Asian Journal of Economics and Banking
ISSN 2615-9821
http://ajeb.buh.edu.vn/Home

Decision Making versus Testing: A Changing of the Guards in Empirical Research?

William M. Briggs[1†], Hung T. Nguyen[2,3]

[1]Independent Researcher, New York, NY, USA
[2]Department of Mathematical Sciences, New Mexico State University, USA
[3]Faculty of Economics, Chiang Mai University, Thailand

Article Info

Received: 4/5/2020
Accepted: 16/6/2020
Available online: In Press

Keywords

Bayesian tests, Hypothesis testing, Loss functions, P-values, Statistical decision theory, Utility functions, Wald statistical Decision functions

JEL classification

A23, B41, C12

MSC2020 classification

62C05, 62C20, 62F03

Abstract

In this paper we will elaborate and give support to Manski's central message in "Statistical decision theory should supplant hypothesis tesing" (Manksi [15]; see also Manski [16]). Decisions should take the place of testing. This review should benefit practioners in empirical research of econometrics and in statistics in general. We emphasize that Wald's Statistical Decision Theory should be used in empirical research instead of "traditional" hypothesis testing (both frequentist and Bayesian testing), since hypothesis testing is only a one-sized-fits-all way to make decisions. The goal of empirical research is to make optimal decisions, and it is possible to make decisions without using testing.

†Corresponding author: William M. Briggs, Independent Researcher, New York, NY, USA. Email address: matt@wmbriggs.com

1 INTRODUCTION

In one sense, this paper is a follow-up on Briggs and Nguyen [4], concerning abandoning the use of P-values in frequentist testing. As spelled out in The American Statistician 73(SI), 2019 (see also Wasserstein et al. [23], [24]), all discussions concerning this critical issue (a bread-and-butter tool in empirical research) clearly stepped on many big toes, and led to the urgent question "Where do we go from now?" The simplest answer to the question "How to test if we must?" might be "Use Bayesian testing". Yet does such a simple answer satisfy all statisticians and econometricians, especially frequentist ones? Is prior information in Bayesian statistics reliable or a problem? It seems that a better question might be "Is there a way to conduct statistical testing which is acceptable in all situations?" The answer could be "Do not test", Briggs [5]. Why?

Let's ask ourself first what the purpose of a test is. We make a test to make a decision. As Briggs [5] said loud and clear "Don't test, decide". So the better question is "Can we do decision-making without testing?".

This brings us to Wald's statistical decision theory (Wald [22]). In fact, among all discussions about the P-values problem in Wasserstein et al. [24], the paper contributed by Charles Manski [15], in our view, is the best. Rather than arguing whether the use of P-values in carrying out frequentist testing is acceptable or not, he addressed a more general problem, namely "Don't test", with the precise message that "statistical decision theory should supplant hypothesis testing".

In this paper, we will elaborare on Manski's paper (Manski [16]) on the one hand, and on the other hand provide what amounts to a tutorial on how to make decisions without using hypothesis testing. Our hope is this. Abandoning P-values will not be a traumatic experience after one learns Wald's statistical decision theory.

This paper is structured as follows. In Section 2, we elaborate on the main goal of empirical research. In Section 3, we revisit Wald's statistical decision theory. In Section 4, we focus on the main question "Can we make decisions without statstical testing?".

2 IT'S ALL ABOUT DECISIONS

Statistics is a man-made "science" used mainly for the social sciences, especially economics, where empirical research aims at providing information for making decisions. The main component of econometrics, as initiated by Haavelmo [9], is statistics where currently we mean by *statistics* the methods for estimation, testing and prediction. "Outcomes" of statistical analyses are "assertions" about unobservable things. Indirectly, such assertions will help to make decisions which are clearly the final goal of all analyses. Without focusing on the "final" goal of an econometric analysis, the practice of statistics is *evidential* rather than *behavioral*. What is a *behavioral statistics*? Do not confuse this question with "statistics for behavioral sciences"! In fact, a behavioral science, like economics, concerns

the study of how people behave (i.e., how they make decisions). Currently, evidential statistics is used to study behavior indirectly. In microeconomics, axiomatic decision theory (e.g., in game theory) provides a direct way to study economic agents' behavior (in making decisions). Thus, even neoclassical microeconomics is behavioral in nature.

The rise of *behavioral economics* reveals that not only social sciences should be behavioral, i.e., focusing directly on decision making under uncertainty, but also that the notion of uncertainty should be modeled appropriately from experimental evidence, rather than just by the "standard probability theory" (Kolmogorov). Remember "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means" (Bertrand Russel, 1929), though this may be better phrased as there is little agreement as to what it means. Discussions about the important problem of uncertainty modeling in social sciences will be for another day.

In this paper, we only focus on the behavioral aspects of, say, economics, when we use statistics. *Behavioral econometrics* should use statistics to reach economic decisions directly: statisticians and econometricians should help us what to do, and not what to say.

What kind of statistics fits this description? It should be *behavioral statistics!*; i.e., a statistical theory which helps us to make decisions. Some might say that the current evidental statistics with which all are familiar does help us to make decisions, such as in the testing of statistical hypotheses. This is correct, but with several important exceptions. First, using testing to reach decisions depends upon the reliability of the testing results. If frequentist testing is used, then the only way to carry out a test is using Fisher's notion of P-values, and it has been clearly stated in the whole statistical community that such a testing procedure is invalid. Secondly, testing, such as in an acceptance sampling, is just an example of a possible axiomatic decision framework (as well as for Bayesian testing). A general and axiomatic decision theory, designed specifically for decision-making in behavioral sciences, is needed to handle general, real applications.

A behavioral statistics is a statistics where statisticians use a decision theory to make decisions! We already have such a decision theory, too. This is Wald's Statistical Decision Theory (SDT), first introduced in 1950. It is a behavioral theory. Thus, if we want a definition of behavioral statistics, we could just say "It is the practice of statistics based upon Wald's SDT". As such, we see at present a renewed interests in SDT because it is consistent with behavioral economics and econometrics.

Why a "renewal" of SDT? It can be said that SDT was dominated by the ease and general use of Neyman-Pearson, theory which was and as the go-to practical theory of statistical inference for empirical research. Wald's SDT is about statistical actions/making decisions, and not directly about inference, where inference means making assertions on the basis of incomplete information. The reappearence of STD

comes at just the right time.

What is the purpose of a decision theory? It seems like a naive question! Not quite. A decision theory can either tell us how to make optimal decisions in prescribed circumstances; e.g., how to choose items to purchase, as in a choice theory. Or it simply tries to describe our behavior when we face decisions to make. A good analogy is the natural sciences. We see clearly that physicists try to understand how nature behaves, rather than to tell nature how to comport herself!

