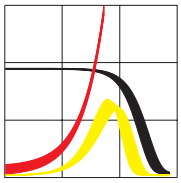

Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2007-037
DECEMBER 2007

The reporting of statistical significance in scientific journals

Jan M. Hoem (hoem@demogr.mpg.de)

JMH/-, 5 April 2001

The reporting of statistical significance in scientific journals

A reflexion by Jan M. Hoem

Scientific journals in most empirical disciplines have regulations about how authors should report the precision of their estimates of model parameters and other model elements. Some journals that overlap fully or partly with the field of demography demand as a strict prerequisite for publication that a *p*-value, a confidence interval, or a standard deviation accompany any parameter estimate.[1] I feel that this rule is sometimes applied in an overly mechanical manner. Standard deviations and *p*-values produced routinely by general-purpose software are taken at face value and included without questioning, and features that have too high a *p*-value or too large a standard deviation are too easily disregarded as being without interest because they appear not to be statistically significant. In my opinion authors should be discouraged from adhering to this practice, and flexibility rather than rigidity should be encouraged in the reporting of statistical significance. One should also encourage thoughtful rather than mechanical use of *p*-values, standard deviations, confidence intervals, and the like. Here is why:

1. The scientific importance of an empirical finding depends much more on its contribution to the development or falsification of a substantive theory than on the values of indicators of statistical significance. It is important that authors be guided by a process of discovery and not blinded by a lack of statistical significance in the description of an empirical pattern. This means that authors should feel free to report findings that appear not to be statistically significant, provided that this fact is also reported. Indicators of statistical significance should not be suppressed, but authors should avoid using them mechanically.

2. Measures of statistical significance may be misleading. When a model has been developed through repeated use of tests of significance to include and exclude covariates, to split or combine levels on categorical covariates, and to determine other model features, the user often loses control over statistical-significance values, and the values computed by standard software may be completely misleading. If one mechanically includes the *p*-values cranked out by standard software, this serves sooner to mislead than to inform.

3. Standard *p*-values can be insufficiently precise indicators of statistical significance, particularly if their values are given only in grouped levels, which are often indicated by asterisks beside parameter estimates ("* = $p<0.1$, ** = $p<0.05$, *** = $p<0.01$", and so on). Then standard deviations and confidence intervals are much more precise when used appropriately. Incidentally, I would discourage the practice of printing standard deviations just underneath estimated

[1] A diametrically opposite position is taken by *Epidemiology*, the Official Journal of the International Society for Environmental Epidemiology. In its Vol.9, No. 3, from May 1998, Kenneth J. Rothman set its policy as follows: "When writing for *Epidemiology*, you can … enhance your prospects if you omit tests of statistical significance. … every worthwhile journal will accept papers that omit them entirely. In Epidemiology, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis… . We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is significant, as if neither chance nor bias could then account for the findings."

parameters in tabular representations, since this can make it hard to see patterns in parameter estimates. If they are included at all, indicators of statistical significance should be presented in a manner that facilitates the interpretation of results, perhaps in separate table columns when appropriate. Significance asterisks are a poor substitute for this.

4. It may be more important for an understanding of demographic behavior or other phenomena studied to know whether the inclusion of a categorical covariate in its entirety contributes significantly to an improvement of the model than to know the significance indicators of each of its levels. Such issues are often checked by means of a test, for instance a likelihood-ratio test. This is where *p*-values have their primary justification as indicators of statistical significance. When used appropriately in such a context, accurate rather than grouped *p*-values should be included. The degree of significance can then be assessed by the reader. Authors should be aware of the possibility of accepting statistical significance at higher *p*-values for small data sets than for large data sets. In particular, there is nothing sacred about a *p*-value limit of 0.05. Much higher *p*-values indicate statistical significance in very small data sets, while for the enormous sets typical of register data for populations with millions of members, much smaller *p*-values than 0.05 may be needed to indicate important features in the data.[2]

5. Standard deviations, when used, should be reported for interesting contrasts, not for features selected automatically by statistical software. In many demographic applications, parameters are contrasts between regression coefficients for the various levels of a categorical regressor, often presented as relative risks in comparisons between the "effect" of one regressor level and a baseline level on the same regressor. Standard software routinely selects the first (or last) level on such a regressor as its baseline level, and parameters measure deviations in the "effect" of having a different level from the baseline on the regressor. Many other comparisons may be of greater substantive importance than the contrast with the mechanically chosen baseline level, and authors should adjust their parameter space accordingly.

[2] Rothman (*op. cit.*) wrote: "Many data analysts appear to remain oblivious to the qualitative nature of significance testing. Although calculations based on mountains of valuable quantitative information may go into it, statistical significance is itself only a dichotomous indicator. As it has only two values, 'significant' or 'not significant', it cannot convey much useful information. Even worse, those two values often signal just the wrong interpretation. These misleading signals occur when a trivial effect is found to be 'significant', as often happens in large studies, or when a strong relation is found 'nonsignificant', as often happens in small studies."