Tidsskriftet
DEN NORSKE LEGEFORENING

Statistical power: Before, but not after!

MEDISIN OG TALL

STIAN LYDERSEN

E-mail: stian.lydersen@ntnu.no
Stian Lydersen, dr.ing. and professor of medical statistics at the Regional Centre for Child and Youth Mental Health and Child Welfare (RKBU Central Norway), Department of Mental Health, Norwegian University of Science and Technology (NTNU).
The author has completed the ICMJE form and declares no conflicts of interest.

Some journals, and some reviewers, request a calculation of statistical power based on the observed effect size after a study has been carried out. This is fundamentally flawed.

Statistical power is the probability of rejecting the null hypothesis in a future study. After the study has been carried out, this probability is 100 % (if the null hypothesis was rejected) or 0 % (if the null hypothesis was not rejected).

Before starting up a study, it is recommended to calculate the statistical power or sample size. This calculation is based on an expected effect size, or on an effect size regarded as clinically important. The statistical power is the probability that the result of a study with a given number of participants will be statistically significant (1). Alternatively, we calculate the number of participants needed to obtain a given statistical power, for example 90 %, for this effect size. The calculations are performed for a given significance level, usually 0.05.

'Observed power' and p-value

After the study, it is generally recommended to report an estimate and a 95 % confidence interval for the effect, as well as a p-value. Sometimes, a journal or a reviewer requests a calculation of the 'observed power' in addition (2). This means a statistical power calculation based on the observed effect, as if it were a future study, and reporting it as if it gives additional information about the study already performed. This is not only fundamentally flawed, but it gives no information in addition to the reported p-value: For every statistical hypothesis test, there is a unique correspondence between the p-value and 'observed power'. For a one-sided test with normally distributed outcome and known variance, this correspondence is particularly simple (2). This is illustrated in Figure 1. Here, 'observed power' is over 50 % if the p-value is less than 0.05, and 'observed power' is under 50 % if p is greater than 0.05.

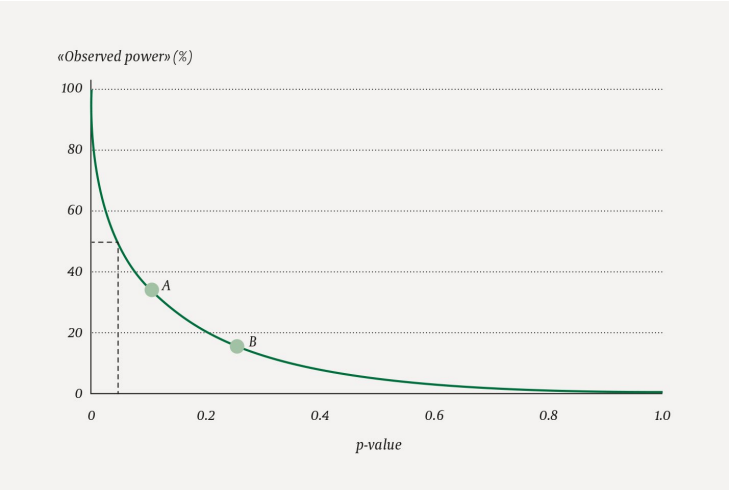

Figure 1 *'Observed power' for a one-sided test with normally distributed outcome variable and known variance, with significance level 0.05. When the p-value equals 0.05, the 'observed power' equals 0.5 (2). Study A has higher 'observed power' than Study B. This does not imply that Study A provides stronger evidence in favour of the null hypothesis: On the contrary, Study A has lower p-value, hence stronger evidence against the null hypothesis.*

'Observed power' provides no additional information beyond the p-value. On the contrary, it can be misleading, something which many researchers seem not to be aware of. Let A and B be two studies where the null hypothesis was not rejected, and the p-values were 0.10 and 0.25, respectively (Figure 1). Study A has higher 'observed power' than Study B. Some may conclude that Study A has strongest evidence in favour of the null hypothesis, which was not rejected, but this is a fallacy. Study A has the lower p-value, and hence, strongest evidence *against* the null hypothesis. This kind of fallacy is called 'the power approach paradox'.

'Observed power' and confidence interval

Other types of retrospective power calculations have been suggested, including this one: Assume a study did not result in the rejection of the null hypothesis. The question is: With the observed variability in the study, what would a hypothetical effect size in a future study need to be to give a certain statistical power, for example 90 %? However, this is also logically flawed, and can lead to a version of the 'power approach paradox', as described in (2). A 95 % confidence interval, on the other hand, indicates the range of effect sizes that are likely, given the observed data.

Reporting statistical power

It is good practice to report a power- or sample-size calculation that was performed before the study was started up. This is recommended in the CONSORT Statement (3) for randomised trials, and helps to document that the study was well planned. After the study has been conducted, a confidence interval and a p-value are appropriate measures of uncertainty. 'Observed power' after the study has been carried out, is both superfluous and misleading.

REFERENCES:

1. Pripp AH. Antalls- og styrkeberegninger i medisinske studier. Tidsskr Nor Legeforen 2017; 137. [PubMed][CrossRef]

2. Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat 2001; 55: 19–24. [CrossRef]

3. The CONSORT 2010 Statement. 7a. Sample size.

http://www.consort-statement.org/checklists/view/32—consort-2010/83-sample-size (30.10.2018)

Published: 28 January 2019. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0847