Statistical Significance

Many scientific papers describe in their results sections for each investigated variable whether it is significant or not significant. The significance of a significant finding is either presented with a specified $P$ value or just as $P<0.05$. Insignificant findings are often denoted n.s. and interpreted as evidence of absence of effect or difference.

This conception of statistical significance represents misleading over-simplification. The purpose of this note is to explain statistical significance and recommend a better practice. Confidence intervals as an alternative to $P$ values will be discussed later.

Interpretation of significance

First, the word significance can be replaced by important. The two words are synonymous. A finding in a study can either be clinically important or not. For example, assume that a certain drug has two side effects: 1) increasing systolic blood pressure by 25–50 mmHg and 2) reducing the growth rate of body hair by 5%. The first effect is clinically important, the second one clinically unimportant. Statistical tests can never be used to assess clinical importance. Whether a finding is clinically important or not is a matter of clinical knowledge and judgment.

In contrast, statistical importance has no relevance for the subjects included in a study. Statistical importance has no relation to clinical importance. A statistically important finding is simply reliable, and cannot be explained by random sampling.

Factors determining statistical significance

Whether a finding is statistically important or not can be calculated using a hypothesis test. Commercial statistical packages contain several tests suitable for testing statistical importance with different variable types and study designs. They all result in $P$ values describing the probability that chance alone caused the tested observation. A statistically significant test will give a $P$ value lower than the predefined significance level (usually 5%).

Three factors determine statistical significance: clinical effect ($E$), normal variability in clinical effect ($V$), and the number of studied subjects ($N$). The relationship can be schematically described:

$$\text{statistical significance} = \frac{E}{(V/\sqrt{N})}$$

Greater clinical effect, greater number of observations, and lower variability in clinical effect increase statistical significance, and vice versa.

However, it should be noted that $P$ values describe the risk of a false-positive finding (or type-1 error) only. The risk of a false-negative finding (or type-2 error) is described by statistical power, not statistical significance. An insignificant finding is thus not evidence of absence, just absence of evidence (1).

One- and two-sided tests

The null hypothesis, that a finding is caused by chance alone, corresponds to a two-tailed test. The alternative hypothesis, then, is that the findings have been caused by a systematic mechanism working in either of two directions: increasing or decreasing a treatment effect or group difference.

When the anticipated outcome of a test is in one direction, it may seem reasonable to include the other direction in the null hypothesis, e.g., when expecting that an investigated drug has only a positive effect, the null hypothesis could be defined: the treatment effect is either negative or caused by chance alone. This null hypothesis would lead to a one-tailed test and a $P$ value of half the size of that with a two-tailed test.

The use of one-tailed tests has been debated in the medical and statistical communities for many years and is still controversial (2). It is, however, clear that a one-tailed test requires the direction of outcome to be specified before data are inspected, otherwise the $P$ value will be corrupted. It is also clear that one-tailed testing can lead to the unusual situation in which a $P$ value that is clearly statistically significant with a two-tailed test (say, $P<0.001$) has to be presented as statistically insignificant with a one-tailed test.

International guidelines for clinical trials (3) recommend that the significance level with one-sided tests should be set at half the conventional significance level used in two-sided tests. This makes the choice between one- and two-sided tests
theoretical and independent of effects on statistical significance.

Multiplicity issues

Two kinds of statistical analysis can be made: exploratory and confirmatory. In exploratory analysis, the aim is to understand data by developing hypotheses about cause and effect. Hypothesis tests may be part of the toolbox for this work. However, hypotheses cannot be generated and tested by the same hypothesis test. Exploratory-generated hypotheses thus need confirmation in new, confirmatory analyses. These are performed in studies with both hypotheses and statistical precision defined in detail prior to the analysis of data.

One problem in confirmatory analysis is that multiple testing creates a discrepancy between the nominal and actual significance level. Two independent tests at the 5% significance level have a combined chance of 9.8% that either one or both will yield a false-positive result. With three tests, the probability of at least one false-positive outcome is 14.3%, with four tests 18.6%, and so on.

To perform a confirmatory analysis with defined statistical precision, multiplicity issues need to be addressed. The best way to do this is to design the study and statistical analysis in such a way that multiple testing is avoided. For example, one primary endpoint can be defined in advance and be confirmatory tested, while other, secondary study endpoints may be exploratory tested.

If multiple testing of several primary endpoints is performed, multiplicity adjustment of P values may be required. One such adjustment technique is known as the Bonferroni method and consists of using a modified significance level by dividing the original level by the number of tests performed. The drawback with multiplicity adjustments is that they reduce statistical power and require greater sample size.

However, not all multiple tests require P-value adjustment. One such situation is when two or more variables are needed to describe relevant effects and statistical significance is needed for all variables. Another situation is when different analysis sets are tested with the purpose of increasing confidence in results obtained from a primary analysis.

Recommendations

When using the word significance, specify whether you are referring to clinical or statistical significance, and avoid the expression "statistical difference," which has no generally accepted definition. Do not use P<0.05 or n.s. as evidence of absence. Avoid dichotomizing results as either statistically significant or not; present all P values numerically. Avoid using one-sided tests unless a one-sided alternative hypothesis has been formally presented in advance (say, in an approved study protocol or ethics committee application) or unless the significance level is halved. When the purpose of a study is confirmation of a predefined hypothesis, specify in detail which part of the study is confirmative and address multiplicity issues in this.

Jonas Ranstam

References

1. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
2. Moyé LA, Tita ATN. Defending the rationale for the two-tailed test in clinical research. Circulation 2002;105:3062.
3. Consensus guideline. ICH topic E9: Statistical principles for clinical trials. CPMP/ICH/363/96. London: CPMP/ICH; 1998.