A common misconception about p-values and its consequences

Jonas Ranstam

Department of Oncology, Lund University, Lund, and the NEPI foundation, Malmö. Correspondence: Jonas Ranstam, Box 746, S-220 07 Lund, Sweden. Tel +46 46-159060. Fax -146868. e-mail ranstam@uni-x.se
Submitted 96-05-27. Accepted 96-07-30

When reading medical journals it soon becomes evident that many investigators misinterpret p-values (Bailar 1986, Oakes 1986, Goodman and Berlin 1994). A common misconception is that an effect exists only if it is statistically significant and that it does not exist if it is not. Investigators with this attitude often screen their datasets with statistical tests to detect 'existing' effects in their dataset. However, such procedures are detrimental to the scientific inference and have consequences which are not usually envisaged.

Statistical tests

It is important to realize that medical research primarily is based on minute samples of patients or animals, hardly ever on an entire population. Therefore one of the main tasks for an investigator is to evaluate the effect that random (sampling) variation has on the experiment. One way of achieving this is by performing a statistical test.

Although it is not always appreciated, a significance test is always based on a study hypothesis and on a negation of this, the null hypothesis. When we compare 2 treatments A and B of a disease, our study hypothesis is that A and B are differently successful, the null hypothesis is that A and B are equally good.

The study hypothesis might seem irrelevant once a null hypothesis has been defined, but it is not. If we investigate a risk factor for a disease, one study hypothesis can be that the risk factor is heterogeneously related to risk, e.g., like beverage consumption (milk, coffee and beer) might be related to risk of a hip fracture. Another possible study hypothesis is that the risk factor is ordinally related to risk, e.g., like dementia (mild, moderate and severe) is related to the risk of hip fracture. The null hypothesis is in both cases the same; the factor is not related to risk. The 2 study hypotheses, however, lead to different tests, the first one to a test for homogeneity and the second one to a test for trend.

To test for a difference in success rate between the 2 treatments A and B, the null hypothesis is assumed to be true. The probability of observing something at least as extreme as the data actually observed is then calculated, this is the p-value. If A and B do differ, the p-value should be small and lead to a rejection of the null hypothesis. The p-value is usually defined as the "long-run relative frequency with which data '*at least* this extreme' would be obtained, given that the null hypothesis is true (Oakes 1986)".

2 errors can be made during the statistical evaluation of a null hypothesis. The first error (type I) occurs if a falsely positive conclusion is reached, i.e., if we reject a true null hypothesis. Given random sampling, the probability that this occurs is called the *significance level*. A significance level of 5% is traditionally used. The p-value concept is, of course, closely related to the significance level and p-values are also sometimes divided into the categories < 0.001, 0.001–0.01, 0.01–0.05, on the one hand, and "NS" (Not Significant), on the other. However, such categorization primarily leads to loss of information and is usually not recommended.

The second error (type II) occurs if a falsely negative conclusion is reached, i.e., if we accept a false null-hypothesis. The probability of avoiding this is known as *statistical power*. The magnitude of the statistical power is determined by the sample size, the variability of the data, the hypothesized size of the tested effect and the significance level. As a traditional rule of thumb, a study should be designed for a power of at least 80%.

Absence of statistical significance can be interpreted as a negative finding only in relation to the hypothesized effect and the statistical power to detect this; the lower the power, the smaller the chance of discovering a true effect, should it exist.

Multiple testing

The significance level and the power have different meanings when the hypothesis was formulated before and after inspection of the data. Statistical tests are designed for use with hypotheses that have been formulated before the inspection of data. Performing a statistical test on a characteristic because observations of data attracted the investigator's interest does not yield a correct p-value; it will look smaller than it should. To understand why, consider the differences in betting on a horse before and after the race.

