What are Null Hypotheses? The Reasoning Linking Scientific and Statistical Hypothesis Testing

Anton E. Lawson
Arizona State University, Tempe, AZ, USA
anton.lawson@asu.edu

Abstract

We should dispense with use of the confusing term *null hypothesis* in educational research reports. To explain why the term should be dropped, the nature of, and relationship between, scientific and statistical hypothesis testing is clarified by explication of (a) the scientific reasoning used by Gregor Mendel in testing specific hypotheses derived from his general inheritance theory and (b) the statistical reasoning used in applying the chi-square statistic to his experimental data. The Mendel example is followed by application of the same pattern of scientific and statistical reasoning to educational examples. A better understanding of the related, but separate, processes of scientific and statistical hypothesis testing, including the role of scientific hypotheses (i.e., proposed explanations) and scientific predictions (i.e., expected test results), not only reveals why null statistical hypotheses and predictions need not be stated, but also reveals how we can improve the clarity of our research reports and improve the quality of the research reported by insuring that alternative scientific hypotheses and theories are in fact tested.

P. Eastwell (personal communication, July 5, 2006) asked readers to consider the confusing and possible misuse of the term *null hypothesis* in the context of research reports. In Eastwell's words:

> We distinguish a prediction (an educated guess about the expected outcome of a test) from a hypothesis (a possible explanation for the observed facts and laws). Does it follow that science education researchers should now dispense with the use of the term null hypothesis in circumstances where it is really a null prediction that is being tested?

I think it would be helpful if science education researchers dispensed with use of the term null hypothesis under any circumstances, primarily because the term comes from the field of statistics and its relationship to scientific hypothesis testing is seldom, if ever, made clear. Hence its use in educational research often leads to confusion and may even limit research quality by restricting the number of scientific hypotheses generated and tested. As will become clear in this paper, the term null prediction is also not required.

Allow me to attempt to clarify by explicating important similarities and differences between scientific and statistical hypothesis testing in the context of a crucial experiment conducted by Gregor Mendel to test his classic inheritance theory. The example will consider Mendel's theory, the reasoning behind how he tested it, and how statistical hypothesis testing could have been used to determine the extent to which departures of Mendel's observed scientific results from his predicted scientific results were due to chance or due to faulty scientific hypotheses. The Mendel example (after Lawson, Oehrtman, & Jensen, 2008, with kind permission of Springer Science and Business Media) will be followed by some educational examples and implications.

Mendel's Experiment and the Reasoning Guiding Scientific Hypothesis Testing

As you may recall, Mendel's theory proposed that dominant and recessive genes exist in pairs (e.g., *YY*, *rr*) and that the genes of a pair separate and pass independently to egg and sperm cells (i.e., the gametes). Then during fertilization, the separated genes (e.g., *Y*, *r*) recombine randomly in zygotes (i.e., in fertilized eggs). To test these theoretical claims (we will call them scientific hypotheses as

they are part of Mendel's more general and complex inheritance theory), Mendel conducted a two-part experiment with pea plants.

During the first part of his experiment, Mendel crossed/mated pure-breeding pea plants that produced yellow-round seeds (presumably with the dominant *YYRR* genotype) with pure-breeding pea plants that produced green-wrinkled seeds (presumably with the recessive *yyrr* genotype). All of the offspring from this cross produced yellow-round seeds (presumably with the mixed *YyRy* genotype). During the second part of his experiment, Mendel crossed the above offspring of the first generation. However, before these plants grew and matured to produce their own seeds, his scientific hypotheses (i.e., his explanatory claims) allowed him to make a very specific prediction (i.e., an expected result of a planned test given that his explanatory claims are correct) about the color and shape of the seeds that should be produced. Specifically his hypotheses led him to expect (predict) that the next generation seeds would appear with a 9:3:3:1 ratio of seed types (i.e., 9 yellow-round: 3 yellow-wrinkled: 3 green-round: 1 green-wrinkled).

When cast in the form of a hypothetico-deductive argument, Mendel's *If/and/then* reasoning looks like this:

> *If* . . . dominant and recessive paired genes pass independently to gametes and recombine randomly in zygotes (scientific hypotheses),
> *and* . . . pea plants presumably with the *RrYy* genotype for seed color and shape are crossed (planned scientific test),
> *then* . . . we should observe a seed color/shape ratio of 9:3:3:1 in their offspring (scientific prediction).

When Mendel collected, observed, and counted the 556 seeds that were produced in these offspring, he found that 315 were yellow-round, 108 were yellow-wrinkled, 101 were green-round, and 32 were green-wrinkled. These numbers constitute his observed scientific result.

