Post-hoc paradoxes

E-mail distributed on 28-07-2025

Dear all,

In a traditional (independent groups) ANOVA, the *F*-test evaluates whether at least one pair of group means differ significantly. However, sometimes the *F*-test turns out non-significant while one or more contrasts are significant. More generally, any omnibus test could disagree with its follow-ups. For the present mailing I want to briefly review such scenarios, why they occur, and what to do about them. Mathematically speaking none of these are truly paradoxes, since from statistical theory it is known exactly under which conditions these tests will disagree. The "paradox" here is more a problem of interpretation and reporting.

1. Omnibus test non-significant, follow-ups significant

Perhaps the most common disagreement is a non-significant omnibus test with some significant follow-ups (e.g., pairwise contrasts following a classical ANOVA). However, in many such cases, the *p*-values of the follow-ups would not survive a correction for multiple testing, in which case the problem of interpretation is sidestepped altogether. Moreover, the general rule is that follow-up tests should not be conducted or reported in any case when the omnibus test is non-significant.[1] That is, unless these follow-ups were the subject of explicit research hypotheses (i.e., so-called "planned" comparisons). Reporting them in this case is permitted though I strongly recommend to keep interpretation at a descriptive level. If the effect size is moderate or large, and/or the direction of the effect is at least consistent with expectation, then this is worth discussing, but inferential conclusions should be avoided and one should not overinterpret the findings in the paper's discussion section.

This disagreement could also be avoided by maintaining a more stringent significance level, such as 0.005. (Benjamin et al., 2017; Gordon et al., 2021; Bogdan, 2025). If the *p*-value of the omnibus test falls between 0.05 and 0.005, there may still be some merit in interpreting it as a trend effect, especially if some of the follow-ups do pass 0.005.

2. Omnibus test significant, follow-ups non-significant

It is possible also for the reverse paradox to occur, where an omnibus test is significant but none of the follow-ups. Interestingly, there are several different scenarios that can bring this about.

[1] The reverse logic of Fisher's least significant difference

Multiple comparison corrections. Most commonly, the paradox occurs because of multiple comparison corrections to *p*-values. When the omnibus test has a *p*-value already close to the significance threshold, it is very likely that none of the follow-ups will survive the multiple comparison correction. While it is possible that the initial omnibus test is a genuine false positive, there is also a legitimate concern that strict corrections may render the follow-ups into false negatives. The choice of correction should therefore be carefully considered. Under Fisher's LSD, the paradox would never occur in this scenario, since it permits uncorrected follow-ups after a significant omnibus test. Unfortunately this procedure may be too liberal, and it is easy to construct cases where it would definitely fail (e.g., mass testing such as in brain data). On the other hand, traditional corrections like Bonferroni are probably too strict, especially if the follow-up tests are not truly independent.[2] If the follow-up tests remain non-significant after the choice of an appropriate correction, one can optionally still interpret them as trend effects, heeding the same cautions from the previous section to not over-interpret results.

Subsetted contrasts. Another common version of this paradox occurs when researchers subset their data to perform pairwise comparisons. In large designs, doing so can substantially reduce degrees of freedom in the resulting *t*-tests, leading to a loss of power compared to the omnibus test. It is therefore recommended to run model-based comparisons whenever possible (e.g., using R packages like emmeans).[3] For example in a 10-group ANOVA with 30 participants per group, the model-based *t*-tests would have DF equal to 290, whereas a subsetted pairwise contrast would only have 58. Although it seems counterintuitive, the model-based contrast "borrows" power from the remaining groups, even when their means are not directly involved in the comparison.

Unbalanced groups. Even without multiple comparison corrections or subsetting of data, it is possible for an ANOVA to be significant and none of the pairwise follow-ups. This can occur when the omnibus effect is weak and the groups are severely unbalanced (e.g., one large group and two small ones). In this case, one interprets the results with the same cautions as in the scenario with multiple comparison corrections.

