Meta analysis of *p*-values

E-mail distributed on 27-10-2025

Dear all,

Meta analysis of *p*-values is generally **not advised**. That is, when the interest is to synthesize multiple studies quantitatively, one should always prefer effect sizes (and their standard errors) over any other statistic, as this will provide the most accurate information on the size, direction, and variance of the effect of interest. *P*-values are immediately flawed for this purpose, since they quantify none of these aspects, only inferential significance.

Nevertheless, the idea to synthesize *p*-values has historically been proposed in statistical literature, including from perspectives other than meta analysis, such as the detection of *p*-hacking. For this mailing I wanted to give a brief overview of these perspectives and discuss their (relative) merits. As per my lead, this is not meant to be an endorsement, merely an overview. I also note that many of the methods cited below share conceptual overlap with multiple comparison corrections, where *p*-values are "combined" in the sense that one *p*-value is adjusted for the presence of others. However, I will omit discussion of the multiple comparisons problem since I have previously addressed it in another mailing.

1. Combining inferences

Suppose we ran two independent studies on the same effect and found $p < 0.0001$ and $p = 0.0600$ (for comparable sample sizes and error variance). While the second effect is not significant at $\alpha = 0.05$, it would seem that the combined *p*-values still suggest overall significance. Fisher (1925) was the first to address this question, proposing a simple formula for aggregating *p*-values (i.e., **Fisher's method**), and showing that the resulting statistic is chi-square distributed with $2 \times k$ degrees of freedom, with k the number of *p*-values. An inferential test on this statistic evaluates a *global null hypothesis*, i.e., that all *p*-values are non-significant, versus at least one significant.

Two drawbacks of Fisher's method are that the *p*-values cannot be weighted, and that they are assumed to be independent. Weighting is attractive to allow heterogeneity between studies, for example, by sample size, measurement error, or some other property. While many variations have been proposed, the most prominent is the **Stouffer-Lipták method** (Stouffer et al., 1949; Lipták, 1958), which first back-transforms the *p*-values to *z*-values, then calculates a weighted average of the *z*-values, and finally submits the average to a basic *z*-test. Weights should be chosen to reflect study heterogeneity, with Lipták recommending, in order of preference, **(a)** effect sizes, **(b)** inverse standard errors, or **(c)** root sample sizes. In R, the `transite`[1] package (Krismer et al., 2020) contains the

[1] Note that this package has to be downloaded through the Bioconductor platform, not CRAN.

function `p_combine` which performs a number of different combination tests, including the Stouffer-Lipták method, with optional use of weights. Alternatively, the `metap` package (Dewey, 2025) features similar functions within an explicit framework of meta analysis (see Section 4).

For dependent *p*-values, numerous extensions to Fisher's method have been proposed that operate either under a known dependence structure (Korn's method) or unknown dependence structure (harmonic mean *p*-value, Kost's method, Cauchy combination method). Implementations in R are scattered over many packages, with `mvMAPIT` (Stamp & Crawford, 2023) offering simultaneously Fisher's method, the harmonic mean *p*-value and the Cauchy combination method.[2] Of note is that many of these packages were developed from the field of genomics, where mass testing on thousands of gene expression features is a common challenge.

2. Bayesian updating

Although Fisher was a trenchant frequentist, one could argue there is a Bayesian intuition lurking behind the idea of combining *p*-values. Suppose this time that we ran ten independent studies intended to replicate the same effect, and found nine times in a row $p < 0.0001$, with only the tenth producing $p = 0.6243$. Despite the high *p*-value, many researchers would consider the earlier nine studies as strong **prior evidence** for a real effect, and use it to downweight the importance of the tenth result. Indeed, while frequentist statistics remain the dominant approach to data analysis of any given study, most researchers probably take a Bayesian perspective to interpret the accumulation of evidence across studies.

Although such reasoning makes sense, applying it to *p*-values would be non-sensical, as *p*-values are inherently frequentist. If the interest is to update current evidence in light of earlier evidence, one should rather opt directly for a Bayesian approach. In R, `BayesRep` (Pawel & Held, 2022) allows the calculation of a "replication Bayes factor" (function `BFr`), which incorporates the effect size of an earlier study as a prior to adjust the Bayes factor for a current study. Package `RBesT` (Weber et al., 2021) generalizes this even further by allowing current data to be adjusted for the entire history of empirical evidence (as opposed to just one earlier study), by taking a multi-stage approach in which the history of evidence is first summarized by a Bayesian meta analytic model, and then converted into a suitable prior for the analysis of the current data.

