LETTER

On the Bayesian interpretation of the harmonic mean *p*-value

Leonhard Held[a,1]

I read with much interest the article by Wilson (1) on the harmonic mean *p*-value (HMP) for combining statistical significance tests. I congratulate the author on a thorough discussion of this proposal. However, I would like to point out that Good (2) had suggested the HMP already in 1958 (see also refs. 3 and 4 and references therein). Good (2) distinguishes tests in parallel from tests in series, where the latter lead to Fisher's method while the former lead to the HMP. Both Good (ref. 2, p. 804) and Wilson (section 5 of supporting information for ref. 1) use the theorem of weighted averages of Bayes factors (5) to derive the HMP. Good's argument is based on his empirical observation that the Bayes factor against the null hypothesis is approximately equal to $1/(\gamma p)$, where p is the *p*-value and γ is between 3⅓ and 30. The weighted average of these Bayes factors then leads to the weighted HMP. However, this approximation is quite crude since the Bayes factor is not necessarily monotonically related to the *p*-value (section 3 of ref. 6) and can even support the null hypothesis when a *p*-value would lead to its rejection (section 4.4 of ref. 7).

Wilson uses a Beta($\xi < 1$, 1) distribution for the *p*-value under the alternative to derive optimal weights for the HMP. This class is suitable for local alternatives and leads to the popular $-1/\{ep\log(p)\}$ upper bound on the Bayes factor against the null (8). However, the derivation in Wilson (1) requires that all tests have good power, where Bayes factors based on simple alternatives are more appropriate and may give larger Bayes factors than $-1/\{ep\log(p)\}$ (6) (Fig. 1). Similar results can be obtained for one-sided *p*-values using the pCalibrate R package.

A promising alternative for well-powered studies is the class of Beta(1, $\kappa > 1$) distributions (section 2.3 of ref. 6). The Bayes factor against the null is then bounded by $-1/\{eq\log(q)\}$ where $q = 1 - p$. This bound is also shown in Fig. 1 and is always above the bounds for the Bayes factors based on simple alternatives. For $p < 0.1$ the $-1/\{eq\log(q)\}$ bound can be well approximated by $1/(ep)$, where $e \approx 2.72$ is remarkably close to Good's lower limit 3⅓ for γ. This bound is a monotone function of the *p*-value and suggests a modification of Good's argument to justify the HMP: If $1/(ep_i)$ is an upper bound for the Bayes factor for all *p*-values p_i considered, then

$$\text{maxBF} = \frac{1}{e}\sum_i \frac{\mu_i}{p_i}$$

is an upper bound for the model-averaged Bayes factor, where μ_i are the prior probabilities of the alternatives. Direct transformation of the HMP $\overset{\circ}{p}$ with weights μ_i to $1/(e\overset{\circ}{p})$ would give the same bound, which shows that the HMP with weights μ_i is compatible with an evidential interpretation of *p*-values using bounds on the Bayes factor.

[1]Email: leonhard.held@uzh.ch.

Fig. 1. Bounds on the Bayes factor against the null as a function of the two-sided *p*-value. The bounds for small sample size *n* are based on the *t*-distribution, for large *n* on the standard normal distribution. Good's range of Bayes factors $1/(\gamma\,p)$ where $3\frac{1}{3} < \gamma < 30$ is represented in gray.

1 Wilson DJ (2019) The harmonic mean *p*-value for combining dependent tests. *Proc Natl Acad Sci USA* 116:1195–1200.
2 Good IJ (1958) Significance tests in parallel and in series. *J Am Stat Assoc* 53:799–813.
3 Good IJ (1983) *Good Thinking: The Foundations of Probability and Its Applications* (Univ of Minnesota Press, Minneapolis).
4 Good IJ (1992) The Bayes/non-Bayes compromise: A brief review. *J Am Stat Assoc* 87:597–606.
5 Good IJ (1950) *Probability and the Weighing of Evidence* (Griffin, London).
6 Held L, Ott M (2018) On p-values and Bayes factors. *Annu Rev Stat Appl* 5:393–419.
7 Spiegelhalter DJ, Abrams KR, Myles JP (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* (Wiley, New York).
8 Sellke T, Bayarri MJ, Berger JO (2001) Calibration of p values for testing precise null hypotheses. *Am Stat* 55:62–71.