The Null Hypothesis

Geoffrey R. Loftus

University of Washington

Send correspondence to: Geoffrey R. Loftus
Department of Psychology, Box 351525
University of Washington
Seattle, WA 98195-1525
gloftus@u.washington.edu
206 543-8874

In many sciences including for example, ecology, medicine, and psychology, null hypothesis significance testing (NHST) is the primary means by which the numbers comprising the data from some experiment are translated into conclusions about the question(s) that the experiment was designed to address. In this entry, I make three main points. First, I provide a brief description of NHST and within the context of NHST, define the most common incarnation of a null hypothesis. Second, I sketch other less common forms of a null hypothesis. Third, I articulate a number of problems with using null hypothesis-based data analysis procedures.

NHST and the Null Hypothesis

Most experiments entail measuring the effect(s) of some number of independent variables on some dependent variable.

An example experiment

In the simplest sort of experimental design, one measures the effect of a single independent variable, say amount of information held in short-term memory on a single dependent variable, say reaction time to scan through this information. To pick a somewhat arbitrary example from cognitive psychology, consider what is known as a *Sternberg experiment,* in which a short sequence of *memory digits* (e.g., “34291”) is read to an observer who must then decide whether a single, subsequently presented *test digit* was part of the sequence. Thus for instance, given the memory digits above, the correct answer would be “yes” for a test digit of “2” but “no” for a test digit of “8”. The independent variable of “amount of information held in short-term memory” can be implemented by varying *set size* which is the number of memory digits presented: in different conditions, set size might be, say, 1, 3, 5 (as in the example), or 8 presented memory digits. The number of different set sizes (here 4) is more generally referred to as the number of *levels* of the independent variable. The dependent variable is the reaction time measured from the appearance of the test digit to the observer’s response. Of interest in general is the degree to which the magnitude of the dependent variable (here, reaction time) depends on the level of the independent variable (here set size).

Sample and population means

Typically, the principal dependent variable takes the form of a *mean*. In this example mean reaction time for a given set size could be computed across observers. Such a computed mean is called a *sample mean*, referring to its having been computed across an observed sample of numbers. A sample mean is construed as an estimate of a corresponding *population mean* which is what the mean value of the dependent variable would be if all observers in the relevant population were to participate in a given condition of the experiment. Generally, conclusions from experiments are meant to apply to population means. Therefore, the measured sample means are only interesting insofar as they are estimates of the corresponding population means.

Notationally, the sample means are referred to as the M_j's while the population means are referred to as the μ_j's. For both sample and population means, the subscript "j" indexes the level of the independent variable; thus in our example M_2 would refer to the observed mean reaction time of the second set-size level, i.e., set size = 3 and likewise, μ_2 would refer to the corresponding, unobservable population mean reaction time corresponding to set size = 3.

Two competing hypotheses

NHST entails establishing and evaluating two mutually exclusive and exhaustive hypotheses about the relation between the independent variable and the dependent variable. Usually, and in its simplest form, the null hypothesis (abbreviated H_0) is that the independent variable has *no effect* on the dependent variable, while the alternative hypothesis (abbreviated H_1) is that the independent variable has *some* effect on the dependent variable. Note an important asymmetry between a null hypothesis and an alternative hypothesis: a null hypothesis an *exact* hypothesis while an alternative hypothesis is an *inexact* hypothesis. By this is meant that a null hypothesis can only be correct in only one way, *viz,* the μ_j's are all equal to one another, while there are an infinite number of ways in which the μ_j's can be different from one another, i.e., an infinite number of ways in which an alternative hypothesis can be true.

Decisions based on data

Having established a null and an alternative hypothesis that are mutually exclusive and exhaustive, the experimental data are used to—roughly speaking; see Point 2 below—decide between them. The technical manner by which one makes such a decision is beyond the scope of this entry, but two remarks about the process are appropriate here.

1. A major ingredient in the decision is the variability of the M_j's. To the degree that the M_j's are close to one another, evidence ensues for possible equality of the μ_j's and, *ipso facto*, validity of the null hypothesis. Conversely, to the degree that the M_j's differ from one another, evidence ensues for associated differences among the μ_j's and, *ipso facto*, validity of the alternative hypothesis.

