nature research

https://doi.org/10.1038/s41593-020-0660-4

Supplementary information

Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence

In the format provided by the authors and unedited

Supplementary note

Continuous testing

When planning experiments, it is often difficult to know in advance how many samples should be collected. The temptation is to start with a small sample of participants (e.g., n=10) and see if results are significant (e.g., p<0.05). If they are, great; if they are not, collect more data until findings are significant or resources run out. Doing so using NHST without correcting for the number of sequential tests is of course deeply flawed, because it will ultimately lead to significant results in all cases, whether the effect is real or not (Figure S1 top row). Corrections for multiple comparisons can protect against the danger to falsely reject H_0 when there is no effect, but the sensitivity of the test is then reduced when there is an effect (Figure S1 second row). What makes frequentist approaches so problematic in this context is that the p-value can only provide evidence against H_0, but not for H_0; continuing to data and test until one has a significant result means continuing until we reject H_0, and hence, eventually we will always end up rejecting H_0.

Bayesian statistics offer a principled way out of this bias by using a *symmetrical* stop rule enabled by the intrinsically symmetrical nature of the Bayes factor. Instead of stopping when p<0.05, you stop including additional cases whenever the evidence favors H_1 *or* H_0. Based on considerations of sensitivity, one can choose a more or less conservative critical value. The bottom two rows in Figure S1 show that with this approach one reaches appropriate decisions in most cases with relatively small sample sizes. A critical value of 3 has the advantage that decisions are reached earlier (in 80% and 90% of cases in samples of ~20 and about ~30, respectively) and with reasonable accuracy (False alarm < 10% and False rejection <14%). Using a stricter critical value of 6 provides later decisions (in 80% and 90% of cases around n=60 and n=90, respectively) but they have higher accuracy (False alarm<7% and False rejection <2%).

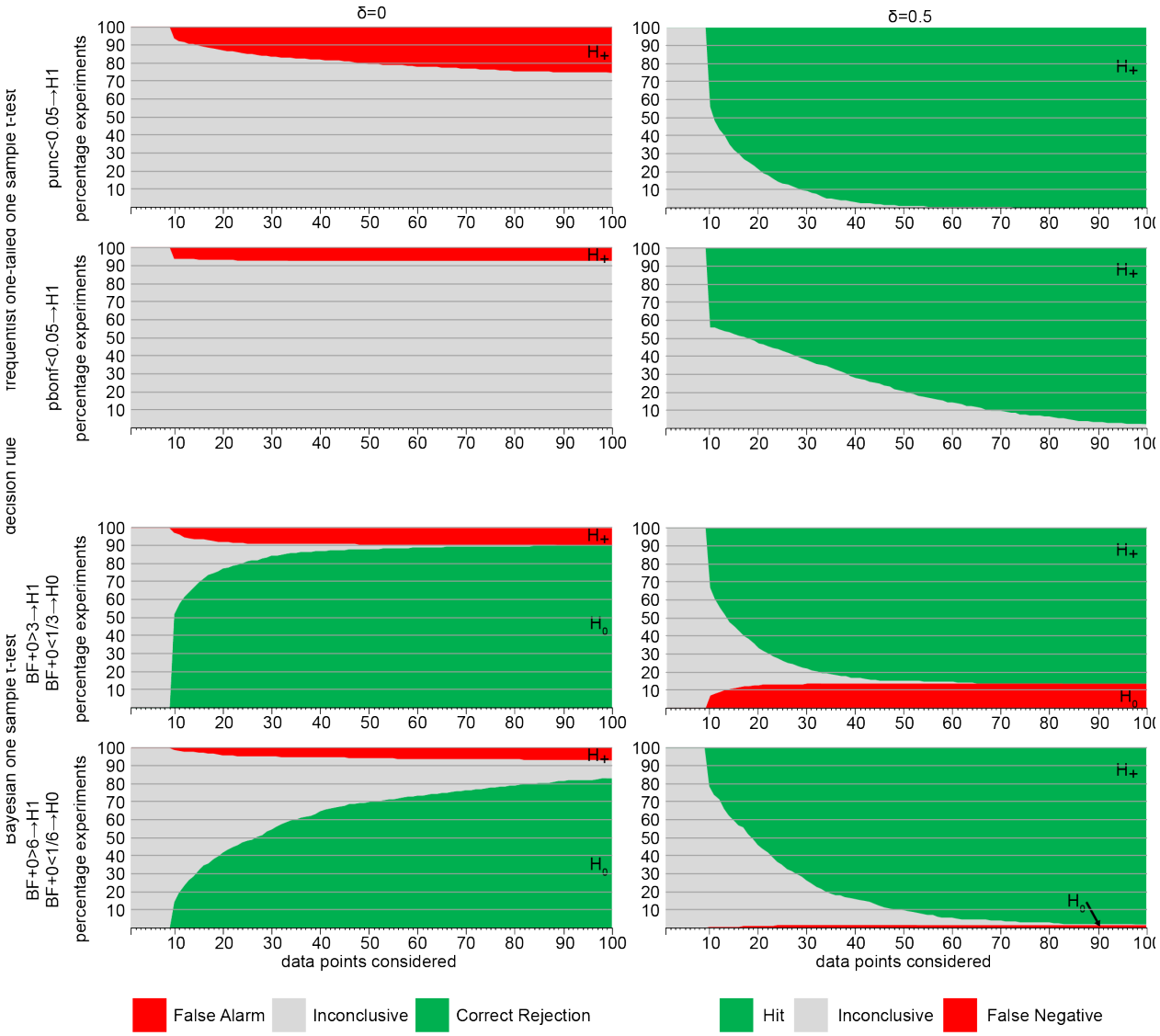

Figure S1 | The effect of decision rule on sequential testing. We simulated 1000 random experiments. For each experiment, we generated 100 random numbers using N(μ=δ,σ=1), where δ is the effect size. After we collected the first n=10 data points, we applied our decision rule for the first time. Four decision rules were used. Top Row: we applied a frequentist one-tailed one-sample t-test. If p<0.05, we classified the experiment as supporting H_+, otherwise as inconclusive. If the test remained inconclusive, we considered one further datapoint, and repeated the rule considering n=11 data points, and so on. Second Row: same but applying a Bonferroni correction for the number of times we used the criterion, so that our criterion becomes p*(n-9)<0.05. Third row, we applied a Bayesian one-tailed one-sample t-test using a critical value of 3. Here, if BF_{+0}<1/3, we classified the experiment as supporting H_0, if BF_{+0}>3 as supporting H_+, otherwise as inconclusive. Fourth row, as third row but using 6 as the critical value. The same decision rules were applied for all 1000 simulated experiments. On each position n on the x-axis, we then display the percentage of the 1000 simulated experiments that were classified as supporting H_0, H_+ or inconclusive. For δ=0, supporting H_0 are displayed as correct rejections (green, only for Bayesian statistics), supporting H_+ as false alarms (red) and remaining inconclusive in gray. For δ=0.5, supporting H_0 are considered false negatives (red), supporting H_1 as hits (green) and remaining inconclusive are shown in gray. Note that for the frequentist approach (top two rows), either too many false alarms (top left) are made, or sensitivity is low (second row, right). Data can be found at https://osf.io/md9kp/.