# **Sampling strategy**

Careful selection of an appropriate sampling approach plays a crucial role in achieving the objectives of the study, as emphasized by Stratton in 2021. For this particular experiment, it is essential to have two distinct, well-structured datasets: one *for individuals seeking employment* and another *for available job listings*. As it was mentioned in previous section the datasets that will be used in the job seeker's dataset will serve as a representation of the participants in the offline experiment, reflecting the demographic characteristics from which the study's subjects will be drawn. In contrast, the job offers dataset will represent the population of 'job listing pool'. This dataset forms the context in which participants will engage in the process of matching and ranking through various vector space NLP architectures.

Below, you’ll find an overview of the selected sampling strategies used for each dataset, along with specific considerations and underlying rationales that inform these choices.

***Job Seekers dataset:*** Before selecting participants to represent the job seeking population, it is critical to consider the following sampling limitations:

1. Difficulty in defining the complete job searchers population*,* whether it refers to an entire nation or a smaller geographical region. This complication arises from the multitude of job seeker’s categories, including those who are actively seeking, passively looking, monitoring, or transitioning between these states due to their dynamic nature (Bortnick, 1992).
2. Unlike other populations such as voters, licensed professionals etc., there is often no unified registry or database that encompasses the complete information of all job seekers, which collectively contribute to the limited accessibility of this population.
3. In this situation, the information provided by participants will only be used to evaluate and compare system architectures and interact with the data in the job listing pool. Any findings and conclusions drawn from the use of the sample won’t be used for generalization of the entire population (Yin, 2009).

Given the limitations indicated above, the *convenience* *technique* would be the most suitable option for selecting participants in the proposed experiment. In addition, it is important to point out that the limitations, as mentioned earlier, fit precisely the qualities defined by Etican et al. in 2016. They describe convenience sampling as a non-probabilistic technique in which participants are chosen based on their accessibility and proximity. It is a convenient and cost-effective method that allows researchers to quickly collect data from voluntarily available individuals ready to engage in the experiment.

Furthermore, they underscore the common use of this method in situations where obtaining a random or representative sample is challenging. However, it is essential to acknowledge that convenience sampling comes with its own set of limitations, as the samples obtained may not precisely mirror the broader population, necessitating caution when interpreting and generalizing results derived from convenience sampling (see Figure 4).

As for the recruitment of volunteers, online social media platforms will be used due to their cost-effectiveness and faster sample acquisition (Antoun et al., 2016). A well-articulated post explaining the research objectives and the expected time commitment will be created and shared on the chosen platform. Once potential volunteers are attracted, detailed information about participation requirements, risks, benefits, and the rights of participants will be provided through personal messages. After obtaining informed consent from those who are willing to join the study, they will undergo *online interviews* using video conferencing platforms. Each volunteer will participate in a structured interview, answering a predetermined set of questions to provide detailed information on the topics illustrated in Figure 5. There is no definite sample size limit and in order to keep costs low, only the most comprehensive and appropriate primary dataset that closely resembles the real-world candidate will be used in the experiment.

***Job Offers dataset:*** Web scraping is the process of automatically extracting data from the internet and structuring it into a well-organized dataset, as discussed by Wickham in 2014. This method offers several advantages, including the rapid and error-minimized collection of large volumes of data. Additionally, it is not a novel concept in the realms of statistics and data science education. Over the last decade, there has been a growing emphasis on computational aspects in statistics and data science education, which has led to increased popularity in the practice of web scraping. A survey conducted by Hardin and colleagues in 2015 revealed that six out of seven schools that introduced data science topics in their statistics curricula included web scraping.

The domain of web scraping offers a lot of different techniques in various formats, each with its own unique advantages. In this research proposal, the Python library Beautiful Soup (see Figure 6) can be used as the data collection tool of job listing pool due to its user-friendly nature and strong community support (Khder, 2021). The objective of using this tool would be to extract publicly available secondary data from the Indeed job platform that can be used in experiment validation. It's important to note that neither random job listings nor the entire job database will be collected from the platform, as this would be impractical. Instead, to create a validation test closely resembling real-world conditions for our system architectures, results of a keyword search based on each participant's specified job title information will be gathered.

For example, if an experiment participant indicates in his or her interview that he or she is looking for opportunities in the nursery domain, all job ads on Indeed that result from a 'nurse' keyword search for the desired area would be collected. As of October 27, 2023, there are 653 nurse-related job advertisements in Dublin and its surrounding locations, as shown in Figure 7. Further, as indicated in Figure 1, the main focus will be on extracting information from these postings which includes the job title, required skills, required experience, and educational qualifications. All other information will be excluded from the experiment for reasons that will be detailed specifically in the proposal's ethical section.