hard sciences is forced to confront a very different reality linked to the psychology and perception of the scientists who adopt and develop them. For instance, Rafi et al. argue that the misuse of statistical significance is primarily a cognitive and semantic problem rather than a statistical issue [2]. Greenland et al. also suggest that the inability to find a straightforward interpretation of the concept of P-value paradoxically favors the proliferation of oversimplified explanations [11]. The same author of this paper has found, during his experience as an editor and peer reviewer in public health-related topics, not only widespread poor knowledge about the difference between the Fisherian and Neyman-Pearson approaches but also a furious resistance to change despite the overwhelming evidence provided. The authors' motivations ranged from "We don't want to make reading complicated" to "We prefer to maintain the traditional use of significance". This scenario is strongly consistent with the concerns raised by internationally renowned statisticians as well as the official statements of the American Statistical Association [22]. In 2014, Professor George Cobb openly denounced

²It must be clear that this scenario is valid for the chosen test; performing a different test could substantially alter these outcomes.

the pivotal role of academic journals and universities in the unwarranted success of $P = 0.05$. Such illogical behaviors are compatible with some phenomena of cognitive psychology whereby modifying a consolidated belief or behavior is highly complex and often temporary when successful (even in scientists) [19, 20]. As early as 1919, Boring emphasized the limitations of the mathematical approach in modeling scientific reality and stressed the impossibility of formulating conclusions based solely on statistical approaches [3]. The fact that, over 100 years later, despite the knowledge accumulated over this period, such concepts elude a significant fraction of the scientific community is signaling more than a formative problem. Accordingly, McShane and Gal observed that dichotomization decreases (but does not eliminate) when researchers are prompted to make decisions based on the evidence, especially if the outcome has personal consequences [16]. Still, that's not all: diabolical academic dynamics such as publish or perish and publication bias push authors to exploit fallacious interpretations of P-values and compatibility intervals to voluntarily exaggerate the apparent degree of evidence found in their studies, thus increasing their chances of being published and cited [4]. Based on this, the purely interpretative aspect of a statistical measure can have very important practical consequences, especially in sectors - such as public health - where errors and overstatements must be weighed on the cost function for stakeholders. In particular, the main objective of this approach is to complete the proposal for replacing P-values with S-values by also requiring the replacement of "confidence" intervals with surprisal intervals. The total abandonment of statistical significance also brings with it the abandonment of all incorrect practices related to erroneous familiarity (e.g., judging a result as non-significant when $P < \alpha$ or is included in the $100(1 - \alpha)\%$ CI, or considering $\alpha = 0.05$ and 95% CIs as some sort of privileged options) and prevents such dichotomies and prejudices at the root. The same term "significance" is inevitably and intrinsically replaced with the term "surprise", thus avoiding unnecessary, dangerous, as well as frequent confusion with practical significance (effect size) or even clinical significance. In order to give the reader an idea of the proportion of these errors in the medical field, a previous study found that only one out of 52 students was able to properly distinguish these concepts [12].

In addition to this, surprisal intervals make the relationship with the measure of surprise of the outcome much clearer and more direct than compatibility intervals do with P-values. First, in addition to terminological consistency, the presentation of results is based on the same statistical quantity, namely, the integer number of consecutive heads when tossing an unbiased coin (bits) rather than a decimal measure of statistical compatibility and a percentage area. Second, the relationship between different intervals is much more intuitive since S-values linearize the behavior of the distribution. For example, an inexperienced user, as often happens, may easily think that the 99|95|91-%CI situation is symmetrical with respect to the central interval when, in fact, this includes very different compatibility requests (since they correspond respectively to 7, 4, and 3 consecutive heads); this cannot happen if the 7|4|3-I notation is used. Third, instead of setting arbitrary thresholds, the user can decide which intervals to show without

exceeding confidence in the result. For example, if $S = 8$ is obtained, it may be useful to show an 8-I in order to understand what is the range of least surprise associated. This solution is highly advantageous because it allows the presentation of the result surprisal and the associated surprisal interval in a single compact form. This also allows for simplifying, both conceptually and operationally, procedures such as adjustment for multicomparison since S-values and S-intervals are no longer separable. Furthermore, while compatibility intervals consent to choose very specific degrees of precision (e.g., 94% or 95% or 96%), in the case of S-intervals the degree of precision is forced to be 1 bit. Ergo, the user is led to evaluate the results in less clear-cut terms (e.g., about 8 consecutive heads).

