A Detailed Review of ‘Statisticians’ Fallacy’ Examples

In my manuscript I wrote: “However, when we survey the literature, we rarely see the viewpoint that *all* approaches to statistical inferences, including *p*-values, provide answers to specific questions a researcher might want to ask. Instead, statisticians often engage in what I call the Statistician’s Fallacy – a declaration of what researchers really ‘want to know’, without limiting a statistical question to any specific context. […] I call these statements a fallacy, which might sound severe, but I believe the arguments provided by these statisticians for their claims about ‘what we want to know’ boil down to nothing more than *wishful thinking*. All authors seem to mean ‘*what I wish you wanted to know*’, or perhaps more normatively, *‘what I think you should want to know’*.”

At the core of the Statisticians Fallacy is a statistician making a statement about what it is ‘we’ want to know that is not sufficiently contextualized – either by research field, research method, phase of the research line, or based on a specific philosophy of science. In this supplement I aim to carefully review the 6 quotes I mention in my article to evaluate if my evaluation of the presence of the Statistician’s Fallacy is justified.

Quote 1: Cohen, 1994: “What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that Ho is true?"

Cohen thinks ‘the very reason an experiment is done’ is to know whether H0 is unlikely. Page 998: “When one rejects Ho, one wants to conclude that Ho is unlikely, say, *p* < .01. The very reason the statistical test is done is to be able to reject Ho because of its unlikelihood! But that is the posterior probability, available only through Bayes's theorem, for which one needs to know P(H()), the probability of the null hypothesis before the experiment, the "prior" probability.”

However, later in the article (p. 999) he writes: “There is a form of Ho testing that has been used in astronomy and physics for centuries, what Meehl (1967) called the "strong" form, as advocated by Karl Popper (1959). Popper proposed that a scientific theory be tested by attempts to falsify it. In null hypothesis testing terms, one takes a central prediction of the theory, say, a point value of some crucial variable, sets it up as the Ho, and challenges the theory by attempting to reject it. This is certainly a valid procedure, potentially even more useful when used in confidence interval form. What I and my ilk decry is the "weak" form in which theories are "confirmed" by rejecting null hypotheses.”

However, this sentence does not seem very consistent. He admits range or point predictions are valid. Then he laments that rejecting the null confirms the alternative. However, this is still done in a range prediction (under a Neyman-Pearson approach, where ‘confirmed’ means ‘accept’ as in ‘act as if H1 is true’ based on specified error rates). It is unclear based on these sentences what Cohen is exactly arguing against, or what is deemed ok. Cohen also seems to suggest the dichotomous nature of a Neyman-Pearson approach is not a good way to do science: “The ritual dichotomous reject-accept decision, however objective and administratively convenient, is not the way any science is done.” And he seems to agree with (and quote, on p. 999) Rozeboom that: “"The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one . . . believes the hypothesis . . . being tested". Cohen seems to argue for a Bayes factor interpretation.

Cohen also believes NHST is asking a fundamentally uninteresting question because H0 is always false: “My work in power analysis led me to realize that the nil hypothesis is always false.” (p. 1000). He then discusses the crud factor by Meehl (for a critical discussion, see Orben & Lakens, 2020) and how statistical significance is not practical significance. Cohen thinks even correct use of p-values does not tell us much: “Even a correct interpretation of *p* values does not achieve very much, and has not for a long time.”

So it is clear NHST is not seen as a solution. It is interesting to examine what Cohen thinks are solutions. He writes (p. 1001) “First, don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist”. He then discusses the values of exploratory data analysis (which I agree with, but it is not a solution to tests of theories). Then he argues to look at effect sizes and confidence intervals. Those are important, but an estimate of an effect size is not of much help if you want to know if you are looking at random noise, or not (the question a hypothesis test answers). In the discussion, it turns out he is not against NHST. Just against badly used NHST (p. 1002): “Even null hypothesis testing complete with power analysis can be useful if we abandon the rejection of point nil hypotheses and use instead "good-enough" range null hypotheses (e.g., "the effect size is no larger than 8 raw score units, or d = .5), as Serlin and Lapsley (1993) have described in detail.” So, he is basically advocating equivalence testing as a good use of NHST.

