InnoRenew CoE

Livade 6, 6310 Izola/Isola, Slovenia, T: +386 40 282 944, E: coe@innorenew.eu, www.innorenew.eu

Improving the representativeness of non-probability samples: A case study of two web surveys

Ana Slavec, PhD

ana.slavec@innorenew.eu

17th Applied Statistics International Conference

Online, 21. 9. 2021

Background

- Traditionally survey methods relied on probability sampling to obtain representative samples
- Web surveys often use non-probability samples (to save resources) and are prone to selection bias
- Even published research is often based on web surveys with questionable sampling approaches
- RQ: How to improve the representativeness of non-probability surveys?

Two case studies

1. Young doctors of Slovenia‘s **survey on compliance with COVID-19 protective measures** (November 2020) 😂
2. University of Maribor student‘s **survey on COVID-19 vaccination attitudes** (December 2020) 💉

MLADI ZDRAVNIKI SLOVENIJE

University of Maribor

COVID-19 SLEDILNIK

Sample

	Large convenience sample (snowball sampling, shared on social media by different organisations)	Smaller representative sample (JazVem online panel from marketing research company Valicon)*
😷 Case study 1: COVID-19 protection measures	5,165 responses	527 responses
💉 Case study 2: COVID-19 vaccination attitudes)	12,042 responses**	1,042 responses***

* It is the largest online panel in Slovenia and three quarters of panalists were included based on **probability sampling**. The representativness of the sample is controled with comparisons with offical data.

We have a list of contacted organisations and we estimated the potential reach to **more than 400,000 users (assuming a zero overlap in followers) which is about 48 % of all social media users according to the data of the Statistical office of Slovenia. However, organic posts usually do not reach all users and according to some estimates the average reach is only between 5-6% of the followers, i.e. between 20,000 and 24,000. However, since there were **45,633 unique clicks** on the survey link in the eleven days that the survey ran, the reach was probably higher. We estimated the response rate is between 3 % (assuming the survey link reached more than 400,000 people) and 26 % (assuming the reach was the same as the number of clicks).

*** 3,143 panelists were invited and 1,423 responded (45 % cooperation rate) but 381 were already vaccinated and were screened out.

Note: all four databases were weighted by gender, age, education and region

Case study 1: COVID-19 protection measures 

	Large convenience sample	Smaller representative sample
% that had two or more symptoms and/or were in a risky contact in the last 7 days	10% of respondents	21% of respondents
% that have been at workplace at least once in the last 7 days	57% of employed respondents	64% of employed respondents
% that visited at least one other household in the last 7 days	44% of respondents	47% of respondents
% wore masks correctly on visits in the last 7 days	59 % of visiting respondents	44 % of visiting respondents
Differences between regions	Sig. differences for almost all indicators (e.g. % that visited other household ranging from 37% to 58%)	Regions too small to calculate differences

Case study 2: COVID-19 vaccination attitudes 

	Large convenience sample	Smaller representative sample
% that will probably or definitely get vaccinated	59% of respondents	50% of respondents
% of medical professionals	17% of respondents	4% of respondents
Differences in vaccination intention between healthcare professionals	Significant differences in between professions (84% physicians and 82% medical students intend to get vaccinated compared to only 51% healthcare students and 50% other healtchare professions)	Not enough healthcare professionals included in sample

