Unicorns *Do* Exist: A Tutorial on "Proving" the Null Hypothesis

David L Streiner, PhD[1]

Introductory statistics classes teach us that we can never prove the null hypothesis; all we can do is reject or fail to reject it. However, there are times when it is necessary to try to prove the nonexistence of a difference between groups. This most often happens within the context of comparing a new treatment against an established one and showing that the new intervention is not inferior to the standard. This article first outlines the logic of "noninferiority" testing by differentiating between the null hypothesis (that which we are trying to nullify) and the "nill" hypothesis (there is no difference), reversing the role of the null and alternate hypotheses, and defining an interval within which groups are said to be equivalent. We then work through an example and show how to calculate sample sizes for noninferiority studies.

(Can J Psychiatry 2003;48:756–761)

Information on author affiliations appears at the end of the article.

Highlights

- Equivalence or noninferiority testing is used when we want to show that one treatment is not significantly worse than another.
- To do this, we reverse the meanings of the null and alternative hypotheses, and similarly the meanings of type I and type II error, and of power.
- In some circumstances, we need many more subjects to prove noninferiority than we need to show a difference.

Key Words: *significance testing, equivalence testing, null hypothesis significance testing*

In Philosophy 101, we learned that, if Person A posits the existence of some phenomenon, it is incumbent on Person A to prove its existence rather than it being Person B's job to disprove it. For example, if I assert that there are pink unicorns with purple polka dots hiding deep in the forest, I cannot simply say, "Well, prove that they don't exist." For the scientific world to accept my claim, I have to bring one back, dead or alive. Is it ever legitimate to violate this injunction and try to prove the nonexistence of something? In this paper, we'll take a look at trying to prove that something does not exist. In particular, we will examine the steps necessary to show that 2 treatments are equivalent—that, in essence, a difference between them does not exist. Let's start off, though, by looking at "normal" science.

In most instances, we design studies to show statistical significance. That is, we want to prove that one treatment is more effective or has fewer side effects than another, or we want to demonstrate that a relation exists between 2 variables, such as a history of sexual or physical abuse and the probability of having a psychiatric diagnosis (1). We begin by positing a null hypothesis that there is no difference between the groups or that there is no correlation between the variables, and then we do everything in our power to disprove it. If fortune deigns to smile upon us, and the statistical test has a P level less than or equal to 0.05, we conclude that we can reject the null hypothesis and therefore accept the alternative—that the treatments do differ or that the variables are correlated. The technical name for this is "null-hypothesis significance testing" (NHST). In NHST, if P is greater than 0.05, we do not say that we can accept (or prove) the null hypothesis; rather, we use the convoluted locution that we "have failed to disprove the null."

The reason for this, as I have said, harkens back to David Hume and the philosophy of science, which asserts that we cannot prove the nonexistence of something (unless, of course, it violates one of the laws of nature, such as the notions of perpetual motion machines, travel that is faster than light, or politicians who tell the truth). To use our original example, no one has ever seen such a unicorn (at least while sober), but we cannot prove that it does not exist. Although it is highly unlikely, a unicorn may walk out of the forest tomorrow, much as the coelacanth was discovered in 1938, after it was thought to have been extinct for millions of years. Using an example closer to home, 6 randomized controlled studies failed to find that ASA had any beneficial effect in preventing reinfarctions. However, Canner's metaanalysis demonstrated that ASA reduced mortality by 10% (2), and it is now inconceivable that post–myocardial infarction patients would not be told to take it. The 6 studies didn't prove that the null hypothesis is true—that there is no difference between ASA and placebo—they simply failed to reject it; that is, all 6 suffered from a type II error, in failing to reject the null hypothesis when in fact it is false. There is a qualitative difference between "highly unlikely" and "impossible" that can never be breached, no matter how many studies have negative outcomes. Therefore, a negative result often means that we should just try harder next time.

