

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 obviously false on a priori grounds!

$$H_0: S_1 = S_2 = S_3 = \cdot \cdot \cdot = S_{15}.$$

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.

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 Estimation Problem
- Second, one must know if the differences are large enough to justify inclusion in a model to be used for inference.
This is a Model Selection Problem

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!

Other examples where the null hypothesis is a trivial strawman:

$$A. \quad H_0: S_J = S_A$$

(juvenile and adult survival probabilities are equal)

$$\mathbf{B. \quad H_0: S_{jC} = S_{jR}}$$

(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.

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 arbitrary and without theoretical basis.

Using a fixed α -level arbitrarily classifies results into meaningless categories

“significant” and “nonsignificant.”

Note, the terms *significant* and *nonsignificant* do not relate to biological importance, only to an arbitrary 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 , χ^2). 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 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.

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 that 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).

Other quotes by well-known statisticians are given below. Also see the web site:

<http://www.indiana.edu/~stigstst/>

for more insights.

WHAT SHOULD BE DONE?

Focus on effect size and its precision.

Stop using the words "significant" and "significance."

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.

Do not report P values or rely on arbitrary α levels

In planning studies, forget the notions of "power" and α and β -- focus on precision of "effect size" as a function of sample size and design.

A Few Quotes Regarding Hypothesis Testing

Dr. Marks Nester <marks@qfri.se2.dpi.qld.gov.au> sent material on hypothesis testing to Ken Burnham at the end of 1996. Ken passed the 2 e-mail files to me. I edited out a few quotes that did not seem that interesting/relevant (e.g., quotes from the Bible), then reformatted and printed in a more readable format. The material (below) starts with the opinions of Dr. Nester regarding hypothesis testing. David Anderson, Jan 2, 1997.

I (i.e., Nester) contend that the general acceptance of statistical hypothesis testing is one of the most unfortunate aspects of 20th century applied science. Tests for the identity of population distributions, for equality of treatment means, for presence of interactions, for the nullity of a correlation coefficient, and so on, have been responsible for much bad science, much lazy science, and much silly science. A good scientist can manage with, and will not be misled by, parameter estimates and their associated

standard errors or confidence limits. A theory dealing with the statistical behaviour of populations should be supported by rational argument as well as data. In such cases, accurate statistical evaluation of the data is hindered by null hypothesis testing. The scientist must always give due thought to the statistical analysis, but must never let statistical analysis be a substitute for thinking! If instead of developing theories, a researcher is involved in such practical issues as selecting the best treatment(s), then the researcher is probably confronting a complex decision problem involving inter alia economic considerations. Once again, analyses such as null hypothesis testing and multiple comparison procedures are of no benefit.

Although some of the following passages have been included for their historical interest, most of the quotations are offered in partial support of my views.

Yates - "the emphasis given to formal tests of significance ... has resulted in ... an undue concentration of effort by mathematical statisticians on investigations of tests of significance applicable to problems which are of little or no practical importance ... and ... it has caused scientific research workers to pay undue attention to the results of the tests of significance ... and too little to the estimates of the magnitude of the effects they are investigating"

Yates - "the occasions ... in which quantitative data are collected solely with the object of proving or disproving a given hypothesis are relatively rare"

Yates - "... the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective"

Hodges, Jr. and Lehmann - "we may formulate the hypothesis that a population is normally distributed, but we realize that no natural population is ever exactly normal"

Hodges, Jr. and Lehmann - "when we formulate the hypothesis that the sex ratio is the same in two populations, we do not really believe that it could be exactly the same"

Anscombe - "Tests of the null hypothesis that there is no difference between certain treatments are often made in the analysis of agricultural or industrial experiments in which alternative methods or processes are compared. Such tests are ... totally irrelevant. What are needed are estimates of magnitudes of effects, with standard errors"

Cochran and Cox - "In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is no difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences."

