



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 172*

Causal Inference in Observational Studies and Experiments: Theory and Applications

MÅRTEN SCHULTZBERG



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019

ISSN 1652-9030
ISBN 978-91-513-0762-6
urn:nbn:se:uu:diva-393810

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, 6 December 2019 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Peng Ding (Department of Statistics, University of California, Berkeley, US).

Abstract

Schultzberg, M. 2019. Causal Inference in Observational Studies and Experiments: Theory and Applications. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 172. 36 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0762-6.

This thesis consists of six papers that study the design of observational studies and experiments.

Paper I proposes strategies to consistently estimate the average treatment effect of the treated using information derived from a large number of pre-treatment measurements of the outcome. The key to this strategy is to use two-level time-series model estimates to summarize the inter-unit heterogeneity in the sample. It is illustrated how this approach is in line with the conventional identifying assumptions, and how sensitivity analyses of several key assumptions can be performed.

Paper II contains an empirical application of the identification strategy proposed in Paper I. This study provides the first causal analysis of the demand response effects of a billing demand charge involuntarily introduced to small and medium sized electricity users.

Paper III proposes strategies for rerandomization. First, we propose a two-stage allocation sample scheme for randomization inference to the units in balanced experiments that guarantees that the difference-in-means estimator is an unbiased estimator of the sample average treatment effect for any experiment, conserves the exactness of randomization inference, and halves the time consumption of the rerandomization design. Second, we propose a rank-based covariate-balance measure which can take into account the estimated relative weight of each covariate.

Paper IV discusses the concept of optimal rerandomization. It is shown that depending on whether inference is to be drawn to the units of the sample or the population, the notion of optimal differs. We show that it is often advisable to aim for a design that is optimal for inference to the units of the sample, as such a design is often near-optimal also for inference to the units of the population.

Paper V summarizes the current knowledge on asymptotic inference for rerandomization designs and proposes some simplifications for practical applications. Drawing on previous work, we show that the non-normal sampling distribution of the difference-in-means test statistic approaches normal as the rerandomization criterion approaches zero. Furthermore, the difference between the correct non-normal distribution and the proposed approximation based on a normal distribution is in many situations negligible even for near optimal rerandomization criteria.

Paper VI investigates and clarifies the relation between the traditional blocked designs and rerandomization. We show that blocking and rerandomization is very similar, and in some special cases identical. Moreover, it is shown that combining blocking and rerandomization is always at least as efficient as using only rerandomization, but the difference is in many cases small.

Keywords: experimental design, identification, observational studies, rerandomization

Mårten Schultzberg, Department of Statistics, Uppsala University, SE-75120 Uppsala, Sweden.

© Mårten Schultzberg 2019

ISSN 1652-9030

ISBN 978-91-513-0762-6

urn:nbn:se:uu:diva-393810 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-393810>)

Dedicated to my parents

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I Schultzberg, M. (2018) Using high frequency pre-treatment outcomes to identify causal effects in non-experimental data.
- II Öhrlund, I., Schultzberg, M. Bartusch, C. (2019) Identifying and estimating the effects of a mandatory billing demand charge.
- III Johansson, P., Schultzberg, M. (2019) Rerandomization strategies for balancing covariates using pre-experimental longitudinal data.
- IV Johansson, P. Rubin, D. B., Schultzberg, M. (2019) On Optimal Rerandomization Designs.
- V Schultzberg, M., Johansson, P. (2019) Asymptotic Inference for Optimal Rerandomization Designs.
- VI Schultzberg, M., Johansson, P. (2019) Rerandomization: a complement or substitute for stratification in randomized experiments?

Reprints were made with permission from the publishers.

Contents

1	Introduction	9
1.1	Why we need the scientific methodology	9
1.2	Experiments	10
1.3	Experiments in complex settings	11
1.3.1	Issues specific to studying humans	12
2	Statistics - A toolbox for separating coincidental spurious relations from causal relations	13
2.1	Potential outcomes and the need for randomized experiments ...	13
2.1.1	Common estimands and estimators	14
2.1.2	The design steps	15
2.2	Observational studies - Trying to deterministically recreate the benefits of randomization	17
2.2.1	Natural experiments	18
2.2.2	Rubin causal models	19
2.3	Experimental Design - Maximizing the utility of each study	21
2.3.1	Blocking	21
2.3.2	Rerandomization	22
2.4	Big data - A blessing and a challenge for the design of scientific studies	24
3	Summary of papers	27
3.1	Paper I	27
3.2	Paper II	27
3.3	Paper III	28
3.4	Paper IV	29
3.5	Paper V	30
3.6	Paper VI	30
3.7	Summary	31
4	Acknowledgements	32
	References	35

1. Introduction

1.1 Why we need the scientific methodology

All humans have a concept of causality, at least in an everyday sense, such as, e.g., ‘If I drink less coffee I’ll sleep better.’. However, the origin of this common notion and the understanding of everyday causality are topics of debate in the field of psychology. Some researchers have claimed that this understanding is innate, while others have suggested that it is something we must learn, or at least discover, through empirical observation (Waldmann, 2017; Leslie and Keeble, 1987). Without taking sides on that matter, it is clear that understanding – and thereby being able to predict with some accuracy – what will happen next if you make one choice as opposed to another is probably key to humans’ ability to plan for the future and alter outcomes in our favor. This is often also the goal of science: Scientists strive to systematically understand which causes and outcomes are related and how, such that we can alter the outcome to better suit our purposes.

Understanding the world is a challenging task, and it frequently happens that explanations that are taken for truths at some point in time are proven to be false on a later occasion. There are at least two reasons for why this happens. The first reason is obvious: The world is complex and difficult to understand and, therefore, easily misunderstood. Second, humans have a tendency to see systematic patterns in randomness, which sometimes leads us to believe that we understand things we in fact do not understand. This tendency is a challenge shared by all humans: During evolution, the human mind seems to have been hard-wired to come up with simple heuristics (Tversky and Kahneman, 1974) and to find and recognize patterns (Foster and Kokko, 2009) that can speed up decisions and actions in real time.

In psychology, there is a cognitive bias known as *apophenia* Mishara (2009). An instance of this cognitive bias is called an ‘apophany,’ which is perhaps easiest understood as a false epiphany, i.e., a false ‘aha’ moment. In other words, an apophany is when you perceive that you understand something, or see a meaningful pattern in something, when in fact your mind has only read in meaningfulness in a randomly occurring sequence of events. This bias is often attributed to humans’ common tendencies to, e.g., keep ‘lucky items,’ believe in paranormal phenomena Simmonds-Moore (2014), and believe in conspiracies Van Elk (2015). From a survival point of view, apophenia is not all bad: Failing to connect a cause with an outcome of great consequences can be devastating. For example, if you do not connect the cause ‘eating poisonous

berries' with the effect 'dying,' you might eat the berries even after watching someone else die after eating them. On the other hand, someone might die after eating perfectly healthy berries, by chance, in which case you would only miss out on a good meal by erroneously connecting the two events. The first mistake would always lead to your death, the second one might lead to your death if food is sparse, but with a probability less than one. For this reason, it makes sense from an evolutionary perspective to have a strong tendency to, when in doubt, favor classifying a relationship as causal rather than not (Foster and Kokko, 2009).