In his reply to comments (Cohen 1995), the full quote is: “To those who rushed to the defense of NHST, I concede that there are circumstances in which the direction, not the size of an effect, is central to the purpose of research. An example is a strictly controlled experiment, such as a clinical trial (although even in a clinical trial, nothing is lost and much may be gained with confidence limits). But the ritual of nil hypothesis testing has so dominated our research practice that it has inhibited our interest in the magnitude of the phenomena we study and the units in which they are measured, the basic stuff of which quantitative sciences are made.”

Cohen’s only remaining problem seems to be that the ritual of a hypothesis test inhibits our interest in effect sizes. I fully agree – but then that is hardly the reason his first article is cited in general! In the reply to comments he also writes “Incidentally, I do not question the validity of NHST, but rather its widespread misinterpretation.” After this reply, I am not really sure what he is actually arguing for or against, beyond that people should not use a tool in an incorrect manner. So all together, I think Cohen 1994 is a good example of the statisticians fallacy. After commentaries, he corrects his statement, even admitting there is a role for NHST and *p*-values. But his original article (which is cited vastly more often than the reply to commentaries) is a fair example of a Statistician’s Fallacy.

Quote 2: Colquhoun, 2017: “*what you want to know is that when a statistical test of significance comes out positive, what is the probability that you have a false positive”.*

On the first page of this pre-print David Colquhoun opens with:

“When you have done an experiment, you want to know whether you have made a discovery or whether your results could have occurred by chance. More precisely, what you want to know is that when a statistical test of significance comes out positive, what is the probability that you have a false positive i.e. there is no real effect and the results have occurred by chance. This probability is defined here as the false positive risk (FPR).”

One of the justifications for why he argues we want to know this statistic is because people often misinterpret the *p*-value as the probability that your results occurred due to chance. He writes: “The most common (mis)interpretations are “the P value is the probability that your results occurred by chance”” and continues with: “The probability that your results occurred by chance is not the P value: it is the false positive risk”. In other words according to Colquhoun the false positive risk is something that people seem to want to know (or at the very least, some people who misinterpret *p*-values this way).

He continues to explain how likelihood ratios are part of the calculation that your results occurred by chance, but concludes that likelihoods are not enough to tell you what you want to know: “However calculating the likelihood ratio still doesn’t tell us what we really want to know, the false positive risk.” and “What we really want to know is the false positive risk, and for that we need al Bayesian approach.”.

His practical recommendation is to continue to provide *p*-values, although it is not clear what they tell us: “Continue to give P values and confidence intervals. These numbers should be given because they are familiar and easy to calculate, not because they are very helpful in preventing you from making a fool of yourself. They do not provide good evidence for or against the null hypothesis. Giving confidence intervals has the benefit of focusing attention on the effect size. But it must be made clear that there is not a 95% chance that the true value lies within the confidence limits you find. Confidence limits give the same sort of evidence against the null hypothesis as P-values, i.e, not much.” He acknowledges that there is value in interpreting the effect size. He thinks “Perhaps the best way of indicating the strength of evidence provided by a single P value is to use the reverse Bayesian method”.

There is no acknowledgement that there are contexts in which *p*-values answer a question of interest. I think it is fair summary to conclude that Colquhoun thinks his false positive risk is the best statistic to report. He does acknowledge we need other information to interpret results (e.g., effect sizes) and that all statistics are limited, and we need replication studies in the end.

Quote 3: Kirk (1996) “*What we want to know is the size of the difference between A and B and the error associated with our estimate”.*

Kirk start his article with explaining three major criticisms on NHST. In the first, he repeats Cohen’s 1994 claim that “The first criticism is that the procedure doesn't tell researchers what they want to know.” The second criticism is that NHST I a “trivial exercise” because the null hypothesis is always false to some degree. Both these points make it clear NHST is not a question of interest, according the Kirk. He writes “As we have seen, the rejection of a null hypothesis is not very informative. We know in advance that the hypothesis is false.” (p. 748)

He the gives an overview of effect sizes and their development, and discusses whether researchers actually complement NHST with an effect size. He then discusses practical significance as an alternative to statistical significance. He makes a strong statement that correctly interpreting *p*-values will not lead to progress. He writes (p. 753):

“As we have seen, the null hypothesis significance test is often misinterpreted. One response to this unfortunate state of affairs is to admonish researchers to clean up their act, start interpreting

significance tests correctly, and get on with the business of science. I believe that even when a significance test is interpreted correctly, the business of science does not progress as it should."

