WHEN IS A HARMONIC MEAN p-VALUE A BAYES FACTOR?

DANIEL J WILSON

Big Data Institute, Nuffield Department of Population Health, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, United Kingdom

I welcome this opportunity [1] to acknowledge Good's papers [2, 3, 4, 5, 6], which I had missed. Good proposed the harmonic mean p-value (HMP) as a classical analog to a model-averaged Bayes factor (BF) which "should be regarded as an approximate tail-area probability [p-value]" [2]. His presentation was amusingly apologetic, e.g. "an approximate rule of thumb is tentatively proposed in the hope of provoking discussion" and "this rule of thumb should not be used if the statistician can think of anything better to do" [2]. I believe my paper dispels these misgivings by formalizing Good's intuitive argument that the HMP is approximately well-calibrated when small (Eq. 5, [7]) and deriving an asymptotically exact test for general use (Eq. 4, [7]). Further, I showed the HMP is a multilevel test procedure (Eq. 6, [7]), demonstrating with examples that it consequently provides a powerful alternative to Bonferroni and Benjamini-Hochberg [8] correction for large-scale multiple testing problems.

Held [1] considers the Bayesian properties of the HMP, which are relevant to its interpretation and power. Applications of the HMP are not limited to model selection problems, however, as it provides a general alternative to Fisher's method [9] for combining tests that are not independent [2, 7].

Good claimed that the HMP is inversely proportional to a model-averaged BF based on his empirical observations that

$$\mathrm{BF} \approx 1/(\gamma\, p), \quad 3\tfrac{1}{3} < \gamma < 30. \tag{1}$$

As he noted, this empirical relationship holds only approximately, it holds better for small p, and γ is not strictly constant in p [2].

Good's claim depends on the density of p-values under the alternative, $f(p|\mathcal{M}_A)$. Random variables with distribution functions regularly varying at zero (RVRV_0) [10] appear to capture Good's empirical observations, producing

$$\mathrm{BF} = \frac{f(p|\mathcal{M}_A)}{f(p|\mathcal{M}_0)} = f(p|\mathcal{M}_A) = \xi\, S(p)\, p^{\xi-1}, \qquad 0 < \xi. \tag{2}$$

This relationship is approximately inversely proportional for small p and tail index $\xi < 1$, but may deviate from strict proportionality through the slowly varying function $S(p)$. Thus the model-averaged BF with prior model probabilities $\mu_1 \ldots \mu_L$ would be

$$\overline{\mathrm{BF}} = \sum_{i=1}^{L} \mu_i \mathrm{BF}_i \approx \sum_{i=1}^{L} (\mu_i \mathrm{BF}_i)^{\frac{1}{1-\xi_i}} = \bar{\xi} \sum_{i=1}^{L} w_i/p_i = \bar{\xi}\, \mathring{p}^{-1}, \tag{3}$$

with weights $w_i = u_i/\bar{\xi}$,

$$u_i = (\mu_i\, \xi_i\, S(p_i))^{\frac{1}{1-\xi_i}}, \tag{4}$$

$\bar{\xi} = \sum_{i=1}^{L} u_i$ and HMP $\mathring{p}$. In the special case that $p|\mathcal{M}_A \sim \mathrm{Beta}(\xi < 1, 1)$, then $S(p) \equiv 1$, and Good's empirical relationship would be considered to hold closely for high-powered tests ($\xi \ll 1$) with $\gamma = \xi^{-1}$.

Held [1] considers whether the class of alternatives $p|\mathcal{M}_A \sim \mathrm{Beta}(1, \kappa > 1)$ supports Good's claim. This is an interesting proposition but I have some reservations. The distribution produces a special case of a RVRV_0 (Eq. 2), in which $\xi = 1$ and $S(p) = \kappa(1-p)^{\kappa-1}$. This yields the relationship $\mathrm{BF} = S(p)$, meaning the BF is a slowly-varying function of p (at zero). Held states that $-1/\{\mathrm{e}\,(1-p)\log(1-p)\} \approx 1/(\mathrm{e}\,p), p < 0.1$, is

an upper bound on this BF. My reservations are first, that a BF slowly varying in p is inconsistent with Good's empirical observations. Second, Held's bound is a regularly-varying function of p, making it a loose bound on the slowly-varying BF for small p and imperfect power ($\kappa < \infty$) (Fig. 1). In conclusion, for BFs slowly varying in p, Good's claim that the HMP is inversely proportional to the model-averaged BF does not hold. Rather than supporting the Bayes factor interpretation of the HMP, Held's example is valuable in showing where it breaks down.

