Against Statistical Inference

Larry Magder

Department of Epidemiology and Preventive Medicine,
University of Maryland

Prologue

- I wrote this up as a commentary and submitted it for publication.
- In the commentary I argued that much of the way statistical methods are used and described in biomedical research is inappropriate.
- The commentary was rejected because, as one referee said...

"This is well known!"

I am against statistical inference.

And so are you!

What do I mean by statistical inference?

I am defining it in a narrow sense:

Specify a null hypothesis and a proposed analysis.

Collect data.

Calculate a p-value.

If the p-value is less than some pre-specified cutpoint (usually .05) we reject the null hypothesis.

The mantra of the beginning biostat student:

"P<.05, therefore we reject the null."

In contrast,

I am in favor of "Scientific Inference"

What do I mean by scientific inference?

After calculating a p-value or other measure of evidence from our research we might make a judgment about a hypothesis based on many considerations such as:

Biologic Plausibility

Previous Research

P-value from our own research

Advantages and Disadvantages of Statistical Inference

Advantages:

- If I use Statistical Inference (at the .05-level), and the null hypothesis is true then I know there is only a 5% chance that I will incorrectly reject the null hypothesis.

  In statistical jargon, I have control over my "Type 1 Error Rate".

- Statistical Inference is objective.

Disadvantages:

- Ignores seemingly relevant considerations.

Advantages and Disadvantages of Scientific Inference

Disadvantages:

- **No control over error rates under specified hypotheses.**
- **Scientific inference is subjective.**

Advantages:

- **Takes account of seemingly relevant considerations.**

- I believe most people would agree that the disadvantages of statistical inference outweigh the advantages

- I think few scientists would automatically reject a null hypothesis simply because they found a $p < .05$

- In my view it would be unethical in a regulatory context to base decisions simply on whether a $p < .05$

So, if we agree, what is the point of this talk?

While we may agree about this, I believe that:

- Many scientists fail to appreciate the *implications* of being against statistical inference.

- The language of statistical inference is pervasive, and leads to scientific conclusions that are inconsistent with scientific inference and defy common sense

Implication 1

The role of statistics in science should be to quantify the strength of evidence in a study so other scientists can integrate the new results with other information to make scientific judgments.

(Reporting exact p-values is one way to summarize the evidence against a null hypothesis, although there are better measures of evidence)

Despite this, there tends to be very little discussion of how to quantify evidence in biostatistical education.

Implication 2

It is not necessary to state

"We considered a p-value of .05 to be statistically significant".

(This statement is clearly unnecessary if exact p-values are reported and the degree of evidence is viewed as a continuum)

Despite this fact, a review of recent papers in major medical journals reveals that a majority of papers contain a statement like this.

Implication 3

Whether to use a one-sided or a two-sided p-value is not an issue.

> (Whether to report a one-sided or two-sided p-value is analogous to the question of whether to report height in inches or centimeters. It doesn't matter, as long as the reader is told which scale is used)

Despite this fact, the difference between one-sided and two-sided tests is still thought of as a key topic in statistical education, and debates regarding which to use still appear in the literature.

Implication 4

It is generally unnecessary to adjust inference for multiple comparisons.

(The common recommendation to do so is concerned with controlling the probability of incorrectly rejecting any null hypothesis in the study. If the role of statistics is simply to quantify evidence, this control is not needed.)

Despite this, statistical guidelines often emphasize the need to adjust for multiple comparisons, and the American Statistical Association considers it *unethical* to fail to adjust for multiple comparisons.

Examples of how the narrow view of statistical inference can lead to scientific conclusions that defy common sense.

Example 1

The language of statistical inference induces scientists to place undue emphasis on whether a p-value crosses the arbitrary threshold of .05.

Mark Nester[1] reports the following:

> "A journal editor has confided that an author's thesis is undoubtedly true, but that the editor must reject the paper because the author's ideas are not supported by statistically significant results."

[1]Nester M. An Applied Statistician's Creed. *Appl. Statist.* 1996;45(4):401-410.

Example 2

Conversely, some scientists seem to have the impression that if a p-value of less than .05 is found, a researcher cannot express the opinion that the association is due to chance[1]:

"There is no justification whatever for the statement in their summary that the few other statistically significant associations between occupation and disease were thought to be due to chance. In making such a pronouncement they automatically destroy the logic of practical statistical inference one of the tenets of which is to say that given a certain probability level we will believe that the results have not arisen by chance."

[1] Dudley H. When is significant not significant. *British Medical Journal*. 1977:47.

Example 3

A study section I was on reviewed a proposal that was designed to address a particular research question.

The applicants proposed to collect information on other outcomes at relatively little cost to address additional research questions.

This seemingly efficient approach was criticized by members of the panel because the inclusion of additional research questions would "diminish the power" for the original research question.

Some complications: What is the best measure of evidence?

- If the role of statistical methods in science is to quantify evidence, then what is the best measure of evidence?
- In this talk, I have suggested that p-values can be used to quantify evidence against a null hypothesis.
- However, Royall and Goodman argue persuasively that in order to quantify the evidence against a particular hypothesis, an alternative hypothesis must be specified.
- They recommend using a likelihood ratio to quantify the relative evidence for the two hypotheses.

Some complications: Confidence Intervals

- This talk has focussed on assessment of hypotheses.
- However, estimation of parameters is another important scientific activity and confidence intervals are can be useful.
- Confidence Intervals can be viewed from the standpoint of Statistical Inference as intervals formed by a procedure that has known probabilities of covering the true parameter.
- Confidence Intervals can be viewed from the standpoint of Scientific Inference as a set of parameter values consistent with the data.

Some interesting papers about these topics

- Teaching hypothesis tests--time for a significant change. Jonathan Sterne, Statistics in Medicine 2002;21:985-994
- P-values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a neglected historical debate, Steve Goodman, AJE, 1993;137:485-96
- Multiple comparisons and related issues in the interpretation of epidemiologic data, Savitz DA and Olshan AF, AJE, 1995, 904-908)

In conclusion, consider the ironic remarks of William Rozeboom[1]

"…Who has ever given up a hypothesis just because one experiment yielded a test of statistic in the rejection region? And what scientist in his right mind would ever feel there to be an appreciable difference between the interpretive significance of data, say, for which one-tailed p=.04 and that of data for which p=.06, even though the point of 'significance' has been set at p=.05? In fact, the reader may well feel undisturbed by the charges raised here against traditional null hypothesis decision procedures because, without perhaps realizing it, he has never taken the method seriously anyway".

[1]Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychological Bulletin*. 1960;57(5):416-428.

- We teach students "P<.05 therefore we reject the null hypothesis".
- But, in the words of Rozeboom, no scientist in his right mind would behave this way.
- It may seem just semantic, but this ritual, and the narrow notion of statistical inference exert an influence on the way we design, execute, interpret, and write about our research projects that is not in the best interest of science.