Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2017/7

Empirical Methods for the Law

Christoph Engel

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2017/7

Empirical Methods for the Law*

Christoph Engel

May 2017

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Empirical Methods for the Law[*]

Christoph Engel

Abstract

To their credit, empirical legal scholars try to live up to the highest methodological standards from the social sciences. But these standards do not always match the legal research question. This paper focuses on normative legal argument based on empirical evidence. Whether there is a normative problem, and whether legal intervention promises to mitigate the problem, requires a decision. If uncertainty cannot be completely removed, the legal decision-maker must weigh the risk of false positives against the risk of false negatives. This may call for an adjustment of the significance level. The fact that all legal choice is historically contingent, that legal problems tend to be ill-defined, and that strategic actors have an incentive to bias the generation of evidence defy frequentist statistics. Yet the law can capitalize on the adversarial principle. Competition among interested parties helps contain the strategic element and spurs the creative search for better evidence. This leads to suggestive, but institutionally contained empirical evidence.

JEL: A12, C01, C11, C12, C18, C81, H41, K00, K41

Keywords: normative claims, frequentist statistics, significance, power, structural equation model, finite mixture, Bayesian statistics, prediction, machine learning

[*] Helpful comments by Jens Frankenreiter and Hanjo Hamann on an earlier version are gratefully acknowledged

1. Introduction

Early predecessors notwithstanding (see Kritzer 2010), the empirical legal movement is a late-born child. When, little more than a decade ago, it gained momentum, economics had already undergone its credibility revolution (Manski 1995, Angrist and Pischke 2008, Angrist and Pischke 2010). To their credit, many of the legal scholars intrigued by the empirical turn of their discipline took the methodological challenge seriously and were adamant on living up to the highest technical standards (see for instance the programmatic piece by Ho and Rubin 2011). They admonished legal academia that causal claims are not to be taken lightly (Epstein and King 2002). Like their most ambitious colleagues from the social sciences, legal scholars are now chasing instrumental variables (see e.g. Brinig and Garnett 2012), hunting random shocks (see e.g. Gazal-Ayal and Sulitzeanu-Kenan 2010), and constructing artificial control groups (see e.g. Hyman, Rahmati et al. 2016). Most empirical legal scholars would readily admit that there is no such thing as a perfect empirical design. Before they draw policy conclusions, they would want to see replications, meta-studies and the results gained from multiple methods (often called triangulation) (Lempert 2008, Hamann 2014).

The law as a discipline has reason to be proud of this state of affairs. Numerous causal claims are centre stage in legal argument and have never been put to the empirical test. Some of these claims are just not on the radar of any social science. Testing other claims requires legal expertise. For a long time to go, the empirical legal movement does not have to dread a dearth of highly relevant research questions. Many of these claims straightforwardly lend themselves to an application of the empirical methods that have been developed in the social sciences. If one wants to know whether the certainty or the severity of punishment is more effective in deterring crime, one better had understand why simple correlations between the frequency of shoplifting being prosecuted, or the mean fine in case of being convicted for theft, with the number of registered convictions are not trustworthy evidence (on the challenges for identifying the effect see Engel 2016).

Yet as I will argue in this paper, applying the tried and tested empirical tools of empirical social sciences, and using their standards for evaluating such evidence, will not always fit the legal research question. Rather than just adhering to the standards developed by other empiricists, empirical legal scholars should go back to the rationale behind the methodological choices made by the social sciences. Ultimately it is not the methods that are critical, but the research questions. Methods are tools meant to help researchers find meaningful answers to these questions. For some research questions, the standard methods from the social sciences do not fit well, but can easily be adjusted. For other research questions, methodological solutions are available, but they require the use of non-standard (or at least less frequently used) empirical methods. And there is a non-negligible set of legal research questions for which the standard frequentist statistics used in the social sciences are inappropriate. But what a legal scholar wants to know does not only challenge empirical methods. The law as a discipline also has the opportunity to muster empirical resources unavailable to the ordinary social scientist. This opens up an avenue to "suggestive, but institutionally contained empirical evidence".

To the best of my knowledge, the match between legal research questions and empirical methods has not been investigated systematically. There are good introductions into empirical methods for lawyers (Lawless, Robbennolt et al. 2010, Epstein and Martin 2014, Towfigh and Petersen 2015), but they focus on introducing lawyers to the empirical toolbox of the social sciences. Others have tried to understand why legal practitioners, legislators, and even legal scholars have often been reluctant to rely on serious empirical evidence (Schneider and Teitelbaum 2006, Rachlinski 2010). Legal theorists have discussed whether empirical evidence is a legitimate contribution to their discourse (Galligan 2010). Tor (2008) comes closest in spirit, and points to some of the challenges this paper is concerned with. But that paper does not interpret these challenges in terms of statistical theory, which is my intended contribution.

In section 2 I remind readers of the basic building blocks of frequentist statistics, and try to define the interface with normative legal research questions. Section 3 defines legal research questions for which technically easy, but substantively non-negligible adjustments of frequentist statistics are feasible. Section 4 isolates typical legal research questions for which methodological solutions exist, but are non-standard. Section 5 focuses on legal challenges that are hard to meet with the help of frequentist statistics. Section 6 explores an additional resource for the generation and the evaluation of empirical evidence for law. It results from the fact that legal decisions are heavily institutionally contained. Section 7 concludes.

2. Frequentist statistics and legal research questions

Empirical legal scholars are versatile statisticians. They eloquently discuss research designs, identification strategies, functional form, dependence structures and workarounds if standard methods are unavailable, for instance for the calculation of standard errors. In all of this they take it for granted that frequentist statistics are the gold standard. But frequentist statistics come with a whole armament of assumptions. In this section, I will remind readers of the essential assumptions, and will then ask whether these assumptions match the quintessential empirical questions a legal scholar asks.

