Subjective attacks on statistical significance

The scientific method depends on the thoughtful use of validated statistical methods.

MICHAEL SWIFT

All too often, statements such as " . . . despite the claims of highly significant differences between the study populations, . . . the numbers seem quite small to support such a claim"* or "the number is too few from which to make a statistically relevant statement" are used to disparage scientific findings, even though the observed difference(s) is (are) statistically significant.

Contemporary biomedical science has relied on the statistical methods developed by "Student,"[1] Sir Ronald Fisher[2], and numerous others[3,4]. One tests a quantitative finding by comparing it with the value expected under the null hypothesis. If the P value is less than 0.05 the observed finding would have occurred by chance less than one time in twenty if the null hypothesis were true[1–4]. Although the sample size enters into the calculation of a P value, this sample size is irrelevant for interpretation of the P value once the calculation is done correctly. Thus, a given P value means exactly the same thing if derived from a study with ten individuals or ten thousand[1–4].

Methods for calculating P values have been developed for a wide variety of experimental designs, and statistical research has established guidelines for the use of each method[3,4]. Since each scientist calculating a P value for a specific data set will obtain the same value under the same assumptions, statistical methods are completely objective.

The term 'small', on the other hand, is wholly subjective in this context. Its use is conveniently flexible. One can object to the size of the numerators in comparing two proportions, or to the size of the entire data set. 'Small' may be used according to one of its dictionary definitions — 'insignificant' — unfortunately confusing the technical meaning of 'significant' with its meaning in everyday usage. Alternatively, someone who criticizes a statistically significant finding as small may be confusing tests of significance with statistical power — the probability that a study will detect a difference of a certain magnitude — which does increase for larger samples. Whatever the justification, 'small' often indicates refusal to accept a statistically significant result.

The adequacy of small samples for statistical tests of significance was shown in the very first published example, 'Student's' t-test[1], in which there were ten observations. The fact that sample size is irrelevant is shown by a hypothetical example in which a new and old drug are compared in the treatment of a serious illness. If the old drug is given to three patients and they die, while the three patients given the new drug survive, the P value equals 0.05 (two-tailed; Fisher's exact test) and 0.025 (one-tailed). In deciding whether to give the new or old drug to a patient with this disease, one would hardly dismiss these findings even though the number of subjects (six) is subjectively 'small'.

Another subjective approach to an objective statistical test is to characterize the confidence limits of an estimate as 'wide' even though the observed difference from the null value is statistically significant. For example: "The author fails to mention the confidence limits for the rate ratios that he quotes. This is a serious omission as they must be very large. The number of cancers he quotes from . . . are very small (8 and 12)". In general, if the P value is close to 0.05 the confidence limits around an observed value will be close to the value that would have been obtained under the null hypothesis. No additional insight is gained by characterizing these limits as 'wide'. Confidence limits are most useful when the precision of an estimate is more important than testing a hypothesis.

Scientific findings can also be disparaged by showing that shifting a small number of outcomes from one category to another changes a finding and its level of statistical significance. For example: ". . . the number of individuals in this study are too few to make a convincing argument. Changes of one or two individuals could make drastic differences in whether this is a statistical finding or not." This comment offers no real insight, as changes in significance level *always* result from shifting outcomes in small studies. When an investigator shifts cases or outcomes in reporting data, it is termed 'scientific fraud'. Why then is this procedure acceptable when a critic does it? Hypotheses are tested by collecting and analysing data as carefully as possible, not inventing or shifting them.

Over-ruling an objective statistical finding through a subjective evaluation is a polemical device often used when the topic is controversial. For example, data showing that ataxia-telangiectasia heterozygotes might be sensitive to breast cancer induction by medical diagnostic X-rays[5] challenged the view that there is no detectable excess of breast cancer after medical diagnostic X-rays. This finding was attacked in a published letter to the editor[6] that said "problems with the study include small sample size . . . ." The author of the letter was subsequently quoted as critical about the 'small' sample size in a newspaper article about the research[7].

While little can be done about the use of subjective quasi-scientific statements in the popular media, remedies do exist within the scientific community. Quantitative findings should always be evaluated by explicit statistical tests. Journal editors can give new or controversial findings a fair hearing if they consistently reject non-scientific downgrading of statistical significance by referees or letter writers. Contemporary biomedical research requires objective statistical analyses.

*All quotations are taken from anonymous reviews for leading biomedical journals.

1. "Student" The probable error of a mean. *Biometrika vi.* 1–25 (1908).
2. Fisher, R.A. *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, 1925).
3. Armitage, P. *Statistical Methods in Medical Research* (Blackwell Scientific Publications, Oxford, 1971).
4. Fleiss, J.L. *Statistical Methods for Rates and Proportions* 2nd edn Wiley & Sons, New York, 1981).
5. Swift, M., Morrell, D., Massey, R.B. & Chase, C.L. Incidence of cancer in 161 families affected by ataxia-telangiectasia. *New Engl. J. Med.* **325**, 1831–1836 (1991).
6. Boice, J.D. & Miller, R.W. Risk of breast cancer in ataxia-telangiectasia (letter). *New Engl. J. Med.* **326**, 1357–1358 (1992).
7. Kolata, G. Study tying gene to cancer risk draws fire, *The New York Times*, December 27, 1991, p. A18.

*Institute for the Genetic
Analysis of Common Diseases
New York Medical College
Hawthorne, New York, 10532, USA*