Statistical Analysis - Representative Samples of Populations

Ole André Hauge
Department of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

***Abstract*—This paper looks at various methods and procedures that can be used to collect representative samples of populations. The information was found using a qualitative literature study of papers, lectures, and sources found on academic search engines and in academic books. The findings show that there are numerous sampling methods that can be used to create a representative sample, however, a random sampling method are said to give the best results. Furthermore, there is no guarantee or complete protection against sampling biases, but researchers have developed methods to mitigate this risk. A questionnaire was used as a case study to reflect on how creating a representative sample, keeping in mind the biases, can be solved. It is evident that researchers have to keep this risk in mind when obtaining samples for research.**

***Index Terms*—Representative sample, population, probability, non-probability, methods, sampling bias, survey**

I. Introduction

Sampling approaches are employed in statistical analysis to gather insights and observations about a population. These strategies are used by researchers to achieve the objectives of their research studies [1, p. 105]. Representative samples are one such sampling methodology, as they are subsets of a population that are intended to correctly reflect the characteristics of the given population. The method has several advantages, including costs, better depth, and the assessment of more variables.

Although representative sampling can increase the survey's content validity [1, p. 115], it is doubtful that every sample will be identical to the population of interest unless every person in the population is in the sample. It is important to remain alert for biases or other errors in the samples, which might lead to a lack of representativeness [2]. Self-selection bias is an example of sampling error in which the sample has an over-representation of persons who are biased towards the question at hand [3]. Another example would be investigating men's physical strength but only taking samples from strongman athletes.

It is easy to understand how the features of the samples might potentially influence the outcomes in these examples. To mitigate this issue researchers take precautions to ensure internal and external validity. For internal validity, double-blind experiments, supplying misleading information, triangulation, and conducting unobtrusive experiments can be used [1, p. 104]. This, however, poses ethical issues [3]. The external validity is further discussed in this paper as it is especially important when creating a representative sample as it inherently should be a generalization of the population it is taken from.

The representative characteristics that best satisfy the study goals can be chosen by the researchers. These traits are often centered on demographic categories such as age, gender, marital status, and other socioeconomic aspects [4]. The amount of features utilized is often influenced by the size of the population under consideration. The more populous the population, the more traits are required to depict it. The size further influences the sample size, depending on the phenomena being examined [1, p. 184].

The purpose of this study is to discuss how to assure the validity of surveys by employing representative samples of populations. The following are the primary contributions of this paper:

1) It explains how to make a sample representative of a population.
2) It finds and describes methods and procedures to achieve this.
3) It discusses how this is solved in the study of *Wayfinding and navigation in the outdoors* [5].

The remainder of this paper is organized as follows: Section II describes the methodology utilized to address the research topic in the study. Section III provides the findings, while Section IV discusses them. Section V concludes the study, followed by the relevant references.

II. Methodology

The topic of this study was explored qualitatively. To provide a framework for the discussion of the research problem, a literature review was conducted. This entailed locating relevant research papers, books, reports, and other textual sources and evaluating their relevance to the issue. To identify relevant academic materials, well-known internet academic search engines and literary databases were employed. This paper is based on the findings of the data that were found.

III. Results

The findings show that a representative sample is the best choice for sampling analysis as it is likely to produce insights and observations that closely match the overall population group [1, p. 183], [6], [7]. Obtaining a good representative sample, on the other hand, might be difficult. There are two

main methods to gather samples [1, p. 177], [3]. This section describes these two methods and their procedures in the two corresponding subsections.

A. Probability methods

Probability methods use random selection to draw strong statistical conclusions about a targeted population. It means that everyone in the population has the same probability of getting chosen for the representative sample [4]. If the goal is to have findings that are typical of the entire population, this is a suitable choice. There are several procedures available to collect data at random, the four main are: random-, systematic-, cluster-, and stratified-sampling [8].

In random sampling, each member of the population has an equal probability of being chosen. It is a simple sampling strategy that a researcher can use to create a representative sample. It is less expensive and uses fewer resources, but comes with a larger chance of sample error, which can result in inaccurate data and false conclusions [3].

Systematic sampling attempts to organize its components by numerically listing all members of a population and selecting them at regular intervals rather than selecting them at random. This sort of sampling might entail selecting every fourth individual from a population list as a sample. Despite the inherent methodical nature of this procedure, it is nevertheless likely to provide a random sample. However, researchers must be cautious of hidden patterns in the list that might bias the results [3], [4].

