American Journal of Epidemiology

Vol. 190, No. 2
DOI: 10.1093/aje/kwaa136
Advance Access publication:
July 10, 2020

Commentary

Surprise!

Stephen R. Cole*, Jessie K. Edwards, and Sander Greenland

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu).

Initially submitted April 3, 2020; accepted for publication July 7, 2020.

Measures of information and surprise, such as the Shannon information value (*S* value), quantify the signal present in a stream of noisy data. We illustrate the use of such information measures in the context of interpreting *P* values as compatibility indices. *S* values help communicate the limited information supplied by conventional statistics and cast a critical light on cutoffs used to judge and construct those statistics. Misinterpretations of statistics may be reduced by interpreting *P* values and interval estimates using compatibility concepts and *S* values instead of "significance" and "confidence."

compatibility; confidence intervals; information; *P* value; random error; *S* value; significance tests; statistical inference

Editor's note: *The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the* American Journal of Epidemiology. *A response to this commentary appears on page 194.*

Measures of information and surprise have a long history (see Good (1), chapter 16) but have seen little use outside the fields of engineering and mathematical statistics. Such measures of information and surprise attempt to quantify the signal present in a stream of noisy data. One such measure is the Shannon information, which, when seeing an event of probability p, is defined as $s = \log_2(1/p) = -\log_2(p)$ and is also known as the binary surprise index or surprisal (2). It has been argued that this measure could aid interpretation of *P* values and interval estimates, especially when the latter are viewed as showing compatibility of data with hypotheses, rather than stronger notions of significance or confidence (3–5). Here we briefly illustrate these ideas in the context of interpreting *P* values as compatibility indices.

THE *P* VALUE AS A COMPATIBILITY INDEX

A *P* value represents the chance of observing a data summary (test statistic) as extreme as or more extreme than what was seen, under a test hypothesis and auxiliary (background) assumptions. Perhaps the most common auxiliary assumptions are that the observed data are randomly sampled or treatment is randomly assigned within observed covariate levels, and that measurement error is negligible (6); regression models add further assumptions. Typically the test hypothesis is that a parameter is 0 (2-sided null) or is no greater than 0 (1-sided null), but other values can and should be tested besides 0 (3–5, 7; also see Rothman et al. (8), chapter 10).

A *P* value is valid if it would have a uniform distribution when sampling data under the tested hypothesis given the auxiliary assumptions used to compute it. Such a *P* value can be interpreted as giving the percentile at which the observed data fell in this distribution. The *P* value can thus be taken as an index of compatibility between the data and the parameter values specified by the tested hypothesis given the auxiliary assumptions, ranging from $p = 0$ (data flatly contradict the hypothesis) to $p = 1$ (data are exactly as expected under the hypothesis) (3–5). A valid 95% confidence interval can be constructed as the set of all parameter values with $p > 0.05$ (see Rothman et al. (8), chapter 10). Therefore, the values of a 95% confidence interval have a compatibility index of 0.05 and above, and they comprise a 5%-or-more compatibility interval (3–5), which can also, like the confidence interval, be abbreviated using "CI." (Some authors define a *P* value as a random variable *P* that is uniform under the test hypothesis and auxiliary assumptions, with p being the value of *P* in the observed data).

Consider a recent randomized trial of lopinavir and ritonavir, versus standard care, in the treatment of severe coronavirus disease 2019 (9) which reported 19 and 25 deaths among 99 and 100 patients, respectively. The authors stated that "no benefit was observed with lopinavir-ritonavir treatment beyond standard care" (9, p. 1787), despite observing a 28-day mortality risk difference of −5.8% (i.e., 19.2% − 25.0%), with a 95% compatibility ("confidence") interval ranging from −17.3% to 5.7%. This interval includes risk differences ranging from −17.3%, which represents a tremendous mortality benefit, to 5.7%, which represents a nontrivial increase in mortality. The statistics leave the hypothesis of no benefit (i.e., a causal risk difference ≥ 0) as reasonably compatible with the data, with 1-sided $p = 0.16$ (from a z score of 0.9780 (−0.0581/0.0587), where 0.0587 (0.0587 = [0.057 − (−0.173)]/3.92) is an approximate standard error). But benefits up to a risk difference of −11.6% are even more compatible with the data, in that they have even higher P values than does no benefit.

