Getting What You Deserve from Data

Paul R. Cohen
Department of Computer Science
Lederle Graduate Research Center
University of Massachusetts, Amherst MA 01003
cohen@cs.umass.edu

The focus of this article will appear at first to be a narrow, prescriptive little corner of the methodological landscape. Data analysis is often dismissed as no more complicated than calculating some means and comparing them with t tests or the like. Consequently, experiments and analyses are inefficient, requiring more data than necessary to show an effect; they waste data, failing to show effects; and they sometimes induce hallucinations, suggesting effects that don't exist. I am the last person to suggest that methodology boils down to statistics [2, p. x], but bad analysis can spoil an entire research program, so warrants attention. I will discuss three common and easily fixed problems:

1. Accepting the null hypothesis, a misuse of statistical machinery.
2. Inadequate attention to sources of variance, leading to insignificant results and failure to notice interactions among factors.
3. Multiple pairwise comparisons, leading to nonexistent effects.

I have constructed a dataset to illustrate these problems. It contains hypothetical assessments of confidence for boys and girls in grades 4, 5, 6 and 7, and is called the *gender dataset*, henceforth. (Real studies of these factors are described in [1]; for real examples from Artificial Intelligence, see [2]; the gender dataset is available from cohen@cs.umass.edu).

	mean	std
boys	4.271	.531
girls	3.996	.888

Table 1: Hypothetical means and standard deviations for confidence scores for boys and girls averaged over students and grade levels.

1 Accepting the Null Hypothesis.

Suppose one has the hypothesis that girls and boys are equally confident. Mean confidence, averaged over grade levels, is shown in Table 1. A t test shows no effect of gender; boys' and girls' confidence levels are not significantly different; therefore, the hypothesis that boys and girls are equally confident is accepted.

This line of reasoning makes nonsense of statistical hypothesis testing. The logic of hypothesis testing is analogous to proof by contradiction: First, formulate a *null hypothesis*, denoted H_0, which is the complement of what you hope to show. Then, derive a *sampling distribution* of all possible sample results, given H_0. Then, if your sample result is very unlikely according to this H_0 sampling distribution, you may reject H_0 and accept the alternative, complement hypothesis.

The probability of incorrectly rejecting H_0, denoted p, is bounded by a parameter denoted α. Conventionally, researchers set $\alpha = .05$, so they will not reject H_0 unless the probability of an observed result given H_0 is $p < .05$.

Here's the catch: H_0 must be an identity, such as, "boys' confidence equals girls' confidence," otherwise, it is impossible to derive the sampling distribution. This means that the alternative hypothesis must be an inequality (e.g., boys are more confident, less confident, or simply not equally confident). And so, you can only *reject* an identity hypothesis; you can never accept one. Failure to reject an identity H_0 does not make the identity true. As we will see, tests fail for reasons that have nothing to do with the veracity of the null hypothesis, notably large sample variance or small sample size.

So how is one supposed to demonstrate identities within the framework of statistical hypothesis testing? For example, how can one show that boys and girls really are equally confident? You cannot *prove* anything statistically, but if you fail to reject the null hypothesis, you can then try to show that boys and girls are very unlikely to have different confidence levels. If you succeed, then you have accrued support for the null hypothesis (in addition to failing to reject it) and you may "accept" it.

One approach is to derive a *confidence interval* for the difference between boys' and girls' confidence. In essence, one attempts to show that the true difference in confidence falls in a narrow range with high probability. For example, given the data in the gender dataset, we can say with 95% confidence that the true difference between boys' and girls' confidence lies in the interval $.275 \pm .413$ (see [2, ch. 4] for formulae, etc.) This confidence interval is not centered around zero, nor is it narrow: Confidence scores in the in the gender dataset range from 2.0 to 5.0, and the width of the interval around the *difference* of boys' and girls' scores is .826. So for the gender dataset, we cannot find support for the identity hypothesis. Although we couldn't reject it, we also cannot accept it.

