Part: Selection of Online samples

Ulrich Rendtel*

Franz Prücklmair†

June 13, 2022

Manuscript under construction

Do not cite

Contents

Abstract

Abstract: Keywords: Sampling frame, Access Panel, River Sampling, Non-probability samples

*FB Wirtschaftswissenschft, Freie Universität Berlin

†Universität Bamberg

1 A taxonomy of online surveys

Samples which are generated via the internet are not non-probability samples per se. There are quite different strategies used in practice to generate so-called "Online Surveys".

We distinguish here three basis types of sampling from the internet:

- Classical sampling and change to online mode
- Sampling from a frame of email addresses
- Sampling from an access panel of volunteers
- River sampling via internet gadgets

Sometimes these basic strategies are mixed in different order, for example, the River sampling is used to create a pool of potential interviewees which serve as an access panel. For the members of this access panel email addresses and basic background variables are known. On the basis of these background variables a stratified sample is generated. Such a strategy uses elements of all above mentioned baseline sampling strategies.

1.1 Classical sampling and change to online mode

In this approach one wants to combine the benefits of a classical probability survey with the gains of cheap and fast online interviewing. A recent example is the use of European Social Survey (ESS) which is a high standard face-to-face survey as the selection basis of an online survey, see Bottoni and Fitzgerald (2021). A special aspect of this cross-national online survey (CRONOS) is the comparability of the internet use across three European countries.

1.2 Sampling from a frame of email addresses

Here sampling agency uses a frame of email addresses and asks in an opening letter via email for participation in an online survey. Sometimes the persons get an personalized token for participation, sometimes it is an open link address. Sometimes reminder letters are send to late interviewees. Examples are surveys among students of a university or single university departments whose email address is known to the university administration.

For general population surveys there is the problem of frame imperfection: not all persons in the population have an email address or use the Internet. Even if there would be a 100 percent coverage of

the population there is the problem of a missing register, where the email addresses can be sampled from. To some extent the coverage problem is similar to the existence of a general frame of telephone numbers. Not all households have a landline access and there is no complete frame of valid landline numbers and mobile phone numbers.

If a valid frame exists then it is in principle possible to draw a probability sample from this frame and we can compute nonresponse rates. It depends on the confidentiality rules of the survey whether respondents and nonrespondents can be distinguished by some individual characteristics which are known from the frame. Usually nonresponse information is very limited. However, this holds also in other survey modes, like telephone interviews or even personal interviews, where often only regional information is at hand. In some cases the owner of the frame of email addresses has some information on the holder of the address. But one can use this information only if responders and nonresponders can be identified in the frame. Sometimes confidentiality arguments deny the identification of responders in the frame.

For this kind of sampling the main difference of online and telephone surveys is the interview mode. In general people answer in telephone interviews or also face-to-face interviews not so extreme as they do in online surveys (Bottoni and Fitzgerals 2021), especially, this holds for questions on attitudes.

1.3 Sampling from Access panels of volunteers

Access panels serve as a pool of volunteers for surveys. Their use has two advantages. First, the response rates are much higher compared to standard surveys. This reduces field costs in personal interviewing due to less nonresponse. Second, for the members of the access panel a lot of information is available, either from the recruitment phase or from earlier surveys. This enables the survey institute to use efficient stratification or the identification of specialized subpopulations of interest. Finally, it is possible here to identify response rates according to some known variables.

The recruitment of access panel members can be done via the internet. However, more traditional approaches are also in use. For example, the German Statistical Institute asked the participants of the German Mircocensus – before they were rotated out after four participation rounds in a mandatory survey – whether they are willing to participate in other non-mandatory surveys of the institute, see Amarov and Rendtel (2013, 2014), Enderle et al. (2013) and Rendtel and Amarov (2014). Often participation in an access panel is incentivated by payments which were advertised by the survey institute. Thus the recruitment is prone of self-section effects, cf. Bethlehem (2010).

1.4 River sampling via internet gadgets

River sampling is a sampling method which is essentially linked to internet users, see AAPOR (2010). For example, readers of some online newspaper, will find a voting gadget whether he/she agrees with a certain law initiative. Typically, such a gadget is positioned within an article on the law initiative. The incentive to answer the question is the direct information how other readers of the article judge this initiative. After this round the readers are asked whether he/she is interested in answering other questions which are not linked to the topic of the article. Before this is done, the interviewee is asked for some basic information on gender, age group and the Zip-code of his/her home. At this level the users are identified by their cookies. Later-on they are asked for an email address for a more direct contact.

Not all gadget votes will enter the final survey result. Often only a stratified subsample of all answers is used for the final survey result. The distribution of the gadget can be also linked to special online platforms and/or to special timepoints. In order to run specialized surveys one can use only topic-specific online platforms.

Compared to the sampling schemes before we don't know any nonresponse rate nor do we have any idea of a sampling probability. The result is a typical non-probability sample, cf. Baker et al. (2013).

2 Modelling selection into online surveys

In order to participate in an online survey one has to have access to the Internet and use it frequently. One has to come across certain gadgets in case of river sampling or certain adds, which ask for participation in a certain survey or in an access panel. For probability based surveys it is often sufficient to have a valid email address.

In the early decades of the Internet a private access was not guaranteed for every household. For example, Valliant (2020) reports for the Michigan Behavioral Risk Factor Surveillance Survey from 2003, which served as a simulation basis for selection from the Internet, has an internet coverage rate of 60 percent. Van Dijk (2005) refers to this lack of coverage as "digital divide". Disparities of usages style manifest a so-called "second-level digital divide" (Hagittai 2002). Nowadays, internet access in private households is almost 100 percent[1]. So it does not make sense to link participation rates in online panels to internet access.