He explicitly rejects the idea that directional tests are sufficient to learn something in science (going against Cohen, 1995 – he does not cite Cohen, 1995, but argues Cohen 1994 is a 'classic'). He writes (in the section from which I have taken his quote, p. 754):

"What does a researcher learn from a failure to reject the null hypothesis? Because all null hypotheses are false, John Tukey (1991) observed that a nonrejection simply means that the researcher is unable to specify the direction of the difference between the conditions. On the other hand, a rejection means that the researcher is pretty sure of the direction of the difference. Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough."

He follows this later by stating: "What we see is a reject - nonreject decision strategy that does not tell us what we want to know and a preoccupation with *p* values that are several steps removed from examining the data." (p. 755)

Nowhere in the article does he admit *p*-values in isolation have the ability to make a useful contribution. Note that I largely agree with his point that practical significance is important and that it is an improvement to supplement NHST with tests based on meaningful effect sizes (e.g., see my work on equivalence tests, Lakens, 2017). But I want to leave it up to researchers to justify the question they are asking, which might, sometimes, just be a directional prediction answered by a *p*-value.

Quote 4: Blume (2011) *'what we really want to know is how likely it is that the observed data are misleading'*.

Blume provides an educational overview of likelihood inferences. He is the most balanced of all quotes included, and admits different questions can be interesting, when discussing 1) strength of evidence and 2) error rates and 3) the probability observed data will be misleading: "Scientists will always look for each of the three evidential quantities; they represent important concepts and they have distinct roles in the scientific process." (p. 499) and "All three quantities are essential to science and to statistics." (p. 494) But he clearly prefers the third question, at least when the data is in: "The insight is this: Once data are collected, EQ2 becomes irrelevant. What 'might have been' becomes irrelevant. It is EQ3, what 'might be', that is the relevant quantity once data are collected." And "Lastly, it is worth reiterating that the likelihood principle indicates that EQ2 is irrelevant once the data have been collected. At that point, EQ1 and EQ3 are the only quantities of interest." Later, he notes the first 2 evidential questions are not that interesting: "EQ1 needs to be explicitly defined and we need to be told how to use it. Also, does is make sense to define EQ2 in a way that is dependent on the prior probabilities? So here too, the lack of an evidential framework leaves the approach so opened ended that its utility is not clear."

To get to the quote used in the manuscript, Blume admits that error control is part of the puzzle, but again stresses this is irrelevant when the data is in (p. 509): "While EQ2 is an important piece of the puzzle, it is irrelevant once observations are collected. At the end of a study, EQ2 is often confused with the probability that the observed evidence is misleading (EQ3). Once we collect data, the design probability is irrelevant — the data are either misleading or not and we do not know which. So what we really want to know is how likely it is that the observed data are misleading." He ends by saying "The Likelihood approach is used to measure the strength of evidence in observed data. It avoids the pitfalls of other statistical paradigms because it has a well-defined evidential framework."

So I think it is fair to interpret his viewpoint as saying that the likelihood is what we want to know. He does admit error control plays a role before the data is in, but his article is a clear argument for why the likelihood approach reflects what we want to know.

Quote 5: Bayarri, Benjamin, Berger, and Sellke (2016) '*we want to know how strong the evidence is, given that we actually observed the value of the test statistic that we did'*.

Bayarri and colleagues propose some changes to current hypothesis testing practices that fix the major issues, as they see them. They write (p. 91): "Our proposal – developed throughout the paper and summarized in the conclusion – is that researchers should report what we call the 'pre-experimental rejection ratio' when presenting their experimental design, and researchers should report what we call the 'post-experimental rejection ratio' (or Bayes factor) when presenting their experimental results"

The quote context is (p. 91):
"For example, it is sometimes incorrectly said that p = 0.05 means that there was only a 5% chance of observing the data under H0. (The correct statement is that p = 0.05 means that there was only a 5% chance of observing a test statistic as extreme or more extreme as its observed value under H0 – but this correct statement is not very useful because we want to know how strong the evidence is, given that we actually observed the value of the test statistic that we did.)"
So the authors directly say that the 'correct' interpretation of a *p*-value is not very useful, because we want to know how strong the evidence is (i.e., have a Bayesian interpretation). I feel this is a fair example of the Statisticians' Fallacy. They could have said 'but there are probably only a limited set of studies where the p-value provides an interesting answer to a question, and we expect there are many studies where what researchers mainly want to know is….'.