What conclusion should Mendel draw from this scientific result? Were his scientific hypotheses supported? A quick calculation reveals that a 9:3:3:1 ratio of seed types should have produced about 313 yellow-round seeds, 104 yellow-wrinkled seeds, 104 green-round seeds, and 35 green-wrinkled seeds. These predicted numbers are very similar to the observed numbers. Therefore, Mendel concluded that the slight departures between his predicted and observed results were random in nature and that his scientific hypotheses (and the more general theory of which they were a part) were supported.

But were the slight departures between his predicted and observed results really due to chance? Or was there in fact something wrong with Mendel's hypotheses? Of course Mendel had no way of knowing because the process of statistical hypothesis testing, the way of knowing, had not been invented in 1865 when Mendel published his results. Consequently, let's briefly consider the reasoning guiding statistical hypothesis testing to see how it can answer this key question.

The Reasoning Guiding Statistical Hypothesis Testing

Consider testing a coin for "fairness." Assuming that one has a fair coin, when tossed, one would predict that it would land heads about half the time and tails the other half. So to test a coin for fairness (i.e., to test the statistical null hypothesis that you have a fair coin), you could toss it 100 times. Suppose it lands heads 47 times and tails 53 times. You probably would not be too bothered by this. Your observed ratio of 47:53 is quite close to the predicted 50:50 ratio. However, what would you conclude if your observed ratio turned out 35:65, if it turned out 5:95? Obviously, there

will be some point when you no longer conclude that the observed result matches your prediction. Would you conclude that a coin that lands heads only 5 out of 100 tosses is fair? You probably would not. Said another way, you would probably reject the statistical null hypothesis that the coin is fair (i.e., that both probabilities are 0.50).

How then can we know when a departure from a predicted scientific result is due to chance or to a faulty scientific hypothesis? Although we can never know for sure, it turns out that, thanks to statistical hypothesis testing, we can nevertheless estimate the likelihood of various departures from predictions. In other words, even though we cannot be certain about the truth or falsity of any particular scientific hypothesis, at least we can estimate our degree of uncertainty.

Mathematicians have invented formulas to generate such uncertainty estimates. One formula, the chi-square formula introduced in 1900 by Karl Pearson (Walker, 1958), can be used in the present context. The chi-square formula calculates a single value (a statistic) that we can compare to values listed in a statistical table to tell us what we need to know. The chi-square value/statistic (i.e., χ^2) is calculated by comparing predicted and observed results. As observed results deviate farther from predicted results, the chi-square values increase. So a relatively large χ^2 value means that the results are probably not due to chance. For example, in a coin toss situation we have two categories of data with predicted numbers of 50 heads and 50 tails and observed numbers of 47 heads and 53 tails. So the χ^2 calculation looks like this:

$$\chi^2 = \frac{(47 \text{ heads} - 50 \text{ heads})^2}{(50 \text{ heads})} + \frac{(53 \text{ tails} - 50 \text{ tails})^2}{(50 \text{ tails})} = \frac{(-3)^2}{50} + \frac{3^2}{50} = 0.36$$

How does one interpret this value of 0.36? Suppose 100 people each have a fair coin. Suppose further that each person flips his/her fair coin 100 times and records the number of heads that turn up. If we now create a graph plotting these numbers versus their frequency, we will end up with a distribution most likely with around 50 heads (or 50 tails) as the modal value. Suppose further that each person calculates a χ^2 value for the results of his/her 100 tosses and we then plot the various χ^2 values versus their frequency. Because the smallest possible value is zero (obtained when observed and predicted numbers are the same), we will end up with a distribution of 100 chi-square values extending to the right of zero with increasingly large values being less and less probable. This is called a sampling distribution. Statisticians have compiled the probabilities associated with several such values and sampling distributions and listed them in statistical tables. Consequently, if we have a new coin and want to know if it is fair, we can toss it 100 times and count the number of times it turns up heads (or tails). We can then use the observed results and the chi-square formula to calculate a χ^2 value and compare it to the values in the appropriate statistical table.

To summarize, we have just tested a descriptive statement (i.e., a statistical null hypothesis) about an unknown parameter. In this case the statistical null hypothesis is that both probabilities are 0.50. And just like in causal scientific hypothesis testing, we used hypothetico-deductive reasoning to do so. That is:

> *If* . . . the probability of landing heads is 0.50 (fair-coin statistical null hypothesis),
> *and* . . . we flip a coin 100 times and compute a chi-square value for the result (planned statistical test),
> *then* . . . the chi-square value should fall well within the sampling distribution as reflected by the values and probabilities that appear in the appropriate statistical table (statistical prediction).

And . . . the calculated value of 0.36 derived from our result of 47 heads and 53 tails does fall well within the sampling distribution. More specifically, the appropriate table tells us that a value of 0.36 will occur due to chance alone between 50% and 70% of the time such a test is conducted (observed statistical result).
Therefore . . . most likely the probability of landing heads (or tails) really is 0.50. Thus, we can be quite confident that the coin is fair (statistical conclusion).

