Important: Please update to **version 3.0** of the `harmonicmeanp` R package to address critical errors in the main function `p.hmp`.

I would like to update the correction I issued on July 3, 2019 to cover a second error I discovered that affects the main function of the R package, `p.hmp`. There are two errors in the original paper:

1. The paper (Wilson 2019 *PNAS* 116: 1195-1200) erroneously stated that the test $\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_{|\mathcal{R}|}\, w_{\mathcal{R}}$ controls the strong-sense family-wise error rate asymptotically when it should read $\overset{\circ}{p}_{\mathcal{R}} \leq \alpha_{L}\, w_{\mathcal{R}}$.
2. The paper incorrectly stated that one can produce adjusted p-values that are asymptotically exact, as intended in the original Figure 1, by transforming the harmonic mean p-value with Equation 4 before adjusting by a factor $1/w_{\mathcal{R}}$. In fact the harmonic mean p-value must be multiplied by $1/w_{\mathcal{R}}$ before transforming with Equation 4.

The `p.hmp` function prior to version 3.0 of the R package was affected by these errors.

In the above,

- 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).

Version 2.0 (released July 2019) of the `harmonicmeanp` R package only addressed the first error, which is why I am now releasing version 3.0 to address both errors. Compared to version 1.0, the main function `p.hmp` has been 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 earlier behaviour. **Please update the R package to version 3.0**.

The tutorial, available as a vignette in the R package and online, is affected quantitatively by both errors, and has been extensively updated for version 3.0.

The second error affects only one line of the corrected paper (issued July 2019). I have updated it to address the second error and two typos in Figure legends 1 and 2: http://www.danielwilson.me.uk/files/wilson_2019_annotated_corrections.v2.pdf. You will need Adobe Reader to properly view the annotations and the embedded corrections to Figures 1 and 2.

I would like to deeply apologise to users for the inconvenience the two errors have caused.

More information follows under the headings:

- **Why does this matter?**
- **How does it affect the paper?**
- **Where did the errors come from?**
- **How do I update the R package?**
- **What if I have already reported results?**

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 than the weak-sense FWER.

The ssFWER is not controlled at the expected rate if:

1. The more lenient threshold $\alpha_{|\mathcal{R}|}$ is used rather than the corrected threshold α_L, both derived via Table 1 of the paper from the desired ssFWER α.
2. Raw p-values are transformed with Equation 4 before adjusting by a factor $w_{\mathcal{R}}^{-1}$, rather than adjusting the raw p-values by a factor $w_{\mathcal{R}}^{-1}$ before transforming with Equation 4.

The `p.hmp` function of the R package suffers both issues in version 1.0, and the second issue in version 2.0. Please update to version 3.0.

Tests in which significance is marginal or non-significant at $\alpha = 0.05$ are far more likely to be affected in practice.

Regarding error 1, individual p-values need to 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 (and indeed small groups of 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 more powerful than Bonferroni for assessing the significance of large *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 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.v2.pdf. Please use Adobe Reader to properly view the annotations and the embedded corrections to Figures 1 and 2.

Where did the errors come from?

Regarding the first error, 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 (an approximate procedure), which is not affected by this correction.

The second error, which was also caused by carelessness on my part, occurred in the main text in the statement "(Equivalently, one can compare the exact *p*-value from Eq. 4 with $\alpha \, w_{\mathcal{R}}$.)" I did not identify it sooner because the corrected version of the paper no longer uses Equation 4 to transform *p*-values in Figure 1.

How do I update the R package?

The R package is maintained at https://cran.r-project.org/package=harmonicmeanp.

In R, test whether you have version 3.0 of the package installed as follows:
```
packageVersion("harmonicmeanp")
```

The online binaries take a few days or weeks to update, so to ensure you install the most recent version of the package install from source by typing:
```
install.packages("harmonicmeanp", dependencies=TRUE, type="source")
```

You can additionally specify the CRAN repository for example:
```
install.packages("harmonicmeanp", dependencies=TRUE, type="source",
repos="https://www.stats.bris.ac.uk/R")
```

After installation, check again the version number:
```
stopifnot(packageVersion("harmonicmeanp")>=3.0)
```

What if I have already reported results?

I am very sorry for inconvenience caused in this case.

As long as the 'headline' test was significant with `p.hmp` under R package versions 1.0 or 2.0, then the weak-sense FWER can be considered to have been controlled. The 'headline' test is the test in which all *p*-values are included in a single combined test. The headline test is not affected by either error, because $|\mathcal{R}| = L$ and $w_{\mathcal{R}} = 1$. The headline test controls the weak-sense FWER, and therefore so does a two-step procedure in which subsets are only deemed significant when the headline test is significant (Hochberg and Tamhane, 1987, *Multiple Comparison Procedures*, p. 3, Wiley).

If the headline test was not significant, re-running the analysis with version 3.0 will not produce significant results either because the stringency is greater for controlling the strong-sense FWER. If the headline test was significant, you may wish to reanalyse the data with version 3.0 to obtain strong-sense FWER control, because this was the criterion the HMP procedure was intended to control.

If some results that were significant under version 1.0 or 2.0 of the R package are no longer significant, you may conclude they are not significant or you may report them as significant subject to making clear that only the weak-sense FWER was controlled.

More information

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