

Summary Report cE3c Advanced Course 2021/2022

**Experimental Design and Reproducibility in Science**

João Belo

Date

The aim of this report is to summarize the information and contents acquired during the course of “Experimental Design and Reproducibility in Science”, lectured by Dr. Inês Fragata, Dr. Leonor Rodrigues and Dr. Diogo Godinho. To accomplish this, I will provide a summary of the daily classes.

**Day 1**

The first day of the course started with three steps that every researcher should follow before going to collect and analyze data: (i) define your question; (ii) define your variables and (iii) define a protocol. These steps must be properly defined as they will determine what are we studying and what and how are we going to collect. This is extremely relevant as it allows researchers to focus on their specific study, rather than dispersing over the high amount of information that is available nowadays. Also, during fieldwork, it can be very useful to focus on the specific variables to collect, rather than start to collect everything that seems to be affecting our study species. Moreover, it is crucial to have in mind how are we going to analyze our data, while defining these steps. Therefore, it was taught that before sampling, we should also know how we are going to analyze samples.

During day 1 and before jumping to each step in detail, we reviewed some basic concepts while using practical examples of simple ecological questions. The first thing was to identify which were the independent and dependent variables in a practical example, as well as the different types of each variable (i.e., quantitative continuous, ordinal discrete and nominal discrete). Afterwards, we talked about two types of experiments: observational – where there is no interference in the conditions and are typically associated with field experiments – and manipulative – where the conditions are controlled, and they are more usual in lab experiments. After reviewing these concepts, we were ready to understand how we should collect the data. Usually, ecological and evolutionary studies do not sample the whole population but rather part of it. However, we should ensure that our sample is representative of the whole population, so that our inferences may be applied at the population level or even at a species level. To accomplish this properly, the first thing to do is to define our sampling unit (an event of data collection that generates one value), and it was highlighted that different questions will probably have different sampling units. For instance, if we want to see how does duration of the pupation of a butterfly (independent variable) affects the wing length (dependent variable), our sampling unit will be one wing of each butterfly. On the other hand,

if we want to understand how do butterfly densities (dependent variable) vary regarding to forest cover (independent variable), our sampling unit will be an area or a defined plot. Moreover, we talked about replication – repetition of the sampling unit- and its importance to sample error estimation. This is very important as it allows us to understand how “far” we can interpret our results. Sample size was also discussed. It was argued that the sample size must have into account the size and variability of the tested population, and there may be trade-offs that one shall make in order to have a sample size big enough to represent such populations parameters, but not to big to the point where there is not much information added but rather a waste of resources.

We then discussed some sampling methods, as well as the main advantages and disadvantages and applied examples of each of those methods. We started by discussing randomization of our sampling design, which may be good to minimize the effects of non-controlled factors and to ensure the independence between the data. Another method is the stratified sampling, which is a good method for comparing species ecologies between different habitats. Here, intra-strata variability must be low and inter-strata variability must be high. Within this method, there are different types of allocations (number of samples per strata) that can be made, depending on our research question and study area. There can be equative allocation (equal number of samples per strata), optimal allocation (equal variance between samples) and allocation proportional to the dimension (more samples if the area is bigger). Cluster sampling is another method discussed during the first day of the course. It is similar to the stratified sample, but we define clusters, and each cluster should be representative of the whole population. We discussed systematic sampling, where sampling is regular in time and/or in space. Despite not being random, we can randomize the order that each sample is taken, and therefore account for external variation. Finally, we also talked about distance sampling, which is commonly used for studying abundance and densities and it consists in sampling across line transects and/or point transects. Afterwards, there were a series of examples where we could apply the concepts learned during the theoretical part. We discussed the importance of a proper control to make sure that our conclusions are reliable and the concept of blocks which is when we organize our experiment in space and in time in groups similar between themselves. Regarding blocks, we analyzed how confounding effects may affect our results. Confounding effects are related to interactions not previously predicted by us.

We had the opportunity to also discuss a factorial design in our experiment. This design is used when we want to investigate the effect of two or more independent variables on the dependent one. Each independent variable as more than one level which in turn leads to an orthogonal design.

