STATISTICAL MODELS: PHENOMENOLOGICAL AND THEORETICAL

Paul Delatte

University of Southern California, *delatte@usc.edu*

We revisit the notion of a statistical model from a functional viewpoint and apply the revised notion to inferential problems framed both with and without unobservable random variables. This approach not only helps clarify statistical practices across fields but also allows us to formalize the difference between phenomenological and theoretical models in statistical inference. Phenomenological statistical models are confined to the observables whose realizations constitute the available data. Theoretical statistical models embed the observables in an explanatory framework by introducing unobservables for which no data is available. The inferential task in a theoretical statistical model is made feasible by mapping it to a phenomenological model in a procedure known as identification. The introduction of unobservables extends the reach of inferential methods from observable random phenomena to an underlying theoretical structure, but it severs the model from the sole support of the data. We illustrate these different constructions by revisiting well-known inferential problems.

Keywords: foundations of statistics; statistical model; econometric model; identification

This version: June 25, 2026. This paper supersedes two working papers previously circulated under the titles "Description and explanation" and "The role of unobservables in statistics and econometrics". We thank Eric Gautier and Jean-Pierre Florens for introducing us to econometric models. We also thank Timothy Armstrong, Julian Duggan, Nicolas Lambert, Roger Moon, Geert Ridder, Martin Weidner, and seminar audiences at the University of Southern California for useful comments.

1 Introduction

Statistical inference is commonly introduced as the problem of recovering the distribution of some random variables whose realizations are observed and recorded in a dataset. Christian Robert in Robert (2007) gives the following characterization of the problem:

> The main purpose of statistical theory is to derive from observations of a random phenomenon an inference about the probability distribution underlying this phenomenon. That is, it provides either an analysis (description) of a past phenomenon, or some predictions about a future phenomenon of a similar nature.

Erich Lehmann and Joseph Romano in Lehmann and Romano (2005) give a similar picture of the problem:

> The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables X whose distribution P_θ is at least partly unknown. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Θ, the parameter space. Statistical inference is concerned with methods of using this observational material to obtain information concerning the distribution of X or the parameter θ with which it is labeled.

To tackle statistical inference rigorously, the hypotheses on the unknown distribution are collected in a set $\{P_\theta : \theta \in \Theta\}$ that is indexed by the statistical parameter $\theta \in \Theta$ to be recovered or inferred from the data. This indexed set forms what is commonly referred to as a statistical model. This definition is ubiquitous and emerges naturally from simple problems where the family has a natural indexation that can be directly used as statistical parameters – for instance, a Gaussian family $\left\{N(\mu, \sigma^2) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)\right\}$. This construction is then informally extended to any family by forcing the statistical parameters on it as an index.

While this approach to statistical modeling is concise and pedagogical, it runs into a few problems. First, it is not always possible to index a set $\mathscr{P}$ of distributions by a given index set Θ of statistical parameters. This supposes, indeed, that there exists a surjective mapping from Θ to $\mathscr{P}$. Many inferential problems fail to satisfy this condition – a typical example is the problem of mean estimation for the class of distributions on $\mathbb{R}$ with finite first moment[1]. Second, the unknown distribution of interest in the inferential problem is usually construed as the distribution of some random variables of which realizations are part of the available data – this is clearly expressed in the two previous excerpts. This restriction to observables leaves open the question of what to do with inferential problems framed with unobservable random variables. Problems of this nature abound in applications, in particular in econometrics and biostatistics. To bridge the gap, ad hoc constructions have been introduced. In econometrics, for instance, the notion of an econometric

[1] Statisticians have addressed this issue by introducing a double parametrization (θ, ψ) where only θ is of statistical interest while ψ is viewed as a nuisance. This reparametrization solves the problem of invalid indexation but leads to a new structure that is not a statistical model proper and requires, as a result, its own frame of analysis.

model has been proposed and broadly accepted. Rosa Matzkin in Matzkin (2007) provides an authoritative definition:

> [W]e define an econometric model by a specification of variables that are observed and variables that are unobserved, variables that are determined within the model and variables that are determined outside of the model, functional relationships among all the variables, and restrictions on the functions and distributions. We will denote by S the set of all vectors of functions and distributions that satisfy the restrictions imposed by the model. We assume that for any element $\zeta \in S$, we can derive the distribution, $F_{Y,X}(\cdot;\zeta)$, of the observable vector of variables that is generated by S. The observable distribution, $F_{Y,X}$, corresponds to the true value ζ^* of ζ.

This formalization implicitly defines a statistical model for the observables in the form of an indexed family $\{P_\zeta : \zeta \in S\}$. However, this definition comes with a few difficulties. The indexation, which may now depend on unobservable quantities outside the model, is left implicit and is assumed to be always valid. The dependence of the model on a preliminary unobservable structure is entirely concentrated on the index $\zeta \in S$ but is abstracted from the statistical model proper. The possibility of computing the statistical parameter from the observables' distribution is then a property of the indexation and its correction when it fails requires stepping outside the statistical model to further constrain the unobservable structure.

All these conceptual difficulties can be dispelled by revisiting the notion of a statistical model. The main idea is to adopt a functional approach and define a statistical model as a pair $(\mathscr{P}, g)$ where $\mathscr{P}$ is a family of distributions and $g\colon \mathscr{P} \to \Gamma$ is a surjective function defining the statistical parameter to be recovered from the data. This approach is ubiquitous in nonparametric statistics, but it has never been used to properly define statistical models before. In Section 2, we make clear a few essential consequences of this approach and connect it properly to the conventional definition of a statistical model as an indexed family of distributions. We show, in particular, that no useful classification of statistical problems can be obtained without some structures on $\mathscr{P}$ in addition to the set-theoretic ones. We reject, in particular, the conventional grouping in terms of parametric, semiparametric, and nonparametric problems in favor of classifications obtained from the regularity of g when $\mathscr{P}$ is endowed with some (pre)metric structure. More clarifications naturally derive from the functional approach to statistical models – for instance, in semiparametric theory where the notions of submodels or parametric curves can be rigorously defined.

We leave these refinements for future work and focus in Section 3 and Section 4 on a conceptual problem of greater importance: the difference between statistical problems framed with and without unobservables. By decoupling $\mathscr{P}$ from g, the functional approach to statistical models can be seamlessly applied to both problems. If only observable random variables are considered, then $\mathscr{P}$ is a set of distributions on the spaces in which the observables take values. If both observable and unobservable random variables are under consideration, then $\mathscr{P}$ is a set of distributions on the product of the respective spaces in which the observables and the unobservables take values. The main difference between the two classes of models is then clearly isolated through g. For models

with observables only, the statistical parameter $\gamma = g(P) \in \Gamma$ is a function of the observables' distribution and can be immediately computed by an oracle with knowledge of P. For models with unobservables, the statistical parameter $\gamma = g(P) \in \Gamma$ is a function of the joint distribution of the observables and unobservables and may not be computable by an oracle with knowledge of the observables' distribution only. To make inference possible in models with unobservables, there needs to exist a function that maps the observables' distribution to the statistical parameter γ. This naturally defines a notion of identification for statistical models independently of any indexation for the family $\mathscr{P}$. This notion can be exactly mapped back to the approach of econometricians to identification – see again Matzkin (2007). The advantages of the functional approach are numerous: models without unobservables are always identified; any identified model can be uniquely mapped to a model without unobservables; identification can be properly considered for models where g is not bijective; identification is made a constructive program whose solution is directly linked to the inferential problem.

The functional approach to statistical models helps clarify many practices that emerge from considering problems with unobservable random variables. However, it leaves open the question of why statisticians consider such problems in the first place (especially if they have to map these problems back to models without unobservables for the sake of inference). The clear partition from the revisited foundations leads to a natural answer: the two classes of problems capture two fundamentally different approaches to statistical inference (and so in spite of the fact that the same data and the same inferential procedures are used). This fundamental difference motivates the following terminology: we call models without unobservables *phenomenological models* and models with unobservables *theoretical models*. In phenomenological models, the restriction to observables without reference to any underlying structure limits the inferential results to a description of the observable random phenomena through the recovery of their distribution. This echoes exactly the picture of statistical inference given in Robert (2007). Because only observables are considered in phenomenological models, the hypotheses collected in $\mathscr{P}$ can possibly be falsified by the data. In theoretical statistical models, the reach of inferential methods is extended from the observable phenomena to an underlying structure supporting them. The unobservable random variables form the necessary bridge that allows statisticians to rigorously feed data into a preliminary theoretical or structural model. The result of inference is not merely a description of the observables, but a characterization of the underlying structure by way of the data. As a consequence of the unobservables' presence, the hypotheses collected in $\mathscr{P}$ can no longer be entirely falsified by the data – only their empirical component under identification can possibly be. A direct implication is that the validity of the characterization of the theory from the inferential results necessarily depends on unverifiable hypotheses that stand beyond the data. This is the natural price to pay for the extension of the statistical scope.

