Applying Classic Concepts of Experimental Design in the Age of Artificial Intelligence

Benton N. Taylor[1][0000-0002-9834-9192] and Andrew B. Davies[1][0000-0002-0003-1435]

1 Harvard University. Department of Organismic & Evolutionary Biology. Cambridge, MA, USA
bentontaylor@fas.harvard.edu

Abstract. As any field advances along the three parallel axes of theory, data, and synthesis, progress along any one of these axes can only outpace the others for a limited time before it must wait for progress along the other axes to catch up. We find ourselves at a time in history where major advances in our ability to gather empirical data and handle those data computationally have, in many fields, created an overabundance of information (data) but a scarcity of knowledge (synthesis). Artificial Intelligence (AI) holds immense promise for uncovering patterns in these data that allow us to make inference about the world around us that would be otherwise impossible. However, patterns occur in data for many reasons, some of which can lead to incorrect or misleading inference if the data were collected in a manner inappropriate for the question of interest. The field of experimental design aims to structure data collection in a way that maximizes the ability of the researcher to make appropriate inference from the data collected, given the question or objective of the study[1]. There are numerous texts devoted to designing effective experiments, many of which are written to provide pertinent advice in a specific field of interest. In this chapter, we cover fundamental concepts of experimental design and provide guidance for how to use and incorporate these concepts into analyses harnessing AI.

Keywords: randomization, replication, significance, study types, variables.

Glossary

Causal Inference: In general terms, the ability to determine that one variable is determining the patterns in another variable (rather than the two variables simply being correlated with one another). This term is also used for the growing field of statistics focusing on new ways to determine direct relationships of cause and effect in data.

Dependent Variable: The variable(s) in a study for which the researcher assesses their response to the independent variable(s). The dependent variable(s) is assessed as a function of the independent variable(s). Also referred to as the "response variable."

Experimental Block: A set of experimental units that are grouped based on similarity in a confounding factor within the block. By replicating experimental units across multiple blocks, one can use a blocking term to account for the confounding factor in a statistical model.

Independent Variable: The variable(s) in a study that can change independently independent of any other variables. In manipulative experiments, researchers actively control the independent variable(s) in order to determine the effects on the dependent variable(s). Also referred to as the “driver variable.”

Manipulative Experiment: A study in which the researcher actively manipulates the independent variable(s) in order to evaluate effects on the dependent variable(s). Also referred to as a “controlled experiment.”

Observational Study: A study in which the researcher measures the dependent variable(s) across naturally occurring variation in the independent variable(s) of interest. Also referred to as a “mensurative experiment.”

Psuedoreplication: Experimental data for which replicates are not statistically independent from each other or experimental designs for which treatment groups are not replicated but sampling within each group is replicated.

Sampling Grain: The spatial/temporal size of the smallest unit of measurement in a study. This can be considered in several dimensions, such as spatial, temporal, demographic.

Sampling Scale: The full extent of the sampling design for a study. This can be considered in several dimensions, such as spatial, temporal, demographic.

Statistical Power: The probability that a statistical test correctly rejects the null hypothesis. In more general terms, the probability that a researcher correctly identifies a pattern in the data.

1 The Fundamentals of Experimental Design

The purpose of experimental design is to develop a method for collecting data that focuses as precisely as possible on the variables of interest in the study while eliminating or accounting for variation in all other variables. With this in mind, the success of any experimental design is first determined by the clarity of the questions, objectives, and hypotheses of the study [2]. The researcher needs to begin by asking questions such as: *Which variables are of interest? Which additional factors that influence your variable of interest do you want to explicitly assess, and which do you simply want to account for their impact? At what spatial and temporal scales do your variables operate and at what scales do you hope to make inference?* Answers to these questions will help frame the questions and objectives of the study.

