

STATISTICAL THEORY

Technical Issues in the Interpretation of S-values and Their Relation to Other Information Measures

Sander Greenland¹  | Zad Rafi² 

¹Department of Epidemiology and
Department of Statistics, University of
California, Los Angeles, CA

²Department of Population Health, NYU
Langone, New York, NY

Correspondence
Sander Greenland

Contact

¹Email: lesdomes@ucla.edu

²Twitter: @Lester_Domes

²Email: zad@lesslikely.com

¹Twitter: @DailyZad

An extended technical discussion of S-values and unconditional information can be found in Greenland [1] and Greenland & Rafi [2]. Here we briefly cover several technical topics mentioned in our [main paper](#) [3]: Different units for (scaling of) the S-value besides base-2 logs (bits); the importance of uniformity (validity) of the P-value for interpretation of the S-value; the combination of the S-value across studies; and the relation of the S-value to other measures of statistical information about a test hypothesis or model.

KEYWORDS

Information Statistics · Statistical Theory · Bayesian Statistics · Statistical Models
Data Interpretation · Hypothesis Tests · P-values · S-values · Statistical Significance

1 | BACKGROUND

An extended technical discussion of S-values and unconditional information can be found in Greenland [1] and Greenland & Rafi [2]. Here we briefly cover several technical topics mentioned in our [main paper](#) [3]: Different units for (scaling of) the S-value besides base-2 logs (bits); the importance of uniformity (validity) of the P-value for interpretation of the S-value; the combination of the S-value across studies; and the relation of the S-value to other measures of statistical information about a test hypothesis or model.

2 | UNITS FOR THE S-VALUE

Other units for measuring information other than bits arise from different choices for the base of the logarithms. For example, using natural (base-e) logs, the S-value becomes $s_e = -\ln(p) = -\log_2(p) \ln(2)$ whose units are called “nats,” while using common (base-10) logs the S-value becomes $s_{10} = -\log_{10}(p) = -\log_2(p) \log_{10}(2)$ whose units are called hartleys, bans, or dits (decimal digits). The ratio of one dit of information to one bit of information is $\log_2(10) = 3.22$ which is similar to the ratio of meters to feet, 3.28. Just as the choice of meters vs. feet does not affect the concepts and methods surrounding length measurement, so choice of dits vs. bits does not affect any of the concepts or methods of information measurement. Bits are most commonly used in communications engineering because the fundamental physical components in electronic information storage are binary and thus their information capacity is one bit. Natural logs are however more mathematically convenient and thus more common in statistical theory (see below), although base-10 logs are also seen.

3 | UNIFORMITY OF THE P-VALUE AND THE INFORMATION IN THE S-VALUE

The decision rule “reject H if $p \leq \alpha$ ” will reject H $100\alpha\%$ of the time under sampling from a model M obeying H (i.e., the Type-1 error rate of the test will be α) provided the random variable P corresponding to p is valid (uniform under the model M used to compute it), but not necessarily otherwise [4]. This is one reason why frequentist writers reject invalid P -values (such as posterior predictive P -values, which highly concentrate around 0.50) and devote considerable technical coverage to uniform P -values [4–6]. A valid P -value (“U-value”) translates into an exponentially distributed S -value with a mean of 1 nat or $\log_2(e) = 1.443$ bits where e is the base of the natural logs.

Uniformity is also central to the “refutational information” interpretation of the S -value used here, for it is necessary to ensure that the P -value p from which s is derived is in fact the percentile of the observed value of the test statistic in the distribution of the statistic under M , thus making small p surprising under M and making s the corresponding degree of surprise. Because posterior predictive P -values do not translate into sampling percentiles of the statistic under the hypothesis (in fact, they are pulled toward 0.5 from the correct percentiles) [5, 6], the resulting negative log does not measure surprisal at the statistic given M , and so is not a valid S -value in our terms.