The two classes of models are respectively illustrated in Section 3.2 and Section 4.2, where the same inferential problems are framed from both angles. The connection of theoretical statistical models to a preliminary theoretical edifice is then illustrated in Section 4.3 with two applications from microeconomics and causal inference.

Literature Overview. The textbook definition of a statistical model (or statistical experiment) has crystallized from the very beginnings of the field and, notably, from the foundational work of Wald, Blackwell, and Le Cam building on earlier ideas of Neyman and Pearson – see, in particular, Wald (1939), Blackwell (1951), Le Cam (1964). The notion has since been used as a natural starting point for most statistical work with few rigorous discussions of the notion, its nature, and its scope. The work of McCullagh (2002) stands out as an exception. Other contributions of a more informal nature include Lehmann (1990), Cox (1990), and Breiman (2001). Statistical models derived from a preliminary theoretical structure have been formalized mostly in economics through the notions of structural models and econometric models. The contributions of Matzkin collected in Matzkin (2007) can be traced back to the early developments of the field and, notably, the foundational work at the Cowles Commission – see, in particular, Koopmans (1949) and Hurwicz (1950). These constructions have rarely (if ever) been rigorously connected to the statistical literature. In the spirit of McCullagh (2002) but in a different direction, our paper revisits the notion of statistical models. We adopt a functional approach that is widespread in theoretical work but has never been used properly to define statistical models and unravel the construction for statistical problems framed with and without unobservables. Our partition of statistical models into the phenomenological and theoretical classes should also be of interest to philosophers of science as it provides a rigorous definition of the empirical component of a theory from within statistics. For the sake of exposition and concision, we do not discuss these issues beyond this point.

Remark 1.1. A consequence of our formalization is to rationalize Bayesian models as theoretical statistical models. Indeed, the Bayesian practice of taking the parameters as random variables falls directly within the class of theoretical statistical models as defined in this paper. We do not discuss this connection further and leave the conceptual fruits it may bear to future research.

The proofs of all propositions in the text are immediate from the definitions and are omitted.

Notation 1.1. Let $k \in \mathbb{N}$. We consider the Euclidean space $\mathbb{R}^k$ with its standard norm, metric, and topology, and endow it with its Borel σ-algebra. We denote by $\mathcal{P}(\mathbb{R}^k)$ the space of all Borel probability distributions on $\mathbb{R}^k$. For any $n \in \mathbb{N}$ and any measures $Q_1, \ldots, Q_n$ defined on measurable spaces $(E_1, \mathcal{E}_1), \ldots, (E_n, \mathcal{E}_n)$, we denote their product measure by $\bigotimes_{i=1}^n Q_i$. For any measure Q on a measurable space $(E, \mathcal{E})$, we denote by $L^2 = L^2(Q)$ the Lebesgue space (of equivalence classes) of measurable functions $f\colon E \to \mathbb{R}$ such that $\|f\|_2 < \infty$ where $\|f\|_2^2 = \int_E |f|^2 \, dQ$. The expectation operator is denoted by $\mathbb{E}[\cdot]$. The standard normal distribution is denoted by $N(0,1)$. Without loss of generality, we assume all random variables to be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ from which we abstract away completely. We generically denote the distribution of a random variable X by P_X and the joint distribution of two random variables X and Y by $P_{X,Y}$. In this case, it is implicit that the subscripts are not indices. We denote by N and M two arbitrary sets. Additional notations are introduced in Notation 2.1 at the end of Section 2.

2 Statistical models

Model-based inference in statistics broadly consists of gaining knowledge about an unknown probability distribution P^* based on a dataset $\mathrm{x} = (x_i : i \in N)$ that is construed as the realizations of some random variables $\mathrm{X} = (X_i : i \in N)$. The unknown distribution of interest P^* is either the distribution P_{X} of the random variables X that constitute the available data or the joint distribution $P_{\mathrm{X,U}}$ of the variables X and some other random variables $\mathrm{U} = (U_j : j \in M)$ for which no realizations are available in the data. The difference between X and U is captured by saying that X is observable and U is unobservable. More generally, we say that a random variable is observable if a realization of it is part of the available data and unobservable if not.

To properly frame the problem of gaining knowledge about P^*, the notion of a statistical model is useful. The following definition is not conventional, but it appears better suited for the task.

Definition 2.1. Let $(E, \mathcal{E})$ be a measurable space, Γ an arbitrary space, and P^* an unknown distribution on $(E, \mathcal{E})$. A statistical model for P^* is a pair $(\mathscr{P}, g)$ where $\mathscr{P}$ is a set of probability distributions on $(E, \mathcal{E})$ and $g\colon \mathscr{P} \to \Gamma$ is a surjective function.

The family $\mathscr{P}$ encapsulates all the restrictions hypothesized on the unknown distribution P^*. The function g delineates the inferential task in the statistical model by defining a statistical parameter $g(P) = \gamma \in \Gamma$ to be recovered from the data. The map g is surjective but not necessarily bijective. This allows the statistical parameter $\gamma = g(P)$ to ignore some of the features of the distribution $P \in \mathscr{P}$ and leads to a partition of the space $\mathscr{P}$ into equivalence classes of distributions defined by the same statistical parameter. If the unknown distribution satisfies $P^* \in \mathscr{P}$, then $g(P^*) := \gamma^* \in \Gamma$ delineates the equivalence class of distributions that P^* belongs to and that we aim to recover from the data. The condition $P^* \in \mathscr{P}$ is known as the correct specification of the statistical model and simply says that the assumptions collected in $\mathscr{P}$ are satisfied by P^*.

Definition 2.2. A statistical model $(\mathscr{P}, g)$ for an unknown probability distribution P^* is said to be correctly specified if $P^* \in \mathscr{P}$.

If g is bijective, then the statistical parameters partition $\mathscr{P}$ into singletons and g defines an index map $h\colon \Gamma \to \mathscr{P}$ by $h = g^{-1}$ so that the set $\mathscr{P}$ can be validly indexed by $\gamma \in \Gamma$ as $\{P_\gamma : \gamma \in \Gamma\}$. This recovers the standard definition of a statistical model as an indexed family of probability distributions. Remark 2.1 makes clear that the family $\mathscr{P}$ can be indexed independently of the choice of a statistical parameter. Remark 2.2 handles statistical models defined by indexation when the index map is surjective but not bijective.

Remark 2.1. Independently of the choice of any statistical parameter, a family of distributions $\mathscr{P}$ can always be indexed as $\{P_\theta : \theta \in \Theta\}$ for some well-chosen index set Θ such that there exists a bijection $h\colon \Theta \to \mathscr{P}$. In this case, it is possible to define a statistical model $(\mathscr{P}, g)$ with $g = h^{-1}$ so that $\gamma = g(P) = h^{-1}(P) = \theta$ and $\Gamma = \Theta$. This choice is common in simple problems where the assumptions on P^* and the choice of a statistical parameter can be easily aggregated. However, it is not always warranted and can lead to nonsensical constructions as seen by noting, for instance, that there is a bijection between $\mathbb{R}$ and the space $\mathcal{P}(\mathbb{R})$ of all Borel probability measures.

Remark 2.2. When defining statistical models by indexation, it is possible to do so based on an index set Θ and an index map $h\colon \Theta \to \mathscr{P}$ that is surjective but not bijective. In this case, the indexation is said to be repetitive: there exist two distinct parameters $\theta_1 \neq \theta_2$ such that $P_{\theta_1} = P_{\theta_2}$. This pathology of the indexation should be a minor problem in statistics since the defect can be simply corrected by deleting the repetitive indices in a bijective indexation of $\mathscr{P}$. This point is made transparent by the functional approach of Definition 2.1. If h is not bijective, the inverse h^{-1} of h is not a function from $\mathscr{P}$ to Θ but a set-valued mapping from $\mathscr{P}$ to a subset of 2^Θ. Choosing h^{-1} still leads to a valid statistical model $(\mathscr{P}, h^{-1})$ according to Definition 2.1. The question is then for the statistician to answer why $h^{-1}\colon \mathscr{P} \to \Gamma \subseteq 2^\Theta$ is of interest as a statistical parameter. The fact that repetitive indexations have any importance in statistics stems from their use to define a notion of significant importance for statistical problems with unobservable random variables called identification. In Section 4.1, we show that this notion can be defined without reference to any indexation and be uniquely linked to problems with unobservables.

