Georgia Journal of Science

Volume 78 *No. 3 Notes, Editorials, Letters, and Errata* Article 1

2020

Editorial, Your Null Hypothesis Must Be False: Test It Anyway

Frank Corotto
University of North Georgia, frank.corotto@ung.edu

Editorial

YOUR NULL HYPOTHESIS MUST BE FALSE: TEST IT ANYWAY

Frank Corotto
Associate Editor, Georgia Journal of Science
Department of Biology, University of North Georgia
Dahlonega, Georgia, 30533
frank.corotto@ung.edu

One of the earlier criticisms of null hypothesis testing came from Berkson,[1] who noted that with large samples everything becomes "significant",[2] because a null hypothesis is hardly ever correct in the first place. Why test a null we know is wrong? It is to use what has been referred to as the null hypothesis test's "sign determination function",[3,4] i.e., to determine the *direction* of a difference.

Your Null Is Wrong and You Should Already Know That

It is rarely appreciated that what makes a null hypothesis a null hypothesis is its infinite precision. Ronald Fisher used his famous example of the lady tasting tea to illustrate. We begin with a woman who says she can tell from the taste whether tea or milk was poured into a cup first. Fisher's null hypothesis was that she cannot tell the difference. Why not test the hypothesis that she *can* tell the difference? Fisher explained as follows.

> *But this last hypothesis, however reasonable or true it may be, is ineligible as a null hypothesis to be tested by experiment, because it is inexact . . . It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the "problem of distribution," of which the test of significance is the solution.*[5]

What Fisher was explaining, badly, was that a probability density function has to be constructed around a specific number, such as a *t* value of zero or an *F* value of unity, and those numbers are infinitely precise. Those precise numbers are predictions based on hypotheses and, because the predictions are infinitely precise, the hypotheses must be too.[6] For example, our null may be that a drug and a placebo have exactly the same effect to an infinite number of decimal places. We would then predict that the two sample means

[1] J. Berkson. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526–542.

[2] "Significance" in inferential statistics means only that results signify something, not that a difference is large or important. See D. Salsburg. 2002. The Lady Tasting Tea. Henry Holt and Company, LLC. p. 98. I use quotation marks to remind the reader of the special meaning of "significance".

[3] R. Harris. 1987. Reforming Significance Testing via Three-valued Logic. *In* L. Harlow, S. Mulaik, and J. Steiger, eds. What if There Were No Significance Tests? Psychology Press.

[4] I follow the CSE style manual when it comes to quotation marks and other punctuation.

[5] R. Fisher. 1971. The Design of Experiments. Hafner Publishing Company, Inc. See p. 16.

[6] P. Meehl. 1967. Theory-testing in psychology and physics: a methodological paradox. Philosophy of Science, 34(2), 103–115.

would be exactly the same and, from that, predict that *t* will be exactly zero. We would construct a probability distribution around that *t* value and take it from there. It is because nulls are infinitely precise that many cannot be true. Can party affiliations be exactly the same in Dade County, Florida, as in Palm Beach County? If 67.645328% of the voters in Dade County consider themselves Democrats, is it possible for it to be the same in Palm Beach County to the same number of decimal places? When large populations are sampled, nulls are certainly incorrect,[7] a fact confirmed when large sample sizes lead to everything being studied being deemed "significant".[8] It has been asserted that all we accomplish by testing these nulls is that we find out if we have sufficient statistical power to reject what we already should know is false.[9]

Things get tricky, though, if we leave sampling studies behind and look at experiments, where we explore the results of manipulating variables. Consider this modification of Frick's null hypothesis,[10] *talking to plants has no effect on their growth*. Suppose we spend an hour a day, talking to our favorite houseplant from 2 m away. The air we exhale will be saturated with water vapor. The partial pressure of carbon dioxide in that air will 28 mm Hg, while in the atmosphere in general it is only 0.3 mm Hg. Our favorite houseplant's growth must be affected. But suppose it is our favorite garden plant, and we are downwind and facing away from it when we talk to it. Here we reach an impasse. If the universe behaves deterministically, the plant's growth will be affected, because everything affects everything else, somehow. If the universe does not behave deterministically, then our null may be correct. Physicists debate whether or not the universe behaves deterministically[11] but, thankfully, it does not matter either way. What we really want to know is not whether talking to plants affects their growth, but *how* it affects their growth. Does talking to plants increase their growth or decrease it? We can test a null hypothesis to find that out.

But I Want to Know If This Difference Is "Significant"!

The meaning of statistical "significance" is widely misunderstood. The word only means that results signify something,[12] i.e., we can exclude sampling error as the sole cause of a difference. "Significance" does not equate with *importance* or *large effect*.[13] Estimates of

[7] D. Bakan. 1966. P The test of significance in psychological research.sychological Bulletin, 66: 423–437; P. Meehl. 1967. Theory-testing in psychology and physics: a methodological paradox. Philosophy of Science, 34(2), 103–115; R. Frick. 1995. Accepting the null hypothesis. Memory & Cognition, 23(1), 132–138.

