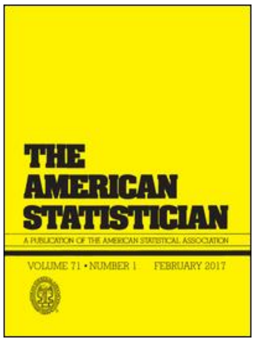

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Response to the ASA's Statement on *p*-Values: Context, Process, and Purpose

Edward L. Ionides, Alexander Giessing, Yaacov Ritov & Scott E. Page

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 1, 88–89
http://dx.doi.org/10.1080/00031305.2016.1234977

Taylor & Francis
Taylor & Francis Group

LETTERS TO THE EDITOR

Response to the ASA's Statement on *p*-Values: Context, Process, and Purpose

Edward L. Ionides[a], Alexander Giessing[a], Yaacov Ritov[a], and Scott E. Page[b]

[a]Department of Statistics, University of Michigan, Ann Arbor, MI; [b]Departments of Complex Systems, Political Science and Economics, University of Michigan, Ann Arbor, MI

The ASA's statement on *p*-values: context, process, and purpose (Wasserstein and Lazar 2016) makes several reasonable practical points on the use of *p*-values in empirical scientific inquiry. The statement then goes beyond this mandate, and in opposition to mainstream views on the foundations of scientific reasoning, to advocate that researchers should move away from the practice of frequentist statistical inference and deductive science. Mixed with the sensible advice on how to use *p*-values comes a message that is being interpreted across academia, the business world, and policy communities, as, "Avoid *p*-values. They don't tell you what you want to know." We support the idea of an activist ASA that reminds the statistical community of the proper use of statistical tools. However, any tool that is as widely used as the *p*-value will also often be misused and misinterpreted. The ASA's statement, while warning statistical practitioners against these abuses, simultaneously warns practitioners away from legitimate use of the frequentist approach to statistical inference.

In particular, the ASA's statement ends by suggesting that other approaches, such as Bayesian inference and Bayes factors, should be used to solve the problems of using and interpreting *p*-values. Many committed advocates of the Bayesian paradigm were involved in writing the ASA's statement, so perhaps this conclusion should not surprise the alert reader. Other applied statisticians feel that adding priors to the model often does more to obfuscate the challenges of data analysis than to solve them. It is formally true that difficulties in carrying out frequentist inference can be avoided by following the Bayesian paradigm, since the challenges of properly assessing and interpreting the size and power for a statistical procedure disappear if one does not attempt to calculate them. However, avoiding frequentist inference is not a constructive approach to carrying out better frequentist inference.

On closer inspection, the key issue is a fundamental position of the ASA's statement on the scientific method, related to but formally distinct from the differences between Bayesian and frequentist inference. Let us focus on a critical paragraph from the ASA's statement: "In view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretical modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct."

Some people may want to think about whether it makes scientific sense to "directly address whether the hypothesis is correct." Some people may have already concluded that usually it does not, and be surprised that a statement on hypothesis testing that is at odds with mainstream scientific thought is apparently being advocated by the ASA leadership. Albert Einstein's views on the scientific method are paraphrased by the assertion that, "No amount of experimentation can ever prove me right; a single experiment can prove me wrong" (Calaprice 2005). This approach to the logic of scientific progress, that data can serve to falsify scientific hypotheses but not to demonstrate their truth, was developed by Popper (1959) and has broad acceptance within the scientific community. In the words of Popper (1963), "It is easy to obtain confirmations, or verifications, for nearly every theory," while, "Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability." The ASA's statement appears to be contradicting the scientific method described by Einstein and Popper. In case the interpretation of this paragraph is unclear, the position of the ASA's statement is clarified in their Principle 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. Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither." Here, the ASA's statement misleads through omission: a more accurate end of the paragraph would read, "The *p*-value is neither. Nor is any other statistical test used as part of a deductive argument." It is implicit in the way the authors have stated this principle that they believe alternative scientific methods may be appropriate to assess more directly the truth of the null hypothesis. Many readers will infer the ASA to imply the inferiority of deductive frequentist methods for scientific reasoning. The ASA statement, in its current form, will therefore make it harder for scientists to defend a choice of frequentist statistical methods during peer review. Frequentist articles will become more difficult to publish, which will create a cascade of effects on data collection, research design, and even research agendas.