Cluster sampling divides the population into subgroups with characteristics that are comparable to those of the total sample. Rather than picking members from the subgroups at random, the subgroups are selected at random. The method is known to perform well with large, scattered populations, but it comes at the cost of a larger risk of sampling errors due to the likelihood of major variance between the clusters [8].

In stratified sampling, the population is separated into groups called strata. A stratum is a demographic subset with at least one common characteristic. This makes it easier to select the right number of people from each stratum to make a representative sample with the same population proportions. While this method requires more upfront information, is more time-consuming, and is costly, the information obtained is generally of higher quality [4], [8].

B. Non-probability methods

Non-probability methods choose population members based on a set of criteria rather than at random. As a result, the obtained sample will not contain every member of the population [8]. These methods, in general, have a higher risk of sample biases, but they are less expensive and easier to use. Non-probability sampling methods are frequently employed in exploratory research where the goal is to obtain an inexpensive approximation of the truth rather than verify a hypothesis [3]. Convenience sampling, judgment sampling, snowball sampling, and voluntary response sampling are four standard methods [3], [8].

Convenience sampling refers to a researcher's selection of the most convenient members or participants. Although it is a simple and inexpensive method of gathering data, there is no way to ensure that the sample is representative of the entire population [8].

Judgment sampling is when researchers utilize their experience to pick a characteristic that is most relevant to the goals of their research. It is frequently employed in qualitative research, especially when the researcher wishes to learn more about a certain phenomenon rather than making statistical inferences, or when the population is small and specific. To be effective, a purposive sample must have explicit inclusion criteria and justification [8].

Snowball sampling is a good way to collect samples for small phenomena. In this method, one participant recruits more participants, resulting in a snowball effect [3].

Finally, voluntary response sampling entails people volunteering to answer questions or participate in studies [8]. An example is people who respond to a poll posted on Facebook. Electronic surveys are frequently helpful in this regard. Cost, rapid access, a large audience, the ability to download data to spreadsheets, and flexibility in the layout are all advantages of electronic surveys. On the other side, there are certain drawbacks, such as privacy and anonymity, as well as a low response rate. It can be beneficial if the purpose is to sample a large or specialized population [3].

IV. Discussion

As can be seen from the findings, a representative sample is likely to produce the best outcomes and may be considered to be a good representation of the population under examination. There are obvious drawbacks to utilizing this sample procedure, such as the fact that it is time-consuming, costly, and resource-intensive. Researchers who want to use the most accurate method, stratified random sampling, must spend time identifying the different characteristics and dividing the population into strata before they can start to choose members for the representative sample.

In general, representative sampling becomes more difficult as the population grows larger. This approach can be challenging to employ if one wants to research an election or a racial problem since it may be difficult to get the desired members to participate in the study. This might happen, for example, if people do not believe they have enough time to participate, resulting in an under-representation in the representative sample. As a result, it's critical to comprehend the benefits and drawbacks of representative sampling, as well as the many data collection methods available [3].

In many circumstances, researchers choose to collect data using non-probability approaches since they are faster and easier. As can be seen from the findings, these approaches make it more difficult to make statistical inferences about the population of interest since one cannot ensure that the sample comprises a representative sample of the population. As a result, if researchers wish to make conclusions that are

typical of a population, especially if it is large, they should employ probability methodologies.

Because probability approaches are rarely perfect, one must be cautious of sampling mistakes and biases in the representative sample. To mitigate this, one must evaluate the sample frame, who all of the potential participants are, as well as how one plans to recruit the study's participants. This will allow researchers to avoid exclusively recruiting individuals from certain subsets of the target demographic, as in the example of the strongman athletes discussed above. The next step is to choose a method for picking members of the sample frame at random. As a result, researchers reduce the risk of selection biases, such as volunteer bias, in which the sample has an over-representation of people who are biased towards the subject at hand. It's critical to ask, "Who actually responds to us?" Those who have been affected by the phenomena are frequently the ones who reply [3].

Whit this in mind, consider the survey conducted by Ole E. Wattne and Frode Volden on *Wayfinding and navigation in the outdoors* [5]. It was part of PhD-research at NTNU to research people's use and perceptions about technologies and phenomena concerning navigating in the outdoors. The survey aimed to understand how one can design to help people navigate. The data was collected anonymously and was protected per the guidelines of the *Norwegian Centre for Research Data*, as stated by the survey [5].