Multiple testing of hypotheses is a problem related to this phenomenon. Since every single test has a type I error rate of 5%, the use of multiple testing will increase the rate of getting at least one type 1 error. If the tested items are independent, the simultaneous type I error rate will in fact be $1-(1-0.05)^m$ when testing m hypotheses. Even with only 10 tests the simultaneous type I error rate is about 0.40 or 8 times the nominal significance level. To avoid a large proportion of false positive findings, it can be reasonable to adjust the nominal level of significance or to use a statistical analysis technique that accounts for simultaneous inference. Both approaches are common; however, the problem is that they reduce the power of the study.

Why power is important

With the exception of reports on the results of randomized clinical trials, statistical power is not often presented in medical papers. Many clinical studies are also based on small samples and low power. Statistical power simply does not seem to be considered important. However, low power is related not only to studies with poor chances of achieving conclusive results but also to high rates of false findings, positive as well as negative.

The commonly used significance level of 5% does not guarantee that, at most, 5% of all statistically significant findings are erroneous; assume that 10,000 statistical tests are performed every year and that 75% of the null hypotheses tested are, indeed, true. With a significance level of 5%, we could expect $0.05 \times 0.75 \times 10{,}000 = 375$ significant but false findings (type I errors). If the power of the tests is 80%, the number of significant true findings (rejected false null hypotheses) will be $0.80 \times 0.25 \times 10{,}000 = 2{,}000$. The percentage of significant false findings will then be $375/(375 + 2{,}000) = 16\%$. If all these statistically significant findings had been published, we would have had 16% (not 5%) of published false findings.

Furthermore, this problem increases with decreasing statistical power. If the statistical power in the previous example had been lower, say 40%, the number of rejected false null hypotheses would have been $0.40 \times 0.25 \times 10{,}000 = 1{,}000$. The proportion of published false findings would then have been $375/(375+1{,}000) = 27\%$.

A new approach is necessary

It is common knowledge that statistically significant findings are easier to publish than results of negative or inconclusive studies. In meta-analysis terms this phenomenon is known as publication bias. If the primary goal is to publish, it might be rational to allocate resources to as great a number of studies as possible and ignore problems with insufficient power. If this really is the case, there are also few arguments for improving the statistical proficiency among investigators to avoid post hoc and multiple testing, and other procedures that result in p-values which overestimate the strength of the statistical significance.

However, we would all benefit from better research and from fewer published false findings: Less financial resources would go astray and fewer patients would be given useless treatments. That the present practice, in fact, is a real problem and not merely a subject for statistical fundamentalists, is clearly shown by the recent discussion about magnesium use after myocardial infarction (Smith and Egger 1995): Several reports have indicated beneficial effects from magnesium. It was even suggested (Yusuf et al. 1993) that "Magnesium treatment represents an effective, safe, simple and inexpensive intervention that should be introduced into clinical practice without further delay." However, all studies indicating an effect of magnesium had low power and no larger studies indicated any effect. Finally, a huge trial ISIS 4 (4th International Study of Infarct Survival) conclusively found that the use of magnesium had no effect.

To aim at lower error rates in published papers, editorial offices and the investigators themselves should place special emphasis on statistical power. However, this is not sufficient: Today, much of the medical literature pays little attention to type II errors. A greater use of estimates and confidence intervals has been proposed as a better alternative than statistical tests (Gardner and Altman 1986). When using confidence intervals, clinical rather than statistical significance is emphasized. Moreover, confidence intervals, by their width, disclose the statistical precision of the results.

The use of confidence intervals is not yet as common as p-values, but once this becomes a standard

procedure considerations about clinically relevant effect size, sample size, and statistical power will inevitably play a greater part in the design and presentation of studies.

References

Bailar J C. Science, statistics and deception. Ann Intern Med 1986; 104: 259-60.

Gardner M J, Altman D G. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746-50.

Goodman S, Berlin J. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994; 121: 200-6.

Oakes M. Statistical inference: A commentary for the social and behavioural sciences. Wiley, Chichester UK, 1986.

Smith D, Egger M. Magnesium use after myocardial infarction. Br Med J 1995; 310: 752-4.

Yusuf S. Koon T, Woods K. Intravenous magnesium in acute myocardial infarction. Circulation 1993; 87: 2043-6.