A Large Scale Survey of Motivation in Software Development and Analysis of its Validity - Population Differences

Idan Amit idan.amit@mail.huji.ac.il
Dror G. Feitelson feit@cs.huji.ac.il
Department of Computer Science
The Hebrew University, Jerusalem 91904, Israel

June 3, 2024

1 Differences Between GitHub and Social Media Participants

We received answers from two different populations. 20% of the participants were developers contributing to a GitHub project, which were recruited via direct emails. The other 80% were reached out to in convenience sampling [2, 1], using messages in social media. Differences between the populations might lead to investigating two different behaviors as an averaged one. It is common to compare populations using a comparison of distributions of demographic variables (e.g., age, gender). However, the relevant questions appeared at the end of our survey, and many developers that did not contribute to open source did not reach this part in our first phase. Instead, we reduced the problem of populations difference into a supervised learning problem, trying to predict the source of the participants using the questions in the first part of the survey. The features included open-source specific questions such as 'I contribute to open source in order to become a better programmer' and 'I contribute to open source due to ideology'. Nevertheless, a decision tree model, suitable to a small number of samples, reached an accuracy of only 78%. Even high-capacity models such as SVM or Neural Networks reached an accuracy of only 80%. Note that since the positive rate is 20%, the majority rule betting that all the developers are from social media would also lead to an accuracy of 80%. Such low predictive power does not mean that the populations are similar. However, it means that there is no big obvious difference, even when considering contribution to open source, based on the personal questions. Therefore, we can analyze both populations together, getting a larger dataset and leading to more robust results.

Though, our response rate of GitHub developers was rather low, raising the threat of response bias [3, 4] and therefore our GitHub developer dataset might

not represent the GitHub developers population. Due to that we cannot claim that our entire sample represents their population too.

References

[1] Acharya, A. S., Prakash, A., Saxena, P., and Nigam, A. Sampling: Why and how of it. *Indian Journal of Medical Specialties 4*, 2 (2013), 330–333.

[2] Etikan, I., Musa, S. A., Alkassim, R. S., et al. Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics 5*, 1 (2016), 1–4.

[3] Gove, W. R., and Geerken, M. R. Response bias in surveys of mental health: An empirical investigation. *American journal of Sociology 82*, 6 (1977), 1289–1317.

[4] Mazor, K. M., Clauser, B. E., Field, T., Yood, R. A., and Gurwitz, J. H. A demonstration of the impact of response bias on the results of patient satisfaction surveys. *Health services research 37*, 5 (2002), 1403–1417.