Are P-values the problem?

Stephen Senn

I welcome the ASA report. In my discussion I shall use *direct interpretation* to refer to P-values as probabilities of (functions of) statistics given hypotheses and *inverse interpretation* to probabilities of hypotheses (or parameters) given statistics.

The ASA report is suitably measured and cautious when discussing P-values, reflecting, in my opinion, a view, expressed by several of those consulted, that many of the problems of which P-values are accused are problems of inference generally, not problems of P-values *per se*.

Amongst these problems of inference are the purely scientific such as

1. Inference is difficult
2. The inferential content of (say) an experiment cannot easily be summarised in one statistic (recognised in the final sentence of the ASA statement)

but also problems of human psychology and society such as

3. An impatience with the necessary nuances of expression that good statistical reporting requires
4. The (usual) prejudice of scientific journals in favour of 'positive' results(Senn 2013)
5. The common habit of transforming shades of grey into either black or white
6. The desire of individual scientists for recognition and reward.

My view is that P-values are statistics that play a definite but limited role in statistical inference(Senn 2001), that inferences will be all the better for recognising their limitations but will be worse if we attempt to replace them rather than supplement them.

In fact, I claim that much current criticism of P-values is misplaced and reflects a false history of how they came about. A common story is as follows

1. For over 125 years scientists were happily calculating posterior probabilities using (what are now called) Bayesian approaches
2. Starting in the 1920s RA Fisher(Fisher 1925) persuaded them to calculate P-values instead
3. Because Fisherian P-values overstate 'significance' they became very popular with scientists
4. We need to return scientists to the path of Bayesian virtue

But this history is false. Tail area probabilities were being calculated well before Fisher but were given an inverse interpretation. Student's famous paper(Student 1908) gives an example. They were also occasionally given the modern direct interpretation provided in the informal definition in the ASA statement. See, for example, Karl Pearson's chi-square paper(Pearson 1900). Fisher pointed out that inverse interpretation was highly dependent on the assumed prior distribution. The subsequent neo-Bayesian revolution proved him right. He stressed the direct interpretation as being safer. Of course he introduced a whole host of statistical techniques, including many tests, but his most influential technical innovation as regards P-values *per se* was to suggest a doubling. This is not uncontroversial, but, since it increases the P-value, cannot be represented as giving significance more easily than what went before(Senn 2015b).

The origin of the modern claim that P-values give significance too easily is that the direct interpretation is compared to an indirect interpretation using a *differen*t Bayesian system: not the one developed by Laplace(Laplace 1951) but that which Harold Jeffreys(Jeffreys 1961) developed between the two wars(Senn 2015a). The response of Jeffreys to the challenge Broad made in showing that the Laplacian formulation could not provide an appreciable *posterior* probability of a scientific law being true(Broad

1918) was to place a lump of *prior* probability on its being true. (An early alternative development by (Haldane 1932) has been noted by (Etz and Wagenmakers 2015)).

The distinction is mainly between testing *point* & *dividing* hypotheses. ((Cox 1977) uses *plausible* for the former.) In the frequentist framework it makes very little difference which you use; in the Bayesian it is crucial(Casella and Berger 1987). However, many (see, for example (Colquhoun 2014)) have implicitly assumed that in re-calibrating direct inferences as inverse ones, no recalibration of significance levels is necessary. In my view this is as false as to argue, in moving to dosing patients by weight limit rather than age limit, that an age group that was 10 years and older should now become 10kg and heavier(Senn 2001).

P-values should be retained for a limited role as part of the machinery of error-statistical approaches. Even within that system they need to be supplemented by other devices(Mayo 1996). This system is valuable precisely because it is independent of the Bayesian one and trying to make it more Bayesian in behaviour misses the point.

I am not arguing against Bayesian inference. I am arguing that it is valuable when those employing it pay very careful attention to appropriate specification of prior distributions making explicit the (prior) evidence on which these rest. This turns out to be much more difficult than many suppose(Senn 2011, 2007). In my opinion, it is advisable to go beyond default Laplacian or Jeffreys (point) prior distributions and certainly there is no point in modifying P-values to make them more ‘Bayesian’.

In short, the problem is less with P-values *per se* but with making an idol of them. Substituting another false god will not help(Gigerenzer and Marewski 2015).

I welcome the ASA statement as a sensible and measured contribution to improving the use of statistical inferential methods.

Acknowledgement

This work was supported by the European Union’s 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. “IDEAL”

References

Broad, C. 1918. "On the relation between induction and probability." *Mind* no. 27:389-404.

Casella, George, and Roger L Berger. 1987. "Reconciling Bayesian and frequentist evidence in the one-sided testing problem." *Journal of the American Statistical Association* no. 82 (397):106-111.

Colquhoun, David. 2014. "An investigation of the false discovery rate and the misinterpretation of p-values." *Royal Society Open Science* no. 1 (3):140216.

Cox, D.R. 1977. "The role of significance tests." *Scandinavian Journal of Statistics* no. 4:49-70.

Etz, Alexander, and Eric-Jan Wagenmakers. 2015. "Origin of the Bayes Factor." *arXiv preprint arXiv:1511.08180*.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Gigerenzer, Gerd, and Julian N Marewski. 2015. "Surrogate Science The Idol of a Universal Method for Scientific Inference." *Journal of Management* no. 41 (2):421-440.

Haldane, JBS. 1932. A note on inverse probability. Paper read at Mathematical Proceedings of the Cambridge Philosophical Society.

Jeffreys, H. 1961. *Theory of Probability*. Third ed. Oxford: Clarendon Press.

Laplace, P.S. 1951. *A philosophical essay on probabilities (English translation)*. Toronto: Dover.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Pearson, K. 1900. "On the criterion that a given sytem of deviation from the probable in a correlated system of variables is such that it can reasonably supposed to have arsien from random sampling." *Philosophical Magazine* no. 50 (5):157-175.

Senn, S.J. 2001. "Two cheers for P-values." *Journal of Epidemiology and Biostatistics* no. 6 (2):193-204.

Senn, S.J. 2007. "Trying to be precise about vagueness." *Statistics in Medicine* no. 26:1417-1430.

Senn, S.J. 2011. "You may believe you are a Bayesian but you are probably wrong." *Rationality, Markets and Morals* no. 2:48-66.

Senn, Stephen. 2013. *Authors are also reviewers: problems in assigning cause for missing negative studies* 20132013]. Available from http://f1000research.com/articles/2-17/v1.

Senn, Stephen. 2016. *Double Jeopardy?: Judge Jeffreys Upholds the Law* 2015a [cited 13 February 2016]. Available from http://errorstatistics.com/2015/05/09/stephen-senn-double-jeopardy-judge-jeffreys-upholds-the-law-guest-post/.

Senn, Stephen. 2016. *The Pathetic P-Value* 2015b [cited 13 February 2016 2016]. Available from http://errorstatistics.com/2015/03/16/stephen-senn-the-pathetic-p-value-guest-post/.

Student. 1908. "The probable error of a mean." *Biometrika* no. 6:1-25.