The functional definition of a statistical model prevents many unfortunate issues caused by the definition by indexation – see again Remark 2.1 and Remark 2.2. However, as for the definition by indexation, nothing in Definition 2.1 immediately prevents the choice of nonsensical statistical parameters: the existence of a bijection $g\colon \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ again illustrates this point. While both definitions fall into the same adversarial trap, the functional approach offers direct and natural guidelines for determining what a *good* statistical parameter should be. By making explicit that a statistical parameter is a function of the distribution, Definition 2.1 leads to a natural desideratum: the map g should preserve some well-chosen structures on $\mathscr{P}$ beyond the set-theoretic ones – the structures can be topological, metric, differentiable, geometric, or algebraic. If g is structure-preserving, then the statistical parameter $\gamma = g(P)$ naturally encodes some (but possibly not all) structure-dependent features shared by all distributions in the equivalence class that P belongs to. Conversely, the choice of appropriate structures on $\mathscr{P}$ is sufficient to obtain encompassing results that link the difficulty of the inferential task to the properties of g. Remark 2.3 provides two fundamental applications that leverage the local variations of g with respect to some well-chosen statistical distances on $\mathscr{P}$. Unfortunately, there does not seem to exist a single structure on probability distributions that is general enough to cover all statistical applications – see Remark 2.4. What is clear, however, is that the mere set-theoretic structure of statistical models is insufficient to obtain any useful classification – see Remark 2.5. This echoes the remarks of Lucien Le Cam and Grace Lo Yang in Section 7.4 of Le Cam and Yang (2000).

Remark 2.3. (Modulus of Continuity) When Γ is a normed vector space with norm $\|\cdot\|_\Gamma$, the local variations of g with respect to some statistical distances provide a useful guide to characterize the difficulty of the inferential task. Given any statistical distance d on $\mathscr{P}$, the modulus of continuity, defined by

$$\omega(\varepsilon, P; g, d) := \sup\left\{\|g(P) - g(Q)\|_\Gamma : Q \in \mathscr{P}, d(P, Q) \leq \varepsilon\right\},$$

for any $\varepsilon > 0$ and any $P \in \mathscr{P}$, appears useful. For instance, the fact that $\omega(\varepsilon, P; g, d_{\mathrm{TV}}) = \infty$ for all $P \in \mathscr{P}$ and all $\varepsilon > 0$, where d_{TV} is the total variation distance, directly leads to a

number of famed impossibility results such as those in Bahadur and Savage (1956). Another fundamental result is provided by Liu and Brown (1993) where singularity, defined by the existence of $P \in \mathscr{P}$ such that $\lim_{\varepsilon \to 0} \omega(\varepsilon, P; g, d_H)/\varepsilon = \infty$ with d_H the Hellinger distance, is linked to an unavoidable asymptotic bias-variance trade-off that prevents rules from achieving $\sqrt{n}$-convergence rates. Similar ideas can be found in Donoho and Liu (1987, 1991a,b). These regularity conditions, which require additional structures on $\mathscr{P}$ but uniquely map to the inferential difficulty of the problem, can then be used to obtain meaningful classifications of statistical models.

Remark 2.4. In the spirit of McCullagh (2002), it is tempting to leverage the surjectivity of g and any topological structure on $\mathscr{P}$ to define a useful classification based on the properties of Γ as a quotient of $\mathscr{P}$. Unfortunately, the resulting structure on Γ is generally insufficient to be usefully linked to the inferential difficulty of the task. As shown in Remark 2.3, specific (pre)metric structures, which do not naturally emerge from quotienting, are better suited for the task.

Remark 2.5. Statistical models defined by bijective index maps $h\colon \Theta \to \mathscr{P}$, where $\gamma = g(P) = h^{-1}(P) = \theta$ and $\Gamma = \Theta$ as in Remark 2.1, are conventionally classified as parametric if $\Theta \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}$ and nonparametric if not. Nonparametric problems are then sometimes subdivided into semiparametric models when $\Theta \subseteq \mathbb{R}^k \times \Psi$ where $k \in \mathbb{N}$ and Ψ is not a subset of any Euclidean space. Unfortunately, this set-theoretic classification is vacuous and nonsensical from a statistical viewpoint. This can be illustrated again by our running example. The statistical model $(\mathcal{P}(\mathbb{R}), \mathrm{id}_{\mathcal{P}(\mathbb{R})})$, where $\mathrm{id}_{\mathcal{P}(\mathbb{R})}$ is the identity function on $\mathcal{P}(\mathbb{R})$, would be reasonably classified as nonparametric by all statisticians. However, there exists a bijection $h\colon \mathbb{R} \to \mathcal{P}(\mathbb{R})$ so that $(\mathcal{P}(\mathbb{R}), \mathrm{id}_{\mathcal{P}(\mathbb{R})})$ is equivalent to the indexed family $\{P_\theta : \theta \in \mathbb{R}\}$ where $\theta = h^{-1}(P)$ for any $P \in \mathcal{P}(\mathbb{R})$. According to the conventional classification, $(\mathcal{P}(\mathbb{R}), \mathrm{id}_{\mathcal{P}(\mathbb{R})})$ is in fact a parametric model: an unfortunate set-theoretic incongruity. The same issue plagues all models defined on standard spaces where an appropriate reparametrization leads them all to be parametric. The explanation is that the set-theoretic structure is too weak to be useful for classification. Additional structures need to be preserved by the inverse $h^{-1} = g$ of the index map for the indexation to be any guide to inferential difficulty. An example of such a structure-preserving property is provided in Remark 2.3.

While the set-theoretic structure of statistical models is too weak to obtain useful classifications of all statistical models in terms of inferential difficulty, it is still strong enough to build relative classifications. For instance, the following notion of a submodel appears useful.

Definition 2.3. A statistical model $(\mathscr{P}, g)$ is said to be a submodel of another statistical model $(\mathscr{P}', g')$ if $\mathscr{P} \subseteq \mathscr{P}'$ and g is the restriction of g' to $\mathscr{P}$.

The previous considerations apply to all statistical models independently of whether P^* is the distribution of some observable random variables or the distribution of some observable and some unobservable random variables. The next two sections make clear that this difference is not vacuous but bears important practical and conceptual consequences. Based on this dichotomy, we can partition the set of statistical models into two classes. These two classes are the object of Section 3 and Section 4, respectively. The following notations are useful for the task.

Notation 2.1. Let $(\mathscr{P}, g)$ be a statistical model for an unknown distribution P^*. We generically denote by $\mathrm{X} = (X_i : i \in N)$ and $\mathrm{U} = (U_j : j \in M)$ all the observable random variables and all the unobservable random variables in the model. If $M = \emptyset$, there are no unobservables and P^* is simply the distribution of the observable X, which we denote by P_{X}. If $M \neq \emptyset$, then P^* is the joint distribution of the observable X and the unobservable U, which we denote by $P_{\mathrm{X,U}}$. We denote by $((\mathcal{X}_i, \mathcal{B}_i) : i \in N)$ and $((\mathcal{U}_j, \mathcal{C}_j) : j \in M)$ the measurable spaces on which the random variables take values. We call $((\mathcal{X}_i, \mathcal{B}_i) : i \in N)$ the observable spaces and $((\mathcal{U}_j, \mathcal{C}_j) : j \in M)$ the unobservable spaces. The space $(E, \mathcal{E})$ on which P^* is defined is either the product space $E = \prod_{i \in N} \mathcal{X}_i$ and $\mathcal{E} = \bigotimes_{i \in N} \mathcal{B}_i$ if $M = \emptyset$ or the product space $E = (\prod_{i \in N} \mathcal{X}_i) \times (\prod_{j \in M} \mathcal{U}_j)$ and $\mathcal{E} = (\bigotimes_{i \in N} \mathcal{B}_i) \otimes (\bigotimes_{j \in M} \mathcal{C}_j)$ if $M \neq \emptyset$. From Definition 2.1, we assume that all distributions in $\mathscr{P}$ are defined on the same space as P^*. We denote by $\pi(\cdot)$ the projection operator that maps $P \in \mathscr{P}$ to its marginal with respect to all the observable spaces $((\mathcal{X}_i, \mathcal{B}_i) : i \in N)$ and call it the observable projection. In particular, if $M = \emptyset$, then $\pi(P) = P$ for all $P \in \mathscr{P}$. Symmetrically, if $M \neq \emptyset$, we denote by $\bar{\pi}(\cdot)$ the projection operator that maps $P \in \mathscr{P}$ to its marginal with respect to all the unobservable spaces $((\mathcal{U}_j, \mathcal{C}_j) : j \in M)$ and call it the unobservable projection. In particular, we have that $\pi(P^*) = P_{\mathrm{X}}$ and $\bar{\pi}(P^*) = P_{\mathrm{U}}$ if $M \neq \emptyset$ where P_{U} denotes the distribution of U. If $N = \{1, 2, \ldots, n\}$, then we denote by $\pi_1(\cdot)$ the projection that maps $P \in \mathscr{P}$ to its marginal with respect to the first observable space $(\mathcal{X}_1, \mathcal{B}_1)$. We extend this notation to any marginal on the observable spaces and any marginal on the unobservable spaces for the respective operator $\bar{\pi}(\cdot)$.

