

# CONTENTS

---

Acknowledgments	xi
Preface	xiii
Acronyms	xv
<b>1 Basic Tools</b>	<b>1</b>
1.1 Goals of inference	1
1.1.1 Population or process?	1
1.1.2 Probability samples	2
1.1.3 Sampling weights	3
1.1.4 Design effects.	5
1.2 An introduction to the data	6
1.2.1 Real surveys	6
1.2.2 Populations	8
1.3 Obtaining the software	9
1.3.1 Obtaining R	9
1.3.2 Obtaining the survey package	10
1.4 Using R	10
	<b>v</b>

1.4.1	Reading plain text data	10
1.4.2	Reading data from other packages	12
1.4.3	Simple computations	12
	Exercises	13
<b>2</b>	<b>Simple and Stratified sampling</b>	<b>17</b>
2.1	Analysing simple random samples	17
2.1.1	Confidence intervals	19
2.1.2	Describing the sample to R	19
2.2	Stratified sampling	21
2.3	Replicate weights	23
2.3.1	Specifying replicate weights to R	24
2.3.2	Creating replicate weights in R	25
2.4	Other population summaries	27
2.4.1	Quantiles	27
2.4.2	Contingency tables	28
2.5	Estimates in subpopulations	31
2.6	Design of stratified samples	34
	Exercises	35
<b>3</b>	<b>Cluster sampling</b>	<b>39</b>
3.1	Introduction	39
3.1.1	Why clusters: the NHANES II design	39
3.1.2	Single-stage and multistage designs	41
3.2	Describing multistage designs to R	42
3.2.1	Strata with only one PSU	43
3.2.2	How good is the single-stage approximation?	44
3.2.3	Replicate weights for multistage samples	44
3.3	Sampling by size	46
3.3.1	Loss of information from sampling clusters	49
3.4	Repeated measurements	51
	Exercises	54
<b>4</b>	<b>Graphics</b>	<b>57</b>
4.1	Why is survey data different?	57
4.2	Plotting a table	58
4.3	One continuous variable	62
4.3.1	Graphs based on the distribution function	62

4.3.2	Graphs based on the density	65
4.4	Two continuous variables	67
4.4.1	Scatterplots	67
4.4.2	Aggregation and smoothing	70
4.4.3	Scatterplot smoothers	71
4.5	Conditioning plots	72
4.6	Maps	75
4.6.1	Design and estimation issues	75
4.6.2	Drawing maps in R	76
	Exercises	79
<b>5</b>	<b>Ratios and linear regression</b>	<b>83</b>
5.1	Ratio estimation	84
5.1.1	Estimating ratios	84
5.1.2	Ratios for subpopulation estimates	85
5.1.3	Ratio estimators of totals	85
5.2	Linear regression	90
5.2.1	The least-squares slope as an estimated population summary	90
5.2.2	Regression estimation of population totals	92
5.2.3	Confounding and other criteria for model choice	97
5.2.4	Linear models in the <code>survey</code> package	98
5.3	Is weighting needed in regression models?	104
	Exercises	105
<b>6</b>	<b>Categorical data regression</b>	<b>109</b>
6.1	Logistic regression	110
6.1.1	Relative risk regression	116
6.2	Ordinal regression	117
6.2.1	Other cumulative link models	122
6.3	Loglinear models	122
6.3.1	Choosing models	124
6.3.2	Linear association models	128
	Exercises	132
<b>7</b>	<b>Post-stratification, raking and calibration</b>	<b>135</b>
7.1	Introduction	135
7.2	Post-stratification	136

7.3	Raking	139
7.4	Generalized raking, GREG estimation, and calibration	141
	7.4.1 Calibration in R	143
7.5	Basu's elephants	149
7.6	Selecting auxiliary variables for non-response	152
	7.6.1 Direct standardization	153
	7.6.2 Standard error estimation	154
	Exercises	154
<b>8</b>	<b>Two-phase sampling</b>	<b>157</b>
8.1	Multistage and multiphase sampling	157
8.2	Sampling for stratification	158
8.3	The case-control design	159
	8.3.1 ★ Simulations: efficiency of the design-based estimator	161
	8.3.2 Frequency matching	164
8.4	Sampling from existing cohorts	164
	8.4.1 Logistic regression	165
	8.4.2 Two-phase case-control designs in R	167
	8.4.3 Survival analysis	170
	8.4.4 Case-cohort designs in R	171
8.5	Using auxiliary information from phase one	174
	8.5.1 Population calibration for regression models	175
	8.5.2 Two-phase designs	178
	8.5.3 Some history of the two-phase calibration estimator	180
	Exercises	181
<b>9</b>	<b>Missing data</b>	<b>185</b>
9.1	Item non-response	185
9.2	Two-phase estimation for missing data	186
	9.2.1 Calibration for item non-response	186
	9.2.2 Models for response probability	189
	9.2.3 Effect on precision	190
	9.2.4 ★ Doubly-robust estimators	192
9.3	Imputation of missing data	193
	9.3.1 Describing multiple imputations to R	195
	9.3.2 Example: NHANES III imputations	196
	Exercises	200

<b>10</b>	<b>★ Causal inference</b>	<b>203</b>
10.1	IPTW estimators	204
10.1.1	Randomized trials and calibration	204
10.1.2	Estimated weights for IPTW	207
10.1.3	Double robustness	210
10.2	Marginal Structural Models	211
Appendix A: Analytic details		217
A.1	Asymptotics	217
A.1.1	Embedding in an infinite sequence	217
A.1.2	Asymptotic unbiasedness	218
A.1.3	Asymptotic normality and consistency	220
A.2	Variances by linearization	220
A.3	Tests in contingency tables	221
A.4	Multiple imputation	223
A.5	Calibration and estimating functions	224
A.6	Calibration in randomized trials and ANCOVA	225
Appendix B: Basic R		229
B.1	Reading data	229
B.1.1	Plain text data	229
B.2	Data manipulation	230
B.2.1	Merging	230
B.2.2	Factors	231
B.3	Randomness	231
B.4	Methods and objects	232
B.5	★ Writing functions	233
B.5.1	Repetition	234
B.5.2	Strings	235
Appendix C: Computational details		237
C.1	Linearization	237
C.1.1	Generalized linear models and expected information	238
C.2	Replicate weights	238
C.2.1	Choice of estimators	238
C.2.2	Hadamard matrices	239
C.3	Scatterplot smoothers	240
C.4	Quantiles	240

**x** CONTENTS

C.5	Bug reports and feature requests	242
Appendix D:	Database-backed design objects	243
D.1	Large data	243
D.2	Setting up database interfaces	245
D.2.1	ODBC	245
D.2.2	DBI	246
Appendix E:	Extending the <code>survey</code> package	247
E.1	A case study: negative binomial regression	247
E.2	Using a Poisson model	248
E.3	Replicate weights	249
E.4	Linearization	251
References		255
Author Index		266
Topic Index		269