Because scientists are also humans (albeit, a special kind), apophanies also occur in their minds. Apophenia in a scientist, so-called *scientific apophenia*, has been studied by, e.g., Goldfarb and King (2013). Humans' proneness towards apophenia causes the scientist's mind to involuntarily fill in the gaps and perceive systematic patterns in coincidental, randomly occurring phenomena. In addition, this tendency is accompanied by many other cognitive biases, such as more general confirmation biases Nickerson (1998) and sunk-cost bias Thaler (1999); MacDonald et al. (2018), that make humans prone to also want to confirm nonsensical patterns once discovered, and then to keep investing time and effort in confirming them.

It is worth noting that these biases are usually not very consequential in the everyday life of humans, and when they are, the consequences are usually restricted to the person in whom the biases occurred. However, because the scientist's goal is to find systematic relations that can be used to inform policymakers in, e.g., medicine, construction, education, etc., these biases may have huge consequences and implications for many individuals. Despite all of these challenges, scientists have been able to make great progress in various fields by relying on *scientific methodologies*. The reason for the success of the scientific method is that it contains numerous checks and balances for the types of biases we humans are prone to have. In the following sections, some key parts of scientific methodology are presented.

1.2 Experiments

One of the most central concepts of the scientific methodology is experimentation. According to many famous researchers, experimenting – as a means of systematically mapping the world – is one of the most important discoveries made by humans. For example, Albert Einstein wrote “*Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance). In my opinion, one has not to be astonished that the Chinese sages have not made these steps. The astonishing thing is*

that these discoveries were made at all.” — Letter to J.S. Switzer, April 23, 1953; Einstein Archive 61-381.

The key aspect of experimenting is going back and forth between coming up with a theory that gives a falsifiable prediction and performing experiments that falsify or do not falsify that prediction. If the prediction is falsified, the theory must be updated. In this way, the direction of the theory is iteratively corrected by the experiments, and pursuing dead ends can be avoided. If the theory can make predictions that are narrow enough that experiments can be constructed to test them, this methodology is powerful enough to make great progress.

The idea of using experiments to confirm or disprove theories was adopted early in physics, where it was used to provide evidence for or against theories. One early example of this type of experiment is when, in 1609, Galileo Galilei found evidence for the heliocentric model, i.e., the notion that the earth orbits the sun. It is not a coincidence that experimentation occurred first in physics and not until much later was adopted by scientists in other fields. In physics, it is often possible to set up fairly simple experiments where the only thing that varies is the independent variable of interest. For example, to test predictions concerning the acceleration of gravity, it is fairly simple to come up with a setting in which few things other than gravity are affecting an object you drop from a certain height. In other words, it is fairly easy to come up with a study design that is robust to alternative explanations. The simpler the setting is, the less likely humans are to let the previously mentioned biases drive the conclusions, as there are fewer randomly occurring things to read patterns into. For this reason, in many of the early studies in physics and chemistry, statistical analysis was not a necessary tool for making progress; the predictions of the theories either worked or they did not, with little leeway for interpretation.

1.3 Experiments in complex settings

In areas such as medicine or psychology, setting up a credible experiment is far more difficult, especially when living organisms are involved in the experiment itself. The first problem with studying phenomena regarding living organisms and their behaviors is the often-present great natural variation. What causes an outcome in one observational unit may not cause it in another, e.g., due to genetic factors, environment, etc. That is, if you were to study another observational unit, different from the one you studied, you might come to another conclusion. If, for some reason, the studied observational units are very atypical, the observed results may be misguided and not apply to units outside the sample. These issues combined with apophania make it difficult to separate randomly occurring atypical results from systematic ones. If one is to make progress by experimenting in such settings, the natural variation among

units and in the outcome must be accounted for. This is exactly what statistics was proposed to do.

1.3.1 Issues specific to studying humans

Besides the great variation that occurs naturally in many phenomena in living things, there are other important issues specific to studying humans. One crux of doing experiments on humans is that they are aware they are being studied, and are affected by it; this is called the Hawthorne effect (McCarney et al., 2007). Moreover, subjects often want to have the treatment rather than placebo, which in some cases causes them disobey the protocols, i.e., they take the treatment even though they should not. Subjects might also influence each other. For example, if your friend is assigned to a particular treatment, you might be inclined to be assigned to the same treatment, etc. Several strategies have been proposed to circumvent these problems: blinding the treatment to the participant, blinding the treatment to the researcher, and recently, blinding the treatment to the data analyst, resulting in what are called triple-blind experiments (Schulz and Grimes, 2002). The treatment is blinded to the researcher and analyst as it is known that the researcher might otherwise affect the results, and an untrained analyst might make choices based on confirmation biases. One famous example is the results in Davenas et al. (1988) concerning homeopathy, which were proven false by Maddox et al. (1988).

In many settings, blinding is not possible for several reasons. For example, consider studying the effect of cognitive behavioral therapy. Clearly, the treatment cannot be blinded for the patient or the therapist. In such settings, experiments might not be the best tool for assessing treatment effects. I will return to this topic later, but to do so, a more formal statistical setup for causal inference is required.

2. Statistics - A toolbox for separating coincidental spurious relations from causal relations

From this point onwards, the focus of this thesis will be on causal relations. Naturally, there are other scientific questions for which statistics is crucial, but these are not discussed further here. Specifically, the attention will be restricted to the concept of a *causal effect* from an intervention, e.g. ‘treatment’ or ‘placebo’ in a medical trail, or the effect of a new labor market training program on the employment rate as compared to the current program.

2.1 Potential outcomes and the need for randomized experiments

As mentioned above, when studying more complex phenomena such as behavior in humans or animals, or the health effects of medicines, several challenges occur. Even if the external setting can be controlled to some extent, the changes in outcomes might depend on which observational units are being studied and which of them are receiving the treatment or control, respectively. Some units might have strong effects, others might have no effect, etc.

In the early 20th century, Neyman wrote his seminal paper proposing the concept of potential outcomes (Neyman, 1923; Rubin and Rubin, 2005), which has helped formalize these issues. The potential outcome framework is essential to being precise about the definition of a causal effect and the challenges of studying it. In the simplest case, with one treatment: The causal effect of a cause on an outcome is the difference between the outcome if the cause occurred and the outcome if the cause did not occur. The two outcomes are the *potential outcomes*. Let Y denote an outcome of interest, $Y_i(0)$ be the outcome that unit i will have if given no treatment, and $Y_i(1)$ be the outcome of unit i if given treatment. The causal effect of the treatment on the outcome Y for unit i is then

$$\tau_i = Y_i(1) - Y_i(0). \quad (2.1)$$

This definition directly opens up the possibility for causal effects to differ from unit to unit, i.e. $\tau_i \neq \tau_j$ for $i \neq j$. As a simple example, the cause of a medicine on your health is the difference in your health if you had been given the medicine and your health if you had not been given the medicine, i.e. the

difference between your potential health status with and without the medicine. The causal effect of the medicine on my health is likely different from the causal effect on yours.

