

1.5 Causal versus Statistical Terminology

This section defines fundamental terms and concepts that will be used throughout this book. These definitions may not agree with those given in standard sources, so it is important to refer to this section in case of doubts regarding the interpretation of these terms.

A **probabilistic parameter** is any quantity that is defined in terms²⁵ of a joint probability function. Examples are the quantities defined in Sections 1.1 and 1.2.

A **statistical parameter** is any quantity that is defined in terms of a joint probability distribution of observed variables, making no assumption whatsoever regarding the existence or nonexistence of unobserved variables.

Examples : the conditional expectation $E(Y|x)$,
the regression coefficient r_{YX} ,
the value of the density function at $y = 0$, $x = 1$.

A **causal parameter** is any quantity that is defined in terms of a causal model (as in (1.40)) and is not a statistical parameter.

Examples : the coefficients α_{ik} in (1.41),
whether X_9 has influence on X_3 for some u ,
the expected value of Y under the intervention $do(X = 0)$,
the number of parents of variable X_7 .

Remark: The distinction between probabilistic and statistical parameters is devised to exclude the construction of joint distributions that invoke hypothetical variables (e.g., counterfactual or theological). Such constructions, if permitted, would qualify any quantity as statistical and would obscure the distinction between causal and noncausal assumptions.

A **statistical assumption** is any constraint on a joint distribution of observed variable; for example, that f is multivariate normal or that P is Markov relative to a given DAG D .

²⁵A quantity Q is said to be *defined in terms of* an object of class C if Q can be computed uniquely from the description of any object in class C (i.e., if Q is defined by a functional mapping from C to the domain of Q).

A **causal assumption** is any constraint on a causal model that cannot be realized by imposing statistical assumptions; for example, that f_i is linear, that U_i and U_j (unobserved) are uncorrelated, or that x_3 does not appear in $f_4(pa_4, u_4)$. Causal assumptions may or may not have statistical implications. In the former case we say that the assumption is “testable” or “falsifiable.”

Remark: The distinction between causal and statistical parameters is crisp and fundamental. Causal parameters can be discerned from joint distributions only when special assumptions are made, and such assumptions must have causal components to them. The formulation and simplification of these assumptions will occupy a major part of this book.

Remark: Temporal precedence among variables may furnish some information about (the absence of) causal relationships—a later event cannot be the cause of an earlier event. Temporally indexed distributions such as $P(y_t|y_{t-1}, x_t)$, $t = 1, \dots$, which are used routinely in economic analysis, may therefore be regarded as borderline cases between statistical and causal models. We shall nevertheless classify those models as statistical because the great majority of policy-related questions cannot be discerned from such distributions, given our commitment to making no assumption regarding the presence or absence of unmeasured variables. Consequently, econometric concepts such as “Granger causality” (Granger 1969) and “strong exogeneity” (Engle et al. 1983) will be classified as statistical rather than causal.²⁶

Remark: The terms “theoretical” and “structural” are often used interchangeably with “causal”; we will use the latter two, keeping in mind that some structural models may not be causal (see Section 7.2.5).

Causal versus Statistical Concepts

The demarcation line between causal and statistical parameters extends as well to general concepts and will be supported by termino-

²⁶Caution must also be exercised in labeling as “data-generating model” the probabilistic sequence $P(y_t|y_{t-1}, x_t)$, $t = 1, \dots$ (e.g. Davidson and MacKinnon 1993, p. 53; Hendry 1995). Causal assumptions of the type developed in Chapter 2 (see Definitions 2.4.1 and 2.7.4) must be invoked before applying such sequences in policy-related tasks.

logical distinction. Examples of *statistical* concepts are: correlation, regression, conditional independence, association, likelihood, collapsibility, risk ratio, odd ratio, and so on. Examples of *causal* concepts are: randomization, influence, effect, confounding, exogeneity, ignorability, disturbance (e.g. (1.40)), spurious correlation, path coefficients, instrumental variables, intervention, explanation, and so on. The purpose of this demarcation line is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators into treating nonstatistical concepts with the proper set of tools.

Some readers may be surprised by the idea that textbook concepts such as randomization, confounding, spurious correlation, or effects are nonstatistical. Others may be shocked at the idea that controversial concepts such as exogeneity, confounding, and counterfactuals *can* be defined in terms of causal models. This book is written with these readers in mind, and the coming pages will demonstrate that the distinctions just made between causal and statistical concepts are essential for clarifying both.