Statistics and Probability Letters 233 (2026) 110662

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Directional replicability: When can the factor of two be omitted

Vera Djordjilović [a,*], Tamar Sofer [b,c,d], Jonathan M. Dreyfuss [e]

[a] Department of Economics, Ca' Foscari University of Venice, Venice, Italy
[b] Cardiovascular Institute, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[c] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[d] Division of Sleep Medicine, Brigham and Women's Hospital, Boston, MA, USA
[e] Bioinformatics & Biostatistics Core, Joslin Diabetes Center, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Composite null hypotheses
Concordant replicability
Directional replicability
Order statistics
Partial conjunction

ABSTRACT

Directional replicability addresses the question of whether an effect studied across n independent studies is present with the same direction in at least r of them, for $r \geq 2$. When the expected direction of the effect is not specified in advance, the state of the art recommends assessing replicability separately by combining one-sided p-values for both directions (left and right), and then doubling the smaller of the two resulting combined p-values to account for multiple testing. In this work, we show that this multiplicative correction is not always necessary, and give a sufficient and necessary condition under which it can be safely omitted.

1. Introduction

Low replicability of scientific findings observed in medicine (Ioannidis, 2005), economics (Camerer et al., 2016) and psychology (Open Science Collaboration, 2015), has motivated great interest in developing formal statistical methods for evaluating replicability. We refer the interested reader to a recent review of statistical methodology for replicability analysis by Bogomolov and Heller (2023).

Consider a phenomenon that is studied in n independent studies. We assume that the object of interest can be represented with a scalar that we refer to as the *effect size* and we let $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ denote a vector of true effect sizes across studies. Let further $n^+ = |\{i : \theta_i > 0\}|$ represent the number of positive elements of θ, and similarly $n^- = |\{i : \theta_i < 0\}|$, the number of negative elements of θ. We are interested in testing the r out of n directional replicability null hypothesis, denoted by $H_{r/n}$, defined as

$$H_{r/n} : n^+ < r \wedge n^- < r,$$

for a given $r \leq n$ against a general alternative $K_{r/n} : n^+ \geq r \vee n^- \geq r$. Rejecting $H_{r/n}$ allows one to conclude that there are at least r effects of the same sign and thus claim r out of n directional replicability. Directional replicability imposes a stronger requirement than merely requiring an effect to be present in at least r studies (i.e. $n^+ + n^- \geq r$) since the latter does not require consistency in the sign of the effect. However, rejecting hypothesis $H_{r/n}$ does not imply complete consistency: consider the case $n = 4, r = 3$ with $n^+ = 3$, and $n^- = 1$. Then, the effect is positive in three studies and $H_{r/n}$ is false; nevertheless, the effect is negative in the fourth study so the sign of the effect is not consistent across studies.

The standard approach to testing $H_{r/n}$ is outlined in Owen (2009). The hypothesis $H_{r/n}$ is seen as the intersection of the two unilateral replicability hypotheses $H^+_{r/n} : n^+ < r$ and $H^-_{r/n} : n^- < r$. Each of the unilateral hypotheses is tested at the significance level $\alpha/2$, for a given $\alpha \in (0, 1)$, and the intersection hypothesis $H_{r/n}$ is rejected if either of the two is rejected.

* Corresponding author.
E-mail address: vera.djordjilovic@unive.it (V. Djordjilović).

https://doi.org/10.1016/j.spl.2026.110662
Received 17 October 2025; Received in revised form 24 January 2026; Accepted 26 January 2026
Available online 30 January 2026

The two unilateral replicability null hypotheses $H^+_{r/n}$ and $H^-_{r/n}$ can be tested with any test statistic suitable for testing partial conjunction hypothesis. Benjamini and Heller (2008) provide a general procedure for constructing valid test statistics starting from p-values of individual studies. See also Bogomolov and Heller (2023) and Wang et al. (2022) for a discussion of p-values for partial conjunction hypotheses. In what follows, we provide a brief overview essential for presenting our main result.

Let us consider a collection of n component null hypotheses $\left\{H^+_1, \dots, H^+_n\right\}$, where $H^+_i : \theta_i \leq 0$ and the associated alternative is $K^+_i : \theta_i > 0$. We assume to have independent normal estimators of components of θ, i.e. we let $T_i \sim N(\theta_i, 1)$ denote an estimator of θ_i, where, for simplicity, it is assumed that its variance is known and equal to 1. Then $p_i = 1 - \Phi(T_i)$ is a valid p-value for H^+_i, with Φ denoting a cumulative distribution function of the standard normal distribution. Independence of $T_i, i = 1, \dots, n$ follows from the assumption of the independence of studies.

