Statistical Science
2017, Vol. 32, No. 3, 349–351
DOI: 10.1214/16-STS590
Main article DOI: 10.1214/16-STS571

Randomization-Based Tests for "No Treatment Effects"

EunYi Chung

Abstract. Although both Fisher's and Neyman's tests are for testing "no treatment effects," they both test fundamentally different null hypotheses. While Neyman's null concerns the average casual effect, Fisher's null focuses on the individual causal effect. When conducting a test, researchers need to understand what is really being tested and what underlying assumptions are being made. If these fundamental issues are not fully appreciated, dubious conclusions regarding causal effects can be made.

Key words and phrases: Fisher's randomization test, Neyman's randomization test, treatment effect, Wilcoxon–Mann–Whitney rank sum test.

I would like to thank Peng Ding for his meticulous investigation to better understand the differences between Fisher's and Neyman's nulls, which have historically been the source of great confusion and contention among researchers. Although both are randomization-based tests for testing "no treatment effects," Neyman's null concerns the *average* causal effect while Fisher's null focuses on the *individual* causal effect. When there is zero individual causal effect, there is zero average causal effect. Thus, Fisher's sharp null logically implies Neyman's null. However, a seemingly paradoxical phenomenon arises because a rejection of Neyman's null does not imply a rejection of Fisher's null in many situations. Ding presents an asymptotic comparison between these two approaches, which provides an explanation as to why such a paradox exists.

In comparing the two approaches, I would like to discuss the importance of understanding: (1) what is being tested and (2) what are the underlying assumptions being made when one is conducting a test. When these fundamental issues are not fully appreciated, they can lead to dubious conclusions regarding causal effects.

1. FISHER'S AND NEYMAN'S TESTS

As randomization-based tests, both Fisher's and Neyman's tests treat the treatment assignment as random while all potential outcomes are fixed. Although both test "no treatment effects," they attempt to answer fundamentally different null hypotheses: Fisher's null deals with the sharp null of no individual causal effect whereas Neyman's null concerns zero average causal effect. In addition, while Neymans' test is (asymptotically) valid regardless of the null hypothesis, the validity of Fisher's test heavily hinges on the sharp null hypothesis. In other words, if the sharp null is not satisfied, Fisher's tests will, in general, fail to control the probability of a Type 1 error.

Under Fisher's sharp null (so Neyman's null is also satisfied), both tests are valid at least asymptotically. It is worth noting that while Fisher's test is an exact test, Neyman's test relies on asymptotics based on two approximations: approximated variance estimation and a normal approximation.

On the other hand, under Neyman's null, Fisher's null is not necessarily satisfied and Fisher's test is, in general, invalid, in the sense that the rejection probability can be far from the nominal level α. It is only in cases where $N_0 = N_1$ or $S_0 = S_1$ that Fisher's and Neyman's tests are asymptotically equivalent. Otherwise, a problem can arise if the sample sizes are unequal and the variances of potential outcomes are different.[1]

EunYi Chung is Assistant Professor of Economics, Department of Economics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA (e-mail: eunyi@illinois.edu).

[1]The same phenomenon can also be observed in permutation tests; the permutation test based on the sample mean difference will be asymptotically valid only if either the two samples are of the same size or the variances of the samples are identical (Romano, 1990). However, this invalidity of the two sample permutation tests

2. WHAT CAN GO WRONG?

Neyman's null clearly does not imply Fisher's sharp null. The treatment may affect the potential outcomes in different ways other than the average. For instance, the treatment may increase the dispersion of the potential outcomes while the means are kept the same. When there is a heterogeneous treatment effect, the comparison between Fisher's test and Neyman's test becomes ambiguous. As discussed in Ding's article, if $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) > 0$, the rejection probability of Fisher's test can be much less than the nominal level, which by continuity implies the test is biased and has little power of detecting a true difference in means. Under such situations, Neyman's test seems to have much higher power than Fisher's test.

However, the reverse case can also occur. To be more concrete, consider the following scenario. Assume that the averages of the potential outcomes are equal, so $\bar{Y}_1 = \bar{Y}_0$, and thus Neyman's null holds. Further assume that $N_0 < N_1$ and $S_0 > S_1$ so that $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) < 0$ and $\hat{V}(\text{Fisher}) < \hat{V}(\text{Neyman})$ for large N. Then Fisher's test will reject more than Neyman's test. This lack of robustness and the increased probability of a Type 1 error can yield misleading conclusions. The rejection of the null may incorrectly be interpreted as rejection of equal means when, in fact, it is caused by unequal variances and sample sizes.