The next step is to clearly define hypotheses – both scientific and statistical. Scientific hypotheses take the form “I hypothesize X will have a positive influence on Y” or “I predict that group 1 will exhibit Y more than group 2” or some similar statement. In classic frequentist statistical frameworks, the statistical hypothesis converts this scientific hypothesis into something that is testable with probability theory. Thus, the above scientific hypotheses are converted to statements such as “I hypothesize that the slope of the relationship between X and Y is different than 0” or “I predict that groups 1 and 2 will differ in characteristic Y more than can be accounted for by chance alone.” More commonly, statistical hypotheses are actually formed in pairs – null and alternative hypotheses – where the null hypothesis would be “differences in variable Y between groups 1 and 2 are no greater than we would expect by random chance” and the alternative hypothesis would be “differences in variable Y between groups 1 and 2 are too large to be accounted for by random chance alone." (See Lehman & Romano[3] for a detailed description of forming effective statistical hypotheses.) While the distinction between scientific and statistical hypotheses may seem subtle, the statistical hypothesis focuses solely on patterns in the data, not the real-world mechanisms driving those patterns. Converting the patterns tested by the statistical hypotheses into answers that support or refute your scientific hypothesis is the job of scientific inference, which depends directly on the quality of your experimental design [4]. Thus, the statistical hypothesis provides the link between the conceptual question of the study and the experimental design that will generate the data to answer that question. With questions and hypotheses in hand, the next step is to define the type of study best suited to test the hypotheses and answer the questions.

1.1 Major Study Types

We find a helpful initial dichotomy for categorizing scientific studies is defining whether to conduct a manipulative (controlled) experiment or a natural observation study (often referred to as a mensurative experiment). In a manipulative experiment, the researcher typically establishes a controlled environment that minimizes variation in outside variables, then explicitly alters the independent variable(s) of interest to assess their influence on the response variable(s). In observational studies, the researcher assesses changes in the response variable(s) across naturally occurring gradients or between groups that differ in the independent variable(s) of interest. One of the most important differences that this distinction creates is whether the researcher accounts for outside variables before data collection (*i.e.* controlled experiments) or after (*i.e.* observational studies). The other primary difference between these study types is that in observational studies, the scale of the independent variable(s) represents at least some portion of the naturally relevant range whereas in manipulative experiments, the researcher can apply differences in independent variables well outside the naturally relevant range. Understanding these differences can guide decisions on which variables need to be controlled for during analysis and the appropriate level of inference that can be made from the data.

Another major distinction in study types focuses on the type of data that comprise the independent variable(s), as this distinction has direct impact on the type of statistical analysis that will eventually be applied to the data. In the broadest sense, studies can be separated into those for which the primary independent variable is categorical vs. those where the independent variable consists of continuous data. This distinction broadly separates the type of statistical test used between tests comparing differences between groups for categorical independent variables *vs.* tests using regression analysis to assess change across a gradient of a continuous independent variable (these tests are actually very closely related, but we find the conceptual delineation is often helpful). It is important to note, however, that the categorical *vs.* continuous nature of data can be somewhat interchangeable. Some categorical independent variables follow an intrinsic order (ordinal data), which can enable the researcher to interpret results across a gradient of ordered categories analogous to a gradient of continuous data. It is also possible to convert continuous independent variables into categorical data by binning continuous data and comparing differences between bins. Taken together, we suggest that the type of data that make up the primary independent variable has a strong influence on the type of analysis conducted but that there are numerous ways to use categorical data in semi-continuous ways and to bin continuous data into discrete categories – steps that can often occur long after data collection is complete.

1.2 Data Considerations

Once the general type of study and the form of data for each variable have been established, a researcher must turn to several considerations of the prospective data to inform exactly how to design the study. First is the question of scale and grain. Scale refers to the full extent of the system you plan to study – how large an area, how long a time scale, how broad a demographic group, etc. Grain refers to the size of the smallest unit of measurement – typically the size (spatial, temporal or otherwise) of an individual replicate in the data. These two aspects of the data will determine the level of detail you can gain on your study system (grain) and the extent to which you can generalize inference from your study system to the broader population (scale). With this in mind, studies that employ the smallest grain and largest scale possible will maximize both the level of detail and the extent of inference in their results. However, constraints on time, effort, and funding often preclude simultaneously reducing grain and increasing scale, so researchers are often faced with optimizing the tradeoff between the two (increasing grain in order to increase scale or reducing scale to allow for smaller grain).