Hogben (a) - "Acceptability of a statistically significant result ... promotes a high output of publication. Hence the argument that the techniques work has a tempting appeal to young biologists, if harassed by their seniors to produce results, or if admonished by editors to conform to a prescribed ritual of

analysis before publication. ... the plea for justification by works ... is therefore likely to fall on deaf ears, unless we reinstate reflective thinking in the university curriculum"

Hogben (a) - "we can already detect signs of such deterioration in the growing volume of published papers ... recording so-called significant conclusions which an earlier vintage would have regarded merely as private clues for further exploration"

Savage - "to make measurements and then ignore their magnitude would ordinarily be pointless. Exclusive reliance on tests of significance obscures the fact that statistical significance does not imply substantive significance"

Savage - "Null hypotheses of no difference are usually known to be false before the data are collected ... when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science"

Cox - "Exact truth of a null hypothesis is very unlikely except in a genuine uniformity trial"

Neyman - "What was the probability (power) of detecting interactions ... in the experiment performed? ... The probability in question is frequently relatively low ... in cases of this kind the fact that the test failed to detect the existence of interactions does not mean very much. In fact, they may exist and have gone undetected."

Kish - "Significance should stand for meaning and refer to substantive matter. ... I would recommend that statisticians discard the phrase 'test of significance' "

Kish - "the tests of null hypotheses of zero differences, of no relationships, are frequently weak, perhaps trivial statements of the researcher's aims ... in many cases, instead of the tests of significance it would be more to the point to measure the magnitudes of the relationships, attaching proper statements of their sampling variation. The magnitudes of relationships cannot be measured in terms of levels of significance"

McNemar - "too many users of the analysis of variance seem to regard the reaching of a mediocre level of significance as more important than any descriptive specification of the underlying averages"

McNemar - "so much of what should be regarded as preliminary gets published, then quoted as the last word, which it usually is because the investigator is too willing to rest on the laurels that come from finding a significant difference. Why should he worry about the degree of relationship or its possible lack of linearity"

Nunnally - "the null-hypothesis models ... share a crippling flaw: in the real world the null hypothesis is almost never true, and it is usually nonsensical to perform an experiment with the sole aim of rejecting the null hypothesis"

Nunnally - "If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data"

Nunnally - "the mere rejection of a null hypothesis provides only meager information"

Nunnally - "We should not feel proud when we see the psychologist smile and say 'the correlation is significant beyond the .01 level.' Perhaps that is the most that he can say, but he has no reason to smile" (I liked this one! DRA)

Rozeboom - "one can hardly avoid polemics when butchering sacred cows"

Rozeboom - "Whenever possible, the basic statistical report should be in the form of a confidence interval"

Rozeboom - "the stranglehold that conventional null hypothesis significance testing has clamped on publication standards must be broken"

Rozeboom - "The traditional null hypothesis significance-test method ... of statistical analysis is here vigorously excoriated for its inappropriateness as a method of inference"

Smith - "One feature ... which requires much more justification than is usually given, is the setting up of unpalatable null hypotheses. For example, a statistician may set out a test to see whether two drugs have exactly the same effect, or whether a regression line is exactly straight. These hypotheses can scarcely be taken literally"

Camilleri - "another problem associated with the test of significance. The particular level of significance chosen for an investigation is not a logical consequence of the theory of statistical inference"

Camilleri - "The precision and empirical concreteness often associated with the test of significance are illusory and it would be a serious error to predicate our actions towards hypotheses on the test of significance as if it were a reliable arbiter of truth"

Edwards et al. - "in typical applications, one of the hypotheses-the null hypothesis-is known by all concerned to be false from the outset"

Yates - "The most commonly occurring weakness ... is ... undue emphasis on tests of significance, and failure to recognise that in many types of experimental work estimates of treatment effects, together with estimates of the errors to which they are subject, are the quantities of primary interest"

Yates - "In many experiments ... it is known that the null hypothesis ... is certainly untrue"

Bakan - "the test of significance has been carrying too much of the burden of scientific inference. It may well be the case that wise and ingenious investigators can find their way to reasonable conclusions from data because and in spite of their procedures. Too often, however, even wise and ingenious investigators ... tend to credit the test of significance with properties it does not have"