The only problem with this philosophical purity is that, as noted, there are times when we do want to demonstrate a lack of difference. This occurs most often in evaluating "me too" drugs—drugs that are supposedly as good as existing ones but that may be cheaper or have fewer side effects. Here, the first task is to show that they are no less effective—in other words, to "prove" the null hypothesis of no difference. Similar situations exist when an outpatient program is compared with an inpatient one, or time-limited therapy with is compared with therapy without a limit on the number of sessions, or a lower dosage of a drug is compared with a higher dosage (3). In all these cases, it would be sufficient to show that the less expensive or less invasive therapies are not worse than the alternative; it is not necessary to prove superiority in terms of outcome for them to be accepted as replacements. As we shall see, showing superiority vs noninferiority or equivalence does not simply demonstrate opposite sides of the same coin.

The Statistical Theory

How do we reconcile the competing demands of wanting to prove equivalence on the one hand with the difficulty, if not impossibility, of proving the null hypothesis on the other? First, we have to correct a common misperception about the null hypothesis (H_0). In almost all situations, the null hypothesis is written as

$$H_0\text{: } \mu_1 = \mu_2 \text{ or } H_0\text{: } \pi_1 = \pi_2 \qquad [1]$$

when we are comparing means (μs) or proportions (πs); that is, the means or proportions of the 2 groups are the same, vs the alternative hypothesis (H_A):

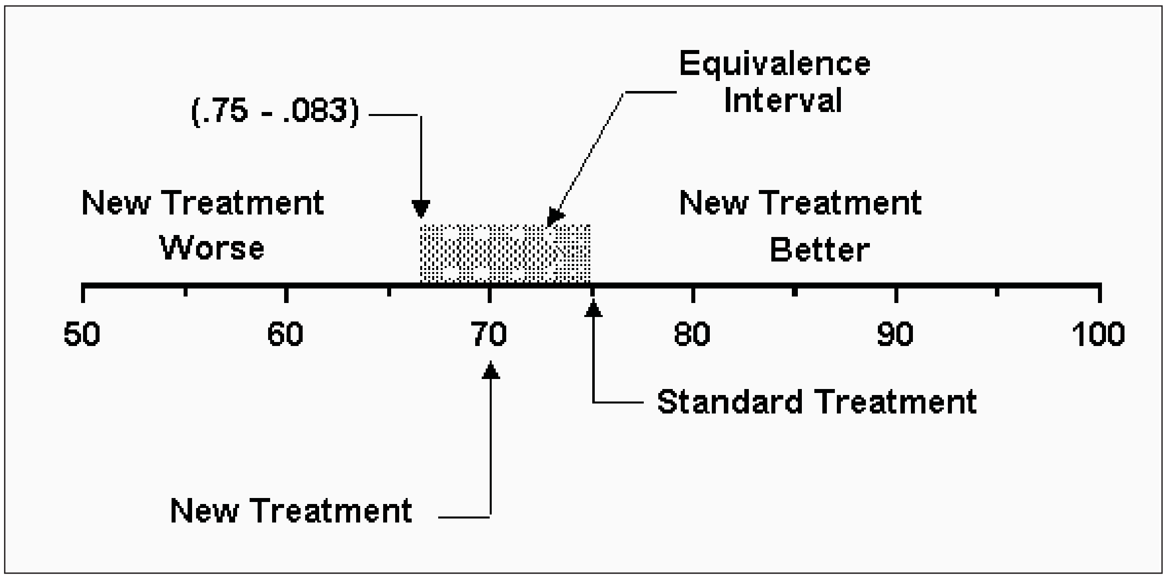
$$H_{A}\text{: } \mu_1 \neq \mu_2 \text{ or } H_A\text{: } \pi_1 \neq \pi_2 \qquad [2]$$

that the 2 means or proportions are different. The mistake is to think that the null hypothesis always has to mean "no difference." In fact, the null hypothesis is the hypothesis to be nullified, or disproven. Cohen refers to the hypothesis of no difference by the delightful name of the "nill hypothesis" (4). In most cases, the null and the nill hypotheses are the same; however, this needn't necessarily be the case, and we will use this distinction in testing for equivalence.

