I would like to issue the following correction to users of the harmonic mean p-value (HMP), with apologies: The paper (Wilson 2019 *PNAS* 116: 1195-1200) erroneously states that the following asymptotically exact test controls the strong-sense family-wise error rate for any subset of p-values $\mathcal{R}$:

$$\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_{|\mathcal{R}|}\, w_{\mathcal{R}}$$

when it should read

$$\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_{L}\, w_{\mathcal{R}}$$

where:

- L is the total number of individual p-values.
- $\mathcal{R}$ represents any subset of those p-values.
- $\overset{\circ}{p}_{\mathcal{R}} = \left(\sum_{i\in\mathcal{R}} w_i\right) / \left(\sum_{i\in\mathcal{R}} w_i/p_i\right)$ is the HMP for subset $\mathcal{R}$.
- w_i is the weight for the ith p-value. The weights must sum to one: $\sum_{i=1}^{L} w_i = 1$. For equal weights, $w_i = 1/L$.
- $w_{\mathcal{R}} = \sum_{i\in\mathcal{R}} w_i$ is the sum of weights for subset $\mathcal{R}$.
- $|\mathcal{R}|$ gives the number of p-values in subset $\mathcal{R}$.
- $\alpha_{|\mathcal{R}|}$ and α_L are significance thresholds provided by the Landau distribution (Table 1).

In version 2.0 of the `harmonicmeanp` R package, the main function `p.hmp` is updated to take an additional argument, `L`, which sets the total number of p-values. If argument `L` is omitted, a warning is issued and `L` is assumed to equal the length of the first argument, `p`, preserving previous behaviour. Please update the R package.

An updated tutorial is available as a vignette in the R package and online here: http://www.danielwilson.me.uk/harmonicmeanp/hmpTutorial.html

Why does this matter?

The family-wise error rate (FWER) controls the probability of falsely rejecting any null hypotheses, or groups of null hypotheses, when they are true. The strong-sense FWER maintains control even when some null hypotheses are false, thereby offering control across much broader and more relevant scenarios.

Using the more lenient threshold $\alpha_{|\mathcal{R}|}$ rather than the corrected threshold α_L, both derived via Table 1 of the paper from the desired ssFWER α, means the ssFWER is not controlled at the expected rate.

Tests with small numbers of p-values are far more likely to be affected in practice. In particular, individual p-values should be assessed against the threshold α_L/L when the HMP is used, not the more lenient α_1/L nor the still more lenient α/L (assuming equal weights). This shows that there is a cost to using the HMP compared to Bonferroni correction in the evaluation of individual p-values. For one billion tests $\left(L = 10^9\right)$ and a desired ssFWER of $\alpha = 0.01$, the fold difference in thresholds from Table 1 would be $\alpha/\alpha_L = 0.01/0.008 = 1.25$.

However, it remains the case that HMP is much more powerful than Bonferroni for assessing the significance of *groups* of hypotheses. This is the motivation for using the HMP, and combined tests in general, because the power to find significant *groups* of hypotheses will be much higher than the power to detect significant *individual* hypotheses when the total number of tests (L) is large and the aim is to control the ssFWER.

How does it affect the paper?

I have submitted a request to correct the paper to *PNAS*. It is up to the editors whether to agree to this request. A copy of the published paper, annotated with the requested corrections, is available here: http://www.danielwilson.me.uk/files/wilson_2019_annotated_corrections.pdf. Please use Adobe Reader to properly view the annotations and the embedded corrections to Figures 1 and 2.

Where did the error come from?

Page 11 of the supplementary information gave a correct version of the full closed testing procedure that controls the ssFWER (Equation 37). However, it went on to erroneously claim that "one can apply weighted Bonferroni correction to make a simple adjustment to Equation 6 by substituting $\alpha_{|\mathcal{R}|}$ for α." This reasoning would only be valid if the subsets of p-values to be combined were pre-selected and did not overlap. However, this would no longer constitute a flexible multilevel test in which every combination of p-values can be tested while controlling the ssFWER. The examples in Figures 1 and 2 pursued multilevel testing, in which the same p-values were assessed multiple times in subsets of different sizes, and in partially overlapping subsets of equal sizes. For the multilevel test, a formal shortcut to Equation 37, which makes it computationally practicable to control the ssFWER, is required. The simplest such shortcut procedure is the corrected test

$$\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_L \, w_{\mathcal{R}}$$

One can show this is a valid multilevel test because if

$$\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_L \, w_{\mathcal{R}}$$

then

$$\overset{\circ}{p} = \left(w_{\mathcal{R}} \, \overset{\circ}{p}_{\mathcal{R}}^{-1} + w_{\mathcal{R}'} \, \overset{\circ}{p}_{\mathcal{R}'}^{-1} \right)^{-1} \leq w_{\mathcal{R}}^{-1} \, \overset{\circ}{p}_{\mathcal{R}} \leq \alpha_L$$

an argument that mirrors the logic of Equation 7 for direct interpretation of the HMP, which is not affected by this correction.

More information

For more information please leave a comment below, or get in touch via the contact page.