p-value Limitations

Because science has for many decades been using *p*-values to confirm novel experimental results, trends of *p*-value misuse have also emerged. Scientific journals, traditional new outlets and a to a lesser extent data science product managers have at least one thing in common—they all need new and exciting findings.

A *p*-value is the probability of finding the observed, or more extreme, results when the null hypothesis (H-zero) of a study question is true.

In other words it is a tool to quantify the evidence against the null hypothesis. The null hypothesis is rejected when this evidence is under some predefined, but arbitrarily defined threshold. It turns out there are many ways to modify the data get the value under the threshold.

- If we re-run the experiment multiple times changing the features until we arrive at the best *p*-value
- If we go back and collect a few more samples to get the *p*-value just over the threshold
- If we remove one or more outliers to ensure that the *p*-value is smaller
- If we set the level of *alpha* after the hypothesis has been tested

These scenarios and a number of others are known as *p*-value hacking and they are carried out because there is a push in academia and in business for novel and impactful findings. If we return to the core idea of *degree of belief* from the module on *scientific thinking* then the notion of **reproducibility** should outweigh the importance of novelty.

In the A/B testing example with two different versions of a website, we used different *p*-values as an investigative tool, not as specific number to base decisions on or draw conclusions from. We also showed another powerful tool that does not use *p*-values at all, that is the posterior distribution. The degree of belief for that experiment was quantified as the posterior distribution, which is a far more informative tool for decision making than the *p*-value itself. Viable alternatives to *p*-values exist through the use of Bayes Factors and Posterior Predictive Checks.

The important thing to remember is that *p*-values themselves are not a source of ground truth, but they are nonetheless quite useful if used appropriately. If the testing of a specific model is business critical then it might be worth taking the time to test the ideas within a Bayesian framework. This can give you more confidence in your conclusions for a given experiment, but whether it is Bayesian or frequentist treatment of the experiment the study still needs to be repeated with newly collected data to ensure reproducibility.

Additional resources

- Wiki article on Data Dredging
- An interactive tool that illustrates p-hacking
- Bayesian Estimation instead of a T-Test
- Bayes Factors in PyMC3
- Posterior Predictive Checks PyMC3
- Nature Methods article about interpreting *p*-values