Review of Vermeesch, "Too tood to be true: Underdispersion in geochronology"

The article presents a comprehensive discussion of underdispersion in the geochronology setting. Discounting cases where underdispersion may be caused by incorrect error propagation (as that is in some cases trivial, and in general already handled by analysis packages), the paper concentrates on outlier removal as the critical phase where underdispersion may be (unwillingly) introduced in the attempt of removing what is perceived as problematic overdispersion, and thus produce overly confident results.

The paper provides compelling arguments that common outlier removal practices may lead to biased distribution of p-values, if performed uncritically. These are supported by analyses of the distribution of p-values from published studies, which reveal that a non-trivial proportion of studies is underdispersed.

The paper shows that plots of p-values vs their empirical cumulative distribution can allow to identify presence of underdispersed samples in meta-analyses of published results. Unfortunately, the same approach cannot be applied by a single practitioner on their study. Perhaps this point should be stressed in the introduction of the paper.

In the biomedical community it is quite common to present sensitivity studies in a research paper, so that the possible dependence of results on some of the assumptions made (for example, imputation strategies or the definition of subgroups) can be evaluated, at least so some extent. Such sensitivity analyses would provide some reassurance that conclusions are not affected by the outlier removal strategy applied.

We hope that the insights provided by the paper will be embraced by the community, through careful application of outlier removal steps, sensitivity analyses and provision of the raw data, so that the extent of outlier removal can be assessed by people external to the work.

Minor comments:

- It may help to provide a definition of statistical power. Given the role it plays in the discussion (especially regarding "powered tests"), having a precise definition would set a firmer ground. I think it's worth spelling out very clearly that there is a direct relationship between statistical power and sample size, so that removal of samples leads to loss of statistical power.
- The MSWD ranges at lines 77–79 are (approximately) valid at $\alpha = 0.05$, not for an arbitrary confidence level, so for clarity that should be stated.
- The arguments presented at the start of Section 5 don't consider correction for multiple testing. Even a plain Bonferroni correction (where the significance threshold is adjusted by dividing it by the number of tests performed) can be enough to correct the distortion in the probability of rejecting a test. With that, the numeric example of line 132 ($\alpha = 0.05$, $n = 14$) evaluates to $1 - (1 - \alpha/n)^n \approx 0.05$ as expected. It may be worth mentioning the issue, as the more tests are performed, the higher the probability of committing a Type I error is (and thus falsely concluding that a sample is overdispersed).
- Line 280: this statement is not clear: "More extreme degrees of underdispersion require that the p-value cutoff is raised from the usual value of $\alpha = 0.05$ to higher values (e.g $\alpha = 0.5$)". As phrased, it seems to imply causation between underdispersion and α (while arguably it's the reverse). Perhaps was it meant that extreme underdispersion corresponds effectively to using $\alpha \gg 0.05$?
- Line 283: the α values corresponding to a MSWD cutoff of 1 would be more clearly derived by reminding that, as MSWD $= \chi^2_\nu/\nu$, then MSWD ≤ 1 corresponds to finding the probability that $\chi^2_\nu \leq \nu$ (where for simplicity $\nu = n - 1$ is taken as the number of degrees of freedom in the numerical examples presented).
- Regarding the fourth point in the Discussion (section 8, lines 454-459): excessive outlier removal not only increases bias, it also decreases precision as the sample size is reduced. So to go back to the bias-variance tradeoff discussed in the two paragraphs before this, while it's possible to improve one term only to the detriment of the other, it's also possible to make both of them worse.

Other suggestions

- Figure 2: Describe the meaning of the arrows in the figure caption, along the lines of: "Arrows show the drift occurring for ratios that failed the homogeneity test before being adjusted".
- Figure 3 and 4: There is no description of the meaning of the green colour. Perhaps it would be clearer to just remove the points corresponding to outliers in the right-most panels, rather than leaving them in without colour fill.
- Figure 7: the caption mention a red curve, but the colour in the plot is yellow.

Typos

- Line 121: "underdisperseion" should read "underdispersion"
- Line 135: It may be clearer to write: "Under the null hypothesis of a non-dispersed [or homogeneous] sample . . . "
- Line 171: rewrite as ". . . is entirely determined by these two values, so that the uncertainties $s[N_s/N_i]$ . . . "
- Line 232: should read as "the intercept (a) and slope (b)"
- Line 240: should read as "where $\hat{a}$ and $\hat{b}$ are the best fit intercept and slope"
- Line 342: "Figure 1b" should be "Figure 8b".
- Line 375: "unscaleable" should be "unscalable".
- Line 405: "The blue ellipses" should presumably be "The red ellipses".

–

Marco Colombo
Heidelberg, July 2026