3 Phenomenological statistical models

3.1 Definition and properties

Definition 3.1. A statistical model $(\mathscr{P}, g)$ for an unknown probability distribution P^* is said to be phenomenological if P^* is the distribution of random variables that are all observable.

Remark 3.1. Following Notation 2.1, we denote all the observable random variables for the model by $\mathrm{X} = (X_i : i \in N)$. This allows us to rewrite the unknown distribution P^* as the distribution of X, that is, $P^* = P_{\mathrm{X}}$. Using Notation 2.1 again and denoting by π the projection operator on the observable spaces, we have the direct equivalence: a statistical model $(\mathscr{P}, g)$ for an unknown probability distribution P^* is phenomenological if and only if $\pi(P^*) = P^*$.

The purpose of phenomenological statistical models is implicit from the restriction to observable random variables. An oracle with unlimited sampling abilities would solve the statistical problem by recovering $P^* = P_{\mathrm{X}}$. The distribution would then provide a complete description of the observable random phenomenon X. A statistician with only limited sampling abilities leverages the model $(\mathscr{P}, g)$ to gain knowledge about $P^* = P_{\mathrm{X}}$. The choice of $\mathscr{P}$ amounts to restricting the set of possible probabilistic descriptions for the observable random phenomenon X. The choice of g amounts to collapsing $\mathscr{P}$ into equivalence classes of probabilistic descriptions. The descriptive nature of the inferential task is inherent to the restriction to observables.

Remark 3.2. In a phenomenological statistical model, an oracle with unlimited sampling abilities would always be able to compute $\gamma^* = g(P_X)$ provided $P_X \in \mathscr{P}$. The inferential task is well-posed in this sense. However, since Definition 3.1 does not impose any structure on $\mathscr{P}$, nothing prevents g from being a nonsensical choice, leading to equivalence classes that cannot be feasibly recovered from the data $\mathrm{x} = (x_i : i \in N)$. It is only when $\mathscr{P}$ is endowed with some structure that the statistician can properly arbitrate the choice of $\mathscr{P}$ and g and get guarantees that the equivalence classes of probabilistic descriptions obtained from $(\mathscr{P}, g)$ can be feasibly recovered from the data. As for general statistical models, there is no single structure on $\mathscr{P}$ that can guide the choice of g in all statistical problems. A few conventional choices are reviewed in the examples of Section 3.2.

Remark 3.3. Independently of the difficulty of the inferential task, the validity of the probabilistic description derived from a phenomenological statistical model $(\mathscr{P}, g)$ depends exclusively on the correct specification of the statistical model, that is, $P^* = P_X \in \mathscr{P}$. A distinctive feature of phenomenological statistical models is that the specification assumption $P^* \in \mathscr{P}$ can be subject to empirical verification within model-based inference by a statistician having access to another draw from the same distribution P^*. This is a direct consequence of the fact that P^* is the distribution of observables only. By embedding the first model $(\mathscr{P}, g)$ in another model $(\mathscr{P}', g')$ where $\mathscr{P} \subset \mathscr{P}'$ and $g' = \mathrm{id}_{\mathscr{P}'}$ is the identity map on $\mathscr{P}'$, the condition $g'(P^*) = P^* \in \mathscr{P}$ falls within the inferential scope. However, the fact that falsification is possible does not prevent the inferential task from being intractable given the data, nor does the task escape the minimal assumptions for the statistical problem to be framed (including the observability and sampling hypotheses).

The nature and purpose of phenomenological statistical models are illustrated in the next section, where we frame five standard statistical problems as phenomenological statistical models. For the sake of exposition, the application of each model to particular problems is not considered and the inferential difficulty of the task in each problem is only briefly discussed.

3.2 Examples

Example 3.1 (Mean Estimation in a Gaussian Family with an i.i.d. Sample)**.** Mean estimation in a Gaussian family with known variance based on an i.i.d. sample is one of the most elementary problems in model-based inference. The problem can be naturally linked to a phenomenological statistical model. The data $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ is construed as the realizations of an i.i.d. sample $\mathrm{X} = (X_1, X_2, \ldots, X_n)$ with identical marginal distribution $P_{X_1} = N(\mu^*, 1)$ where $\mu^* \in \mathbb{R}$. The statistical model $(\mathscr{P}, g)$ for the problem is then obtained by taking

$$\mathscr{P} = \left\{ \bigotimes_{i=1}^{n} N(\mu, 1) : \mu \in \mathbb{R} \right\},$$

and

$$g \colon \mathscr{P} \to \mathbb{R},\ P \mapsto \int_{\mathbb{R}} x \, d\pi_1(P)(x).$$

In this case, g is bijective and the statistical parameter corresponds to the indexation of the family $\mathscr{P}$, that is, $g(\bigotimes_{i=1}^{n} N(\mu, 1)) = \mu$ for all $\mu \in \mathbb{R}$. By virtue of the constraints on $\mathscr{P}$ and the choice of g, the inferential problem of recovering $\mu = g(\bigotimes_{i=1}^{n} N(\mu, 1))$ from $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ comes with exact finite-sample guarantees that make it one of the simplest problems in statistics.

Example 3.2 (Mean Estimation in a Nonparametric Family with an i.i.d. Sample)**.** The Gaussian assumption of Example 3.1 can be relaxed by considering a much larger family for $\mathscr{P}$. The observability and sampling assumptions are identical, but $\mathscr{P}$ is now taken to be

$$\mathscr{P} = \left\{ \bigotimes_{i=1}^{n} Q : Q \in \mathcal{P}(\mathbb{R}), \int_{\mathbb{R}} |x| \, dQ(x) < \infty \right\},$$

and g is still the expectation operator but defined on the first marginal for the new family $\mathscr{P}$. The family $\mathscr{P}$ cannot be meaningfully indexed and the map g is no longer injective. The inferential task amounts to recovering the equivalence class of distributions in $\mathscr{P}$ defined by the same mean $g(P_{\mathrm{X}}) = \int_{\mathbb{R}} x \, d\pi_1(P_{\mathrm{X}})(x)$ as P_{X}. The difficulty of the task changes drastically compared to the task in Example 3.1: the modulus of continuity of g with respect to the total variation distance is no longer finite – this leads to a number of impossibility results for the problem as shown in Bahadur and Savage (1956).

Example 3.3 (Nonparametric Regression with an i.i.d. Sample)**.** Nonparametric regression based on an i.i.d. sample is another fundamental problem of model-based inference. The problem can be naturally linked to a phenomenological statistical model. The data $\mathrm{x} = ((z_1, y_1), (z_2, y_2), \ldots, (z_n, y_n))$ is construed as the realizations of an i.i.d. n-sample $\mathrm{X} = ((Z_1, Y_1), (Z_2, Y_2), \ldots, (Z_n, Y_n))$ with identical marginal distribution P_{Z_1, Y_1} such that Y_1 has finite second moment. The statistical model $(\mathscr{P}, g)$ for the problem is obtained by taking

$$\mathscr{P} = \left\{ \bigotimes_{i=1}^{n} Q : Q \in \mathcal{P}(\mathbb{R}^2), \int_{\mathbb{R}^2} y^2 \, dQ(z, y) < \infty \right\},$$

and

$$g \colon \mathscr{P} \to \bigcup_{P \in \mathscr{P}} L^2(\pi_1(P)(\cdot \times \mathbb{R})), \; P \mapsto \operatorname*{arg\,min}_{m \in L^2(\pi_1(P)(\cdot \times \mathbb{R}))} \int_{\mathbb{R}^2} \big(y - m(z)\big)^2 \, d\pi_1(P)(z, y).$$

By the projection theorem, the function g is well-defined and corresponds to the conditional expectation operator. The family $\mathscr{P}$ cannot be meaningfully indexed and g is not injective. The inferential task amounts to recovering the equivalence class of distributions in $\mathscr{P}$ defined by the same conditional expectation $\mathbb{E}[Y_1 | Z_1 = \cdot]$ as P_{X}. In spite of the information collapsed by g from P_{X} to the conditional expectation, the task remains notoriously difficult. From the modulus of continuity characterization of Remark 2.3, the problem is singular, which leads to slower-than-$\sqrt{n}$-convergence rates – see, for instance, Györfi, Kohler, Krzyzak, and Walk (2002).

