Book Proposal:

Theory of Agnostic Statistics

Peter M. Aronow and Benjamin T. Miller*

February 13, 2016

1 Rationale and Scope

Reflecting a sea change in how empirical research has been conducted over the past decade, *Theory of Agnostic Statistics* is a foundational text that develops a formalization of statistical theory for the social and health sciences without invoking strong modeling assumptions. The book represents the first unified, book-length treatment of probability theory, statistical inference and nonparametric identification for applied topics (including missing data and causal inference). We provide a formal justification and a cogent argument for the use of *agnostic statistics*, a perspective on empirical research that implies minimal assumptions and emphasizes credible inference. We believe that our book is uniquely suited to teach readers the foundational logic of statistics in a manner that prepares them for critical engagement with advanced statistical inference and research methods. Although there exist many textbooks that address nonparametric statistics or causal inference, we believe there is no other book accessible to first-year graduate students that builds from first principles to allow students to understand cutting edge statistical methods.

Our book begins with a comprehensive treatment of probability theory and asymptotic statistics in the frequentist paradigm. We then develop the core idea of *plug-in estimation*, a means of estimating features of probability distributions with little to no additional statistical assumptions. We develop this idea to explain complex methods in simple terms, including the bootstrap, kernel estimation, penalized regression, sieve estimation, and doubly robust estimators. We proceed to derive the operating characteristics of linear regression as a plug-in estimator, without using additional assumptions about the underlying process that gives rise to the data. Using these results, we then tackle the related challenges of missing data and causal inference, showing that when

*Peter M. Aronow (contact author) is Assistant Professor, Departments of Political Science and Biostatistics, Yale University, 77 Prospect St., New Haven, CT 06520 (Email: peter.aronow@yale.edu). Benjamin T. Miller is Graduate Student, Department of Political Science, Yale University, 115 Prospect St., New Haven, CT 06520 (Email: benjamin.miller@yale.edu).

simple statistical intuitions and methods are wedded to substantive assumptions, researchers can obtain principled approximations to real-world problems. We conclude with a chapter on parametric models, the traditional approach to inference, and show how standard models can be profitably used to approximate more complex inferential targets without assuming their validity. Together, our results provide a unified treatment of statistical inference for researchers unwilling to make assumptions beyond what they or their audience would find credible.

Theory of Agnostic Statistics is novel in many respects, but is most innovative in its accessible, yet formal, treatment of technical topics. To this end, we provide new proofs for well-known theorems using minimal mathematics, allowing even novice readers to understand seemingly complex ideas. The book makes minimal use of matrix algebra, which we have found consumes a great deal of students' time without clarifying core ideas. We believe our book to be the first to provide a formal justification for so many advanced topics without demanding a level of mathematical sophistication unreasonable to expect of first-year students in the social and health sciences.

We expect that our book will prepare students to engage with books and articles on more specialized topics, arming them with the ability to derive and understand any new methods under credible assumptions. *Theory of Agnostic Statistics* is designed as a foundational text, and accordingly there is much that the book does not cover in depth, namely, Bayesian inference, model-based inference (e.g., model testing, model diagnostics, and graph-based models), and some recent techniques in causal inference (e.g., instrumental variables, regression discontinuity designs). We believe all of these topics have been covered well in other textbooks, while our book provides a set of tools for understanding these methods building from the basics of probability theory.

2 Prospective Readers and Market

Theory of Agnostic Statistics is designed to be used in conjunction with a one-year sequence in statistics for graduate students or perhaps advanced undergraduates in the social and health sciences. The book is also suitable for a one-semester course if Chapters 4-6 are not addressed in depth in the classroom. We expect that the book will be used by many graduate courses, as a foundational text before moving on to more specialized books. Table 1 lists a sampling of courses (restricting our attention to the overall top-10 research universities, as ranked by the 2015 U.S. News and World Report) that could use our coursebook.

The audience of the book goes beyond the classroom, however. We believe the book to be appropriate and useful for all applied researchers who seek to develop a foundational understanding of statistics to support empirical research. Much like Angrist and Pischke (2009)'s *Mostly Harmless Econometrics*, we expect the book to be an important reference on the bookshelf of a very wide range of scholars working in the empirical social and health sciences.