Gelman and Shalizi (2013) provided a relevant discussion of the distinction between deductive reasoning (based on deducing conclusions from a hypothesis and checking whether they

CONTACT Edward L. Ionides *ionides@umich.edu* Department of Statistics, University of Michigan, Ann Arbor, MI 48109.

can be falsified, permitting data to argue against a scientific hypothesis but not directly for it) and inductive reasoning (which permits generalization, and therefore allows data to provide direct evidence for the truth of a scientific hypothesis). It is held widely, though less than universally, that only deductive reasoning is appropriate for generating scientific knowledge. Usually, frequentist statistical analysis is associated with deductive reasoning and Bayesian analysis is associated with inductive reasoning. Gelman and Shalizi (2013) argued that it is possible to use Bayesian analysis to support deductive reasoning, though that is not currently the mainstream approach in the Bayesian community. Bayesian deductive reasoning may involve, for example, refusing to use Bayes factors to support scientific conclusions. The Bayesian deductive methodology proposed by Gelman and Shalizi (2013) is a close cousin to frequentist reasoning, and in particular emphasizes the use of Bayesian p-values.

The ASA probably did not intend to make a philosophical statement on the possibility of acquiring scientific knowledge by inductive reasoning. However, it ended up doing so, by making repeated assertions implying, directly and indirectly, the legitimacy and desirability of using data to directly assess the correctness of a hypothesis. This philosophical aspect of the ASA statement is far from irrelevant for statistical practice, since the ASA position encourages the use of statistical arguments that might be considered inappropriate.

A judgment against the validity of inductive reasoning for generating scientific knowledge does not rule out its utility for other purposes. For example, the demonstrated utility of standard inductive Bayesian reasoning for some engineering applications is outside the scope of our current discussion. This amounts to the distinction Popper (1959) made between "common sense knowledge" and "scientific knowledge."

References

Calaprice, A. (2005), *The New Quotable Einstein*, Princeton, NJ: Princeton University Press. [88]

Gelman, A., and Shalizi, C. R. (2013), "Philosophy and the Practice of Bayesian Statistics," *British Journal of Mathematical and Statistical Psychology*, 66, 8–38. [88]

Popper, K. (1963), *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Routledge and Kegan Paul. [88]

Popper, K. R. (1959), *The Logic of Scientific Discovery*, London: Hutchinson. [88,89]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [88]

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 1, 89–90
http://dx.doi.org/10.1080/00031305.2016.1265586

Kittaneh, O. A., Khan, M. A. U., Akbar, M., Bayoud, H. A. (2016), "Average Entropy: A New Uncertainty Measure with Application to Image Segmentation," *The American Statistician*, 70, 18–24.

Average Entropy Does Not Measure Uncertainty

Kittaneh et al. (2016) proposed a new quantity, the average entropy, to serve as an alternative measure of uncertainty. They demonstrated that the average entropy preserves several desirable properties of Shannon entropy (1948), even when applied to continuous probability density functions (pdfs).

However, in the process of satisfying these properties, the average entropy fails to meet the most fundamental requirement of all, namely, measuring uncertainty: The more spread out the outcome probabilities, the greater the uncertainty, and hence the greater the entropy should be. For example, in the case of k possible outcomes and equal probabilities $p_i = \frac{1}{k}$, the entropy is a monotonic increasing function of k (Shannon 1948). But the average entropy does not satisfy this requirement in general. I illustrate first with a discrete random variable (geometric), then with two continuous ones (exponential and normal).