Equation 2.1 implies that the individual causal effect may never be observed, as only one potential outcome can be observed at any time point. This is solved by moving to the group level, i.e. using the difference in the average of the potential outcomes of those treated and those not treated as an estimate of the causal effect. In a sample of units, the researcher can, e.g., assign half to treatment and half to control and use the difference in the mean of their outcomes as an estimate. However, this is when matters become somewhat complicated. Because both potential outcomes of all units are likely to differ, and perhaps be unique for to units, the estimate $\hat{\tau}$ will be highly depend on which units from the population are in the sample, and which units in the sample are assigned to treatment, respectively. These two steps, i.e., sampling subjects from a population and treatment allocation, are central to the validity, and therefore to the design, of any experiment. Before going through these steps properly, it is beneficial to set up some common estimands and their estimators formally.

2.1.1 Common estimands and estimators

The potential outcome framework makes it possible to precisely define the entities of interest that we wish to estimate, i.e., the *estimands*. For example, it is common to be interested in the Population Average Treatment Effect (PATE), which is the average of the individual causal effects of all units in a population of interest. Consider a population of N units, then the PATE is defined as

$$PATE = \frac{1}{N} \sum_{i=1}^N Y_i(1) - Y_i(0). \quad (2.2)$$

This effect is often of great interest to policymakers who are to decide on whether or not to adopt a policy; what would the average effect be on the outcome of interest if the policy were applied to the population? Sometimes it is useful to assume that the population is infinitely large, in which case the average is replaced by an integral over the distribution of the outcomes.

If a sample s of size n from the population is available for the experiment, where $n \leq N$, and we are only interested in the effect of the treatment in this sample, the Sample Average Treatment Effect (SATE) is defined as

$$SATE_s = \frac{1}{n} \sum_{i \in s} Y_i(1) - Y_i(0), \quad (2.3)$$

where s contains the indexes for the units from the population that are in the sample s . Note that SATE has a subscript, s , as it will vary depending on which

units are in the sample. For a population of size N , there are $\binom{N}{n}$ different SATE for sample size n .

Again, because only one of the potential outcomes of a unit can be observed at the same time, neither of these estimands can ever be observed directly. Instead, they are estimated, often by comparing the group means of the treated and the controls. This implies splitting the sample into two groups. Let the size of the treated and control groups be n_1 and n_0 , respectively, such that $n_1 + n_0 = n$. There are $\binom{n}{n_1}$ ways of splitting the sample. The difference-in-mean estimator is given by

$$\hat{\tau}_j = \frac{1}{n_1} \sum_{i:W_i^j=1} Y_i(W_i^j) - \frac{1}{n_0} \sum_{i:W_i^j=0} Y_i(W_i^j), \quad (2.4)$$

where W_i^j is a treatment indicator taking the value one if unit i is assigned treatment and zero otherwise, according to allocation j , where $j = 1, \dots, \binom{n}{n_1}$.

Note that this estimator is used for estimating both SATE and PATE. This might seem odd at first, given how different the estimands are. However, regardless of the estimand of interest, typically, only the information from the sample is available.

Returning to the last paragraph of the previous section, the observed estimate from a study will be one of the $\binom{N}{n} \times \binom{n}{n_1}$ possible estimates. Some of these estimates are close to its $SATE_s$, and some are close to $PATE$. In some cases, $SATE_s$ is close to $PATE$, in which case there are estimates that are close to both. In other words, some estimates are very typical, and/or describe the sample and or the population well, but other estimates do not. To say something about the properties of this estimator, we must know how the sample is selected and how the treatment is allocated. In the following section, these key steps of design are addressed again but with the difference-in-means estimator in mind.

2.1.2 The design steps

The first step of a study is the *sampling* of units from the population to the sample. Assume that there exists a group of units for which we wish to investigate whether an intervention has a casual effect on an outcome of interest. If this group is large, it is often difficult, or impossible, to conduct the study on the whole group. The solution to this problem is to study a smaller group, a sample, from the population of interest and hope that this sample is representative of the larger population, i.e. that $SATE_s$ is close to $PATE$. For example, assume that the population consists of 10,000 individuals in a small city. The population consists of all sorts of individuals, young and old, healthy and sick, etc. A researcher might, e.g., only have the resources to study 100 individuals. The question now is how to choose the 100 individuals such that the results in

the sample best reflect the results in the population, i.e., if all individuals had been studied.

Scientists have long been aware of the importance of sampling, but it was not until the work of Peirce (1884), and then more formally by Neyman (1934) and Fisher (1935), that *randomization* was proposed as a procedure. By drawing a *random sample* from a population, it is ensured that the sample is representative of the population *in expectation*. Here, *in expectation* means that if we were to draw samples repeatedly, they would on average have the same composition of individuals as the population. One way to understand this more formally is by noting that

$$\frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} SATE_s = PATE. \quad (2.5)$$

The relation between SATE and PATE can be further understood as follows: Because, under random sampling, all individuals have exactly the same probability of being sampled, there is no specific type of individual that will be sampled more often than others on average, i.e., all samples $s = 1, \dots, \binom{N}{n}$ are equally probable. Of course, in any specific study, only one sample will be drawn and, for this sample, it might be the case that $SATE_s$ is far from $PATE$.

The second step of an experiment is the *treatment allocation*. Treatment allocation is one of the main focuses of this thesis, and the topic will be returned to several times and discussed in detail. The basic goal of the treatment allocation step is to establish that the observed difference in the outcome of interest, i.e. treatment effect, between the treated group and the control group in the sample is in fact the casual effect of the treatment, and not due to anything else. Returning to the example, say that the drawn sample of 100 individuals consists of 50 young and 50 old individuals. Say that a researcher is interested in discovering whether a labor market training program increases the employment rate. If, e.g., all old individuals are assigned to the program and all young individuals are assigned to some placebo program, any observed difference, or lack of difference, in employment rate might be explained by the difference in age, not the program itself. For example, the rate may be underestimated if there is age discrimination and overestimated if employers value experience.

Again, by randomly assigning the treatment to the sample, old and young individuals will be balanced across the groups in expectation. In fact, using some randomized treatment allocation where all units have the same probability of receiving treatment, e.g., flipping a coin for each individual, any such factor that might affect the outcome, observed or unobserved, will be balanced across the treatment and control groups on average. Coin-flip assignments are more formally called Bernoulli trials and were used extensively historically when random numbers were more difficult to generate. The most common randomized treatment assignment mechanism used today is so-called *complete randomization*, where the treatment group sizes are fixed and one of all

possible allocations is randomly drawn. In a sample of n units, there are $\binom{n}{n_1}$ possible ways of splitting the sample into one group of size n_1 and one group of size $n - n_1$, assuming $n_1 < n$. Note that randomizations, e.g., Bernoulli or complete randomization, are sometimes viewed as designs, as they allow for unbiased estimation of causal effects. However, in this thesis, only procedures where data are allowed to influence the treatment allocation are referred to as designs. That is, randomization is considered an identification strategy that can be combined with experimental designs.

By introducing randomness, effect estimates that are due to something other than the treatment of interest are avoided, *on average*. Formally, this means that the difference-in-mean estimator is unbiased under randomized treatment allocation, i.e.,

$$\frac{1}{\binom{n}{n_1}} \sum_{j=1}^{\binom{n}{n_1}} \hat{\tau}_j = SATE_s. \quad (2.6)$$

In other words, if the treatment assignment within the sample is random, the difference-in-mean estimator is an unbiased estimator of SATE, and, if the sample is randomly drawn from the population, it is also an unbiased estimator of PATE. Unbiasedness for PATE follows directly by plugging Equation 2.6 into Equation 2.5, i.e.,

$$\frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} SATE_s = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} \frac{1}{\binom{n}{n_1}} \sum_{j=1}^{\binom{n}{n_1}} \hat{\tau}_j = PATE. \quad (2.7)$$

Although random sampling and treatment allocation are fairly simple concepts, it is often difficult to perform either of them in practice. Due to the practice of obtaining informed consent, it is often not possible to draw truly random samples from a population, which makes it difficult to draw inference to the units of the population from the sample.