Similar to Blume (2001), Bayarri et all acknowledge that before the data is in, other aspects of the design are relevant. They write (p. 92): "We want to know: if we run the experiment, what are the odds of correct rejection of the null hypothesis to incorrect rejection? We call this quantity the 'pre-experimental rejection odds' (sometimes dropping the word 'rejection' for brevity)." But this is only something we want to know when we present the design of our study – when we present the results, the post-experimental odds are, according to the authors, what we want to know (p. 92): These recommendations can be boiled down to two: researchers should report the pre-experimental rejection ratio when presenting their experimental design, and researchers should report the post-experimental rejection ratio when presenting their results.

The authors are not saying there is only one thing we want to know. They write: "A number of alternative statistical methods have been proposed, including several in this special issue, and we are sympathetic to many of these proposals. In particular we are highly sympathetic to efforts to wean the scientific community away from an over-reliance on hypothesis testing, with utilization of often-more-relevant estimation and prediction techniques". But if we test a hypothesis, they present their alternative to *p*-values without acknowledging that *p*-values in themselves might ever be the answer you care about, and refer in an unspecified manner to what 'we want to know'. I think this lack of context in which we want to know their preferred statistical question warrants categorizing this as the Statistician's Fallacy.

Quote 6: Mayo (2018) '*We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out?*'

The main reason I believe I am doing Mayo justice is because she said so on Twitter after reading my preprint. https://twitter.com/learnfromerror/status/1115637135846068224. Mayo uses the sentence 'we want to know' three times in her book. The full quote I use in the manuscript is (p. 300):

"We don't seek a probabilist assignment to a hypothesis or model. We want to know what the data say about a conjectured solution to a problem: What erroneous interpretations have been well ruled out? Which have not even been probed? The warrant for these claims is afforded by the method's capabilities to have informed us of mistaken interpretations. Statistical methods are useful for testing solutions to problems when this capability/incapability is captured by the relative frequency with which the method avoids misinterpretations."

For Mayo, severity is the underlying goal of all statistical methods (and it is what should get us beyond the statistics wars, in her book). It is clear she does not think a *p*-value by itself will be enough, after the data is in. Before the data is collected, power is important. After data is in, Mayo argues we want a severe test:

"Why object to applying the severity analysis by changing the null hypothesis, and doing a simple P-value computation? P-values, especially if plucked from thin air this way, are themselves in need of justification. That's a major goal of this journey. It's only by imagining we have either a best or good test or corresponding distance measure (let alone assuming we don't have to deal with lots of nuisance parameters) that substituting different null hypotheses works out. Pre-data, we need a test with good error probabilities (as discussed in Section 3.2). That assures we avoid some worst case. Post-data we go further. For a claim H to pass with severity requires not just that (S-1) the data accord with H, but also that (S-2) the test probably would have produced a worse fit, if H were false in specified ways."

She acknowledges long run error rates are important – but we also want to evaluate the data at hand (p. 14): "I do not mean to disparage the long-run performance goal – there are plenty of tasks in inquiry where performance is absolutely key. Examples are screening in high-throughput data analysis, and methods for deciding which of tens of millions of collisions in high-energy physics to capture and analyze. New applications of machine learning may lead some to say that only low rates of prediction or classification errors matter. Even with prediction, "black-box" modeling, and non-probabilistic inquiries, there is concern with solving a problem. We want to know if a good job has been done in the case at hand."

The main goal of Mayo's book is to propose an alternative (the title of her book is 'Moving Beyond the Statistics Wars') to Bayesian and Frequentist approaches that exist. It is not surprising she argues for an alternative to what we want to know – but again, there is little discussion of whether there are contexts where other questions are more important.