Example 3.4 (Linear Regression with an i.i.d. Sample)**.** Linear regression is a subproblem

of nonparametric regression considered in Example 3.3. The problem is of utmost importance in applications and can be naturally linked to a phenomenological statistical model. The observability and sampling assumptions are unchanged, leading to the same set $\mathscr{P}$ as in Example 3.3, but the function g is now defined as

$$g\colon \mathscr{P} \to \{\{x\} : x \in \mathbb{R}\} \cup \{\mathbb{R}\},\ P \mapsto \operatorname*{arg\,min}_{b\in\mathbb{R}} \int_{\mathbb{R}^2} \big(y - bz\big)^2\, d\pi_1(P)(z, y).$$

The map g is well-defined but not necessarily singleton-valued. If we restrict $\mathscr{P}$ to $\mathscr{P}'$ by taking $\mathscr{P}' = \{\bigotimes_{i=1}^n Q \in \mathscr{P} : Q(\{0\}\times\mathbb{R}) < 1\}$, then $g(\mathscr{P}') = \{\{x\} : x \in \mathbb{R}\}$ and so g is singleton-valued. This restriction is conventionally enforced in applications. The family $\mathscr{P}'$ cannot be meaningfully indexed and g is not injective. The inferential task amounts to recovering the equivalence class of distributions in $\mathscr{P}'$ defined by the same linear approximation of the conditional expectation $g(P_{\mathrm{X}}) = \{\mathbb{E}\,[Y_1 Z_1]/\mathbb{E}\,[Z_1^2]\}$ as P_{X}. The choice of g compared to Example 3.3 makes the inferential problem more tractable. From the modulus of continuity characterization, the problem is no longer singular and conventionally admits rules achieving $\sqrt{n}$-convergence rates.

Example 3.5 (Autocorrelation Estimation with a Stationary Sample)**.** The previous examples all relied on an i.i.d. sample. To move beyond this assumption, we consider the standard problem of autocorrelation estimation with a mean-zero stationary sample and link it to a phenomenological statistical model. The data $\mathrm{x} = (x_1, x_2, \ldots, x_T)$ is construed as the realizations of some random variables $\mathrm{X} = (X_1, X_2, \ldots, X_T)$. The statistical model $(\mathscr{P}, g)$ for the problem is obtained by taking

$$\mathscr{P} = \left\{P \in \mathcal{P}(\mathbb{R}^T) : \int_{\mathbb{R}^T} x\, dP(x) = 0,\ \int_{\mathbb{R}^T} xx^\top\, dP(x) \in V^T\right\}$$

and

$$g\colon \mathscr{P} \to (-1, 1),\ P \mapsto \left(\int_{\mathbb{R}} x^2\, d\pi_1(P)(x)\right)^{-1} \int_{\mathbb{R}^2} xy\, d\pi_{1,2}(P)(x, y)$$

where V^T is the subset of symmetric positive definite $T\times T$ matrices S_{++}^T defined as $V^T = \{\Sigma \in S_{++}^T : \exists\, \sigma^2 > 0, \kappa \in \mathbb{R} \text{ s.t. } \Sigma_{t,t} = \sigma^2 \text{ for all } t \in \{1, \ldots, T\} \text{ and } \Sigma_{t,t+1} = \kappa \text{ for all } t \in \{1, \ldots, T-1\}\}$. The function g is well-defined and corresponds to the first-order autocorrelation functional. The family $\mathscr{P}$ cannot be meaningfully indexed and the function g is not injective. The inferential task amounts to recovering the equivalence class of distributions in $\mathscr{P}$ defined by the same first-order autocorrelation $g(P_{\mathrm{X}}) = \mathbb{E}\,[X_1 X_2]/\mathbb{E}\,[X_1^2]$ as P_{X}. Based on the modulus of continuity characterization with respect to the total variation distance, the inferential problem is seen to exhibit similar pathologies as the problem of nonparametric mean estimation from Example 3.2.

4 Theoretical statistical models

4.1 Definition and properties

Definition 4.1. A statistical model $(\mathscr{P}, g)$ for an unknown probability distribution P^* is said to be theoretical if P^* is the distribution of random variables that are not all observable.

Remark 4.1. Following Notation 2.1, we denote all the observable random variables by $\mathrm{X} = (X_i : i \in N)$ and all the unobservable random variables by $\mathrm{U} = (U_j : j \in M)$. This allows us to rewrite the unknown distribution P^* as the joint distribution $P_{\mathrm{X,U}}$, that is, $P^* = P_{\mathrm{X,U}}$. Using Notation 2.1 again and denoting by π the projection operator on the observable spaces, we have the direct equivalence: a statistical model $(\mathscr{P}, g)$ for P^* is theoretical if and only if $\pi(P^*) \neq P^*$.

The presence of unobservable random variables in statistical problems is surprising at first as it comes with a significant cost: the unknown distribution $P^* = P_{\mathrm{X,U}}$ is never accessible to an oracle with unlimited sampling abilities with respect to the observables only. It follows that the inferential problem delineated by $(\mathscr{P}, g)$ may not even be well-posed in the sense that the same oracle may not be able to compute $\gamma^* = g(P_{\mathrm{X,U}})$ when $P_{\mathrm{X,U}} \in \mathscr{P}$. This impossibility unavoidably applies to a statistician with limited sampling abilities with respect to the observables.

The reason statisticians consider such problems is that the introduction of unobservables allows them to significantly extend the scope of statistical methods from observable random phenomena to a hypothesized structure supporting and explaining the observations. In practice, the unobservables and the assumptions linking them to the observables originate from a field-specific theory that precedes the statistical model and gives a proper meaning to the joint distribution $P_{\mathrm{X,U}}$ and the statistical parameter $\gamma^* = g(P_{\mathrm{X,U}})$. The choice of $\mathscr{P}$ no longer amounts to restricting the set of probabilistic descriptions for X but to restricting the set of possible explanations for them in terms of an underlying structure captured by U. The choice of g does not amount simply to collapsing $\mathscr{P}$ into equivalence classes but to recovering a theoretical object whose meaning depends on the nature of the unobservables. This constitutes the best (and only) way for the statistician to leverage the tools of model-based inference to rigorously feed data into an existing theoretical or structural model. The result is not merely a description of some observable random phenomenon, but an attempt at an explanation based on some underlying structure.

The main issue is that the inferential problem in a theoretical statistical model may not be well-posed in the sense defined above that an oracle with unlimited sampling abilities with respect to the observables may not be able to compute the statistical parameter $\gamma^* = g(P_{\mathrm{X,U}})$. This fact leads to a property for statistical models known as identification. This property partitions the set of models into two classes: identified models, where the inferential problem only depends on the observables, and non-identified models, where the inferential problem does not.

Definition 4.2. A statistical model $(\mathscr{P}, g)$ is said to be identified if there exists a function $G\colon \pi(\mathscr{P}) \to \Gamma$ such that $g(P) = G(\pi(P))$ for all $P \in \mathscr{P}$.

The property can be equivalently defined by introducing an intuitive equivalence relation on the distributions in $\mathscr{P}$. Given a statistical model $(\mathscr{P}, g)$, we say that any two distributions P and Q in $\mathscr{P}$ are observationally equivalent if $\pi(P) = \pi(Q)$. This directly defines an equivalence relation on $\mathscr{P}$ induced by the surjective map $\pi(\cdot)$.

Proposition 4.1. *A statistical model $(\mathscr{P}, g)$ is identified if and only if g is constant on each equivalence class of $\mathscr{P}$ induced by $\pi(\cdot)$.*

The problem of identification in statistical models emerges from the presence of unobservable random variables and is thus the distinctive feature of theoretical statistical models. An immediate consequence of the definition is that phenomenological statistical models are always identified. Another useful characterization of identified theoretical statistical models is that they can be associated uniquely with phenomenological statistical models. This congruence can then be taken as a formal definition of the empirical content of theoretical models.

Proposition 4.2. *A phenomenological statistical model is identified.*

Proposition 4.3. *For any identified theoretical statistical model $(\mathscr{P}, g)$, there exists a unique phenomenological statistical model $(\pi(\mathscr{P}), G)$ such that $G(\pi(P)) = g(P)$ for all $P \in \mathscr{P}$. In this case, $(\mathscr{P}, g)$ and $(\pi(\mathscr{P}), G)$ are said to be congruent.*

If a theoretical statistical model $(\mathscr{P}, g)$ is not identified, then it is always possible to modify the model by either restricting $\mathscr{P}$ or modifying g so as to obtain an identified statistical model.

Remark 4.2. If the modifications consist of restricting $\mathscr{P}$ to a new set $\mathscr{P}' \subseteq \mathscr{P}$, then the additional assumptions in $\mathscr{P}'$ are commonly referred to as identification assumptions for $(\mathscr{P}, g)$. We can more simply say that the new model $(\mathscr{P}', g)$ is identified. Similarly, if changing g to g' for a given set $\mathscr{P}$ leads to identification, then we simply say that the new model $(\mathscr{P}, g')$ is identified. This avoids introducing specific terminology such as set identification.