For simplicity, we have assumed that at least an approximately valid P -value can be derived for testing M . This is so in typical regression analyses in health and medical sciences, but not always. To deal with exceptions, P -values are often said to be “conservatively valid” for testing M when under M they stochastically dominate a unit-uniform distribution, i.e., under M the probability that P exceeds a given p is at least p , and for some p exceeds p . Typical exact P -values from discrete data are conservatively valid but approach uniformity with increasing sample size. In cases for which we can only deduce that P is conservatively valid (as when its observed value p is an upper bound rather than a direct tail probability), we would interpret its corresponding S -value as conservatively valid in the sense of representing the minimum information against M supplied by the test.

The coin-toss interpretation we have used to physically gauge this information assumes that the only alternative to fairness is in the direction of loading for heads. The S -value it produces thus corresponds to a P -value for the 1-sided hypothesis $\Pr(\text{heads}) \leq 1/2$; nonetheless, this interpretation applies even if the original observed P -value p was 2-sided. This translation from a 2-sided P -value to a 1-sided S -value parallels the transformation of P -values into 1-sided sigmas (Z -scores) in physics, in

which for example a P -value of 0.05 from a two-sided test would become a sigma of 1.645, the upper one-sided 5% cutoff for a standard-normal deviate [7].

4 | COMBINATION OF S-VALUES ACROSS STUDIES

Preprints of our article referred to the S -value as “additive” over independent sources, which is incorrect insofar as the combined refutational information from independent tests of the same model \mathbf{M} is a subadditive function of the separate S -values; only the combination of the latter into a summary test statistic is additive. More precisely, suppose we have K studies, each contributing an independent valid S -value S_k for \mathbf{M} . Under \mathbf{M} , each S_k has an expected value of 1 nat, which can be viewed as the expected “noise” contribution to the S -value (Good [8, 9] dealt with this factor by what in our case reduces to subtracting 1 nat from each surprisal, which however creates problematic negative values when $s < 1$ nat; we have instead chosen to follow the subsequent theoretical literature and not do so). The sum $S_+ = \sum_k S_k$ will thus have an expectation of K “noise” nats under \mathbf{M} . Furthermore, the distribution of $2S_+$ will be χ^2 on $2K$ degrees of freedom [10]; hence the summary S -value $S_\&$ derived from the sum S_+ will be the negative log of the P -value from comparing $2S_+$ to a $2K$ df χ^2 distribution. Under \mathbf{M} this summary- χ^2 $S_\&$ has an expectation of 1 nat and thus will on average be $K - 1$ nats smaller than S_+ , with even larger discrepancies under violations of \mathbf{M} that the test is sensitive to.

More generally, if the test model \mathbf{M}_k varies across studies, the S -summation test just given is valid for testing the conjunction (intersection) hypothesis $\mathbf{M}_\& = \mathbf{M}_1 \& \dots \& \mathbf{M}_K$, but is rarely optimal because it makes no use of homogeneity or other relations among the models. In particular, the test remains valid when as usual each study model \mathbf{M}_k incorporates a shared target or focal hypothesis \mathbf{H} across studies (e.g., no association of a given treatment and disease), combined with different sets of background assumptions \mathbf{A}_k (e.g., due to differences in study designs), so that $\mathbf{M}_k = \mathbf{H} \& \mathbf{A}_k$ and $\mathbf{M}_\& = \mathbf{H} \& \mathbf{A}_1 \& \dots \& \mathbf{A}_K = \mathbf{H} \& \mathbf{A}_\&$. When however a homogeneity assumption is correct, this general test will have lower power (sensitivity) for violations of \mathbf{H} given the background $\mathbf{A}_\&$ than the usual summary tests (which use homogeneity in deriving the study-specific contributions to their summary statistics). On the other hand, those usual tests can have lower power if homogeneity is very wrong. In any case, the S -values from the two tests can differ considerably due to the additional (and possibly incorrect) homogeneity information used in the usual tests.

Some insight into these results may follow from reviewing the traditional parallel procedure for combining Z -scores Z_k (e.g., standardized residuals) by squaring and summing them. The resulting sum of squares $\sum_k Z_k^2$ has a K df χ^2 distribution with expectation K if no cross- k infor-

mation is used to compute the Z_k . More generally however the df are reduced by the number of cross-study sharp constraints used. For example, if H is a hypothesis that a mean difference δ_k is 0, the test using the assumption that the δ_k are constant across studies imposes $K - 1$ constraints ($\delta_1 = \dots = \delta_K$) to derive the Z_k , and so will have only 1 df instead of K df; this reduction in df produces considerably more power if the assumption is correct, but not necessarily otherwise. See Greenland and Rafi [2] for details and examples.