However, the frequency of internet use may be a good indicator that the users comes across a gadget

[1]See the development for Germany under `https://de.statista.com/statistik/daten/studie/153257/umfrage/haushalte-mit-internetzugang-in-deutschland-seit-2002/` (Access 10.2.2022)

or an add. Here the ESS questionnaire asks whether a person uses the Internet daily (y/n) and if yes, how long on a typical day. As the daily use of the Internet has become a standard in German households [2]. Therefore one should look on the duration of the internet use on a typical day. Here we tried different thresholds for the selection into an online survey. Alternatively on may use a continuous selection probability according to the length of the internet access [3]

An important feature of participation in online surveys is the individual usage of the Internet. There may be persons who use the Internet only for reading or the download of documents or there are persons who use the internet quite active. There are many reports that active internet users are over-represented in online panels (Chang and Krosnick 2009, Dever, Rafferty and Valliant 2008 and Malhotra and Krosnick 2007). The ESS asks its respondents: Have you posted or shared anything about politics online in the last 12 month? (y/n). Therefor we used this Blog variable for the selection mechanism.

Apart from the Internet access and the individual style to use it motivational aspects of self-selection (Bethlehem 2010) may be important. Often gadgets are posted within online newspaper articles. Here the users are asked for their opinion on the topic where the article is about . The reward for the participants is that they immediately see how other people have voted. Here topical self-selection (Lehdonvirta et al. 2020) comes into play. For example, Chang and Krosnick (2009) reported that political participation is 10 percent higher than in RDD-samples or probability internet samples. This motivationel aspect will especially high in case of political dssatisfaction, where online participants can express their anger with political circumstances via a gadget voting. The ESS offers some variables which could be used in this context, see Table 1

Table 1: Variables in the ESS with relevance for participation in an online panel

Politics	
How interested in politics	1,...,4
Trust in political parties	0 (low),1, ..., 10 (high)
Posted or shared anything about politics last 12 month	y/n
Placement on left/right political Scale	0 (left), ... , 10 (right)
How satisfied with national government	0 (low), ... , 10 (high)
Subjective well being	
How happy are you?	0 (low), ... , 10 (high)
Subjective general health	1 (low), ... , 5 (high)
Belong to a minority ethnic in country	y/n

Beyond motivational variables there may be an economic self-selection if persons in an access panel are

[2] According to the IKT-Survey 2018 89 percent of users have access "almost every day", see Statistisches Bundesamt (2018??)

[3] However, most of the reported durations we entire hours.

payed for their participation.

Of course, there are some correlations of the variables with demographic variables, for example, internet use and age. In his simulations Valliant (2020) used only age to simulate self-selection in his artificial universe. Here, we use the above listed variables to simulate participation in an online survey. Later on we will try to compute quasi-randomisation or calibration weights from demographic variables, which are known from a probability sample or whose totals in the population are known. There is a special group of variablesv which refer to the internet use whose population totals are known. Such "webograghics" (Schonlau, van Sost and Kapteyn 2007) can be useful to augment the standard demographic variables.

References

American Association for Public Opinion Research AAPOR) 2010: *Report on Online Panels* `https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels.aspx` (Accessed 26.1.2022)

Amarov, Boyko; Rendtel, Ulrich (2013): The Recruitment of the Access Panel of German Official Statistics from a Large Survey in 2006: Empirical Results and Methodological Aspects, Survey Research Methods, 7, 103–114

Baker R., Brick J. M., Bates N. A., Battaglia M., Couper M. P., Dever J. A., Gile K. J. and Tourangeau R. (2013) *Report of the AAPOR Task Force on Non-Probability Sampling*

Bethlehem J. (2010): Selection Bias in Web Surveys. *International Statistical Review*, 78 161-188.

Bottoni, G., and R. Fitzgerald,(2021): Establishing a baseline: bringing innovation to the evaluation of cross-national probability-based online panels. *Survey Research Methods* 15, 115-133. `https://doi.org/10.18148/srm/2021.v15i2.7457`

Chang, L., and J.A. Krosnick (2009): National Surveys Via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly* 73, 641–678.

Dever, J.A., A. Rafferty, and R. Valliant (2008): Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? *Survey Research Methods* 2, 47–62.

Enderle T., R. Münnich and Ch. Bruch (2013): On the Impact of Response Patterns on Survey Estimates from Access Panels *Survey Research Methods* 7, 91-101.

Hargittai, E. (2002): Second Level Digital Divide: Mapping Differences in People's Online Skills. *First Monday* 7 (4). http://firstmonday.org/article/view/942/864

Lehdonvirta V., A. Oksanen , P. Räsänen , and G. Blank (2020): Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research. *Policy & Internet*, 13, 134 - 155. https://doi.org/10.1002/poi3.238

Malhotra, N., and J.A. Krosnick. (2007): The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys With Nonprobability Samples. *Political Analysis* 15, 286–323.

Rendtel, Ulrich; Amarov, Boyko (2014): The Access Panel of German Official Statistics as a Selection Frame, In: Engel, Jann, Lynn, Scherpenzeel, Strugis (eds): Improving Survey Methods, Taylor & Francis / Routledge, Chapter 20, 236–249

Schonlau, M., V A. Soest, and A. Kapteyn (2007): Are "Webographic" or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?, *Survey Research Methods* 1, 155–163.

Valliant, R. (2020): Comparing alternatives for estimation from nonprobablility samples. *Journal of Survey Statistics and Methodology* 8, 231 - 263.

Van Dijk, J. (2005): *The Deepening Divide: Inequality in the Information Society* London: Sage