Note on the choice between one-tailed and two-tailed significance tests

Chris Phillips
1 February 2023

1. Introduction

This note addresses the question of whether to use a one-tailed or two-tailed test in the statistical significance testing of experimental data. The purpose of the test is to distinguish between a null hypothesis and an alternative hypothesis. Two situations are considered, corresponding to two classes of alternative hypotheses: (1) hypotheses in which there is a wholly directional effect, and (2) hypotheses in which there is an effect that may act in either the positive or the negative direction (though not necessarily with the same probability).

For simplicity, it will be assumed that on the null hypothesis the result of the experiment can be expressed as a single normally distributed statistic. Without loss of generality, this distribution can be assumed to be standard normal, with a mean of 0 and a variance of 1 respectively. (Note that this will describe approximately any experiment in which a large number of independent sample measurements are taken, and the individual measurements are then added together to produce a single statistic.)

For the first class of alternative hypotheses, it will be assumed that the distribution of the statistic remains normal, but that the mean is shifted by a specified amount, while the variance remains unaltered. The size of the shift can be thought of as an overall effect size for the experiment. If the statistic is the sum of N sample measurements, and the effect size for each individual sample is ϵ, then the overall effect size for the experiment is $E = N^{1/2}\epsilon$.

In the second class of alternative hypotheses, the shift in the mean will be either positive, with a specified probability, or else negative.

In each case, a hypothesis test may be constructed in the usual way, by specifying a range (or combination of ranges) of target values for the statistic. If the observed value of the statistic lies within the target range(s), the result will be considered positive, meaning that it favours the alternative hypothesis. Otherwise it will be negative, favouring the null hypothesis.

In such a test, there are two possible types of error. The test may give a positive result even though the null hypothesis is true - that is, a false positive result (known as a Type I error). Or it may give a negative result even though the alternative hypothesis is true - a false negative result (or Type II error). Given the null hypothesis (here a standard normal distribution), the probability of a false positive result can be calculated. This is known as the significance level of the test, and is usually denoted by α. Similarly, given the alternative hypothesis, the probability of a false negative result can be calculated. This is usually denoted by β. It is simply related to the statistical power of the test, defined as the probability of a positive result given that the alternative hypothesis is true, and is equal to $1 - \beta$.

2. Which test should be used when the alternative hypothesis is wholly directional?

Clearly, in the terms outlined above, it is desirable that the error rates α and β should be small, or equivalently that α should be small and that the statistical power should be close to 1.

The form of the test that maximises the statistical power for a specified value of α is given by a well known mathematical result known as the Neyman-Pearson lemma. It can be expressed in terms of the likelihood ratio - that is, the ratio of the probability of the observed result given that the alternative hypothesis is true, to the probability of the observed result given that the null hypothesis is true. The most powerful test is equivalent to applying a simple threshold to the likelihood ratio. If the ratio is above a specified threshold, the result of the test is positive, and otherwise it is negative. (The value of the threshold will depend on the value specified initially for α .)

It is easy to show that for the first class of alternative hypotheses considered here, for which there is a specified positive effect size, the most powerful test based on the likelihood ratio is always equivalent to a simple one-tailed test applied to the experimentally observed statistic.[1] This is illustrated numerically by the curves shown in Figure 1. These show the variation of statistical power as a function of effect size (for the experiment as a whole, in the sense explained above), for both the one-tailed and the two-tailed test, when $\alpha = 0.05$.

