Statistical Fallacies

Seminar: How Do I Lie with Statistics?

Mustafa Fuad Rifet Ibrahim

WS 2019/20

Contents

1 Introduction

A scientific study has many parts to it, starting with the sampling and data gathering and going all the way to reporting the findings and presenting possible interpretations. On all these levels there are multiple opportunities for biases and errors to creep in and reduce the overall validity of the study. There are of course ways to intentionally exploit these fallacies and so to guard oneself and to keep up the general qualitiy of research in the scientific community one has to understand each facet of a scientific study and how it relates to biases, errors and manipulation. The purpose of this report is to analyze the first parts of a scientific study, namely the sampling and the measurement.

2 Sampling

Sampling refers to the process of picking a subset of elements from the target population, that is the actual object of interest, in such a way that it is as representative of the target population as possible (Figure 1). The measurements and subsequent analysis are then done on that subset and the goal is to infer information about the target population. The purpose of this procedure is to save resources in the data collection process while maintaining a certain degree of validity. There are two fundamentally different ways of sampling, namely probability sampling and non-probability sampling. The following subsections will clarify the important properties of these two categories as well as give examples of methods that fall under them.

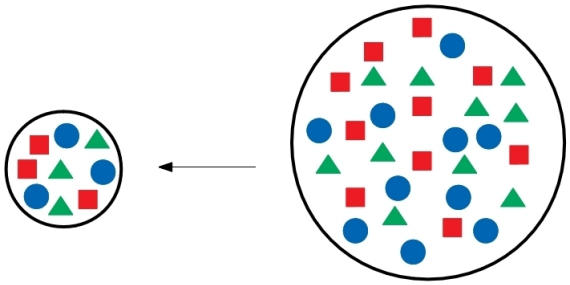


Figure 1: Visualization of the basic sampling procedure. Here, the target population is represented by the bigger circle containing the elements of interest (basic geometric shapes) and the sample is represented by the smaller circle.

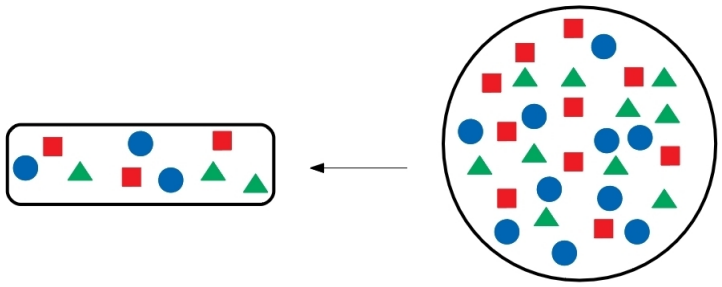
2.1 Probability Sampling

In probability sampling every single element of the target population has a non-zero chance of being picked and this probability can be accurately determined. This allows for the use of statistical theory to estimate sampling errors and infer from sample to target population. There are several different methods that fall under this category and in the following their strengths and weaknesses will be explored[1].

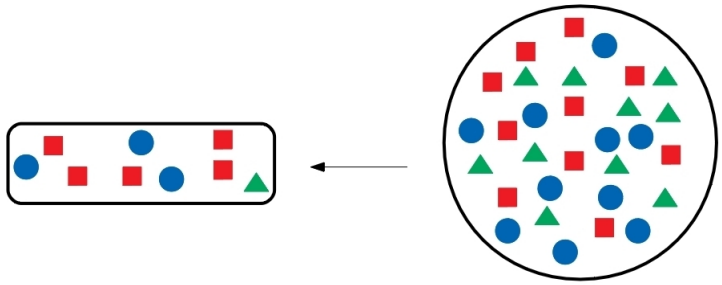
2.1.1 Simple Random Sampling

Simple Random Sampling (SRS) refers to a way of sampling in which any two subsets of equal size have the exact same chance to be picked from the target population (Figure 2). This is the simplest kind of probability sampling since it doesn't require any additional knwoledge about the target population that could be used to subdivide it further. The downside however is twofold. For one, the simplicity comes at the cost of efficiency since this sampling method requires the ability to do repeated random draws from the entire target population of

interest. For instance if the target population was the population of a city, a random sample could lead to high travel costs. The second problem is the fact that the method doesn't ensure representativeness of any single sample drawn since these samples are picked at random (Figure 2). This means that it is possible that the relative proportion of the different element types in the sample doesn't resemble their relative proportion in the target population[2]. However, the representativeness can be ensured by drawing many samples using simple random sampling and averaging over them or increasing the sample size[3].



