Comment on D. Mayo: "The statistics wars and intellectual conflicts of interest"

1 Introduction

In her editorial "The statistics wars and intellectual conflicts of interest" in Conservation Biology, D. Mayo discusses and ultimately rejects demands to abandon claims of statistical significance. In particular, she addresses the "no-threshold view" that demands the avoidance of thresholds on p-values to distiguish whether tested hypotheses are significantly rejected or not.

Here I will argue that that the debate about statistical significance and p-values reflects deep and essential difficulties with statistical and probabilistic reasoning that need to be acknowledged in order to appropriately understand uncertainty in the reasoning from data. These difficulties are manifest in all statistical reasoning, and I believe, in line with Mayo, that they cannot be resolved or made to disappear by abandoning a well established approach based on existing misunderstanding and misuse.

2 Difficulty

Statistics is hard. Well-trained, experienced and knowledgeable statisticians disagree about standard methods. Statistics is based on probability modelling, and probability modelling in data analysis is essentially about whether and how often things that did not happen could have happened, which can never be verified. The very meaning of probability, and by extension of every probability statement, is controversial.

There is also agreement among statisticians, and what they agree about provides the most reliable guideline statisticians have to offer when it comes to questions like the one Mayo raised in her editorial: "How should journal editors react to heated disagreements about statistical significance tests in applied fields, such as conservation science, where statistical inferences often are the basis for controversial policy decisions?"

Here are statements regarding p-values that are generally agreed among statisticians, as far as I discern. They are largely in line with the 2016 ASA Statement on p-values, Wasserstein & Lazar, 2016:

i) p-values do not indicate the probability that the null hypothesis is true.

ii) An insignificant p-value does not mean that the null hypothesis is true.

iii) A significant p-value does not mean that the alternative hypothesis is true; apart from a type I error it may also be caused by violation of model assumptions.

iv) Multiple testing and data dependent selection of tests invalidate p-values and error probabilities of tests, unless they are explicitly and appropriately adjusted.

v) A p-value does not measure the strength of an effect.

On the positive side, the basic idea of statistical tests is that a statistical model can be deemed incompatible with the data if an event happens that is very unlikely to happen given the model. This is a simple, direct, and intuitive idea that I have hardly ever seen disputed among statisticians. Generally, the null model should not believed to be true (and neither should any other model). A p-value is surely informative; regarding given data, compatibility is the best that models can ever achieve, keeping in mind that many models can be compatible with the same data.

Statisticians also agree that misunderstanding and misuse of tests and p-values are endemic. There are issues with tests and p-values about which there is disagreement even among the most proficient experts. For example, there are no agreed guidelines regarding when and how exactly corrections for multiple testing should be used, or under what exact conditions a model can be taken as "valid". Such decisions depend on the details of the individual situation, and there is no way around personal judgement.

The fact that p-values (and statistical reasoning in general) regard idealized models that are different from reality seems to be hard to stomach and easy to ignore; contrarily sometimes this is interpreted as testifying the uselessness of p-values ("Why would we test a null hypothesis that we do not believe to be true anyway?") or frequentist statistical inference in general. It seems more difficult to acknowledge how models can help us to handle reality without being true, and how finding an incompatibility between data and model can be a starting point of an investigation how exactly reality is different and what that means. For this, a test gives a rough direction (such as "the mean looks too large"), which can be useful, but is certainly limited as information.

I do not think that these are defects specific to p-values and tests. The task of quantifying evidence and reasoning under uncertainty is so hard that problems of these or other kinds arise with all alternative approaches as well, acknowledging that there are specific misunderstandings concerning certain approaches, and there is useful empirical research concerning their occurrence (e.g., Coulson et al. 2010).

3 Tension

A much bigger problem is the tension between the difficulty of statistics and the demand for it to be simple and readily available. Data analysis is essential for science, industry, and society as a whole. Not all data analysis can be done by highly qualified statisticians, and society cannot wait with analysing data for statisticians to achieve perfect understanding and agreement. On top of this there are incentives for producing headline grabbing results, and society tends to attribute authority to those who convey certainty rather than to those who emphasize uncertainty. Statistics provides standard model based indications of uncertainty, but on top of that there is model uncertainty, uncertainty about the reliability of the data, and uncertainty about appropriate strategies of analysis and their implications. A statistician who emphasizes all of these will often meet confusion and disregard.