2.2 Observational studies - Trying to deterministically recreate the benefits of randomization

Observational studies are used for several reasons. It may be unethical to randomize the treatment, e.g., in medical studies on ill patients, or it may not be possible for practical, legal, or logistical reasons, e.g. when comparing the effects of socioeconomic status on health. There are several ways of *identifying*, i.e., assuring that an possible observed effect outcome is in fact the causal effect of the treatment on average and not due to other between-group differences. To understand identification properly, it is useful to introduce the concept of *confounders*. Again, let $Y_i(W_i)$ be the potential outcomes if unit i is

given treatment w_i . To be able to identify the SATE, it must hold that

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i. \quad (2.8)$$

That is, the treatment assignment mechanism must be independent of the potential outcomes. This is called the unconfoundedness assumption (UA) (Rubin, 1990). In fact, confounders have already been discussed; in the labor market training program example, age was an example of a possible confounder. Assume, in that example, that a treatment mechanism that favors treating old individuals is used, and old individuals get employed less than young individuals. In this case, the treatment mechanism, W , would not be independent of the potential employment rates, as both the treatment assignment and the employment are influenced by age. Clearly, if the treatment assignment is completely random UA is fulfilled on average. However, if randomization is not an option, some other strategy must be found to fulfill this assumption.

A useful taxonomy is to categorize all scientific studies where the treatment assignment is not under the control of the researcher as *quasi-experiments*. There are two large subclasses of quasi-experiments; *Rubin causal models* (Rubin and Rubin, 2005) and *natural experiments*. Rubin causal models are studies in which the randomized experiment is reconstructed using observed data. Natural experiments are studies where randomization into treatment occurs naturally, as explained below. The focus in this thesis is on Rubin causal models. However, given the importance of natural experiments as a tool for identifying causal estimands, a short description of some common natural experiments is given here for completeness.

2.2.1 Natural experiments

In some settings, natural experiments occur, i.e., randomization into treatment occurs naturally. In other contexts, the manner in which a treatment was applied, e.g., policy that is implemented from a certain date, may give rise to ‘local’ randomization. If, for example, it is decided that all children born after a certain date will receive a vaccine, the individuals born the week before and after this date may be considered randomized into treatment and control. It is not likely that they differ systematically in any other respect than the treatment status on average, and the causal effect can be studied by comparing these groups of children. This identification strategy is called discontinuity design, or regression discontinuity design (Lee and Lemieux, 2010). The formal identifying assumption of a discontinuity design is

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i | i \text{ Close to the discontinuity.} \quad (2.9)$$

Another common identification strategy is the so-called method of *instrumental variables* (Angrist and Krueger, 2001). The idea of this strategy is that there might exist some variable that affects treatment assignment, but that is

independent of the potential outcomes, even if the treatment assignment is not independent of the potential outcomes. Returning to the labor market training example, assume that participation is voluntary and that old individuals have a harder time getting employed and therefore are more prone to participate. Clearly, the treatment assignment mechanism is not independent of the potential outcomes. However, say now that we have the addresses and therefore can calculate the distance from the training office to the individual's home. If it is reasonable to assume that proximity to the office increases the proneness to participate, but that it is independent of other factors that might affect the chance of employment, then this could be an instrument. The trick here is that if we can view the distance from the office to the individuals' homes as random with respect to the employment rate, but systematically affecting the treatment mechanism, this randomness can be utilized.

2.2.2 Rubin causal models

We now return to the Rubin causal models, i.e., observational studies where the observed sample consists of one group that has received treatment and one that has not and where there is no apparent natural design or instrumental variables available. In the labor market training example, assume that the labor market training program in the city of interest is voluntary and we have no idea what the decision to participate is based on. Such treatment groups are often expected to differ systematically for several reasons. If the treatment is voluntary, it is likely that participation is based on expected gain from the treatment, which leads to so-called selection into treatment, i.e. individuals who think they will have large $Y_i(1) - Y_i(0)$ are more likely to participate. For example, if old individuals believe that they have greater benefits from participating than the benefits imagined by young individuals, the groups may differ in age, which may imply that the estimated effect is not the average effect across all ages. Of course, as discussed in previous sections, if we are to claim that it is the treatment, and not something else, that *causes* a difference in the outcome of interest, we must ensure that the treated and untreated groups are similar in all respects except for the treatment they received. In this case, the possibility of identifying a causal effect relies completely on the observed data. That is, we need to find groups of treated and untreated individuals who are very similar on observed covariates, and the observed covariates must be sufficient to ensure that any possible observed effect is due to the treatment and nothing else. In other words, all confounders must be observed, which leads to the updated version of the UA, sometimes called the *weak unconfoundedness assumption*, given by

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i | \mathbf{X}_i, \quad (2.10)$$

where \mathbf{X}_i is the vector of observed covariates. That is, there can be no unobserved unbalanced covariate that is able explain the difference in outcome besides

a causal effect. However, it is not enough that sufficient covariates are observed, there must also be units in both groups with all values of these covariates, so-called *overlap*. That is, if the vector of all possible confounders is known and observed, and we can find individuals among the treated and controls who have the same observed covariate vectors, then the causal effect can be identified. Returning to the labor program example, if age is the only factor that affecting treatment assignment and the potential outcomes and age are observed, the causal treatment effect can be identified if there are in fact individuals of all ages in both groups. This means that if there are no young individuals in the treatment group, we cannot identify the PATE even if age is the only confounder and it is observed. In this case, however, it is possible to identify other estimands, a topic discussed in detail in Paper I.

If the available data are sparse, the UA is a strong assumption. However, if there are rich data sources such as the Swedish national registers, or if the outcomes are observed for all individuals repeatedly before anyone received the treatment, this is quite an appealing approach. In contrast, complete randomization gives balance in all covariates, observed and unobserved, but only achieves this on average. With rich data, observational studies give balance in the observed covariates in each study, but the balance in unobserved covariates are not dealt with at all. This means that, contrary to what many might believe, evidence from a single observational study designed from rich data should sometimes have far more weight than a single randomized experiment. This is especially true in studies on humans, where there are many things that might bias even well-designed randomized studies, as discussed in Section 1.3.1. Many of these issues can be circumvented by using observational data. For example, if register data can be used after some intervention has occurred, Hawthorne effects and cheating by the researcher (e.g, giving extra care to treated individuals) can be avoided.

In practice, *matching* is often used to achieve balanced groups. There are numerous available matching algorithms (Stuart, 2010; Rosenbaum, 2019). The main differences between these algorithms lie in the way in which the ‘distance,’ or dissimilarity, between two individuals is measured and what algorithm is used to minimize it. One critical part, which I will return to in Section 2.4, is the dimensionality of the available data. There are two opposing aspects here: Rich data are needed to make the UA plausible, however, if the data are ‘too rich,’ the matching runs into the curse of dimensionality problem. The curse of dimensionality in this case is a matter of overlap. As the dimensions of the covariate space increase, the parts of the space with few or no observations increase exponentially, and it quickly becomes extremely difficult to find matching groups with overlap in all dimensions.

