That's Not a Two-Sided Test! It's Two One-Sided Tests!

Mark Rubin
Durham University

Citation: Rubin, M. (2022). That's not a two-sided test! It's two one-sided tests! *Significance*, *19*(2), 50-53. https://doi.org/10.1111/1740-9713.01619

Abstract

When reporting tests of significance, researchers might claim to have conducted a two-sided test when in fact they have conducted two one-sided tests. **Mark Rubin** explains the confusion and how to avoid it.

Correspondence concerning this article should be addressed to Mark Rubin at the Department of Psychology, Durham University, South Road, Durham, DH1 3LE, UK. Tel: +61 0407 949785. E-mail: Mark-Rubin@outlook.com

Do men have higher, lower, or the same level of self-esteem as women? To answer this question, researchers might sample a number of men and women and measure their self-esteem. They might then compare the self-esteem scores of their male and female participants. However, they need some way of distinguishing random differences in self-esteem from "significant" differences that are much less likely to have arisen if there is no genuine gender difference. Significance tests allow researchers to make this distinction. Importantly, there are two types of significance test: *two-sided* tests and *one-sided* tests. Two-sided tests yield two-sided *p* values that align with conclusions about *non-directional* hypotheses (e.g., "men have *different* self-esteem than women"). One-sided tests yield one-sided *p* values that align with conclusions about *directional* hypotheses (e.g., "men have *higher* self-esteem than women"). Problematically, researchers sometimes use the two-sided *p* values from two-sided tests to make claims about directional hypotheses. As I explain below, two-sided *p* values do not align with directional claims.

Two-sided *p* values do not align with directional claims

Imagine that a researcher conducts an independent samples *t* test in which participants' gender (male/female) is the independent variable, and participants' self-esteem scores are the dependent variable. Further imagine that the researcher conducts a two-sided significance test, obtains a significant two-sided *p* value (i.e., $p < 0.05$), and notes that male participants have higher

self-esteem scores than female participants. In this case, the researcher may reject a directional null hypothesis (i.e., that "men have *the same or lower* self-esteem than women" – shown as Direction 1 H_0 in Figure 1) and claim support for a directional alternative hypothesis (i.e., that "men have *higher* self-esteem than women" – shown as Direction 1 H_1).

However, if the researcher has undertaken a two-sided test and obtained a two-sided *p* value, then their significant result does not refer to a directional null hypothesis. Instead, it refers to a non-directional, two-sided, joint null hypothesis. In this example, it refers to the joint null hypothesis that "men have *neither* (i) higher self-esteem *nor* (ii) lower self-esteem than women" (shown as Joint H_0 in Figure 1).[1]