3. *P*-hacking

Under the null hypothesis, *p*-values have a uniform distribution. That is, when the effect is truly absent, one is equally likely to obtain, e.g., $p = 1.0000$ as $p = 0.0001$. This appears to be counterintuitive to many researchers, but can easily be verified, e.g., for the independent samples *t*-test:

```
p.null <- vector()
p.alt <- vector()
for(i in 1:1000) {
  p.null <- c(p.null,t.test(rnorm(100,0,1),rnorm(100,0,1))$p.value)
  p.alt <- c(p.alt,t.test(rnorm(100,0,1),rnorm(100,0.5,1))$p.value)
}
```

[2] For a weighted version of the Cauchy combination test see the `CCT` function in package `GRAB`.

```
par(mfrow=c(1,2)) ; hist(p.null) ; hist(p.alt,xlim=c(0,1))
```

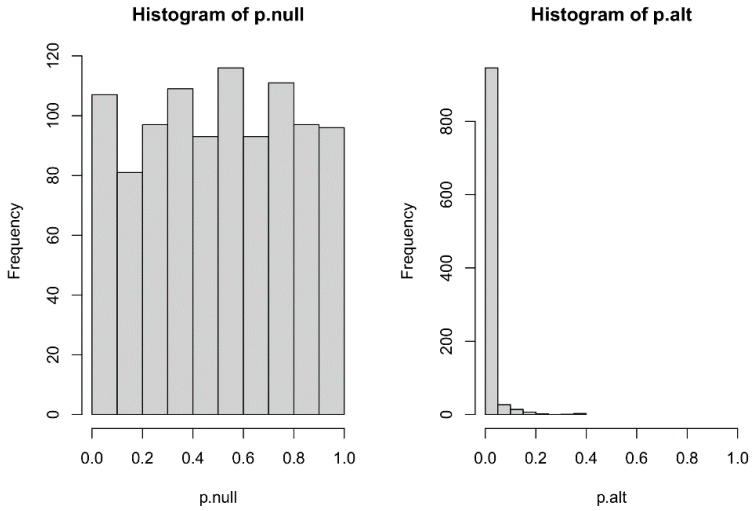

Figure 1. Distribution of *p*-values of a *t*-test under the null hypothesis (left) and the alternative hypothesis (right).

P-value combination methods—including Fisher's method—typically rely on this property in deriving the distribution of their test statistic, hence its assumption of a global null hypothesis. Another area where this reasoning has been applied is for the detection of publication bias or *p*-hacking. Two such methods that have gained some popularity in recent years are the ***p*-curve** and ***p*-uniform** methods (Van Aert, Wicherts, & van Assen, 2016).

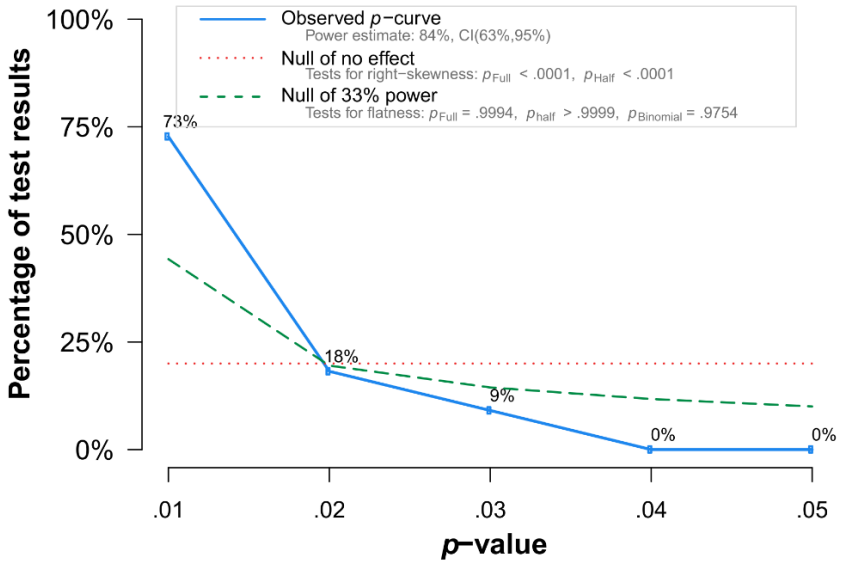

Figure 2. Example of a *p*-curve analysis from the `dmetar` package.

Both of these methods consider only significant *p*-values for synthesis, and test whether their distribution deviates from a uniform distribution. If not, it may constitute evidence of "*p*-hacking", although I generally recommend avoiding that term, since bias in *p*-values may be present for many

reasons other than deliberate manipulation or fraud. In R, the package `dmetar`[3] (Harrer et al., 2019) offers the function `pcurve` to return *p*-curve and *p*-uniform tests (see Figure 2).The *p*-uniform method differs from the *p*-curve in that it is also used to derive effect sizes and confidence intervals for meta analysis, whereas *p*-curve is purely an analysis of bias. Nevertheless, caution should be exercised when applying these methods, as they have been criticized on conceptual, statistical and practical grounds (Morrey & Davis-Stober, 2025). Like many distributional tests, their statistical properties are poor under many conditions (e.g., small samples, large samples, heterogeneity of effects). For the purpose of assessing publication bias in meta analyses, priority should be given to visual methods such as funnel plot inspection (Afonso et al., 2023).