Papers I and II deal with the topic of observational designs, specifically how to use longitudinal data to create balanced groups. Drawing on the ideas from those papers, Paper III covers designing randomized experiments using longitudinal data.

Intuitively, if we can also use observed information in experimental design to create more comparable groups, we should. However, because the UA assumption is strong without randomization, we would like to use the observed information without losing the benefits of randomization. The following section introduces two common design strategies that are central to this thesis.

2.3 Experimental Design - Maximizing the utility of each study

Given that resources are limited, it is important to reduce the uncertainty in each study. As mentioned in the previous sections, the theoretical justification for random sampling and treatment allocation is based on taking expectations over several random samples and treatment allocations, which implies that, in principal, a large number of similar studies must be conducted to ensure valid conclusions. However, in many cases, only one or a few studies are performed in practice, which makes the quality of each study important. Here, the focus will be on the treatment allocation step. The sampling step is also important, but is beyond the scope of this thesis.

In the treatment allocation step, the goal is to obtain treatment groups that are similar in all relevant respects, such that the observed effect is due to the treatment, and nothing else. Random assignment ensures this on average. However, in any given experiment, an unlucky albeit random unbalanced allocation may be drawn. If important covariates are observed, this can be detected by the researcher after randomization. Returning to the labor market training example in an experimental setting, if the age of the individual is known and the random allocation assigns all old individuals to the program, then clearly the groups are not comparable. It is of small comfort to the researchers that the results will be credible if they just repeat the experiment a large number of times. As a consequence, treatment allocation strategies based on pre-experimental data, i.e., experimental designs, have been proposed, the aim being to avoid unlucky random allocations. Experimental designs can be seen as a combination of the benefits of randomized designs and observational study designs. Traditionally, particularly before the age of fast computers, the ways in which this could be achieved were limited, although some of the traditional designs are so powerful that they are still used. The following sections present, first, the most used traditional designs, and second, the new designs that are possible thanks to the development of computational abilities.

2.3.1 Blocking

Perhaps the most common experimental design is the *Block design*, also called *blocking*, *blocked randomization*, *stratified randomization*, or simply *stratification*. The idea of blocking is to group units into groups based on similarity

on observed covariates. In the example with young and old individuals and labor market training, assume that sex is also observed and considered important to the employment rate. In this case, four groups, i.e. blocks, are created by the researcher: young females, young males, old females, old males. Within each block, treatment is then randomly assigned using, e.g., complete randomization. This ensures that units from all blocks are represented in both the treatment and the control groups. because the treatment within blocks is randomized, blocking ensures balance also on unobserved covariates on average. In addition, the design can in expectation only be improved, but not made worse, by blocking (Imbens and Rubin, 2015) as opposed to randomization without blocking. Blocking on a covariate does only improve the design if the covariate is somehow related to the outcome or the treatment response. If, in the labor example, the sample were blocked on some irrelevant covariate such as favorite color, then that would not improve the design. In many empirical settings, it may be less obvious which of the observed covariates are indeed important, i.e., related to the outcome and the treatment assignment.

Although blocking is a very powerful design strategy, it has clear limitations. Blocking builds on the ability to create meaningful groups based on the covariates, something that is often difficult, especially with continuous covariates. On the other hand, it is the grouping that made it plausible to use these designs in the pre-computer era, as the subset of allocations fulfilling the blocking, i.e. the set of allocations from which one should be randomly drawn, is known without computation. The following section introduces rerandomization, which is a more general experimental design strategy. In rerandomization, the subset of allocations in which to randomize is not known and must instead be found using computational intensive algorithms.

2.3.2 Rerandomization

Informal variants of the rerandomization design strategy have been used for a long time by researchers when they obtain an unlucky randomization. That is, if a randomization has been found to have large imbalances in observed covariates, researchers have simply *rerandomized* the treatment allocation until the groups are better balanced according to the observed covariates. This procedure may seem quite arbitrary and ad hoc. However, the idea, properly formalized, turns out to be very useful for proper experimental designs.

Morgan and Rubin (2012) proposed a rigorous framework for rerandomization. They showed that by using a pre-specified covariate balance criterion as a stopping rule for the rerandomization, probability theory enables valid statistical analysis with improved efficiency as compared to complete randomization. One great benefit of this framework is that it allows use of continuous covariates in the design without discretizing. Just as with blocking, rerandomization

Table 2.1. *Number of possible ways to split a sample in two for various sample sizes together with approximate time consumption for calculating a covariate balance measure for all allocations, assuming that one million allocations can be considered per second.*

Sample size	Number of allocations	Time
10	252	0 hours
20	184,756	0.00005 hours
40	137,846,528,820	38.29 hours
60	118,264,581,564,861,152	3750 years
80	107,507,208,733,336,251,928,280	3409031225 years

based on covariates that are unrelated to the potential outcomes cannot theoretically make the design worse than complete randomization.

The possibilities created by rerandomization also come at a potentially high computational cost: It may take a very large number of randomly drawn treatment allocations before one is found that achieves the balance criterion. Because there is a fixed number of possible allocations, one might be tempted to let a computer go through all possible allocations. However, not even with modern computers is this possible. Just to give some perspective on how quickly the number of possible allocations grows, some examples are presented in Table 2.1. For a sample size of 80, it would take an average modern computer many millions of years to go through all possible allocations.

Luckily, efficient designs can be obtained by considering much smaller numbers of allocations, and it is also possible to estimate the number of allocations that must be considered to obtain an allocation that fulfills a certain balance criterion, making it possible to evaluate the plausibility of rerandomization from case to case. Papers III-VI cover the topic of rerandomization.

Mahalanobis-based Rerandomization

This section briefly motivates the focus of Papers IV-VI: rerandomization based on the Mahalanobis distance covariate balance measure. That is, the rerandomization criterion is specified based on the Mahalanobis distance between the covariate mean vectors of the treated and controls. The focus on this measure is motivated by its mathematical properties. It is an affinely invariant multivariate distance measure, which implies that the scaling of the covariates does not affect the measure. By rerandomizing on the Mahalanobis distance measure, the variation in the projection of the potential outcomes on the covariate space is decreased, i.e., it is closely related to controlling for covariates in a multiple regression estimated with OLS. Additionally, the distribution of the Mahalanobis distances across treatment allocations is asymptotically Chi-square, which enables the derivation of various important results, including the expected variance reduction in the difference-in-mean estimator and its asymptotic sampling distribution.

2.4 Big data - A blessing and a challenge for the design of scientific studies

Big data – or really, digitalization and how it has made it possible to collect and store large amounts of data – have a lot to offer to scientific study designs. In the previous sections, it has been illustrated how the designs of both experimental and observational studies can potentially be improved by using observed covariates.