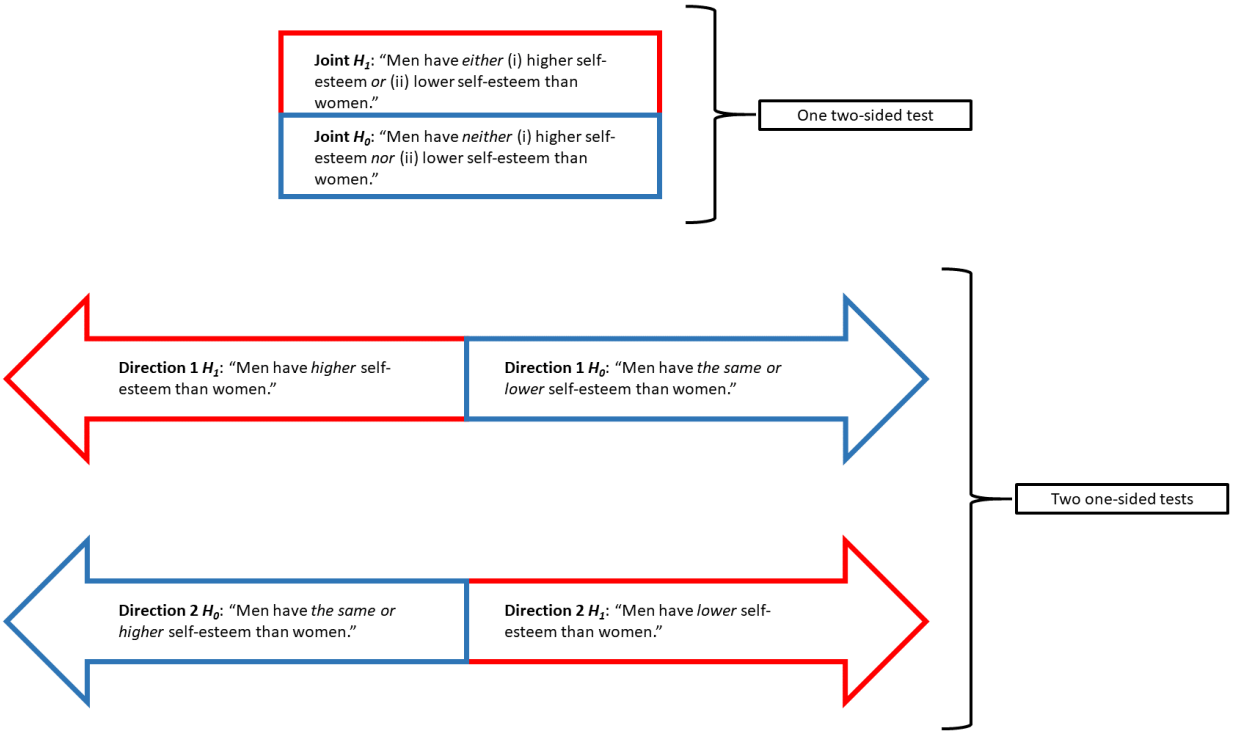

FIGURE 1 Illustration of the hypotheses involved in two-sided and one-sided testing.

The two-sided joint null hypothesis in Figure 1 can be restated more simply as the *nil null hypothesis* that "men have the same self-esteem as women". If the researcher rejects this nil null hypothesis using a significant two-sided *p* value from a two-sided test, then they can make the non-directional claim that "men have *either* (i) higher self-esteem *or* (ii) lower self-esteem than women" (shown as Joint H_1) or, more simply, that "there is a gender difference in self-esteem".

Importantly, a significant two-sided *p* value from a two-sided test does not align with the directional claim that "men have higher self-esteem than women", because two-sided *p* values refer to non-directional two-sided null hypotheses, not directional one-sided null hypotheses. If a researcher conducts a two-sided test and then wishes to make a directional claim about their observed effect, then they should halve their two-sided *p* value in order to obtain a one-sided *p* value (assuming a symmetrical sampling distribution). In this case, it would be more appropriate for the researcher to describe their test as "two one-sided tests" rather than as a two-sided test (for related points, see Georgi Z. Georgiev's onesided.org, Richard D. Morey's *Medium* post (http://bit.ly/3dNk2sx), and papers by Cortina and Dunlap,[2] Cox,[3] Meehl,[4] and Tukey[5]). Under this interpretation, the one-sided *p* value for the observed effect is equal to half the value of the two-

sided *p* value (i.e., *p*/2), and the one-sided *p* value for the effect in the opposite direction is equal to 1 – *p*/2.

The reinterpretation of a two-sided test as two one-sided tests and the associated halving of the two-sided *p* value is important because, if a researcher uses two-sided *p* values to make decisions about directional null hypotheses, then (a) their evidence will be weaker than it should be, and (b) their Type II ("false negative") error rate will be higher than necessary (for the same view, see Georgiev's onesided.org). For example, if a researcher rejects a directional null hypothesis on the basis of a two-sided *p* value of 0.04 and an alpha level (significance threshold) of 0.05, then the evidence for their claim will be weaker than if they had used a one-sided *p* value of 0.02. In addition, if a researcher uses a two-sided *p* value of 0.07 to make a decision about a directional null hypothesis, then they will fail to reject that hypothesis using an alpha level of 0.05. However, if they halve their two-sided *p* value to produce a one-sided *p* value, then they will obtain a significant result (i.e., $p = 0.035$) that allows them to reject the directional null hypothesis.