Remark 4.3. The notion of identification in Definition 4.2 can be mapped back to the notion of identification commonly encountered in the econometric literature (when it is well-defined[2]). The econometric approach to identification is to index the family $\pi(\mathscr{P})$ for the distribution of the observables as $\{P_\theta : \theta \in \Theta\}$ based on the possibly non-injective index map $h\colon \theta = g(P) \mapsto \pi(P)$. Identification is then defined as the fact that the indexation is not repetitive, that is, that h is injective. When g is bijective so that h is surjective, this is equivalent to Definition 4.2. Indeed, if h is injective, then it is invertible and we can take its inverse h^{-1} for G in Definition 4.2: we directly have $h^{-1}(\pi(P)) = g(P)$ for all $P \in \mathscr{P}$. Beyond the problem of invalid implicit parametrizations, this approach suffers from two other drawbacks: it does not uniquely link the problem of identification to unobservables since it is always possible to index a family in a non-injective way even in the absence of any unobservables – see Remark 2.2; it is not operational as it leaves implicit the practical problem of finding G from which inference is made feasible.

Remark 4.4. In a theoretical statistical model $(\mathscr{P}, g)$, the family $\mathscr{P}$ of possible probability distributions for $P^* = P_{\mathrm{X,U}}$ is often constrained by a set of stochastic models between random variables taking values in the same spaces $((\mathcal{X}_i, \mathcal{B}_i) : i \in N)$ and $((\mathcal{U}_j, \mathcal{C}_j) : j \in M)$ as X and U. These models are the stochastic translation of the field-specific structural models supporting the theory. We follow standard practices and express these stochastic models directly in terms of X and U. This is an abuse of notation: writing the stochastic models in this fashion only constrains $\mathscr{P}$ and

[2]We assume in Remark 4.3 that g is bijective so that there is no risk of invalid parametrization through $g(P) \mapsto P$. As for general statistical models defined by implicit parametrization from indexation, the problem of choosing $\theta = g(P)$ in a non-surjective way is generally ignored since the unindexed sets $\mathscr{P}$ and $\pi(\mathscr{P})$ are rarely characterized in practice.

does not imply that the stochastic model holds true for X and U. In particular, it is still possible that the statistical model is incorrectly specified in the sense that $P_{X,U} \notin \mathscr{P}$.

Remark 4.5. In theoretical statistical models, the map g still delineates equivalence classes of distributions in $\mathscr{P}$. Because the distributions now include an unobservable component, the equivalence classes induced by g get an extra meaning from the field-specific theory supporting the model. This difference is often highlighted by calling the statistical parameters $\gamma = g(P)$ structural parameters. While the preliminary structure naturally guides the selection of g, it does not guarantee that g is a sensible statistical choice that would lead under identification to tractable inferential problems. Indeed, Definition 4.1 does not impose any structure on $\mathscr{P}$ and Definition 4.2 does not impose any regularity condition on G. As for general statistical models, it is only by adding structures on $\mathscr{P}$ that the statistician can better arbitrate the choice of $\mathscr{P}$ and g (with the implied function G under identification). Compared to phenomenological statistical models, these choices are further constrained by the preliminary theoretical edifice and lead in practice to a delicate arbitrage between theoretical validity and inferential feasibility.

Remark 4.6. Independently of the inferential problem, the validity of the theoretical model is entirely captured by the specification assumption $P^* = P_{X,U} \in \mathscr{P}$. Compared to phenomenological models, this assumption cannot be subject to empirical verification within model-based inference. This is a direct translation of the necessary *leap of faith* in any theoretical model that goes from the observables to an underlying structure removed both logically and empirically from them. Breaking away from this hard constraint would require another draw of the data but under different observability rules – the unobservables would have to be observable. While the specification assumption $P^* = P_{X,U} \in \mathscr{P}$ cannot be entirely subject to falsification, the problem of identification exactly makes clear what part of the model is subject to empirical falsification: it is the phenomenological model that is congruent to the theoretical one under identification. This phenomenological model, which can be viewed as the empirical component of the theoretical one, is subject to empirical falsification within model-based inference as made clear in Remark 3.3.

We illustrate the structure of theoretical statistical models in Section 4.2 by unearthing standard stochastic models that lead to theoretical statistical models congruent under identification to the phenomenological statistical models of Section 3.2. For these models, the connection to a preliminary theoretical structure is left implicit. To illustrate the origin of theoretical statistical models in a preliminary theoretical edifice, we introduce in Section 4.3 two well-known applications and their respective origins in microeconomics and causal inference.

4.2 Examples

Example 4.1 (Gaussian White Noise Model)**.** The phenomenological model of Example 3.1 can be linked to a theoretical statistical model obtained from a well-known stochastic model – the Gaussian white noise model. The data $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ is still construed as the realizations of an i.i.d. sample $\mathrm{X} = (X_1, X_2, \ldots, X_n)$, but the statistical model $(\mathscr{P}, g)$ now targets the joint distribution $P_{X,U}$ of X and some unobservable random variables $\mathrm{U} = (U_0, U_1, U_2, \ldots, U_n)$. The

set $\mathscr{P}$ of hypothesized distributions for $P_{\mathrm{X,U}}$ is constrained by a set of stochastic models obtained by assuming that $U_0 = \mu \in \mathbb{R}$ is constant, $(U_1, U_2, \ldots, U_n)$ are i.i.d. with $U_1 \sim N(0,1)$, and

$$X_i = \mu + U_i$$

for all $i \in \{1, 2, \ldots, n\}$. The theoretical statistical model is completed by taking

$$g\colon \mathscr{P} \to \mathbb{R},\ P \mapsto \mu.$$

In this case, the family induced by $\mathscr{P}$ for the observables' distribution is

$$\pi(\mathscr{P}) = \left\{ \bigotimes_{i=1}^{n} N(\mu, 1) : \mu \in \mathbb{R} \right\}.$$

The statistical model $(\mathscr{P}, g)$ is thus identified by taking $G\colon \pi(\mathscr{P}) \to \mathbb{R}$ as the expectation operator with respect to the first marginal $\pi_1(P)$, and the identified model $(\mathscr{P}, g)$ is then congruent to the phenomenological statistical model of Example 3.1.

Example 4.2 (White Noise Model)**.** The Gaussian white noise model can be relaxed to obtain a theoretical statistical model that can be linked to the phenomenological statistical model of Example 3.2. The observability assumptions, sampling assumptions, and the stochastic model are the same as in the Gaussian white noise model except for the assumption on the i.i.d. variables $(U_1, U_2, \ldots, U_n)$, which is relaxed from normality to $\mathbb{E}[U_1] = 0$. This defines a new theoretical statistical model $(\mathscr{P}, g)$ with $g\colon P \mapsto \mu$ now defined on the expanded set $\mathscr{P}$. The family induced by $\mathscr{P}$ for the observables' distribution is now

$$\pi(\mathscr{P}) = \left\{ \bigotimes_{i=1}^{n} Q : Q \in \mathcal{P}(\mathbb{R}), \int_{\mathbb{R}} |x|\, dQ(x) < \infty \right\}.$$

The statistical model is thus identified by taking $G\colon \pi(\mathscr{P}) \to \mathbb{R}$ as the expectation operator with respect to the first marginal $\pi_1(P)$. The identified model $(\mathscr{P}, g)$ is then congruent to the phenomenological statistical model of Example 3.2.