A second way to demonstrate identities depends on the *power* of statistical tests. Power is the probability that a test will reject the null hypothesis if it is false. Several factors affect power. Some tests are intrinsically more powerful than others. For example, one might test whether the means or the medians of two groups are significantly different, but a test of means will generally be more powerful than a test of medians because the mean summarizes more information about a group than the median. Power is also affected by the sample size and sample variance; in general, as the former increases and the latter decreases, a given test is increasingly likely to reject the null hypothesis correctly. Now, suppose an extremely powerful test fails to find a difference between boys' confidence and girls'. Then, you could argue that the test would have found a difference if one exists, and it didn't, so you "accept" the null hypothesis that boys and girls are identical. On the other hand, if the test is weak (which it is, in the gender dataset) then the failure to find a difference between boys and girls does not mean they are identical. Computing the power of a test is more involved than computing confidence intervals; see [2, sec. 4.9] for details.

2 Inadequate Attention to Variance.

Suppose our null hypothesis, H_0, is that boys and girls are equally confident, and our alternative hypothesis, H_1, is that boys are more confident. As noted above, a t test of the gender data fails to reject H_0, so we cannot conclude H_1. Many analyses published in the AI literature stop here—with the result of a t test—but in fact, this result is very misleading. To understand why, it will help to review how t tests (indeed, all statistical tests) work. Test statistics such as t compare the magnitude of an observed effect to the variance of the sampling distribution given H_0. This

Source	dof	Sum of Squares	Mean Square	F	p
Gender	1	.907	.907	2.407	.1287
Grade	3	6.312	2.104	5.581	.0027
Interaction	3	3.247	1.082	2.871	.0482
Error	40	15.08	.377		

Table 2: Analysis of variance showing a significant main effect of grade level and a significant interaction effect.

Grade	4	5	6	7
Boys	4.467	4.450	4.167	4.000
Girls	3.750	4.650	4.450	3.133

Table 3: Mean confidence for boys and girls at four grade levels.

variance is called the *standard error*. In general, increasing the sample size decreases the standard error and makes the observed effect more significant; whereas increasing sample variance increases the standard error and decreases significance. Thus, three factors affect whether an observed effect is statistically significant: the magnitude of the effect (e.g., a difference of .275 in mean confidence), sample size, and sample variance. The researcher controls sample size, and has indirect control over sample variance. Obviously, if the researcher controlled all three factors, there would be no point to running an experiment.

One may boost an observed effect to significance by collecting a very large sample, but this tactic is wasteful of data. More importantly, it neglects the most informative cause of insignificant results, sample variance. Sometimes, sample variance is large and truly random, and nothing can be done about it. But usually, sample variance reflects the combined influence of several factors. If you can tease these influences apart, you can get statistically significant results with no additional data, and a better understanding of the data, as well.

To illustrate, Table 2 shows a two-way analysis of variance of the gender dataset. Two-way analysis of variance decomposes the sample variance into four parts: two represent the effects of the factors (called *main effects*), one represents the interaction between the factors (called the *interaction effect*), and one is due to random chance (called *error*). The *mean square* column in Table 2 gives the relative magnitudes of these components of variance. (Mean squares are just summed, squared deviations divided by degrees of freedom, both listed in Table 2, i.e., they are variances.) F statistics are used to test whether the main effects and interaction effect are large relative to the effect of random chance. The main effect of grade level is highly significant ($p = .0027$) whereas the main effect of gender is insignificant ($p = .1287$). There is a tantalizing interaction effect ($p = .0482$): Apparently, the effect of grade-level on confidence depends on gender, or conversely, the effect of gender on confidence depends on grade.

The means of each grade-gender combination complete the picture (Table 3). Boys' confidence decreases gradually from 4.467 in fourth grade to 4.0 in seventh grade (all children start out overconfident), whereas girls' confidence starts lower (3.75), peaks in fifth grade, and then drops. In short, boys' confidence follows a different developmental pattern than girls'. The interaction effect in Table 2 picks up this difference.