5 | OTHER MEASURES OF STATISTICAL INFORMATION ABOUT A TEST HYPOTHESIS OR MODEL

A common measure for evaluating a hypothesis or model restriction H under background assumptions or unrestricted model A is the maximum-likelihood ratio (MLR), which is the value of the likelihood function at its maximum under A alone, divided by its (restricted) maximum when the test hypothesis H is additionally imposed [11, 12]. The MLR defined this way is always above 1; it is however sometimes confused with the posterior odds against the tested value H given A , which it equals only under very special (and usually unrealistic) conditions. The MLR does however show the most extreme increase in posterior odds against H that the data could produce given A . The corresponding information measure paralleling the S-value is the deviance difference or likelihood-ratio (LR) statistic for H given A , $2 \ln(MLR)$, which is itself a test statistic for H given A . The change in the Akaike Information Criterion (without small-sample adjustment) from adding H to the background model is $2 \ln(MLR) - 2d$ where d is the dimension (degrees of freedom) of H [11, 13].

Now consider a sharp constraint hypothesis H with a P -value less than $1/e = 0.368$. Bayarri & Berger [14] and Sellke et al. [15] show that $b = -e \cdot p \cdot \ln(p) = e \cdot p \cdot s_e$ is a sharp lower bound on the Bayes factor for H under A , where A now includes strong restrictions on the alternatives to H . (A Bayes factor is the ratio of posterior data probabilities under H and an alternative, given A .) Thus, given A and the data, b is a lower bound on the reduction in odds for H given A in moving from a prior to a posterior, and $1/b$ is an upper bound on the increase in odds against H given A . Simple numeric examples show that the latter bound is much lower than the MLR. The strength of the restrictions added to A is indicated for example by the fact that for $p = 0.05$ the MLR in Table 1 of our main paper [3] is 6.83, while $1/b$ is only 2.46. Sellke et al. [15] also discuss how $\frac{1}{1+1/b}$ is the Type-1 error rate for a particular type of conditional decision rule.

Grünwald et al. [16] introduce a general concept they call an S-test statistic (where “S” stands for “safe”) for H given A , defined as any random variable S satisfying $E_M(S) \leq 1$ under any model M obeying H and A . They also call this S an “S-value”. As noted above, our binary S-value $S_2 =$

$-\log_2(P)$ can be redefined using natural logs and thus rescaled to units of nats instead of units of bits, via $S_e = -\ln(P) = -\log_2(P) \ln(2)$. S_e is then an example of their S-value, since $E_M(S_e) \leq 1$ when the random P -value P is valid or conservatively valid (uniform or dominated by a uniform random variable under \mathbf{M}); it is also an example of a betting score [17] (hence “S” can also be taken as “information score”). Grünwald et al. [16] discuss other S-values, including those based on Bayes factors.

Finally, consider a 1-dimensional continuous parameter μ with test hypotheses \mathbf{H} of the form $\mu \leq \mu_0$ and a specified alternative $\mu \geq \mu_1$ (or $\mu = \mu_0$ with alternative $\mu = \mu_1$) where $\mu_0 < \mu_1$. In this context, yet another S-word, “severity”, has been used to refer to the P -value $p(\mu \geq \mu_1)$ for $\mu \geq \mu_1$ (the lower tail of the test statistic $m - \mu_1$ for a 1-sided test of $\mu = \mu_1$ when using the estimate m of μ), which decreases as μ_1 increases; see p. 345 and Fig. 5.5 of Mayo [18]. Since the complement $p(\mu \leq \mu_1) = 1 - p(\mu \geq \mu_1)$ is the P -value for $\mu \leq \mu_1$, we find that (whatever the base) the corresponding S-value function $s(\mu \leq \mu_1) = -\log(p(\mu \leq \mu_1))$ measuring the information against $\mu \leq \mu_1$ increases as $p(\mu \geq \mu_1)$ increases; thus $p(\mu \geq \mu_1)$ varies directly with the information $s(\mu \leq \mu_1)$ against $\mu \leq \mu_1$ (the case with alternative $\mu_1 < \mu_0$ is handled symmetrically). This so-called “severity” of the test of the original \mathbf{H} ($\mu \leq \mu_0$) is not in fact a function of μ_0 and so is identical for all μ_0 . Furthermore, it incorporates no information about background assumptions (e.g., whether treatment assignment was blinded) which bear heavily on practical notions of severity. We thus conclude that it is misleading to label $p(\mu \geq \mu_1)$ as a severity measure, and it instead should be recognized and treated as the P -value function it is.