Example 4.3 (Nonparametric Regression with Additive Errors)**.** The problem of nonparametric regression considered in Example 3.3 from the point of view of a phenomenological statistical model can be equivalently linked to a theoretical statistical model. The data $\mathrm{x} = ((z_1, y_1), (z_2, y_2), \ldots, (z_n, y_n))$ is still construed as the realizations of an i.i.d. n-sample $\mathrm{X} = ((Z_1, Y_1), (Z_2, Y_2), \ldots, (Z_n, Y_n))$, but the statistical model $(\mathscr{P}, g)$ now targets the joint distribution of X and some unobservable random variables $\mathrm{U} = (U_0, U_1, \ldots, U_n)$. The set $\mathscr{P}$ of hypothesized distributions for $P_{\mathrm{X,U}}$ is constrained by a set of stochastic models obtained by assuming that $U_0 = m(\cdot) \in L^2(P_{Z_1})$ is constant, $(U_1, U_2, \ldots, U_n)$ are i.i.d. with $\mathbb{E}[U_1^2] < \infty$, and

$$Y_i = m(Z_i) + U_i$$

for all $i \in \{1, 2, \ldots, n\}$. The theoretical statistical model is completed by taking

$$g \colon \mathscr{P} \to \bigcup_{P \in \mathscr{P}} L^2(\pi_1(P)(\cdot \times \mathbb{R})),\ P \mapsto m(\cdot).$$

Without further assumptions, the theoretical statistical model $(\mathscr{P}, g)$ is not identified. For instance, taking $m_1(z) = z$ and $U_1 \sim N(0,1)$ independent of Z_1 and $m_2(z) = z - c$ and $U_1 \sim N(c,1)$ independent of Z_1 with $c > 0$ yields the same observables' marginals but different images for g, hence making it impossible to unearth a function G. A standard set of identification assumptions consists of restricting $\mathscr{P}$ to $\mathscr{P}'$ by further assuming that $\mathbb{E}[U_1|Z_1] = 0$ P_{Z_1}-a.s. in the definition of the stochastic models. The family induced by $\mathscr{P}'$ for the observables' distribution is then

$$\pi(\mathscr{P}') = \left\{ \bigotimes_{i=1}^{n} Q : Q \in \mathcal{P}(\mathbb{R}^2), \int_{\mathbb{R}^2} y^2 \, dQ(z,y) < \infty \right\}$$

and so the statistical model $(\mathscr{P}', g)$ is identified by taking

$$G(\pi(P)) = \underset{m \in L^2(\pi_1(P)(\cdot \times \mathbb{R}))}{\arg\min} \int_{\mathbb{R}^2} \big(y - m(z)\big)^2 \, d\pi_1(P)(z,y)$$

for any $P \in \mathscr{P}'$. The identified model $(\mathscr{P}', g)$ is then congruent to the phenomenological model of Example 3.3.

Example 4.4 (Linear Regression with Additive Errors)**.** The stochastic model considered in Example 4.3 can be strengthened to obtain a theoretical statistical model equivalent to the phenomenological model of Example 3.4 (up to a homeomorphism). The stochastic model is strengthened by assuming that $U_0 = b \in \mathbb{R}$ and

$$Y_i = bZ_i + U_i$$

for all $i \in \{1, \ldots, n\}$. This restriction is not sufficient for the model $(\mathscr{P}, g)$ to be identified. Restricting it further by adding the assumption that $\mathbb{E}[U_1 Z_1] = 0$ and $0 < \mathbb{E}[Z_1^2] < \infty$ leads to an identified model $(\mathscr{P}', g)$ since

$$\pi(\mathscr{P}') = \left\{ \bigotimes_{i=1}^{n} Q : Q \in \mathcal{P}(\mathbb{R}^2), \int_{\mathbb{R}^2} y^2 \, dQ(z,y) < \infty, Q(\{0\} \times \mathbb{R}) < 1 \right\}$$

so that we can take

$$G(\pi(P)) = \left(\int_{\mathbb{R}^2} z^2 \, d\pi_1(P)(z,y) \right)^{-1} \int_{\mathbb{R}^2} zy \, d\pi_1(P)(z,y)$$

for any $P \in \mathscr{P}'$. In particular, if $P_{\mathrm{X,U}} \in \mathscr{P}$, then $G(P_{\mathrm{X}}) = (\mathbb{E}[Z_1^2])^{-1}\mathbb{E}[Y_1 Z_1]$. The identified model $(\mathscr{P}', g)$ is directly seen to be congruent to the phenomenological model considered in Example 3.4 (up to a standard homeomorphism between $\{\{x\} : x \in \mathbb{R}\}$ and $\mathbb{R}$).

Example 4.5 (AR(1) Model)**.** We introduce a theoretical statistical model based on a standard time

series stochastic model and show that it is equivalent to a strict submodel of the phenomenological model introduced in Example 3.5. The data $\mathrm{x} = (x_1, x_2, \ldots, x_T)$ is still construed as the realizations of some random variables $\mathrm{X} = (X_1, X_2, \ldots, X_T)$, but the statistical model $(\mathscr{P}, g)$ now targets the distribution of X and some unobservable random variables $\mathrm{U} = (U^0, (U_t : t \in \mathbb{Z}), (X_t : t \in \mathbb{Z} \setminus \{1, \ldots, T\}))$. The set $\mathscr{P}$ of possible distributions for $P_{\mathrm{X,U}}$ is constrained by hypothesizing a set of stochastic models obtained by assuming that $U^0 = \rho \in \mathbb{R}$ is constant, $\mathbb{E}[U_t] = 0$ and $\mathbb{E}[U_t^2] = \omega^2 < \infty$ for all $t \in \mathbb{Z}$, and

$$X_t = \rho X_{t-1} + U_t$$

for all $t \in \mathbb{Z}$. The theoretical statistical model is completed by taking

$$g \colon \mathscr{P} \to \mathbb{R},\ P \mapsto \rho.$$

Without additional assumptions, the model is not identified. For instance, if $(X_t : t \in \mathbb{Z})$ is an i.i.d. sequence with $X_1 \sim N(0, 1)$, then taking $\rho_1 = 0$ and $U_{t,1} = X_t$ and taking $\rho_2 \neq 0$ and $U_{t,2} = X_t - \rho_2 X_{t-1}$ both satisfy the constraints in $\mathscr{P}$ and lead to the same observables' distribution. We can define a restricted set $\mathscr{P}'$ based on further restricting the stochastic models by assuming that $|\rho| < 1$ and $\mathbb{E}[U_t X_{t-1}] = 0$ for all $t \in \mathbb{Z}$. The set $\mathscr{P}'$ then leads to an identified model $(\mathscr{P}', g)$. Indeed, we have

$$\pi(\mathscr{P}') = \left\{ P \in \mathcal{P}(\mathbb{R}^T) : \int_{\mathbb{R}^T} x \, dP(x) = 0,\ \int_{\mathbb{R}^T} xx^\top \, dP(x) \in W^T \right\}$$

where $W^T = \{\Sigma \in S_{++}^T : \exists \omega^2 > 0, \rho \in (-1, 1) \text{ s.t. } \Sigma_{i,j} = \omega^2 \rho^{|i-j|}/(1 - \rho^2) \text{ for all } i, j \in \{1, \ldots, T\}\}$. Identification is then achieved by taking G as

$$G(\pi(P)) = \left(\int_{\mathbb{R}} x^2 \, d\pi_1(P)(x) \right)^{-1} \int_{\mathbb{R}^2} xy \, d\pi_{1,2}(P)(x, y)$$

for all $P \in \mathscr{P}'$. Since $W^T \subset V^T$, we directly have that $(\mathscr{P}', g)$ is congruent to a strict submodel of the phenomenological statistical model considered in Example 3.5.

4.3 Applications

Example 4.6 (Random Utility Model)**.** Utility maximization is a standard theory in microeconomics used to explain choices made by individuals among a finite set of alternatives. Microeconomists hypothesize the existence of a theoretical quantity, called utility, that any individual enjoys from choosing an alternative. Microeconomists then assume that any individual chooses the option that provides the most utility. Utility is a theoretical quantity: it is not directly observable, but microeconomists map it, at least partially, to observable quantities. To rigorously bring this theory of choice to the data, statistical models of a particular nature need to be considered. Since utility as a theoretical quantity is not directly observable, these statistical models are inevitably theoretical according to Definition 4.1. To illustrate this, we introduce a simple example with only two

options. Many more complicated models exist – see Galichon (2026) for a comprehensive review. In our simple setting, the data $\mathrm{x} = ((y_1, v_1^a, v_1^b), \ldots, (y_n, v_n^a, v_n^b))$ is viewed as the realizations of some i.i.d. random triples $\mathrm{X} = ((Y_1, V_1^a, V_1^b), \ldots, (Y_n, V_n^a, V_n^b))$, where $y_i \in \{0, 1\}$ corresponds to the choice of the representative agent and $(v_i^a, v_i^b) \in \mathbb{R}^2$ to the observed components of the utility derived for each option $\{a, b\}$. The random choices $(Y_1, \ldots, Y_n)$ are then hypothesized to be the result of utility maximization, where utility is assumed to be the sum of the observable variables $((V_1^a, V_1^b), \ldots, (V_n^a, V_n^b))$ and some other variables $\mathrm{U} = ((\varepsilon_1^a, \varepsilon_1^b), \ldots, (\varepsilon_n^a, \varepsilon_n^b))$ that are taken as unobservable. For any $i \in \{1, \ldots, n\}$, it is further assumed that $(\varepsilon_i^a, \varepsilon_i^b)$ is independent of (V_i^a, V_i^b). This structure naturally delineates a set of stochastic models (which are part of what are known as random utility models) given by

$$Y_i = \begin{cases} 1 & \text{if } V_i^a + \varepsilon_i^a > V_i^b + \varepsilon_i^b \\ 0 & \text{if } V_i^a + \varepsilon_i^a \leq V_i^b + \varepsilon_i^b \end{cases}$$

for all $i \in \{1, \ldots, n\}$. From the statistical viewpoint, these models constrain a set $\mathscr{P}$ of possible distributions for the joint distribution $P_{\mathrm{X,U}}$ of the random choices and utilities. Several possible statistical parameters can then be chosen to form valid theoretical statistical models. For simplicity, we can directly consider the marginal distribution of the random choices, that is,

$$g \colon \mathscr{P} \to \left\{Q(\cdot \times \mathbb{R}^2) : Q \in \pi_1(\mathscr{P})\right\},\ P \mapsto \pi_1(P)(\cdot \times \mathbb{R}^2).$$

The model $(\mathscr{P}, g)$ is then trivially identified by taking $G(\pi(P)) = \pi_1(P)(\cdot \times \mathbb{R}^2)$ for all $P \in \mathscr{P}$, and the phenomenological statistical model that $(\mathscr{P}, g)$ is congruent to can be used for inference. In this case, the inferential results do not simply provide a probabilistic description of the observable choices but characterize them empirically as the outcome of utility maximization.

