Rianne de Heide
Faculty of EEMCS
University of Twente
r.deheide@utwente.nl

Column Rianne grijpt haar kans

Optional stopping

Rianne de Heide zal op regelmatige basis in dit blad een column schrijven over een actueel statistisch onderwerp.

Suppose you are doing a trial on 70 subjects. The *p*-value is promising but just not significant $(\mathrm{p} = 0.06)$. Your boss says there is some more money for adding 10 more subjects to the trial. What do you do?

A. You add 10 subjects to the study, and you calculate a new *p*-value based on the total data, i.e. 80 subjects.
B. You calculate a new *p*-value for the 10 new subjects, and you multiply that *p*-value by 0.06, the *p*-value from the first 70 subjects.
C. You say to your boss: sorry, this is not possible. You cannot conclude anything other than that you cannot conclude anything.
D. You calculate a new *p*-value for the 10 new subjects, and you use a method they also use in meta-analyses to combine the *p*-values.

This is a question I included in my talks for applied statisticians the last months. Although I am always happy to see that (almost) no-one wants to multiply two *p*-values, the high number of people in the audience choosing answer A is a concern! Let's get to the bottom of this!

First, a reminder on the definitions of a *p-value* and a *test* is always nice. A *p*-value p is a $[0,\infty)$-valued random variable satisfying $\mathrm{P}(\mathrm{p} \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ and for all $\mathrm{P} \in \mathrm{H}_0$, that is, for all distributions in the collection of distributions that form the null hypothesis.

Typically *p*-values take values in $[0,1]$. *P*-values with the last inequality replaced by an equality are called *exact*. A *binary test* ϕ is a $\{0,1\}$-valued random variable, and has Type I error $\mathbb{E}_{\mathrm{P}}[\phi]$, again for $\mathrm{P} \in \mathrm{H}_0$. The Type I error is thus the probability of rejecting the null when it is true: a false positive. Lots of frequentist statistics focusses on bounding the Type I error. Often this is done by prespecifying the number α, what is called the (*significance*) *level* of the test, and then designing a test which guarantees a proportion of Type I errors of at most α. Classical hypothesis testing bases these tests on *p*-values, which are, conveniently, even defined in terms

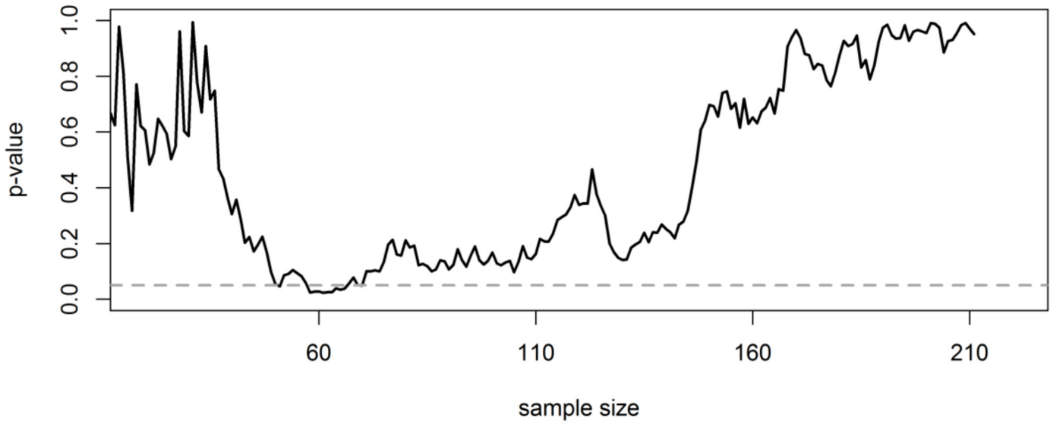

Figure 1 *P*-values under optional stopping

of a Type-I error guarantee!

However, in real life, working with *p*-values is pretty restricted: the whole *sampling plan* needs to be specified upfront. One practice that is pertinently not allowed, is what is happening in answer A, and this has a name: *optional stopping*. Informally this means 'looking at the results so far to decide whether or not to collect some more data', and it invalidates *p*-values and their error guarantees. Why is that?

In Figure I have plotted *p*-values under optional stopping. I used the most common test in applied statistics: the *t*-test. In a *t*-test, the null hypothesis is that the mean of a Gaussian with arbitrary variance is 0, and the alternative hypothesis is that the mean is different from zero, again with arbitrary variance. For this picture, I generated data points from a standard normal distribution $\mathrm{N}(0,1)$, thus, the null hypothesis is true: the data-generating distribution is indeed a Gaussian with mean zero. I started with two data points, and I calculated the *p*-value based on the *t*-test. Next, I generated a third data point, again from $\mathrm{N}(0,1)$, and I calculated the *p*-value based on the three data points I had. I plotted the *p*-value and drew a line between the two *p*-values. I continued to do this for 210 data points, and that is how Figure 1 came about.

What do we see? The *p*-values go a bit up, a bit down, a bit up again. Can we quantify how often the *p*-values are in some subinterval of $[0,1]$? Well, yes: as we can see from the definition, exact *p*-values are distributed uniformly under the null, i.e. when the null hypothesis is true — the true data generating process is a distribution in the null hypothesis — which is the case in the picture. That means that the *p*-value is guaranteed to dive below the dashed line $\alpha = 0.05$, but of course this holds for any $\alpha \in (0,1)$ we choose. In a formula: $\mathrm{P}(\exists t \in \mathbb{N}: p_t \leq \alpha) = 1, \forall P \in H_0$: the probability that there exists a sample size for which the *p*-value dives below the α-line, is one. And it will even do so infinitely often. We are guaranteed to find a significant *p*-value while the null hypothesis is true: so we are guaranteed to falsely reject the null if we do optional stopping with *p*-values. This is why optional stopping with *p*-values is not allowed. And many researchers do this anyway. A study among psychologists [1] shows that the percentage of them doing optional stopping is even as high as 55%.

As to answer D: in meta-analyses, just as in this example, studies are not independent. The second study, or our experiment on the additional sample of 10 subjects, would not have been performed if the first part did not yield promising results. Therefore, there are complicated dependencies between the two studies, and no valid *p*-value tests can be constructed [2].

So, the right answer was answer C: nothing can be done, when one has decided to work with *p*-values. However, modern research provides a solution for this problem, and the paradigm including, among others, tests that can handle optional stopping is called *any-time valid inference*.

Referenties

1 John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5), 524-532. doi: 10.1177/0956797611430953

2 Ter Schure, J., Grünwald P. (2019) Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000Res.* June 25;8:962. doi: 10.12688/f1000research.19375.1.