Determining the appropriate scale and grain of data collection requires the researcher to consider the goals of the study, the logistical constraints of data collection within the study system, and the potential variability of the data to be collected [5]. If the goal of the study is to infer broad patterns, maximizing scale should be the highest priority, but if the goal is to accurately characterize the details of patterns in the data, then minimizing grain should be prioritized. To some degree, the study system itself dictates appropriate targets and limits to scale and grain. For example, using a spatial grain of a 1m x 1m sampling plot is far too small for a study on elephants but may be entirely appropriate for a study on soil bacteria. Finally, estimating the potential variability in the data that will be collected can help constrain appropriate grain and scale. If you anticipate high variability in your data relative to the strength of the pattern that you are trying to assess, decreasing grain and reducing the scale of the study to increase replication may be the best course of action. Together, scale and grain make up important considerations for how to structure data collection in your study. The next decisions are concerned with what to sample.

1.3 Independent variables, dependent variables, and confounding factors

While the scope of this chapter is sufficiently broad to preclude discussion of particular approaches to data collection, several general considerations may be useful as you plan your data collection. One of the most common mistakes that researchers make is not fully anticipating the needs of their statistical analyses prior to data collection. Foreseeing the exact columns of data for independent and dependent variables allows the researcher to ask such questions as: *What are the most likely sources of error/uncertainty when collecting each variable? If my variables are calculated metrics, do I have all the raw data necessary to make those calculations? Do independent and dependent variables need to be collected at the same scale?* Asking these questions before data collection can dramatically improve the efficiency and success of a study. To this end, we strongly suggest constructing a mock dataset with each variable of interest and running initial statistical tests to determine if any additions or adjustments to your data collection plans are needed. With these independent and dependent variables of interest planned, attention can then turn to anticipating confounding factors in your study system.

One of the most difficult but critical aspects of effective experimental design is the identification of potential confounding factors in your study system – factors that are not of direct interest to the study but that are correlated with variables of interest and thus may influence their behavior. For example, imagine a hypothetical study investigating the link between alcohol consumption and heart disease. Because alcohol consumption is correlated with the likelihood of smoking cigarettes and cigarette smoking impacts heart disease, the direct relationship between alcohol consumption and heart disease is confounded by the effect of cigarette smoking. That is, some portion of the perceived impact that alcohol consumption has on heart disease is actually the impact of smoking, which just happens to be correlated with alcohol consumption. In this hypothetical case, the relationship between alcohol consumption and heart disease would differ between a group of individuals that smoke and those that do not. These types of confounding factors are extremely common and, if not accounted for, can dramatically impact the results of a study. Fortunately, numerous approaches have been developed to deal with confounding factors before and after data collection.

Ideally, the researcher can identify important confounding factors ahead of time and design their study to account for these factors prior to data collection. In manipulative experiments, confounding factors are often negated by creating tightly controlled environments where only the independent variable(s) of interest are allowed to vary. This approach, when done well, eliminates confounding factors and assigns any change in the dependent variable to variation in the only factor that is allowed to vary – the independent variable. Observational studies are typically conducted in less controlled settings, but confounding factors can often still be addressed prior to data collection. One of the most common and effective ways to deal with confounding factors in observational studies or field experiments is to create experimental blocks [6], each of which contain a set of replicates that are confined to the same condition of the confounding factor (also see Gotelli & Ellison 2004 for alternative experimental block designs). By establishing multiple experimental blocks and including this blocking factor in statistical models, the researcher can account for any variation between the blocks caused by differences in the confounding factor, thus statistically isolating the relationship between independent and dependent variables. Finally, if your replicates span across the range of a continuous confounding variable, you can take data on this variable and include it as a factor in your statistical model to account for its confounding effect.

Accounting for confounding factors becomes extremely difficult, and often impossible, if data on the confounding factor are not collected during the study. It is occasionally possible to obtain data on a confounding factor from outside data sources (*e.g.* publicly available data for climate, human demographics, economics, etc.) post-hoc as long as the data are of the appropriate spatial/temporal scale to match with your experimental data, but matching data from different sources in this way can lead to unforeseen issues in data comparability. As a general rule, sophisticated statistics to account for confounding factors post-hoc are a poor replacement for a well-designed study that accounts for the confounding factors prior to data collection.

