Does significance matter?

The name of this magazine was the subject of fierce debate. Although "significance" is a word often associated with statistics, it is very often misapplied or misunderstood. Its technical implications differ from the everyday use of the word and competing schools of thought within statistics have varying views on the correct usage. As the magazine is launched, **R. Allan Reese** examines some of these ideas from the viewpoint of a pragmatic data analyst.

Significance *versus* importance

A quick definition of significance is that it is a measure (*P* for probability) of how likely an observed result, or something more extreme, was to have occurred, *on the basis of a set of assumptions.* Finding an unlikely result should lead us to question our assumptions, but what figure should be considered "significant", and what should we do with "non-significant" findings? Clarke and Cooke[1] give a clear explanation of the usual steps in constructing a hypothesis test, one step being to "fix a significance level, e.g. 0.05". That approach restricts the thinking of most users and is too theoretical to be useful in their applications. In particular, it creates in students the view that statistical analysis is about "looking for significance". Regardless of any warning from their teachers, they view lack of significance as indicating lack of importance—how often are textbook exercises based on finding non-significance?

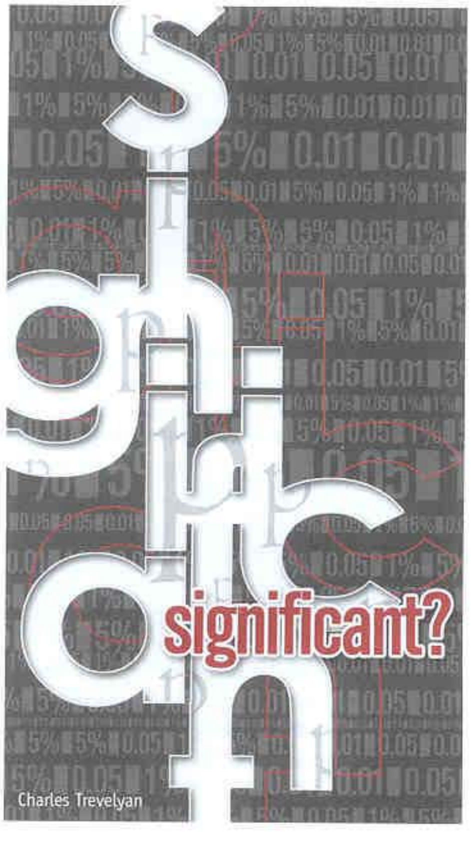
Equating significance with importance presumes that a "good" result is one that disproves a null hypothesis. Most statisticians now despise the practice—dating from before computers—of computing a test statistic, comparing its value with the tabulated critical values, and reporting the result only as "significant" or "not significant". Unfortunately, that mechanical approach is retained by editors of some academic journals. Coolidge[2] reports that the American Psychological Association publication manual still "claims that if results do not meet the 0.05 level of significance, then they are to be interpreted as chance findings". Such an attitude prompted this plaintive request on an e-mail list: "A two-tailed test has come out as significant at $P = 0.08$; can I halve this so it becomes significant at $P < .05$ as a one-tailed test?". (Reference withheld but it exists!)

The sacred 0.05?

Comparison with a critical value was an arbitrary, if reasonable, choice in the 1930s at a time when percentiles of test statistics had to be laboriously tabulated. Researchers could check only whether a test statistic for a particular experimental result was above or below the tabulated value. The background was agricultural science, where individual experiments can be large, long term and expensive. 1 in 20 is a nice round figure, though significance might have been more widely understood if R. A. Fisher had gone for the bookies' method of quoting odds and used 20:1 (rather than 19:1) as the yardstick. Asterisks in output also date from times when the most advanced office technology was a typewriter.

Significance as described above is based on a model of a single trial, with the test defined in advance of examining the data. Different adjustments for multiple tests are available, based on a variety of mathematical assumptions that need to be carefully examined. It is a common mistake to think that the assumptions underlying a calculation are somehow validated because the calculation seems to fit the data.

Coolidge hints at an alternative view: "Does God really love the 0.05 level of significance more than the 0.06 level?" and reports that "sometimes researchers will report 'trends' in their data". (The word "trend" is misleading here, as it suggests a series of values. Coolidge should have written "tendency".) Small scale projects that can be carried out by one researcher are likely to have low power (i.e. probability of detecting "true" differences), and it is not surprising that their results only *tend* towards significance. It is important for scientific advance that such "non-significant" results should be recorded to prompt further investigation, but the attitude of the American Psychological Association and similar academic groups militates against this.



Charles Trevelyan

"Use of a pass–fail critical value confuses the reporting and interpretation of data with the need to make a decision based on the findings"

My own view is that "significance" should be taught and understood in a wider context of data analysis as a process for continuous refinement of knowledge: what John Nelder calls "statistical science"[3]. David Moore writes[4]: "Some hesitation about the unthinking use of significance tests is a sign of statistical maturity." Use of a pass–fail critical value confuses the reporting and interpretation of data with the need to make a decision based on the findings.

Charles Trevelyan

Not significant: doesn't matter?