Bakan - "a priori reasons for believing that the null hypothesis is generally false anyway. One of the common experiences of research workers is the very high frequency with which significant results are obtained with large samples"

Bakan - "there is really no good reason to expect the null hypothesis to be true in any population ... Why should any correlation coefficient be exactly .00 in the population? ... why should different drugs have exactly the same effect on any population parameter"

Bakan - "if the test of significance is really of such limited appropriateness ... we would be much better off if we were to attempt to estimate the magnitude of the parameters in the populations"

Bakan - "When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to common sense"

Bakan - "we need to get on with the business of generating ... hypotheses and proceed to do investigations and make inferences which bear on them, instead of ... testing the statistical null hypothesis in any number of contexts in which we have every reason to suppose that it is false in the first place"

Meehl - "in psychological and sociological investigations involving very large numbers of subjects, it is regularly found that almost all correlations or differences between means are statistically significant"

Skipper Jr., Guenther and Nass - "The current obsession with .05 ... has the consequence of differentiating significant research findings and those best forgotten, published studies from unpublished ones, and renewal of grants from termination. It would not be difficult to document the joy experienced by a social scientist when his F ratio or t value yields significance at .05, nor his horror when the table reads 'only' .10 or .06. One comes to internalize the difference between .05 and .06 as 'right' vs. 'wrong,' 'creditable' vs. 'embarrassing,' 'success' vs. 'failure'

"Skipper Jr., Guenther and Nass - "blind adherence to the .05 level denies any consideration of alternative strategies, and it is a serious impediment to the interpretation of data"

Lykken - "Unless one of the variables is wholly unreliable so that the values obtained are strictly random, it would be foolish to suppose that the correlation between any two variables is identically equal to 0.0000... (or that the effect of some treatment or the difference between two groups is exactly zero)"

Lykken - "the finding of statistical significance is perhaps the least important attribute of a good experiment"

Lykken - "The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied"

Lykken - "Editors must be bold enough to take responsibility for deciding which studies are good and which are not, without resorting to letting the p value of the significance tests determine this decision"

Morrison and Henkel - "In addition to important technical errors, fundamental errors in the philosophy of science are frequently involved in this indiscriminate use of the tests [of significance]"

Morrison and Henkel - "we usually know in advance of testing that the null hypothesis is false"

Nelder - "multiple comparison methods have no place at all in the interpretation of data"

Tversky and Kahneman - "Significance levels are usually computed and reported, but power and confidence limits are not. Perhaps they should be."

Tversky and Kahneman - "The emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance."

Box - "all models are wrong"

Chew - "the research worker has been oversold on hypothesis testing. Just as no two peas in a pod are identical, no two treatment means will be exactly equal. ... It seems ridiculous ... to test a hypothesis that we a priori know is almost certain to be false"

GRAYBILL - "when making inferences about parameters ... hypothesis tests should seldom be used if confidence intervals are available ... the confidence intervals could lead to opposite practical conclusions when a test suggests rejection of H_0 ... even though H_0 is not rejected, the confidence interval gives more useful information"

Kemphorne - "no one, I think, really believes in the possibility of sharp null hypotheses -- that two means are absolutely equal in noisy sciences"

Pratt - "tests [of hypotheses] provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful"

Pratt - "And when, as so often, the test is of a hypothesis known to be false ... the relevance of the conventional testing approach remains to be explicated"

Pratt - "This reduces the role of tests essentially to convention. Convention is useful in daily life, law, religion, and politics, but it impedes philosophy"

Barndorff-Nielsen - "Most of the models considered in statistics are but rough approximations to reality"

Cox - "Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned"

Cox - "there are considerable dangers in overemphasizing the role of significance tests in the interpretation of data"

Cox - "statistical significance is quite different from scientific significance and ... therefore estimation ... of the magnitude of effects is in general essential regardless of whether statistically significant departure from the null hypothesis is achieved"

Healy - "The commonest agricultural experiments ... are fertilizer and variety trials. In neither of these is there any question of the population treatment means being identical ... the objective is to measure how big the differences are"