1.4 Replication Do's and Don'ts

One of the most common questions that arises as researchers design experiments is *How many replicates do I need?* Unfortunately, there is no one right answer to this – the required number of replicates for a study depends on the type of analysis that will be used, the variability in the broader population of observations, and the size of the effect that is being assessed. While larger sample sizes will always provide a better representation of the population that is being described, logistical constraints typically force researchers to compromise between the number of variables to consider, the number of treatment groups (for categorical independent variables), the range of the population to sample, and the number of replicates at each level of sampling. A straightforward, pragmatic approach to these decisions is to establish the total number of samples that funding and logistics will allow you to collect and then decide how to distribute those across your experiment (number of treatments, replication within each treatment, etc.). These compromises are unique to each study, but in our experience, a common replication pitfall is for researchers to include too many treatments and/or independent variables at the expense of replication within each group. The danger of this decision is that it often leads to insufficient sample sizes to make conclusions about any of the treatments or variables. Thus, when facing these decisions, we think it a useful exercise to ask oneself *Is this additional treatment that I would love to include in my experiment worth the risk of the entire experiment being invalidated due to lack of replication?* Fortunately, tools called Power Analyses have been developed to determine necessary sample sizes *a priori* [7,8], but these analyses require the researcher to make reasonable estimates of the effect size and variability in the data that they are likely to encounter.

Beyond the question of how many replicates to sample, there are different ways to structure replication in an experiment - some good, some bad. A critical underlying assumption of all traditional frequentist statistics is that samples are independently pulled from the larger population of interest. Meeting this assumption is often harder than it first may seem (and many poorly designed studies fail to do so). A common culprit is that samples can be autocorrelated – that is, two samples are more similar than we would expect from random chance, typically due to proximity in space and/or time. For example, imagine a researcher interested in the relationship between sunlight and photosynthesis throughout a forest who has the logistical capacity to sample 100 leaves. Sampling a single random leaf from 100 trees represents independent samples, but if, in an effort to save time, the researcher samples 10 leaves from 10 trees, then the samples from within a single tree are likely to be more similar to each other for reasons other than just the amount of sunlight they receive. This describes a scenario of pseudoreplication [9], where despite having taken 100 leaf samples, the researcher has actually only engaged in 10 independent sampling events. In this case, the best course of action would be to take average values for the 10 leaves on each tree and use those averages for their statistical analyses – resulting in a sample size of 10 rather than the original 100 samples taken. (It is important to note, however, that a single leaf from an entire tree may not be a representative sample of that tree, so the most ideal approach in this example is likely to take multiple leaves from as many trees as possible within the relevant logistical constraints.) In general, the best approach to avoiding pseudoreplication is to structure sampling such that each sample is as independent as possible even if that makes the sampling process more logistically challenging. Randomly choosing samples from the larger population (another underlying assumption of frequentist statistical models) often helps decrease the risk of autocorrelation in your sampling effort.

To the new experimentalist, the myriad decisions to consider when designing an experiment can be overwhelming. We suggest a few guiding principles that can help avoid common pitfalls. 1) The more specific you make your hypotheses, the easier it will be to design an experiment to test those explicit expectations. 2) When possible, plan your statistical analyses while designing your experiment so that you know exactly what information you will need to take during the study. 3) Take time to identify all potential confounding factors and make a plan to control or account for them. 4) Randomize everything unless there is a strong reason not to – this includes taking samples from a population, the spatial arrangement of experimental units, the timing of sampling, etc. 5) More replicates in fewer treatments is generally better than fewer replicates in many treatments. In addition to these principles, interactive tools such as the Experimental Design Assistant [10] have been developed to help researchers design effective experiments, plan statistical analyses, and calculate required sample sizes.