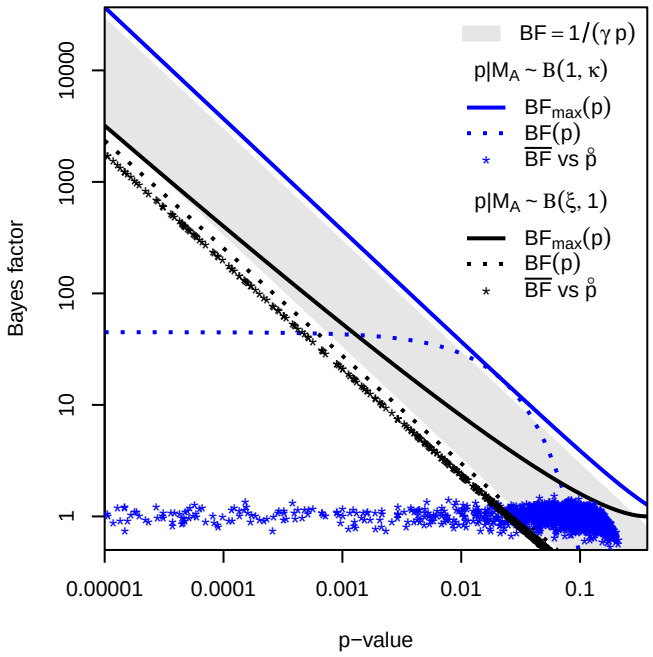

FIGURE 1. The HMP is strongly inversely proportional to $\overline{\text{BF}}$ when $p|\mathcal{M}_A \sim \text{Beta}(\xi < 1, 1)$ but not when $p|\mathcal{M}_A \sim \text{Beta}(1, \kappa > 1)$, despite the maximum BF argument. $\text{BF}(p)$ was calculated from each beta density and $\text{BF}_{\max}(p)$ from the respective upper bound [1, 11]. $\overline{\text{BF}}$ and $\mathring{p}$ were calculated by simulation: $L = 1000$ p-values per simulation, $(L-1)$ Uniform(0,1) and one Beta(ξ,1), Beta(1,κ) or Uniform(0,1) with equal probability, assuming equal weights for $\mathring{p}$. $\xi = 0.0352$ and $\kappa = 44.9$ were chosen to achieve 90% test power at $\alpha = 0.05$.
R code: https://doi.org/10.6084/m9.figshare.7699955

Acknowledgments

D.J.W. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant 101237/Z/13/Z). D.J.W. is supported by a Big Data Institute Robertson Fellowship.

References

[1] L. Held, "On the Bayesian interpretation of the harmonic mean p-value," *Proceedings of the National Academy of Sciences*, 2019.

[2] I. J. Good, "Significance tests in parallel and in series," *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 799–813, 1958.

[3] I. J. Good, "C192. One tail versus two-tails, and the harmonic-mean rule of thumb," *Journal of Statistical Computation and Simulation*, vol. 19, no. 2, pp. 174–176, 1984.

[4] I. J. Good, "C193. Paired versus unpaired comparisons and the harmonic-mean rule of thumb," *Journal of Statistical Computation and Simulation*, vol. 19, no. 2, pp. 176–177, 1984.

[5] I. J. Good, "C213. a sharpening of the harmonic-mean rule of thumb for combining tests "in parallel"," *Journal of Statistical Computation and Simulation*, vol. 20, no. 2, pp. 173–176, 1984.

[6] I. J. Good, "C214. the harmonic-mean rule of thumb: some classes of applications," *Journal of Statistical Computation and Simulation*, vol. 20, no. 2, pp. 176–179, 1984.

[7] D. J. Wilson, "The harmonic mean p-value for combining dependent tests," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1195–1200, 2019.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[9] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, fifth ed., 1934.

[10] T. Mikosch, *Regular variation, subexponentiality and their applications in probability theory*. Eindhoven University of Technology, Eindhoven, The Netherlands, 1999.

[11] T. Sellke, M. J. Bayarri, and J. O. Berger, "Calibration of p values for testing precise null hypotheses," *The American Statistician*, vol. 55, no. 1, pp. 62–71, 2001.