Where Do Statistical Models Come From?
Revisiting the Problem of Specification

Aris Spanos
Department of Economics
Virginia Tech,
Blacksburg, VA 24061

ABSTRACT

R. A. Fisher founded modern statistical inference in 1922 and identified its fundamental problems to be: *specification*, *estimation* and *distribution*. Since then the problem of *statistical model specification* has received scant attention in the statistics literature. The paper traces the history of statistical model specification, focusing primarily on pioneers like Fisher, Gosset, Egon Pearson, Neyman, and more recently Lehmann and Cox, and proposes a new approach to specification, known as the *Probabilistic Reduction* (PR) approach, as a synthesis and extension of these views.

The PR approach is based on the premises that empirical models constitute an amalgam of theoretical and statistical information. Hence, it distinguishes between a *structural* (substantive) and a *statistical* (empirical) *model*, viewing the former as a *reparameterization/restriction* of the latter. That is, a *structural model* gains statistical *operational meaning* when *embedded* into a statistical model. A pre-specified statistical model is viewed as a subset of the set $\mathbb{P}$ of all statistical models that could have (potentially) given rise to the data $\mathbf{Z}:=(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ in question. Each statistical model $\mathcal{M}$ in $\mathbb{P}$ is characterized by the probabilistic (reduction) assumptions on $D(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n;\boldsymbol{\theta})$, the *joint distribution* of the observable stochastic processes $\{\mathbf{Z}_t, t\in\mathbb{T}\}$, that would give rise to this model via a *reduction*; hence the term PR. For instance, the **Linear Regression model**:

$$y_t=\boldsymbol{\beta}'\mathbf{x}_t+u_t,\ t\in\mathbb{T},$$

based on $D(y_t|\ \mathbf{x}_t;\boldsymbol{\varphi}_1)$, is characterized by the reduction assumptions: $\{\mathbf{Z}_t, t\in\mathbb{T}\}$ is a Normal, IID process. These assumptions give rise to the reduction:

$$D(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n;\boldsymbol{\theta}) = \prod_{t=1} D(\mathbf{Z}_t;\boldsymbol{\varphi})= \prod_{t=1} D(y_t|\ \mathbf{x}_t;\boldsymbol{\varphi}_1)\ D(\mathbf{x}_t;\boldsymbol{\varphi}_2).$$

By choosing different reduction assumptions from three basic categories, *Distribution*, *Dependence*, and *Heterogeneity*, one can specify numerous statistical models which include all the well-known models in statistics and other applied fields. This way of viewing statistical models renders specification the cornerstone of statistical modeling because it defines the *premises* of inference and brings out the problem of securing the *reliability* and *precision of inference*, by ensuring the *soundness* and the *informational content* of the premises.

By mapping out the 'territory' surrounding a statistical model, the set $\mathbb{P}$ provides the framework in the context of which all other phases of modeling, including **Estimation** (estimating the unknown parameters)**, Misspecification Testing** (probing its underlying assumptions thoroughly for possible departures)**, Respecification** (choosing a more appropriate statistical model), **Identification** (relating the statistical to the structural model), **Hypothesis Testing** and **Prediction**, take place. A primary advantage of this framework is that it enables the modeler to secure the validity of the pre-specified statistical model (**statistical adequacy**) in order to use data **Z** to draw reliable inferences concerning the phenomenon of interest.