A second point is that not all differences are created equal, and there are some we can safely ignore. Because of sampling error, there will always be a difference between groups, no matter how similar they may be. Further, if we simply increase the sample size sufficiently, we will always be able to show that this difference is statistically significant. For example, let's assume that School 1 has a mean IQ score of 100, School 2 has a mean IQ score of 103, and the standard deviation is the usual 15 points. Most people would agree that this 3-point difference is clinically trivial. However, if we draw a sample of 400 students from each school, we will probably find that this difference is statistically significant. With larger sample sizes, we can find statistical significance with even smaller differences.

These 2 points—that the null hypothesis doesn't always mean no difference and that some differences may be statistically significant but clinically trivial—form the basis of testing for equivalence. First, rather than saying that the 2 means (or proportions, or whatever parameter we're interested in) have to be absolutely identical, we establish an equivalence interval within which we would say that the groups are "close enough." For instance, let's assume that, for sociophobic patients, Treatment A results in a mean score of 10 on a scale of social comfort (that is, $M_1 = 10$), where a higher score reflects greater comfort. How much lower can the score be with a different therapy (Treatment B) for us to say that the difference between the groups (which we call delta, or δ) is clinically unimportant—1 point lower? 2 points? 3 points? This is not a statistical question but, rather, a clinical one, based on our knowledge of the condition, the scale, and the intervention. If the new treatment is significantly faster, cheaper, or—if it's a drug—has a better side effect profile, we may be willing to accept a lower score (that is, somewhat poorer adjustment) than if the new therapy does not offer these advantages. As Kendall and others (5) point out, though, there's a trade-off in the choice of this interval. The smaller its value, the more similar the treatments must be but the harder it is to demonstrate equivalence statistically. Conversely, it is easier to show equivalence with wider intervals, but then we

Figure 1 Hypothetical results of a new and standard therapy



have to accept bigger differences between the 2 groups and still say they're not different.

There are 2 approaches to equivalence testing. The 2-tailed approach tries to show that the 2 means or proportions are similar to each other; that is, that one is neither much larger nor much smaller than the other. The 1-tailed method is far more common and tests whether the second mean or proportion is different only in 1 direction. This is also referred to as noninferiority testing, because it is often used to see whether a new therapy isn't any worse than usual treatment. We don't care whether it's better—in fact, we'd be ecstatic—we merely want to insure that it's not significantly worse. The 2-tailed method is certainly theoretically important. However, on a practical level, it is much more likely that we would be interested in showing that the new treatment is not worse than the standard (noninferiority testing), so we will restrict ourselves to that situation.

The first step is to use our clinical judgement to define the equivalence interval, which we designate as δ. Using the previous example, assume that we'd accept a difference of 20% at most, which translates into δ = 2 points. That means that the mean for the new treatment (M_2) cannot be less than 8 if it is to be deemed noninferior.

Now let's bring the first point into play and redefine the null hypothesis. Instead of the usual nill hypothesis of no difference, we say that the null hypothesis is

$$H_0: \mu_1 - \mu_2 > \delta \qquad [3]$$

(or in English, the first mean is more than δ points greater than the second), and the alternative hypothesis is

$$H_A: \mu_1 - \mu_2 \leq \delta \qquad [4]$$

(that is, the difference between the means is less than δ, which also covers the possibility that μ_2 is larger than μ_1). Note that if δ = 0, these are simply the null and alternative hypotheses for a 1-sided *t*-test.