The sampling was conducted by personal invitations for participation via mail or other channels, but not publicly announced. This was done to ensure that the researchers could describe the invited population and document the response rates. This is a non-probability method and can be interpreted as a hybrid of snowball sampling and judgment sampling, as students at NTNU were asked to email at least 15 people asking them to participate in the study. Letting the students pick who they forwarded the survey invitation to might have resulted in the recipients being more likely to respond to the survey as a result of our previous relationships. I did for example only forward the survey to other students at NTNU, which can be a group that is more entailed to understand the need for the study and thus more likely to answer the survey. They might also feel compelled to answer the survey as they know that they might have to send out surveys in the future and hence have to depend on other students' participation as well.

This is an example of self-selection bias and can be a drawback of using this sampling method. However, the reason for the choice of this sampling method might be to save time and keep costs low, which are the advantages of using this method. As the researchers are aware of this weakness, they are more likely to compensate for that or at least keep it in mind when analyzing the data and looking at the results. This statement is further strengthened as the researchers asked the students, whom they used to propagate the invites through and to describe the method they used including the number of invites they sent.

Another bias that is self-evident in the questionnaire is the volunteer bias, which can be seen as a subset of self-selection bias. The survey is answered voluntarily which is clearly stated on the first pages consent from: "It is voluntary to answer the questionnaire [...]" [5]. This might mean that the people who answer the questionnaire are biased toward the topic or the idea of conducting or participating in such surveys.

The questionnaire asked the participants about their age and gender. This is most likely done to give the researchers an understanding of how the samples represented the population. It then proceeded to ask general questions about the participant to profile how often they were outdoors and what activities they did. They also included an option to freely type in outdoor activities to capture as much data as possible. The second and third parts focused on questions about navigation, use of maps, landmarks, and place names. Here the researchers implemented an *Agree / Disagree survey*, which allows respondents to answer more precisely and provided the researchers with more nuanced survey responses to analyze. It also enables the researchers to see if the participants are focused on answering or if they are rushing through the questions. For example, if a participant answered *Strongly agree* to the following two statements: "I have a good sense of place" and "I often get lost in the outdoors", it could indicate that the respondent did not pay sufficient attention to the questions when answering.

Therefore, looking at the survey it is evident that the researchers have made a calculated decision weighing the advantages of a non-probabilistic method against its downfalls. The implementation of contradicting questions also shows that they thought of a method to measure the focus of the participant to further be able to say something about the reliability and validity of the data.

V. Conclusion

The findings show that it is complicated to attain a perfectly representative sample of a population. Several methods, either probability- or non-probability-based can be used for sampling where a random selection is viewed as the most accurate alternative. The findings showed that there is no guarantee or complete protection against sampling biases. It is therefore important to keep this in mind when conducting research. Looking at the *Wayfinding and navigation in the outdoors* survey one can see how some of these methods, like contradicting questions and asking for age and gender, are used to mitigate sampling errors. It is evident that the researcher has made a calculated decision weighing the advantages of a non-probabilistic method against its downfalls. The biggest takeaway is that researchers have to be aware of the various challenges, like sampling biases, when they research in order to mitigate the negative impact on the study's findings, results, and validity.

References

[1] P. D. Leedy and J. E. Ormrod, *Practical Research: Planning and Design, 11th Edition*. Pearson Education Limited, Edinburgh Gate, Harlow, Essex CM20 2JE, England: Pearson, 2019. ISBN: 9780133741322.

[2] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974. DOI: 10.1126/science.185.4157.1124.

[3] F. Volden, “Lecture 5: Surveys and representativity of participants.” Cloud.swivl.com, 13.10.2020. [Sound recording]. Available: https://cloud.swivl.com/v/ca0e734238f749afc85ba4fdaf91c79c. [Accessed on: 20.10.2021].
[4] J. Young, “Representative sample.” https://www.investopedia.com/terms/r/representative-sample.asp, 08 2021. [Accessed on: 23.10.2021].
[5] F. Volden, “Wayfinding and navigation in the outdoors / veifinning og navigering i naturen.” https://nettskjema.no/a/223986#, 2021. [Accessed on: 03.11.2021].
[6] S. Solutions, “What is a representative sample?.” https://www.statisticssolutions.com/what-is-a-representative-sample/, 2021. [Accessed on: 23.10.2021].
[7] B. D’Exelle, *Representative Sample*, pp. 5511–5513. Dordrecht: Springer Netherlands, 2014. DOI: 10.1007/978-94-007-0753-5_2476.
[8] S. McCombes, “What is a representative sample?.” https://www.scribbr.com/methodology/sampling-methods/, 09 2019. [Accessed on: 23.10.2021].