Figure 1 should of course be familiar from textbooks on elementary statistics. One is interested in identifying the causal effect of X on outcome variable Y. One assumes that this relationship holds for an entire population. One however only observes a sample. Ideally this sample consists of a series of independent random draws from the population. By another random draw, part of this sample is exposed to X, while the remainder is not. One compares Y in the treatment group with Y in the control group. The null hypothesis posits that X has no effect. Then the difference between Y in the treatment group and Y in the control group should be 0. The statistical test focuses on the magnitude of the observed mean difference between both groups. One rejects the null hypothesis if it is sufficiently unlikely that a difference of this magnitude is itself random. In the social sciences the conventional cutoff is 5%. 5% is of course a relative measure. In the figure below it is a percentage of the area under the curve of the Gaussian normal distribution. If the estimation is unbiased, this curve has a symmetric shape. But its width has itself to be estimated from the data. The 5% cutoff is a necessary precaution since one assumes that there are unobserved causes for variance in the treatment effect, but that these causes are uncorrelated with X. Since the data are assumed to be noisy, it is possible that the randomly drawn sample suffers from selection bias. Now usually interest is not in the null hypothesis, but in the alternative hypothesis. The fact that the null hypothesis is rejected only provides support for the alternative hypothesis if there is theory. This theory must be strong enough to exclude any alternative explanation for the treatment difference. The result of the exercise is a prediction: were one to draw new samples (of the same size) from the population, with replacement, one would reject the null hypothesis in no less than 95% of them. If one so desires, one can also make out of sample predictions: if one were to change some independent variable by some degree, which change in the dependent variable would one expect to find?

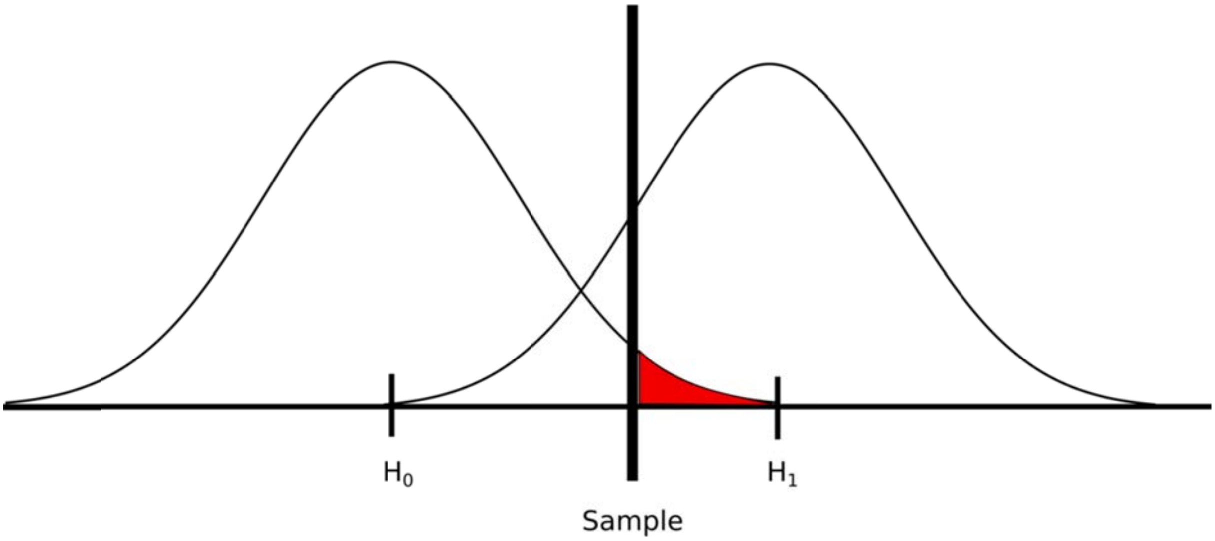

Figure 1
Basic Structure of Frequentist Statistical Inference

For a social scientist, the causal relationship is itself the research question. She has theory predicting the relationship, and uses data as an attempt to falsify her theory. For a legal scholar, the causal relationship is an argument. Her research question is normative. She cares about the causal relationship since, if it was true, she would take a different normative decision than in case it was wrong. Such arguments are routinely made if ambiguous legal rules are interpreted purposively (“teleologically”). The rule is assumed to serve a purpose. At the highest level of generality, this purpose is social betterment. The rule intervenes into individual freedom if (a) in the absence of the rule the authority predicts a normatively undesirable outcome and (b) were the rule interpreted as proposed, this outcome would not obtain, or this would at least be less likely. Hence causal claims are critical for assessing whether (1) there is a social problem that calls for legal intervention and (2) intervention is likely to mitigate the social problem. It is no coincidence that this definition of the legal research question closely resembles the principle of proportionality (Lerche 1999, Clérico 2001, Harbo 2010). The doctrine of proportionality would call (1) the definition of the legitimate aim, and would bring (2) under the rubric of the intervention being conducive to attaining this aim (and potentially also least intrusive and not out of proportion, given the resulting intrusion into constitutionally protected freedom).

3. Adjusting frequentist statistics to the legal research question

Ultimately, lawyers make decisions. They decide themselves if they are judges, public authorities or legislators. If they represent a party in court or in administrative proceedings, they try to influence the decision maker. Same for lawyers representing interest groups in the legislative process. Legal scholars have a choice. They may adopt the role of a mere observer (see the famous distinction by Hart 1961). Much like any other social scientist, such legal observers try to make sense of what they see. They just focus on a specific raw material: legal text or action meant to apply or shape the law. Such legal scholars are free to define the field of observation such that it fits the established empirical methods of the social sciences. But many legal scholars are not content with observation. They want to make a contribution to the evolution of the law. They argue for a different interpretation of a rule, or they propose a new rule. This turns them from observers to (advisors of) actors. In what follows I focus on this branch of legal scholarship, on legal scholars arguing normatively.

In retrospect, a decision maker may regret her choice. There are two reasons for regret, Table 1: it would have been desirable to intervene, but the decision-maker has not done so; or it would have been desirable not to intervene, but the decision-maker has taken action. If the decision-maker perfectly understands the choice problem, and if she is perfectly informed about the facts, she can avoid either mistake with certainty. But if there is conceptual or factual uncertainty, at the meta-level, a second normative decision has to be taken. The need for this decision is illustrated by Figure 1. If one is strict with avoiding false positives, and cannot increase the degree of certainty (in the figure: narrow down the width of the distribution by increasing sample size), one must accept a higher risk of false negatives.

	no reason for intervention	reason for intervention
no intervention	true negative	**false negative**
intervention	**false positive**	true positive

Table 1
Ex Post Assessment of Decisions

As the parallel to Figure 1 makes obvious, this normative choice is at the core of frequentist statistics. It becomes patent as soon as significance testing (avoiding false positives) is coupled with power calculations (avoiding false negatives). By convention, empirical social scientists are just more concerned about claiming a causal relationship that might actually not be true, rather than not making a causal claim if there is a causal relationship in the population. The conventional α level (accepted risk of a false positive) is 5%. In many social science publications, the β level is not made explicit. If power calculations are added, one usually accepts a risk of 20% of missing a true causal effect.

For most of social science, these conventions make good sense. Social scientists do not want to be accused of producing "bogus evidence". Not so rarely, legal scholars have good reason to follow the convention. Any legal intervention curtails the freedom of its addressees. It often is normatively well founded to do this only if it is very unlikely that the intervention is actually unnecessary. For one of its core activities, the legal system even makes this choice explicit. In criminal procedure, almost all legal orders of the world adopt a standard of proof akin to the US standard "beyond a reasonable doubt". It implies the normative decision that convicting an innocent defendant is much worse than acquitting guilty defendants (for background see Engel 2009).[1] Yet in other contexts, the law has good reason to be more concerned about false negatives. A classic is the precautionary principle (O'Riordan 1994). While this has never been uncontested (Sunstein 2005), most legal scholars would agree that there are risks so severe that intervention is justified even if, in retrospect, intervention was not necessary. The reactions of legal orders to the Fukushima incident are a case in point.

Adjusting frequentist statistics to this concern is straightforward. If empirical legal scholars want to make a contribution to the normative legal debate, they should not only report significance, but also power. But ultimately, the 5% cutoff is only conventional. Whether it is appropriate depends on the legal research question. For some research questions, it may be far too liberal to accept a 5% risk of false positives. And for other research questions, the 5% level may be too restrictive. It may be normatively relevant to alert legal authorities to a risk, even if there is a 10 or 20% probability that it doesn't exist. Of course, the normative research question should not be an excuse for sloppy empirical design. If a harder significance criterion can be met, this is desirable. But if the available evidence is limited, and if false negatives are normatively troublesome, the fact that the result is not significant at the 5% level should not prevent empirical legal scholars from publishing it.

If legal scholars evaluate or propose how to shape legal intervention, they often do not only want to learn whether it works at all. They want to compare the performance of alternative interventions. For the moment, I bracket that these comparisons tend to be multidimensional and assume that there is some accepted unidimensional metric, say the efficacy of the intervention, or its cost benefit ratio. Adjusting frequentist statistics to such legal research ques-

[1] Strictly speaking, this application does not fit here, however. It is fraught with the additional challenge of assessing likelihoods for an individual case.

tions is easy. One simply defines the null hypothesis differently. One compares the outcome variable of choice for one intervention with that same outcome variable for a competing intervention. But the significance level again becomes relevant. If there is a serious normative problem, not doing anything because one cannot choose between two interventions may not be acceptable. Of course, better evidence would be desirable. But if it is not available (within the politically relevant timeframe), it may still be relevant information that the risk of making a mistake when preferring one intervention over the other is smaller than, say, 10%, or 15%.

Usually, new law is not designed from scratch. The prototypical situation is institutional reform. The legislator replaces the previous regime with a new one. Even if the legal order does not give the courts power to make new law (as in continental Europe), courts and administrative agencies effectively change the law through reinterpretation. But this is even more an incremental approach. Empirical legal scholars may want to evaluate such changes. There are many issues to worry about that are dealt with in the econometrics textbooks (an excellent guide is Blundell and Costa Dias 2009). But in principle, assessing the desirability of institutional change lends itself to the classic "difference in difference" approach. One needs a jurisdiction that has not been affected by the change, and another that has. One needs a sufficiently long panel so that one can assess whether, before the intervention, both jurisdictions have been on the same track. Identification comes from the fact that the intervention puts the jurisdiction that has been affected on a substantially different track. Technically, identification comes from the significance of the interaction effect. With this legal research question, it is less defensible to adjust the significance level. Of course, if one sticks to the conventional 5% level, this leads to conservatism. It becomes more likely that the previous regime remains in force even if adopting the new regime would have been an improvement. But any legal reform is fraught with uncertainty while the existing regime has stood the test of time.

4. Legal research questions calling for more elaborate frequentist tools

Other legitimate legal research questions call for the use of statistical tools that are not standard in at least some branches of the empirical social sciences. A first challenge has already been alluded to. It is most frequent if legal scholars want to evaluate interventions. Whether the intervention is desirable is not only a question of efficacy. If thieves lose a hand, one may have reason to believe that this is an effective deterrent. But if the intervention is already that drastic for a rather minor crime, the scope for discriminating between crimes of different severity shrinks. This may put thieves on the track for more severe crime. The legal order may dread the impact on others obeying the law if the legal order is associated with cruelty. If the country is a welfare state, it will have to feed the thieves for the rest of their lives since they can no longer work.

All of these could be dealt with as separate causal claims. Is theft less frequent in those Islamic countries that still execute this sanction? Is more serious crime more frequent in these countries? Do thieves in these countries more frequently engage in violent crime? Is petty crime, say jaywalking, more frequent in these countries? Is a higher fraction of the population on welfare? Each result would be a valuable contribution to the normative debate. But ultimately, the most critical part of the exercise would then be left to the intuitive assessment by policy-makers or doctrinal lawyers. Sometimes, intuition is the lender of last resort for legal decision-making. But if quantitative results are ultimately fed into intuitive decision-making, chances are that the advantages in terms of precision and control are lost.

If each of the dimensions of normative interest is measured for each individual observation, there is an elegant quantitative way of aggregating the information. One can treat the observations for each dimension as an indicator variable for a latent variable (effectiveness or desirability, depending on the design of the investigation). One would again compare untreated with treated cases, but would now learn whether the intervention is overall preferable. In a structural equation model, the entire path diagram of Figure 2 can be jointly estimated. One learns how likely it is that, overall, the alternative intervention outperforms the status quo, or another comparison group (for technical details see Kaplan 2008, Westland 2015).

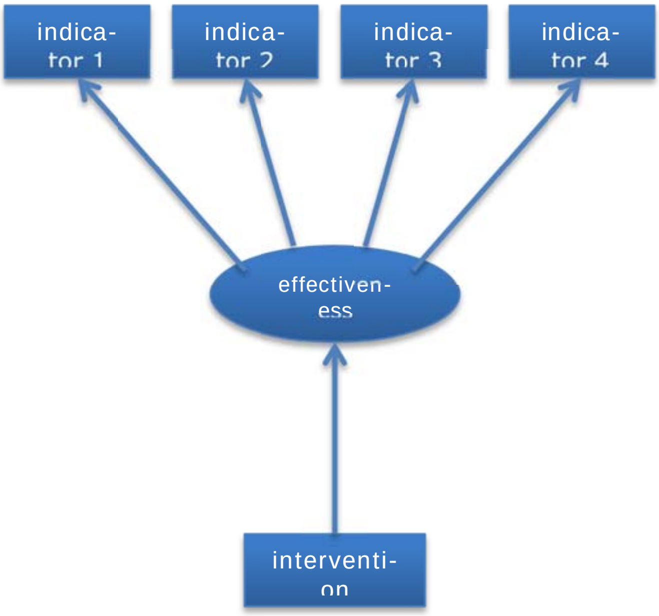

Figure 2
Path diagram for a structural model capturing multiple normative dimensions

Elementary statistics assume a homogeneous population. As Figure 1 illustrates, this does not imply that every member of the population is supposed to be identical. But one assumes that differences within the population are not systematically correlated with the intervention. For many legal research questions, this assumption is strong. Happily, most of the time most individuals obey the law. But some do not. Regulators may be tempted to target these exceptional individuals. But such interventions may be backfiring. While the intervention indeed affects the exceptional individuals who did not care about the law in the first place, the willingness of many others to abide by the law just because it is in force may erode. Such crowding out effects are an active area of behavioral research (Fehr and Gächter 2001, Frey and Jegen 2001, Borges and Irlenbusch 2007). If she suspects crowding out, an empirical legal researcher would not be well advised to just compare population means. She might correctly conclude that the intervention is not as effective as expected. But she would not see the cause of the problem. It results from the heterogeneity of the effect. In the case of crowding out, the effects do even have opposite sign. While the effect is positive for the target group, it is negative for the majority of the population.

Sometimes the researcher is in the happy situation of having theory about the heterogeneity of the effect, and additional data for classifying individuals. In that case, she can estimate a richer statistical model that explains outcomes with the intervention, group membership, and the interaction of both variables. Otherwise she must try to simultaneously estimate group membership and the treatment effects conditional on group membership. A finite mixture model can achieve that, but one needs sufficiently many observations and sufficiently rich data (for detail see McLachlan and Peel 2004).

5. Legal research questions hard to tackle with frequentist statistics

A third group of empirical legal research questions are hard to tackle with frequentist statistics. As Douglass North once aptly put it: institutions are lumpy responses to lumpy perceived problems (North 1990). From a normative perspective, it is ultimately not important whether some intervention increases the awareness for a normative problem, whether it raises the opportunity cost of breaking the rule, or whether it makes those tempted to break the rule more confident that others will not get away with violating the rule either. What counts is whether overall "the rule does the trick". For an empiricist, this is troublesome. As explained when introducing Figure 1, frequentist statistics critically hinge on theory. The fact that the null hypothesis is rejected is not meaningful. Only theory can tell the researcher that therefore the alternative hypothesis is supported. Now if one is exclusively interested in effectiveness, there is a way out. One may confine the claim to the statement that "the intervention works" or that "this intervention works better than that intervention". But one learns nothing about the cause of the effect and therefore is also not in a position to predict which changes in framework conditions would make the intervention ineffective.

Douglass North's statement is related to an even greater challenge. As I have stressed throughout this paper, legal research questions are normative. Legal scholars give advice for normative choice. Those in the legislature or in court who receive this advice have to decide. They will be personally held responsible for the decisions they make. It is therefore out of the question for them to neglect seemingly legitimate concerns just because it is difficult to test them empirically. Even if a concern is not yet conceptually clear, but intuitively plausible, legal authorities will try not to neglect it. The very fact that decisions must be made causes the typical legal research question to be fairly complex, if not simply ill-defined (more on this challenge from Gigerenzer and Engel 2006).

Frequentist statistics at best offer partial responses. If the data are sufficiently rich, one may stress test a finding with sensitivity analysis. One may – either randomly or systematically – perturb control variables and check whether the treatment effect survives. If one is at least able to make the elements of the concern explicit, one may complement the statistical analysis of the observed data with agent-based simulations (for detail see Helbing 2012). One then systematically varies the quality of potential influences, and their magnitude, and checks whether the outcome from the observed data changes. But often the researcher will remain uncertain whether she has truly understood the social problem, and the ways in which the intervention in question affects its incidence. This is when it may be helpful to remember that the ultimate goal for normative research is not explanation, but prediction.

Prediction based on explanation has the big advantage of facilitating justification. Since the law engages sovereign power, in constitutional states mere justifiability is normally not enough. The intervention must be accompanied by explicit justification. But justification need not follow from first principles. The fact that the presence of a social problem, or its mitigation by some intervention, have not been tested with the methods of frequentist statistics does not per se make the intervention illegitimate. It may still be intuitively appealing. Or the intervening authority may purely rely on what political scientists have called input legitimacy (Easton 1965, Scharpf 1999). The authority has come into power in a democratically controlled way, and the exercise of its power comes under regular and effective scrutiny. More importantly even, the legitimacy of the intervention may be assessed ex post by the fact "that it has worked". This is called output legitimacy in this literature.

From this perspective, it may be conceivable to give legal authorities academic advice that exclusively focuses on minimizing prediction error. This is what companies like Google and Facebook engage in very effectively. They are not working on a grand theory of consumer choice. Rather they want to be in a position to predict as precisely as possible which consumer is likely to buy which product under which circumstances. Quite likely they also give advice to their business customers on how to design offers such that target consumers are more likely to accept them. Along the same lines, legal authorities may want to predict which of their addressees is likely to react in the normatively desirable direction, to which intervention and under which circumstances. While the precise algorithms used by Google and Facebook are their business secrets, the machine learning methods aimed at finding the most predictive cue patterns have been published long ago (for detail see Anzai 2012), and could be used for legal research (for a recent application see Kleinberg, Lakkaraju et al. 2017).

A final challenge results from the normative character of empirical legal research. As repeatedly stressed, the ultimate goal of this research is not explanation, but decision-making. The more empirical research becomes important for the development of legal doctrine or of legislation, the more interested parties will try to influence the study design, its evaluation, and the introduction of the results into legal decision-making. Biasing empirical research is easy. Statistical theory is chiefly concerned with bias resulting from selection: the sample is not representative for the population. Selection may be strategic. Interested parties may try to gear empirical researchers towards a part of the population that is biased in their favor. Yet another risk looms even larger. If such parties know, or guess, why evidence is collected, they may try to influence the choices of those individuals that are studied. This risk is particularly pronounced with an otherwise appealing empirical method: experimental legislation. Exposing different parts of the population to different regimes, and randomly assigning them to either regime, can legally only be justified if there is serious uncertainty about the performance of either regime. This, however, makes it very difficult for the experimenting legislator not to publicly discuss the motive for experimentation. This public debate alerts interested parties and guides them towards the most effective channel for manipulation.

While the empirical researcher and her intended doctrinal or legislative audience should be aware of this possibility, the concern is difficult to address with standard empirical methods. Researchers are of course safe if they ignore data as long as its manipulation cannot be ruled out. But that would often imply closing the only available source of empirical evidence. If the fact of manipulation, and its degree, were known, one might adjust the estimated (non-)effect of the intervention. But empirical researchers will rarely have this information. The only possibility is adjusting the significance level. But that would at most avoid finding an effect where there is actually none, or not finding the effect where there is actually one. One would still not know how strongly the estimate is biased. And the empirical researcher is unlikely to know for a fact that an interested party has manipulated the collection of evidence. Shall mere suspicion suffice to adopt a more stringent standard?

6. Legal institutions as a resource for empirical research

The empirical legal movement has been fuelled by the availability of data from the legal system. In this respect, legal institutions are an obvious resource for empirical research. This is, however, not the kind of relevance of legal institutions I am interested in in this section. Rather I want to discuss in which ways the fact that empirical research is intended to be used for doctrinal or legislative purposes is not only a challenge, but also an opportunity for this research.

A first opportunity is straightforward: courts may demand the generation of empirical evidence, or they may make it clear to the parties that their chances to win hinge on the availability of hard empirical evidence. The legislator may order an empirical investigation before taking action. Most importantly it may use the legislative process itself to generate empirical evidence. From a scientific perspective the already mentioned experimental legislation is often less than perfect. Treatment and control group do not match very well. Participants are not randomly assigned. But even an imperfect legislative test gives researchers additional data.

With a bit of luck they can remedy the deficiencies, for instance by finding an acceptable instrumental variable.

In principle, the fact that legal institutions are rarely designed from scratch is a challenge for empirical research. This challenge is compounded by the fact that there is not only a dynamic dimension. Both the segment of social life into which the legal rule intervenes, and this rule itself, are historically contingent. They exist in one given society, at one given moment of time, and one often has good reason to doubt that it would be legitimate to abstract from this historical context. For frequentist statistics, historicity is a severe obstacle. If historicity cannot be neglected, there is strictly speaking no longer a population of reasonably comparable situations.

Yet there is an alternative statistical approach. Bayesian statistics do not require a population of independently drawn identical observations. Bayesian statistics use any available information to update prior expectations (for an excellent introduction see McElreath 2016). This is why historicity can be an advantage for empirical research. If one has good reason to see a trace of the past, the empirical investigation need not start from the assumption that nothing is known. Rather the existing knowledge, or even widely accepted theory, can be used as a starting point. New evidence is used to adjust this prior expectation, or to increase confidence in it. For the intended legal audience, the Bayesian approach has a further advantage. The degree of confidence in the empirical claim, and the degree by which empirical judgement has been swayed by fresh evidence, become transparent. This makes it easier to assess whether the evidence is important enough to change normative legal choice. The Bayesian approach also fits the model most lawyers have for the development of the law. They tend to see this as an evolutionary process. It usually starts with a concrete conflict raising awareness for a potential social problem. Initially courts and administrative agencies tend to cautiously explore the concern. It usually needs a striking case to start the process of shaping a legal reaction to this concern. Over time, the legal system better understands the confines of the problem, and the power of possible reactions. Often the legislator waits for the courts to collect these experiences before a more encompassing regime is drafted. A Bayesian approach can exactly match this process. It can inform the law about the increasing degree of confidence as evidence accumulates.

Court procedure is adversarial. The same holds for many administrative procedures. While the sides to the dispute are less institutionalized in legislation, here too usually stakeholders compete for influence. In the previous section I have explained why the resulting strategic influence on the generation of empirical evidence is a challenge for empirical research. But this challenge too can be turned into an opportunity. Provided the competing parties all have sufficient access to data, and to researchers who are able and willing to generate and analyze this data, the legal system can extend the adversarial principle to the generation and evaluation of empirical evidence. Each side to a dispute is invited to present the best available evidence to support her cause. The legal decision-maker is in the comfortable situation of comparing evidence with counter evidence. This adversarial approach not only helps to contain strategically induced bias. It also musters the parties' creativity and their entrepreneurial spirit to generate or find hitherto unavailable evidence. This is particularly valuable if one suspects the normative decision problem to be overly complex or ill-defined. Much like competition in the market (von Hayek 1945, von Hayek 1969), the competitive nature of legal conflict spurs creativity and increases chances for uncovering as yet unknown, critical evidence.

The adversarial generation and evaluation of empirical evidence has a further advantage. It is best illustrated with court procedure since there the procedure is institutionalized. But the advantage also extends to the less institutionalized battle over new law. All one needs are credible representatives of either side: those advocating a change in the law, or one specific change for that matter, and those opposing it.

To illustrate, take a dispute between plaintiff and defendant over a merchandise that defendant has bought from plaintiff. Defendant refuses to pay on the argument that the product is defective. It suffices for plaintiff to maintain that defendant has bought and received the product. The court will not investigate whether a deal has been struck, or whether defendant was under age, as long as defendant does not challenge the validity of the contract. If there is a valid contract, in principle defendant has to pay the agreed upon price. But there is a remedy if the product is defective. In principle it is for defendant to claim the remedy. If she does and plaintiff does not object, the lawsuit is over and plaintiff loses. Let's assume that the defect itself is not contested. But defendant has only detected it after having used the product for a couple of days. Plaintiff ventures the possibility that the cause of the defect is not production, but the improper use of the product. Now things get interesting. For a consumer it is genuinely difficult to prove that a producer has not taken sufficient care when designing or making a product. Legal orders have reacted in different ways (Howells 2012). Some legal orders go as far as shifting the burden of proof to the producer. She has to prove, to the requisite standard, that she has *not* caused the product to be defective. Other legal orders are more cautious. Rather than completely shifting the burden of proof, they rely on the technique of *prima facie* evidence (for background see Herlitz 1994, Schweizer 2015). The less it is likely that the specific defect has been caused by misuse, the higher the burden on plaintiff. But plaintiff need not prove that she has taken any conceivable care when producing the goods. She may instead also bring forth specific features of the case that make it more likely that there was misuse.

This is a powerful technique for assessing empirical claims. It is tailored to narrowing down the actual investigation on the features of a much richer situation that (a) are critical for the normative decision to be taken and (b) contested. The definition of the investigation program is not exogenous. It is in the hands of the contestants. If the opponent is silent, mere contestation suffices. It is now for the opponent to prove the contested claim. But the better a claim is substantiated, the more the opponent has to advance for shifting the burden of the proof. The *prima facie* principle repeats this logic at the proof stage. If one party has proven a *prima facie* claim, her opponent may no longer content herself with simply contesting the validity of the evidence. She has to claim, and if necessary prove, facts to show that this is a special case, so that the *prima facie* evidence loses its probative power.

Now this paper is not concerned with specific evidence (has defendant been on the crime scene?) but with generic evidence (does the prospect of having to compensate victims for a loss deter torts?). But the logic may be extended to the generation and evaluation of generic evidence. As soon as contestants are defined and considered reliable, one may leave it to their interaction to carve out the area of contestation. One may use the burden of substantiation to further narrow down the investigation program. One may replace full proof with the assessment of typicity if there is sufficiently reliable evidence on a more general class of phenomena to which the concrete issue arguably pertains.

7. Conclusion

Legal argument is rife with causal claims. But this does not turn legal scholarship into empirical social science. The core of legal scholarship remains normative. Causal claims are made because they inform the law about a normative problem, or about the instrumentality of legal intervention as a remedy for this normative problem. Normative legal scholarship gives advice to legal authorities that have power to interpret the existing law, or to make new law. Hence normative legal scholarship prepares decisions that are taken with sovereign authority. If empirical research is to be instrumental, it has to reflect this ultimate normative purpose of the exercise. In principle, scientifically sound empirical evidence is a valuable contribution to taking these normative decisions. But the conventions of empirical social science focus on preventing unreliable evidence of one specific kind: false positives. It depends on the normative purpose whether avoiding false positive decisions is indeed paramount, or whether false positives and false negatives have to be balanced out differently.

It is relatively easy to adjust the assessment of empirical evidence to some characteristic features of normative legal problems. Lawyers tend to choose between alternative interventions. This can be reflected by choosing an appropriate null hypothesis. Lawyers tend to change earlier rules, rather than designing a regime from scratch. This can be reflected by a difference-in-difference approach. Other features of the typical legal problem require more elaborate statistical approaches. The characteristic multidimensionality of normative argument can be captured by a latent variable in a structural equation model. The characteristic patterned heterogeneity of expected reactions to a new rule can be captured by the estimation of heterogeneous treatment effects. Yet other challenges are pushing the limits of frequentist statistics. In the words of Douglass North institutions are lumpy responses to lumpy perceived problems. This makes it difficult to generate a theoretical claim that allows to derive support for an alternative hypothesis from rejecting the null hypothesis. Historical embeddedness defies the idea that identical cases are randomly drawn from a population. Bayesian methods are better suited to exploit this historical contingency. Ultimately lawyers decide on people's lives. This is why they may not neglect an intuitively relevant concern just because there is no convincing theory, or no reliable empirical evidence. A radical, but potentially preferable way out is completely shifting away from explanation to prediction, and use machine learning techniques.

The normative legal purpose of the exercise is not only a challenge for empirical research. Legal decision-making is densely institutionalized. This institutional framework can be used as an additional resource for empirical research. There is room for suggestive, but institutionally contained evidence. The most important feature of legal decision-making is its adversarial nature. This gives a handle on an otherwise thorny problem: if interested parties anticipate that empirical evidence will become relevant for legal decisions that affect them, they have an incentive to strategically perturb the generation and evaluation of this evidence. If they know that their opponent has an incentive to do the same, the two biases may cancel each other out. More importantly, much like competition in the market, the competitive nature spurs creativity and helps legal decision-makers uncover hitherto neglected, but normatively relevant evidence, and detect normatively critical dimensions of a problem only partly understood previously.

This paper has focused on the arguably most challenging, but also most important interface between empirical research and legal scholarship: the generation of evidence for interpreting or designing legal rules. But doctrinal lawyers also need empirical evidence when the facts of a case are disputed. Since the case is singular by definition, frequentist statistics do only have a minor role to play in this. One may argue that the individual case pertains to a class of cases for which generic knowledge is available. By contrast Bayesian statistics are also able to quantify the confidence in the assessment of individual cases.

In the terminology of Hart (1961), when they engage in advising legal decision-makers, legal scholars are in the role of actors. In principle, traditional frequentist statistics are a better fit when legal scholars assume the role of a scientific observer. But even then, additional methodological challenges have to be addressed. Often legal scholars rely on their professional expertise to generate the data. They translate qualitative observations, usually taken from legal text, into quantitative data. If they do so in person, the art of coding is required (for an elaborate treatment see Epstein and Martin 2014). But they can instead also use machine learning methods, or combine the professional training of algorithms with their later implementation (for an example see Talley and O'Kane 2012).

Legal scholars are frequently faced with problems of “small N”. This in particular holds for the comparison of the solutions taken by different jurisdictions, different courts, or different decision bodies within judicial or administrative agencies. Legal scholars should resist the temptation to neglect the resulting independence problem. Observing the same country for several years does not mean that one has several independent observations from this country. The country stays the same, and influences from the past play themselves out throughout the entire period of observation (more on the challenges inherent in quantitative comparative law from Spamann 2015). Empirical legal scholars should rather rely on the insights from political science that frequently faces the same problem (King, Keohane et al. 1994). The problem of independence is exacerbated if empirical legal scholars investigate the jurisprudence of a single court. Frequentist statistics are inappropriate for this exercise in the first place. The whole line of jurisprudence is a single observation. There is nothing to be tested. This limitation should not be swept under the carpet. If the legal scholar has the good fortune of observing the complete output of the court, there is no need for statistical inference in the first place though. She observes the complete population (of a single independent observation, that is). The descriptives are the message. If the court only publishes some of its decisions, there is a problem of selection. All one can do is searching for smoking guns indicating that the available evidence is not only incomplete, but biased in a way that is relevant for one's research question.

The law as a discipline is in the fortunate situation of having opened itself up to serious empirical investigation when other disciplines had long ago filled a rich methodological toolbox. Not so rarely, applying the best tools of the empirical social sciences also serves the law best. But as I have argued in this paper, this need not be the case. This is certainly no justification for basing legal choice, or legal scholarship for that matter, on poor evidence. But legal scholars would be well advised to carefully reflect which empirical methods fit their research questions best. This quest for the most appropriate empirical tools may ultimately even lead to the development of new, law specific empirical methods.

References

ANGRIST, JOSHUA D AND JÖRN-STEFFEN PISCHKE (2008). *Mostly Harmless Econometrics. An Empiricist's Companion*. Princeton, Princeton University Press.

ANGRIST, JOSHUA D. AND JÖRN-STEFFEN PISCHKE (2010). "The Credibility Revolution in Empirical Economics. How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* **24**: 3-30.

ANZAI, YUICHIRO (2012). *Pattern Recognition and Machine Learning*, Elsevier.

BLUNDELL, RICHARD AND MONICA COSTA DIAS (2009). "Alternative Approaches to Evaluation in Empirical Microeconomics." *Journal of Human Resources* **44**(3): 565-640.

BORGES, GEORG AND BERND IRLENBUSCH (2007). "Fairness Crowded Out by Law. An Experimental Study of Withdrawal Rights." *Journal of Institutional and Theoretical Economics* **163**: 84-101.

BRINIG, MARGARET F. AND NICOLE STELLE GARNETT (2012). "Catholic Schools and Broken Windows." *Journal of Empirical Legal Studies* **9**(2): 347-367.

CLÉRICO, LAURA (2001). *Die Struktur der Verhältnismäßigkeit*. Baden-Baden, Nomos.

EASTON, DAVID (1965). *A Systems Analysis of Political Life*. New York, Wiley.

ENGEL, CHRISTOPH (2009). "Preponderance of the Evidence Versus Intime Conviction. A Behavioral Perspective on a Conflict Between American and Continental European Law." *Vermont Law Review* **33**: 435-467.

ENGEL, CHRISTOPH (2016). "A Random Shock is not Random Assignment." *Economics Letters* **145**: 45-47.

EPSTEIN, LEE AND GARY KING (2002). "The Rules of Inference." *University of Chicago Law Review* **69**: 1-133.

EPSTEIN, LEE AND ANDREW D MARTIN (2014). *An Introduction to Empirical Legal Research*, Oxford University Press.