Benjamini and Heller (2008) show that, since $H^+_{r/n}$ is true if there are at most $r-1$ arbitrarily large positive effects, a valid p-value is obtained by ignoring the $r-1$ smallest p-values and combining the remaining $n-r+1$ right sided p-values with a combining function that leads to a p-value stochastically larger or equal to the uniform distribution under $H^+_{r/n}$. Various combining functions could be employed. A simple method for combining p-values under arbitrary dependence is the Bonferroni method. Let $p_{(1)}, \dots, p_{(n)}$ be a sequence of p-values in a non-decreasing order. The Bonferroni method leads to the following combined p-value.

Definition 1. Bonferroni partial conjunction p-value is $p^+_{r/n} = (n - r + 1)p_{(r)}$.

The factor in Definition 1 corresponds to the usual Bonferroni correction applied to a collection of $n-r+1$ p-values obtained by ignoring the $r-1$ smallest p-values. It may be instructive to consider a special case of a global null hypothesis $H^+_{1/n}$ in which the above correction reduces to the well known factor n, and $H^+_{1/n}$ is rejected if $p_{(1)} \leq \alpha/n$.

Consider now the collection $\left\{H^-_1, \dots, H^-_n\right\}$, where $H^-_i : \theta_i \geq 0$ with the associated alternative $K^-_i : \theta_i < 0$. Then, given the continuity of the distribution of T_i, the p-value for H^-_i is $q_i = \Phi(T_i) = 1 - p_i$. As a consequence, the Bonferroni p-value for $H^-_{r/n}$ is $p^-_{r/n} = (n - r + 1)q_{(r)} = (n - r + 1)\left(1 - p_{(n-r+1)}\right)$.

2. Directional replicability when r is large

In general, the procedure for testing $H_{r/n}$ is based on a double of the smaller p-value pertaining to unilateral replicability hypotheses, i.e. in this case $p_{r/n} = 2\min\left\{p^-_{r/n}, p^+_{r/n}\right\}$, see Owen (2009), Wang et al. (2022) and Jaljuli et al. (2023). The following Theorem indicates that when r is large enough with respect to n, and the combining function is Bonferroni, the correction factor of two is unnecessary.

Theorem 1. *Let $T_i \sim N(\theta_i, 1)$ be an estimator of θ_i for $i = 1, \dots, n$, and assume that $T_1, \dots, T_n$ are independent. Define $p_i = 1 - \Phi(T_i)$ and let $p_{(1)} \leq \dots \leq p_{(n)}$ denote the ordered p-values. Let r be such that $(n+1)/2 < r \leq n$. Let further*

$$p^-_{r/n} = (n - r + 1)\left(1 - p_{(n-r+1)}\right), \quad p^+_{r/n} = (n - r + 1)p_{(r)}$$

be Bonferroni p-values for testing $H^-_{r/n} : n^- < r$ and $H^+_{r/n} : n^+ < r$, where n^- and n^+ indicate the number of positive and negative components of $\theta = (\theta_1, \dots, \theta_n)$, respectively. Then $p_{r/n} = \min\left\{p^-_{r/n}, p^+_{r/n}\right\}$ is a valid p-value for $H_{r/n} : n^- < r \wedge n^+ < r$.

The intuition behind Theorem 1 relies on observing that for the majority of points in the null parameter space, the probability of rejecting $H^-_{r/n}$ or $H^+_{r/n}$ is much lower than α. Specifically, the probability of rejecting $H^-_{r/n}$ approaches α only in the presence of $r-1$ very strong negative effects, and analogously, the probability of rejecting $H^+_{r/n}$ approaches α only in the presence of $r-1$ very strong positive effects. When r is above the median number of studies, the simultaneous presence of $r-1$ positive and $r-1$ negative effects is impossible, making the correction factor of 2 unnecessary. The proof of Theorem 1 is given in Supplementary material.

3. Directional replicability when r is small

In this section we present a counterexample that shows that for smaller values of r a certain correction factor is necessary.