4. CONCRETE APPLICATION EXAMPLES

4.1. Example 1

The design of this study is intentionally non-optimal for didactic purposes. The aim is to show a proper application of the S-interval concept as well as the potential and limitations of the statistical approach. Let's suppose we have developed a long-term treatment to reduce blood pressure in hypertensive individuals. We convince 10 patients with clinically similar conditions to adopt this treatment for 3 months. We measure blood pressure levels before and after the treatment, obtaining the data in Table 2.

Patient	Before (mmHg)	After (mmHg)	Differences (mmHg)
1	140	132	-8
2	150	145	-5
3	130	128	-2
4	135	139	+4
5	145	140	-5
6	138	132	-6
7	142	143	+1
8	128	125	-3
9	152	148	-4
10	134	130	-4

Table 2: Hypothetical blood pressure data before and after treatment: case 1.

Since we are searching for a reduction, we decide to apply a one-sided one-sample t-test. To do so, in addition to assuming that all experimental procedures have been executed correctly (including random sampling), we need to check the compatibility of the data with the following assumptions of the test: i) normal distribution (including the absence of outliers), ii) independence of observations,

iii) interval or ratio data. Since the data (column “Differences”) represents continuous real number values of blood pressure from independent patients, we can reasonably consider assumptions ii) and iii) to be validated. To investigate the compatibility of the data with the assumption of normality, we observe that the data reasonably follows the Q-Q plot line (readers can easily verify this independently). So, we can apply the one-sample t-test with reasonable confidence in its interpretability. The mean value is $\bar{x} = -3.2$ (SD 3.5). By choosing the null one-sided assumption $h \geq 0$, the largest experimental result is $r = \bar{x} - \min\{h\} = -3.2$ and the associated test result is $t_9 = -2.9$, which implies $S = 6.8$ (i.e., the test result is as surprising as just under 7 consecutive heads). Can we conclude anything? The quick answer is no. Indeed, we have no idea how the degree of surprise of our result varies compared to other hypotheses. To remedy this, we construct the following S-intervals: 6|5|4|3-I. The goal is to understand the “rapidity” at which statistical surprise diminishes to less surprising levels. We obtain 6|5|4|3-I = $(-\infty, -0.4 | -\infty, -0.8 | -\infty, -1.3 | -\infty, -1.8)$. In practice, we lose 1 bit (head) for every 0.5 mmHg, meaning statistical surprise diminishes very rapidly in relation to target hypotheses that predict tiny variations in blood pressure. This indicates that our results are highly unstable. Therefore, we cannot conclude anything other than “these results are too uncertain to properly inform a scientific conclusion.”

4.2. Example 2

The scenario is supposed to be the same as the previous example, but in this case, we refer to the data in Table 3. Let’s take all the statistical and non-statistical background assumptions for granted (the statistical ones can be easily investigated, as shown in the previous example).

Patient	Before (mmHg)	After (mmHg)	Difference (mmHg)
1	140	127	-13
2	150	140	-10
3	130	123	-7
4	135	136	+1
5	145	135	-10
6	138	141	+3
7	142	138	-4
8	128	120	-8
9	152	143	-9
10	134	125	-9

Table 3: Hypothetical blood pressure data before and after treatment: case 2.

