Bibliographical Notes for Chapter 4

The definition of counterfactuals as derivatives of structural equations, Eq. (4.5), was introduced by Balke and Pearl (1994a,b), who applied it to the estimation of probabilities of causation in legal settings. The philosopher David Lewis defined counterfactuals in terms of similarity among possible worlds Lewis (1973). In statistics, the notation $Y_x(u)$ was devised by Neyman (1923), to denote the potential response of unit u in a controlled randomized trial, under treatment $X = x$. It remained relatively unnoticed until Rubin (1974) treated Y_x as a random variable and connected it to observed variable via the consistency rule of Eq. (4.6), which is a theorem in both Lewis's logic and in structural models. The relationships among these three formalisms of counterfactuals are discussed at length in Pearl (2000, Chapter 7) where they are shown to be logically equivalent; a problem solved in one framework would yield the same solution in another. Rubin's framework, known as "potential outcomes," differs from the structural account only in the language in which problems are defined, hence, in the mathematical tools available for their solution. In the potential outcome framework, problems are defined algebraically as assumptions about counterfactual independencies, also known as "ignorability assumptions." These types of assumptions, exemplified in Eq. (4.15), may become too complicated to interpret or verify by unaided judgment, In the structural framework, on the other hand, problems are defined in the form of causal graphs, from which dependencies of counterfactuals (e.g., Eq. (4.15)), can be derived mechanically. The reason some statisticians prefer the algebraic approach is, primarily, because graphs are relatively new to statistics. Recent books in social science (e.g., Morgan and Winship 2014) and in health science (e.g., VanderWeele 2015) are taking the hybrid, graph-counterfactual approach pursued in our book.

The section on linear counterfactuals is based on Pearl (2009, pp. 389–391). Recent advances are provided in Cai and Kuroki (2006) and Chen and Pearl (2014). Our discussion of ETT (Effect of Treatment on the Treated), as well as additive interventions, is based on Shpitser and Pearl (2009), which provides a full characterization of models in which ETT is identifiable.

Legal questions of attribution, as well as probabilities of causation are discussed at length in Greenland (1999) who pioneered the counterfactual approach to such questions. Our treatment of PN, PS, and PNS is based on Tian and Pearl (2000) and Pearl (2009, Chapter 9). Recent results, including the tool kit of Section 4.5.1 are given in (Pearl 2015a).

Mediation analysis (Sections 4.4.5 and 4.5.2), as we remarked in Chapter 3, has a long tradition in the social sciences (Duncan 1975; Kenny 1979), but has gone through a dramatic revolution through the introduction of counterfactual analysis. A historical account of the conceptual transition from the statistical approach of Baron and Kenny (1986) to the modern, counterfactual-based approach of natural direct and indirect effects (Pearl 2001; Robins and Greenland 1992) is given in Sections 1 and 2 of (Pearl 2014a). The recent text of VanderWeele (2015) enhances this development of new results and new applications. Additional advances in mediation, including sensitivity analysis, bounds, multiple mediators, and stronger identifying assumptions are discussed in Imai et al. (2010) and Muthén (2011).

The mediation tool kit of Section 4.5.2 is based on Pearl (2014a). Shpitser (2013) has derived a general criterion for identifying indirect effects in graphs.