THE *S* VALUE

The S value provides a reinterpretation of the P value using a familiar mechanical framework for calibrating intuitions, one that is simpler and less abstract than effect estimation from statistical models. Envision a coin-tossing setup that we want to check for bias toward heads (as we might be advised to do if we were going to wager on tails from this setup). We check by tossing the coin s times. If we observe heads on *every* toss, the exact P value for the hypothesis of no bias toward heads is 0.5^s, a special case of the fact that, for m heads in n tosses, the exact P value for the 1-sided hypothesis that "the probability of heads is no greater than μ" is $\sum_{k=m}^{n} \binom{n}{k} \mu^k (1-\mu)^{(n-k)}$. The Shannon measure of the information against this hypothesis is then the binary surprisal $-\log_2(0.5^s) = s$, the number of heads in a row observed. Because s is computed using base-2 logs, its units are said to be bits (binary digits) of information (2, p. 32); other base units are possible (3).

A key benefit of the S value is that it provides a simple coin-tossing framework for interpretation of P values and confidence intervals. Returning to the coronavirus example, the 1-sided P value of 0.16 for the no-benefit hypothesis yields an S value of 2.6 ($-\log_2(0.16) = 2.6$). To place this result into our coin-tossing framework, a result of all heads in 3 fair tosses has a 0.125 (1 in 8) chance of occurring and thus does not seem terribly surprising (albeit it is more surprising than 2 heads in a row, where $p = 0.25$, and less surprising than 4 heads in a row, where $p = 0.0625$). Therefore we say that, if there is no treatment benefit, the observed $p = 0.16$ is less surprising than seeing 3 heads in a row in 3 fair tosses (because $2.6 < 3$).

The P value for a benefit of 11.6% is equal to the P value for the no-benefit hypothesis, meaning that the data are equally compatible with (and would be equally surprising under) both hypotheses. These data would be even less surprising under risk differences between 0 and 11.6%. Viewing the compatibility interval of –17.3% to 5.7%, the data supply at most 4.3 bits of information ($-\log_2(0.05) = 4.3$) against treatment effects ranging from a 17.3% reduction to a 5.7% increase in mortality, and all risk differences in this interval make the data about the same as or less surprising than seeing 4 heads ($-\log_2(0.05) \approx 4$) in 4 fair tosses.

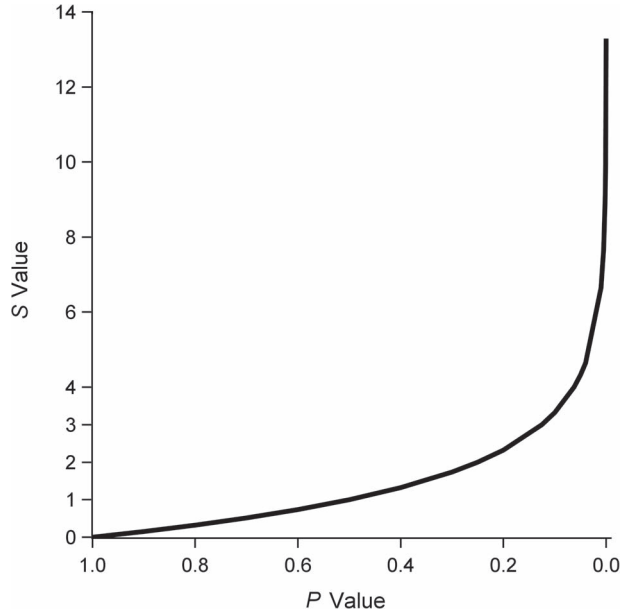

Figure 1. The S value as a function of the P value.

