Naïve Notes on Hypothesis Testing

James S. Wolper
Department of Mathematics
Idaho State University

[A work in progress]
revised 17 November 2005

Hypothesis testing is a widely-accepted tool in statistical inference, and one which many students find confusing. I suspect that many practitioners are confused about hypothesis testing as well, and that it has become a mask to hide sloppy reasoning. A lot is written about hypothesis testing, but most of what I have seen consists of straightforward directions, as might be found in a textbook. Works (*e.g.*, [C]) addressing the philosophy of hypothesis testing are rare, and almost have the feel of a cult topic, like those on non-boolean logic.

I am not a statistician, I am a mathematician. I do have some limited experience as a practitioner of statistics, mostly in the form of stochastic models from engineering, and I only read these, rather than write them. I have to entertain the hypothesis that what I find objectionable is my ignorance. If so, I am not alone: statistics is commonly taught by mathematics professors who have little practical experience with the subject: mathematics does not have data! What mathematics has is logical reasoning.

What is a hypothesis test? Start with a random sample and a statistic calculated from the sample. The statistic is a random variable, that is, a variable whose value depends on some random event. It is impossible to calculate any properties of a random variable without knowing the underlying distribution, so one hypothesizes a distribution, and calculates. The hypothesis test decides, based on the probability of seeing the sample statistic, whether this hypothesis was reasonable.

A hypothesis test does not directly answer interesting questions like "do statins reduce cholesterol?" or "do blondes have more fun?" The hypothesis is not a *scientific* hypothesis. Instead, the test is a statistical construction for the purpose of evaluating evidence. The fundamental error is to confuse these two meanings of hypothesis. There is no guarantee that the statistical hypothesis corresponds to the scientific hypothesis. But this is widely misunderstood: one widely-respected website, run by a physicist [De], states "In the context of hypothesis testing, a hypothesis is a tentative explanation." This is incorrect: the hypothesis is a distribution.

For example, consider a problem from a recent textbook [M, 8.34].

A study of chromosome abnormalities and criminality examined data on 4124 Danish males ... [e]ach ... was classified as having the normal male XY chromosome pair or one of the abnormalities XYY or XXY. Of the 4096 men with normal chromosomes, 381 had criminal records while 8 of the 28 men with chromosome abnormalities had criminal records. Some experts believe that chromosome abnormalities are associated with increased criminality. Do these [sic] data support that belief?"

The textbook solution to this problem uses a hypothesis test. One assumes (H_0) that the populations have the same proportion of people with criminal records, and does a "difference of means" z-test (using the normal approximation to the binomial distribution). The probability that a simple random sample would have such a large difference is tiny, so H_0 is rejected.

But what scientific hypothesis was tested here? Here is one abstraction of this problem: if we assume that populations with different genotypes are different, it is very unlikely that the proportions of the population having some property are not different. There is no news here; no knowledge has been gained. In logic, "False Implies False" and "False Implies True" are *both* true. The effect of the hypothesis test was to prove that the impossible is merely improbable.

Here is a less abstract but logically equivalent problem. Researchers studied apples and oranges and determined the proportion of blemishes that broke the skin for each fruit. The abstract structure is the same: comparing proportions from two populations with different genotypes.

Now, considering that two populations are hypothesized to have equal proportions, we can choose to pool the data, that is, treat it as if it came from one population. In other words, hypothesis testing encourages us to mix apples and oranges!

This problem in [M] is based on an article from *Science* which, given the stature of that journal, leads one to conclude that the study is considered to be good science. Some comments on the original paper by Witkins [W] seem to some indicate favorable response, although there is some controversy. For example, [L, p. 10] claims that many of the crimes committed by those with abnormal genes were petty. I suspect that a lot of effort has gone into refuting the Witkins claim.

But this need not have been, if someone had only examined the logical structure of the hypothesis test. There is no way that any useful information could have been derived using hypothesis testing.

Here is another example, drawn from the same text [M, 8.68]: "As part of one study, children in different age groups were compared on their ability to sort new products into the correct product category" Of course, the 6– and 7–year olds did much better than the 4– and 5– year olds. This is no surprise: the null hypothesis was that the performance of the older and younger children was equal, which is logically equivalent to hypothesizing that children do not develop cognitively as they age. The article on which this problem is based has been cited in the academic marketing literature, at least in passing. Close study of Piaget's work on child development would have been more satisfactory and a less expensive.