This means that, if we can reject the null hypothesis, we are left by default with the alternative hypothesis that the difference between the means (or proportions) is probably correct. The test for this (a *t*-test if the sample sizes are small or a *z*-test if they are above 10 or so) looks very similar to the usual one, with the exception of δ in the numerator:

$$t(df) = \frac{(M_1 - M_2) - \delta}{S_{M_1 - M_2}} \qquad [5]$$

where $S_{M_1 - M_2}$ is the standard error of the difference:

$$S_{M_1 - M_2} = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right] \times \left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \qquad [6]$$

and where $df = (n_1 + n_2 - 2)$, the n is the sample size in each group, and the s is the standard deviation. If we are dealing with proportions rather than means, then we simply replace M_1 with p_1 and M_2 with p_2 in Equation [5] (the proportions in each group), and use Equation [7] for the standard error

$$S_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \qquad [7]$$

An Example

Let's work through an example. Assume that we specified ahead of time that we would consider 2 treatments for sociophobia equivalent if the new one worked for at least 85% as many patients as did the usual therapy. What we actually find is that, with 20 patients per group, 75% improve on the standard therapy, A (that is, $p_A = 0.75$), and 70% improve with the new treatment, B ($p_B = 0.70$). Since 15% of 0.75 is 0.083, we set δ to be 0.083. Thus, the null hypothesis is

$$H_0 : p_A - p_B > 0.083$$

and the alternative hypothesis is

$$H_A : p_A - p_B \le 0.083$$

Spelled out, the null hypothesis is that the proportion of successful patients in the standard therapy group is more than 0.083 better than the proportion of successful patients in the new treatment group; the alternative hypothesis is that the difference in proportions is less than or equal to 0.083. This is shown in Figure 1.

Using Equation [7], we find that the standard error of the difference between these 2 proportions is 0.141, and therefore

$$t(38) = \frac{(0.75 - 0.70) - 0.083}{0.141} = 0.234$$

Since this is smaller than the critical value of 1.645 that we would need to reject a 1-tailed hypothesis at the 0.05 level, we cannot reject the null hypothesis, and we have to conclude that the 2 treatments are not equivalent.

Sample Size and Power

In the example we just worked through, it would seem at first glance that the 2 treatments should have come out as equivalent. We said that we would accept a 15% difference from the effectiveness of the standard treatment or roughly 0.083 less effectiveness than that demonstrated for Treatment A, 0.75. The success rate for Treatment B, 0.70, seems to be this much less; in fact, the difference is only 1 subject per group (that is, 15/20 for A vs 14/20 for B). Why do these results seem to be counterintuitive?

Simply examining raw differences overlooks 2 important points. First, we cannot just look at the difference between the 2 groups. As is always the case, the means or proportions that emerge from a study are sample estimates of the true population parameters. Because of this, they deviate from the real values to some degree. The amount of this deviation is related to the variability in what is being measured (for example, the standard deviation) and the sample size, and these have to be taken into account when we test to see whether the difference is statistically significant.

The second point is that, in testing for equivalence, we reverse the usual meanings of the null and alternative hypotheses. This means that we have to alter both our interpretations of type I and type II errors and what we mean by power. In both NHST and equivalence testing, a type I error occurs when we conclude that the null hypothesis is false when in fact it is true; a type II error occurs when we erroneously conclude that the null hypothesis is true when it is not. Power is the ability to reject the null hypothesis when it is false.

In noninferiority testing, though, the null hypothesis is that the standard treatment is better than the new one. This means that

1. A type I error occurs when we say that the 2 treatments are equivalent, when in fact the standard treatment is better.

2. A type II error occurs when we conclude that the standard treatment is better, when it fact the treatments are equivalent.

3. Power is the probability of accepting that the groups are equivalent when in fact they are equivalent (6).

The issue, then, is the power of the test. As we would expect, with only 20 subjects per group, the power of the tests we just ran is low. The reality is that equivalence testing is at times less powerful than testing for a difference. That is, we would need more subjects to test when a given difference is within the equivalence interval than when we test to see whether the 2 groups differ from each other.