Consider r such that $2 \leq r \leq (n+1)/2$. The main difference with respect to the setting of the previous section is the possibility of simultaneous presence of $r-1$ positive and $r-1$ negative effects. Given that the problem is invariant with respect to the permutations of θ, we can without loss of generality consider the null parameter space Θ_0, where:

$$\Theta_0 = \left\{\theta \in \mathbb{R}^n : \begin{array}{ll} \theta_i \geq 0, & i = 1, \dots, r-1; \\ \theta_i \leq 0, & i = r, \dots, 2r-2; \\ \theta_i = 0, & i = 2r-1, \dots, n. \end{array}\right.$$

Consider a test, that for a given $\alpha \in (0, 1/2)$, rejects $H_{r/n}$ if $\min\left\{p^-_{r/n}, p^+_{r/n}\right\} \leq \alpha$ as before. The two events leading to Type I error are no longer disjoint and thus the probability of Type I error, after performing simple set operations, can be expressed as

$$c(\theta) = pr_\theta(T_{(n-r+1)} \geq t) + pr_\theta(T_{(r)} \leq -t,\, T_{(n-r+1)} \leq t), \tag{1}$$

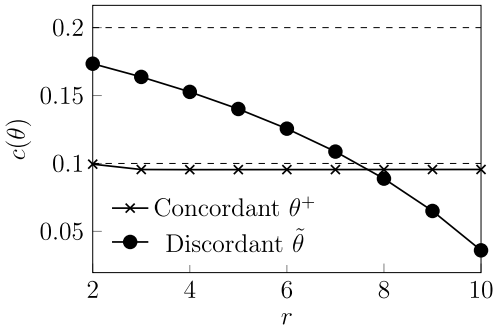

Fig. 1. Type I error probability as a function of r for two specific parameter configurations: "concordant" representing the presence of $r-1$ positive effects and "discordant" representing the presence of $r-1$ strong positive effects and $r-1$ strong negative effects. Dashed lines representing the nominal level of the test $\alpha = 0.1$, as well as the level of the standard corrected test 2α, are added for reference. The total number of studies is $n = 20$.

where $t = \Phi^{-1}\left[1-\alpha/(n-r+1)\right]$. Let θ^+ represent the setting with $r-1$ strong positive effects:

$$\theta^+ = \begin{cases} \theta_i^+ = \infty & i = 1,\dots,r-1, \\ \theta_i^+ = 0 & i = r,\dots,n. \end{cases} \tag{2}$$

Let $\tilde{\theta}$ represent the setting with $r-1$ strong positive and negative effects:

$$\tilde{\theta} = \begin{cases} \tilde{\theta}_i = \infty, & i = 1,\dots,r-1, \\ \tilde{\theta}_i = -\infty, & i = r,\dots,2r-2, \\ \tilde{\theta}_i = 0, & i = 2r-1,\dots,n. \end{cases} \tag{3}$$

Then it can be shown that

$$c(\theta^+) = 1 - \Phi(t)^{n-r+1} + \sum_{k=r}^{n-r+1} \binom{n-r+1}{k} \{1-\Phi(t)\}^k \{2\Phi(t)-1\}^{n-r+1-k}, \tag{4}$$

$$c(\tilde{\theta}) = 1 - \{2\Phi(t)-1\}^{n-2r+2}. \tag{5}$$

These probabilities can be evaluated numerically. Fig. 1 shows $c(\theta^+)$ and $c(\tilde{\theta})$ as a function of r for $n = 20$. For $r \in \{2,\dots,7\}$, the probability of Type I error at $\tilde{\theta}$ exceeds that of θ^+, and more importantly exceeds $\alpha = 0.1$. As a consequence, the size of the test, defined as a supremum of $c(\theta)$ over Θ_0, will exceed α. Interestingly, the situation is reversed for $r \in \{8,9,10\}$.

Remark 1. Theorem 1 states then when $r > (n+1)/2$, the factor of 2 can be omitted, while the previous example illustrates that for some values of r below this threshold, i.e. $r \in \{2,\dots,7\}$, a correction factor is necessary. The following theorem states that the least favorable configuration under the null hypothesis is either θ^+ or $\tilde{\theta}$ and thus allows us to determine the smallest necessary correction factor for any pair (n,r).