Drafts of the book have been used in the year-long doctoral-level methodology sequence (PL500 and PL503) in the Political Science department at Yale University over two years (AYs 2013-2014, 2014-2015). These courses included many students from public health, biostatistics, and

University	Course Name	Course Number	Department
Harvard	Introduction to Quantitative Methods I	GOV2000	Government
Harvard	Intermediate Quantitative Research Methods	SOCIOL202	Sociology
Harvard	Statistical Inference	STAT211	Statistics
Harvard	Probability Theory and Applications II	BIOSTAT250	Biostatistics
Harvard	Econometrics I	ECON2110	Economics
Harvard	Advanced Quantitative Methods I: Statistics	API209	Kennedy School
Harvard	Basics of Statistical Inference	BIO222-01	Public Health
Harvard	Intro to Statistical Methods	BIO201-01	Public Health
Harvard	Statistical Inference I	BIO231-01	Public Health
Yale	Statistics	PLSC500	Political Science
Yale	Quantitative Methods	PLSC503	Political Science
Yale	Econometrics (IDE)	ECON558	Economics
Yale	Multivariate Statistics	PSYC518	Psychology
Yale	Methods in Quantitative Sociology	SOCY580	Sociology
Yale	Intro to Statistics: Political Science	STAT502	Statistics
Yale	Intro to Statistics: Social Sciences	STAT503	Statistics
Princeton	Quantitative Analysis I	POL571	Politics
Princeton	Quantitative Analysis II	POL572	Politics
Princeton	Quantitative Analysis in Psych Research I	PSY503	Psychology
Princeton	Econometric Theory I	ECO517	Economics
Princeton	Econometric Theory II	ECO518	Economics
Princeton	Advanced Social Statistics	SOC504	Sociology
Princeton	Quantitative Analysis: Basic	507b	Woodrow Wilson School
Princeton	Quantitative Analysis: Advanced	507c	Woodrow Wilson School
Stanford	Political Methodology I: Regression	POLISCI350a	Political Science
Stanford	Statistical Methods for Behavioral and Social Sciences	PSYCH252	Psychology
Stanford	Sociological Methodology I: Introduction	SOC381	Sociology
Stanford	Sociological Methodology II: Principles of Regression Analysis	SOC382	Sociology
Stanford	Introduction to Statistical Inference	STATS200	Statistics
Stanford	Introduction to Probability and Statistics for Epidemiology	HRP259	Health Research and Policy
Stanford	Applied Econometrics for Public Policy	PUBLPOL 303D	Public Policy
Penn	Advanced Statistical Analysis	SM692	Political Science
Penn	Statistics for Psychologists	STAT500	Psychology
Penn	Intro to Nonparametric Methods	STAT501	Psychology
Penn	Statistical Inference I	BSTA621	Biostatistics
Penn	Intro to Statistics for Health Policy	HPR604	Health Policy
Penn	Intro to Biostats	PUBH501	Public Health
Penn	Applied Econometrics I	STAT520	Statistics
Penn	Statistical Methodology	STAT541	Statistics
Columbia	Quantitative Political Research	W4290	Political Science
Columbia	Intro to Stats in Psychology	G6006	Psychology
Columbia	Statistical Inference	W4107	Statistics
Columbia	Intro to Econometrics	W3412	Economics
Columbia	Intro to Econometrics II	G6412	Economics
Columbia	Quantitative Methods in Program Evaluation	SIPAU8500	SIPA
Columbia	Intro to Biostats Methods II	P6104	Biostatistics
Columbia	Theory of Statistical Inference I	P9109	Biostatistics
Chicago	Statistical Methods of Research II	SOCI30005/01	Sociology
Chicago	Probability and Statistics	STAT32400/01	Statistics
Chicago	Statistical Methods and Foundations for Public Policy I	PPHA31031/01	Public Policy
Chicago	Mathematical Statistics for Public Policy I	PPHA31200/01	Public Policy
Chicago	Applied Econometrics I	PPHA42400/01	Public Policy
Chicago	Applied Econometrics II	PPHA42001/01	Public Policy
Duke	Advanced Quantitative Research Methods	POLSCI748	Political Science
Duke	Applied Multivariate Statistics	PSY720	Psychology
Duke	Social Statistics I	SOC722	Sociology
Duke	Social Statistics II	SOC723	Sociology
Duke	Intro to Econometrics	ECON608	Economics
Duke	Statistical Inference	STA732	Statistics
Duke	Intro to Statistical Theory and Methods I	BIOSTA701	Biostatistics
Duke	Intro to Statistical Theory and Methods II	BIOSTA704	Biostatistics
MIT	Quantitative Research in Political Science II	17.872	Political Science
MIT	Quantitative Research Methods: Multivariate	17.874	Political Science
MIT	Statistical Methods in Economics	14.381	Economics
Caltech	Econometrics	222	Social Sciences