Calculating and Interpreting a Chi-Square Value for Mendel's Results

Let's now return to Mendel's experiment and use his predicted and observed results to calculate a chi-square value and see if the departures are likely due to chance. The calculated χ^2 value turns out to be 0.51. A quick check of the appropriate statistical table shows this value (with three degrees of freedom) associated with probabilities 0.90 and 0.95. This means that between 90% and 95% of the time, chance variations would result in a greater departure from a true 9:3:3:1 distribution than do Mendel's results. In other words, it seems safe to conclude that the difference between Mendel's observed and predicted results are due to chance. Therefore, not only is the descriptive statistical null hypothesis supported, but so are Mendel's causal scientific hypotheses and his general inheritance theory.

Table 1 summarizes both scientific and statistical hypotheses in terms of the *If/and/then* arguments in which hypotheses are tested through the generation of specific predictions. As you can see, both processes involve prediction generation followed by data collection and the comparison of predicted and observed results. However, the goal of scientific hypothesis testing is to test scientific hypotheses, which are causal in nature, while the goal of statistical hypothesis testing is to test statistical null hypotheses, which are descriptive in nature.

Note also that the scientific prediction (i.e., we should observe a seed color/shape ratio of 9:3:3:1 in the offspring plants) and the statistical null hypothesis (i.e., a seed color/shape ratio of 9:3:3:1 exists in the offspring) sound much the same. In the former case, however, we have a statement about how a scientific test should turn out assuming that a causal scientific hypothesis is correct, while in the latter case we have a descriptive statistical hypothesis about the nature of seed colors and shapes.

Educational Examples and Implications

In the first edition of their classic statistics textbook, Glass and Stanley (1970) discuss the evaluation of three teaching methods (i.e., textbook, programmed textbook, and computer-level program) on reading comprehension. The evaluation involves random assignment of students into three treatment groups, one group for each teaching method. Students are then administered a posttest to determine which method was most effective. During their discussion, Glass and Stanley state the experiment's null hypothesis as "the population means for the three teaching methods are equal." (p. 411)

As discussed, this statement represents a descriptive statistical hypothesis; not a causal scientific hypothesis. Unfortunately, in their example Glass and Stanley (1970) do not offer any causal scientific hypotheses. If scientific hypotheses were discussed, they would provide reasons/causes for the possible superiority of one treatment over the other(s) (e.g., programmed texts are better because they include frequent questions that provoke students to reflect on what they have read). Thus, an unmentioned hypothetico-deductive argument might go something like this:

If . . . provoking students to reflect on what they have read increases comprehension (scientific hypothesis),
and . . . some students read standard text while others read programmed text or a computer-level program and the three groups are then tested (planned scientific test),
then . . . mean test score of the programmed text students should be higher than those of the other two groups (scientific prediction). Or, stated as a statistical null hypothesis, the population means for the three teaching methods are equal.

Table 1
The Reasoning Guiding Scientific and Statistical Hypothesis Testing (Lawson, Oehrtman & Jensen, 2008)

Aspect of reasoning	Process	
	Scientific hypothesis testing	Statistical hypothesis testing
Hypotheses: *If* . . .	Dominant and recessive gene pairs pass independently to gametes and recombine randomly in pea plant zygotes (scientific hypotheses).	A seed color/shape ratio of 9:3:3:1 exists in the offspring (statistical null hypothesis).
Planned tests: *and* . . .	Cross pea plants presumably with the *RrYy* genotype for seed color and shape (planned scientific test).	Collect a sample of seeds and compute the value of our selected statistic (planned statistical test).
Predictions: *then* . . .	We should observe a seed color/shape ratio of 9:3:3:1 in the offspring (scientific prediction).	The value of the statistic should fall well within the sampling distribution (statistical prediction).
Results: *And/But* . . .	Of the 556 seeds, 315 were yellow-round, 108 were yellow-wrinkled, 101 were green-round, and 32 were green-wrinkled (observed scientific result).	The value for Mendel's observed results (Chi-square = 0.51, df = 3) falls well within the sampling distribution (observed statistical result).
Conclusions: *Therefore* . . .	Mendel's scientific hypotheses for pea plants and his general inheritance theory are supported (scientific conclusion).	The departure of Mendel's observed scientific results from the predicted scientific results are most likely due to random variation, so the statistical null hypothesis is supported (statistical conclusion).

This argument adds a critical component to Glass and Stanley's (1970) example; namely, a reason that one treatment is predicted to be superior to the other(s). Without such a reason, even if only implicitly held, the researchers would most likely not have conducted the experiment in the first place. Hence, by omitting discussion of possible reasons (i.e., scientific hypotheses), Glass and Stanley not only omit a critical aspect of the research process, they also fail to differentiate scientific hypothesis testing from statistical hypothesis testing.