Theorem 2. *Consider the procedure proposed in Theorem 1, now with $r \le (n+1)/2$. Let $\Theta_0 \subset \mathbb{R}^n$ be the null parameter space containing all values of θ such that the hypothesis $H_{r/n}$ is true. Let $c(\theta)$ denote the probability of making a Type I error for a given significance level $\alpha \in (0,1/2)$. Then*

$$\sup_{\theta\in\Theta_0} c(\theta) = \max\left\{c(\theta^+), c(\tilde{\theta})\right\}, \tag{6}$$

where $\theta^+, \tilde{\theta}$, $c(\theta^+)$ and $c(\tilde{\theta})$ are defined in (2)–(5). Furthermore, if $c(\tilde{\theta}) > \alpha$, then $\sup_{\theta\in\Theta_0} c(\theta) = c(\tilde{\theta})$, otherwise $\sup_{\theta\in\Theta_0} c(\theta) \le \alpha$.

A straightforward consequence of Theorem 2 is the necessary and sufficient condition for removing the factor of 2 when testing $H_{r/n}$.

Corollary 1. *Testing procedure proposed in Theorem 1 is valid for $r \le (n+1)/2$ if and only if*

$$c(\tilde{\theta}) = 1 - \{1 - 2\alpha/(n-r+1)\}^{n-2r+2} \le \alpha. \tag{7}$$

The case of $n = 3$ and $r = 2$ may be of particular practical interest.

Corollary 2. *Consider the procedure of Theorem 1 for $n = 3$ and $r = 2$. Then $p_{r/n}$ is a valid p-value for $H_{r/n}$.*

4. Data adaptive choice of r

We have so far assumed that r is chosen prior to analysis. In many circumstances there are no substantive considerations indicating which value of r should be preferred. In those situations, it may be desirable to determine r in a data adaptive manner. This can be achieved by sequential testing of a collection of nested null hypotheses $H_{k/n}, H_{(k+1)/n}, \dots, H_{n/n}$, where k is the smallest integer satisfying (7) for given n and α (Maurer, 1995). The procedure starts from $H_{k/n}$ and tests each hypothesis at level α. If the hypothesis is rejected, the testing proceeds to the next, if not the procedure terminates. This procedure controls familywise error rate at level α. This is trivially true when all considered hypotheses are false. Otherwise, let h be an index of the first true hypothesis, with $k \le h \le n$ and let R_i be an indicator variable indicating whether $H_{i/n}$ has been rejected ($R_i = 1$) or not ($R_i = 0$) for $i = k, \dots, n$. Then the event of making at least one Type I error, $[\sum_{i=h}^{n} R_i > 0]$, coincides with the event $[R_h = 1]$, since a Type I error can only be made when $H_{h/n}$ is rejected. Therefore, the familywise error rate is equal to the probability of falsely rejecting $H_{h/n}$, which is bounded by α.

As a consequence, we can interpret the outcome of this procedure in terms of lower confidence bounds. Let l be the index of the last rejected hypothesis, set to zero if $H_{k/n}$ is not rejected. Then l is a $(1-\alpha)$ lower confidence bound for the maximum number of effects of the same sign, that is $l \le \max\{n^+, n^-\}$ with probability at least $1-\alpha$, or equivalently, there are at least l effects in the same direction with probability at least $1-\alpha$.

5. Discussion

Bonferroni method is a simple way to combine p-values under arbitrary dependence; here however we assume data coming from different studies to be independent. In that case, the power of the Bonferroni method can be improved by methods that exploit independence. In particular, it can be easily shown that for $r > (n+1)/2$, the result of Theorem 1 remains valid when the Bonferroni correction is substituted by the Šidák (1967) correction that assumes independence. Other combining functions that assume independence include the Simes method

$$p^+_{r/n} = \min_{i=1,\dots,n-r+1} \left\{ \frac{n-r+1}{i} p_{(r-1+i)} \right\},$$

and the Fisher combining function

$$p^+_{r/n} = pr\left\{ \chi^2_{2(n-r+1)} \ge -2\sum_{i=r}^{n} \log p_{(i)} \right\}.$$

An interesting question that awaits future research is whether the result presented in this work can be extended to these combining functions.

In this work, we have focused on a single directional replicability hypothesis, while in many applications, numerous features are studied simultaneously. In those situations, the power for identifying replicating signals can be increased by a careful consideration of the multiple testing aspect. Two existing approaches include filtering based on conditioning (Wang et al., 2022; Dickhaus et al., 2021). To obtain directional replicability claims, one can follow Owen (2009) and apply proposed procedures over the set of features twice, once for each direction at the significance level $\alpha/2$ and declare as replicated features belonging to the union of the two rejection sets. An open question that awaits future research is whether the correction factor of 2 can be removed in this case, extending the results reported in Dreyfuss et al. (2024) for $n = 2$.