Table 1: Sample of courses offered in AY 2015-2016 at top-10 universities (according to the 2015 U.S. News and World Report) for which *Theory of Agnostic Statistics* would be appropriate.

economics. The current manuscript is being used in a doctoral-level methodology course in the Political Science department at UC Berkeley in fall 2015.

3 Related Books

It is difficult to find a published book that, as a whole, covers the full range of statistical topics, much less provide a formal statement of the "agnostic" philosophy to statistics. This was in fact our impetus for writing the manuscript. However, some elements of our book are found elsewhere, and we believe that our book is highly complementary to these (rather successful) books, as it provides a unifying treatment.

Perhaps closest to our book in spirit is Goldberger (1991)'s *A Course in Econometrics*, now a classic textbook in econometrics. While beloved by many academics (including us), the book is now quite dated. In our experience teaching, students struggle with Goldberger (1991), for both notational issues and because it is quite light on exposition. Furthermore, Goldberger notably omits (i) any treatment of nonparametric identification of causal quantities as considered in the contemporary setting—which is not a fault of the book but simply a fact of the timing of its writing—and (ii) Goldberger spends little time developing the fundamentals of probability theory, which we believe is essential for providing a solid grounding in its ideas. We believe that our book resolves these issues, in large part by updating, clarifying and augmenting the core intuitions of Goldberger (1991) for modern audiences.

Also closely related to our book is Angrist and Pischke (2009). In many ways, *Mostly Harmless Econometrics* begins where our book ends, with some overlap in our discussion of regression and causal inference. We believe *Theory of Agnostic Statistics* naturally leads into *Mostly Harmless Econometrics*, as our book provides statistical foundations for many of the ideas advocated by Angrist and Pischke. The same holds for Angrist and Pischke (2014), *Mastering 'Metrics*, an undergraduate text that also does not develop statistical foundations.

Another related book is Wasserman (2004)'s *All of Statistics*. Our Chapters 1 and 2, particularly with respect to plug-in estimation, mirror much of the discussion in Wasserman. In our experience, Wasserman (2004) is far too technical for the vast majority of students in the social and health sciences. Our treatment of regression is considerably different from Wasserman (focusing on its properties as plug-in estimator), and thus naturally links to the later topics of missing data and causal inference more clearly.

Less related are recent books on causal inference, including Gerber and Green (2012), Hernan and Robins (2016), Morgan and Winship (2014), and Imbens and Rubin (2015). Again, like Angrist and Pischke (2009), these books presume a prior knowledge of statistical inference. Our book is designed to lead into these books, given that we develop a foundational basis for inference and provide some overlap in our treatment of causal inference.

Finally, we note a number of other introductory books that cover statistical inference at the graduate-level for social scientists: e.g., Freedman (2009), Wooldridge (2010), Greene (2011), and Gailmard

(2014). These books are principally concerned with the theory and practice of statistical modeling. There is a rather strong contrast to the philosophy expressed by our book, which at no point assumes the existence of an underlying parametric (or semiparametric) model to be estimated. Accordingly, our nonparametric treatment of statistics naturally leads into discussions of causal inference or missing data all while imposing minimal assumptions that social or health scientists might find credible.

4 Table of Contents

Chapter 0: Introduction (pp. 1–3)

This short introduction outlines the broad philosophy expressed in the book, expectations of the reader, what will and will not be covered, and acknowledgements.

Chapter 1: Probability Theory (pp. 4–65)

In Chapter 1, we develop probability theory beginning from elemental axioms and set theory under the frequentist paradigm. We then proceed to develop the theory of random variables, and summaries thereof. The chapter is novel in that it immediately sets a basis for the use of regression by expressing the conditional expectation function and best linear predictor as features of a probability distribution. At no point do we assume that this random variable must follow a given distribution, and it is thus a nonparametric approach. One unusual feature of the ordering of the book is that we consider probability theory *before* data; this "population first" approach has been used in Goldberger (2004) and Angrist and Pischke (2009), but neither provides a detailed treatment of the first principles of probability theory.

- Derivation of probability spaces under the Kolmogorov axioms
- Properties of probability spaces
- Frequentist interpretation of probability
- Random variables, random vectors and distribution functions
- Characterizing multivariate relationships
- Summarizing random vectors (expected values, moments, correlations)
- Mean squared error
- Conditional expectation functions and best linear predictors

Chapter 2: Learning from Random Samples (pp. 66–92)

Chapter 2 addresses inference from random samples from a given random variable. We address standard results, e.g., law of large numbers, central limit theorem, confidence intervals, hypothesis testing. However, the chapter is unusual in that it uses the logic of plug-in estimation to derive a general theory for estimation of any statistical functional (loosely speaking, a feature of a probability distribution). Thus we are able to address traditionally complex topics, such as kernel density estimation and the bootstrap, all within a unified and straightforward framework.

- Properties of the sample mean and variance
- Asymptotic properties of estimators (laws of large numbers, continuous mapping theorem, Slutsky's theorem, central limit theorem)
- Sampling from finite populations
- Estimation theory (unbiasedness, consistency, asymptotic normality)
- Plug-in estimation and empirical distribution functions
- Kernel density estimation
- Standard errors, confidence intervals, p-values
- The bootstrap as plug-in estimator
- Properties of clustered random samples

Chapter 3: Regression (pp. 93–119)

Chapter 3 derives ordinary least squares linear regression from our nonparametric framework. We show that the regression estimator is a plug-in estimator of the sort derived in Chapter 2, and thus its properties are easily understood without assuming any further model. We provide multiple ways to compute the linear regression estimator and standard errors thereof, including approaches that require no matrix algebra. We proceed to discuss advanced topics in and extensions of regression, including nonlinear methods and penalized methods (such as the lasso).

- Nonparametric derivation of linear regression (as plug-in estimator)
- Regression with matrix algebra or with the Frisch-Waugh-Lowell theorem
- Standard errors for regression, including intuitions under collinearity and different specifications
- Properties of regression with clustered random samples

- Nonlinearity in regression (polynomial approximation, interactions and saturated models)
- Extensions to advanced methods: penalized regression (e.g., the lasso) and sieve estimation

Chapter 4: Missing Data (pp. 120–140)

Chapter 4 develops, in a simple case, the logic of *identification* – or, what can be learned from an unobservable random variable given an observable one. We consider the case of missing data – what can we learn about a random variable when some of its elements cannot be seen in the data that we could collect? We begin with simple bounding results (Manski 2003), then proceed to develop results under various types of as-if random assignment of missingness. One distinguishing feature of this chapter is a focus on a separation between substantive and statistical assumptions: we show that, once identification is achieved, the plug-in estimators developed in Chapters 2 and 3 directly facilitate estimation. We show that popular estimators for missing data, including regression, matching, weighting, and doubly robust estimators, can all be understood as plug-in estimators.

- Nonparametric bounds for population means
- Nonparametric identification under ignorability
- Regression for missing data
- The role of the propensity score
- Hot deck imputation-based estimators
- Weighting estimators
- Doubly robust estimators

Chapter 5: Causal Inference (pp. 141–171)

Chapter 5 extends the ideas of Chapter 4 to causal inference. We first show that causal inference is logically equivalent to a missing data problem, and proceed to directly mirror the language and methods used in Chapter 4. After considering bounds, we develop identification results under various types of (potentially as-if) random assignment. We again develop regression, matching, weighting, and doubly robust estimators under the plug-in principle. We then consider more advanced ideas, including balance testing and further details on the mechanics of regression for causal inference. We conclude the chapter with a very brief discussion of how these ideas extend to popular strategies involving instrumental variables, regression discontinuities, and longitudinal data.

- Potential outcomes and causality

- Nonparametric bounds for causal effects
- Nonparametric identification under ignorability
- Regression for causal inference
- The role of the propensity score
- Matching estimators
- Weighting estimators
- Doubly robust estimators
- Balance testing
- Details on linear regression for causal inference: covariance adjustment, collinearity, and regression weighting
- Brief discussion of extensions to instrumental variables, regression discontinuity designs, and longitudinal data

Chapter 6: Parametric Models (pp. 172–196)

Chapter 6 contains an agnostic derivation of parametric models and maximum likelihood estimation. We begin with a standard definition of parametric models (statistical models that presume that the generative process is governed by a finite number of parameters). We then show examples of these models (i.e., the classical linear model, binary choice models) and standard results on maximum likelihood estimation when the model is correct. However, we then delve into the properties of these methods when the models are not correct. We show that maximum likelihood estimation is itself a plug-in estimator, and it satisfies a "best approximation" property even when the model is wrong. We further show how to conduct inference with parametric models even when the model is believed to be incorrect, and how to wed parametric models to nonparametric causal identification assumptions.

- Definition of parametric models
- Classical linear model
- Binary choice models
- Maximum likelihood estimation under proper specification
- Maximum likelihood estimation under misspecification
- Inference from maximum likelihood estimates

- Nonparametric interpretation of maximum likelihood estimates, including calculation of the conditional expectation function and average partial derivatives
- Mixture models

Chapter 7: Conclusion (pp. 197–198)

This short conclusion discusses the implications of the agnostic philosophy to statistics. We suggest general guidelines for producing and consuming research, focusing on the separation of statistical and substantive assumptions.

Appendices: Glossary of Mathematical Notation and References (pp. 199–203)

We provide a mathematical glossary for terms not otherwise defined in text, largely relating to set theoretic notation. We also provide a reference list.

5 Manuscript Length and Timeline

A full draft of the manuscript is completed, and is submitted (at the editor's request) with the proposal. The manuscript is currently 203 pages (single-spaced, A9 page size), including 18 figures embedded in text.

6 Author Qualifications

Peter M. Aronow is an assistant professor in the departments of Political Science and Biostatistics at Yale University; he is also affiliated with the university's Institution for Social and Policy Studies, Institute for Network Science, and Operations Research Doctoral Program. His work has appeared in the *Annals of Statistics*, *Biometrika*, *American Journal of Political Science*, *Journal of Politics*, *Political Analysis*, *Survey Methodology*, *Journal of Survey Statistics and Methodology*, *Journal of Causal Inference*, *Sociological Methods and Research*, among other journals in the social sciences and in statistics. He currently teaches the first year (two-semester) seminar in statistics and quantitative methods for the Yale doctoral program in Political Science. Professor Aronow holds a PhD in Political Science from Yale University.

Benjamin T. Miller is a doctoral student in the Department of Political Science at Yale University. He served as a teaching fellow, working with Professor Aronow, for a doctoral-level seminar in quantitative methods. Mr. Miller holds a BA in Economics and Mathematics from Amherst College.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics*. Princeton, NJ: Princeton University Press.

Freedman, David A. 2009. *Statistical Models: Theory and Practice*. New York, NY: Cambridge University Press.

Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Gailmard, Sean. 2014. *Statistical Modeling and Inference for Social Science*. New York, NY: Cambridge University Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: WW Norton.

Greene, William H. 2011. *Econometric Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Hernan, Miguel A. and James M. Robins. 2016. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC, forthcoming.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York, NY: Springer-Verlag New York, Inc.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. New York, NY: Cambridge University Press.

Wasserman, Larry. 2004. *All of Statistics*. New York, NY: Springer Science+Business Media, Inc.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.