The role of background assumptions in severity appraisal

@Lakens

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

NWO Nederlandse Organisatie voor Wetenschappelijk Onderzoek

If we have a statistics war, what are the underlying causes? Economic gain, territorial gain, religion, or nationalism?

religion
(or philosophy of science)

Thoughts about how people should use statistics follow from a philosophy of science. So what is the **goal** of statistical inferences?

The conventional wisdom behind the approach goes something like this: The logic begins, more or less, with the proposition that one does not want to accept a hypothesis that stands a fairly good chance of being false (i.e., one ought to avoid Type I errors). The logic goes on to state that one either accepts hypotheses as probably true (not false) or one rejects them, concluding that the null is too likely to regard it as rejectable. The .05 alpha is a good fail-safe standard because it is both convenient and stringent enough to safeguard against accepting an insignificant result as significant. The argument, although not beyond cavil (e.g., Bakan, 1967), affords a systematic approach that many researchers would insist has served scientists well. We are not interested in the logic itself, nor will we argue for replacing the .05 alpha with another level of alpha, but at this point in our discussion we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?

[Rosnow & Rosenthal, 1989]

as far as a particular hypothesis is concerned, no test based upon the theory of probability* can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x \leqq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leqq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.

[Neyman & Pearson, 1933]

Establishing a Corpus of Findings

Rozeboom (1960) suggested that scientists should not be making decisions about claims, they should be calculating and updating the probability of these claims. However, this does not seem practical. If there were only a handful of potential claims in any given area of psychology, it would be feasible to assign them probabilities, to be constantly updating the probabilities, and to expect experimenters to keep track of these ever-changing probabilities. In fact, just the number of claims in psychology is overwhelming. It would probably be impossible for human beings to keep track of the probability for each claim, especially if these probabilities were constantly changing. In any case, scientists do not assign probabilities to claims. Instead, scientists act like the goal of science is to collect a corpus of claims that are considered to be established (Giere, 1972).

[Frick, 1996]

Science needs mechanisms for the accumulation of sound conclusions (sensu [Tukey, 1960]). A major rival for Evidentialism as a philosophically sound (in our eyes) system for the advancement of science is the "Error Statistical" brand of Neyman-Pearson analysis promoted by Deborah Mayo.

We dismiss Bayesianism for its use of subjective priors and a probability concept that conceives of probability as a measure of personal belief. Bayesianism is held by many philosophers as the most appropriate method of developing personal knowledge. This may be, but is irrelevant to the task at hand. Science depends on a public epistemology not a private one. The Bayesian attempts to bridge the gap between the private and the public have been tortured.

It is not that we believe that Bayes' rule or Bayesian mathematics is flawed, but that from the axiomatic foundational definition of probability Bayesianism is doomed to answer questions irrelevant to science. We do not care what you believe, we barely care what we believe, what we are interested in is what you can show.

[Taper & Lele, 2011]

Researchers like to make claims in titles and conclusions. It's how we draw attention to our work (and get others annoyed enough so that they will try to prove us wrong).

Claims can be correct or wrong. If we would use a coin flip as a methodological procedure to make claims, we would be right 50% of the time.

To be right a bit more often, hypothesis testing procedures were developed that control error rates in the long run.

3. DISTINCTIVE TYPES OF APPLICATION

3.1. Preliminary Remarks

We now distinguish four broad situations in which the calculation of *p*-values is potentially useful. The first corresponds closely to the considerations underlying the Neyman–Pearson theory of testing hypotheses (Neyman & Pearson 1928, 1967)

[Cox, 2020]

After p < alpha, we should formally conclude:

"We **claim** there is an effect, while acknowledging that if scientists make claims using this **methodological procedure**, they will be misled at most **alpha% or beta% of the time**, which we **deem acceptable.** Let's **for the foreseeable future** (until new data emerges that proves us wrong) **assume our claim is correct**."

Finally, it might be argued that in making an inference we are 'deciding' to make a statement of a certain type about the populations and that therefore, provided that the word decision is not interpreted too narrowly, the study of statistical decisions embraces that of inference.

[Cox, 1958]

I hope to show that its constituent processes fall under three headings: (i) visualization of several possible sets of hypotheses relevant to the phenomena studied, (ii) deductions from these sets of hypotheses, and (iii) an act of will or a decision to take a particular action, perhaps to assume a particular attitude towards the various sets of hypotheses mentioned under (i).