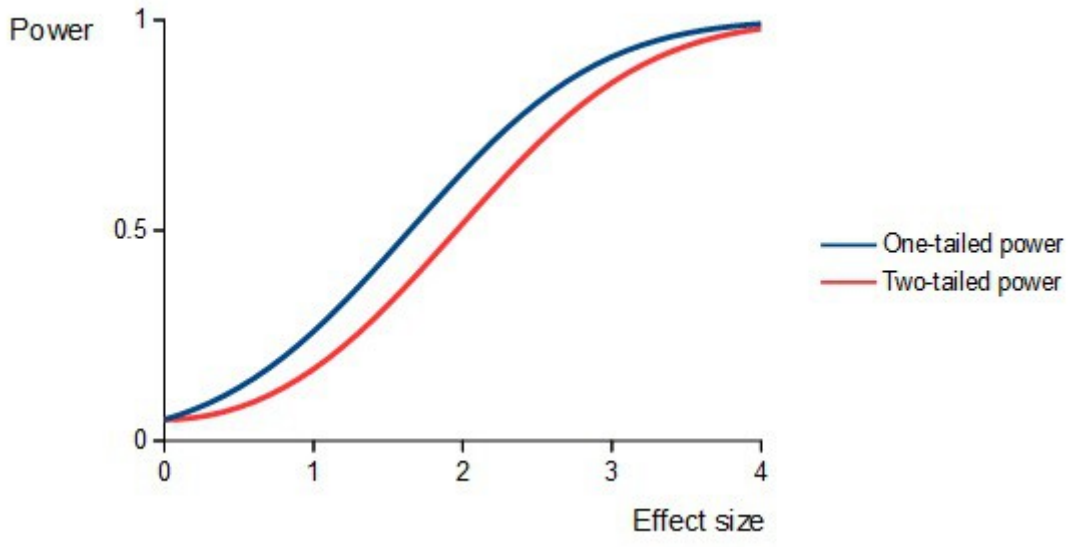

FIGURE 1. When the alternative hypothesis is that there is an effect which acts only in the positive direction, the plot shows the statistical power as a function of the effect size (for the experiment as a whole), for both one-tailed and two-tailed tests, with a significance level of 0.05.

3. Why do people use two-tailed tests for wholly directional hypotheses?

If the hypothesis really is wholly directional, it is very difficult to understand why anyone should use a two-tailed test rather than a one-tailed test, given its inferior statistical power. But below are some arguments that might be advanced in favour of two-tailed tests.

(1) *The null hypothesis has no preferred direction, so the test should have no preferred direction.* This argument really depends on a purely 'Fisherian' perspective, in which only a null hypothesis is

1 More generally, it is straightforward to show that the same conclusion follows if the alternative hypothesis is not a single specified shift in the mean, but a range of possible shifts with arbitrary probabilities, provided that all the possible shifts are positive.

considered, and no alternative hypothesis. In contrast, the situation dealt with here is explicitly one in which there is a (wholly directional) alternative hypothesis, and that is what produces the directional nature of the test.

But beyond that it should be noted that the null hypothesis in isolation is not capable of saying anything at all about which test is the *best* one to use. Any number of different statistical tests could be applied to investigate the consistency of experimental results with the null hypothesis. Without the introduction of an alternative hypothesis, there is no basis on which to choose between them. To give a concrete example, consider an alternative hypothesis in which - instead of the mean of the standard normal distribution being shifted while its variance is left unchanged - the variance is reduced and the mean is left unchanged. In that case, neither the one-tailed nor the two-tailed test makes any sense, because both are more likely to produce a positive result under the null hypothesis than under the alternative hypothesis.

(2) *The availability of both one-tailed and two-tailed tests offers a "degree of freedom" to the experimenter that may be abused.* In other words, an unscrupulous experimenter eager to publish a positive result might look at the results before deciding which test to use. That would raise the probability of a false positive result from the nominal value of 0.05 to a real value of 0.075 (namely, the 0.05 of the one-tailed test, plus an extra 0.025 from the lower tail of the two-tailed test).

This is a valid argument in favour of pre-registration of experimental studies, and particularly specification of the statistical analysis techniques to be used, but no argument at all for discarding a more powerful test just because in the absence of pre-registration it could be abused. If a study is intended to be evidential rather than exploratory, there is no excuse these days for not pre-registering it, and if a study is pre-registered this argument becomes irrelevant.

(3) *The two-tailed test is "more conservative" than the one-tailed test.* This is quite often claimed, and may initially sound plausible. But what does "more conservative" mean here? The only sensible way of comparing one-tailed and two-tailed tests is first to fix a common significance level. But once the significance level is fixed, the false positive rate is identical for both the tests. In that sense, neither is "more conservative" than the other.

What really seems to be meant here by "more conservative" is that *even when the alternative hypothesis is true* the two-tailed test is less likely to give a positive result than the one-tailed test. Put in those terms, it is difficult to imagine anyone arguing that this is a desirable feature.