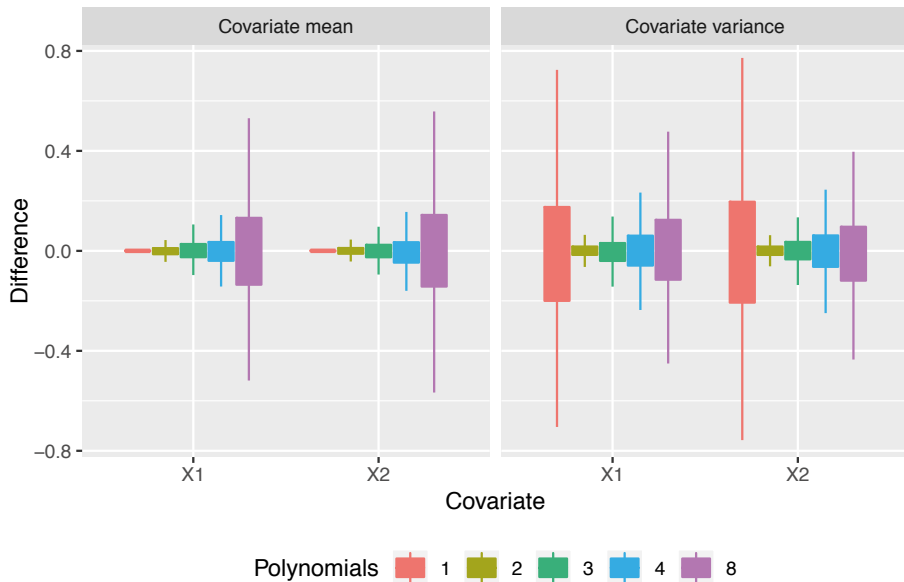
With a lot of available information in the design phase of a study, there is great potential for increasing the validity of any given study. However, large amounts of data also pose challenges for the designer. If a large number of covariates are ‘thrown’ into a matching or rerandomization algorithm, the curse of dimensionality will soon make the results useless. That is, if covariates that are highly related to the outcome are mixed with covariates that are essentially independent of the outcome, the independent ones will diffuse the algorithm such that balance may not be achieved in any covariate. The question is how a large dataset should be utilized, such that the design is as good as possible without imposing the curse of dimensionality.

As a simple example of the curse of dimensionality, consider the following. Two continuous covariates are observed, both believed to be highly related to the potential outcomes, however, the structural relation between the potential outcomes and the two covariates is not known. This means that it is not clear which of the covariate-distribution moments should be balanced to reduce the variance in the outcome the most. Higher order polynomials can be included in the rerandomization, but this also makes it more difficult to find an allocation with a small imbalance.

The tradeoff in the above example can be illustrated in a small simulation. Data are generated for 1000 samples of 100 units each. For each sample, two independent standard normal covariates are generated. Mahalanobis-based rerandomization is performed in five ways: based on the first, first-second, first-third, first-fourth, and first-eighth, polynomials of the covariates, respectively. This implies 2, 4, 6, 8 and 16 rerandomization covariates, respectively. 10,000 random allocations are considered for each sample, and the one with the smallest Mahalanobis distance is kept for each sample. For the kept allocation, the group-difference in the means and variances are recorded for both covariates. Because the covariates are independent and normally distributed, the marginal means and variances are sufficient, i.e., for this setup the covariates are completely balanced if they are balanced on the first and second moments. This implies that rerandomization on the first and second order polynomials should be sufficient.

Figure 2.1 displays the differences in means and variances over the 1000 random samples. When only the first order covariates are included, the means are well balanced, but the variances are not at all balanced. As expected, by including the second order of the covariates the variances of the covariates as

Figure 2.1. Example of mean and variance covariate differences under rerandomization based on 1st, 2nd, 3rd, 4th, and 8th order polynomials of the covariates.



well are similar between the treatment and control groups. However, as the number of polynomials is increased after two, both the imbalances in the means and the variances increase again. In a real application, the true multivariate distribution of the covariates is not known, which means that the researcher does not know which polynomials should be balanced. For this reason, the researcher might want to add more than two polynomials and interactions of the covariates. In other words, in a real application, the tradeoff between good balance and the curse of dimensionality is present, but often not possible to evaluate.

If pre-treatment outcomes are available, it may be possible to estimate the relative importance of each covariate. That is, if a covariate is found to be unrelated to the pre-treatment outcome, it is likely weakly related to the post-treatment outcome as well and can be discarded from the design. However, here too, the risk of apophanies is present. If, e.g., the relation between a pre-treatment outcome measure and a large number of covariates is explored, it is likely that some coincidental relation occurs. In fact, danah Boyd and Crawford (2012) pointed out that the risk of apophany is extra pressing with big data: *“Too often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions. In one notable example, Leinweber (2009) demonstrated that data mining techniques could show a strong but spurious correlation between the changes in the S&P 500 stock index and*

butter production in Bangladesh.”(danah Boyd and Crawford, 2012, pp.668). In other words, big data offer us a possibility to improve our designs, but they simultaneously introduce new potential pitfalls.

In some sense, this brings the thesis, thus far, full circle: Ee are prone to see patterns in randomness and therefore need scientific methodology to help us avoid these biases. If we use randomized controlled trials, we can avoid many biases, but progress will be slow because we need many studies to obtain conclusive results. We can use data to improve the naïve randomized controlled trials by using data-driven randomized designs to speed up progress, but in this design step as well our proneness to see patterns is an obstacle and it might be difficult to use large amounts of data efficiently.

Fortunately data driven designs such as rerandomization cannot, if properly used, give us less efficient studies than complete randomization in expectation. In other words, with rich data, we have a great possibility to improve study designs, but we must also be aware of the fact that, in study design, sophisticated data-driven strategies using massive amounts of data are never substitutes but complements to reasoning and substantive knowledge.

3. Summary of papers

3.1 Paper I

In Paper I, a new strategy for identifying the Average treatment effect of the treated (ATET) using time-intensive longitudinal data is proposed. The typical way of identifying the ATET from observational data is to match on observed covariates. Arguably, the most important covariates are pre-treatment observations of the outcome, and due to technological developments, this type of data is becoming more prevalent and reasonable to collect. If there are differences in the characteristics of subjects that make their pre-treatment outcomes different, then the pre-treatment outcome should contain information about these characteristics, at least the heterogeneity in these characteristics. The idea here is to extract information about the characteristics of an observational unit by studying the time series of the outcome prior to treatment assignment. Even if the unobserved characteristics themselves may not be retained, sufficient information about the heterogeneity may.

Specifically, the idea in this paper is to match on the estimated parameters from parametric time-series models, but instead of fitting one model for each unit, flexible two-level models are used with a large number of random coefficients for capturing the heterogeneity. When a matched control group has been found, the success of the strategy, both the fit of the time-series and the matching on the parameter estimates, can be evaluated by looking at the test statistic, e.g., the difference-in-mean for all pre-treatment time periods. If a match can be found such that the groups are not substantially different from each other for a large number of pre-treatment time periods, then it is possible to non-parametrically identify causal effects in the post-treatment outcomes using any estimator of choice.

3.2 Paper II

In Paper II, the identification strategy from Paper I is applied to an observational study of the effect of dynamic electricity grid fees on electricity consumption. Given that humans are using increasingly large amounts of electricity, it has become more important for grid suppliers to balance the load on the grid over the 24 hours of the day and especially to avoid large peaks. Because of this increasing challenge, many grid suppliers have tried to alter the consumption behavior of their customers by interventions such as economic incentives

to move load from busy hours to less busy ones. More recently, dynamic tariffs have been proposed. In this paper, we investigate the effect of a tariff where the grid fee is based on the highest peak of the month. That is, if a consumer's highest peak is large, the grid fee is high as opposed to if the highest peak is low. In other words, great variation is punished.