In this case, we have an average value $\bar{x} = -6.6$ (SD 5.1). By choosing

the null one-sided assumption $h \geq 0$, the largest experimental result is $r = \bar{x} - \min\{h\} = -6.6$ and the associated test result is $t_9 = -4.1$, which implies $S = 9.5$ (i.e., the test result is more surprising than 9 consecutive heads). To assess the surprise decrease rapidity against different target assumptions, we construct the following S-intervals: $9|7|5|3\text{-I} = (-\infty, -0.4| -\infty, -1.8| -\infty, -3.2| -\infty, -4.6)$. In this case, the decrease could be acceptable. Thus, can we say we have proven the effectiveness of the treatment? Absolutely not. As mentioned earlier, statistics is a limited component of scientific inquiry. So, can we at least say we have found evidence in favor of the treatment's effectiveness? No. At best, we have found evidence compatible with the effectiveness of the treatment. However, this evidence is also compatible with other equally valid hypotheses. For example, the absence of a control group prevents us from establishing the impact of atmospheric variations and changes in the patients' physical activity and dietary habits over these 3 months (e.g., with the onset of summer, patients might spend more time outdoors and be inclined toward a more Mediterranean diet). Bias analysis is extremely important in this regard [14]. Alongside this, the sample is arguably too small to be representative of the entire population. At the ethical and scientific level, we must assess the invasiveness of the therapy. For instance, does the effect size justify any potential physical and/or psychological adverse events? Not only that, statistics deals with numbers, i.e., it is unable to encompass the clinical complexity of each individual patient. Indeed, patients 4 and 6 even recorded an increase in blood pressure that should be investigated clinically. Furthermore, the dataset exhibited high variability (percentage variation coefficient=77%). Nonetheless, admitting that there are valid biochemical reasons to suspect the effectiveness of the treatment, in the event that the latter has not yielded negative consequences for the patients, these results could justify further research.

5. Discussion

The adoption of surprisal intervals completes the evaluative approach of statistical surprise, avoiding any reliance on statistical significance and confidence (topics that are much more complex and cryptic even for expert statisticians). S-intervals finally make explicit the relationship between surprise and effect size and, in light of the uncertainties that affect the testing of a statistical hypothesis, prevent the adoption of excessively sharp and senseless statistical significance thresholds. This interpretation is reinforced by the definition of S-intervals, which only permits reporting intervals at least 1 bit apart. For this reason, the conventions and methodologies suggested in this paper never allow for the statistical rejection or acceptance of a single target hypothesis since the researcher is urged to reason only in terms of greater or lesser surprise (also in relation to effect size estimation intervals). Indeed, any concrete action of this type (e.g., to promote a drug for commercialization) must be made solely based on a careful evaluation of the quality of the evidence available and a detailed analysis of biases, costs,

and benefits for stakeholders since no mere statistical criterion can ever automatically demonstrate causation nor answer the question, “Is it worth it?” The final decision must, therefore, be informed by the union of evidence of various kinds (e.g., statistical tests, proven chemical-biological mechanisms, clinical reports, etc.). Such scientific practice, known as decision analysis, is central for public health [7, 11, 14, 18]. In addition to this, the compact formulation of multiple intervals can provide a much more complete and clearer overview than that described by a traditional confidence/compatibility interval without excessively burdening the reading, i.e., remaining suitable to be used in summary sections such as the abstract. Although the problems related to statistical testing are numerous and go beyond the scope of this manuscript (e.g., arbitrary multiple comparisons adjustments, p-hacking, statistical power misconceptions, and publication bias), the interpretation of test results is fundamental or integral to each of these [1–18, 22]. Surprisal intervals, in conjunction with surprisals, can provide great assistance to the scientific community in framing research problems, especially in the field of public health where errors regarding statistical significance are as frequent as they are dangerous. In fact, comparing test results to a perceptually familiar phenomenon, such as the number of consecutive successes (heads) when flipping an unbiased coin, not only greatly simplifies the evaluation of the statistical weight of the event under consideration but also contributes to avoiding overstatements. Consequently, it is highly recommended that surprisal intervals be adopted in future scientific investigations based on statistical testing.