For instance, if we want to know if a plant inhabiting polluted areas are more tolerant to pollutants, we firstly need to sample environments with and without pollutants. Inside of each environment we shall have two locations and inside of each location two/three plan species. Within this example, the independent variable is the pollution level and the dependent one is the species tolerance. The environment may be used as a fixed factor and the location as a random factor in our analysis. We then discussed several scenarios (e.g. lack of time, lack of money, etc.) that imposed us restrictions in the sampling, and we needed to cut somewhere in order to answer to the same question. We also discussed the nested design, where one variable is always “inside” of the other. Taking the example of plant tolerance to pollution, if we sample location A in a polluted environment and in not polluted environment in day 1; and sample location B in a polluted environment and in not polluted environment in day 2, then each location is nested inside of each day and we would not be able to compare statistically locations among them, as different days would have different conditions and thus our results may be caused by these differences between days rather than between locations (which we do not control). However, one alternative would be to use the day as random factor, so that our model may “know” that the different days may affect the results. We then did several exercises where we could apply these concepts and discuss alternatives to improve some experimental designs.

We finished day 1 by understanding which are the “ingredients” to calculate sample size. The first to have in account will be the effect size (magnitude of the effect under the alternative hypothesis). The effect size is the difference between the mean response of the control and the mean response of the treatment, divided by the standard deviation. Usually, the higher the effect size, the easier it gets to detect it. The problem is that usually we do not have this effect size, but we can use previous/pilot studies or even use pre-defined approximations for each statistical test (each test will have different effect sizes). Also, in order to calculate the sample size, we have two more parameters that are independent of the sampling. The first one is the power, which is the probability of rejecting the null hypothesis when it is false. The higher the power the more accurate is the rejection. The second one is the significance level, which is the probability of rejecting the null hypothesis even if it is true (usually it is 0.05). We then learned how to calculate the sample size by defining threshold for the abovementioned parameters in R software.

We had the opportunity to read and interpret an article published in 2000 and criticize it, particularly the methods section in light of what was learned on that day.

**Day 2**

We started day 2 of the course by reading and criticizing an article, based on what we had discussed during these two days. This is a crucial step in a research career: do not take what other authors write for granted. In my group, we were able to identify some problems regarding the methods, particularly the insufficient and unbalanced number of replicates and lack of independence between samples, which in turn affected the conclusions that the authors made. By the end of that day, we presented our ideas and identified the problems with this article and discussed them with the lecturers and colleagues.

**Day 3**

During day 3, we addressed reproducibility in science. We defined it as the capacity by other researchers to obtain the same results as we did, by reproducing the same methods that we did. In order to have reproducibility in science, we need to explain in the highest detail possible the experimental setup (variables, locations, times, etc.) the collection of data (i.e., instruments used, methodologies, etc.), data processing and data analysis (code, program, etc.). This of extreme relevance as it allows transparency and confidence on the conclusions and reviewers can judge appropriately if our experiment was well conducted. Reproducibility is also important for ourselves, as many times there is a time lag between analyzing data and publishing. So, we need to ensure that we understand what we did months before and also, it can be useful to reuse several steps and analysis in order to save time. We then distinguish between reproducibility – when someone uses the same data, workflow, programs and analyses and reaches the same results – and replicability – when someone performs an independent study with similar experimental design and reaches the same results.

Afterwards, we reviewed the steps that make our research reproducible, both for ourselves and for others. The first step is to annotate, where we should write the explanations of each step of our workflow. This can be done by using notebooks and version control. Next step will be to automate. By automating our workflow, we avoid manual manipulation as it is more prone to errors. This may be attained using notebooks, pipelines or scripts. Following this, we need to share our work

through work collaborative environments or even among our colleagues, through online repositories, code (e.g., GitHub) and papers (e.g., bioRxiv). Finally, the last step is to store the information by doing regular backups with the maximum detail possible. This can be done by using backup software, version control (e.g., git) and online repositories (e.g., GitHub).

Afterwards, we went the back to the same article of day 2 and tried to see if the workflow followed by the authors was reproducible. We found some problems, as the authors did not provide the data nor the code that they used to reach their conclusions.

**Day 4 & 5**

During day 4 of the course, we learned through several exercises how to use RMarkDown in R. We also did some group exercises using git repository and GitHub in order to understand how these repositories work. Particularly, we created repositories with a Version Control, and used branches so the other members of our work could work on the same documents.

The final day of the course (Day 5) was dedicated to applying all the concepts learned during the course. This exercise was divided in two distinct tasks: (i) planning an experiment based on pre-defined questions, identifying the variables to be measured and a calendar with all the activities needed; and (ii) creating a git repository with all the information from task one.