The Causal Foundation of Applied Probability and Statistics

Sander Greenland, 2021

Introduction

- Scientific inference is a branch of causality theory.
- Statistical theory has been developed focusing on (recognized as) a branch of applied mathematics such as probability theory, but causal theory is essential for statistical science.
- The application of statistics should consider causation if it is to represent underlying reality from observed data.
  - Any real-world data analysis has a causal component with underlying network generating the data, even for the descriptive analysis.
  - Decision analysis showing the effects of the decisions requires design strategies to rule out alternative explanations of the observations.

Introduction

- Causality is not extra-statistical but instead is a logical foundation of statistical analysis.
  - e.g. claims of random (unbiased/ignorable) sampling requires causal actions to block unwanted causal effects on the sample patterns.
  - w/o such causal analysis, independence can only be treated as a subjective assumption that has to claim no confounders affecting selection and outcomes.
- Incorporation of causality into introductory statistics from study design to data analysis is urgently necessary.
  - Probability cannot be an adequate foundation for applied statistics.
  - Statistical practice integrates logic, context, and probability into scientific inference and decision, using **causal narratives** to explain diverse data.

Causality is central even for purely descriptive goals

- Causal descriptions encode the information and goals that lead to concerns about associations.
- Consider a survey such as proportion of voters who would vote for a given candidate.
  - Some characteristics C may affect both survey participation (S=1) and voting intent V
  - $[S=1] \leftarrow C \rightarrow V$
    - [S=1] indicates the observations are conditioned on S=1.
  - If the distribution of C is different b/w observed sample (S=1) and population, then distribution of V from the data is biased: $P(V=v|S=1) \neq P(V=v)$.
    - Causal relation $C \rightarrow S$ causes bias even for the descriptive study.

Causality is central even for purely descriptive goals

- We can remove the bias by reweighting the sample using target-population ethnicity distribution.
  - It is also a causal process: need to obtain target-weighting data and program the reweighting to make the adjusted estimate.
  - $[S=1] \leftarrow C \rightarrow V \leftarrow W \leftarrow C$
    - New causal diagram with reweighting procedure.
    - W is a weighting intervention to adjust target-population distribution and obtain unbiased distribution of V: $P(V=v)$

The strength of probabilistic independence demands physical independence

- Data generating does not only mean some abstract structural equation.
- Consider coin tossing and its causal diagram: $Y_1, \cdots, Y_N$
  - N isolated (unconnected) nodes for independent identical distribution
    - No arrows b/w nodes
  - iid is not a single assumption, but a set of assumptions: N! dependency pattern b/w Y_i.
  - Suppose that we have N observational data about the coin tossing.
  - The dependency grows much father than the number of observations N.
  - The amount of deductive information in this assumption set is beyond data alone could contain; only contextual (background and design) information can supply enough information to warrant the assumptions.

The strength of probabilistic independence demands physical independence

- Consider again descriptive or decision analysis decision analysis
- Only the physical action of blocking all causal effects on selection or treatment can provide deductive justification for the entire set of assumptions corresponding to independence.

The Superconducting Supercollider of Selection

- In human field studies, a selection indicator node S should be considered as part of the data generating process.
  - S may be influenced by study variable.
  - By definition, only samples with S=1 are observed; it is always conditioned on
  - If S is affected by more than one variable, it is a conditioned collider: a source for a potential bias.
  - However, this fact is often ignored in studies.

The Superconducting Supercollider of Selection

- Selection bias may arise even when S is not a collider: [S=1]←C→V
  - C opens back-door path from S to V as a confounder.
  - The bias can be adjusted by condition on C.
  - The marginal (C-unconditional) distribution of V, P(V=v), can be obtained by integrating C out.
- A parallel example of selection bias in treatment w/o collider bias: consider the effect of treatment X on outcome Y w/ a modifier C.
- [S=1]←C→Y← X
  - C is independent of X, and Y is independent of S given C, but the path S←C→Y causes bias in estimating the treatment effect given S=1 (condition on selection).

Data and algorithm are causes of reported results

- Valid statistical analysis is causal to the core; the core problem is about factors causing differences in distributions of those targeted and those observed: voter and survey responders; patients with a given indicator and patients in a trial.
    - w/o a causal model, we do not have basis to connect probability calculations of the observations and the world.
- Statistics is laying out the causal sequence leading from data to inferences and decisions.
    - The outputs of statistical analysis is physically justified only when it is deduced form the causal structure of the generator.

Getting causality into statistics by putting statistics into causal terms from the start

- Causal explanations provide the contextual justifications for the probability models used in the analysis, displaying information about study features that physically constrain data generation.
- Researchers should understand/learn causal thinking before probability and mathematical statistics.
  - Introductory statistics should cover basic logic and its causal extensions before mathematical statistics theory.

Causation in 20th-century statistics

- Foundations of causation date back to early 20th century
  - Neyman [1923] proposed potential outcome
  - Potential outcome (counterfactual) models entered statistics journals by the 1930's
- Statistical developments in the 20th century were foremost concerned with causal inferences derived from physical randomization
  - Fisher laid out potential outcomes clearly
  - Formalized by others into potential-outcome model

Causal analysis vs. traditional statistical analysis

- Causal theories can include important mistakes even while successfully predicting intervention effects
  - e.g. malaria: coming from parent Italian meaning bad air
  - Malaria rates were higher near swamps; attributed that to toxic air
    - swamp → toxic air → malaria
  - Led to successful intervention such as draining swamps and building elevated houses
  - Missed the actual causal structure
    - swamp → mosquito exposure → malaria
  - Swamp intervention can test only swamp → malaria effect, not the intermediate pathways

Causal analysis vs. traditional statistical analysis

- An intervention experiment provides evidence only on classes of mechanisms, not specific mechanisms.
- Even randomized controlled trial may not identify causal effects.
- It applies even more strongly to passive observational (non-experimental) studies.
  - Extraction of information about target (treatment) effect in observational studies requires causal models for physical data generation that include nonrandom variation (bias) sources beyond the treatment.

Relating causality to traditional statistical philosophies and “objective” statistics

- Frequentism and Bayesianism are incomplete both as learning theories and as philosophies of statistics, and in sufficient for all sound applications
    - Causal justifications are the foundation for classical frequentism
    - And for Bayesianism
- Many statisticians assign primacy to objective model components
    - Derivable from observed mechanisms, such as random number generators
    - Strong assumptions such as randomness can be deduced from the physical data-generating mechanisms, not from observed frequencies or other purely associational information

Conclusion

- Statistical science requires realistic causal models, to analyze data, for the generation of the data and the deduction of their empirical consequences.
- Decision analysis requires further causal analysis to explain various pathways.
- Causal foundation is essential for the teaching and applications of statistics.