To determine why this is so, let's take a look at the equations to calculate sample size (7). For the equivalence of 2 means, the equation is

$$n = \frac{2(z_\alpha + z_{\beta/2})^2 s^2}{[(M_1 - M_2) - \delta]^2} \qquad [8]$$

and for the equivalence of 2 proportions, it is

$$n = \frac{(z_\alpha + z_{\beta/2})^2 [p_1(1 - p_1) + p_2(1 - p_2)]}{[(p_1 - p_2) - \delta]^2} \qquad [9]$$

(If we are testing whether the 2 means or proportions are identical, then the denominator becomes simply δ^2.) These are very similar to the usual equations for sample size determination (8) with 2 differences, one that has a small effect on the sample and one that has a potentially large effect. The difference with the small effect is that we now want to minimize the type II error rather than the type I error, as in NHST. Consequently, the values of α and β are reversed, in that we set β at 0.05 and α at 0.10 or 0.20. The difference that has a potentially large effect is the δ in the denominator.

If we use Equation [9] to figure out the required sample size for the example (setting $\alpha = 0.20$, $\beta = 0.05$, and therefore power = 0.95), we will find it to be 2255 subjects per group! Conversely, with only 20 subjects per group, the power of the test to detect a difference of 0.083 between these proportions is less than 30%.

The sample size to test for equivalences is not always larger than that for testing for differences; again, it depends on the value of δ. Table 1 gives the sample sizes needed to test for noninferiority for various combinations of proportions in the standard treatment group, p_s, and the experimental group, p_e,

Table 1 Sample size per group for 1-tailed equivalence testing on proportions

p_s	p_e	δ				$p_s > p_e$
		0.10	0.15	0.20	0.25	
0.90	0.90	117	54	31	21	–
	0.85	545	140	64	37	540
	0.80		625	159	72	157
	0.75			691	175	79
	0.70				745	49
0.80	0.80	200	89	51	33	–
	0.75	860	216	96	54	862
	0.70		914	229	102	231
	0.65			956	238	109
	0.60				984	64
0.70	0.70	260	116	65	42	–
	0.65	1080	270	120	67	1084
	0.60		1109	276	122	281
	0.55			1126	280	128
	0.50				1130	74
0.60	0.60	296	132	74	47	–
	0.55	1203	300	133	74	1208
	0.50		1207	300	133	305
	0.45			1199	298	136
	0.40				1178	77
0.50	0.50	309	137	77	49	–
	0.45	1288	306	135	76	1233
	0.40		1207	300	133	305
	0.35			1174	292	134
	0.30				1130	74
0.40	0.40	296	132	74	47	–
	0.35	1154	288	128	71	1159
	0.30		1109	276	122	281
	0.25			1053	262	120
	0.20				984	64
0.30	0.30	260	116	65	42	–
	0.25	982	246	109	62	986
	0.20		914	229	102	231
	0.15			835	209	95
	0.10				745	49

Notes: p_s = proportion in standard group; p_e = proportion in experimental group. α = 0.20, β = 0.05 for equivalence testing; α = 0.05, β = 0.20 for $p_s > p_e$.

with $\alpha = 0.20$ and $\beta = 0.05$. For comparison, the last column is the sample size required for the traditional NHST that $p_s > p_e$. When $\delta = 2\ (p_s - p_e)$, the sample sizes are about equal for both types of tests. When $\delta < 2\ (p_s - p_e)$, the sample size for equivalence testing is larger than for difference testing. When $\delta > 2\ (p_s - p_e)$, it is smaller. Note that, when δ is larger than $2\ (p_s - p_e)$, the change in sample size is relatively small. However, when δ is smaller than $2\ (p_s - p_e)$, the sample size increases rapidly and exponentially. The same relation holds for testing the noninferiority of means, with M_s and M_e replacing p_s and p_e.