How conservative a test should be is a matter of choice. But the sensible way of making a test more conservative is simply to reduce the significance level - for example from $\alpha = 0.05$ to 0.01.[2] That will reduce the probability of a false positive result. But having fixed the value of α, it makes no sense to use a less powerful test when a more powerful one is available. All that does is to raise unnecessarily the probability of a false negative result.

4. What about hypotheses that are not wholly directional?

Of course, the situation is quite different if - rather than the effect being wholly directional - the effect may act with equal probability in either the positive or the negative direction. In that case the most powerful test based on the likelihood ratio is equivalent to a symmetrical two-tailed test applied to the experimentally observed statistic.[3] This is illustrated numerically by the curves shown

2 The significance levels used in hypothesis tests vary enormously between different areas of science. In particle physics, a commonly used criterion is 5 standard deviations from the mean, which is equivalent to $\alpha = 0.0000003$.

3 More generally, the same conclusion follows if the alternative hypothesis is not a single specified shift, acting with

in Figure 2. These show the variation of statistical power as a function of effect size (for the experiment as a whole, as above), for both the one-tailed and the two-tailed test, when $\alpha = 0.05$. In this case, as the effect sizes rises beyond about 1 the two-tailed test becomes much more powerful. When the effect size is large, its power is about twice that of the one-tailed test, because the one-tailed test will nearly always fail when the effect is in the negative direction, and that will happen about half the time.

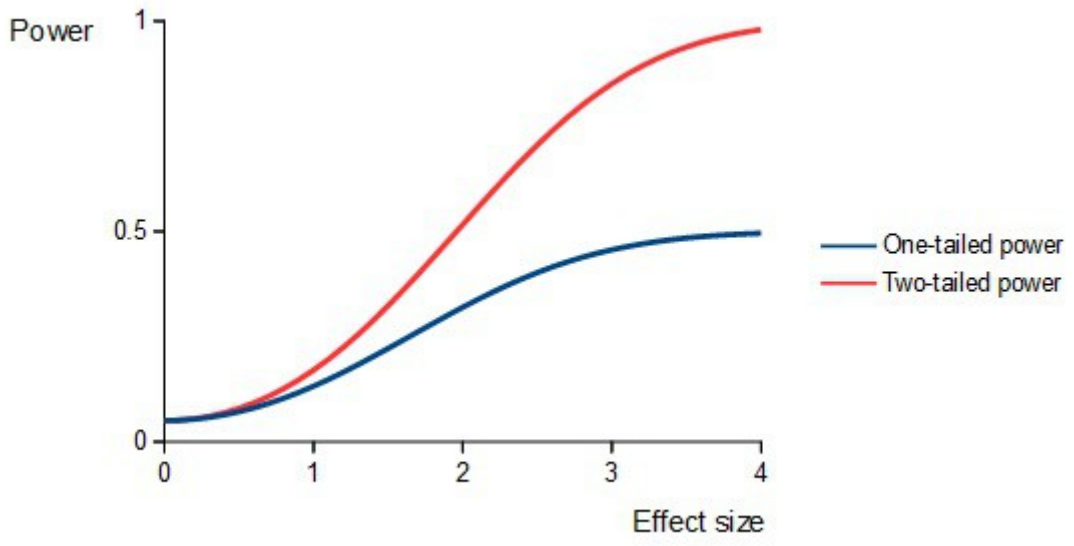

FIGURE 2. When the alternative hypothesis is that there is an effect which acts in the positive or negative direction with equal probability, the plot shows the statistical power as a function of the effect size (for the experiment as a whole), for both one-tailed and two-tailed tests, with a significance level of 0.05.

A more interesting question arises when the effect does have a preferred direction, but its direction is reversed for some of the time. For example, in parapsychology it has been suggested that in some cases there may be an effect in the direction opposite to that expected ("psi missing"). In the simple model considered here, suppose that the effect is positive with probability P, and negative with probability $1-P$. As shown above, when $P=1$ the effect is wholly directional and the one-tailed test is more powerful, but when $P=0.5$ there is no preferred direction and the symmetrical two-tailed test is more powerful.[4] As the value of P increases from 0.5 to 1, at some point the one-tailed test will become more powerful than the two-tailed test. The crossover value of P, for which the two tests are equally powerful, will depend on the effect size (for the experiment as a whole, as above). This dependence is shown in Figure 3, when $\alpha=0.05$.

equal probability in either the positive or the negative direction, but a range of possible positive and negative shifts with arbitrary probabilities, provided the probability distribution is symmetrical between the positive and negative directions.