Unplanned contrasts. In large or complex factorial designs, researchers often restrict follow-ups to a limited number of planned contrasts. The omnibus test considers *all* contrasts, however, which means there is a possibility that the contrasts which contributed to the significant *F*-test are not among the planned ones. A similar scenario can occur for an interaction between a continuous and a categorical predictor (e.g., X×G). With the overall interaction test significant, one could choose to test X slopes within groups, or test group differences within *selected* values of X. While both approaches should be consistent with the significant omnibus test, in the second approach it is possible that values of X are selected where none of the group differences are significant. The point where the slopes diverge significantly may even be at the limits of the observed data range. In this case, one may have to conclude that, in a practical sense, the interaction is not actually present.

Multicollinearity. In linear regression, it may occur that the model's overall *F*-test is highly significant but none of the *t*-tests of the individual regressors. In this case the cause is multicollinearity between regressors. That is, even though some of them may be highly predictive of the outcome, collinearity between them inflates standard errors and reduces the individual *t*-values to non-

[2] Which they rarely are

[3] Especially in complex models like multilevel regression. Here it is crucial that the pairwise contrasts maintain the same random effects structure as the overall model, and that degrees of freedom follow the same correction rule.

significance. It is recommended therefore to routinely run collinearity diagnostics (e.g., variance inflation factors) on your regression models.

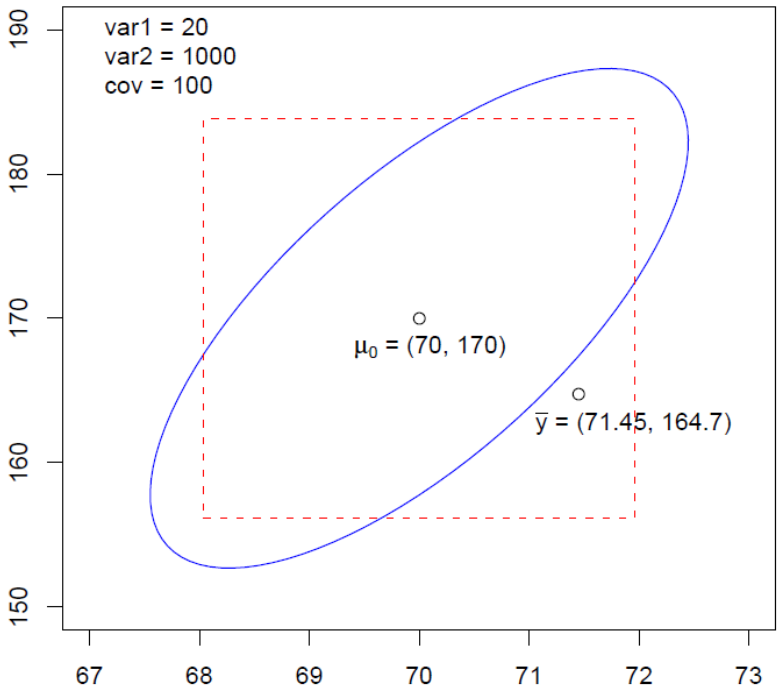

Figure 1. Multivariate versus univariate rejection region for a vector of observed means against a vector of null-hypothesis means (Rosseel, 2008).

3. MANOVA and ANOVA

A special case of the preceding two paradoxes is a multivariate omnibus test (e.g., MANOVA) followed by univariate tests for each outcome separately. Figure 1 depicts a case where we are comparing a vector of 2 observed means to a vector of 2 null-hypothesis means (Rosseel, 2008). Each mean could be compared univariately to its corresponding null-mean, or multivariately to both simultaneously. Crucially, the rejection regions of these two tests (red and blue) have a different shape, with non-overlapping parts in both directions. Whenever the observed means are within the non-overlapping regions, a disagreement between the tests will occur (e.g., within the square but not the ellipse the multivariate test will reject the null-hypothesis, whereas the univariate tests will not). Only when the vector of observed means exceeds both spaces will both tests reject the null hypothesis.

Figure 1 also illustrates that the degree of overlap partially depends on the strength of the correlation between the two outcomes, with generally less overlap for more strongly correlated outcomes. For uncorrelated variables, the multivariate rejection region would be circular in shape.