(a) Unbiased Sample



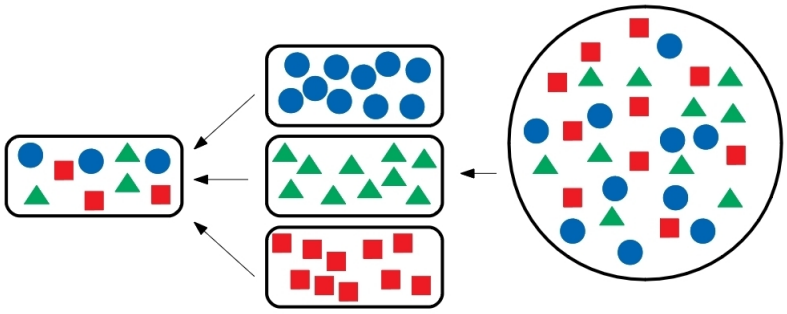
(b) Biased Sample

Figure 2: Visualization of simple random sampling. In picture a) the sample drawn is a perfect representation of the target population. In picture b) the sample drawn is biased towards the red square elements. In this case the colour and shape of the elements can be thought of as either being the actual variable of interest or correlating with the variable of interest.

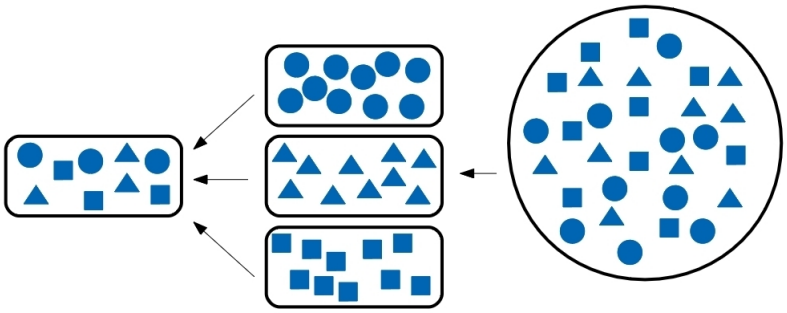
2.1.2 Stratified Sampling

Stratified Sampling consists of two parts. First the target population is subdivided into groups according to a variable that correlates with the variable of interest. This creates homogenous groups, i.e. the intragroup correlation is maximized while the intergroup correlation is minimized. Then, a random sample is drawn from each group relative to the proportion of the group in the target population (Figure 3). This method has the advantage of ensuring that samples drawn are not biased towards certain groups in the context of the vari-

able used for subdiving the target population. In addition to that it can make the process of measurement cheaper and more efficient depending on the specific tools or methods used to measure the paramters of interest. The downside of this method is that it can't be properly utilized when there is no such variable according to which one can subdivide the target population (Figure 3).



(a) Stratification possible



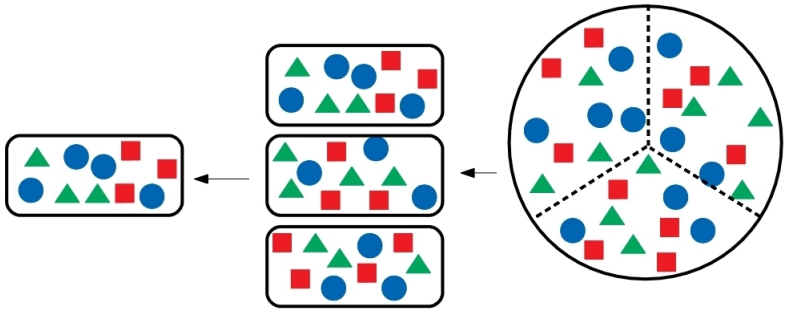
(b) Stratification impossible

Figure 3: Visualization of stratified sampling. In this example the variable of interest is the shape of the elements and the variable that correlates with it is the colour of the elements. In picture a) the elements are grouped according to the colour and then a random sample is drawn from each group relative to the proportion of that group in the target population. In picture b) there is no variable available that correlates with the variable of interest and so stratified sampling can't be applied here. In this case it essentially resembles cluster sampling.