Kruskal - "statistical significance of a sample bears no necessary relationship to possible subject-matter significance"

Kruskal - "it is easy to ... throw out an interesting baby with the nonsignificant bath water. Lack of statistical significance at a conventional level does not mean that no real effect is present; it means only that no real effect is clearly seen from the data. That is why it is of the highest importance to look at power and to compute confidence intervals"

Kruskal - "Another criticism of standard significance tests is that in most applications it is known beforehand that the null hypothesis cannot be exactly true"

Kruskal - "Because of the relative simplicity of its structure, significance testing has been overemphasized in some presentations of statistics, and as a result some students come mistakenly to feel that statistics is little else than significance testing"

Chew - "I have tried to steer them [agricultural researchers] away from testing H_0 . I maintain that on a priori physical, chemical and biological grounds, H_0 is always false in all realistic experiments, and H_0 will always be rejected given enough replication"

Chew - "As Confucius might have said, if the difference isn't different enough to make a difference, what's the difference?"

Kruskal - "the traditional table [analysis of variance table] with its terminology and seductive additivities has in fact often led to superficiality of analysis"

Cox and Snell - "Models are always to some extent tentative"

Little - "The idea that one should proceed no further with an analysis, once a non-significant F-value for treatments is found, has led many experimenters to overlook important information in the interpretation of their data"

Cox - "It is very bad practice to summarise an important investigation solely by a value of P".

Cox - "The criterion for publication should be the achievement of reasonable precision and not whether a significant effect has been found"

Preece - "over-emphasis on significance-testing continues"

Preece - "the norm should be that only a standard error is quoted for comparing means from an experiment"

Box - "The resultant magnification of the importance of formal hypothesis tests has inadvertently led to underestimation by scientists of the area in which statistical methods can be of value and to a wide misunderstanding of their purpose"

Bryan-Jones and Finney - "Of central importance to clear presentation is the standard error of a mean"

Bryan-Jones and Finney - "In interpreting and in presenting experimental results there is no adequate substitute for thought - thought about the questions to be asked, thought about the nature and weight of evidence the data provide on these questions, and thought about how the story can be told with clarity and full honesty to a reader. Statistical techniques must be chosen and used to aid, but not to replace, relevant thought"

Good - "A large enough sample will usually lead to the rejection of almost any null hypothesis ... Why bother to carry out a statistical experiment to test a null hypothesis if it is known in advance that the hypothesis cannot be exactly true"

Jones - "There is a rising feeling among statisticians that hypothesis tests ... are not the most meaningful analyses"

Jones - "preoccupation with testing 'is there an interaction'" in factorial experiments, ... emphasis should be on 'how strong is the interaction?' "

Jones - "Reporting of results in terms of confidence intervals instead of hypothesis tests should be strongly encouraged"

Preece - "Statistical 'recipes' are followed blindly, and ritual has taken over from scientific thinking"

Preece - "The ritualistic use of multiple-range tests-often when the null hypothesis is a priori untenable ...- is a disease"

Altman - "Somehow there has developed a widespread belief that statistical analysis is legitimate only if it includes significance testing. This belief leads to, and is fostered by, numerous introductory statistics texts that are little more than catalogues of techniques for performing significance tests"

Chatfield - "differences are 'significant' ... nearly always ... in large samples"

Chatfield - "Within the last decade or so, practising statisticians have begun to question the relevance of some Statistics courses ... However ... Statistics teaching is still often dominated by formal mathematics"

Chatfield - "tests on outliers are less important than advice from 'people in the field' "

Chatfield - "significance tests ... are also widely overused and misused"

Chatfield - "Rather than ask if these differences are statistically significant, it seems more important to ask if they are of educational importance"

Chatfield - "it has ... become impossible to get results published in some medical, psychological and biological journals without reporting significance values even when of doubtful validity"

Cormack - "Estimates and measures of variability are more valuable than hypothesis tests"

Nelder - "the grotesque emphasis on significance tests in statistics courses of all kinds ... is taught to people, who if they come away with no other notion, will remember that statistics is about tests for significant differences. ... The apparatus on which their statistics course has been constructed is often worse than irrelevant, it is misleading about what is important in examining data and making inferences"

Chernoff - "Analysis of variance ... stems from a hypothesis-testing formulation that is difficult to take seriously and would be of limited value for making final conclusions."