REFERENCES

- [1] AMRHEIN, V., GREENLAND, S., and MCSHANE, B. (2019). Scientists rise up against statistical significance, *Nature*, **567**, 7748, 305-307.
- [2] BIAU, D.J., JOLLES, B.M., and PORCHER, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers, *Clinical Orthopaedics and Related Research*, **468**, 3, 885–892.
- [3] BORING, E.G. (1919). Mathematical vs. scientific significance, *Psychological Bulletin*, **16**, 10, 335–338.
- [4] FRIESE, M., and FRANKENBACH, J. (2020). P-Hacking and publication bias interact to distort meta-analytic effect size estimates, *Psychological Methods*, **25**, 4, 456–471.
- [5] GELMAN, A., and STERN, H. (2006) The difference between “significant” and “not significant” is not itself statistically significant, *The American Statistician*, **60**, 4, 328-331.
- [6] GREENLAND, S. (2017). Invited commentary: The need for cognitive science in methodology, *American Journal of Epidemiology*, **186**, 6, 639–645.
- [7] GREENLAND S. (2021). Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons, *Paediatric and Perinatal Epidemiology*, **35**, 1, 8-23.

- [8] GREENLAND, S. (2023) Connecting simple and precise P-values to complex and ambiguous realities (includes rejoinder to comments on “Divergence vs. decision P-values”, *Scandinavian Journal of Statistics*, **50**, 3, 899-914.
- [9] GREENLAND S. (2023). Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not, *Scandinavian Journal of Statistics*, **50**, 1, 54-88.
- [10] GREENLAND, S., MANSOURNIA, M.A., and JOFFE, M. (2022) To curb research misreporting, replace significance and confidence by compatibility: A preventive medicine golden jubilee article, *Preventive Medicine*, **164**, 107127.
- [11] GREENLAND, S., SENN, S.J., ROTHMAN, K.J., CARLIN, J.B., POOLE, C., GOODMAN, S.N., and ALTMAN, D.G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, *European Journal of Epidemiology*, **31**, 4, 337–350.
- [12] KÜHBERGER, A., FRITZ, A., LERMER, E., and SCHERNDL, T. (2015). The significance fallacy in inferential statistics, *BMC Research Notes*, **8**, 54-88.
- [13] LAKENS, D. (2021) The practical alternative to the P value is the correctly used P value, *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, **16**, 3, 639-648.
- [14] LASH, T.L., FOX, M.P., MACLEHOSE, R.F., MALDONADO, G., MCCANDLESS, L.C., and GREENLAND, S. (2014). Good practices for quantitative bias analysis, *International Journal of Epidemiology*, **43**, 6, 1969–1985.
- [15] MCSHANE, B.B., BRADLOW, E.T., LYNCH, J.G., and MEYER, R.J. (2023) EXPRESS: “Statistical significance” and statistical reporting: moving beyond binary, *Journal of Marketing*, **0**, ja.
- [16] MCSHANE, B.B., and GAL, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence, *Management Science*, **62**, 6, 1707-1718.
- [17] RAFI, Z., and GREENLAND, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise, *BMC Medical Research Methodology*, **20**, 1, 244.
- [18] ROVETTA, A. (2023). Common statistical errors in scientific investigations: A simple guide to avoid unfounded decisions, *Cureus*, **15**, 1, e33351.
- [19] ROVETTA, A., and CASTALDO, L. (2022). Are we sure we fully understand what an infodemic is? A global perspective on infodemiological problems, *JMIRx Med*, **3**, 3, e36510.
- [20] SWIRE-THOMPSON, B., DOBBS, M., THOMAS, A., and DEGUTIS, J. (2023). Memory failure predicts belief regression after the correction of misinformation, *Cognition*, **230**, 105276.
- [21] XIE, M., and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter – a review (with discussion), *International Statistical Review*, **81**, 3-39.
- [22] WASSERSTEIN, R.L., and LAZAR, N.A. (2016). The ASA’s statement on P-values: Context, process, and purpose, *The American Statistician*, **70**, 2, 129-133.