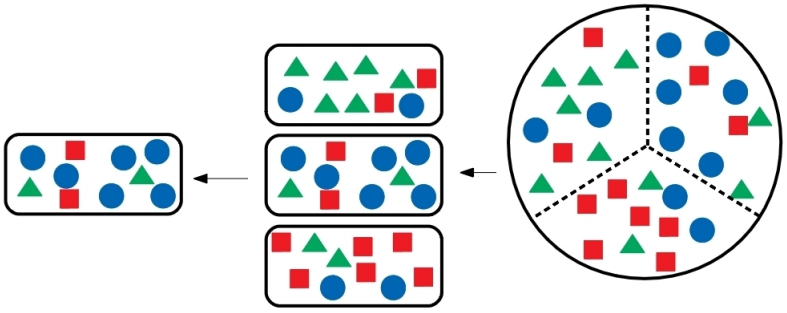
2.1.3 Cluster Sampling

Cluster sampling consists of two parts. First the population is subdivided into groups (clusters) based on some variable that doesn't correlate with the variable of interest and then a random cluster is picked with simple random sampling and used to measure the parameters of interest (Figure 4). This creates heterogenous groups, i.e. the intragroup correlation is minimized while the intergroup corre-

lation is maximized. This method of sampling can be used to minimize resource cost. Using the example target population of a city again, cluster sampling allows to cluster the population into e.g. districts of the city and so the variable used to subdivide would in this case be the physical location of the people with respect to the districts of the city. Then, after picking a random district all people in that district can be used to measure the parameters of interest, saving time and money. This is of course under the assumption that the location is completely independent of the variabel of interest. Otherwise a bias would be introduced. The downside of cluster sampling lies in the fact that in the real world finding a variable that is truly independent of the variabel of interest is hard. Thus, intragroup correlation can often not be minimized which causes a sampling error that increases with the strength of this correlation (Figure 4).



(a) Unbiased Sample

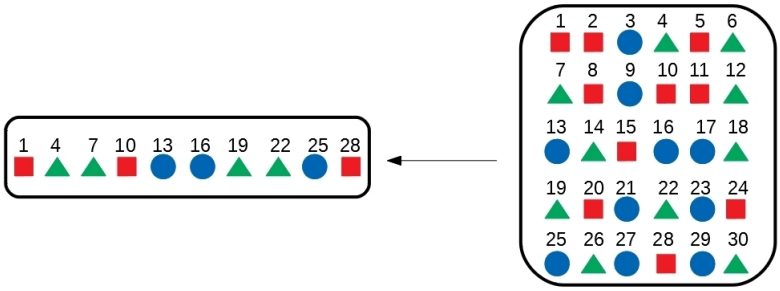


(b) Biased Sample

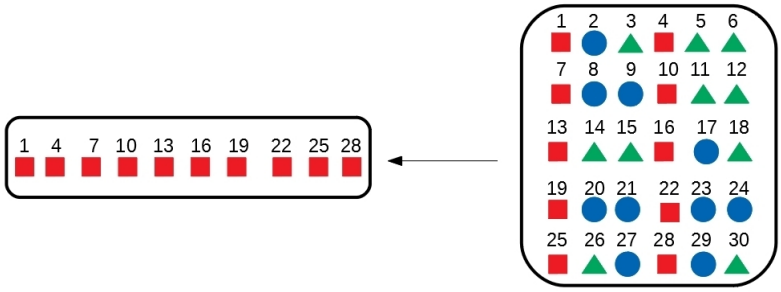
Figure 4: Visualization of cluster sampling. In this case the circle that resembles the target population was subdivided into three roughly equally big regions, i.e. the variable of the subdivision used was the location according to this arbitrary subdivision of the circle area. As can be seen from the picture a) this location does not correlate with the variable of interest which in this case can be either the shape or the colour. This yields an unbiased sample. In contrast the picture b) shows a target population where the location with respect to this arbitrary subdivision of the circle area correlates with the variable of interest (shape or colour), producing a biased sample.

2.1.4 Systematic Sampling

Systematic Sampling is a way of sampling in which the target population is first ordered according to a variable and then elements are picked from this list starting at a random starting point and stepping through with a certain step size (Figure 5). Depending on the desired sample size and the size of the target population, the step size is given by $\frac{N}{n}$, where N is the target population size and n is the sample size. When the variable according to which the elements of the target population are ordered is correlated with the variable of interest, then this method of sampling resembles the stratified sampling method because the ordering essentially induces a stratification. If that variable is not correlated with the variable of interest, then this method essentially resembles a more efficient simple random sampling but without the guarantee that subsets of equal size have the same probability of being chosen. Regardless of this distinction, the weakness of this method lies in hidden patterns of the target population. These can introduce a bias (Figure 5).