For several reasons, the treatment, i.e. the dynamic tariff, was not randomized. Instead, one small city was used as the treatment group, in which all costumers were assigned the new dynamic tariff. Another larger city was proposed as a control group. To identify the Average Treatment Effect of the Treated (ATET), the strategy from Paper I was used. The matching was performed on the parameter estimates from a two-level autoregressive model based on daily electricity consumption data from the year prior to treatment assignment. The strategy could successfully find a matching control group from the pool of controls for which the pre-treatment assignment difference in outcome did not differ. The conclusion from the study is that the dynamic tariff significantly lowered group level consumption over the observed two following years.

This paper illustrates how time-intensive time series data on the pre-treatment outcome can be utilized in observational studies, something that is very common in electricity studies, and is likely to become more and more prevalent also in other fields.

3.3 Paper III

Paper III proposes alternative balance measures and strategies for choosing and weighting covariates in rerandomization designs. Specifically, the focus is on the situation where the outcome is observed repeatedly before the experiment. This type of data poses special challenges for constructing multivariate balance measures, as the covariates are often highly correlated. For example, the Mahalanobis distance measure can be intractable due to non-positive definite covariance matrices. In addition, with longitudinal data, it is not surprising if the number of pre-treatment measurements is greater than the number of subjects, in which case the covariance matrix is singular by definition. In Paper III, several strategies for dealing with this situation are proposed. An alternative rank-based balance measure is proposed which is robust to outliers in small samples and which can handle the situation with more covariates than units. By introducing the concept of mirror allocations, unbiasedness under rerandomization using the proposed balance measure is easily proved. Moreover, this measure has explicit weights for each covariate, which enables estimating the weights based on their relative importance. It is illustrated how the last pre-treatment outcome before the treatment assignment can be used as a proxy for the outcome under no treatment after treatment assignment. That is, by regressing this pre-treatment outcome on covariates and/or other

pre-treatment outcomes, the relative weights can be extracted. We illustrate several estimation techniques for this purpose, including penalized-estimation procedures such as cross-classified LASSO regression.

In addition to the balance measures, we also propose an alternative algorithm for finding the best possible allocations according to a given balance measure, in a given time frame. We discuss why exact inference is always unbiased and valid under this algorithm and how the precision can be probabilistically bounded by setting the number of considered allocations accordingly.

In a simulation study we show that the proposed measure gives similar results to Mahalanobis-based rerandomization in settings with few, equally important weights, and continues to give power improvements when there are large numbers of covariates and/or pre-treatment outcomes available. Using empirical data from an electricity consumption study, we demonstrate that the power gains from the proposed rerandomization strategy, as compared to complete randomization, can be as large as 80 % under a moderately strong hypothetical treatment effect.

3.4 Paper IV

Paper IV discusses the concept of optimal rerandomization designs. We show that, depending on whether inference is to be drawn to the sample or the population, i.e., SATE or PATE, the notion of optimal design differs.

In a Mahalanobis-based rerandomization design, the efficiency gain is a function of the rerandomization criterion. The closer the criterion is to zero, the greater the improvements. For this reason, it is only natural to, when reading about the rerandomization framework, consider making this criterion as small as possible. In other words, finding the allocation with the smallest covariate imbalance and deterministically assigning the treatment accordingly. In Kallus (2018), the author proposed several algorithms for doing exactly this, in addition to alternative balance measures.

In Paper IV, we discuss how such deterministic designs affect the possibilities to draw inference. Particularly, we show that probabilistic inference to units of the sample is only possible if the treatment is randomized within a set of allocations with sufficiently large cardinality. Moreover, some misconceptions regarding rerandomization, and Mahalanobis-based rerandomization in particular, are addressed by reanalyzing an example from Kallus (2018).

The conclusion of Paper IV is that it is often advisable to aim for a design that is optimal for inference to SATE, as such a design is often near-optimal also for PATE. The inference to SATE can then be made exact, without any assumptions.

3.5 Paper V

Paper V summarizes the current knowledge on asymptotic inference for rerandomization designs and proposes some simplifications for practical applications. Drawing on Li et al. (2018), we show that the non-normal sampling distribution of the difference-in-mean test statistic approaches normal as the rerandomization criterion approaches zero. Furthermore, the difference between the correct non-normal distribution and the proposed approximation based on a normal distribution is in many situations negligible even for close-to-zero criterion. These results imply that standard Gaussian inference can be used in many rerandomization designed studies. A strategy for evaluating how well the approximation works in any given setting is proposed and illustrated. The results presented in this paper simplify the usage of asymptotic inference for rerandomization designs, and should facilitate the application of rerandomization in experiments in general.

3.6 Paper VI

Paper VI investigates and clarifies the relation between the traditional blocked designs and rerandomization. Several scholars stand behind the recommendation, formulated by D. Rubin during a seminar, 'Block on what you can, rerandomize on what you cannot'. However, others have viewed rerandomization as an alternative, rather than a substitute, and have for that reason, among others, not recommended rerandomization over blocking. Paper VI shows, both theoretically and with simulations, why it is most often a good idea to first block on categorical covariates and then rerandomize on continuous covariates, a design that we refer to as *stratified rerandomization*. The comparison is achieved by first showing that stratification and, therefore, also stratified rerandomization, are special cases of rerandomization. This means that the theoretical properties of rerandomization can be used for comparisons.

In the paper, we demonstrate that stratified randomization can only be less efficient in small sample settings ($N < 20$) where the categorical covariates are weakly related to the outcome. In all other settings, stratified rerandomization is as good as rerandomization regardless of the importance of the categorical covariates. Asymptotically they give the same result.

In summary, if only categorical covariates are available, blocking and rerandomization give the same efficiency gains if the rerandomization criterion is close to zero, for a larger criterion, blocking is more efficient. If only continuous covariates are available, rerandomization is always more efficient than blocking, as blocking implies discretizing which means information loss. If both categorical and continuous covariates are available, stratified rerandomization and rerandomization give the same efficiency gain if the rerandomization criterion is small enough. Stratified rerandomization is always as good or better than only blocking or only rerandomizing.

3.7 Summary

Papers I and II propose and showcase a strategy for utilizing time-intensive longitudinal data to identify causal effects in observational data. These ideas then inspire Paper III, where rerandomization strategies use longitudinal pre-treatment outcome data to improve the randomized design in experiments. Papers IV and V discuss optimal rerandomizations designs, and how inference, both exact and asymptotic, must be updated to account for the rerandomization. Finally, Paper VI compares classical blocked design with rerandomization designs to shed light on when to use what, concluding that rerandomization is generally preferable, as it encompasses traditional designs but is far more general.

4. Acknowledgements

There are so many people who have made the completion of this thesis possible. As much as I would like to, it is not possible to list everyone here. I would, however, like to extend my special thanks to the following people.

I am so very grateful to my supervisor Professor Per Johansson. You have given me a research environment in which I have been totally free to pursue my ideas, but above all, you have given me so much of your time. The countless hours of discussions in your office have without doubt been the most valuable part my university studies, and I have learned so much from you. You have been available to me and my questions at all times, during workdays, weekends, and vacations, which has made it possible for me to move forward a lot faster than I could have otherwise. I have had so much fun working with you during these years. Thank you!