REFERENCES

- [1] Greenland, S. Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician* **73**, 106–114 (2019). <https://doi.org/10.1080/00031305.2018.1529625>.
- [2] Greenland, S. & Rafi, Z. To Aid Scientific Inference, Emphasize Unconditional Descriptions of Statistics. *arXiv:1909.08583 [stat.ME]* (2020). <https://arxiv.org/abs/1909.08583>.
- [3] Rafi, Z. & Greenland, S. Semantic and Cognitive Tools to Aid Statistical Science: Replace Confidence and Significance by Compatibility and Surprise. *BMC Medical Research Methodology In Press* (2020). <https://arxiv.org/abs/1909.08579>.
- [4] Kuffner, T. A. & Walker, S. G. Why are P-values controversial? *The American Statistician* **73**, 1–3 (2019). <https://doi.org/10.1080/00031305.2016.1277161>.
- [5] Robins, J. M., van der Vaart, A. & Ventura, V. Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association* **95**, 1143–1156 (2000). <https://doi.org/10.1080/01621459.2000.10474310>.
- [6] Bayarri, M. J. & Berger, J. O. P Values for Composite Null Models. *Journal of the American Statistical Association* **95**, 1127–1142 (2000). <https://doi.org/10.2307/2669749>.
- [7] Cousins, R. D. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese* **194**, 395–432 (2017). <https://doi.org/10.1007/s11229-014-0525-z>.
- [8] Good, I. J. The surprise index for the multivariate normal distribution. *The Annals of Mathematical Statistics* **27**, 1130–1135 (1956). <https://doi.org/10.1214/aoms/1177728079>.
- [9] Good, I. J. Corrections to "The Surprise Index for the Multivariate Normal Distribution". *The Annals of Mathematical Statistics* **28**, 1055–1055 (1957). <http://www.jstor.org/stable/2237075>.
- [10] Cox, D. R. & Hinkley, D. V. Chapter 3, Pure significance tests. In *Theoretical Statistics*, 64–87 (Chapman and Hall/CRC, 1974). <https://doi.org/10.1201/b14832>.
- [11] Cummings, P. *Analysis of Incidence Rates* (CRC Press, 2019). <https://doi.org/10.1201/9780429055713>.
- [12] Jewell, N. P. *Statistics for Epidemiology* (CRC Press, 2003). https://books.google.com/books/about/Statistics_for_Epidemiology.html?id=if4K02ui2p4C.
- [13] Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer-Verlag, New York, 2002), second edn. <https://doi.org/10.1007/b97636>.
- [14] Bayarri, M. J. & Berger, J. O. Quantifying Surprise in the Data and Model Verification. *Bayesian Statistics 6*, 53–82 (1999). <https://www.semanticscholar.org/paper/Quantifying-Surprise-in-the-Data-and-Model-Bayarri-Berger/e0795969e605e80c62c4120000766280bcd9ab83>.
- [15] Sellke, T., Bayarri, M. J. & Berger, J. O. Calibration of p values for testing precise null hypotheses. *The American Statistician* **55**, 62–71 (2001). <https://doi.org/10.1198/000313001300339950>.
- [16] Grünwald, P., de Heide, R. & Koolen, W. Safe Testing. *arXiv:1906.07801 [cs, math, stat]* (2019). <https://arxiv.org/abs/1906.07801>.
- [17] Shafer, G. The Language of Betting as a Strategy for Statistical and Scientific Communication. *Journal of the Royal Statistical Society in press* (2020).
- [18] Mayo, D. G. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (Cambridge University Press, 2018). <https://doi.org/10.1017/9781107286184>.