[Neyman, 1957]

When we “accept” or “reject” a hypothesis in a Neyman-Pearson approach to statistical inferences, we do not communicate any belief or conclusion about the substantive hypothesis.

Instead, we utter a Popperian **basic statement**, based on a prespecified decision rule and given background assumptions, that the observed data reflect a certain state of the world. This basic statement **corroborates** our prediction, or **not**.

The alternative to using a methodological procedure that **makes claims with a known error rate**, is to use a methodological procedure that **does not make claims**, or **makes claims with unknown error rates**.

Without *p*-values we either **do not build a collectively agreed upon corpus of findings**, or we build a corpus of findings that, given bias, **is wrong more often than we want**.

One attractive property of this methodological approach to make scientific claims is that the scientific community can **collectively agree** upon the **severity** with which a claim has been tested.

If we design a study with 99.9% power for the smallest effect size of interest and use a 0.1% alpha level, **everyone agrees the risk of an erroneous claim is low**.

Or can we? Is the evaluation of severity always simply a matter of quantified error rates? In practice, it seems not.

As researchers have started to preregister, it turns out they often preregister uninformed predictions, and change their analysis plan.

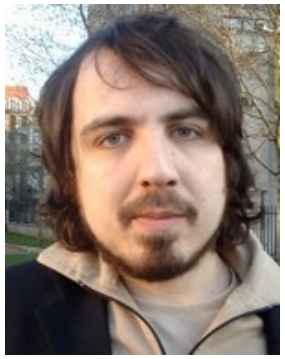
Deviations can be improvements (as Meehl says: Don’t make a mockery of honest ad-hoccery). Deviations trade guaranteed error control against a subjective evaluation of higher *validity*.

From a methodological falsificationist philosophy, high validity is conceptually strongly related to severity. They might even be the same thing.

The following is based on work with:

Aline Claessen

Wolf Vanpaemel



Noah van Dongen

Many of the crises in psychology (measurement crisis, generalizability crisis, theory crisis) are specific instances of problems with severity.

Let's define severity as the ratio of the probability of observing evidence (**e**) given the hypothesis (**H**) and background knowledge (**B**), to observing evidence (**e**) given only the background knowledge (**B**).

$$S(e,H,B) = Pr(e|H,B) / Pr(e|B).$$

[Popper, 1963]

Colloquially: Getting it right when you are right and getting it wrong when you are wrong.

In the last decade most of the focus has been on statistics. Problems related to low statistical power, flexibility in the data analysis, and publication bias all impact severity.

The severity of statistical tests can be quantified. But error rates depend on assumptions, and different researchers can disagree on these assumptions.

For example, researcher A preregisters a Student's t-test, while researcher B believe the homogeneity assumption will be violated and wants to see a Welch's t-test.

The severity of statistical tests can be quantified (if we agree on assumptions). But how severely a claim is tested does not only depend on the test result.

The theory crisis is caused by a lack of a clear specification of a theory's universality and precision.

[Glöckner & Betsch, 2011; Oberauer & Lewandowsky, 2019]

Just as flexibility in the Type 1 error rate reduces the probability of being wrong when you are wrong, lack of specificity reduces the possibility of a theory being proven wrong.

Performing a direct replication in such cases is rather uninteresting because the results would have little to no consequence for our evaluation of the theory.

The measurement/validity crisis is a problem with the severity of claims.

For example, lack of measurement invariance increases Pr(e|B) relative to Pr(e|H,B). It can become relatively easy to observe a difference, even if the theory is false.

Similarly, measurement error and lack of construct validity reduce the severity of tests as you are less likely to get it right when you're right or wrong when you're wrong.

If we want severely tested claims, researchers need to reach agreement on assumptions about data, measurement, and theories.

This requires an extremely systematic approach to hypothesis testing (a *systematic replications framework*) to narrow down all auxiliary hypotheses.

[Uygun-Tunç & Tunç 2020]

Summary

For some researchers, science is about making claims.

If we make claims, we want these to be severely tested.

Summary

How severely as statistical test is can be quantified (under assumptions).

But claims depend on more than the statistical result.

Summary

All of the crises in psychology (and other sciences) can be viewed from a severity perspective.

The solution is to systematically test auxiliary hypotheses to increase test severity.

Thanks!

@Lakens

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY