

Complex Surveys

A Guide to Analysis Using R

Thomas Lumley

*University of Washington
Department of Biostatistics
Seattle, Washington*



A John Wiley & Sons, Inc., Publication

CONTENTS

Acknowledgments	xi
Preface	xiii
Acronyms	xv
1 Basic Tools	1
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	6
1.2 An introduction to the data	6
1.2.1 Real surveys	7
1.2.2 Populations	8
1.3 Obtaining the software	9
1.3.1 Obtaining R	10
1.3.2 Obtaining the survey package	10
1.4 Using R	10
	v

1.4.1	Reading plain text data	10
1.4.2	Reading data from other packages	12
1.4.3	Simple computations	13
	Exercises	14
2	Simple and Stratified sampling	17
2.1	Analyzing simple random samples	17
2.1.1	Confidence intervals	19
2.1.2	Describing the sample to R	20
2.2	Stratified sampling	21
2.3	Replicate weights	23
2.3.1	Specifying replicate weights to R	25
2.3.2	Creating replicate weights in R	25
2.4	Other population summaries	28
2.4.1	Quantiles	28
2.4.2	Contingency tables	30
2.5	Estimates in subpopulations	32
2.6	Design of stratified samples	34
	Exercises	36
3	Cluster sampling	39
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	46
3.3	Sampling by size	46
3.3.1	Loss of information from sampling clusters	50
3.4	Repeated measurements	51
	Exercises	54
4	Graphics	57
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	73
4.6.1	Design and estimation issues	73
4.6.2	Drawing maps in R	76
	Exercises	80
5	Ratios and linear regression	83
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 survey package	98
5.3	Is weighting needed in regression models?	104
	Exercises	105
6	Categorical data regression	109
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	123
6.3.1	Choosing models.	124
6.3.2	Linear association models	129
	Exercises	132
7	Post-stratification, raking and calibration	135
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	154
	7.6.2 Standard error estimation	154
	Exercises	154
8	Two-phase sampling	157
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	181
	Exercises	182
9	Missing data	185
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

10	★ Causal inference	203
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	211
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	221
A.2.1	Subpopulation inference	221
A.3	Tests in contingency tables	223
A.4	Multiple imputation	224
A.5	Calibration and influence functions	225
A.6	Calibration in randomized trials and ANCOVA	226
Appendix B: Basic R		231
B.1	Reading data	231
B.1.1	Plain text data	231
B.2	Data manipulation	232
B.2.1	Merging	232
B.2.2	Factors	233
B.3	Randomness	233
B.4	Methods and objects	234
B.5	★ Writing functions	235
B.5.1	Repetition	236
B.5.2	Strings	238
Appendix C: Computational details		239
C.1	Linearization	239
C.1.1	Generalized linear models and expected information	240
C.2	Replicate weights	240
C.2.1	Choice of estimators	240
C.2.2	Hadamard matrices	241
C.3	Scatterplot smoothers	242

x CONTENTS

C.4	Quantiles	242
C.5	Bug reports and feature requests	244
Appendix D:	Database-backed design objects	245
D.1	Large data	245
D.2	Setting up database interfaces	247
D.2.1	ODBC	247
D.2.2	DBI	248
Appendix E:	Extending the package	249
E.1	A case study: negative binomial regression	249
E.2	Using a Poisson model	250
E.3	Replicate weights	251
E.4	Linearization	253
References		257
Author Index		269
Topic Index		271