Contents

List of Figures 13

1 Introduction 15
1.1 Sources of randomness 16
1.1.1 Stochastic model 16
1.1.2 Design approach 17
1.2 Surveys 17
1.2.1 Characteristics of surveys 17
1.2.2 Sampling frame 19
1.2.3 Probability sampling 19
1.2.4 Sampling and inference 19
1.3 Specific designs 20
1.3.1 Simple random sampling 20
1.3.2 Stratified sampling 20
1.3.3 Cluster sampling 21
1.4 Outline of the book 21
1.5 Exercises 23

2 Introduction to R 25
2.1 Some R basics 26
2.1.1 Object orientation 26
2.1.2 Dataframes 26
2.1.3 Sequences, replications, conditions and loops 27
2.1.4 Matrices 30
2.1.5 Storing and reading data files 31
2.1.6 Probability distributions 32
2.1.7 Graphics 33
2.1.8 Linear regression 34
2.2 Sampling from a population 36
2.2.1 Enumeration of samples 36
2.2.2 The sample() function 37
2.3 Exercises 38

Bibliografische Informationen
http://d-nb.info/1061211851

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

3 Inclusion probabilities **41**
3.1 Introduction 42
3.2 Some notation 42
3.3 Inclusion indicator I 43
3.4 A small example 44
3.5 Inclusion probabilities π 45
3.6 Obtaining inclusion probabilities with R 46
3.7 Simple random sampling (SI) 47
3.8 Properties of the inclusion indicator 50
3.8.1 The expected value of the inclusion indicator 51
3.8.2 The variance of the inclusion indicator . . . 51
3.8.3 The covariance of the inclusion indicator . . 51
3.8.4 Properties of the covariance 52
3.8.5 Covariance matrix and sums of sums 53
3.9 Exercises 54

4 Estimation **57**
4.1 Introduction 58
4.2 Estimating functions and estimators 58
4.3 Properties of estimation functions 58
4.4 The π-estimator 59
4.4.1 Properties of the π-estimator 59
4.4.2 Expected value and variance of the π-estimator 59
4.4.3 An alternative expression of the variance . 62
4.4.4 The Yates-Grundy variance of the total . . 62
4.5 Estimation using R 64
4.5.1 A small numerical example 64
4.5.2 An empirical example: PSID 68
4.6 Generating samples with unequal inclusion probabilities 70
4.6.1 Probabilities proportional to size (PPS) . . 70
4.6.2 The Sampford-algorithm 72
4.7 Exercises 73

5 Simple sampling **75**
5.1 Introduction 76
5.2 Some general estimation functions 76
5.2.1 The π-estimator for the total 76
5.2.2 The π-estimator for the mean 76
5.2.3 The π-estimator for a proportion 77

5.3 Simple random sampling 77
5.3.1 The π-estimator for the total (SI) 78
5.3.2 The π-estimator for the mean (SI) 79
5.3.3 The π-estimator for a proportion (SI) . . . 80
5.4 Some examples using R 81
5.5 Exercises 86

6 Confidence intervals **89**
6.1 Introduction 90
6.2 Chebyshev-inequality 92
6.2.1 Derivation of the Chebyshev-inequality . . . 92
6.2.2 Application of the Chebyshev-inequality 94
6.3 Confidence intervals based on a specific sample . . 94
6.4 Some general remarks 97
6.4.1 No approximate normality 97
6.4.2 Simplified variance estimators 97
6.4.3 Effect of simplification in the simple random sampling case 98
6.4.4 Effect of simplification for the general π-estimator 99
6.4.5 Effect of simplification in stratified and clustered samples 99
6.4.6 Bootstrap 100
6.5 Exercises 101

7 Stratified sampling **103**
7.1 Introduction 104
7.2 Some notation and an example 104
7.2.1 Notation 104
7.2.2 Example: Sectors of employment as strata . 105
7.3 Estimation of the total 108
7.3.1 Simple random sampling within strata . . . 108
7.3.2 Example: simple random sampling within sectors 110
7.4 Choosing the sample size for individual strata . . . 110
7.4.1 The minimization problem 111
7.4.2 The Cauchy-Schwarz inequality 112
7.4.3 Solving the minimization problem 113
7.4.4 Example: optimal sampling size within sectors 115

7.5 Sample allocation and efficiency 116
7.5.1 Variance comparisons based on the variance decomposition 117
7.5.2 No stratification versus proportional allocation 118
7.5.3 Proportional allocation versus optimal allocation 118
7.5.4 No stratification versus optimal allocation . 119
7.5.5 An efficiency comparison with R 120
7.6 Exercises 123

8 Cluster sampling **125**
8.1 Introduction 126
8.2 Notation 126
8.2.1 Clustering the population 126
8.2.2 Artificially clustering the PSID sample . . . 127
8.2.3 Sampling clusters 127
8.2.4 Inclusion probabilities 128
8.3 Estimating the population total 129
8.3.1 The π-estimator of the population total . . 129
8.3.2 Variance of the π-estimator of the population total 130
8.4 Simple random sampling of clusters (SIC) 131
8.4.1 The π-estimator of the population total . . 131
8.4.2 The π-estimator in the PSID example . . . 131
8.4.3 Variance of the π-estimator of the population total 132
8.4.4 Variance of the mean estimator in the PSID example 133
8.5 Exercises 135

9 Auxiliary variables **137**
9.1 Introduction 138
9.2 The ratio estimator 139
9.2.1 Example of the ratio estimator using PSID data 140
9.2.2 Taylor series expansion 142
9.2.3 The approximate variance of the ratio estimator 145
9.2.4 Estimating the approximate variance of the ratio estimator using PSID data 146

9.2.5 Comparison of the ratio estimator with the simple π-estimator 148
9.2.6 The ratio estimator in the regression context 149
9.2.7 The linear regression model under specific heteroskedasticity assumption 151
9.3 The difference estimator 152
9.3.1 The difference estimator using regression notation 152
9.3.2 Properties of the difference estimator 153
9.3.3 The difference estimator of average wage using the PSID data 154
9.4 Exercises 158

10 Regression **161**
10.1 Introduction 162
10.1.1 Regression without intercept 162
10.1.2 Regression with intercept 165
10.1.3 Multivariate linear regression with intercept 167
10.2 Variance of the parameter estimators 169
10.2.1 Linear approximation of the π-estimator 169
10.2.2 The variance of the linear approximation of the π-estimator 171
10.2.3 Simple regression through the origin 175
10.2.4 Simple regression with intercept 176
10.2.5 Simple regression with intercept and simple random sampling 177
10.2.6 Wage regression with PSID data and simple random sampling 177
10.3 Exercises 180

Indices **183**
Functions Index 183
Subject Index 185