

Contents

List of Tables	xiii
List of Figures	xv
Preface	xvii
Preface to the Original Book	xix
1 INTRODUCTION	1
1.1 Overview of the Book	3
1.2 How to Use this Book	7
1.3 Introduction to Stata	8
1.3.1 Arithmetic Operations	9
1.3.2 Variables	10
1.3.3 Labels	16
1.3.4 Describing the Data	18
1.3.5 Data Files	20
1.3.6 Merging Data Sets in Stata	21
1.3.7 Packages	23
1.3.8 Programming and Learning Tips	25
1.4 Summary	26
1.5 Exercises	27
1.5.1 Bias in Self-Reported Turnout	27
1.5.2 Understanding World Population Dynamics	28
2 CAUSALITY	32
2.1 Racial Discrimination in the Labor Market	32
2.2 Subsetting the Data in Stata	38
2.2.1 Relational Operators	38
2.2.2 Logical Operators	39
2.2.3 Simple Conditional Statements and Variable Creation	40
2.2.4 Subsetting Using Conditions	43
2.2.5 Preserving and Transforming Data Sets	44
2.3 Causal Effects and the Counterfactual	47

2.4	Randomized Controlled Trials	49
2.4.1	The Role of Randomization	50
2.4.2	Social Pressure and Voter Turnout	51
2.5	Observational Studies	56
2.5.1	Minimum Wage and Unemployment	56
2.5.2	Confounding Bias	60
2.5.3	Before-and-After and Difference-in-Differences Designs	63
2.6	Descriptive Statistics for a Single Variable	67
2.6.1	Quantiles	67
2.6.2	Standard Deviation	72
2.7	Summary	75
2.8	Exercises	75
2.8.1	Efficacy of Small Class Size in Early Education	75
2.8.2	Changing Minds on Gay Marriage	77
2.8.3	Success of Leader Assassination as a Natural Experiment	79
3	MEASUREMENT	81
3.1	Measuring Civilian Victimization during Wartime	81
3.2	Handling Missing Data in Stata	84
3.2.1	Missings Package	85
3.3	Visualizing the Univariate Distribution	86
3.3.1	Bar Plot	87
3.3.2	Histogram	89
3.3.3	Box Plot	92
3.3.4	Printing and Saving Graphs	94
3.4	Survey Sampling	96
3.4.1	The Role of Randomization	96
3.4.2	Nonresponse and Other Sources of Bias	101
3.5	Measuring Political Polarization	105
3.6	Summarizing Bivariate Relationships	106
3.6.1	Scatterplot	106
3.6.2	Correlation	109
3.6.3	Quantile–Quantile Plot	114
3.7	Clustering	117
3.7.1	The <i>k</i> -Means Algorithm	117
3.8	Summary	121
3.9	Exercises	122
3.9.1	Changing Minds on Gay Marriage: Revisited	122
3.9.2	Political Efficacy in China and Mexico	123
3.9.3	Voting in the United Nations General Assembly	125
4	PREDICTION	128
4.1	Predicting Election Outcomes	128
4.1.1	Macros	129
4.1.2	Loops	131
4.1.3	Poll Predictions	133

4.2	Linear Regression	144
4.2.1	Facial Appearance and Election Outcomes	144
4.2.2	Correlation and Scatterplots	146
4.2.3	Least Squares	148
4.2.4	Regression toward the Mean	154
4.2.5	Model Fit	160
4.3	Regression and Causation	167
4.3.1	Randomized Experiments	167
4.3.2	Regression with Multiple Predictors	171
4.3.3	Heterogeneous Treatment Effects	177
4.3.4	Regression Discontinuity Design	184
4.4	Summary	190
4.5	Exercises	190
4.5.1	Prediction Based on Betting Markets	190
4.5.2	Election and Conditional Cash Transfer Program in Mexico	193
4.5.3	Government Transfer and Poverty Reduction in Brazil	195
5	PROBABILITY	197
5.1	Probability	197
5.1.1	Frequentist versus Bayesian	197
5.1.2	Definition and Axioms	199
5.1.3	Permutations	202
5.1.4	Sampling with and without Replacement	205
5.1.5	Combinations	208
5.2	Conditional Probability	210
5.2.1	Conditional, Marginal, and Joint Probabilities	210
5.2.2	Independence	218
5.2.3	Bayes' Rule	226
5.2.4	Predicting Race Using Surname and Residence Location	228
5.3	Random Variables and Probability Distributions	241
5.3.1	Random Variables	241
5.3.2	Bernoulli and Uniform Distributions	241
5.3.3	Binomial Distribution	246
5.3.4	Normal Distribution	249
5.3.5	Expectation and Variance	256
5.3.6	Predicting Election Outcomes with Uncertainty	260
5.4	Large Sample Theorems	264
5.4.1	The Law of Large Numbers	264
5.4.2	The Central Limit Theorem	266
5.5	Summary	271
5.6	Exercises	272
5.6.1	The Mathematics of Enigma	272
5.6.2	A Probability Model for Betting Market Election Prediction	274

6	UNCERTAINTY	276
6.1	Estimation	276
6.1.1	Unbiasedness and Consistency	277
6.1.2	Standard Error	285
6.1.3	Confidence Intervals	290
6.1.4	Margin of Error and Sample Size Calculation in Polls	296
6.1.5	Analysis of Randomized Controlled Trials	301
6.1.6	Analysis Based on Student's <i>t</i> -Distribution	304
6.2	Hypothesis Testing	307
6.2.1	Tea-Tasting Experiment	307
6.2.2	The General Framework	313
6.2.3	One-Sample Tests	316
6.2.4	Two-Sample Tests	322
6.2.5	Pitfalls of Hypothesis Testing	327
6.2.6	Power Analysis	329
6.3	Linear Regression Model with Uncertainty	336
6.3.1	Linear Regression as a Generative Model	337
6.3.2	Unbiasedness of Estimated Coefficients	343
6.3.3	Standard Errors of Estimated Coefficients	345
6.3.4	Inference about Coefficients	348
6.3.5	Inference about Predictions	350
6.4	Summary	356
6.5	Exercises	357
6.5.1	Sex Ratio and the Price of Agricultural Crops in China	357
6.5.2	Filedrawer and Publication Bias in Academic Research	359
6.5.3	The 1932 German Election in the Weimar Republic	361
7	DISCOVERY	364
7.1	Network Data	364
7.1.1	Marriage Network in Renaissance Florence	365
7.1.2	Undirected Graph and Centrality Measures	367
7.1.3	Twitter Following Network	375
7.1.4	Directed Graph and Centrality	376
7.2	Spatial Data	386
7.2.1	The 1854 Cholera Outbreak in London	386
7.2.2	Spatial Data in Stata	389
7.2.3	United States Presidential Elections	393
7.2.4	Expansion of Walmart	395
7.2.5	Animation in Stata	397
7.3	Textual Data	400
7.3.1	The Disputed Authorship of <i>The Federalist Papers</i>	400
7.3.2	Topic Discovery	404
7.3.3	Document–Term Matrix and Clusters	411
7.3.4	Authorship Prediction	413
7.3.5	Cross Validation	417

7.4	Summary	420
7.5	Exercises	420
7.5.1	International Trade Network	420
7.5.2	Mapping US Presidential Election Results over Time	422
7.5.3	Analyzing the Preambles of Constitutions	424
8	NEXT	429
	General Index	433
	Stata Index	439
	Stata Command Abbreviation List	443