The authors define the average entropy as $A(X) = -E[\log \frac{f(X)}{E[f(X)]}]$, or

$$A(X) = H + \log E[f(X)], \tag{1}$$

where H is the Shannon entropy, $H = -E[\log f(X)]$. For the geometric pdf $f(y) = p(1-p)^{y-1} = pq^{y-1}$, defined on $y = 1, 2, \ldots,$ calculate

$$H = -E[\log p + (Y-1)\log q] = -\frac{p\log p + q\log q}{p}, \tag{2}$$

and the expected value of $f(Y)$

$$E[f(Y)] = \frac{p^2}{q^2}\sum_{y=1}^{\infty} q^{2y} = \frac{p^2}{q^2}\frac{q^2}{1-q^2} = \frac{p^2}{1-q^2}. \tag{3}$$

Insert (2) and (3) into (1):

$$A(x) = -\frac{p\log p + q\log q}{p} + \log\frac{p^2}{1-q^2}. \tag{4}$$

The geometric entropy should approach infinity for small p, when probability is spread out across many values of y (illustrated for $p = 0.1$ in Figure 1(a)), and zero as p approaches 1,

CONTACT Susan E. Hodge *susan.hodge@nationwidechildrens.org* Battelle Center for Mathematical Medicine, The Research Institute, Nationwide Children's Hospital & Department of Pediatrics, The Ohio State University, Research Building III, 575 Children's Crossroad, Columbus, OH 43215, USA.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.

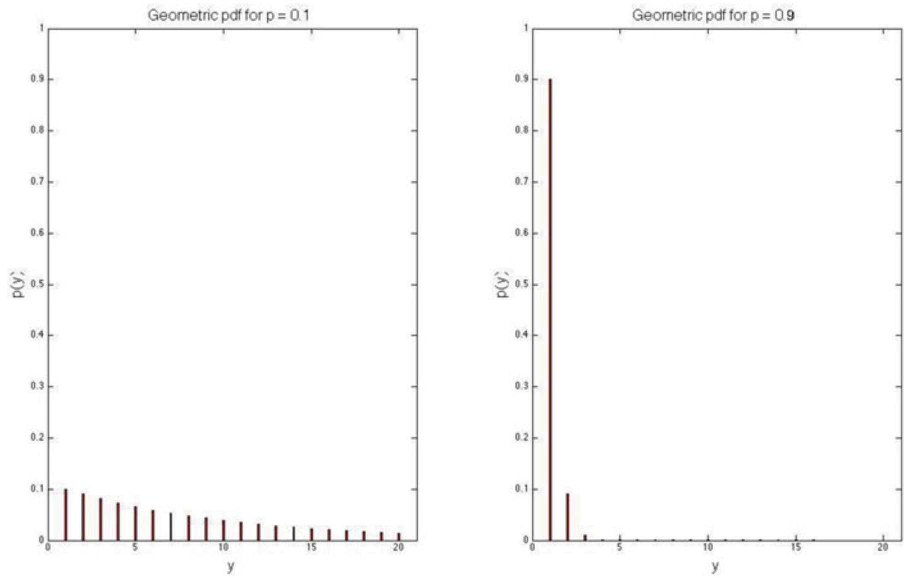

Figure 1. Geometric pdf as a function of y. (a) $p = 0.1$. (b) $p = 0.9$.