(a) Unbiased Sample



(b) Biased Sample

Figure 5: Visualization of systematic sampling. In picture a) the sample is drawn by starting at position 1 and using a step size of 3 which yields a sample size of 10. In that case no hidden pattern introduces a bias. In picture b) the same exact pattern of sampling is used, however in this case there is a hidden pattern in the target population that causes the sample to be extremely biased.

2.2 Non-Probability Sampling

In non-probability sampling the criterion for probaibility sampling is simply not met, i.e. there is at least one element of the target population that has zero chance of being picked or the probability of being selected is not determinable. These types of sampling techniques are not suitable for inference of parameters of the target population and sampling errors can not be calculated. They are instead meant for exploration and when a probability sampling technique is infeasable. They do however have the advantage of being cheaper than probability sampling methods and sometimes they are the only feasible option. There are multiple examples of methods that fall under this category. The following sections look at a few examples of this fundamentally different way of sampling since it is widly used[1].

2.2.1 Convenience Sampling

In convenience sampling elements from the target population are drawn based on how convenient it is to draw them. For example asking people at the nearest and easy to reach location where a lot of people can be found. This can obviously reduce costs but also introduce extreme biases depending on what the variable of intererst is and how the variables that were implicitly and explicitly used to determine the convenience correlate with that variable of interest.

2.2.2 Snowball Sampling

In snowball sampling the sample is picked by recruiting people through the social network of a few people that are initially contacted directly by the people who designed the study. The study thus spreads through the social graph and the number of participants can grow exponentially. Again, the advantage of this method is the simplcity and low cost, especially if the study is spread through the internet. However, this method of sampling is inherently biased towards people that have more social connections since they will receive the opportunity to partake in the study with a higher likelihood than people with less social connections. If this variable of social connections somehow directly or indirectly correlates with the variable of interest, the sampling error will increases.

2.2.3 Judgemental Sampling

Another form of non-probability sampling is judgmental sampling in which the participants are picked by how well they are suited to participate in the study. This is done purely on a subjective basis by the people conducting the study. A well known example for this are case studies.

3 Measurement

Measurement is the process of gathering the data on the parameters of interest. From using the bare human senses to tools and machines that allow us to go beyond the human limits, measurement is an essential part of a study. As is the case with sampling, measurment also allows for errors and manipulation to decrease the validity of the study. The following sections will give an overview over different aspects of measurement and how they can be problematic[4].

3.1 Definition of Measured Quantities

At the base of all measurement is the actual definition of the quantities one is about to measure. When the definitions are clear and universally agreed on, as is usually the case in the hard sciences like physics, there will be no confusion amongst the scientific community and the public as far as this particular aspect of the study goes. However, in areas like sociology or psychology definitions tend to get more vague since the topics that are dealt with in those areas are more complex and harder to break down and quantify. A good example of this are unemployment estimates (Figure 6).

Agency Preparing Estimate	*Estimate of Number Unemployed*
The National Industrial Conference Board	9,177,000
Government Committee on Economic Security	10,913,000
The American Federation of Labor	10,077,000
National Research League	14,173,000
Labor Research Association	17,029,000

Figure 6: Differences in unemployment estimates[4]. These unemployment estimates are from reputable agencies from the year 1935. The drastic differences in the numbers are the result of differences in the definition of an "unemployed person". Some of the points in which the definitions differed include people that just left school and are looking for a job, people that have found jobs but have not yet reported for work, people that work part-time and many more.

3.2 Precision

The precision of measurements is another point which can be exploited. In an ideal case the precision of a measurement is stated with respect to the specific measurement tool used and the results are presented with standard deviations. Doing this ensures that there is clarity in where potential errors or mistakes could come from. Moreover the rules for significant digits are not violated to roughly maintain significance throughout the calculations. The more rigorous option would be to calculate the proper propagation of uncertainty for the function or functions at hand. However the very trust in this system can be abused

by stating results with a precision that is not justified by the data or tools used. This practice is referred to as spurious precision or false precision. This can evoke a sense of competence and trust in the reader of the study because high precision is generally associated with correctness. This can also occur when converting between units by using more significant digits in the converted number than were present in the original number.

4 References

References

[1] Robert M. Groves et al Survey Methodology 2009

[2] Tyson H. Holmes Ten categories of statistical errors: a guide for research in endocrinology and metabolism Am J Physiol Endocrinol Metab 286: E495–E501, 2004

[3] David J. Slutsky Statistical Errors in Clinical Studies J Wrist Surg 2013;2:285–287

[4] Stephen K. Campbell Flaws and Fallacies in Statistical Thinking 1974