Now consider P values of 0.10, 0.05, 0.01, and 0.005. The corresponding S values are 3.3, 4.3, 6.6, and 7.6, so with rough rounding, these P values should seem about as surprising as seeing 3, 4, 7, or 8 heads in a row from fair coin-tossing. Figure 1 provides the mapping from P values to S values. One may feel that $p > 0.05$ is unsurprising if the test hypothesis is correct, given $s < 4.3$. That judgment is fine; nonetheless, effect sizes with higher P values than the test hypothesis exhibit more compatibility with the data and have less information against them than does the test hypothesis. Thus, $p > 0.05$ is not a sufficient basis for claiming or acting as if the results support the test hypothesis or do not support alternatives, since such dichotomizations mask important distinctions.

The S value is based on the same assumptions as those used to compute its source P value, and thus introduces no new technical or validity issues. While the computations are objectively determined by data and assumptions, their interpretations are subject to the limitations of human cognition. One should expect an event with chance 1 in 10 to happen in one-tenth of our observations, on average. If one hypothesizes that the event is as likely as not (i.e., chance 1 in 2), then one ought to feel no surprise if one sees 1 event in 2 tries (2-sided $p = 1, s = 0$). The extent of our surprise ought to grow, as does the S value, as the data diverge from the hypothesis. Specifically, the S value grows by a unit for every halving of the P value. Being a continuum, there is no particular S value cutpoint above which one ought to be

"surprised." Use of the *P* value or *S* value as a continuum is not as arbitrary as making a dichotomous comparison, say $P < 0.05$. A key point here is that the *S* value maps directly onto a standard game of coin-tossing, providing the highly heterogeneous set of human observers with an easily taught reference system, to help gauge the information content of studies.

In conclusion, we advise that misinterpretations which remain standard in the medical literature can be reduced by reinterpreting *P* values and confidence intervals as indicators of compatibility with data (rather than as indicating significance, confidence, or support). In the above example (9), the authors used confidence intervals as significance tests, concluding that "no benefit was observed" because the 95% confidence interval contained the null value (equivalent to a null $p > 0.05$). But interpreting the *P* values and confidence intervals as compatibility values and intervals instead of significance tests shows that the results are 1) most compatible with a modest benefit and 2) imprecise and therefore highly compatible with a wide range of effects. We thus conclude that compatibility interpretations and *S* values can help communicate the limited information supplied by conventional statistics and can cast a critical light on the cutoffs used to judge and construct those statistics.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Stephen R. Cole, Jessie K. Edwards); Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California (Sander Greenland); and Department of Statistics, College of Physical Sciences, University of California, Los Angeles, Los Angeles, California (Sander Greenland).

This work was supported in part by National Institutes of Health grants K01 AI125087 and R01 AI157758.

Conflict of interest: none declared.

REFERENCES

1. Good IJ. *Good Thinking*. Minneapolis, MN: University of Minneapolis Press; 1983.
2. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–424, 623–656.
3. Greenland S. Valid *P*-values behave exactly as they should: some misleading criticisms of *P*-values and their resolution with *S*-values. *Am Stat*. 2019;73(suppl 1):106–114.
4. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *Am Stat*. 2019;73(suppl 1):262–270.
5. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol.* 2020; 20:Article 244.
6. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1(6):421–429.
7. Poole C. Beyond the confidence interval. *Am J Public Health*. 1987;77(2):195–199.
8. Rothman KJ, Greenland S, Lash T. *Modern Epidemiology*. 3rd ed. New York, NY: Lippincott-Raven; 2008.
9. Cao B, Wang Y, Wen D, et al. A trial of lopinavir-ritonavir in adults hospitalized with severe Covid-19. *N Engl J Med*. 2020; 382(19):1787–1799.