Jones and Matloff - "We recommend that authors display the estimate of the difference and the confidence limit for this difference"

Jones and Matloff - "at its worst, the results of statistical hypothesis testing can be seriously misleading, and at its best, it offers no informational advantage over its alternatives"

Jones and Matloff - "all populations are different, a priori"

Jones and Matloff - "The only remedy ... is for journal editors to be keenly aware of the problems associated with hypothesis tests, and to be sympathetic, if not strongly encouraging, toward individuals who are taking the initial lead in phasing them out"

Lindley - "estimation procedures provide more information [than significance tests]: they tell one about reasonable alternatives and not just about the reasonableness of one value"

Perry - "A confidence interval certainly gives more information than the result of a significance test alone ... I ... recommend its use [standard error of each mean]"

Warren - "the word 'significant' could be abolished ... Based on a dictionary definition, one might expect that results that are declared significant would be important, meaningful, or consequential. Being 'significant at an arbitrary probability level,' ... ensures none of these"

Casella and Berger - "In a large majority of problems (especially location problems) hypothesis testing is inappropriate: Set up the confidence interval and be done with it!"

Hinkley - "for problems where the usual null hypothesis defines a special value for a parameter, surely it would be more informative to give a confidence range for that parameter"

Finney - "rigid dependence upon significance tests in single experiments is to be deplored"

Finney - "The primary purpose of analysis of variance is to produce estimates of one or more error mean squares, and not (as is often believed) to provide significance tests"

Finney - "A null hypothesis that yields under two different treatments have identical expectations is scarcely very plausible, and its rejection by a significance test is more dependent upon the size of an experiment than upon its untruth"

Finney - "I have failed to find a single instance in which the Duncan test was helpful, and I doubt whether any of the alternative tests [multiple range significance tests] would please me better"

Finney - "Is it ever worth basing analysis and interpretation of an experiment on the inherently implausible null hypothesis that two (or more) recognizably distinct cultivars have identical yield capacities?"

Chatfield - "We all know ... that the misuse of statistics and an overemphasis on p values is endemic in many scientific journals"

Finney (b)- "the Blind need frequent warnings and help in avoiding the multiple comparison test procedures that some editors demand but that to me appear completely devoid of practical utility"

Healy - "it is a travesty to describe a p value ... as 'simple, objective and easily interpreted' ... To use it as a measure of closeness between model and data is to invite confusion"

Kruskal and Majors - "We are also concerned about the use of statistical significance-P values-to measure importance; this is like the old confusion of substantive with statistical significance"

Moore and McCabe - "Some hesitation about the unthinking use of significance tests is a sign of statistical maturity"

Moore and McCabe - "It is usually wise to give a confidence interval for the parameter in which you are interested"

Hunter - "How about 'alpha and beta risks' and 'testing the null hypothesis'? ... The very beginning language employed by the statistician describes phenomena in which engineers/physical scientists have little practical interest! They want to know how many, how much, and how well ... Required are interval estimates. We offer instead hypothesis tests and power curves"

Preece - "I cannot see how anyone could now agree with this [Fisher's 1935 quote about experiments and null hypotheses]"

Street - "Fisher ... appears to have placed an undue emphasis on the significance test"

Street - "in many experiments it is well known ... that there are differences among the treatments. The point of the experiment is to estimate ... and provide ... standard errors. One of the consequences of this emphasis on significance tests is that some scientists ... have come to see a significant result as an end in itself"

Matloff - "the test is asking whether a certain condition holds exactly, and this exactness is almost never of scientific interest"

Matloff - With regard to a goodness-of-fit test to answer whether certain ratios have given exact values, "we know a priori this is not true; no model can completely capture all possible genetical mechanisms"

Matloff - "the number of stars by itself is relevant only to the question of whether H_0 is exactly true-a question which is almost always not of interest to us, especially because we usually know a priori that H_0 cannot be exactly true."