So, does gender have an effect on confidence? The t test says no, but the analysis of variance,

which decomposes sample variance further, says the effect of gender is felt through its interaction with grade level. This interaction effect is invisible to the t test: it becomes clear only when two factors are analyzed for their independent and joint effects. The additional main effect and interaction effect "soak up" some of the sample variance that made the original t test insignificant. By concentrating on sample variance, we see that the effect of gender on confidence changes with age. Had we attempted to boost the effect of gender in the original t test by collecting a larger sample, we would have wasted data and missed this important dependency.

Multiple Comparisons. At this juncture in an analysis, researchers may be tempted to compare individual mean scores. For example, to see whether boys have significantly higher scores at each grade level than girls, one would compare 4.467 to 3.750, 4.450 to 4.650, and so on, with individual t tests. This tactic leads to the problem of multiple comparisons. Recall that every statistical test has a probability, $p < \alpha$, of incorrectly rejecting H_0. Note that α refers to a *single* test; we'll mark this fact by adding a subscript—"c" for "comparison"—to α. Suppose we conduct two tests with $\alpha_c = .05$. What is the probability that at least one test incorrectly rejects H_0? Clearly, it is $1 - (1 - \alpha_c)^2 = .0975$. In general, if we conduct n tests, then the probability that at least one incorrectly rejects H_0 is

$$\alpha_e \approx 1 - (1 - \alpha_c)^n.$$

This "experimentwise error," α_e, is not precisely known because the tests are generally not independent; see [2, p. 190]. If we compare boys and girls at all four grade levels, setting $\alpha_c = .05$, then the probability of at least one error is approximately $1 - (1 - .05)^4 = .185$. In other words, the probability of detecting a difference between boys and girls where none exists is roughly one in five. I have reviewed papers in which authors report dozens of pairwise comparisons, virtually guaranteeing that some apparently significant results are spurious. Unfortunately, there is no way to know which of the apparent results are wrong.

Clearly, any solution to the problem of multiple comparisons involves a tradeoff between α_e and α_c. One can favor α_e, but this requires reducing α_c, making it harder to reject H_0 on a given test, which means that some weaker effects are no longer significant. Or, one can favor α_c, resulting in elevated probabilities of one or more spurious tests. I recommend a hybrid approach, where one conducts n tests with a stringent α_c, designed to give $\alpha_e \approx .05$, and then one conducts all the tests again with the usual $\alpha_c = .05$. Finally, one compares the results: Which tests were significant with $\alpha_c = .05$ but not with the more stringent α_c? These are the tests that might be spurious, and if one cares deeply about any of them, then one might attempt to reduce variance or increase sample size to boost them to significance (see [2, pp. 195-205] for details).

3 Summary.

I have described three errors in data analysis and how to fix or compensate for them. I selected these three because they are common, easy to fix, and because they can ruin one's research. These are not trifling errors. You should not accept the null hypothesis simply because you cannot reject it; you must provide additional support for it. If you fail to notice interactions between factors, then you might conclude that a factor has no effect, when its effect is actually realized through interactions with another factor. If you run multiple comparisons without correcting α_c, then some will probably be spurious. I have noticed that many AI researchers worry about relatively subtle aspects of statistical practice but neglect the issues I have raised here. For instance, one colleague ran scores of uncorrected pairwise comparisons because she thought she wasn't allowed to run an

analysis of variance. True, the analysis of variance assumes normally-distributed populations (for that matter, so does the t test that was used for the pairwise comparisons) but any errors that might have been induced by violating this assumption are miniscule compared with the errors almost certainly induced by multiple uncorrected comparisons.

4 References

1. Beal, C. R. Boys and Girls: The Development of Gender Roles. New York: McGraw-Hill. 1994.
2. Cohen, P. R. Empirical Methods for Artificial Intelligence. MIT Press. 1995.