4 When P is strictly between 0.5 and 1, the most powerful test based on the likelihood ratio is equivalent to an asymmetrical two-tailed test applied to the experimentally observed statistic. In general, the locations of the tails will depend on the effect size and on P. But in practical terms it seems preferable to use either a one-tailed or a symmetrical two-tailed test, avoiding this dependence on the details of the alternative hypothesis

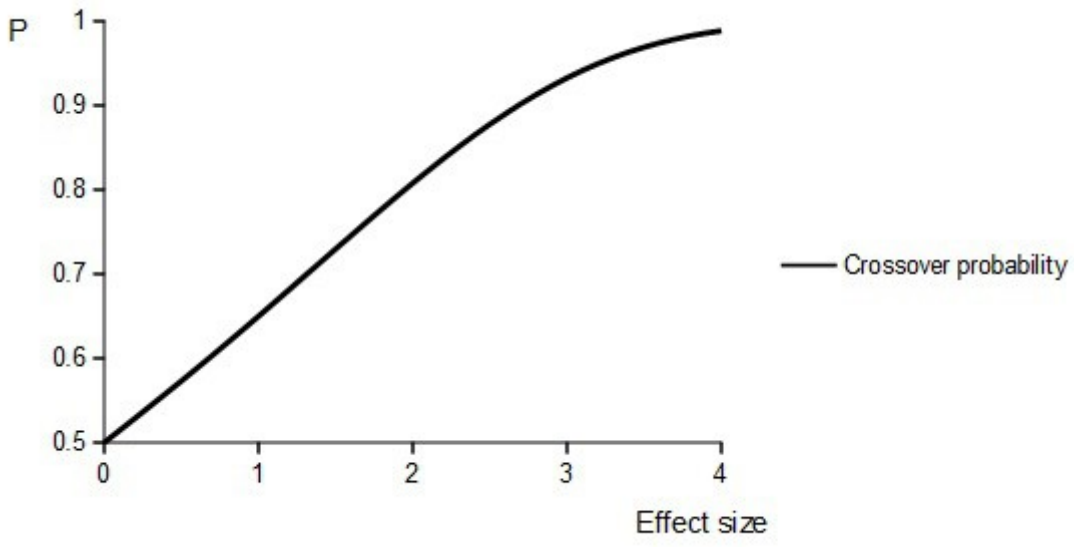

FIGURE 3. When the alternative hypothesis is that there is an effect which acts in the positive direction with probability *P* or the negative direction with probability $1-P$, the plot shows the value of *P* for which the one-tailed and two-tailed tests have equal statistical power, as a function of the effect size (for the experiment as a whole) when the significance level is 0.05. For values of *P* above the curve, the one-tailed test is more powerful, and for values of *P* below the curve, the two-tailed test is more powerful.

In this case, for values of the effect size above about 3 - which correspond to the experiment having what is often regarded as adequate statistical power, of around 0.8 or more - the one-tailed test is preferable only when *P* rises above 0.9 or so. This means that in a well powered experiment, even a relatively small probability that the direction of the effect will be reversed (for example, a small probability of psi missing) will render the two-tailed test more powerful than the one-tailed test. (Moreover, Figure 1 shows that in such a well powered experiment, even when the effect is wholly directional, the power of the two-tailed test is not much less than that of the one-tailed test.)

5. Conclusions

(1) When the alternative hypothesis is wholly directional, the one-tailed test is preferable, because for a given significance level it offers higher statistical power (Figure 1).

(2) When there is no preferred direction, the symmetrical two-tailed test is preferable (Figure 2).

(3) When there is a preferred direction but the direction is sometimes reversed, which test is preferable will depend on the effect size. But for a well powered experiment, the two-tailed test will be preferable even when the probability of reversal is relatively small (Figure 3).