# Problems with the Hypothesis Testing Approach

Over the past several decades (e.g., since Berkson 1938) people have questioned the use of hypothesis testing in the sciences. The criticisms apply to both *experimental* data (control and treatment(s), random assignment of experimental units, replication, and some “design") and *observational* data (some of the above, but at least 2 groups to be compared or contrasted). While we should focus on experiments in ecology, we are often left with only observational studies. Some of the problems with null hypothesis testing are given below (several more problems exist, but these tend to be somewhat technical and, thus, are not given here):

**1.** The most glaring problem with the use of hypothesis testing is that nearly all null hypotheses are <u>obviously</u> false on a priori grounds! Consider the example where the null hypothesis states that the probability of survival (*S*) of an animal is equal over a 15 year study period,

$$\mathbf{H}_9\mathbf{:}\ \ \mathbf{S}_{"} = \mathbf{S}_{\#} = \mathbf{S}_{\$} = \dagger\ \dagger\ \dagger = \mathbf{S}_{"\&}\mathbf{.}$$

This is a trivial “strawman." Why test this? It is obviously false. The “rejection" hardly advances science nor does it give meaningful insights for management.

Other examples where the null hypothesis is a trivial strawman:

**A.** $\mathbf{H}_9\mathbf{:}\ \boldsymbol{S}_{\mathrm{N}} = \boldsymbol{S}_{\mathrm{E}}$

(juvenile and adult survival probabilities are equal)

**B.** $\mathbf{H}_9\mathbf{:}\ \boldsymbol{S}_{4\mathrm{G}} = \boldsymbol{S}_{4\mathrm{V}}$

(survival probabilities for birds fitted with a Radio transmitter are equal to Control birds without a radio transmitter in each year *j*). Any other parameter could be substituted for survival probability in these null hypotheses. I have found that a quick scan of most journals in the biological sciences indicates that such null hypotheses are the rule, not the exception.

**The central issues here are twofold:**

- **First, one must estimate the *magnitude* of the differences and their precision.**
  **The “effect size" – trivial, small, medium, large.**
  **This is an <u>Estimation</u> <u>Problem</u>**

ì **Second, one must know if the differences are large enough to justify inclusion in a model to be used for inference.**
**This is a <u>Model</u> <u>Selection</u> <u>Problem</u>**

**These central issues are not one of Hypothesis Testing.**

"We do not perform an experiment to find out if two varieties of wheat or two drugs are equal. We know in advance, without spending a dollar on an experiment, that they are not equal." (Deming 1975). How could the application of (say) nitrogen on a field have *no* effect on yield? Even the application of sawdust must have *some* effect!

People seemed stunned at the prevalence of testing null hypotheses. Dr. William Thompson (pers. comm.) estimated that a recent volume of *Ecology* contained over 8,000 null hypothesis test results! He felt nearly all of these null hypotheses were false on *a priori* grounds; that is, no one really believed the null. Why are resources spent so carelessly? Why is this practice so common? How can we be so unthinking?

**2.** The alpha level (nearly always 0.1, 0.05, or 0.01) is <u>arbitrary</u> <u>and</u> <u>without</u> <u>theoretical</u> <u>basis.</u>

Using a fixed **!**-level <u>arbitrarily</u> classifies results into relatively <u>meaningless</u> categories

"significant" and "nonsignificant."

Note, the terms *significant* and *nonsignificant* do not relate to biological importance, only to an arbitrary statistical classification. This seems simply stupid. We have been brainwashed!

**3.** Likelihood ratio tests between models that are not nested do not exist. This makes comprehensive analysis problematic. How many results appear in the literature based on a likelihood ratio test between models that are not nested?

**4.** In observational studies, the distribution of the test statistic under the null hypothesis is not known.

We often, mistakenly, hope/think that the distribution is that same nominal distribution as if a true experiment had been conducted (e.g., $F$, $t$, $z$, $x^{\#}$). If hypotheses are formed after looking at the data (data dredging) then the ability to make valid inference is severely compromised (e.g., model-based standard errors are not a valid measure of precision).

**5.** Biologists are better advised to pursue Chamberlain's concept of “Multiple Working Hypotheses" – this seems like superior science.

However, this leads to the *multiple testing problem* in statistics and arbitrariness in defining the null hypotheses. Furthermore, the notion of a null hypothesis presents a certain asymmetry in that the null is favored and has an “advantage." The framework of a *null hypothesis* seems to be of relatively little use.

**6.** Presentation of only test statistics, degree of freedom and P-values limits the effectiveness of (future) meta-analyses. There is a strong “publication bias" whereby only “significant" P-values get reported (accepted) in the literature.

It is important to present parameter estimates and their precision – these become the relevant “data" for a meta-analysis.

**7.** We generally lack theory for testing hypotheses when the model includes nuisance parameters (e.g., the sampling probabilities in capture-recapture models).

One must be very careful in trying to infer something about a P-value (say 0.11 or 0.02) as the *strength of evidence* for the null hypothesis.

In a real sense, the P-value overstates the evidence against the null hypothesis. The standard likelihood ratio (not the likelihood ratio *test*), based on likelihood theory, provides a more realistic basis for such evidence.

**8.** Model selection using hypothesis tests is particularly problematic. Even if there are only 5-8 models being considered, there is no formal (valid) rule that indicates how the results of the various tests might be used to lead to a best model. In fact, the properties of a model selected using a testing approach has unknown properties.

Even among the models that are nested, the test results are not independent and typically have differing power. How are such result to be combined to aid in model selection? The methods in use are *ad hoc*, even though well known in some cases.

In multiple regression is differing models are selected when using step-up, vs. step-down vs. stepwise strategies. In addition, differing models are selected dependent upon on the arbitrary value of **!** used. The various *ad hoc* testing approaches to model selection are commonly taught in statistics classes. In fact, many such courses focus primarily on testing methods and this seems relatively unproductive. While hypothesis testing may be on a decline, the rate of the decline is slow, due in large part to the way that statistics is taught in universities.

**Some famous quotes:**

Wolfowitz (1967), writing about the woes of hypothesis testing states – “The practical statisticians who now accept useless theory should rebel and not do so any more."

“No one, I think, really believes in the possibility of sharp null hypotheses that two means are absolutely equal in noisy sciences." (Kempthorne 1976)).

Nelder (1996) notes the “grotesque emphasis on significance tests in statistics courses of all kinds."

Nester (1996) states, “I contend that the general acceptance of statistical hypothesis testing is one of the most unfortunate aspects of 20th century applied science."

Many believe (erroneously) that a P-value is the probability that the null hypothesis is true!

Approximately 400 references (this number could be quite low) now exist in the quantitative literature than warn of the limitations of hypothesis testing. Harlow et al. (1997) provide a recent edited book entitled, *What If There Were No Significance Tests?* (Lawrence Erlbaum Associates, Publishers, London).

Also see the web site:

**http://www.indiana.edu/~stigstst/**

for more insights.

# WHAT SHOULD BE DONE?

1. Focus on effect size and its precision.

2. Avoid using the words “significant" and “significance" in the publication of research results. In particular avoid any computation of *post-hoc* power.

3. Do not rely on statistical hypothesis tests in the analysis of data from observational studies. With strictly experimental data, use the usual methods (e.g., ANOVA and CANOVA) but focus on the estimated treatment means and their precision, without an emphasis on the F and P values.

4. In planning studies, focus on the “effect size" and its precision as a function of sample size and design and minimize the arbitrary notions of **!** and **"**.