4. Differences in ordered means

Another paradox along similar lines is a case where we have a set of ordered means (A < B < C), where A-B is non-significant, B-C is non-significant, but A-C is significant. While we understand that the means

are too close in this case for the inner contrasts to reach significance, it does create something of a paradox of interpretation, since intuitively if A-B is equal and B-C is equal, then A-C should be equal. In mathematics this is known as transitivity. However, it does not hold for significance testing.

An additional complication occurs when B represents a neutral or baseline category. For example, A could be a drug to lower blood pressure, B a placebo, and C a drug to raise blood pressure. How could the drugs differ significantly if neither differs significantly from the placebo? In emotion this scenario is sometimes encountered with valence conditions (e.g., A=negative, B=neutral, and C=positive stimuli). Here again, how does one interpret a significant A-C contrast, when neither level is significantly different from "neutral"? As in the other paradox examples, the solution lies in cautious reporting. For the ordered means the direction of the effects will at least be consistent with expectations, so descriptively the result still makes sense. Most likely the paradox occurred because of a weak manipulation, or low power in small samples.

5. Skipping the omnibus test?

The above may call into question the utility of an omnibus test in certain scenarios. Indeed, some statisticians have suggested to skip overall tests completely, for example in the case of a traditional ANOVA, instead proceeding straight to pairwise comparisons (Wilcox, 2017).[4] However, I believe that in general they do have utility, and can contribute to reducing false positive results. In complex designs such as multi-way ANOVA, insisting on a significant omnibus test before proceeding to a lower level of testing can be an effective way to control the multiple testing problem. That said, there has been a historical tradition among researchers to apply this logic selectively. In multiple regression, the overall *F*-test of the model is often ignored, and instead tests on individual effects are immediately reported. Likewise, in the case of multiple outcome variables, a multivariate test could be run before proceeding to univariate tests, but this extra layer of testing is often omitted, or thought to be excessive.[5]

How many layers of omnibus testing one considers ultimately comes down to individual choice and is not inherently right or wrong. However, in other scenarios, the omission of an omnibus test is a real error. If an effect X was significant in group A but not group B, some researchers would conclude that the effects differed significantly, but such claims cannot be made without testing the X×Group interaction. Niewenhuis and colleagues (Niewenhuis, Forstmann, Wagenmakers, 2011) found that this mistake was surprisingly pervasive in many neuroscience papers (e.g., comparing effects between brain regions without testing the interaction of X×Region), and most likely in social sciences in general. In some of these cases the omission may have been intentional, because of a non-significant interaction test, though the authors of the aforementioned study speculate that more often the mistake was based on fallacious reasoning. As they noted, "when making a comparison between two effects, researchers should report the statistical significance of their difference rather than the difference between their significance levels."

---

[4] Under the caution that proper multiple comparison controls are implemented

[5] For example, consider a study with a 3×4 between-subjects design and 5 outcome variables. The first omnibus test could be a MANOVA of the 3×4 conditions on all outcomes. When significant, the second layer of omnibus tests would be 5 overall *F*-tests for each model separately. For any significant, the third layer of omnibus tests would be the 3×4 interaction for each model. For any significant, the fourth layer of omnibus tests would be 3-level ANOVAs within the 4 levels of the other factor, for each model. For any significant, finally, we would be allowed to run pairwise contrasts of the first factor's levels within the 4 levels of the other factor, for each model.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., … & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2(1)*, 6–10.

Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *Plos One*, *16(4)*, e0248780.

Bogdan, P.C. (2025). One decade into the replication crisis, how have psychological results changed? *Advances in Methods and Practices in Psychological Science*, *8(2)*.

Maxwell, S.E., & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective (2nd Edition)*. Mahwah, NJ: Larwence Erlbaum Associates.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14(9)*, 1105–1107.

Rosseel, Y. (2008). *Data Analysis II (2008–2009)*. Course syllabus. University of Ghent (Belgium).

Wilcox, R.R. (2017). *Understanding and Applying Basic Statistical Methods Using R*. Hoboken, NJ: Wiley

--

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