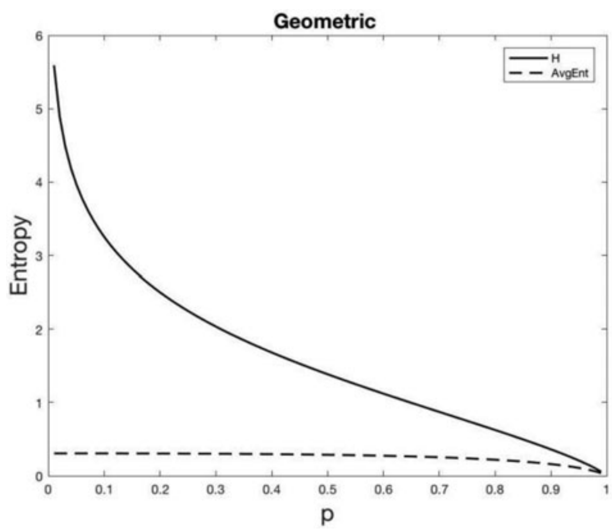

Figure 2. Shannon entropy (H) and average entropy (AvgEnt) as functions of parameter p in geometric pdf.

when all the probability is concentrated near $y = 0$ ($p = 0.9$ in Figure 1(b)). Indeed, the Shannon entropy in (2) behaves in exactly this way (solid curve in Figure 2). However, the average entropy in (4) is practically flat as a function of p (dashed curve in Figure 2), capturing nothing of the extent of uncertainty in the geometric distribution.

The exponential function, $f(y) = \lambda e^{-\lambda y}$, represents a continuous analog to the geometric, and its average entropy is actually *constant* with respect to the parameter λ: $H = 1 - \log \lambda$, $\log\ E[f(Y)] = \log\frac{\lambda}{2}$, and thus average entropy $= 1 - \log 2$. (This quantity, $1 - \log 2$, also equals the limit as $p \to 0$ of the average entropy in Figure 2.) Similarly, the normal function, $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$, also has constant average entropy with respect to the parameter σ^2: $H = 0.5[\log(2\pi\sigma^2) + 1]$, $\log E[f(Y)] = \log\frac{1}{2\sqrt{\pi\sigma^2}}$, and average entropy $= \frac{1-\log 2}{2}$, that is, one-half the constant value for the exponential pdf. Since the average entropy is not a function of parameters for these two pdfs, it cannot give any information about how spread out the probabilities is as a function of the parameters.

There do exist some pdfs for which the average entropy reflects the spread of probabilities. However, as these examples show, the average entropy cannot function as a general measure of uncertainty.

Funding

This research was supported by The Research Institute, Nationwide Children's Hospital, Columbus, OH.

References

Kittaneh, O. A., Khan, M. A. U., Akbar, M., and Bayoud, H. A. (2016), "Average Entropy: A New Uncertainty Measure with Application to Image Segmentation," *The American Statistician*, 70, 18–24. [89]

Shannon, C. E. (1948), "A Mathematical Theory of Communications," *Bell System Technical Journal*, 27, 379–423. [89]

Susan E. Hodge
Battelle Center for Mathematical Medicine, The Research Institute, Nationwide Children's Hospital & Department of Pediatrics, The Ohio State University, Columbus, OH

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 1, 91
http://dx.doi.org/10.1080/00031305.2016.1269484

Reply

The letter directs some comments and criticisms regarding Kittaneh et al. (2016). Unfortunately, the author(s) reject the average entropy (A) as a measure of uncertainty because it has different behavior and properties from the well-known Shannon entropy (H).

- The author(s) mention that the more spread out the outcome probability, the greater the value of the entropy should be:

Which entropy is meant here? Shannon entropy is one member of a big family of entropies. For example, the well-known Tsallis entropy (1988) is not in general increasing function of k, if we have k outcomes with equal probabilities $p_i = \frac{1}{k}$.

In fact entropy measures quantify to some extent the distance between probability distribution and the uniform distribution (the noninformative distribution or the state of complete ignorance). In particular, Shannon entropy attains its maximum when the distribution is uniform, however, average entropy is minimum (zero) for uniform distributions.

- It is also mentioned that the entropy of the geometric distribution

$$f(y) = p(1-p)^{y-1},\ y = 1, 2, \ldots. \tag{1}$$

should go to infinity as p goes to zero.