Consider another example that shows how far textbook authors are willing to go to encourage sloppy reasoning. This is from [DP], problem 10.55. In this problem we are told that a sample of some kind has determined that the mean attention span of teenage boys is 4 minutes. The student is told to assume that the sample standard deviation was 1.4,

and that the sample size is 50, and asked to test the hypothesis that $\mu = 5$. It seems to me that the null hypotheses involves *three* hypotheses:

$$H_0: \qquad \mu = 5 \text{ AND } s = 1.4 \text{ AND } n = 50$$

The purpose of the null hypothesis is to give us a distribution with which to calculate, and we need all three parameters to calculate: They are all hypothesized.

The hypothesis that the mean attention span is 5 minutes is silly, given that the evidence is that the mean is 4 minutes. The only possible result of the hypothesis test is rejection of the null hypothesis, and, of course, the statistical evidence for rejection is quite strong.

The difficulty here is that we cannot really know *why* we are rejecting the null hypothesis. There were three components to this hypothesis, and the failure of any one of the three invalidates their conjunction. Is any of the hypothesis supported by the evidence? Can something be salvaged.

These problems bear a structural resemblence to a false technique of mathematical proof that is unfortunately still taught at the high school level. Imagine two long expressions in the trigonometric functions $\sin(x)$ and $\cos(x)$. One of the beauties of trigonometry is the large number of identities involving long expressions. Students are taught to set the two expressions equal and then to apply operations to both sides of the resulting equation until they reach the equation $0 = 0$. But this is not a valid proof: for one thing, we already know that $0 = 0$, so we have reduced our trigonometrical problem to one of the logical forms "False implies True" or "True implies True". *Both of these are true!* Thus, we have neither proven nor disproven the hypothesis about the two expressions. (With further work, this proof technique can be made valid, although this is not emphasized.)

The primary mistake made in each of these problems, including the trigonometric identity, is that of paying insufficient attention to the hypotheses. This is not a new issue: It was addressed in 1963 by John R. Platt in an address to the American Chemical Society later published in *Science* as "Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others". Platt cited Molecular Biology (in other words, DNA) and High-Energy Physics as two areas of science that practiced strong inference. The essence of the practice is asking "What experiment could *dis*prove your hypothesis?" and "What hypothesis does your experiment disprove?" In the case of the problems above, the answer to both questions is that no tenable hypothesis could have been disproved by either.

Hypothesis testing has become part of the normal practice of science, but it may be that this should not be so. Hypothesis testing is appropriate if we need to make a decision, but we seem to have lost track of that message. Many hypothesis tests are really parameter estimates, but it seems that society doesn't value parameter estimation, except in physical science ("What is the speed of light in a vacuum?"). Hypothesis testing is also used in some kinds of automation.

But, as someone wrote, too many theses and tenure decisions live or die on "statistical significance", not reasoning. But this may mean that academic careers advance when an author deduces that an unreasonable hypothesis really was unreasonable.

Nor is advanced mathematics the solution. The field of mathematical statistics is growing, interesting, and useful. But, to paraphrase [M], no amount of fancy mathematics can make a test with a poorly-chosen hypothesis worth doing.

We have let "statistical significance" enable sloppy reasoning. We must be careful to avoid sloppy reasoning, and to point it out where we see it. The key method is very old, but not worn-out: form good hypotheses. To do so requires deep understanding of the underlying problems and facility with reasoning, as well as the willingness to reason. In some cases, reasoning alone enables one to determine the truth of a hypothesis. But if we are too ignorant to decide on the absolute truth, we may resort to hypothesis testing to assess the *likelihood* of the truth. Hypothesis testing is, in fact, a confession of ignorance; hypothesis testing should be our final, not our first, attack.

[C] Cohen, Paul R., "Getting What You Deserve from Data", *IEEE Transactions on intelligent Systems*, December, 1996 (vol 11, No. 6)

[De] Denker, av8n.com

[DP] Devore & Peck

[M] Moore & McCabe

[L] Lindsay *et. al.*, *Offenders with Developmental Disabilities*, Wiley

[W] Witkin, H.A., Mednick, S.A., Schulsinger, F., Bakkestrom, F., Christiansen, E., Goodenough, K.O., Hirschhorn, D.R., Lundstein, K., Ownen, C., Phillip, J., Ruben, D. & Stocking, M. (1977). "Criminality, aggression and intelligence among XYY and XXY men." In S. Mednick & K. Christiansen (eds) *Biosocial Bases of Criminal Behaviour.* New York: GardenerPress.