Consider a second educational example. Suppose you are a high school biology teacher and have just taught a unit on Mendelian genetics. Upon testing your students you find that some of them did very well on the test while others did very poorly. Piagetian theory argues that intellectual development occurs in stages and that formal stage reasoning patterns are needed to understand theoretical concepts, such as many of those embedded in Mendelian genetics. Based on Piagetian theory, you suspect that some of your students may not yet have

developed the presumably necessary formal reasoning patterns. Consequently, you generate the following causal scientific hypothesis, planned test, and scientific prediction:

Scientific hypothesis. Formal stage reasoning patterns are necessary to understand Mendelian genetics.
Planned test. Assess students' stages of intellectual development and compare their stages with their understanding of Mendelian genetics as measured by test performance.
Scientific prediction. The concrete operational students should be the ones who fail the test, while the formal operational students should be the ones who pass the test.

In terms of statistics, you are predicting that the collective scores of the formal students will be significantly higher than those of the concrete students, where significantly refers to statistical significance. When stated in the null form, we get the following: The mean test scores of the formal and concrete students should be equal. If we conduct the appropriate statistical test and find that the mean test score of the formal students is in fact statistically higher than that of the concrete students, we can reject the statistical null hypothesis. This in turn allows us to accept the causal scientific hypothesis. In other words, we have support for the scientific hypothesis that formal stage reasoning patterns are needed to understand Mendelian genetics.

What then would you make of a report in which the author states:

> The following hypotheses were postulated and computed at the 0.05 level of significance:
> *Hypothesis 1*: Students grouped in heterogeneous cooperative groups will perform significantly higher than those grouped in friendship cooperative groups.
> *Hypothesis 2*: Students grouped in friendship cooperative groups will perform significantly higher than those grouped in traditional groups.

Are these scientific or statistical hypotheses? Clearly they are statistical hypotheses. Accordingly, it becomes incumbent upon the author to clearly state what the scientific theories and/or hypotheses are and why they led him/her to predict such a statistical outcome. In short, a solid research effort and a well-crafted research report should clearly identify:

1. The puzzling observation in need of explanation.
2. The general theory or theories that may offer a possible explanation(s).
3. Specific hypotheses derived from those theories that the study aims to test.
4. The research design, including the *If/and/then* argument identifying the reasoning linking the scientific hypothesis and the design (i.e., planned test) to clearly stated scientific prediction(s).
5. In the case of quantitative research, the specific statistic(s) used to determine the match between the scientific prediction(s) and the result(s). Note that there is no need to state statistical null hypotheses as doing so is likely to confuse readers.
6. The research results and the extent to which they match the scientific prediction(s).
7. A conclusion about the status of the tested scientific hypothesis/theory (i.e., supported, contradicted) including, if possible, ad hoc scientific hypotheses and suggestions for future research.

The next time you read, or perhaps write, an educational report, see if it spells out these critical elements and their connections. The implication is that becoming more conscious of how to conduct and report research aimed at testing scientific theories and hypotheses--not just statistical hypotheses and statistical null hypotheses--should improve the way science educators conceive of,

carry out, conduct, and report their research. This in turn should better inform and improve practice.

Acknowledgements

I would like to thank Mike Oehrtman and Jamie Jensen for their help in explicating the similarities and differences between scientific and statistical hypothesis testing and Peter Eastwell and several anonymous reviewers for their helpful comments on earlier versions of this manuscript.

References

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Lawson, A. E., Oehrtman, M., & Jensen, J. (2008). Connecting science and mathematics: The nature of scientific and statistical hypothesis testing. *International Journal of Science and Mathematics Education*, *6*, 405-416.

Walker, H. M. (1958). The contributions of Karl Pearson. *Journal of the American Statistical Association*, *53*(281), 11-22.

The Science Education Review (ISSN 1446 - 6120) is published by Science Time Education, "Willow Downs," M/S 623, Warwick, Queensland 4370 Australia.

The Science Education Review (*SER*) is an international, peer-reviewed periodical aiming to provide primary and high school teachers with the latest, and best, ideas in science education from around the world. *SER* also publishes original articles, including research articles, and much more.

Please visit ***The Science Education Review*** at http://www.ScienceEducationReview.com . ***SER On-Line*** offers individuals and institutions password and/or IP authenticated access to the content of all issues of *SER*.

Contributions are welcome, as are expressions of interest in joining the Editorial Review Board. The latter role requires the periodic review of submitted contributions, and is not onerous. Comments, questions, and article proposals may be sent to The Editor, Dr Peter H. Eastwell, at editor@ScienceEducationReview.com .

* * *