[8] D. Bakan. 1966. The test of significance in psychological research.Psychological Bulletin, 66: 423–437; P. Meehl. 1967. Theory-testing in psychology and physics: a methodological paradox. Philosophy of Science, 34(2), 103–115;

[9] G. Loftus. 1991. On the tyranny of hypothesis testing in the social sciences. A review of G. Gigerenzer, Z. Swijink, T. Porter, L. Daston, J. Beatty, and L. Krüger. The Empire of Chance: How Probability Changed Science and Everyday Life. Contemporary Psychology, 36(2), 102–105.

[10] R. Frick. 1995. Accepting the null hypothesis. Memory & Cognition, 23(1), 132–138.

[11] For example, see R. Colbeck and R. Renner. 2011. No extension of quantum theory can have improved predictive power. Nature Communications, 2:411, doi:10.1038/ncomms1416; and G. Ghirardi and R. Romano. 2013. Onthological models predictively inequivalent to quantum theory. Physical Review Letters, 110(17), 170404.

[12] D. Salsburg. 2002. The Lady Tasting Tea. Henry Holt and Company, LLC. p. 98.

[13] D. Bakan. 1966. The test of significance in psychological research. Psychological Bulletin, 66: 423–437. G. Loftus. 1991. On the tyranny of hypothesis testing in the social sciences. A review of G. Gigerenzer, Z. Swijink, T. Porter, L. Daston, J. Beatty, and L. Krüger. The Empire of Chance: How Probability Changed Science and Everyday Life. Contemporary Psychology, 36(2), 102–105.

the latter, which is best referred to as *effect size*, may take the form of a difference between two sample means, or something fancier like Cohen's *d*. Although *P* values are influenced by effect size, *P* values also depend on sample size and, in the case of numerical data, scatter; *P* values are not a pure indicator of effect size. Bakan emphasized that *P* values tell us about our data, not about anything at the population level.[14]

We can blame Ronald Fisher for this misleading term, "significance". Not only was he infatuated with the word, but he spoke of degrees of significance, which runs counter to the true meaning of the word in this context—the results signify *something*.

Drawing Conclusions About Direction

Our "proper statistics books" end some of their examples with the decision to accept the alternative (or alternate) hypothesis that something other than the null is correct. This hypothesis is often expressed as follows, $\mu_1 \neq \mu_2$, meaning that the two population means differ. Is that really where we stop? Of course not. We accept the direction of the difference in sample means as the direction of the difference in population means. This is exactly what Fisher did. Did he exclude chance as the sole cause of a change in crop yields, while withholding judgement as to whether yields increased or decreased? No. He described "significant" *decreases* in crop yields.[15] Was that after performing a one-tailed test? No. It was after performing analysis of variance, which is nondirectional.

In an often-cited, influential, and yet bizarre paper published in 1960, Kaiser[16] stated that we cannot draw a conclusion regarding direction after accepting the alternative hypothesis that $\mu_1 \neq \mu_2$. We can only draw a conclusion regarding direction if we perform a one-tailed test or what he described as a "directional" two-tailed test. What makes a directional two-tailed test different from a regular two-tailed test is the alternative hypothesis or, rather, hypotheses. In place of this, $\mu_1 \neq \mu_2$, we have these, $\mu_1 < \mu_2$ and $\mu_1 > \mu_2$. How can that possibly change the conclusions we can draw when none of those three alternatives are used for anything?[17] What Kaiser did not know is where the alternative hypothesis comes from. To Fisher, there was no alternative, and he unhesitatingly drew conclusions regarding direction. The idea of alternatives stems from Neyman and Pearson's method of hypothesis testing.[18] Neyman and Pearson thought we should be distinguishing between several hypotheses, a main one and one or more alternatives. All of those hypotheses would be *specific*. With their method, we would decide on α and β and determine the minimum sample size needed to detect a minimum effect size (the distance between the peaks of the distributions). Unfortunately, parts of Neyman and Pearson's approach became combined with parts of Fisher's. When the idea of an alternative got transplanted into Fisher's method, where there was no alternative, it turned into this $\mu_1 \neq \mu_2$, which is useless because it is the polar opposite of Neyman and

[14] D. Bakan. 1966. The test of significance in psychological research. Psychological Bulletin, 66: 423–437.

[15] R. Fisher. 1921. Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. Journal of Agricultural Science, 11, 107–135. See p. 110.

[16] H. Kaiser. 1960. Directional statistical decisions. Psychological Review, 67(3), 160–167.

[17] As if this were not enough, Kaiser went on to say that we can never draw conclusions regarding direction from analyses of variance of chi-squared tests, because *F* and X^2 are not directional statistics. They can only be positive, never negative. Fisher made directional decisions after calculated both of those statistics.

[18] For the best comparison of Fisher's approach with Neyman and Pearson's that I know of, see J. Perezgonzalez. 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, 6, article 223.