3. WILCOXON–MANN–WHITNEY RANK SUM TEST

The fact that Fisher's test and Neyman's test become asymptotically equivalent when $N_0 = N_1$ or $S_0 = S_1$ holds is a unique occurrence for the difference-in-means statistic. The situation can be even worse when basing a test on a different statistic, in the sense that even if the variances are equal and the sample sizes are the same, the asymptotic rejection probability under Neyman's null can be very far from the nominal level α. We will investigate this fact using the Wilcoxon–Mann–Whitney rank sum statistic.

Consider the test statistic of interest, which is given by

$$\hat{\theta} = \frac{1}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} I\big(Y_i(0) < Y_j(1)\big),$$

which can be evidently viewed as an estimator of $\theta = P(Y_i(0) \le Y_j(1))$.

Under Fisher's null, the Wilcoxon–Mann–Whitney rank sum test for testing Fisher's sharp null results in exact level α in finite sample cases. However, this test under Neyman's null may result in faulty conclusions. Under Neyman's null, rejecting the null does not necessarily imply a zero average treatment effect. Just like the difference-in-means case, the null can be rejected even if the average treatment effect is zero. Even with balanced sample sizes and equal variances of the potential outcomes, the limiting variance of Fisher's test and that of Neyman's are not generally equal.

The intuitive reasoning of such a result stems from the fact that the statistic is a rank statistic, that is, its variance of the statistic under Fisher's null is solely determined by the sample sizes N_0 and N_1, regardless of the distributions of the potential outcomes. For example, the limiting variance of the test statistic is $1/3$ when $N_0 = N_1 = 4$, 0.3472 when $N_1 = 12$ and $N_0 = 18$, and 0.375 when $N_1 = 50$ and $N_0 = 100$.

In contrast, the variance of the statistic under Neyman's null will depend on the true nature of the distributions of the potential outcomes. In fact, under Neyman's null, the rejection probability of the test can be far away from the nominal level α. More importantly, the Wilcoxon–Mann–Whitney rank sum test statistic is not suitable for detecting divergence of Neyman's null since the test statistic is more appropriate as an estimator of $\theta = P(Y_i(0) \le Y_j(1))$, not the mean difference.

4. CONCLUSION

So, which test is better? The answer depends on what one is really trying to test and what assumptions one is willing to make. When one is interested in making an inference about the average treatment effect, the first thing one needs to ask oneself is whether or not one is willing to make an additional assumption about the distribution of the potential outcomes. While Neyman's test is asymptotically valid, Fisher's test under Neyman's null can fail to control the Type 1 error, even asymptotically unless one assumes a shift model where the disparity between the potential outcomes is accompanied by a shift in means.

In contrast, if one is interested in Fisher's null, both Fisher's test and Neyman's test are valid (at least asymptotically for Neyman's test). For the Fisher's test, although *any* choice of statistic will be valid to use in the sense that the rejection probability under the null is exactly the nominal level, when it comes to the

can be overcome by using a test statistic which is appropriately transformed to be asymptotically pivotal. See Chung and Romano (2013, 2016) for more details.

power of a test, the choice of statistic plays an important role. The key element for higher power is to choose a statistic that can sensitively distinguish between the null and sensible alternatives. As discussed earlier, under the same sharp null, a statistic based on the mean difference can only detect the difference of the mean while not being able to detect any differences other than the mean. For instance, even if there was a significant causal effect while maintaining the mean, the statistic based on the mean difference will fail to detect such a difference. Therefore, when testing Fisher's sharp null, it is advisable that one uses a more omnibus statistic such as the Kolmogorov–Smirnov statistic or the Cramér–von Mises statistic. In doing so, the statistic captures the differences of the entire distribution as opposed to a particular aspect of the distribution.

Without controlling the level of tests, comparing the power of tests has less credibility. One possible solution to overcome the increased probability of a Type 1 error for Fisher's test under Neyman's null is to studentize the test statistic so that the imbalance of the samples does not cause a failure in controlling the probability of a Type 1 error. In addition, a test statistic should be deliberately chosen so that it will detect the divergence from the null.

REFERENCES

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

CHUNG, E. and ROMANO, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plann. Inference* **168** 97–105. MR3412224

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85** 686–692. MR1138350