In the social sciences, consumers play the role of nature, of course. Here, it is possible to tell people what to do in taking actions and making decisions. But *how* should they be told? Should we first try to figure out their behavior? Like in the natural sciences, understanding consumers' behavior will help, say, marketing economists to do a better job.

This is precisely the spirit of behavioral economics, in which experimental evidence is the guide to understand consumer behavior. Such spirit will lead to new improved models in social sciences like economics, over those we use at present. For example, consumer behavior could be described by a "choice wave model" inspired from an analogy with Schrodinger's wave equation in quantum mechanics; e.g., Johnson [12]. This is a quantum model for social sciences.

Thus, a classical choice theory (see a Text like Kreps [13]) or a newly developed decision theory both should serve the aim of making-decision. In other words, it is all about decisions.

Now comes to the basic question: How to make a decision?

Consider the case where we can gather data to help us to make decisions. This is seen as the traditional territory of statistics. Yet is the concept of "decision" universal and well understood and "self evident" to everybody? The answer to this question is important, as it leads to how to best make a decision. It turns out that the answer is no, as spelled out in complete details in an old philosophical paper by Birnbaum (Foundations of Probability and Statistics; [2]).

In common language, making a decision is choosing an action to do something. There is a concrete literal sense as in acceptance sampling (to be elaborated shortly). It is the *behavioral interpretation* of the notion of decision. A constrasting interpretation of decision is *evidential* , which is based on Neyman-Pearson testing theory; i.e., a statistical decision (a decision based on statistical data) is a statistical test whose outcome dictates the decision to make, regardless of circustance. The behavioral perspective found many early critics. Birnbaum said it "must be criticized and rejected, in the view of many investigators and statisticians". Yet it turns out that today, when we turn to behavioral economics, which is supplanting neoclassical economics, it is the "behavioral" interpretation of everything that matters! So why do statisticians still prefer the evidential interpretation of decision? Because they *equate* decision theory with hypothesis testing theories!

Here is what Freedman et al. ([8], p. 562-563) said regarding hypothesis testing "Often, tests of significance

turn out to answer the wrong question. Therefore, many problems should be addressed not by testing but by estimation...Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory...This sounds so impressive, and there is so much mathematical machinery clanking around in the background, that tests seem truly scientific – even when they are complete nonsense."

Let's look at a decision-making problem in a so-called "acceptance sampling plan". A manufacturing lot (a population of a large size N) consists of items, say, from a manufacturing process, which are good or defective. The proportion of defective items in the lot is $\theta = \frac{D}{N}$ which is unknown since the number of defectives D is unknown. The producer believes that $\theta \leq \theta_o$ (a known number). If that is true, the producer considers this as a good lot and expects to sell it. A consumer is unwilling to buy the lot if there are too many defectives in it, say, when $\theta \geq \theta_a$ (a known number greater than θ_o). How does the consumer decides whether or not to buy the lot? He might insist on examining a sample of items, say, of size n. Suppose the producer and the consumer agree that the consumer can refuse the lot if the number of defective items X in the sample exceeds some threshold c. Thus, (n, c) constitues a sampling plan. How to specify an acceptable sampling plan? Clearly, the producer and consumer's "risks" are, respectively,

$$P(\text{consumer refuses the lot} \mid \text{the lot is good}) = P(X \geq c|\theta \leq \theta_o),$$

$$P(\text{consumer buys the lot} \mid \text{the lot is not good}) = P(X < c|\theta \geq \theta_a).$$

Of course, each wants their own risk to be as small as possible, say, bounded by α, β, respectively. Thus, an acceptance sampling is (n, c) such that

$$P(X \geq c|\theta \leq \theta_o) \leq \alpha,$$
$$P(X < c|\theta \geq \theta_a) \leq \beta.$$

Once an acceptance sampling is determined, the consumer's decision is dictated by the observed data. This "procedure" of making decisions can be formulated in the context of hypothesis testing. However, rejection of H_o does not mean that H_o is false, and not rejecting H_o does not mean that H_o is true. We can prove nothing either way with statistics.

Remark. To actually determine (n, c), observe that X follows a hypergeometric distribution with density being a non-decreasing function of θ or D when $X = x$ is large, so that to find $(n.c)$, it suffices to solve the simpler inequalities:

$$P(X \geq c|\theta = \theta_o) \leq \alpha,$$
$$P(X < c|\theta = \theta_a) \leq \beta.$$

The decision flows simply from the probabilty.

Can we make decisions without testing?

Here is a typical situation. The problem of treatment choice in clinical trials is a decision-making problem: choosing the status quo or an innovative treatment. How to decide which one to use? Facing this problem, statisticians use testing to decide. The outcome of

the testing problem dictates the decision to make. It is one possible way to make decisions. The difficulty is that it is wrong logically since, as it is well-known by now, the use of P- values to carry out a (frequentist) test is a form of "proof by contradiction" which, unfortunately, is not valid outside binary logic. If you do not buy it, then read Ioannidis [11] or Briggs [3]! The checking of results using testing to make decisions in the Ioannidis paper has triggered much ado with P-values, see Wasserstein et al. ([23], [24]), also Hurlbert and Lombardi [10]. So, while it is a possible way to make decisions as a conventional practice, hypothesis testing has many clear and serious deficiencies, and hence it is about time to abandon it. Another clear difficulty with hypothesis testing is that it cannot take into account the costs and losses of making correct and incorrect decisions. Hypothesis testing makes one-size-fits-all decisions, where the "all" is taken quite literally.

Remark. One more thing we should point out to those who used to use frequentist hypothesis testing to make decisions. It is precisely the use of P-values to carry out tests that statisticians can do P- hacking as they please, as exemplified, e.g., in causal inference in economics, in Brodeur et al. [6]. It is far too easy to "game" a decision in the direction the test maker wants, and not in the direction dictated by the evidence. Such bad practices can be avoided if decision theory supplants testing. In this way, the same set of data and the same model of that data can lead to different decisions, because opposite testing different loss functions can be incorporated.

Note also that, in an effort to "improve" standard hypothesis testing, another alternative, called "No-decision classification", was proposed in [1].

So how to make decisions (directly) without testing? Well, as Minski and Tetenov (2016) spelled it out loudly (see also Manski ([15], [16]), we can make decisions directly by using Wald's statistical decision theory, Wald [22]. We will elaborate on this after briefly recalling Wald's decision theory.

