Response Paper

Generalizability of heterogeneous treatment effect estimates across samples

Coppock et al. PNAS December, 2018 vol. 115 no. 4912441–12446

Summary

In this article, Coppock et al. investigate the phenomenon that survey results obtained from non-representative convenience samples are often generalizable to target populations, despite obvious differences between the groups.

The authors at first propose two possible explanations for this:
A) Effect homogeneity across participants, such that sample characteristics are irrelevant.
B) Effect heterogeneity that is approximately orthogonal to selection.

To test which of these two might explain the comparability of convenience samples to the target population, the authors analyzed subgroup conditional average treatment effects (CATEs) on 27 original surveys performed on target populations and their convenience replication, encompassing 101745 individual survey responses in total.

Surveys encompassed topics in political science, psychology, public health, communication, business, sociology, law, education, and public policy.

Surveys already performed on target populations were replicated on convenience samples with the MTurk service. MTurk provides nonprobability, fully opt-in samples of participants who complete online tasks for financial compensation. The MTurk service is described as a provider of high quality survey results.

The authors picked six attributes commonly measured in nearly all the 27 original studies. To measure these attributes in the same way as in the original study, the authors coarsened them, e.g. age (18 to 39, 40 to 59, 60+), partisanship (Democrat, Independent, Republican).

The authors compare the slope of the original survey CATE with respect to the replication CATE estimate, corrected for measurement error via a generalized Deming regression.

The CATE comparison shows that the overall “significance match” rate is around 70%. So, results of convenience replication studies correspond with the results of the original target population.

The authors explain this similar average treatment effect with a low treatment effect heterogeneity, sample characteristics seem not to affect the results, so initial explanation A) seems to be the case.

Despite some caveats in their study design and suggestions for improvements, the authors conclude that convenience samples can be indeed representative for a target population, despite differences in their characteristics, the determinant factor thereby is a low treatment effect heterogeneity, which makes them comparable.

Reflection and thoughts

The article begins with stating that surveys are often met with skepticism of how their results generalize, which I think is true.

Especially in politics, surveys on a tiny part of the population are regularly used to portray the current intentions of all voters. A new survey is always worth an article, often with a quite exaggerated headline. “Austrians want X”, or “Austrians don’t want Y”, although just a small fraction of the population was asked.

The US presidential election in 2016 is often mentioned to undermine the meaning of surveys. Despite Hillary Clinton leading in nearly any survey for months, she lost in the end.

Funnily enough, a good proportion of politicians only considered results of surveys relevant, if they are in their favor, if not, the survey is not trustworthy, not relevant or simple speculation.

What kind of challenges surveys are facing, I already learned in the Nases et al. paper.

Of course, I am aware, that this article is a scientific one published in a reviewed journal and is not really referring to, or analyzing data from the weekly Sunday political survey.

However, I think that there is a trend that surveys have lost in public perceived value in recent years, many people seem to be highly skeptical. There is a general awareness of how surveys can be twisted and tweaked to meet certain intentions, especially if they are asking for political views. Although also here might apply the same concept as for politicians, if the results fit your own worldview, they are more easily accepted, if not, they must be fishy.

Another layer is how survey results are portrayed, which can be independent of the intentions of the conductors and is not always under their control. Results might get cherry-picked, misrepresented, or verbally exaggerated.

The authors state, that they regret the choice to have coarsened certain features too much, such as white vs. non-white. In none of the papers I have reads so far, which were all biology themed, I have encountered the word regret. To admit a “mistake” that openly, is quite unusual for me to read. Maybe biologists never regret something.

The article is very open about their caveats and things they didn’t consider, which makes their discussion more trustworthy and valuable. Reading about problems and drawbacks is refreshing and adds to the understanding of current possibilities and limitations to overcome. In general, I think the scientific community should be more open about their shortcomings, so far this seems to me is more discussed in political science than in biology.
However, I can imagine that being too open about possible shortcomings might risk your chances for publications, so it is tied to a certain risk.

Given all the skepticism I mentioned before, the article does give the impression that in certain cases, convenience samples are indeed representative and can give clues about the overall target population.

Also, political surveys often follow the set of standards to achieve representative results and are performed professionally, still, their predictive power often lacks.

Interesting for me was also reading about the MTurk service, I have never heard before and didn’t expect that the scientific community would use such services, I wonder how much something like this costs.