To be clear, the results of two-sided tests *can* be used to make directional claims. However, to do so, the two-sided *p* value should be halved to obtain a one-sided *p* value (assuming a symmetrical sampling distribution), and this one-sided *p* value should be conceptualised as the result of one of two one-sided tests. So, the critical point here is that there must be a logical consistency between the type of claim that is made (directional or nondirectional) and the type of *p* value that is used to support that claim (one-sided or two-sided). For example, it would be inappropriate to halve a two-sided *p* value in order to support a nondirectional claim.

BOX

How to report the results of two-sided and one-sided tests

A two-sided test

A **two-sided** independent samples *t* test found that men had significantly **different** self-esteem scores ($M = 32.51$, $SD = 46.97$) than women ($M = 22.89$, $SD = 52.89$), $t(479) = 2.11$, $\boldsymbol{p = 0.036}$, $d = 0.19$.

Here, a two-sided p *value is reported, and the claim is non-directional.*

A one-sided test

A **one-sided** independent samples *t* test found that men had significantly **higher** self-esteem scores ($M = 32.51$, $SD = 46.97$) than women ($M = 22.89$, $SD = 52.89$), $t(479) = 2.11$, $\boldsymbol{p = 0.018}$, $d = 0.19$.

Here, a one-sided p *value is reported, and the claim is directional.*

Two one-sided tests of directional hypotheses do not require an alpha adjustment

Researchers often explore their data for interesting effects and then attempt to explain those effects in a *post hoc* manner by drawing on prior theory in the literature.[6,7] In this exploratory mode, researchers may use two one-sided tests in which one test is used to investigate a putative effect in one direction (e.g., "do men have *higher* self-esteem than women?"), and the other test is used to investigate a putative effect in the other direction (e.g., "do men have *lower* self-esteem than women?"). Do researchers need to adjust their alpha level to take into account the multiple (dual) testing that occurs in this situation? The answer to this question depends on the type of hypothesis that the researcher is testing: non-directional or directional.

If a researcher uses two one-sided tests to test a *non-directional* null hypothesis (e.g., "men have the same self-esteem as women"), then it is necessary for them to reduce the alpha level for each of their tests (e.g., from 0.05 to 0.025), because they are undertaking multiple (dual) tests of the same joint null hypothesis. However, if a researcher uses two one-sided tests to test two *directional* null hypotheses, then it is not necessary for them to reduce their alpha level, because their decision about rejecting each directional hypothesis is based on a single test result, rather than multiple (dual) test results (for a similar view, see http://bit.ly/2Jxyqav).

To illustrate this point, let us first consider the case in which a researcher wishes to make a decision about a *non-directional* joint null hypothesis (such as Joint H_0 in Figure 1) using two one-sided tests. In this case, the joint null hypothesis consists of two constituent directional null hypotheses (Joint H_0(i) and Joint H_0(ii)).[1] One of these constituent directional null hypotheses posits the absence of an effect in one direction (e.g., "men do not have higher self-esteem than women"), and the other posits the absence of an effect in the other direction (e.g., "men do not have lower self-esteem than women"). The researcher's two one-sided tests would then test each of these directional constituent null hypotheses and yield a one-sided *p* value in each case. If the researcher aims to reject the non-directional joint null hypothesis because one of these two one-sided *p* values is significant (i.e., *union–intersection testing*),[1,8] then they should halve the alpha level for each one-sided test. For example, if their alpha level for the non-directional joint null hypothesis (α_{Joint}) is 0.05, then they should set the alpha level for each of the two directional constituent null hypotheses ($\alpha_{\text{Constituent}}$) at 0.025. The researcher needs to make this alpha adjustment to compensate for the multiple (dual) testing of the non-directional joint null hypothesis and maintain the Type I ("false positive") error rate for the joint null hypothesis at its nominal level (i.e., 0.05).[1]