The special case of n out n directional replicability corresponds to testing whether all components of θ are non-zero and of the same sign. In that case, the test of Theorem 1 rejects the null hypothesis if either $p_{(n)} \le \alpha$ or $p_{(1)} \ge 1-\alpha$, or equivalently if either $T_i \le -t$ for all $i = 1, \dots, n$, or $T_i \ge t$, for all $i = 1, \dots, n$, with $t = \Phi^{-1}(1-\alpha)$. This is a special case of a problem studied by Sasabuchi (1980) in the context of testing hypotheses pertaining to multivariate normal means (equation 4.3 in Section 4). Sasabuchi (1980) has derived the test of Theorem 1 as a likelihood ratio test for a slightly different null hypothesis, while Berger (1989) showed that the result remains valid for the null hypothesis $H_{n/n}$. Our work can be seen as an extension of these results to $r < n$.

In general, rejection of $H_{r/n}$ allows one to conclude that there are at least r effects sharing a common sign, without indicating whether that sign is positive or negative. Inferring the sign from data after rejecting the null hypothesis can lead to Type III error, an issue of concern in directional inference, see Heller and Solari (2024). It is easily checked that the procedure presented in Theorem 1 controls Type III error (Proposition 1 in Supplementary material) and thus allows one to infer the sign of the replicated effects post-hoc: the sign is positive if $p^+_{r/n} < \alpha$, and the sign is negative if $p^-_{r/n} < \alpha$.

For $n > 2$, multiple values of r are possible, each corresponding to a different replicability claim. Larger values of r represent stronger claims and are therefore of greater practical relevance. However, the power to reject $H_{r/n}$ decreases as r increases, so choosing r requires balancing these opposing considerations. We hope that our contribution – demonstrating how power can be improved for larger values of r – will be useful when addressing this trade-off.

Acknowledgments

This work was supported in part by United States National Institute on Aging grants R01AG080598 and R03AG102038 and by the European Union - Next Generation EU, Mission 4 Component 2, within the National Recovery and Resilience Plan project "Age-It - Ageing well in an ageing society" (PE0000015 - CUP H73C22000900006).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2026.110662.

Data availability

No data was used for the research described in the article.

References

Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64 (4), 1215–1222.
Berger, R.L., 1989. Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. J. Amer. Statist. Assoc. 84 (405), 192–199.
Bogomolov, M., Heller, R., 2023. Replicability across multiple studies. Statist. Sci. 38 (4), 602–620.
Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al., 2016. Evaluating replicability of laboratory experiments in economics. Science 351 (6280), 1433–1436.
Dickhaus, T., Heller, R., Hoang, A.-T., Rinott, Y., 2021. A procedure for multiple testing of partial conjunction hypotheses based on a hazard rate inequality. arXiv preprint arXiv:2110.06692.
Dreyfuss, J.M., Djordjilović, V., Pan, H., Bussberg, V., MacDonald, A.M., Narain, N.R., Kiebish, M.A., Blüher, M., Tseng, Y.-H., Lynes, M.D., 2024. ScreenDMT reveals DiHOMEs are replicably inversely associated with BMI and stimulate adipocyte calcium influx. Commun. Biol. 7 (1), 996.
Heller, R., Solari, A., 2024. Simultaneous directional inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 86 (3), 650–670.
Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2 (8), e124.
Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O.A., Heller, R., 2023. Quantifying replicability and consistency in systematic reviews. Stat. Biopharm. Res. 15 (2), 372–385.
Maurer, W., 1995. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypothesis. Biomed. Chem.-Pharm. Ind. 6, 3–18.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716.
Owen, A.B., 2009. Karl Pearson's meta-analysis revisited. Ann. Statist. 37 (6B), 3867–3892.
Sasabuchi, S., 1980. A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. Biometrika 67 (2), 429–439.
Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Amer. Statist. Assoc. 62 (318), 626–633.
Wang, J., Gui, L., Su, W.J., Sabatti, C., Owen, A.B., 2022. Detecting multiple replicating signals using adaptive filtering procedures. Ann. Stat. 50 (4), 1890.