Tidsskriftet Tidsskrift for Den norske legeforening

Why are one-sided hypothesis tests rarely used?

MEDICINE AND NUMBERS

STIAN LYDERSEN

E-mail: stian.lydersen@ntnu.no
Stian Lydersen, PhD, 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.
The author has completed the ICMJE form and declares no conflicts of interest.

Many hypotheses in medical research are in principle one-sided, for example in a randomised, controlled trial that investigates whether a new type of clinical treatment has a better effect than treatment as usual. So why are two-sided hypothesis tests used?

Let us assume, for example, that we register the number of successful outcomes, meaning the number of patients who recover from the disease, in two separate treatment groups. The null hypothesis (H_0) is that the probability of success is the same in both groups. But what is the alternative hypothesis? This is a trial that seeks to investigate whether the new treatment has a better effect than the standard treatment, that is, a superiority trial. One might assume that the alternative hypothesis is precisely this. This is called a one-sided alternative hypothesis and the appurtenant hypothesis test and p-value are referred to as one-sided. However, if we choose a two-sided alternative hypothesis, that is, that the new treatment produces an effect which is different from that of the standard treatment, we have a two-sided hypothesis test and an appurtenant two-sided p-value.

Greater power of one-sided tests?

An argument in favour of choosing a one-sided test is that it has greater statistical power than the corresponding two-sided test. Let us assume that we are planning a randomised, controlled trial and want a high probability of claiming a difference in effect if the probabilities of success with the standard treatment and new treatment are 0.6 and 0.8, respectively. If we plan to use a two-sided test, we would need 82 patients in each group to achieve statistical power of 80 % at a significance level of 0.05. If we plan to use a one-sided test, on the other hand, 64 patients in each group will be sufficient.

Let us assume that this trial was undertaken with 100 patients in each group. In the group that received the standard treatment 64 patients recovered, while in the group with the new treatment 76 recovered. The estimated difference in the probability of success is 76/100–64/100 = 0.12. Pearson's chi-square test gives a two-sided p-value of 0.064, meaning that the difference is not statistically significant at a significance level of 0.05. However, if the alternative hypothesis were one-sided, the p-value would be half of this, i.e. 0.032. In general, a two-sided p-value is equal to twice the corresponding one-sided p-value.

Around the 1990 s, there was some debate on the choice of one-sided versus two-sided tests in medical statistics (1, 2). However, one issue has always remained beyond dispute: the choice of a one-sided or two-sided hypothesis test must be made in advance. This rule seems to have been frequently disregarded. In his textbook from 1991, Altman wrote: 'The small number of one-sided tests that I have seen reported in published papers have usually yielded P values between 0.025 and 0.05, so that the result would have been non-significant with a two-sided test. I doubt that most of these were pre-planned one-sided tests' ((3), p. 171)

Effects in both directions are possible

One could argue in favour of using a one-sided test only if an effect in the opposite direction is impossible or of no interest. However, we can rarely be certain that an effect in the opposite direction is impossible. If so, we would not need to conduct the trial (3), p. 171). There is, however, a type of trial in which an effect in the opposite direction is of no interest, namely a non-inferiority trial. The new treatment will be adopted if it is at least as effective as the standard treatment. It does not need to be better, and the *relevant* hypothesis is one-sided ((4).

Could a one-sided test be used in a superiority trial if this is decided in advance? That would be problematic if the effect turned out to go in the opposite direction, meaning that the new treatment had a poorer effect than the standard treatment. In this case, it would have to be attributed to chance, irrespective of how great the difference was. One-sided tests have greater

statistical power in one direction, but exclude the possibility of claiming any effects in the opposite direction. This is illustrated in Figure 1. This and other arguments in favour of using two-sided tests are elaborated in (1).

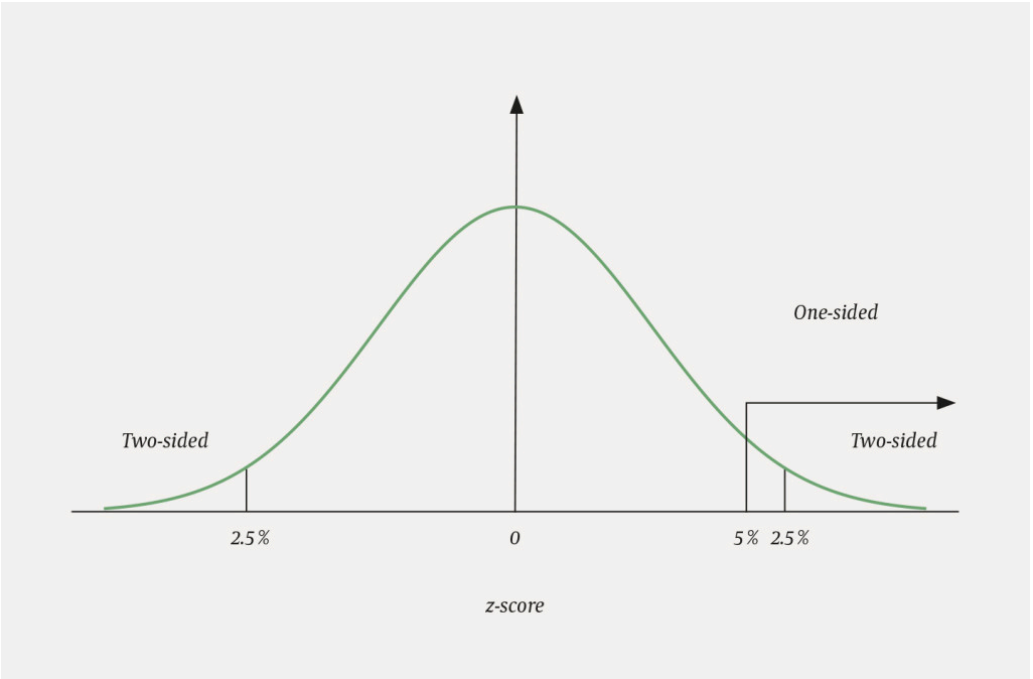

Figure 1 One-sided or two-sided test in a superiority trial

Consensus on two-sided tests

Today, there seems to be a consensus on using two-sided tests in medical research. This applies to intervention studies as well as observational studies. The only important exception is non-inferiority trials, where it is appropriate to use one-sided tests.

LITERATURE

1. Moyé LA, Tita ATN. Defending the rationale for the two-tailed test in clinical research. Circulation 2002; 105: 3062–5. [PubMed][CrossRef]

2. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309: 248. [PubMed][CrossRef]

3. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.

4. Skovlund E. Hvordan vise likhet? Tidsskr Nor Legeforen 2017; 137. doi: 10.4045/tidsskr.17.0668. [PubMed][CrossRef]

Publisert: 7 June 2021. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.21.0111