Summary

At times, despite all philosophical injunctions to the contrary, we have to prove that there are no unicorns. The solution, as we've seen, is to reverse the meanings of the null and alternate hypotheses and try to show that the null hypothesis of a difference can be rejected. This leaves us, by elimination, with the alternative—that there is no difference (or at least, that the difference is small enough for us to ignore). The issue is that the closer the groups must be to be considered equivalent, the larger the sample size required. This is entirely analogous to the situation for the traditional NHST: larger sample sizes are needed to detect smaller differences between groups. In both cases, sample size is like magnification with a microscope: the smaller the object that's being observed, the more magnification we need.

Acknowledgements

I am deeply grateful to Mr Malcolm Binns and Dr Elizabeth Linn for their careful reading of an earlier draft and for their many helpful comments.

References

1. MacMillan HL, Boyle MH, Wong MY, Duku EK, Fleming JE, Walsh CA. Slapping and spanking in childhood and its association with lifetime prevalence of psychiatric disorders in a general population sample. CMAJ 1999:161;805–9.
2. Canner PL. Aspirin in coronary heart disease: comparison of six clinical trials. Isr J Med Sci 1983:19;413–23.
3. Bollini P, Pampallona S, Tibaldi G, Kupelnick B, Munizza C. Effectiveness of antidepressants. Meta-analysis of dose-effect relationships in randomised clinical trials. Br J Psychiatry 1999:174;297–303.
4. Cohen J. The earth is round (p < .05). Am Psychol 1994;12:997–1003.
5. Kendall PC, Marrs-Garcia A, Nath SR, Sheldrick RC. Normative comparisons for the evaluation of clinical significance. J Consult Clin Psychol 1999;67:285–99.
6. Hatch JP. Using statistical equivalence testing in clinical biofeedback research. Biofeedback and Self-Regulation 1996;21:105–19.
7. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. Psychol Bull 1993;113:553– 65.
8. Streiner DL. Sample size and power in psychiatric research. Can J Psychiatry 1990;35:616–20.

Manuscript received May 2003 and accepted June 2003.
This is the 23rd article in the series on Research Methods in Psychiatry. For previous articles, please see Can J Psychiatry 1990;35:616–20, 1991;36:357–62, 1993;38:9–13, 1993;38:140–8, 1994;39:135–40, 1994;39:191–6, 1995;40:60–6, 1995;40:439–44, 1996;41:137–43, 1996;41:491–7, 1996;41:498–502, 1997;42:388–94, 1998;43:173–9, 1998;43:411–5, 1998;43:737–41, 1998;43:837–42, 1999;44:175–9, 2000;45:833–6, 2001;46:72–6, 2002;47:68–75, 2002;47:262–6, 2002;47:552–6.

[1]Director, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care; Professor, Department of Psychiatry, University of Toronto, Toronto, Ontario.
Address for correspondence: Dr DL Streiner, Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, ON M6A 2E1
e-mail: dstreiner @ klaru-baycrest.on.ca

Résumé : Les licornes existent vraiment : une leçon consistant à « confirmer » l'hypothèse nulle

Les cours d'introduction aux statistiques nous enseignent que nous ne pouvons jamais confirmer l'hypothèse nulle; nous ne pouvons que la rejeter ou refuser de la rejeter. Cependant, il y a des occasions où il est nécessaire de tenter de prouver l'inexistence d'une différence entre les groupes. Cela se produit le plus souvent lorsqu'on compare un nouveau traitement avec un traitement établi, et que l'on démontre que la nouvelle intervention n'est pas inférieure à la régulière. Cet article présente d'abord la logique des tests de « non-infériorité » en distinguant l'hypothèse nulle (celle que nous tentons d'annuler) de l'hypothèse « rien » (il n'y a aucune différence), en inversant les rôles des hypothèses nulles et autres, et en définissant un intervalle dans lequel on détermine que les groupes sont équivalents. Nous présentons ensuite un exemple et indiquons comment calculer les tailles d'échantillons pour des études de non-infériorité.