2 Applying Artificial Intelligence to Experimental Data

2.1 Understanding Your Data

When an early-career experimentalist sets out to test a new question, a common piece of advice is for them to learn as much as they can about their study system ahead of time, as this knowledge almost always improves their experimental design. Understanding their study subject or the environment in which that subject exists helps the experimentalist streamline data collection, predict variability and necessary sample sizes, and identify potential confounding factors in their experiment. For the early-career data analyst using AI to ask questions with an existing data set, the analogous advice is to become as familiar with the dataset as possible before asking the scientific question of interest. Understanding how the data were collected, the quality of each variable in the data set, and the structure of those data is critical to the effective implementation of AI on a data set. A classic example of the value of understanding one's data is Simpson's Paradox, where the overall pattern in a data set is completely reversed when assessing the same pattern in subsets of the data. For example, even if tax rates for all income classes go down over a given period of time, the overall tax rate of the entire population can (and often does) actually increase so long as a larger portion of the population moves into higher tax brackets during the time period [11]. Similar oddities and apparent paradoxes of patterns are more common in data sets than one might think, and spending as much time as possible probing a dataset for these patterns and abnormalities at the outset of a study is often critical for interpreting the eventual results of AI analyses. Even in relatively straight-forward data sets, understanding the quality of data collected for each variable and the likely uncertainty in those data can help the researcher propagate that uncertainty in their final analysis. For these reasons, we strongly encourage researchers to thoroughly probe and vet their data sets, even (and maybe especially) if they did not directly collect the data themselves.

2.2 Statistical vs. Scientific Significance

When applying AI approaches to empirical data, one of the most important distinctions to consider is the difference between statistical significance (*i.e.* verifiable non-random patterns) and scientific significance (*i.e.* patterns that have real-world importance)[12]. Statistical significance is a valuable way for researchers to evaluate the probability that patterns in their data are not simply due to random chance. However, many studies find patterns in data that, while statistically significant, are so slight that they have no real-world importance. Other studies find patterns in data that have important real-world implications but are not statistically significant - often due to high variability in the data and insufficient sample sizes – warranting further data collection to verify the results. We often find that interrogating effect sizes first and P values second can be a helpful way to maintain focus on the scientifically relevant results in your analyses. Understanding what influences statistical and scientific significance can help you weigh the relative importance of each [13]. While the assignment of statistical significance varies substantially between statistical frameworks, three factors strongly influence statistical significance: the strength of the pattern (effect size), variability in the data, and sample size. Strong patterns and large sample sizes paired with low variability in the data all increase the probability of accurately identifying a pattern in the data - a probability referred to as "statistical power."

In the era of "big data," one of the most useful applications of AI is building algorithms to detect patterns in datasets characterized by many explanatory variables and massive numbers of observations. But with great computational power comes great statistical and scientific responsibility. Having access to many thousands or millions of observations that can have information for dozens of variables often makes it very easy to find statistical relationships between those variables – even if those relationships are scientifically irrelevant, driven by confounding variables, or simply represent autocorrelation in the data. Given these risks, we strongly recommend identifying targeted questions and hypotheses – even for analyses on data collected by other researchers – rather than looking for all possible correlations in a dataset and then creating mechanistic explanations for those patterns post-hoc. As discussed above, we highly recommend becoming as familiar as possible with both the study system and the method of data collection when analyzing data collected by others, as this can often help you determine the appropriate applications and limitations of each variable in the dataset. Similarly, for researchers using AI on data collected by others, a deep understanding of the study system can help avoid drawing erroneous conclusions from patterns in a dataset. For these reasons, we recommend that data analysts using AI collaborate with those who collected the data or with experts in the field of their study system.

2.3 Determining Pattern vs. Process

AI has become one of the most powerful tools at researchers' disposal to detect patterns in data. These patterns are often the first critical step in a line of inquiry, and in many cases, AI detects patterns in data that the researcher had not initially been looking for, leading to entirely new pathways of investigation. Patterns in data are the primary way that researchers make sense of their study system, and the ability of AI to detect these patterns – even in highly complex data sets – makes AI a valuable tool in almost any field of inquiry. For many researchers, the next logical step is to identify the processes and mechanisms underpinning the patterns observed in their data. Isolating mechanisms and identifying processes in a study system are tasks for which AI is notably less well suited. The impressive ability of AI to search for correlations between variables in a data set requires the researcher to vigilantly remember the cornerstone adage of scientific inference: "correlation does not equal causation." In our experience, researchers are often tempted to assign processes to the patterns in their data based on assumptions and the researchers' own knowledge of their study system. While knowledge of one's study system is often a researcher's most valuable asset, this knowledge should be paired with rigorous assessments of causality if one wishes to make strong inference about the processes driving patterns in their data.