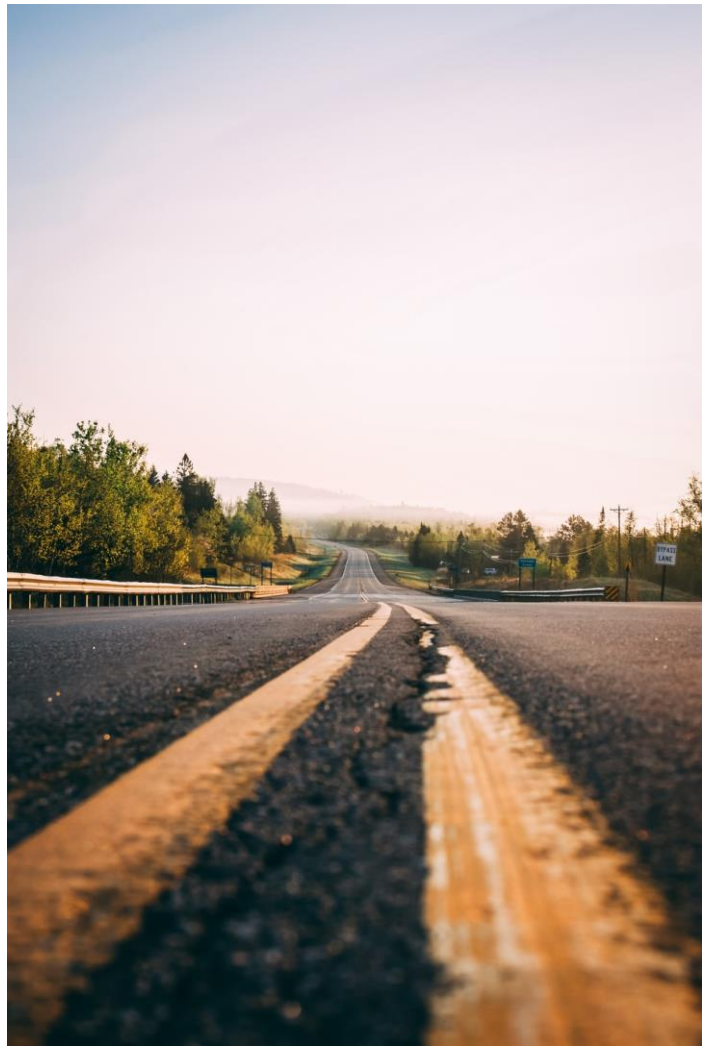
Discussion

- Convenience samples are not representative and do not allow making inferences to the population
- Using more representative samples (e.g. marketing panels) gives us less biased results but it is more expensive to conduct
- Using convenience samples may provide accurate correlations so it can be used to study relationships between variables

- Limitations of the approach:
  - Not enough context information to calculate response rates
  - Online panel is not (fully) based on probability sampling and it is affected by non-response and panel attrition

Future research

- Focus on improvements in data collection approaches to convenience sampling:
  - More control and tracking over where and how the survey link is shared (custom links)
  - Using auxiliary data of users of websites/social media pages where the survey link was shared
  - Developing standards for reporting data collection details in scientific publications
- Using propensity score weighting to match the characteristics of the more representative sample



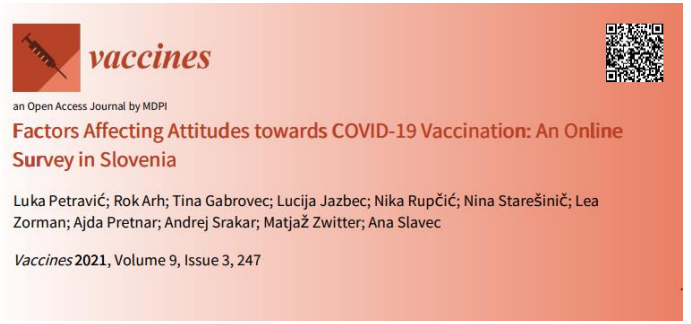
Resources

Compliance with COVID-19 protective measures 😷

- Medium posts (in Slovenian):
  - Večina še vedno na delovnih mestih
  - Obiski in širjenje virusa
  - Kako upoštevamo zaščitna priporočila?
  - Bolj verjetni prenašalci virusa
  - Samozaščitni ukrepi – razlike med regijami?
- Data in the process of being deposited to the Slovenian Social Science Data Archive

Attitudes towards COVID-19 vaccination 💉

- Medium posts (in Slovenian):
  - Kateri dejavniki vplivajo na odnos do cepljenja?
  - Koliko zdravstvenih delavcev se bo cepilo?
  - Dejavniki, ki vplivajo na odnos do cepljenja proti COVID-19
- Research paper (in English): DOI



- Data available through the Slovenian Social Science Archive: SARSPR20

InnoRenew CoE

Livade 6, 6310 Izola/Isola, Slovenia, T: +386 40 282 944, E: coe@innorenew.eu, www.innorenew.eu

Thank you for your time.