Many of the students that I work with carry out questionnaire surveys, and their survey forms contain blocks of questions relating to broad topics. Students are dismayed when they tabulate the responses for the study groups and find that no differences are significant, especially after adjusting for the multiple tests (e.g. Bonferroni corrections). My advice is that they should report the results as found, and cautiously interpret any patterns that might not have been anticipated. Questionnaires are not designed in ignorance; when several questions all show "non-significant" results that are nevertheless in the direction expected, it is not reasonable to write these off as zero effects.

Another pragmatic point when interpreting test statistics is to question both all the assumptions of the calculation, and the data themselves, rather than simply accepting or rejecting the "null hypothesis". Even routine data collection is subject to unanticipated changes, extreme values or simple errors. Checking is not just about avoiding errors; spotting an exception may be like Fleming's discovery of penicillin.

Most of us now obtain *P*-values directly from software as part of a packaged analysis, and rarely refer to printed tables. Software often prints *P*-values with many digits of precision, but the interpretation generally depends only on the order of magnitude. An answer that is presented as a specific figure always runs the risk of being considered more precise than was intended. Another consequence of having computers on tap is that our datasets are typically larger than those used by the pioneers who relied on tables. Reporting results based on several hundred observations and significance, say 0.000001, as "$P < 0.05$" is disingenuous or naive. What matters is the intelligent interpretation of the results, putting them in the context of the questions that prompted the research.

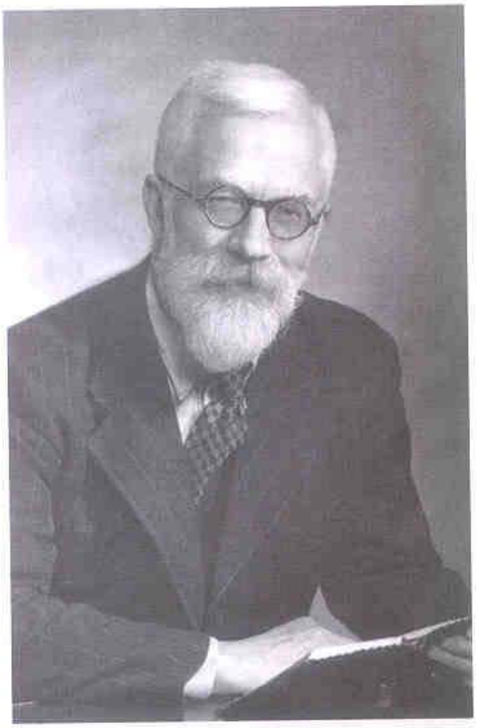
For example, data are often collected at the start of a study in order to confirm the equivalence of study groups. Researchers ask whether they need to report or even to carry out comparisons on the baseline values which are almost certain to be non-significant. Certainly they should carry out the tests and report very briefly that they have. The outcome of the baseline tests affects the subsequent analysis.

Knowledge worth having

Significance tests, as described in textbooks, refer to a default assumption (the "null hypothesis") to be tested against a sample of data from a population. It is assumed that we can describe the population and the process of selection, and that the null hypothesis is a useful comparison. But real life is never as simple as a textbook example. The choice of what to measure and how to select the sample is based on existing knowledge. Far from being an unprejudiced enquiry testing for an unknown parameter, data collection is usually a conscious quest for a particular result. That is not necessarily a criticism. Without a theoretical basis or strong methodology, blind searching is unlikely to uncover meaningful patterns. But it does require us to look at the assumptions when applying a model derived from blind sampling of balls from an urn.

One consequence is that you should evaluate any significance level in relation to the prior expectation. The Bayesian approach formalises this, but requires further calculations and, in my opinion, could alienate researchers who fear the mathematics of statistics. The intelligence researchers should bring to interpretation is to consider how unlikely they truly find the test result. In other words, have they really found results that surprise? For example, within most organisations salaries for males are on average higher than those for females, so any study that reports this result as "significant" could be criticised as stating common knowledge. On the other hand, a powerful study that adjusted for various factors and found no significant difference, and therefore no evidence of gender discrimination, would be of great interest.

Calculating statistical significance is a tool, a step in the process of analysis. The interpretation of a result requires the researcher's knowledge, in particular to put the new data in the context of previous scientific knowledge—a quasi-Bayesian approach rather than by mathematical formulae. What is important, or "significant" in the English sense, may be statistically significant or non-significant.



The great R. A. Fisher wrote in 1926: "Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance." (Quoted in Moore, 1979 edition).

It is the fate of a guru that what he sees as a convenient but arbitrary option is taken by followers as written in stone. But it is a philosophy that must be abandoned.

Let us wish *Significance* the magazine a successful and influential future.

References

1. Clarke, G. M. and Cooke, D. (1983) *A Basic Course in Statistics*, 2nd edition. London: Arnold.

2. Coolidge, F. L. (2000) *Statistics: a Gentle Introduction*. London: Sage.

3. Nelder, J. A. (1999) From statistics to statistical science. *Statistician*, **48**, 257–269.

4. Moore, D. S. (1997) *Statistics: Concepts and Controversies*, 4th edition. New York: Freeman.

Allan Reese was employed for 25 years in universities, advising researchers on data handling and analysis. He recently quit the education sector to move into a research environment.