Pearson's specific alternatives. How could replacing $\mu_1 \neq \mu_2$ with $\mu_1 < \mu_2$ and $\mu_1 > \mu_2$ make any difference when none of those hypotheses are used for anything? The mathematical procedure is unchanged. Bakan justifiably ridiculed Kaiser as follows.

> *One really needs to strike oneself in the head! If Sample Mean A is greater than Sample Mean B, and there is reason to reject the null hypothesis, in what other direction can it reasonably be? What kind of logic is it that leads one to believe that it could be otherwise than that Population Mean A is greater than Population Mean B? We do not know whether Kaiser intended his paper as a reduction ad absurdum, but it certainly turned out that way.*[19]

In spite of Bakan's reasonable opinion, a body of literature arose around Kaiser's directional two-tailed tests as well as his type III errors, which occur when we are correct to reject the null, but we get the direction wrong.[20] Harris[21] brought the question of direction to the fore in 1987, proposing Kaiser's "three-valued logic" as the answer to some of the criticisms of null hypothesis testing. I see no value in Kaiser's alternative hypotheses—alternatives are only useful in the framework of Neyman and Pearson's method—but Harris at least highlighted the importance of testing null hypotheses to determine direction.

Though the idea of testing for direction has been around for some time, John Tukey gives us the best quotation to justify the practice. The setting is an experimental one, not a sampling study, and Tukey assumes the universe behaves deterministically.

> *All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking "Are the effects different?" is foolish. What we should be answering first is "Can we tell the direction in which the effects of A differ from the effects of B?" In other words, can we be confident about the direction from A to B? Is it "up," "down," or "uncertain"?*[22]

Traditional Null Hypothesis Testing

Some dogmatically insist that the null can never be correct. Tukey called the null hypothesis "fiction".[23] Consider also the quotation above. But some nulls can be correct and sometimes we just want to know if they are or if they are not. An example I read somewhere is *he has no extrasensory perception*. Obviously, psychological phenomena may exist or may not. The result of a test when they do exist, however, would be a difference in a certain direction. My own example is *these two genes are on separate*

[19] D. Bakan. 1966. The test of significance in psychological research.Psychological Bulletin, 66: 423–437.

[20] Just to show this literature exists, here are two examples conveniently in front of me. L. Leventhal and C-L. Huynh. 1996. Directional decisions for two-tailed tests: power, error rates, and sample size. Psychological Methods, 1(3), 278–292; P. McDonald. 1999. Type I, and type III error rates of parametric and nonparametric statistical tests. The Journal of Experimental Education, 67(4), 367–379.

[21] R. Harris. 1987. Reforming Significance Testing via Three-valued Logic. *In* L. Harlow, S. Mulaik, and J. Steiger, eds. What if There Were No Significance Tests? Psychology Press.

[22] J. Tukey. 1991. The philosophy of multiple comparisons. Statistical Science, 6(1), 100–116.

[23] L. Jones and J. Tukey. 2000. A sensible formulation of the significance test. Psychological Methods, 5(4), 411–414.

chromosomes. That is a perfectly good null because it leads to the prediction that the offspring of a certain type of mating will consist of precisely equal numbers of four different types of fruit flies. And if we reject that null, we are done. The genes share a chromosome.

How Often Do We Get the Direction Wrong?

Jones and Tukey[24] say that if we want to be wrong about direction 5% of the time, we should set α to 0.10, presumably because if we are wrong to make a decision, we would still get the direction correct half of the time by chance. In fact, they are correct only when the null hypothesis is true. When nulls are false, studies have shown that statistical power must be very low for errors of direction to be a problem.[25] Bakan was correct to ridicule Kaiser. We could set α to something like 0.15 and still make errors of direction only about 5% of the time. The other option is to keep α at 0.05 and know we will hardly ever make errors of direction. If you *really* want to understand errors of direction, read Chapter 17 in *Understanding Null Hypothesis Tests, and Their Wise Use*, which is available here. https://digitalcommons.northgeorgia.edu/bio_facpub/2/. I explain why a type I error is something we approach to different degrees. I explain how a statistic, like *t*, can have different balances between a meaningless component, created by sampling error, and a meaningful component, created by effect size. I explain why we get the direction wrong less than 2.5% of the time.

A Note Added Postpublication on 21 May 2020

Since publishing this editorial, I discovered that, in addition to the authorities I cite here, Hurlbert and Lombardi[26] also advocate that we test nulls to determine direction. In their excellent review of the topic, they cite still more authorities that agree.

[24] L. Jones and J. Tukey. 2000. A sensible formulation of the significance test. Psychological Methods, 5(4), 411–414.

[25] A. Gelman and J. Carlin. 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641–651; J. Lu, Y. Qiu, and A. Deng. 2018. A note on type S/M errors in hypothesis testing.

[26] S.H. Hurlbert and C.M. Lombardi. 2009. Final collapse o the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici, 46(5), 311–349.