Matloff - "problems stemming from the fact that hypothesis tests do not address questions of scientific interest"

Upton - "The experimenter must keep in mind that significance at the 5% level will only coincide with practical significance by chance!"

Wang - "the tyranny of the N-P [Neyman-Pearson] theory in many branches of empirical science is detrimental, not advantageous, to the course of science"

BOARDMAN - "He [W. E. Deming] went on to suggest that the problem lay in teaching 'what is wrong.' The list of evils taught in courses on statistics ... is a long one. One of the topics included hypothesis testing. Personally I have found few, if any, occasions where such tests are appropriate."

Inman - "Like many working scientists since, Buchanan-Wollaston professed a belief that commonly used statistical tests were either obvious or irrelevant to the scientific problem of interest"

McCloskey - "scientists care about whether a result is statistically significant, but they should care much more about whether it is meaningful"

Additional information can be found the the following web site:

<http://www.indiana.edu/~stigstst/>

REFERENCES (also see those given in the book by Harlow et al. (1997).

Altman, D. G. (1985). Discussion of Dr Chatfield's paper. *J. R. Statist. Soc. A* 148, Part 3 : 242.

Anscombe, F. J. (1956). Discussion on Dr. David's and Dr. Johnson's Paper. *J. Roy. Statist. Soc. B* 18 : 24-27.

Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society* 23 : 186-190.

Bakan, D. (1967). The test of significance in psychological research. From Chapter 1 of *On Method*, Jossey-Bass, Inc. (San Francisco). Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).

- Barndorff-Nielsen, O. (1977). Discussion of D. R. Cox's paper. *Scand. J. Statist.* 4 : 67-69.
- Beaven, E. S. (1935). Discussion on Dr. Neyman's Paper. *Journal of the Royal Statistical Society, Supplement 2* : 159-161.
- Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association* 82(397) : 112-122.
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Ass.* 33 : 526-536.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association* 37(219) : 325-335.
- Boardman, T. J. (1994). The statistician who changed the world: W. Edwards Deming, 1900-1993. *The American Statistician* 48(3) : 179-187.
- Box, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Ass.* 71 : 791-799.
- Box, G. E. P. (1983). An apology for ecumenism in statistics. In *Scientific Inference, Data Analysis, and Robustness*, G. E. P. Box, T. Leonard and C. F. Wu (eds.), Academic Press, Inc. : 51-84.
- Braithwaite, R. B. (1953). *Scientific Explanation. A Study of the Function of Theory, Probability and Law in Science.* Cambridge University Press.
- Bryan-Jones, J. and Finney, D. J. (1983). On an error in "Instructions to Authors". *HortScience* 18(3) : 279-282.
- Buchanan-Wollaston, H. J. (1935). The philosophic basis of statistical analysis. *Journal of the International Council for the Exploration of the Sea* 10 : 249-263.
- Camilleri, S. F. (1962). Theory, probability, and induction in social research. *American Sociological Review* 27 : 170-178. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Casella, G. and Berger, R. L. (1987). Rejoinder. *Journal of the American Statistical Association* 82(397) : 133-135.
- Chatfield, C. (1985). The initial examination of data (with discussion). *J. R. Statist. Soc. A* 148, Part 3 : 214-253.

