The (Practical) Importance of SE Experiments

Tore Dybå

Much experimental SE research involves testing a hypothesis regarding a relationship or difference between two variables. Typically, a null hypothesis (H_0) of a zero correlation or no difference between the means of the two populations is posited. The standard way of reporting results from such statistical hypothesis testing is by presenting *p*-values or information about the rejection or acceptance of H_0.

However, in an applied discipline such as SE, it is not enough to find out that one method or technique is better than another – we need evidence of *how much* better. Such evidence can be expressed in terms of an *effect size*. So, whereas *p*-values reveal whether a finding is *statistically* significant, effect size indices are measures of *practical* importance. Interpreting such effect sizes is critical, because it is possible for a finding to be statistically significant but not meaningful, and *vice versa*.

In an ongoing systematic review at the Simula Research Laboratory [2] we found that more than two thirds of SE experiments did not report any effect size measure. This lack of effect size reporting can lead to serious inferential problems and effectively reduces the practical utility of experimental results. For those experiments that reported effect sizes, or included enough descriptive statistics for effect size indices to be calculated, the median effect size was $d = 0.6$. At the same time a quantitative assessment of statistical power revealed that the experiments, on average, only had a one-thirds chance of detecting phenomena with such medium effect sizes [1]. These results indicate that SE experiments currently are both underpowered and underreported.

To further advance the field of empirical software engineering we must not only address relevant topics, we must also plan for acceptable power and report the results of our studies in a manner that could be put to use. What we regard as practically important effect sizes vary depending on the goal of the research and the fields within which its results are applied. However, in order to seriously discuss these issues and inform judgement about practical importance, effect size, or sufficient descriptive statistics for relevant effect size indices to be calculated, must be reported as part of our experimental results.

References

[1] Tore Dybå, Vigdis By Kampenes, and Dag Sjøberg. A Systematic Review of Statistical Power in Software Engineering Experiments. *Information and Software Technology*, vol. 48, no. 8, August 2006, pp. 745-755.

[2] Vigdis By Kampenes, Tore Dybå, Jo Hannay, and Dag Sjøberg. A Systematic Review of Effect Size in Software Engineering Experiments. Simula Research Laboratory, work in progress.

V. Basili et al. (Eds.): Empirical Software Engineering Issues, LNCS 4336, p. 129, 2007.