Assigning causality and process to patterns in data is both a pillar of classical statistics and an area of rapidly new development – largely due to the emergence of AI. The classic approach to determining process in data is to conduct a manipulative experiment. One of the most effective implementations of this approach is to pair an observational study that assesses naturally occurring patterns in one's study system with targeted manipulative experiments designed to isolate the mechanisms and processes underlying those patterns. For many, this method of paired observational and manipulative studies still serves as the gold standard for inferring the link between pattern and process. Yet, this approach often suffers from drawbacks of being logistically challenging and yielding relatively small data sets, and only well-designed manipulative experiments can conclusively determine process. Moreover, in many cases the researcher is working with data for which it is impossible to conduct follow-up manipulative experiments.

In recent decades, the emergence of the statistical field of Causal Inference has added a valuable new set of tools for determining process behind the patterns in data[14]. Largely based on the fundamentals of Baye's Theorem, methods of Causal

Inference allow a researcher to identify a set of mutually exclusive hypotheses for different processes that could underpin patterns in their data, and to then determine the probabilities that each hypothesized process is the true driver of the patterns they observe. This approach has the notable advantage of not requiring additional experimentation – a researcher can often apply methods of Causal Inference to a data set that is already in hand. Because this approach relies upon searching for and comparing patterns in a dataset, AI has become an increasingly common method for assessing Causal Inference.

3 Conclusions

As AI becomes an increasingly widespread tool for data collection and analysis, it becomes increasingly important to bridge the concepts of strong experimental design with the analytical power of AI. Even for researchers using data collected by others, having a strong understanding of the principles of experimental design can help one better understand the structure, possibilities, and limitations of the data, and can help avoid making false inference from patterns in the data. Almost all general texts on experimental design highlight the inextricable relationship between designing an experiment and analyzing the data. Such statements are often aimed at the empiricist that hopes to start their experiment before learning the fundamentals of data analysis; however, this concept is equally true for the sharp data analyst that hopes to dive into a dataset before learning about the design of the study from which the data came. A strong understanding of both experimental design and AI provides the researcher an incredibly powerful tool set for identifying patterns and making accurate scientific inference in the big data era.

References

1. Myers, J. L. Fundamentals of Experimental Design. (Allyn & Bacon, 1972).
2. Gotelli, N. J. & Ellison, A. M. A Primer of Ecological Statistics. (Sinauer Associates, 2004).
3. Lehmann, E. L. & Romano, J. P. Testing statistical hypotheses. (Springer, 2005).
4. Easterling, R. G. Fundamentals of statistical experimental design and analysis. (John Wiley & Sons, 2015).
5. Casler, M. D. Fundamentals of experimental design: Guidelines for designing successful experiments. Agron. J. 107, 692–705 (2015).
6. Addelman, S. The generalized randomized block design. Am. Stat. 23, 35–36 (1969).
7. Kraemer, H. C. & Blasey, C. How many subjects? Statistical power analysis in research. (Sage Publications, 2016).
8. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. (Taylor & Francis, 1988).
9. Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54, 187–211 (1984).
10. Percie du Sert, N. et al. The Experimental Design Assistant. PLoS Biol. 15, 1–9

(2017).

11. Wagner, C. H. Simpson's paradox in real life. Am. Stat. 36, 46–48 (1982).
12. Rothman, K. J. Disengaging from statistical significance. Eur. J. Epidemiol. 31, 443–444 (2016).
13. McShane, B. B. & Gal, D. Statistical significance and the dichotomization of evidence. J. Am. Stat. Assoc. 112, 885–895 (2017).
14. Pearl, J., Glymour, M. & Jewell, N. P. Causal inference in statistics: A primer. (John Wiley & Sons, 2016).