- Chatfield, C. (1989). Comments on the paper by McPherson. *Journal of the Royal Statistical Society, Series A*, 152 : 234-238.
- Chernoff, H. (1986). Comment. *The American Statistician* 40(1) : 5-6.
- Chew, V. (1976). Comparing treatment means: a compendium. *HortScience* 11(4) : 348-357.
- Chew, V. (1980). Testing differences among means: correct interpretation and some alternatives. *HortScience* 15(4) : 467-470.
- Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*. 2nd ed. John Wiley & Sons, Inc.
- Cormack, R. M. (1985). Discussion of Dr Chatfield's paper. *J. R. Statist. Soc. A* 148, Part 3 : 231-233.
- Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics* 29 : 357-372.
- Cox, D. R. (1977). The role of significance tests. (With discussion). *Scand. J. Statist.* 4 : 49-70.
- Cox, D. R. (1982). Statistical significance tests. *Br. J. Clinical. Pharmac.* 14 : 325-331.
- Cox, D. R. and Snell, E. J. (1981). *Applied Statistics Principles and Examples*. Chapman and Hall.
- Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70 : 193-242.
- Finney, D. J. (1988). Was this in your statistics textbook? III. Design and analysis. *Expl Agric.* 24 : 421-432.
- Finney, D. J. (1989a). Was this in your statistics textbook? VI. Regression and covariance. *Expl Agric.* 25 : 291-311.
- Finney, D. J. (1989b). Is the statistician still necessary? *Biom. Praxim.* 29 : 135-146.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd (London).
- Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd (Edinburgh).
- Gauch Jr., H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics* 44 : 705-715.

- Geary, R. C. (1947). Testing for normality. *Biometrika* 34 : 209-242.
- Good, I. J. (1983). *Good Thinking. The Foundations of Probability and Its Applications*. University of Minnesota Press (Minneapolis).
- Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury Press (Massachusetts).
- Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press.
- Hahn, G. J. (1990). Commentary. *Technometrics* 32(3) : 257-258.
- Healy, M. J. R. (1978). Is statistics a science? *J. R. Statist. Soc. A* 141, Part 3 : 385-393.
- Healy, M. J. R. (1989). Comments on the paper by McPherson. *Journal of the Royal Statistical Society, Series A*, 152 : 232-234.
- Hinkley, D. V. (1987). Comment. *Journal of the American Statistical Association* 82(397) : 128-129.
- Hodges Jr., J. L. and Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B*, 16 : 261-268.
- Hogben, L. (1957a). The contemporary crisis or the uncertainties of uncertain inference. *Statistical Theory*, W. W. Norton & Co., Inc. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Hogben, L. (1957b). Statistical prudence and statistical inference. *Statistical Theory*, W. W. Norton & Co., Inc. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Hunter, J. S. (1990). Commentary. *Technometrics* 32(3) : 261.
- Inman, H. F. (1994). Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from *Nature*. *The American Statistician* 48(1) : 2-11.
- Jones, D. (1984). Use, misuse, and role of multiple-comparison procedures in ecological and agricultural entomology. *Environmental Entomology* 13(3) : 635-649.

- Jones, D. and Matloff, N. (1986). Statistical hypothesis testing in biology: a contradiction in terms. *Journal of Economic Entomology* 79(5) : 1156-1160.
- Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association* 61(313) : 11-34.
- Kempthorne, O. (1976). Of what use are tests of significance and tests of hypotheses. *Commun. Statist. - Theor. Meth A5* (8) : 763-777.
- Kish, L. (1959). Some statistical problems in research design. *American Sociological Review*, 24 : 328-338. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Kruskal, W. H. (1978). Significance, Tests of. In *International Encyclopedia of Statistics*, eds. W. H. Kruskal and J. M. Tanur, Free Press (New York) : 944-958.
- Kruskal, W. (1980). The significance of Fisher: a review of R. A. Fisher: The Life of a Scientist. *Journal of the American Statistical Association* 75(372) : 1019-1030.
- Kruskal, W. and Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician* 43(1) : 2-6.
- Lindley, D. V. (1986). Discussion. *The Statistician* 35 : 502-504.
- Little, T. M. (1981). Interpretation and presentation of results. *HortScience* 16(5) : 637-640.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin* 70 : 151-159. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Matloff, N. S. (1991). Statistical hypothesis testing: problems and alternatives. *Environmental Entomology* 20(5) : 1246-1250.
- McCloskey, D. N. (1995). The insignificance of statistical significance. *Scientific American* 272(4) : 104-105.
- McNemar, Q. (1960). At random: sense and nonsense. *American Psychologist* 15 : 295-300.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34 : 103-115. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).