I would also like to extend my gratitude to my second supervisor, Mattias Nordin. You have on every possible occasion met me with infectious enthusiasm and the perfect amount of sarcasm to keep my ego in check. Your combination of openness to ideas and opinions, and high levels of skill and knowledge in various fields, is truly inspiring. You have both been truly awesome supervisors, I wish for every PhD student to have at least one supervisor like you!

I wish to thank my external mentor, my uncle Bengt Muthén. You have inspired and guided my interest and work in statistics long before, under, and hopefully after, my Ph.D. studies. You have successfully balanced humbleness in response to my ignorance with inspiring high expectations. By letting me take a small part in the Mplus team, you have shifted my references for what can be achieved and what being ambitious means. I am forever grateful for all the time, thought, and care that you, and of course also Linda, have invested in my statistical endeavors and personal development in general.

I would like to thank all my colleagues in the Department of Statistics. Notably, Thommy Perlinger, who has been a great inspiration for teaching, as well as a really good friend. Our tough-love relationship has been a pure energy source for me, ever since I started working as a TA. The importance of our relationship for my wellbeing at work is probably greater than you think Thommy.

Another colleague who has impacted me more than people might think is Rauf Ahmad. Your ambition to never settle for anything less than excellence and your unprecedented constant hard work have moved the boundaries in my head for what is possible. Almost every weekend I have spent at Ekonomikum,

you have been here before I arrived and left after me, each time reminding me that ‘hard’ work is a matter of reference.

Among my colleges are of course my fellow Ph.D. students. I am grateful to all of you for sharing your experiences and advice; I wish you all luck in your continued work! It has been so nice to have a gang of peers at the department. I do wonder how quiet the PhD meetings will be without me though, perhaps someone else will finally have a chance to talk? Alexander, you are a lot more than a colleague to me. We have studied statistics together since the first semester of our undergraduate studies (2012!). We have seen different eras come and go together. Notable eras are: youth, the before and the after of the Master’s inference course, getting a small office when we started as TAs, getting moved to an equally small office when we became PhD students. Our relationship, with a strong common understanding of respect and work discipline, has been extremely valuable to me during this whole time. Much of my work and study discipline can be derived from us pushing each other when facing seemingly undoable challenges at both the Bachelor’s and Master’s level. Also, without our table tennis breaks, I would have gone mad. Thank you for sometimes letting me win!

There is a team of administrative personnel who have provided support whenever needed, always with a super helpful attitude! I would like to thank Evorna (Eva Enefjord and Eva Karlsson) for always solving last-minute problems with travel and payments. Anna Henriksson for always spreading a positive “no-problem” vibe. Finally, Christian, you are probably the most efficient person in the world when it comes to solving things. Whenever you say that you’ll fix something, nothing is more certain than that it will be fixed! All of you have made all of my practical problems disappear, which has made my work a lot easier, I really appreciate that!

I would like to extend my gratitude to Isak Öhrlund. Your willingness to let me in on your ambitious projects has been one of the most important factors helping me finish my thesis. Almost all papers I have written can be derived from ideas I got directly or indirectly from working on data from our projects! Also, you are such a great person to work with and I have enjoyed every minute of our collaborations!

Outside of the academic sphere there are also many people that I would like to thank. I would like to thank my dear friends and band members for giving me music and all the joy that comes with it. Nothing makes me forget all my research problems faster than playing with you. I would like to thank my family – my brother Tore and my sister Anja – for always being there whenever I need you. My girlfriend Sofie – thank you for putting up with my ridiculous number of work hours and always supporting me when I’m struggling. My friend Erik Hedin for always being there with good advise and support.

Finally, I would like to thank my parents. I don’t know how you did it, but you managed to inspire me and motivate me, from a study-reluctant teenager to a PhD in statistics. Well played. Mom, seeing you organizing and leading

with your incredible efficiency has inspired me to challenge myself, especially during my PhD studies. I am so impressed and inspired by your abilities. Dad, your attitude towards working hard has always inspired me, and I will always remember what you said to me when I was younger, about learning difficult things, freely translated: “There is no lack of talent that you can’t make up for with hard work.” Also, you are the kindest and most considerate person I know; everybody I know who knows you loves you. I only hope to become more like you.

Mårten Schultzberg
Ekonomikum 2019-09-20

References

- Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *The Journal of Economic Perspectives*, 15(4):69–85.
- danah Boyd and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679.
- Davenas, E., Beauvais, F., Amara, J., Oberbaum, M., Robinzon, B., Miadonnai, A., Tedeschi, A., Pomeranz, B., Fortner, P., Belon, P., Sainte-Laudy, J., Poitevin, B., and Benveniste, J. (1988). Human basophil degranulation triggered by very dilute antiserum against IgE. *Nature*, 333(6176):816–818.
- Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.
- Foster, K. R. and Kokko, H. (2009). The Evolution of Superstitious and Superstition-like Behaviour. 276(1654):31–37.
- Goldfarb, B. D. and King, A. A. (2013). Scientific Apophenia in Strategic Management Research. *Ssrn*, (March 2018).
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.
- Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(1):85–112.
- Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity in Economics. *Journal of economic literature*, 20(1):281–355.
- Leinweber, D. J. (2009). Stupid Data Miner Tricks. *The Journal of Investing*, 16(1):15–22.
- Leslie, A. M. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288.
- Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157 – 9162.
- MacDonald, A. W., Schmidt, B. J., Seeland, K. D., Abram, S. V., Redish, A. D., Sweis, B. M., and Thomas, M. J. (2018). Sensitivity to “sunk costs” in mice, rats, and humans. *Science*, 361(6398):178–181.
- Maddox, J., Randi, J., and Stewart, W. W. (1988). "High-dilution" experiments a delusion.
- McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., and Fisher, P. (2007). The Hawthorne Effect: A randomised, controlled trial. *BMC Medical Research Methodology*, 7:1–8.
- Mishara, A. L. (2009). Klaus Conrad (1905–1961): Delusional Mood, Psychosis, and Beginning Schizophrenia. *Schizophrenia Bulletin*, 36(1):9–13.
- Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.

- Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. *Annals of Agricultural Sciences*.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558.
- Nickerson, R. R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2):175–220.
- Peirce, C. S. (1884). A Theory of Probable Inference. Idiana University press.
- Rosenbaum, P. R. (2019). Modern Algorithms for Matching in Observational Studies. *Annual Review of Statistics and Its Application*.
- Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292.
- Rubin, D. B. and Rubin, B. (2005). Inference Using Potential Outcomes : Design , Modeling , Decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Schulz, K. F. and Grimes, D. A. (2002). Blinding in randomised trials : hiding who got what. *Lancet*, 359:696–700.
- Simmonds-Moore, C. (2014). Exploring the perceptual biases associated with believing and disbelieving in paranormal phenomena. *Consciousness and Cognition*, 28(1):30–46.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21.
- Thaler, R. (1999). Mental accounting matters. *Journal of Behavioral decision*, 12:183–206.
- Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty : Heuristics and Biases. *Science*, 185(4157):1124–1131.
- Van Elk, M. (2015). Perceptual biases in relation to paranormal and conspiracy beliefs. *PLoS ONE*, 10(6):1–15.
- Waldmann, M. (2017). *The Oxford Handbook of Causal Reasoning*. Oxford handbooks online. Oxford University Press.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 172*

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-393810



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2019