Another important tension exists between the requirement for individual decision-making depending on the specifics of a situation, and the demand for automated mechanical procedures that can be easily taught, easily transferred from one situation to another, justified by appealing to simple general rules (even though their applicability to the specific situation of interest may be doubtful), and investigated by statistical theory and systematic simulation. Any threshold for p-values will seem inadequate in many situations, yet a threshold is required to enable theoretical statements about error probabilities; furthermore language is discrete and any interpretation of a p-value in words (let alone potential binary decisions to be made based on data) will implicitly rely on thresholds.

p-values are so elementary and apparently simple a tool that they are particularly suitable for mechanical use and misuse. To have the data's verdict about a scientific hypothesis summarized in a single number is a very tempting perspective, even more so if it comes without the requirement to specify a prior first, which puts many practitioners off a Bayesian approach. As a bonus, apparently well established thresholds allow to make a binary "accept or reject" statement. Of course all this belies the difficulty of statistics and a proper account of the specifics of the situation.

Alternative statistical approaches have their merits and pitfalls, too, always including the temptation to over-interpret their implications, often by taking the assumed model as a truth rather than a model (also a Bayesian model of belief should not just be believed). The pessimistic belief seems realistic that the general popularity and spread of any statistical approach will correspond to its capacity of being mechanically used, misused, and over-interpreted, making it easy for its opponents to criticize it. Once more looking at the Bayesian alternative to p-values, supposedly "objective" priors that do not encode existing information are most popular, depriving the Bayesian approach of a major benefit. Vast amounts of applied Bayesian literature make no or only a very deficient attempt to motivate the prior from existing information, probably due to the difficulty of specifying and justifying "subjective" informative priors, in contrast to the simplicity of using a readily available default.

4 Dilemma

As statisticians we face the dilemma that we want statistics to be popular, authoritative, and in widespread use, but we also want it to be applied carefully and correctly, avoiding oversimplification and misinterpretation. That these aims are in conflict is in my view a major reason for the trouble with p-values, and if p-values were to be replaced by other approaches, I am convinced that we would see very similar trouble with them, and to some extent we already do.

Ultimately I believe that as statisticians we should stand by the complexity and richness of our discipline, including the plurality of approaches. We should resist the temptation to give those who want a simple device to generate strong claims what they want, yet we also need to teach methods that can be widely applied, with a proper appreciation of pitfalls and limitations, because otherwise much data will be analyzed with even less insight. This may be seen as a somewhat contradictory message, as advertising and criticizing an approach at the same time, and may not necessarily increase the public trust in statistics. But I think that it is more genuine than using agreed issues with one approach to advertise another one as solving all the problems.

5 Conclusion

When it comes to a representative association such as ASA, I think that the approach taken in the 2016 statement followed this ideal and was as such valuable. I would have hoped that the assertions made could be accepted by a vast majority of statisticians despite much existing disagreement, maybe tolerating disagreement with certain details of the statement. The "2019 editorial" (Wasserstein et al. 2019) had a different spirit by recommending to "abandon" methodology that a substantial number of statisticians routinely use and defend. This was obviously not something that could hope for broad agreement, and I think it was quite damaging for the profession. If we see ourselves as flag bearers of the acknowledgement and communication of uncertainty (and I think we should define ourselves in this way), this task alone puts us in a difficult position with a public who expect certainty and quick results. Regarding methodological controversies within our profession, we should be pluralist and open for the arguments of each side, rather than trying to shut one side out.

What we should like to see is scientists (and other statistics users) who are aware of the many sources of uncertainty and misunderstanding, and interpret their results keeping this in mind. As statisticians we should not convey the impression that whether things are done right or wrong is a matter of the chosen statistical approach as long as it finds support within the statistics community. Instead it is a matter of awareness of the limitations of whatever they do.

References

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical testing. Frontiers in Psychology, 1. DOI=10.3389/fpsyg.2010.00026

Mayo, D. G. (2021). The statistics wars and intellectual conflicts of interest. Conservation Biology. DOI: 10.1111/cobi.13861

Wasserstein, R., & Lazar, N. (2016). The ASA’s statement on p-values: Context, process and purpose. American Statistician, 70(2), 129–133.

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond “p < 0.05”. American Statistician, 73(S1), 1–19.