- Moore, D. S. and McCabe, G. P. (1989). Introduction to the Practice of Statistics. W. H. Freeman and Company (New York).
- Morrison, D. E. and Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist* 4 : 131-140. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Morrison, D. E. and Henkel, R. E. (Eds.) (1970). *The Significance Test Controversy - A Reader*. Aldine Publishing Company (Butterworth Group).
- Nelder, J. A. (1971). Discussion on papers by Wynn, Bloomfield, O'Neill and Wetherill. *Journal of the Royal Statistical Society, Series B*, 33 : 244-246.
- Nelder, J. A. (1985). Discussion of Dr Chatfield's paper. *J. R. Statist. Soc. A* 148, Part 3 : 238.
- Neyman, J. (1958). The use of the concept of power in agricultural experimentation. *Journal of the Indian Society of Agricultural Statistics* 9(1) : 9-17.
- Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231 : 289-337.
- Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement* XX(4) : 641-650.
- Pearce, S. C. (1992). Data analysis in agricultural experimentation. II. Some standard contrasts. *Expl Agric.* 28 : 375-383.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated systems of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series V*, 1 : 157-175.
- Pearson, K. (1935a). Statistical tests. *Nature* 136 : 296-297. (Not sighted, reproduced in H. F. Inman (1994). Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from *Nature*. *The American Statistician* 48(1) : 2-11.)
- Pearson, K. (1935b). Statistical tests. *Nature* 136 : 550. (Not sighted, reproduced in H. F. Inman (1994). Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from *Nature*. *The American Statistician* 48(1) : 2-11.)

- Perry, J. N. (1986). Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology* 79(5) : 1149-1155.
- Pratt, J. W. (1976). A discussion of the question: for what use are tests of hypotheses and tests of significance. *Commun. Statist.-Theor. Meth.* A5(8) : 779-787.
- Preece, D. A. (1982). The design and analysis of experiments: what has gone wrong? *Utilitas Mathematica* 21A : 201-244.
- Preece, D. A. (1984). Biometry in the Third World: science not ritual. *Biometrics* 40 : 519-523.
- Preece, D. A. (1990). R. A. Fisher and experimental design: a review. *Biometrics* 46 : 925-935.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin* 57 : 416-428. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Savage, I. R. (1957). Nonparametric statistics. *J. Amer. Statist. Ass.* 52 : 331-344.
- Skipper Jr., J. K., Guenther, A. L. and Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist* 2 : 16-18. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company (Butterworth Group).
- Smith, C. A. B. (1960). Book review of Norman T. J. Bailey: *Statistical Methods in Biology*. *Applied Statistics* 9 : 64-66.
- Street, D. J. (1990). Fisher's contributions to agricultural statistics. *Biometrics* 46 : 937-945.
- "Student" (1908). The probable error of a mean. *Biometrika* 6 : 1-25.
- Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* 76(2) : 105-110.
- Upton, G. J. G. (1992). Fisher's exact test. *J. R. Statist. Soc. A* 155(3) : 395-402.
- Vardeman, S. B. (1987). Comment. *Journal of the American Statistical Association* 82(397) : 130-131.
- Venn, J. (1888). Cambridge anthropometry. *Journal of the Anthropological Institute* 18 : 140-154.
- Wang, C. (1993). *Sense and Nonsense of Statistical Inference*. Marcel Dekker, Inc.

Warren, W. G. (1986). On the presentation of statistical analysis: reason or ritual. *Can. J. For. Res.* 16 : 1185-1191.

Yates, F. (1951). The influence of Statistical Methods for Research Workers on the development of the science of statistics. *Journal of the American Statistical Association* 46 : 19-34.

Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics* 20 : 307-321.

Zeisel, H. (1955). The significance of insignificant differences. *Public Opinion Quarterly* 17 : 319-321. Reprinted in *The Significance Test Controversy - A Reader*, Eds. D. E. Morrison and R. E. Henkel, 1970, Aldine Publishing Company