2. The asymmetry between the null hypothesis (which is exact) and the alternative hypothesis (which is inexact) sketched above implies an associated asymmetry in conclusions about their validity. If the M_j's differ sufficiently, one "rejects the null hypothesis" in favor of accepting the alternative hypothesis. However if the M_j's do not differ sufficiently, one does not "accept the null hypothesis", but rather one "fails to reject the null hypothesis". The reason for the awkward, but logically necessary, wording of the

latter conclusion is that, because the alternative hypothesis is inexact, one cannot generally distinguish a genuinely true null hypothesis on the one hand from an alternative hypothesis entailing very small differences among the μ_j's on the other hand.

Multifactor designs: Multiple null hypothesis-alternative hypothesis pairings

So far I have described a simple design in which the effect of a single independent variable on a single dependent variable is examined. Many, if not most experiments, utilize multiple independent variables, and are known as *multifactor designs* ("factor" and "independent variable" are synonymous). Continuing with the example experiment, imagine that in addition to measuring effects of set size on reaction time in a Sternberg task, one also wanted to simultaneously measure effects on reaction time of the test digit's visual contrast (informally, the degree to which the test digit stands out against the background). One might then *factorially combine* the four levels of set size (now called "Factor 1") with, say, two levels, "high contrast" and "low contrast," of test-digit contrast (now called "Factor 2"). Combining the four set-size levels with the two test-digit contrast levels would yield 4 x 2 = 8 separate conditions. Typically, three independent NHST procedures would then be carried out, entailing three null hypothesis-alternative hypothesis pairings. They are:

1. For the set size main effect:

H_0: Averaged over the two test-digit contrasts, there is no set-size effect

H_1: Averaged over the two test-digit contrasts, there is a set-size effect

2. For the test-digit contrast main effect:

H_0: Averaged over the four set sizes, there is no test-digit contrast effect

H_1: Averaged over the four set sizes, there is a test-digit contrast effect

3. For set-size x test-digit contrast *interaction*:

Two independent variables are said to *interact* if the effect of one independent variable depends on the level of the other independent variable. As with the main effects, interaction effects are immediately identifiable with respect to the M_j's; however again as with main effects, the goal is to decide whether interaction effects exist with respect to the corresponding μ_j's. As with the main effects, NHST involves pitting a null hypothesis against an associated alternative hypothesis.

H_0: With respect to the μ_j's, set size and test-digit contrast do not interact.

H_1: With respect to the μ_j's, set size and test-digit contrast do interact.

The logic of carrying out NHST with respect to interactions is the same as the logic of carrying out NHST with respect to main effects. In particular, with interactions as with main effects, one can reject a null hypothesis of no interaction, but one cannot accept a null hypothesis of no interaction.

Non-“Zero-Effect” Null Hypotheses

The null hypotheses described above imply “no effect” of one sort or another—either no main effect of some independent variable, or no interaction between two independent variables. This kind of “no-effect” null hypothesis is by far the most common null hypothesis to be found in the literature. Technically however, a null hypothesis can be any *exact hypothesis*; that is the null hypothesis of “all μ_j’s are equal to one another” is but one special case of what a null hypothesis can be.

To illustrate another form, let us continue with the first, simpler Sternberg-task example (set size is the only independent variable), but imagine that prior research justifies the assumption that the relation between set size and reaction time is *linear*. Suppose further that research with *digits* has yielded the conclusion that reaction time increases by 35 ms for every additional digit held in short-term memory; i.e., that if reaction time were plotted against set size, the resulting function would be linear with a slope of 35 ms.

Now let us imagine that the Sternberg experiment is done with words rather than digits. One could establish the null hypothesis that “short-term memory processing proceeds at the same rate with words as it does with digits”, i.e., that the slope of the reaction time versus set-size function would be 35 ms for words just as it is known to be with digits. The alternative hypothesis would then be “for words, the function’s slope is anything other than 35 ms.” Again the fundamental distinction between a null and alternative hypothesis is that the null hypothesis is exact (35 ms/digit), while the alternative hypothesis is inexact (anything else). This distinction would again drive the asymmetry between conclusions, articulated above: a particular pattern of empirical results could logically allow “rejection of the null hypothesis; i.e., acceptance of the alternative hypothesis” but not “acceptance of the null hypothesis”.