Let's now turn to the more common case in which a researcher wishes to make decisions about two *directional* null hypotheses (such as Direction 1 H_0 and Direction 2 H_0 in Figure 1) using two one-sided tests. In this case, the researcher is not interested in making non-directional claims by rejecting a non-directional joint null hypothesis. Consequently, they do not need to adjust their alpha level for either of their one-sided tests. Instead, they may conduct these two one-sided tests using a conventional unadjusted alpha level in each case (i.e., $\alpha = 0.05$) in order to make two separate decisions about rejecting two directional null hypotheses.[9–11] For example, they may decide whether to reject Direction 1 H_0 in Figure 1 based on a single one-sided *p* value from a single one-sided test using an unadjusted alpha level. In this case, the Type I error rate only refers to the null hypothesis that "men have *the same or lower* self-esteem than women." It does not refer to the Direction 2 H_0 that "men have *the same or higher* self-esteem than women" or the Joint H_0 that "men have *neither* (i) higher self-esteem *nor* (ii) lower self-esteem than women".

In summary

Researchers sometimes conduct two-sided tests and then use the resulting two-sided *p* values to make directional claims. However, this approach is inappropriate because two-sided *p* values refer to non-directional null hypotheses, not directional null hypotheses. This approach also produces weaker evidence and a higher Type II error rate than necessary because two-sided *p* values are twice the size of the one-sided *p* value for the observed effect (assuming a symmetrical sampling distribution). In order to make directional claims, researchers should halve their two-sided *p* values to obtain one-sided *p* values. In addition, in exploratory research situations, researchers who wish to make directional claims should use two one-sided tests that test effects in each direction. In this case, they do not need to adjust their alpha level to account for multiple (dual) testing because they are not testing a non-directional joint null hypothesis.

Author bio

Mark Rubin is a professor in the Department of Psychology at Durham University, UK. For more about Rubin's work, visit http://bit.ly/rubinpsyc.

Note

This article is a revised version of a previously published article. The revisions took into account helpful comments from Wei Hong. Dr Hong's comments and my response are available here: https://doi.org/10.1111/1740-9713.01620. I would also like to thank Georgi Georgiev for his advice on the issues covered in this article.

References

1. Freedman, L. S. (2008) An analysis of the controversy over classical one-sided tests. *Clinical Trials*, **5**(6), 635–640.

2. Cortina, J. M. and Dunlap, W. P. (1997) On the logic and purpose of significance testing. *Psychological Methods*, **2**, 161–172.

3. Cox, D. R. (1977) The role of significance tests. *Scandinavian Journal of Statistics*, **4**, 49–63.

4. Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, **34**, 103–115.

5. Tukey, J. W. (1991) The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.

6. Rubin, M. (2017) Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, **21**, 269–275.

7. Rubin, M. (2022) The costs of HARKing. *British Journal for the Philosophy of Science, 73*. https://doi.org/10.1093/bjps/axz050

8. Roy, S. N. (1953) On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, **24**, 220–238.

9. Savitz, D. A. and Olshan, A. F. (1995) Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology*, **142**, 904–908.

10. Tukey, J. W. (1953) *The problem of multiple comparisons*. Princeton, NJ: Princeton University.

11. Wilson, W. (1962) A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin*, **59**, 296–300.

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.

Correction Notice from Publisher

This article is a corrected version of the article originally published in the June 2020 issue of the magazine (https://doi.org/10.1111/1740-9713.01405). It contains multiple changes and corrections made after its initial publication. The changes were made by the author after issues were raised by a reader. To summarise the changes, the author has: rewritten some text to refer to one- and two-sided *p*-values rather than one- and two-sided tests; removed the point that two-sided tests do not warrant directional claims, replacing it with the point that two-sided *p*-values do not align with directional claims; clarified that the results of two-sided tests can be used to make directional claims and that, to do so, a two-sided *p*-value should be halved to obtain a one-sided *p*-value (assuming a symmetrical sampling distribution); clarified that it would be inappropriate to halve a two-sided *p*-value to support a nondirectional claim; and removed the misleading claim that two-sided tests require an alpha adjustment. The decision to take this approach to making the correction was taken by the editors of *Significance*. Wiley retains its commitment to the standard methods of correcting published content, described here: https://authorservices.wiley.com/ethics-guidelines/index.html#10.