Example 4.7 (Potential Outcome Model)**.** Causal inference is a multidisciplinary approach that aims at explaining observable phenomena through the prism of causality. Causality is a complex metaphysical notion that is commonly tackled by means of intermediary theoretical quantities known as potential outcomes (or counterfactuals). A potential outcome for some action is defined as what an observable outcome would have been had this action been taken. By its very nature, a potential outcome may not be observable. This theoretical edifice can be used to answer causal queries but, in order to bring it rigorously to the data, some statistical models of a particular nature need to be considered. Since some potential outcomes are necessarily unobservable, these statistical models are inevitably theoretical according to Definition 4.1. To illustrate this, we introduce a simple example based on a binary treatment. More complicated models can be considered – see, for instance, Ding (2024) for additional examples. In our simple setting, the data $\mathrm{x} = ((y_1, d_1), \ldots, (y_n, d_n))$ is construed as the realizations of some i.i.d. random pairs $\mathrm{X} = ((Y_1, D_1), \ldots, (Y_n, D_n))$, where $y_i \in \mathbb{R}$ represents the outcome of interest and $d_i \in \{0, 1\}$ the treatment status. If $d_i = 1$, the treatment was administered. If $d_i = 0$, it was not. The potential outcomes are introduced as the i.i.d. unobservable random pairs $\mathrm{U} = ((Y_1(1), Y_1(0)), \ldots, (Y_n(1), Y_n(0)))$ and are linked to the

observables by means of the simple rule $Y_i = Y_i(d)$ if $D_i = d$ for all $i \in \{1, \ldots, n\}$. This general rule simply says that a potential outcome for some action is equal to the observable outcome if this action was taken. In the binary case, it is equivalent to the simple stochastic model given by

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

for all $i \in \{1, \ldots, n\}$. From the statistical viewpoint, these models constrain a set $\mathscr{P}$ of possible distributions for the joint distribution $P_{\mathrm{X,U}}$ of the observable outcomes, treatment statuses, and potential outcomes. A plethora of statistical parameters can be considered to form valid theoretical statistical models. For simplicity, we can consider the most standard statistical parameter in causal inference – the average treatment effect – which corresponds to the difference in expectation for the potential outcomes with and without treatment, that is,

$$g \colon \mathscr{P} \to \mathbb{R},\ P \mapsto \int_{\mathbb{R}^2} (y_1 - y_0)\, d\bar{\pi}_1(P)(y_1, y_0).$$

In particular, if $P_{\mathrm{X,U}} \in \mathscr{P}$, then $g(P_{\mathrm{X,U}}) = \mathbb{E}[Y_1(1)] - \mathbb{E}[Y_1(0)]$. However, the theoretical statistical model $(\mathscr{P}, g)$ is not identified. A counterexample is given by taking $D_i \sim \mathrm{Bernoulli}(1/2)$ and $Y_i = 0$. Then $P' \in \mathscr{P}$ with $Y_i'(0) = Y_i'(1) = 0$ and $P^\dagger \in \mathscr{P}$ with $Y_i^\dagger(1) = 2(1 - D_i)$ and $Y_i^\dagger(0) = 0$ satisfy the model. These choices yield $\pi(P') = \pi(P^\dagger)$, but $g(P') = 0 \neq 1 = g(P^\dagger)$. A conventional path to identification is to restrict $\mathscr{P}$ to $\mathscr{P}'$ based on further constraining the potential outcome model by assuming that $P(D_i = 1) \in (0, 1)$, $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)]$, and $\mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)]$ for all $i \in \{1, \ldots, n\}$. The last two conditions hold, in particular, when $Y_i(1)$ and $Y_i(0)$ are independent of D_i. This captures the idea that the individuals who receive the treatment do not differ from the individuals who do not receive it. The model is then identified by taking

$$G(\pi(P)) = \frac{\int_{\mathbb{R}\times\{1\}} y\, d\pi_1(P)(y, d)}{\pi_1(P)(\mathbb{R} \times \{1\})} - \frac{\int_{\mathbb{R}\times\{0\}} y\, d\pi_1(P)(y, d)}{\pi_1(P)(\mathbb{R} \times \{0\})}$$

for all $P \in \mathscr{P}'$. In particular, if $P_{\mathrm{X,U}} \in \mathscr{P}$, then $G(P_{\mathrm{X}}) = \mathbb{E}[Y_1|D_1 = 1] - \mathbb{E}[Y_1|D_1 = 0]$. The phenomenological model that $(\mathscr{P}', g)$ is congruent to can then be used for inference. In this case, the inferential results do not simply provide a probabilistic description of the outcomes and treatment statuses but an empirical characterization of what is interpreted as the average causal effect of the treatment on the outcome.

Remark 4.7. In Example 4.7, the mean independence assumption for the potential outcomes is often motivated by invoking some form of randomization in the treatment assignment. None of these hypotheses on the unobservable structure can be directly tested. Only the congruent phenomenological model can be subject to falsification provided another draw from the same distribution $\pi(P^*) = P_{\mathrm{X}}$ is available – see again Remark 4.6. In causal inference, the impossibility of directly verifying the independence assumption on the potential outcomes is the main *leap of*

faith from the observables to the underlying causal structure. As with any theoretical statistical model, it is the price to pay for the extension of the statistical reach.

References

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *The Annals of Mathematical Statistics*, 27(4), 1115–1122.

Blackwell, D. (1951): "Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, pp. 93–102. University of California Press, Berkeley.

Breiman, L. (2001): "Statistical Modeling: The Two Cultures," *Statistical Science*, 16(3), 199–231.

Cox, D. R. (1990): "Role of Models in Statistical Analysis," *Statistical Science*, 5(2), 169–174.

Ding, P. (2024): *A First Course in Causal Inference*. CRC Press.

Donoho, D. L., and R. C. Liu (1987): "Geometrizing Rates of Convergence, I," Discussion Paper 137, Dept. Statistics, Univ. California, Berkeley.

——— (1991a): "Geometrizing Rates of Convergence, II," *The Annals of Statistics*, pp. 633–667.

——— (1991b): "Geometrizing Rates of Convergence, III," *The Annals of Statistics*, pp. 668–701.

Galichon, A. (2026): *Discrete Choice Models*. Princeton University Press.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer.

Hurwicz, L. (1950): "Generalization of the Concept of Identification," in *Statistical Inference in Dynamic Economic Models*, vol. 10, pp. 245–57. Wiley New York.

Koopmans, T. C. (1949): "Identification Problems in Economic Model Construction," *Econometrica*, pp. 125–144.

Le Cam, L. (1964): "Sufficiency and Approximate Sufficiency," *The Annals of Mathematical Statistics*, pp. 1419–1455.

Le Cam, L., and G. L. Yang (2000): *Asymptotics in Statistics*. Springer.

Lehmann, E. (1990): "Model Specification: The Views of Fisher and Neyman, and Later Developments," *Statistical Science*, 5(2), 160–168.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses*. Springer.

Liu, R. C., and L. D. Brown (1993): "Nonexistence of Informative Unbiased Estimators in Singular Problems," *The Annals of Statistics*, pp. 1–13.

Matzkin, R. L. (2007): "Nonparametric Identification," in *Handbook of Econometrics*, vol. 6B, pp. 5307–5368. Elsevier.

McCullagh, P. (2002): "What Is a Statistical Model?," *The Annals of Statistics*, 30(5), 1225–1310.

ROBERT, C. P. (2007): *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*. Springer.

WALD, A. (1939): "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *The Annals of Mathematical Statistics*, 10(4), 299–326.