Problems with NHST

No description of NHST in general, or a null hypothesis in particular is complete without at least a brief account of serious problems that accrue when NHST is the sole statistical technique used for making inferences about the μ’s from the M_j’s. Very briefly, three of the major problems involving a null hypothesis as the centerpiece of data analysis are these.

A null hypothesis cannot be literally true

In most sciences it is almost a self-evident truth that any independent variable must have some effect, even if small, on any dependent variable. This is certainly true in psychology. In the Sternberg task, to illustrate, it is simply implausible that set size would have literally *zero* effect on reaction time, i.e., that is that the μ_j’s corresponding to the different set sizes would be *identical* to an infinite number of decimal places. Therefore, rejecting a null hypothesis—which, as noted, is the only strong conclusion that is possible within the context of NHST—tells the investigator nothing that the investigator should have been able to realize was true beforehand. Most investigators do not recognize this, but that does not prevent it from being so.

Human nature makes acceptance of a null hypothesis almost irresistible

Earlier I articulated why it is logically forbidden to accept a null hypothesis. However, human nature dictates that people do not like to make weak yet complicated conclusions such as "We fail to reject the null hypothesis." Scientific investigators, generally being humans, are not exceptions. Instead, a "fail to reject" decision, dutifully made in an article's results section, almost inevitably morphs into "the null hypothesis is true" in the article's discussion and conclusions sections. This kind of sloppiness, while understandable, has led to no end of confusion and general scientific mischief within numerous disciplines.

NHST emphasizes barren, dichotomous conclusions

Earlier, I described that the pattern of population means—the relations among the unobservable μ_j's—are of primary interest in most scientific experiments, and that the observable M_j's are estimates of the μ_j's. Accordingly, it should be of great interest to assess how *good* are the M_j's as estimates of the μ_j's. If, to use an extreme example, the M_j's were perfect estimates of the μ_j's there would be no need for statistical analysis: the answers to any question about the μ_j's would be immediately available from the data. To the degree that the estimates are less good, one must exercise concomitant caution in using the M_j's to make inferences about the μ_j's.

None of this is relevant within the process of NHST, which does not in any way emphasize the degree to which the M_j's are good estimates of the μ_j's. In its typical form, NHST allows only a very limited assessment of the nature of the μ_j's: Are they all equal or not? Typically, the "no" or "not necessarily no" conclusion that emerges from this process is woefully insufficient to evaluate the totality of what the data might potentially reveal about the nature of the μ_j's.

An alternative that is gradually emerging within several NHST-heavy sciences—an alternative that is common in the natural sciences—is the use of *confidence intervals* which assess directly how good is a M_j as an estimate of the corresponding μ_j. Very briefly, a confidence interval is an interval constructed around a sample mean that, with some pre-specified probability (typically 95%), includes the corresponding population mean. A glance at a set of plotted M_j's with associated plotted confidence intervals provides immediate and intuitive information about (a) the most likely pattern of the μ_j's and (b) the reliability of the pattern of M_j's as an estimate of the pattern of μ_j's. This in turn provides immediate and intuitive information both about the relatively uninteresting question of whether some null hypothesis is true, and about the much more interesting questions of what the pattern of μ_j's actually *is* and how much belief can be placed in it based on the data at hand.

Further readings

Fidler, F.. & Loftus, G.R. (in press). Why hypothesis testing is misunderstood: Hypotheses and Data.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. Current Directions in Psychological Science, 161-171.

Loftus, G.R. & Masson, M.E.J. (1994) Using confidence intervals in within-subjects designs. Psychonomic Bulletin & Review, 1, 476-490.

Other relevant entries

Alpha

Analysis of Variance (ANOVA)

Beta

Chi-squared Test

Confidence Intervals

Contrasts

Decision Rule

Directional Hypotheses

F Test

Hypothesis

Hypothesis Testing

Inference (Inductive and Deductive)

Level of Significance

Logic of Scientific Discovery, The (Popper)

Nonsignificance

Population

Power Analysis

p-value

Research Hypothesis

Significance (Statistical Significance)

Significance Level

Simple Main Effects

Statistical Power Analysis for the Behavioral Sciences (Cohen)

Two-tailed Test

Type I Error

Type II Error