3 WALD'S STATISTICAL DECISION THEORY REVISITED

It should be noted that while Neyman-Pearson's decision theory (i.e., their hypothesis testing) can be considered as an example of Wald's decision theory, but only as an example. See e.g., Spanos [18]. Putting it differently, Wald's decision theory is much more general and we should not use his theory via squeezed into just this one example of equating decision theory with hypothesis testing. Unfortunately, that was exactly what happened historically in statistics. As spelled out in Manski [16], "Neyman - Pearson hypothesis testing does not provide satisfactory guidance for decision making". The point is this that it is time, in view of Manski's work, to look at Wald's decision theory for what it really is. And that is important since it could provide statisticians with the correct tool to do trusted empirical research.

In fact, new perspectives on statistical decisions surface in current renewed interest in statistical decision theory,

e.g., in the context of decision making under ambiguity, Knight uncertainty in economics, especially in the face of partial identification problem. See, e.g., Stoye [19]. Note that by ambiguity we mean the type of uncertainty which does not behave like standard probabilistic uncertainty, e.g., not additive or not commutative (like quantum probability).

Wald's Statistical Decision Theory (SDT), Wald [22], provides a general framework for the use of sample data to make decisions. SDT has no direct connection to statistical inference, though it can be used to make inferences. It is essentially based on the intuitive notion of a statistical decision function (SDF) which maps data to actions. Examples of SDF are point estimators of parameters, statistical inference procedures and forecasting methods. It can be said that SDT seems to have vanished from statistical practices since statisticians viewed their objectives in empirical research as comprising almost wholly estimation and testing, tasks whose foundations are distinct from the decision problem, for example, results from tests are remote from decision-making. This is odd because the point of the analyses was to provide guidance on how to make decisions.

There is now a renewed interest in SDT for a variety of reasons. Our emphasis here is on a very practical goal: It is not just that making decisions where more general procedures than testing should be considered, but also because SDT, considered even as a frequentist approach, could supplant frequentist testing when frequentist statisticians do not know what to do when testing is no longer a trusted tool to do empirical research (again, because of the use of P-values to carry out tests). If one is determined to be a committed frequentist, decision theory should be used over testing because decision theory eliminates many of the errors and inconsistencies found in testing.

Here is an example to illustrate how a general decision framework can be used for decision making in incomplete information situations, such as in robust Bayesian analysis, using a classic view of probability. We do not necessarily recommend this view, but present this analysis in awareness that many do. For simplicity, suppose the state of nature in a given decision problem is a finite set Θ, and the prior information about the true state of nature $\theta_o \in \Theta$ is given, not as a specific probability distribution on Θ, but as a subset $\mathcal{P}$ of the set Λ of all probability distributions on Θ. For example, let $\{\Theta_1, \Theta_2, ..., \Theta_k\}$ be a finite partition of Θ; and $\mathcal{P} = \{P \in \Lambda : P(\Theta_j) = \alpha_j,\ j = 1, 2, ..., k\}$, where $\alpha_j \geq 0, \sum_{j=1}^{k} \alpha_j = 1$, and these weights are given. Consider the decision problem of choosing the best action among the set of feasible actions $\mathbb{A}$ with unknown state of nature θ_o. Given a utility function $u : \mathbb{A} \times \Theta \to \mathbb{R}$, where $u(a, \theta)$ the the "payoff", we use the expected value $E_Q u(a, .)$, with $Q \in \mathcal{P}$, to guide our decision of choose the action. For a given $Q \in \mathcal{P}$, the optimal action (in this context) is $\arg\max_{a \in \mathbb{A}} E_Q[u(a, .)]$. Now, Q is only known to be in $\mathcal{P}$, we could consider the optimal action as $\arg\max_{a \in \mathbb{A}} \inf_{Q \in \mathcal{P}} E_Q[u(a, .)]$. But it can be shown that $\inf_{Q \in \mathcal{P}} E_Q[u(a, .)]$ is

in fact $\min_{Q\in\mathcal{P}} E_Q[u(a,.)]$, i.e., the infimum, with respect to $Q \in \mathcal{P}$, is attained at some $Q' \in \mathcal{P}$, say. The upshot is that $E_{Q'}[u(a,.)]$ can be computed from the knowledge of $\mathcal{P}$, as follows.

First, observe that if we let $g(.) : 2^\Theta \to [0,1]$ be $g(B) = \inf\{P(B) : P \in \mathcal{P}\}$, then it can be verified that $\mathcal{P} = \{P \in \Lambda : P(.) \geq g(.)\}$.

Next, the set function $g(.)$ is increasing, and hence it can be used to define a Choquet integral for $u(a,.)$, in general, as

$$C_g[u(a,.)] = \int_0^\infty g[u(a,.) > t]dt + \int_{-\infty}^0 (g[u(a,.) > t] - 1)dt$$

Note that for $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$, this Choquet integral is very simple: Assuming that $u(a,\theta_1) \leq u(a,\theta_2) \leq ... \leq u(a,\theta_n)$, then

$$C_g[u(a,.)] = \sum_{i=1}^n u(a,\theta_i)[g(\{\theta_i, \theta_{i+1}, ..., \theta_n\}) - g(\{\theta_{i+1}, \theta_{i+2}, ..., \theta_n\})]$$

It remains simply by verifying that $E_{Q'}[u(a,.)] = C_g[u(a,.)]$, from which the optimal decision problem can be solved by optimization. For more details, see Nguyen (2006).

Now, for easy of discussions on the main supporting message of this paper, namely we should use decision theory to make decisions, and abandon the "professional habit" of making decisions by hypothesis testing, let's review briefly Wald's Statistical Decision Theory (SDT), Wald [22]. See also Ferguson [7].

The elementary example of estimating the parameter of a Bernoulli model, based on a random sample, is in fact an example of a general statistical decision problem.

Let X be a Bernoulli random variable with unknown "probability of sucess" $\theta_o \in \Theta = [0,1]$, the space of states of nature. By tossing the (biased) coin n times, we observe the data $(X_1, X_2, ..., X_n)$ from the sample space $\mathcal{X} = \{0,1\}^n$. Suppose we wish to estimate θ_o after seeing the data. This estimation problem is viewed as a decision-making problem, a natural context, because it assumes more data than just the sample will be seen. If only the data in the sample will ever exist, then the "decision" is simple: we just look at what happened.

Using an estimate of θ_o is equivalent to taking an action in the space of possible actions $\mathbb{A} = [0,1]$. An estimator of θ_o is a statistic $\delta(X_1, X_2, ..., X_n) \in \Theta$, i.e., a map $\delta(.) : \mathcal{X} \to \mathbb{A}$ which acts like a *decision rule*. Of course, among many possible ways to estimate θ_o,we seek the "best" (optimal) one! But then, "best" should be defined. In the context of estimation, it is the notion of estimation error that is used for comparison. If the state of nature is θ, and we use the estimator (decision rule) $\delta(X)$ to estimate it, then the error could be, but need *not* be, of the form $(\delta(X) - \theta)^2$ which is viewed as a "loss", so that a *loss function* is chosen, such as $L(.,.) : \mathbb{A} \times \Theta \to \mathbb{R}^+$, $L(a,\theta) = (a-\theta)^2$. The loss could, and should, be tailored to the individual decision maker's situation. To compare estimators, we use the concept of *risk* associated with a chosen loss function, such as $E_\theta(\delta(X) - \theta)^2 = E_\theta L(\delta(X), \theta) = r(\delta, \theta)$, where the risk function $r(.,.) : D \times \Theta \to \mathbb{R}^+$ is defined in terms of a chosen loss function L, and D is the set of

all feasible decision rules (estimators).

With respect to the above mean squared error concept, an estimator δ^* is preferred to δ if $r(\delta^*, \theta) \leq r(\delta, \theta)$, for all $\theta \in \Theta$, with at least strict inequality for one θ. A decision rule δ is inadmissibe if there is another δ^* which is preferred to it. An *admissible* decision rule (estimator, here) is one which is not inadmissible. Therefore, looking for the best estimator (in a given framework) is an optimal decision making problem.

We can view Wald's statistical decision theory as a general formulation of the above setting for probability models of any stripe.

Consider a decision problem consisting of making a decision by choosing an action in a given set of feasible actions $\mathbb{A}$ based on the information provided by the data observed from a random variable X taking values in a set $\mathcal{X}$. The law governing the random evolution of X is a probability measure on $\mathcal{X}$. For technical rigor, each space $\mathbb{A}, \mathcal{X}$, is equiped with a $\sigma-$ field $\mathcal{B}(\mathbb{A}), \mathcal{B}(\mathcal{X})$, respectively. Of course there is also a probability space $(\Omega, \mathcal{A}, P)$ in the background, where $X(.) : \Omega \to \mathcal{X}$. The probability law of the observable X (the DGP) is only specified as one among a set of possible probability measures on $\mathcal{B}(\mathcal{X})$, indexed by the set of states of nature $(\Theta, \mathcal{B}(\Theta))$, namely $\mathcal{P} = \{P_\theta : \theta \in \Theta)$, the true state of nature, denoted as $\theta_o \in \Theta$, is unknown. The problem is how to choose an "optimal" action in $\mathbb{A}$ when observing the data from X (informative for θ_o). Mathematically, this boils down to choose a (measurable) function $\delta(.) : \mathcal{X} \to \mathbb{A}$, called a decision rule. To investigate the optimal choice of $\delta(.)$, we consider a loss function $L(.,.) : \mathbb{A} \times \Theta \to \mathbb{R}^+$, and its associated risk $r(.,.) : \mathbb{D} \times \Theta \to \mathbb{R}^+$, $r(\delta, \theta) = E_\theta L(\delta(X), \theta)$, where $\mathbb{D}$ is the set of all feasible decision rules. As in the estimation setting, admissibility of decision rules is defined similarly, as well as optimal decision solutions. A popular procedure for obtaining optimal decisions is Wald's minimax : Minimizing the maximum risk, i.e., $\arg\min_\delta \max_\theta r(\delta, \theta)$.

Wald's SDT can be viewed as frequentist but behavioral. Specifically, *behavioral statistics* is defined as the science of decision making under uncertainty, rather than just estimation and testing. Thus, when considering today behavioral economics, it is Wald's SDT with its general decision theoretic viewpoin, which will form the core of the associated *behavioral econometrics*.

4 HOW TO MAKE DECISIONS WITHOUT TESTING ?

From the above discussions, we know the answer to this question, namely using a decision theory! Before elaborating in some more details on how practically that can be carried out, let us be clear about why the question arises.

In empirical research, statisticians or econometricians focus on statistical hypothesis testing (frequentist or Bayesian), although the purpose of doing so is to make decisions. For example, a Dickey-Fuller (frequentist) test is meant to ascertain whether the autoregressive model of a time series data has a unit root. What is this test for specifically? It is for deciding whether the

time series is stationary or not. The decision making (under uncertainty) process is left to hypothesis testing. This might seem an adequate procedure. Not quite, since it depends on whether testing is a valid procedure to use or not. While there seems to be few critiscisms regarding Bayesian testing based on Bayes factors, which require priors, most acknowledge problems with frequentist hypothesis testing, or more specifically, as we all know by now, with the way to carry out frequentist tests (using P-values).

Ss far as frequentist statistics is concerned, since testing is no longer valid, we can no longer say "to make decisions, we just do testing". But we still need to make decisions! Obviously, when a situation such as this happens, statisticians must look for other alternatives. Rather than trying to find alternative *tests*, why don't we ask "Can we make decisions without testing?". It turns out we already have the answer "Use Wald's statistical decision theory"(SDT) which was designed specifically to make directly decisions in frequentist statistics (without the need of testing, although a testing setting can be viewed, formally, as an example of SDT format).

Thus, not only does SDT provide a way to make decisions in general situations, but practically, it is the "alternative procedure" to make decisions without P-values!

Here is an example of empirical research on causal inference which does not need hypothesis testing.

Classically, what is said to be the causal effect of an intervention is usually analysed as follows. If two groups t (treatment) and c (control) are completely "similar" probabilistically, then the difference in (sample) means (of course, assuming that both populations have finite means and variances), such as normals, are typically used to declare the existence of causal effect. This is the case of experimental design called the "golden standard" of causal inference.

Most observational studies involve human subjects in psychology, for instance, so that we cannot attain this "golden standard". Something else is needed. A "new" design to replace random samples, and to produce observational data for investigating causal effects, is called the Regression Discontinuity Design (RDD), invented by the psychologists D. Thistlethwaite and D. Campbell, in 1960 (see Thistlethwaite and Campbell [20]).

Recently, in view of L'Hopital's rule in calculus, RDD is extended to Regression Kink Design (RKD). This is the modern setting for causal effect analysis. For more details, see the Appendix.

Note that there is no need of *testing* to identify causal effects. What happens if we use testing? In the case of Granger causality for example, Thursman and Fisher [21] investigated the question "Eggs and chickens, which came first?" By using Granger causality tests (a frequentist test using P-values!), they concluded that "the egg came first"!

Wald's Statistical Decision Theory (SDT) does not necessarily involve statistical inference (i.e., statistical testing or estimation). It provides a direct way to make decisions without resorting to statistical inference. Of course, SDT can be implemented as a statisti-

cal inference, providing an indirect way to reach decisions, if statistical inference procedures so used are valid! Using testing as an indirect way to make decisions is what Manski [16] called *inference-based* procedures. Thus, we should not ask "Is it possible to make decisions under (statistical) uncertainty without using statistical hypothesis testing?"!

A general decision making under uncertainty scenario is this. A decision making problem in a given situation you face is for some specific purposes, i.e., you know why and what kinds of decisions you need to make or to suggest (recommend) to a policy maker. You need to choose an action in a set (collection) of feasible actions $\mathbb{A}$ to acheive some specific goal. The action you are going to choose is related to an environment, called nature. The true state of nature θ_o is unknown to you (this need not be a mathematical parameter in a probability model, but an actual state of nature). You only specify that θ_o is in a set Θ, called the state space. How to choose an "optimal" action in such a situation?

If you do not have any additional sample information (data), then you will base your decisions only on your objective. Look at your decision-making problem carefully. What is its objective? You must have this. Suppose your objective is, say, to increase your "welfare", and that welfare function depends not only on your action, but also on the state of nature, i.e., your welfare function is $\psi(.,.) : \mathbb{A} \times \Theta \to \mathbb{R}$. If you can specify your welfare function, then, of course, your optimal action is $\arg\max_{a\in\mathbb{A}} \psi(a, \theta_o)$. But, unfortunately, you do not know θ_o ! Well, you could use criteria such as *Maxmin* criterion to choose an optimal action which is

$$\arg\max_{a\in\mathbb{A}} \min_{\theta\in\Theta} \psi(a, \theta)$$

, or *Minimax-Regret* yielding

$$\arg\min_{a\in A} \max_{\theta\in\Theta} [\max_{b\in\mathbb{A}} \psi(b, \theta) - \psi(a, \theta)]$$

.

Remark. If you use a Bayesian approach in treating the unknown "parameter" θ_o as a random variable, then you could try to see whether you can obtain a (subjective) reliable prior probability measure π (on Θ) for it. In this case, your decision could be $\arg\max_{a\in\mathbb{A}} E_\pi \psi(a, \theta)$.

Consider now the case where you have sample data which can help you to make decisions. With sample data, the decision making problem is called a statistical decision problem, for which Wald's Statistical Decision Theory (SDT) was designed specifically to "solve" it.

A typical situation is this. The sample data come from a random variable X whose distribution (probability measure on the range $\mathcal{X}$ of X) is P_{θ_o}, where you only know that $\theta_o \in \Theta$, and not θ_o itself, i.e., you only know its model $\{P_\theta : \theta \in \Theta\}$. The observation X is the information (data) you have to assist you to make your desired decision, in the sense that your decision rule is a function $\delta(.) : \mathcal{X} \to \mathbb{A}$. Your decision problem boils down to "How to construct $\delta(.)$?" Note that since $\delta(.)$ will be constructed without (Bayesian) subjective prior information about θ_o, SDT is a frequentist approach to decision making.

Suppose you specify your welfare function $\psi(.,.) : \mathbb{A} \times \Theta \to \mathbb{R}$. Now, instead of choosing directly an action $a \in \mathbb{A}$, you do that according to your additional observed data X, i.e., by a statistical decision function $\delta(.) : \mathcal{X} \to \mathbb{A}$, to be chosen, so that if your data is X, then your chosen action is $\delta(X)$, resulting in your welfare of $\psi(\delta(X), \theta)$ when the state of nature is $\theta \in \Theta$.

So far, in our quest to make a decision by choosing an action according to our observed data, where we have our own specified "welfare" function $\psi(.,.)$ (e.g., returns of an investment prospect), and our statistical model $\{P_\theta : \theta \in \Theta\}$, we ask next: "Which decision rule $\delta(.)$ to choose?"

The answer is: the best one! In terms of welfare, if we choose another decision rule $\eta(.) : \mathcal{X} \to \mathbb{A}$ to guide our decision making process, yielding an welfare of $\psi(\eta(X), \theta)$ in state θ, then of course, we will prefer $\eta(.)$ than $\delta(.)$, in state θ, if $\psi(\eta(X), \theta) \geq \psi(\delta(X), \theta)$, almost surely (as random variables). This can be checked by looking at their probability distribution functions. If $F(.)$ and $G(.)$ denote the distribution functions of $V = \psi(\eta(X), \theta)$ and $W = \psi(\delta(X), \theta)$, respectively, then $F(.) \leq G(.) \Longrightarrow V \geq W$, almost surely. Indeed, let $V^* = F^{-1}(U)$, $W^* = G^{-1}(U)$, where U is uniformly distributed on $[0, 1]$, then, in one hand, $V^* = V$, $W^* = W$, in distribution, and on the other hand, $F(.) \leq G(.) \Leftrightarrow G^{-1}(.) \leq F^{-1}(.)$, so that $W^* \leq V^*$.

Remark. Of course, if $V \geq W$ (a.s.) then $F(.) \leq G(.)$, i.e., $V \geq W$ (a.s.) is equivalent to $F(.) \leq G(.)$. This (partial) order relation between random variables is known as First-Order stochastic dominance in risk analysis.

Unfortunately, stochastic dominance is not a total order relation. Therefore, we must proceed differently. Consider $E_\theta[\psi(\delta(X), \theta)]$, where E_θ denotes the expectation of the random variable $\psi(\delta(X), \theta)$, when the probability distribution of X is P_θ. "Good" statistical decision functions are those which are *admissible* in the following sense. $\delta(.)$ is inadmissible if there is $\eta(.)$ such that $E_\theta[\psi(\eta(X), \theta)] \geq E_\theta[\psi(\delta(X), \theta)]$ for all $\theta \in \Theta$, with at least one θ with strict inequality. An admissible statistical decision function is one which is not inadmissible. admissibility is, unfortunately, a weak criterion. We must focus on admissible statistical decision functions, but how to choose the "best" among them? It may happen that there is no best single one.

We can see this from a simple parameter estimation problem (an example in statistical inference formulated in SDT form). Let X be a random variable with model $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$. Given a random sample $X_1, X_2, ..., X_n$ drawn from X, a (point) estimator of some function $\varphi(\theta)$ is a statistic $T(X_1, X_2, ..., X_n)$. The natural question is "which statistic should we use to estimate $\varphi(\theta)$?". Of course, the answer is "The best one!" But then we have to spell out what we mean by "the best", and then ask "Do such estimators exist?".

Using the natural notion of error in estimation to compare estimators, suppose we consider the concept of mean squared error (MSE) $MSE_\theta(T) =$

$E_\theta[(T-\varphi(\theta))^2]$. With this criterion for comparing estimators, the best estimator S is the one such that , for any estimator T, we have $MSE_\theta(S) \leq MSE_\theta(T)$ for all $\theta \in \Theta$ (Noting that the condition "for all $\theta \in \Theta$" is required since the true $\theta_o \in \Theta$ is unknown). Suppose $\varphi(\theta_1) = \theta_1 \neq \theta_2 = \varphi(\theta_2)$. Take $T_1(X_1, X_2, ..., X_n) \equiv \theta_1$ and $T_2(X_1, X_2, ..., X_n) \equiv \theta_2$. Then $MSE_{\theta_1}(T_1) = MSE_{\theta_2}(T_2) = 0$. These (degenerate) statistics both have the smallest possible MSE but at different points of Θ, therefore there does not exist one T which has smallest possible MSE for all points of Θ. The other way around: for each $T(X_1, X_2, ..., X_n)$, there is an estimator $T^* \equiv \theta'$ which has smaller MSE than T at $\theta = \theta'$ but larger MSE than $T^{**} \equiv \theta'' \neq \theta'$. These "pathological" estimators are such that, for $j = 1, 2$, $E_\theta(T_j) \neq \varphi(\theta)$ for some $\theta \in \Theta$, and statisticians give them the name: biased estimators (just a name). It turns out that, we we do not take "biased" into account, then it is possible to obtain the best estimator (in the MSE sense) among the "unbiased estimators". Note that, as we know today, biased estimators do not mean "bad" estimators!

The difference between inference-based and direct decision making is spelled out in complete details in Manksi (2020). Roughly speaking, it is the difference between, e.g., $\arg\max_{a\in\mathbb{A}} \psi(a, T(X))$, where $T(X)$ is a consistent estimator of θ_o (for large sample size), and $\arg\max_{\delta(.)\in\Delta} \min_{\theta\in\Theta} \psi(\delta(X)$, where Δ is a set of feasible statistical decision functions. The technical difference is in the functional optimization of the latter where the calculus of variations needs to be used, but the data could be with any sample size. Thus, for realistic decision making problems, althought SDT direct method seems difficult technically, empirical researchers must now face it!

In summary, as stated as the beginning of this paper, we aim simply to remind empirical researchers that they need to be aware of advances in statistical science to make their empirical research more trusted. This is especially crucial in an era with a known reproducibility crisis and the known weaknesses and inconsistencies of hypothesis testing. Among quite a lot of such advances reported in the literature, we only focus here on the decision making problem, which is the goal of most of statistical and econometrical issues. We repeat the message of Manski [15] that frequentist econometricians should "switch" their traditional habit of using hypothesis testing to make crude and universal decisions, to the direct method of flexible and extensible decision making based upon the Wald's Statistical Decision theory.

References

[1] Berg, N. (2004). No-decision classification: an alternative to testing for statistical significance, *the Journal of socio-Economics, 33*, 631-650.

[2] Birnbaum, A. (1977). The Neyman-Pearson theory as decision theory , and as

inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory, *synthese 36*(1), 19-49.

[3] Briggs, W. M. (2016). *Uncertainty: The Soul of Modeling, Probability and Statistics,* Springer.

[4] Briggs, W. M., and Nguyen, H. T. (2019). Clarifying ASA'a view on P-values in hypothesis testing, *Asian Journal of Economics and Banking 3*(2), 1-16.

[5] Briggs, W. M. (2020). Don't test: decide, in *Behavioral Predictive Modeling in Econometrics,* Springer Series "Studies in Computional Intelligence", to appear.

[6] Brodeur, A., Cook, N., & Heyes, A. (2018). Methods matter: P-hacking and causal inference in economics, *Z A Institute of Labor Economics*, *IZA DP No 11796.*

[7] Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach,* Academic Press.

[8] Freedman, D., Pisani, R., and Purves, R. (2002). *Statistics,* W.W. Norton & Company.

[9] Haavelmo, T. (1944). The probability approach in econometrics, *Econometrica, 12,* Supplement, 1-115.

[10] B Hurlbert, S. H. and Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and the rise of the neofisherian, *Ann. Zool. Fennici 46*, 311-349.

[11] Ioannidis, J. P. (2005). Why most published research findings are false, *PLoS Med. 2(8) e124.*

[12] Johnson, R. (2016). A probabilistic demand application in the American cracker market, *Intern. J. Food and Agricultural Economics 4*(3), 49-61.

[13] Kreps, D. M. (1988). *Notes on The Theory of Choice,* Westview Press.

[14] Manski, C. F., & Tetenov, A. (2016). Sufficient size to inform clinical practice, *Proc. Nat. Acad. Sciences, 113*(38), 10518-10523.

[15] Manski, C. F. (2019). Treatment choice with trial data: statistical decision theory should supplant hypothesis testing, *The American Statistician, 73*, 296-304.

[16] Manski, C. F. (2020). Econometrics for decision making, Google.

[17] Owen, A. B. (2001). *Empirical Likelihood,* Chapman and Hall / CRC Press.

[18] Spanos, A. (2016). Why the decision-theoretic perspective mispresents frequentist inference: Nuts and bolts vs learning from data, Google.

[19] Stoye, J. (2011). New perspectives on statistical decisions under ambiguity, Google.

[20] Thistlethwaite, D., & Campbell, D. (1960). Regression discontinuity analysis: an alternative to the ex post facto experiment, *Journal of Educational Psychology, 51,* 309-317.

[21] Thursman, W. N., & Fisher, M. E. (1988). Chickens, eggs, and causality, or which came first?, *american Journal od Agricultural Economics, 70*(2), 237-238.

[22] Wald, A. (1950). *Statistical Decision Functions,* J. Wiley.

[23] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on P-values: Context, process, and purpose, *The American Statistician, 70*, 120-133.

[24] Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Editorial: Moving to a world beyond $p < 0.05$, *The American Statistician, 73*(51), 1-19.

An Appendix on Causal Inference

A transition from conventional association /correlation inference to causal inference could address more meaningful questions of greater interest, especially in the area of risk management.

For applied statisticians as well as for econometricians, it is about time to focus more on causal inference, i.e., asking questions concerning either the causes of effects, or effects of the causes (causal effects), rather than keep asking questions about correlations among variables only, remembering that "association is not causation".

As it is well known, the main difficulties with causal inference are twofold:

(i) Are we concerned with "causes of effects" or "effects of the causes"? Clearly, the problem of "causes of effects" (such as "are eggs the cause of chickens" or is it the other way around? or "does smoking cause lung cancer?) is somewhat delicate since, even in philosophy, where the very meaning of a cause is not universally accepted. For example, in econometrics, Granger defined an "econometric causality concept".

Let's consider the other problem of "causal effects", i.e., to figure out whether some "intervention" (treatment) will have an effect on some response variable of interest.

(ii) Statistical inference is based on data. Unlike association inference, we face often the missing data problem in causal inference. Data are often observational as opposed to experimental.

A now popular design of observational studies to discuss causal inference is called Regression Discontinuity Designs (both sharp and fuzzy).

After problems (say, a question concerning causal effect of some treatment), and collected data (from an observational study, like RDD), we need to look for statistical tools to finish our job! Since more "assumptions" on statistical models only get us far away from realities, we should strike to employ more general and robust statistical method in making inferenced (estimation/point and confidence regions, testing, forecasting) about the possible causal effect under study. To complete the picture, we invoke the not so well-known (but powerful) nonparametric method of inference, known as Empirical Likelihood.

In associational inference, via, say regression analysis, we gather data on individuals of a population U on a regressor variable X and a response variable Y, on each unit i of a, say, random sample, of size n, from U resulting in the observed data $\{(X_i, Y_i) : i = 1, 2, ...n\}$. Note that data of this type is obtained from a random experiment.

Using MSE as a concept of error, we "regress" Y on X (i.e., moving backward to the mean) by looking as $E(Y|X)$. With the data available, we could run a nonparametric regression to estimate $E(Y|X)$, from which we can assess the correlation (which may not be causal) between X and Y, or its strength (via Pearson

correlation coefficient) and predict future values of Y from observed X.

Knowing that there is an association between X and Y does not always imply that X causes Y (X is a cause of Y), in view of possible confounders. In causal inference, specifically, in causal effect inference, we ask another types of questions, such as:

(i) What should be done to increase mathematical ability of students?

(ii) How to evaluate the effectiveness of a new medical drug?

Causal inference concerns what would happen to an "outcome" (response) Y as a result of a "treatment" or "intervention". More specifically, it concerns the comparison of a treatment with something else, say "no treatment" or "control". In this context, the treatment is viewd as a possible "cause" of the effect Y.

The question is: how to figure out whether there is a causal effect? We need data!

Suppose we have a sample (random or not) of size n of units from a population U to spit into two groups t (treatment) and c (control). Unlike associational inference, each unit i now has two potential outcomes: Y_{1i} and Y_{0i}, representing the outcome on the unit i if it is exposed to t and c, respectively.

The individual treatment effect is obviously the difference $Y_{1i} - Y_{0i}$. However, we cannot observe both Y_{1i} *and* Y_{0i}, but can only observe one of them.

When $i \in t$, we may wish to substitute to the unoberved Y_{0i} by some $j \in c$ which is "similar" to i, say, in terms of other characteristcs. This is possible if the observation study is can be conducted by a random process which tends to "balance out" similarity , so that counterfactuals can be answered.

If we let D be the assignment rule: $D_i = 1$ or 0 according to $i \in t$ or $i \in c$, then the "regression" observed model is

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

In applications, unfortunately, data in causal inference are often observational rather than experimental. The investigator has no control over the way the selected sample is spit into groups. She just watched what happend. Units in two groups are not similar at all (due to concomittants which are confounders not affected by the separation of groups, but can affect the outcome variable). And this is the difficulty. Of course, as for the problem of missing data, there exist statistical techniques to approach this situation.

Regression Discontinuity Design (RDD) was initiated in Psychology in 1960 but only became popular in econometrics since 2001 because of the emergence of a "theory" for it. Remember "without theory, we are blind with data"!.

The typical problem is this. We wish to evaluate the effect of a program such as giving scholarships to a group of students in the hope to increase their academic performance.

For this purpose, we give a test to a group of students and set up the assignment rule as: if the test score X_i of student i exceeds a threshold x_o (known), then give

a scholarship to i (i.e., assign i to t), otherwise assign $i \in c$ (no scholarship). some years later, a "post" test is given with observed Y_i to all students in both groups.

Here the assignement rule D is not random and is not under the control of the evaluator.

$$D_i = 1_{(X_i \geq x_o)}$$

In such a situation, how to "identify" the treatment effect? and how to estimate it? We need first some "theories" which will guide us to do the "right" things! The observed model is

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

If we plot the data, then we see two pictures

a) Plotting X_i versus D_i: there is a "sharp" jump of the assignment at the cutoff point $x_o : P(D_i = 1|X_i)$ jumps from 0 to 1 at x_o,

b) Plotting X_i versus Y_i : there is a discontinuity of $E(Y_i|X_i)$ at x_o which could be used to determine a causal effect

Specifically, the question is: How to estimate the causal effect $E(Y_1 - Y_0)$ from data $(D_i, Y_i : i = 1, 2, ..., n)$? Let $a_i = Y_{0i}$, and $b_i = Y_{1i} - Y_{0i}$, then our observed model is

$$Y_i = a_i + b_i D_i$$

Consider first the "sharp" design where the assignment rule $D = 1_{(X \geq x_o)}$ where X is a concomittant variable.

The population parameter b_i is said to be (nonparametrically) identifiable if we can express it uniquely in an "estimable" fashion. A special case is when treatment effect is constant throughout the population, i.e., when $b_i = b$ for all i. In this case, the following condition (to be tested) is sufficient for identification:

Condition (A1): The function $x \to E(Y_{0i}|X_i = x)$ is continuous at x_o.

Theorem 1: Under (A1), the constant treatment effect b is identified as $b = Y^+ - Y^-$ where

$$Y^+ = \lim_{x \to x_o +} E(Y_i|X_i = x)$$

$$Y^- = \lim_{x \to x_o -} E(Y_i|X_i = x)$$

Proof:

For $\partial > 0$, we have

$$E(Y_i|X_i = x + \partial) - E(Y_i|X_i = x - \partial) =$$
$$E(a_i + bD_i|X_i = x + \partial) - E(a_i + bD_i|X_i = x - \partial) =$$

$$E(a_i|X_i = x+\partial) + bE(D_i|X_i = x+\partial) - E(a_i|X_i = x-\partial) - bE(D_i|X_i = x-\partial) =$$

$$b[E(D_i|X_i = x+\partial) - E(D_i|X_i = x-\partial)] + E(a_i|X_i = x+\partial) - E(a_i|X_i = x-\partial)$$

When $\partial \to 0$, the last term goes to zero by A1, the other terms go to Y^+, Y^- and the announced result follows.

When the treatment effect varies across units, additional conditions are needed for identification.

Theorem 2:

Under (A1) and

(A2): $x \to E(Y_{1i} - Y_{0i}|X_i = x)$ is continuous at x_o

(A3) : D_i is independent of $Y_{1i} - Y_{0i}$ conditional on X_i near x_o

we have

$$E(b_i|X_i = x_o) = Y^+ - Y^-$$

Proof: $\partial > 0$

$$E(Y_i|X_i = x+\partial) - E(Y_i|X_i = x-\partial) =$$
$$E(a_i + b_iD_i|X_i = x+\partial) - E(a_i + b_iD_i|X_i = x-\partial) =$$

$$E(a_i+|X_i = x+\partial) - E(a_i+|X_i = x-\partial) + E(b_iD_i|X_i = x+\partial) - E(b_iD_i|X_i = x-\partial)]$$

For small ∂, by (A3),

$$E(b_iD_i|X_i = x+\partial) = E(b_i|X_i = x+\partial)E(D_i|X_i = x+\partial)$$

$$E(b_iD_i|X_i = x-\partial) = E(b_i|X_i = x+\partial)E(D_i|X_i = x-\partial)$$

Thus, by (A1) and (A2),

$$Y^+ - Y^- = E(b_i|X_i = x_o)[\lim_{x\to x_o+} E(D_i|X_i = x_o) - \lim_{x\to x_o-} E(D_i|X_i = x_o)] =$$
$$E(b_i|X_i = x_o)$$

since $D^+ = \lim_{x\to x_o+} E(D_i|X_i = x_o) = 1$, and $D^- = \lim_{x\to x_o-} E(D_i|X_i = x_o) = 0$.

Note that, for Fuzzy RDD, $\lim_{x\to x_o+} E(D_i|X_i = x_o) - \lim_{x\to x_o-} E(D_i|X_i = x_o)$ is different than zero, and we have

$$E(b_i|X_i = x_o) = (Y^+ - Y^-)/(D^+ - D^-)$$

Under appropriate conditions of the RDD, treatment effect can be estimated, locally around the cutoff point from observed data, just like in a random experiment.

Clearly, the estimate of the treatment effect near the cutoff point is obtained as a plug-in estimator. Specifically, it suffices to estimate Y^+, Y^-, D^+, D^-. Now observe that these parameters are conditional means. As such, nonparametric regression method is used for estimation. However, beside point estimators, the problem of variance estimation for confidence interval estimation is complicated. A novel nonparametric method for confidence intervals, known as empirical likelihood {Owen is therefore called for, since this method avoids variance estimation and provides confidence regions based soly on data.

This nonparametric method seems unfamiliar to applied statisticians (and to some proportion of econometricians), let's introduce it. It can be used in a variety of situations, especially for parameters in moment condition models.

Consider the simplest (standard) setting: let $X_1, X_2, ..., X_n$ be i.i.d. drawn from a population X with unknown distribution function F_o. Since the (nonparametric) papameter space for F_o is the set (or a subset) $\mathcal{F}$ of all distribution functions, a likelihood of $F \in \mathcal{F}$, given the observations is

$$L(F|X_1, X_2, ..., X_n\ ) = \prod_{i=1}^{n}[F(X_i) - F(X_i-)] = \prod_{i=1}^{n} p_i$$

This likelihood is "consistent" with the fact that the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(X_i \le x)},$$

maximizes it. Note that

$$L(F_n|X_1, X_2, ..., X_n\ ) = (\frac{1}{n})^n$$

so that the likelihood ratio

$$r(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i$$

Suppose our parameter of interest is $\theta = T(F)$. Then the profile likelihood is

$$R(\theta) = \sup\{r(F) : F \in \mathcal{F} \cap T^{-1}(\theta)\}.$$

We reject the hypothesis $H_o : \theta = \theta_o$ when $R(\theta_o) < c$ for some c determined by the specified level of the test, and the associated confidence interval for θ_o is $\{\theta : R(\theta) \ge c\}$.

The asymptotic test is similar to classical Wilk's theorem, namely: $-2\log R(\theta_o) \xrightarrow{D} \mathcal{X}_d^2$ as $n \to \theta$.

This concept of (nonparametric) likelihood is particularly useful for setting up natural confidence intervals in moment condition models, frequently encountered in econometrics.

As an example, consider the linear model

$$Y = \theta X + e$$

where X and the error e are uncorrelated. Then we have the moment condition

$$E[X(Y - \theta X)] = 0,$$

which is of the form $E[g((X, Y), \theta] = 0$.

For a moment condition model of the form $E[g(X, \theta)] = 0$, we estimate (p, θ), where $p = (p_1, p_2, ..., p_n) \in S_{n-1}$, by $\max \prod_{i=1}^{n} p_i$ subject to $p = (p_1, p_2, ..., p_n) \in S_{n-1}$ and $\sum_{i=1}^{n} p_i g(X_i, \theta) = 0$.

Note that this optimization can be carried out in two simpler steps: First, fix $\theta \in \mathcal{O}$, and $\max \prod_{i=1}^{n} p_i$ subject to $p = (p_1, p_2, ..., p_n) \in S_{n-1}$ and $\sum_{i=1}^{n} p_i g(X_i, \theta) = 0$, resulting in a "profile likelihood" $\mathcal{L}(\theta)$. Then $\max \mathcal{L}(\theta)$ subject to $\theta \in \mathcal{O}$.

As another example, EL can be applied to Quantile Regression : For $a \in (0, 1)$, consider the linear QR at level a:

$$Y_i = \theta_a X_i + e_i$$

where the error term e is such that

$$P(e_i \leq 0 | X_i) = a$$

The data is $(X_i, Y_i), i = 1, 2, ..., n$

This model has a moment condition

$$Eg((X_i, Y_i), \theta_a) = 0$$

where

$$Eg((X_i, Y_i), \theta_a) = [1_{(Y_i \leq \theta_a X_i) - a]} X_i$$

For causal effect estimation with RDD, the EL method is applied as follows.

The estimation of Y^+ for example is by nonparametric regression of the form

$$\min_{a,b} \sum_{i: X_{ii} \geq c} K(\frac{Xi - c}{h_n})[Y_i - a - b(X_i - c)]^2$$

with first order condition of the form $\sum_{i=1}^{n} g(Y_i, Y^+) = 0$. Thus, considering this as a moment condition for the parameter Y^+, we apply empirical likelihood method to estimate Y^+ as well as setting confidence intervals.