Actually, this is applicable to some entropy measures like Shannon entropy but not all; again, we can check that Tsallis entropy of degree 2 goes to one as p goes to zero. So can we say that Tsallis entropy does not measure uncertainty because of this. Actually, each entropy measure quantifies the uncertainty in different way.

- The author(s) point out that the average entropy is constant and equal to 1-log2 for the exponential distribution, and this value is also the limit of the average entropy for the geometric distribution with parameter p as p goes to zero.

It is a well-known fact that the geometric distribution is the discrete analog of the exponential distribution. In particular, $X \sim \text{Geom}(p = 1/n)$ converges asymptotically to $Y \sim \text{Exp}(1/n)$. In this case, Shannon entropy for X and Y are, respectively, equal to $H(X) = \log n + (n-1)\log(1-1/n)$ and $H(Y) = 1 + \log n$. Also the average entropy for X and Y are, respectively, equal to $A(X) = (n-1)\log(1-1/n) - \log(2-1/n)$ and $A(Y) = 1 - \log 2$. Note that both $H(X)$ and $H(Y)$ go to the same value (infinity), and both $A(X)$ and $A(Y)$ go to the same value$(1 - \log 2)$, as n goes infinity. Both results are acceptable (consistent). However, if we start with an exponential distribution $Z \sim \text{Exp}(\lambda)$ with PDF $f(z) = \lambda e^{-\lambda z}$,$z \in [0, \infty)$,and divide its support into bins of equal lengthΔ, then the corresponding empirical distribution say Z^Δ has PMF

$$P(Z^\Delta = z_k) = \int_{k\Delta}^{(k+1)\Delta} f(x)dx, \\ = (1-e^{-\lambda\Delta})e^{-\lambda\Delta k} = \Delta f(z_k), \quad k = 0, 1, 2, \ldots \tag{2}$$

which is exactly a geometric PMF with parameter $p = 1 - e^{-\lambda\Delta}$, but with different support from that of the geometric distribution. The support here is the set of values say $z_0, z_1, z_2, \ldots$ chosen from each subinterval $[k\Delta, (k+1)\Delta)$ for $k = 0, 1, 2, \ldots$. It can be easily checked that $H(Z) = 1 - \log\lambda < \infty$ and $H(Z^\Delta) \to \infty$ as $\Delta \to 0$. How can this be interpreted from a statistical point of view?

On the other hand, for the average entropy, $A(Z^\Delta) \to A(Z) = 1 - \log 2$. This is exactly the meaning of consistency for the average entropy, which is proved in general in Theorem 3.1 in Kittaneh et al. (2016).

- Finally, the author(s) indicate that the average entropy is constant for particular distributions.

In fact, many entropies (including Shannon entropy) are free of the distribution's parameters for some probability distributions. Consider the following triangular distribution with parameter $(0 < c < 2)$

$$f(x) = \begin{cases} x/c, & 0 < x < c \\ (2-x)/(2-c), & c < x < 2 \end{cases}, \quad \text{and zero otherwise.}$$

It can be checked that Shannon entropy is equal to 1/2, which does not depend on the parameter c. At the end I would like to say that information theoreticians always look at uncertainty through the window of Shannon entropy. Whereas, considering entropy as a statistical measure requires its validity for both discrete and continuous distributions, in addition to the consistency when moving between the two spaces.

References

Kittaneh, A. O., Khan, M. A., Akbar, M., and Bayoud, H. A. (2016), "Average Entropy: A New Uncertainty Measure with Application to Image Segmentation," *The American Statistician*, 70, 18–24. [91]

Tsallis, C. (1988), "Possible Generalization of Boltzmann-Gibbs Statistics," *Journal of Statistical Physics*, 52, 470–487. [91]

Omar A. Kittaneh
Electrical and Computer Engineering Department
Effat University, Jeddah, Saudi Arabia

CONTACT Omar A. Kittaneh *okitanneh@effatuniversity.edu.sa* Electrical and Computer Engineering Department, Effat University, Jeddah 22332, Saudi Arabia.