4. Meta analysis

Finally, there have been proposals for the formal meta analysis of *p*-values within the meta analysis framework, including many discussed earlier (Hedges, 1992). This has been motivated primarily for situations where **(a)** the effects of interest are too heterogeneous (or incomparable) across studies, or **(b)** the desired statistics are missing and only *p*-values are systematically reported. However, even in these situations, meta analysis of *p*-values should be considered a last resort. Heterogeneity should be modelled rather than ignored (e.g., with meta regression), and rules-of-thumb could be applied to transform incomparable effect sizes to a common metric (e.g., odds ratio to Pearson correlation), for example as implemented in the R package `effectsize` (Ben-Shachar, Lüdecke & Makowski, 2020). Moreover, if effects across studies are considered to be incomparable, one should reflect on whether coherent quantitative conclusions can be drawn at all from a meta analysis. In this case, a purely qualitative review may be a better choice for a study.

If truly no other statistics but *p*-values are available, one should still keep in mind their intrinsic weaknesses. Foremost, *p*-values only quantify significance, not the size or direction of an effect. In fact, the use of *p*-values for meta analysis runs somewhat contrary to meta analytic philosophy, in that its purpose is generally descriptive rather than inferential. The goal is normally to quantify the effect of interest and its precision (with a confidence interval) without regard to its overall significance. Secondly, *p*-values are sensitive to sample size, and therefore not comparable between small and very large samples. In order to make sense, sample sizes should be incorporated into the meta analysis (e.g., as in the Stouffer-Liptak method). Finally, the insight from combination tests such as described in Section 1 is rather limited, as a significant result only reflects that *at least one* effect is non-zero.

In R, the package `metap` (Dewey, 2025) collects many of the methods discussed in this newsletter, as well as plotting and interpretation functions. The package vignettes clarify the theoretical background and how to use and interpret these functions practically. Those who have run meta analysis on effect sizes (e.g., with the metafor package) will find this framework vastly more limited than the modern approach with effect sizes and random effects models. It is best recommended as a back-up analysis or as an intermediate step for deriving approximate effect sizes.

For a more thorough review on the meta analysis of *p*-values and its drawbacks, I strongly recommend the paper by Hedges (1992), which remains an extremely lucid and accessible treatment of the subject.

[3] Not (yet) available on CRAN. Must be loaded externally from Github.

References

Afonso, J., Ramirez-Campillo, R., Clemente, F. M., Büttner, F. C., & Andrade, R. (2024). The perils of misinterpreting and misusing "publication bias" in meta-analyses: An education review on funnel plot-based methods. *Sports medicine*, *54*(2), 257–269.

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56), 2815.

Dewey, M. (2025). *metap: Meta-Analysis of Significance Values*. R package version 1.12.

Fisher R. A. (1925). *Statistical Methods for Research Workers (1st ed.).* London, England: Oliver & Boyd.

Harrer, M., Cuijpers, P., Furukawa, T. & Ebert, D. D. (2019). *dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R'.* R package version 0.1.0. URL http://dmetar.protectlab.org/

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, 17(4), 279–296.

Krismer, K., Bird, M.A., Varmeh, S., Handly, E.D., Gattinger, A., Bernwinkler, T., Anderson, D.A., Heinzel, A. Joughin, B.A., Kong, Y.W., Cannell, I.G., & Yaffe, M.B. (2020). Transite: A computational motif-based analysis platform that identifies RNA-binding proteins modulating changes in gene expression. *Cell Reports*, 32(8), 108064.

Lipták, T. (1958). On the combination of independent tests = független mozgó szintes próbák összevont értékeléséről. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, *3*(3-4), 171–197.

Morey, R. D., & Davis-Stober, C. P. (2025). On the poor statistical properties of the P-curve meta-analytic procedure. *Journal of the American Statistical Association*, 1–19.

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(3), 879–911.

Stamp ,J., & Crawford, L. (2023). *mvMAPIT: Multivariate Genome Wide Marginal Epistasis Test*. R package version 2.0.3.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The american soldier: adjustment during army life. *Studies in Social Psychology in World War II, Vol. 1*.

Van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11(5), 713–729.

Weber, S., Li, Y., Seaman, J.W., Kakizume, T., & Schmidli, H. (2021). Applying meta-analytic-predictive priors with the R bayesian evidence synthesis tools. *Journal of Statistical Software*, 100(19), 1–32.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79