FEHR, ERNST AND SIMON GÄCHTER (2001). Do Incentive Contracts Crowd Out Voluntary Cooperation?

FREY, BRUNO AND RETO JEGEN (2001). "Motivation Crowding Theory. A Survey of Empirical Evidence." *Journal of Economic Surveys* **15**: 589-611.

GALLIGAN, D.J. (2010). Legal Theory and Empirical Research. *Oxford Handbook of Empirical Legal Research*. P. Cane and H. Kritzer. Oxford, Oxford University Press**:** 976-1001.

GAZAL-AYAL, OREN AND RAANAN SULITZEANU-KENAN (2010). "Let My People Go. Ethnic In-Group Bias in Judicial Decisions—Evidence from a Randomized Natural Experiment." *Journal of Empirical Legal Studies* **7**(3): 403-428.

GIGERENZER, GERD AND CHRISTOPH ENGEL, Eds. (2006). *Heuristics and the Law*. Boston, MIT Press.

HAMANN, HANJO (2014). *Evidenzbasierte Jurisprudenz: Methoden empirischer Forschung und ihr Erkenntniswert für das Recht am Beispiel des Gesellschaftsrechts*, Mohr Siebeck.

HARBO, TOR-INGE (2010). "The Function of the Proportionality Principle in EU Law." *European Law Journal* **16**(2): 158-185.

HART, HERBERT LIONEL ADOLPHUS (1961). *The Concept of Law*. Oxford, Clarendon Press.

HELBING, DIRK (2012). Agent-based Modeling. *Social self-organization*. D. Helbing, Springer**:** 25-70.

HERLITZ, GEORG NILS (1994). "The Meaning of the Term Prima Facie." *Louisiana Law Review* **55**: 391-408.

HO, DANIEL E AND DONALD B RUBIN (2011). "Credible Causal Inference for Empirical Legal Studies." *Annual Review of Law and Social Science* **7**: 17-40.

HOWELLS, GERAINT (2012). Product Liability. *Elgar Encyclopedia of Comparative Law*. J. M. Smits. Cheltenham, Elgar**:** 716-725.

HYMAN, DAVID A, et al. (2016). "Medical Malpractice Litigation and the Market for Plaintiff-Side Representation. Evidence from Illinois." *Journal of Empirical Legal Studies* **13**(4): 603-636.

KAPLAN, DAVID (2008). *Structural Equation Modeling. Foundations and Extensions*, Sage Publications.

KING, GARY, et al. (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton, N.J., Princeton University Press.

KLEINBERG, JON, et al. (2017). Human Decisions and Machine Predictions, National Bureau of Economic Research.

KRITZER, HERBERT (2010). The (Nearly) Forgotten Early Empirical Research. *The Oxford Handbook of Empirical Legal Studies*. P. Cane and H. Kritzer. Oxford, Oxford University Press**:** 875-900.

LAWLESS, ROBERT M, et al. (2010). *Empirical Methods in Law*, Aspen Publishers.

LEMPERT, RICHARD (2008). "Empirical Research for Public Policy. With Examples from Family Law." *Journal of Empirical Legal Studies* **5**(4): 907-926.

LERCHE, PETER (1999). *Übermaß und Verfassungsrecht. Zur Bindung des Gesetzgebers an die Grundsätze der Verhältnismäßigkeit und der Erforderlichkeit*. Goldbach.

MANSKI, CHARLES (1995). *Identification Problems in the Social Sciences*. Cambridge, Harvard University Press.

MCELREATH, RICHARD (2016). *Statistical Rethinking. A Bayesian Course with Examples in R and Stan*, CRC Press.

MCLACHLAN, GEOFFREY AND DAVID PEEL (2004). *Finite Mixture Models*, Wiley-Interscience.

NORTH, DOUGLASS CECIL (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge ; New York, Cambridge University Press.

O'RIORDAN, TIMOTHY (1994). *Interpreting the Precautionary Principle*, Earthscan.

RACHLINSKI, JEFFREY J (2010). "Evidence-based Law." *Cornell L. Rev.* **96**: 901-924.

SCHARPF, FRITZ WILHELM (1999). *Governing in Europe. Effective and Democratic?* Oxford ; New York, Oxford University Press.

SCHNEIDER, CARL E AND LEE E TEITELBAUM (2006). "Life's Golden Tree. Empirical Scholarship and American Law." *Utah Law Review*: 53-106.

SCHWEIZER, MARK (2015). *Beweiswürdigung und Beweismaß*. Tübingen, Mohr.

SPAMANN, HOLGER (2015). "Empirical Comparative Law." *Annual Review of Law and Social Science* **11**: 131-153.

SUNSTEIN, CASS R (2005). *Laws of Fear. Beyond the Precautionary Principle*, Cambridge University Press.

TALLEY, ERIC AND DREW O'KANE (2012). "The Measure of a MAC. A Quasi-Experimental Protocol for Tokenizing Force Majeure Clauses in M&A Agreements." *Journal of Institutional and Theoretical Economics* **168**: ***.

TOR, AVISHALOM (2008). "The Methodology of the Behavioral Analysis of Law." *Haifa Law Review* **4**: 237-327.

TOWFIGH, EMANUEL V AND NIELS PETERSEN (2015). *Economic Methods for Lawyers*, Edward Elgar Publishing.

VON HAYEK, FRIEDRICH-AUGUST (1945). "The Use of Knowledge in Society." *American Economic Review* **35**: 519-530.

VON HAYEK, FRIEDRICH-AUGUST (1969). Der Wettbewerb als Entdeckungsverfahren. *Freiburger Studien. Gesammelte Aufsätze von Friedrich-August von Hayek*. F.-A. von Hayek. Tübingen, Mohr**:** 249-265.

WESTLAND, J CHRISTOPHER (2015). *Structural Equation Models. From Paths to Networks*.