A Systematic review of statistical power in software engineering experiments

Tore Dyba, Vigdis By Kampenes,
Dag I.K. sjoberg

Presentation by Yi Li
• Statistical power

• There is inadequate attention to statistical power in the ESE literature.

• Purpose of this paper:
  – Review of the statistical power of ESE research
  – Discuss the implications
  – Techniques to increase the statistical power
• Type I and Type II errors

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ is true</th>
<th>$H_0$ is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept $H_0$</td>
<td>1-$\alpha$: Correct</td>
<td>$\beta$: Type II error</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>$\alpha$: Type I error</td>
<td>$1-\beta$: Correct (power)</td>
</tr>
</tbody>
</table>

• Traditionally, $\alpha$ is set to .05, $\beta$ is set to .20.

• The distribution of risk between Type I and Type II errors should be appropriate to the situation at hand.
Determinants of statistical power

- Sample size ($N$)
- Significance criterion ($\alpha$)
- Population effect size (ES)

Statistical power ($1-\beta$)
• Larger $\alpha$ -> greater power
• Directionality of the significance criterion affects the power of a statistical test.

• Larger $N$ -> greater power
• Effect Size:
  – Larger ES -> greater power
  
  – ES measures the true size of the difference between $H_0$ and $H_A$.

  – Three levels: small, medium and large.
• The power of experimental SE research falls markedly below the levels attained by IS research.

• The considerations of statistical power are underemphasized in ESE research.

• Low statistical power makes replication and meta-analysis troublesome.

• Researchers should report the effect sizes for readers to judge whether the statistically significant findings have practical importance.
• Ways to increase statistical power:
  – Increase the size of the sample
  – Relax the significance criterion
  – Choose powerful statistical tests
  – Reduce measurement error and subject heterogeneity
  – Balance groups
  – Investigate only relevant variables
Questions to discuss

• Is statistical power the only criterion of judging the quality of an experiment?
• What is the rationale of guarding four times more against Type I errors than Type II errors? Is the traditional value .05 appropriate in most of the studies?
• Will the relative seriousness of Type I and Type II errors become an excuse for data manipulation (misuse of statistics)?
• Is post-hoc power analysis useful? Is the result trustworthy?
• Is the Comparison with IS research valid? Considering the possibly different definition for effect sizes.
• Can “expert judgment” on behavioral research be applied to software engineering research?